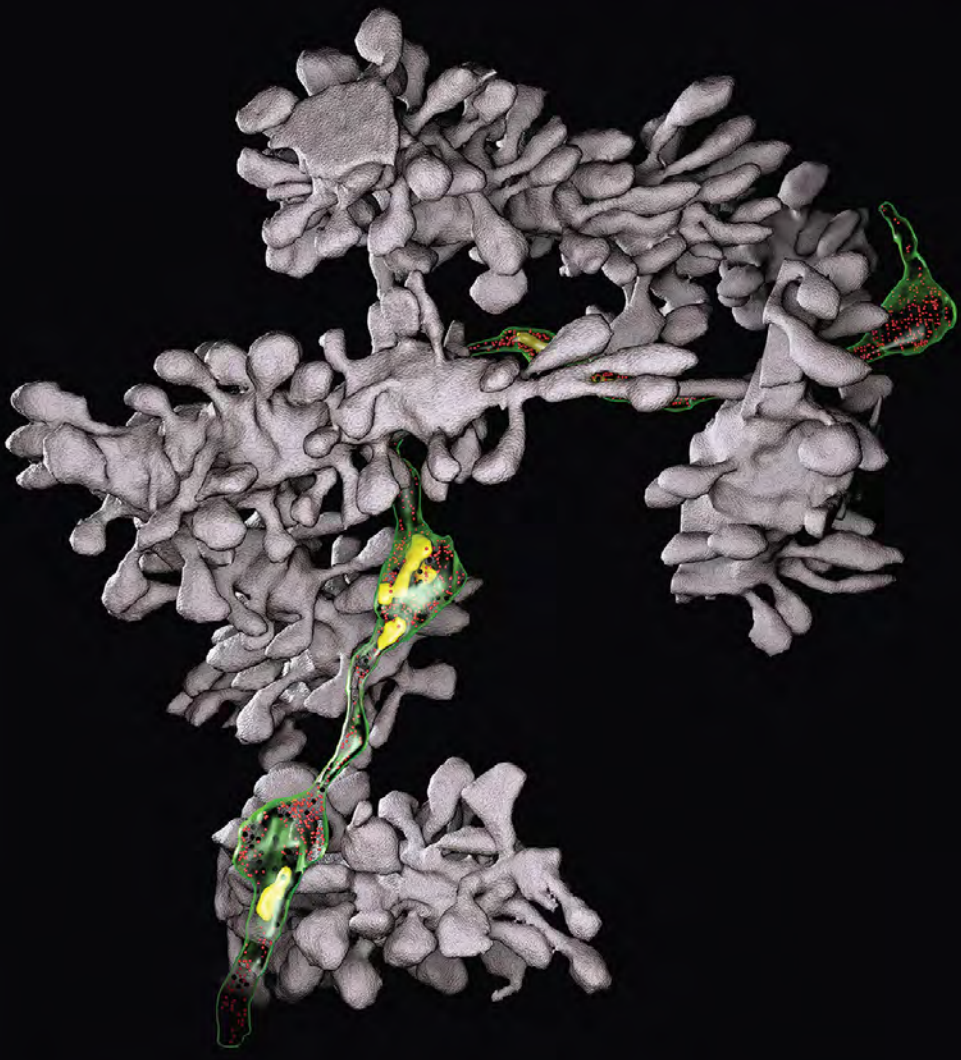
Principles of Neural Design



Peter Sterling and Simon Laughlin

Principles of Neural Design

Principles

Compute with chemistry

Compute directly with analog primitives

Combine analog and pulsatile processing

Sparsify

Send only what is needed

Send at the lowest acceptable rate

Minimize wire

Make neural components irreducibly small

Complicate

Adapt, match, learn, and forget

Principles of Neural Design

Peter Sterling and Simon Laughlin

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Stone Sans and Stone Serif by Toppan Best-set Premedia Limited.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data
Sterling, Peter (Professor of neuroscience), author.
Principles of neural design / Peter Sterling and Simon Laughlin.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-262-02870-7 (hardcover : alk. paper)
I. Laughlin, Simon, author. II. Title.
[DNLM: 1. Brain—physiology. 2. Learning. 3. Neural Pathways. WL 300]
QP376
612.8′2—dc23

2014031498

10 9 8 7 6 5 4 3 2 1

For Sally Zigmond and Barbara Laughlin

Contents

Preface

Neuroscience abounds with stories of intellectual and technical daring. Every peak has its Norgay and Hillary, and we had imagined telling some favorite stories of heroic feats, possibly set off in little boxes. Yet, this has been well done by others in various compendia and reminiscences (Strausfield, 2012; Glickstein, 2014; Kandel, 2006; Koch, 2012). Our main goal is to evince some principles of design and note some insights that follow. Stories deviating from this intention would have lengthened the book and distracted from our message, so we have resisted the natural temptation to memoir-ize.

Existing compendia tend to credit various discoveries to particular individuals. This belongs to the storytelling. What interest would there be to the Trojan Wars without Odysseus and Agamemnon? On the other hand, dropping a name here and there distorts the history of the discovery process—where one name may stand for a generation of thoughtful and imaginative investigators. Consequently, in addition to forgoing stories, we forgo dropping names—except for a very few who early enunciated the core principles. Nor do the citations document who did what first; rather they indicate where supporting evidence will be found—often a review.

Existing compendia often pause to explain the ancient origins of various terms, such as cerebellum or hippocampus. This might have been useful when most neuroscientists spoke a language based in Latin and Greek, but now with so many native speakers of Mandarin or Hindi the practice seems anachronistic, and we have dropped it. Certain terms may be unfamiliar to readers outside neuroscience, such as physicists and engineers. These are italicized at their first appearance to indicate that they are technical (*cation channel*). A reader unfamiliar with this term can learn by Googling in 210 ms that "cation channels *are pore-forming proteins that help establish and control the small voltage gradient across the plasma membrane of all living cells . . .*"

(Wikipedia). So rather than impede the story, we sometimes rely on you to Google.

Many friends and colleagues long aware of this project have wondered why it has taken so long to complete. Some have tried to encourage us to let it go, saying, "After all, it needn't be perfect . . ." To which we reply, "Don't worry, it isn't!" It's just that more time is needed to write a short book than a long one.

Acknowledgments

For reading and commenting on various chapters we are extremely grateful to: Larry Palmer, Philip Nelson, Dmitry Chklovskii, Ron Dror, Paul Glimcher, Jay Schulkin, Glenn Vinnicombe, Neil Krieger, Francisco Hernández-Heras, Gordon Fain, Sally Zigmond, Alan Pearlman, Brian Wandell, and Dale Purves.

For many years of fruitful exchange PS thanks colleagues at the University of Pennsylvania: Vijay Balasubramanian, Kwabena Boahen, Robert Smith, Michael Freed, Noga Vardi, Jonathan Demb, Bart Borghuis, Janos Perge, Diego Contreras, Joshua Gold, David Brainard, Yoshihiko Tsukamoto, Minghong Ma, Amita Sehgal, Jonathan Raper, and Irwin Levitan. SL thanks colleagues at the Australian National University and the University of Cambridge. Adrian Horridge, Ben Walcott, Ian Meinertzhagen, Allan Snyder, Doekele Stavenga, Martin Wilson, Srini (MV) Srinivasan, David Blest, Peter Lillywhite, Roger Hardie, Joe Howard, Barbara Blakeslee, Daniel Osorio, Rob de Ruyter van Steveninck, Matti Weckström, John Anderson, Brian Burton, David O'Carroll, Gonzalo Garcia de Polavieja, Peter Neri, David Attwell, Bob Levin, Aldo Faisal, John White, Holger Krapp, Jeremy Niven, Gordon Fain, and Biswa Sengupta.

For kindly answering various queries on specific topics and/or providing figures, we thank: Bertil Hille, Nigel Unwin (ion channels), Stuart Firestein, Minghong Ma, Minmin Luo (Olfaction); Wallace Thoreson, Richard Kramer, Steven DeVries, Peter Lukasiewicz, Gary Matthews, Henrique von Gersdorff, Charles Ratliff, Stan Schein, Jeffrey Diamond, Richard Masland, Heinz Wässle, Steve Massey, Dennis Dacey, Beth Peterson, Helga Kolb, David Williams, David Calkins, Rowland Taylor, David Vaney, (retina); Roger Hardie, Ian Meinertzhagen (insect visual system); Nick Strausfeld, Berthold Hedwig, Randolf Menzel, Jürgen Rybak (insect brain); Larry Swanson, Eric Bittman, Kelly Lambert (hypothalamus); Michael Farries, Ed Yeterian, Robert Wurtz, Marc Sommer, Rebecca Berman (striatum); Murray Sherman,

Al Humphreys, Alan Saul, Jose-Manuel Alonso,Ted Weyand, Larry Palmer, Dawei Dong (lateral geniculate nucleus); Indira Raman, Sacha du Lac, David Linden, David Attwell, John Simpson, Mitchell Glickstein, Angus Silver, Chris De Zeeuw (cerebellum); Tobias Moser, Paul Fuchs, James Saunders, Elizabeth Glowatzki, Ruth Anne Eatock (auditory and vestibular hair cells); Jonathan Horton, Kevan Martin, Deepak Pandya, Ed Callaway, Jon Kaas, Corrie Camalier, Roger Lemon, Margaret Wong-Riley (cerebral cortex).

For long encouragement and for skill and care with the manuscript and illustrations, we thank our editors at MIT: Robert Prior, Christopher Eyer, Katherine Almeida, and Mary Reilly, and copy editor Regina Gregory.

Introduction

A laptop computer resembles the human brain in volume and power use—but it is stupid. Deep Blue, the IBM supercomputer that crushed Grandmaster Garry Kasparov at chess, is 100,000 times larger and draws 100,000 times more power (figure I.1). Yet, despite Deep Blue's excellence at chess, it too is stupid, the electronic equivalent of an idiot savant. The computer operates at the speed of light whereas the brain is slow. So, wherein lies the brain's advantage? A short answer is that the brain employs a hybrid architecture of superior design. A longer answer is this book—whose purpose is to identify the sources of such computational efficiency.

The brain's inner workings have been studied scientifically for more than a century—initially by a few investigators with simple methods. In the last 20 years the field has exploded, with roughly 50,000 neuroscientists applying increasingly advanced methods. This outburst amounts to 1 million person-years of research—and facts have accumulated like a mountain. At the base are detailed descriptions: of neural connections and electrical responses, of functional images that correlate with mental states, and of molecules such as ion channels, receptors, G proteins, and so on. Higher up are key discoveries about mechanism: the action potential, transmitter release, synaptic excitation and inhibition. Summarizing this Everest of facts and mechanisms, there exist superb compendia (Kandel et al., 2012; Purves et al., 2012; Squire et al., 2008).

But what if one seeks a book to set out principles that explain how our brain, while being far smarter than a supercomputer, can also be far smaller and cheaper? Then the shelf is bare. One reason is that modern neuroscience has been "technique driven." Whereas in the 1960s most experiments that one might conceive were technically impossible, now with methods such as patch clamping, two-photon microscopy, and functional magnetic resonance imaging (fMRI), aided by molecular biology, the situation has reversed, and it is harder to conceive of an experiment that can*not* be done.

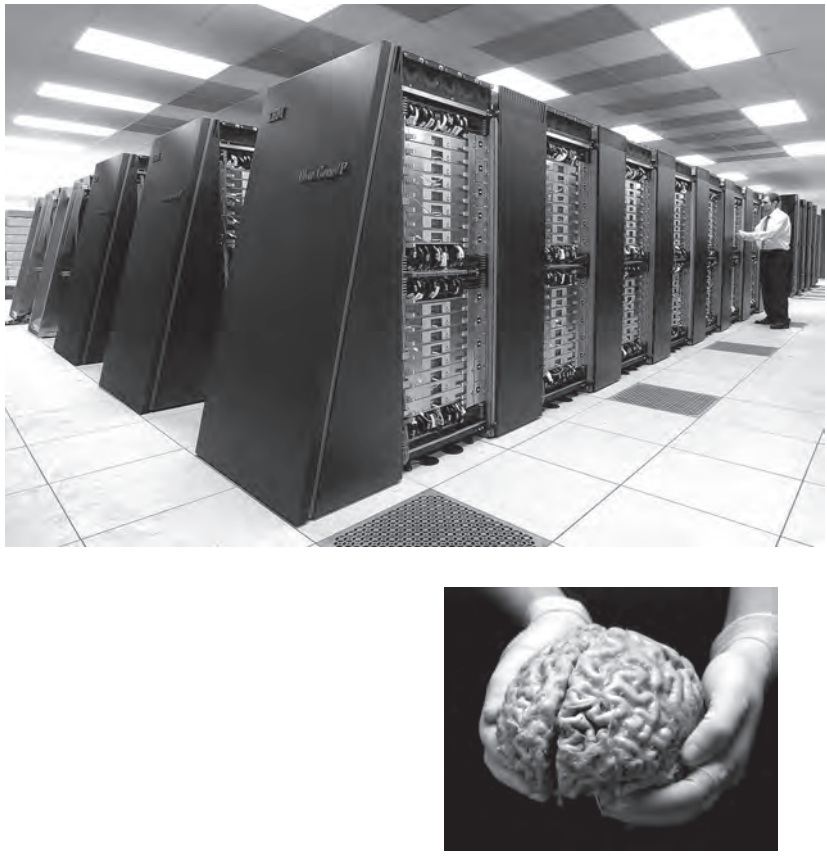


Figure I.1
How do neural circuits use space and power so efficiently? Computer: Image http://upload.wikimedia.org/wikipedia/commons/d/d3/IBM_Blue_Gene_P_supercomputer.jpg. Brain: Photo by UW-Madison, University Communications © Board of Regents of the University of Wisconsin System.

Consequently, the idea of pausing to distill principles from facts has lacked appeal. Moreover, to many who ferret out great new facts for a living, it has seemed like a waste of time.

Yet, we draw inspiration from Charles Darwin, who remarked, "My mind seems to have become a kind of machine for grinding general laws out of large collections of facts" (Darwin, 1881). Darwin, of course, is incomparable, but this is sort of how our minds work too. So we have written a small book—relative to the great compendia—intending to beat a rough path up "Data Mountain" in search of organizing principles.

Principles of engineering

The brain is a physical device that performs specific functions; therefore, its design must obey general principles of engineering. Chapter 1 identifies several that we have gleaned (not being engineers) from essays and books on mechanical and electrical design. These principles do not address specific questions about the brain, but they do set a context for ordering one's thoughts—especially helpful for a topic so potentially intimidating. For example, it helps to realize that neuroscience is really an exercise in "reverse engineering"—disassembling a device in order to understand it.

This insight points immediately to a standard set of questions that we suppose are a mantra for all "reverse engineers": *What does it do? What are its specifications? What is the environmental context?* Then there are commandments, such as *Study the interfaces* and *Complicate the design*. The latter may puzzle scientists who, in explaining phenomena, customarily strive for simplicity. But engineers focus on designing effective devices, so they have good reasons to complicate.[1] This commandment, we shall see, certainly applies to the brain.

Why a brain?

To address the engineer's first question, we consider why an animal should need a brain—what fundamental purpose does it serve and at what cost to the organism? Chapter 2 begins with a tiny bacterium, *Escherichia coli* which succeeds *without* a brain, in order to evaluate what the bacterium can do and what it cannot. Then on to a protozoan, *Paramecium caudatum*, still a single cell and brainless, but so vastly larger than *E. coli* (300,000-fold) that it requires a faster type of signaling. This prefigures long-distance signaling by neurons in multicellular organisms.

The chapter closes with the tiny nematode worm, *Caenorhabditis elegans*, which does have a brain—with exactly 302 neurons. This number is small in absolute terms, but it represents nearly one third of the creature's total cells, so it is a major investment that better turn a profit, and it does. For example, it controls a multicellular system that finds, ingests, and digests bacteria and that allows the worm to recall for several hours the locations of favorable temperatures and bacterial concentrations.

Humans naturally tend to discount the computational abilities of small organisms—which seem, well . . ., mentally deficient—nearly devoid of learning or memory. But small organisms *do* learn and remember. It's just

that their memories match their life contexts: they remember only what they need to and for just long enough. Furthermore, the mechanisms that they evolved for these computations are retained in our own neurons—so we shall see them again.

The progression bacterium → protozoan → worm is accompanied by increasing computational complexity. It is rewarded by increasing capacity to inhabit richer environments and thus to move up the food chain: protozoa eat bacteria, and worms eat protozoa. As engineering, this makes perfect sense: little beasts compute only what they must; thus they pay only for what they use. This is equally true for beasts with much larger brains discussed in chapter 3.

Why a bigger brain?

The brain of a fruit fly (*Drosophila melanogaster*) is 350-fold larger than *C. elegans*', and the brain of a human (*Homo sapiens*) is a million-fold larger than the fly's. These larger brains emerge from the same process of natural selection as the smaller ones, so we should continue to expect from them nothing superfluous—only mechanisms that are essential and pay for themselves. We should also expect that when a feature works really well, it will be retained—like the wheel, the paper clip, the aluminum beer can, and the transistor (Petroski, 1996; Arthur, 2009). We note design features that brains have conserved (with suitable elaborations) across at least 400 million years of natural selection. These features in the human brain are often described as "primitive"—reptilian—reflecting what are considered negative aspects of our nature. But, of course, any feature that has been retained for so long must be pretty effective.

This chapter identifies the core task of all brains: it is to regulate the organism's internal milieu—by responding to needs and, better still, by anticipating needs and preparing to satisfy them before they arise. The advantages of omniscience encourage omnipresence. Brains tend to become universal devices that tune all internal parameters to improve overall stability and economy. "Anticipatory regulation" replaces the more familiar "homeostatic regulation"—which is supposed to operate by waiting for each parameter to deviate from a "set point," then detecting the error and correcting it by feedback. Most physiological investigations during the 20th century were based on the homeostatic model—how kidney, gut, liver, pancreas, and so on work independently, despite Pavlov's early demonstration of the brain's role in anticipatory regulation (Pavlov, 1904). But gradually anticipatory control has been recognized.

Anticipatory regulation offers huge advantages.[2] First, it matches overall response capacity to fluctuations in demand—there should always be enough but not too much. Second, it matches capacity at each stage in the system to anticipated needs downstream, thus threading an efficient path between excess capacity (costly storage) and failure from lack of supplies. Third, it resolves potential conflict between organs by setting and shifting priorities. For example, during digestion it can route more blood to the gut and less to muscle and skin, and during exercise it can reverse this priority. This allows the organism to operate with a smaller blood volume than would otherwise be needed. Finally, it minimizes errors—which are potentially lethal and also cause cumulative damage.

Anticipatory regulation includes behavior

An organ that anticipates need and regulates the internal milieu by overarching control of physiology would be especially effective if it also regulated behavior. For example, it could reduce a body's need for physiological cooling (e.g., sweating—which costs energy and resources—sodium and water) by directing an animal to find shade. Moreover, it could evoke the memory of an unpleasant heatstroke to remind the animal to take anticipatory measures (travel at night, carry water). Such anticipatory mechanisms are driven ceaselessly by *memories* of hunger, cold, drought, or predation: *Pick the beans! Chop wood! Build a reservoir! Lock the door!*

The memories of danger and bad times that shape our behavior can be our own, but often they are stored in the brains of our parents and grandparents. We are reared with *their* nightmares—the flood, the drought, the famine, the pogrom. Before written history, which spans only 6,000 years, all lessons that would help one anticipate and thus avoid a lethal situation could be transmitted only by oral tradition—the memory of a human life span. Given that the retention of memories in small brains corresponds to their useful span, and that retention has a cost, human memory for great events should remain vivid with age whereas recent memories of lesser events should fade (chapter 14).

The most persistent dangers and opportunities, those extending far beyond a few generations, eventually become part of the neural wiring. Monkeys universally fear snakes, and so do most humans—suggesting that the response was encoded into brain structure before the lines split—on the order of 35 million years. But beyond alertness for predators, primate societies reserve their most acute observations and recall for relationships within the family and the troop. The benefit is that an individual's chances

for survival and reproduction are enhanced by the *group's* ability to anticipate and regulate. The cost is that the individual must continuously sense the social structure—in its historical context—to receive aid when needed and to avoid being killed or cast out (Cheney & Seyfarth, 2007).

Consequently, primate brains have expanded significantly in parts concerned with social recognition and planning—such as prefrontal cortex and amygdala. Humans greatly expand these areas and also those for social communication, such as for language, facial expression, and music. These regions serve both the cooperative and the competitive aspects of anticipatory regulation to an awesome degree. They account for much of our brain structure and many of our difficulties.

Flies too show anticipatory behavior—to a level consonant with their life span and environmental reach. A fly need not wait for its blood sugar to fall dangerously low, nor for its temperature to soar dangerously high, before taking action. Instead its brain expresses prewired commands: *Find fruit! In a cool spot!* Anticipatory commands are often tuned to environmental regularities that predict when and where a resource is most likely to appear—or disappear. Thus, circadian rhythms govern foraging and sleep. Seasonal rhythms, which broadly affect resource availability, govern mating and reproduction. Consequently, specific brain hormones tuned to day length send orders to prewired circuits: *Court a mate! Intimidate a competitor!*

What drives behavior?

To ensure that an organism will execute these orders, there are neural mechanisms to make it "feel bad" when a job is undone and "feel good" when it has succeeded. These are circuits whose activity humans experience, respectively, as "anxiety" and "pleasure." Of course, we cannot know what worms or flies experience—but the same neurochemicals drive similar behaviors. This is one wheel that has certainly been decorated over hundreds of millions of years, but not reinvented.

To actually accomplish a task is vastly complicated. Reconsider Deep Blue's task. Each side in chess has 16 pieces—that move one at a time, slowly (minutes), and only in two dimensions. Each piece is constrained to move only in certain ways, and some pieces repeat so that each side has only six different types of motion. This relatively simple setup generates so many possible moves that to evaluate them requires a Deep Blue.

But the organ responsible for anticipatory regulation takes continuous data from every sensory neuron in the organism—both internal and

external—plus myriad hormones and other chemicals. While doing so, it is calculating in real time—milliseconds—how to adjust every body component inside and out. It is flying the fly, finding its food, shade, and mate; it is avoiding predators and intimidating competitors—all the while tweaking every *internal* parameter to match what is about to be needed. Thus, it seems fair to say that Deep Blue is stupid even compared to a fruit fly. This defines sharply the next engineering question: what constrains the design of an effective and efficient brain?

What constrains neural design?

When Hillel was asked in the first century B.C.E. to explain the whole Torah while standing on one leg, he was ready: "That which is hateful to you, do not unto another. The rest is commentary—and now go study."

There is a one-leg answer for neural design: "As information rate rises, costs rise disproportionately." For example, to transmit more information by spikes requires a higher spike rate. Axon diameter rises linearly with spike rate, but axon volume and energy consumption rise as the diameter squared. Thus, the essence of neural design: "Send only information that is needed, and send it as slowly as possible" (chapter 3). This key injunction profoundly shapes the brain's macroscopic layout, as explained in chapter 4. We hope that readers will . . . go study.

If spikes were energetically cheap, their rates would matter less. However, a 100-mV spike requires far more current than a 1-mV response evoked by one packet of chemical transmitter. Obviously then, it is cheaper to compute with the smaller currents. This exemplifies another design principle: minimize energy per *bit* of information by computing at the finest possible level. Chapter 5 identifies this level as a change in protein folding on the scale of nanometers. Such a change can capture, store, and transmit one bit at an energetic cost that approaches the thermodynamic limit. Chapter 6 explains how proteins couple to form intracellular circuits on the scale of micrometers, and chapter 7 explains how a neuron assembles such circuits into devices on a scale of micrometers to millimeters.

It emerges that to compute most efficiently in space and energy, neural circuits should *nanofy*:

1. Make each component irreducibly small: a functional unit should be a single protein molecule (a channel), or a linear polymer of protein subunits (a microtubule), or a sandwich of monomolecular layers (a membrane).

2. Combine irreducible components: a membrane to separate charge and thus permit a voltage, a protein transporter to pump ions selectively across the membrane and actually separate the charges (charge the battery), a pore for ions to flow singly across the membrane and thus create a current, a "gate" to start and stop a current, an amplifier to enlarge the current, and an adaptive mechanism to match a current to circumstance.
3. Compute with *chemistry* wherever possible: regulate gates, amplifiers, and adaptive mechanisms by binding/unbinding small molecules that are present in sufficient numbers to obey the laws of mass action. Achieve speed with chemistry by keeping the volumes small.
4. For speed over distance compute *electrically*: convert a signal computed by chemistry to a current that charges membrane capacitance to spread passively up to a millimeter. For longer distance, regenerate the current by appropriately clustered voltage-gated channels.

Design in the visual system

Having discussed protein computing and miniaturization as general routes to efficiency, we exemplify these points in an integrated system—phototransduction (chapter 8). The engineering challenge is to capture light reflected from objects in the environment in order to extract informative patterns to guide behavior. Transduction employs a biochemical cascade with about half a dozen stages to amplify the energy of individual photons by up to a million-fold while preserving the information embodied as signal-to-noise ratio (S/N) and bandwidth. We explain why so many stages are required.

The photoreceptor signal, once encoded as a graded membrane voltage, spreads passively down the axon to the synaptic terminal. There the analogue signal is digitized as a stream of synaptic vesicles. The insect brain can directly read out this message with very high efficiency because the distance is short enough for passive signaling (chapter 9). The mammal brain can*not* directly read out this vesicle stream because the distance is too great for passive signaling. The mammal eye must transmit by action potentials, but the photoreceptor's analogue signal contains more information than action potentials can encode. Therefore, on-site retinal processing is required (chapters 10, 11).

Principles at higher levels

The principles of neural design at finer scales and lower levels also apply at larger scales and higher levels. For example, they can explain why the first

visual area (V1) in cerebral cortex enormously expands the number and diversity of neurons. And why diverse types project in parallel from V1 to other cortical areas. And why cortex uses many specific areas and arranges them in a particular way. The answers, as explained in chapter 12, are always the same: diverse circuits allow the brain to send only information that is needed and to send it at lower information rates. This holds computation to the steep part of the benefit/cost curve.

Wiring efficiency

Silicon circuits with very large-scale integration strive for optimal layout—to achieve best performance for least space, time, and energy. Neural circuits do the same and thereby produce tremendous diversity of neuronal structure at all spatial scales. For example, cerebellar output neurons (*Purkinje cells*) use a two-dimensional dendritic arbor whereas cerebral output neurons (*pyramidal cells*) use a three-dimensional arbor. Both circuits employ a layered architecture, but the large Purkinje neurons lie *above* a layer of tiny neurons whereas the large pyramidal neurons lie *below* the smaller neurons. Cerebellar cortex folds intensely on a millimeter scale whereas cerebral cortex on this scale is smooth.

Such differences originate from a ubiquitous biophysical constraint: the irreducible electrical resistance of neuronal cytoplasm. Passive signals spread spatially and temporally only as the square root of dendritic diameter ($\sqrt{d}$). This causes a second law of diminishing returns: a dendrite, to double its conduction distance or halve its conduction delay, must quadruple its volume. This prevents neural wires from being any finer and prevents local circuits from being any more voluminous. In both cases conduction delays would grow too large. The constraint on volume drives efficient layout: equal lengths of dendrite and axon and an optimum proportion of wire and synapses. Chapter 13 will explain.

Designs for learning

All organisms use new information to better anticipate the future. Thus, learning is a deep principle of biological design, and therefore of neural design. Accordingly, the brain continually updates its knowledge of every internal and external parameter—which means that learning is also a brain function. As such, neural learning is subject to the same constraints as all other neural functions. It is a design principle that must obey all the others.

To conserve space, time, and energy, new information should be stored at the site where it is processed and from whence it can be recalled without

further expense. This is the synapse. Low-level synapses relay short-term changes in input, so their memories should be short, like that of a bacterium or worm. These synapses should encode at the cheapest levels, by modifying the structure and distribution of proteins. High-level synapses encode conclusions after many stages of processing, so their memories deserve to be longer and encoded more stably, by enlarging the synapse and adding new ones.

A new synapse of diameter (d) occupies area on the postsynaptic membrane as d^2 and volume as d^3. Because adding synapses increases costs disproportionately, learning in an adult brain of fixed volume is subject to powerful space constraints. For every synapse enlarged or added, another must be shrunk or removed. Design of learning must include the principle "save only what is needed." Chapter 14 explains how this plays out in the overall design.

Design and designer

This book proposes that many aspects of the brain's design can be understood as adaptations to improve efficiency under resource constraints. Improvements to brain efficiency must certainly improve fitness. Darwin himself noted that "natural selection is continually trying to economize every part of the organization" and proposed that instincts, equivalent in modern terms to "genetically programmed neural circuits," arise by natural selection (Darwin, 1859). So our hypothesis breaks no new conceptual ground.

A famous critique of this hypothesis argues that useless features might survive pruning if they were simply unavoidable accompaniments to important features (Gould & Lewontin, 1979). This possibility is undeniable, but if examples are found for neural designs, we expect them to be rare because each failure to prune what is useless would render the brain less efficient—more like Deep Blue—whereas the brain's efficiency exceeds Deep Blue's by at least 10^5.

So what do we claim *is* new? The energy and space constraints have been known for a while, as have various principles, such as "minimize wire." The present contribution seems to lie in our gathering various rules as a concise list and in systematically exemplifying them across spatial and functional scales. When a given rule was found to apply broadly with constant explanatory power, we called it a "principle." Ten are listed as a round number. As with the Biblical Commandments and the U.S. Bill of Rights, some readers

will find too many (redundancy) and others too few. We are satisfied to simply set them out for consideration.

Some readers may object to the expression "design" because it might imply a design*er*, which might suggest creationism. But "design" can mean "*the arrangement of elements or details*," also "*a scheme that governs functioning*." These are the meanings we intend. And, of course, there *is* a designer—as noted, it is the process that biologists understand as natural selection.[3]

Limits to this effort

Our account rests on facts that are presently agreed upon. Where some point is controversial, we will so note, but we will not resort to imagined mechanisms. Our goal is not to explain how the brain *might* work, but rather to make sense of what is already known. Naturally what is "agreed upon" will shift with new data, so the story will evolve. We gladly acknowledge that this account is neither complete nor timeless.

We omit *so* much—many senses, many brain regions, many processes—and this will disappoint readers who study them. We concentrate on vision partly because it has dominated neuroscience during its log growth phase, so that is where knowledge goes deepest at all scales. Also we have personally concentrated on vision, so that is where our own knowledge is deepest. Finally, to apply principles across the full range of scales, but keep the book small, has required rigorous selection. We certainly hope that workers in other fields will find the principles useful. If some prove less than universal and need revision, well, that's science. The best we can do with Data Mountain really is just to set a few pitons up the south face.

1 What Engineers Know about Design

During the Cold War, the Soviets would occasionally capture a U.S. military aircraft invading their airspace, and with comparable frequency a defecting Soviet pilot would set down a MiG aircraft in Japan or Western Europe. These planes would be instantly swarmed by engineers—like ants to a drop of honey—with one clear goal: to "reverse engineer" the craft. This is the process of discovering how a device works by disassembling and analyzing in detail its structure and function. Reverse engineering allowed Soviet engineers to rather quickly reproduce a near perfect copy of the U.S. B-29 bomber, which they renamed the Tu-4. Reverse engineering still flourishes in military settings and increasingly in civilian industries—for example, in chip and software development where rival companies compete on the basis of innovation and design.

The task in reverse engineering is accelerated immensely by prior knowledge. Soviet engineers knew the B-29's purpose—to fly. Moreover, they knew its performance specifications: carry 10 tons of explosive at 357 mph at an altitude of 36,000 feet with a range at half-load of 3,250 miles. They also knew how various parts function: wings, rudder, engines, control devices, and so forth. So to grasp how the bomber must work was straightforward. Once the "how" of a design is captured, a deeper goal can be approached: what a reverse engineer really seeks is to understand the *why* of a design—*why* has each feature been given its particular form? And *why* are their relationships just so? This is the step that reveals principles; it is the moment of "aha!"—the thrilling reward for the long, dull period of gathering facts.

Neuroscience has fundamentally the same goal: to reverse engineer the brain (O'Connor, Huber, & Svoboda, 2009). What other reason could there be to invest 1 million person-years (so far) in describing so finely the brain's structure, chemistry, and function? But neuroscience has been somewhat handicapped by the lack of a framework for all this data. To some degree we

resemble the isolated tribe in New Guinea that in the 1940s encountered a crashed airplane and studied it without comprehending its primary function. Nevertheless, we can learn from the engineers: we should try to state the brain's primary goal and basic performance specifications. We should try to intuit a role for each part. By placing the data in some framework, we can begin to evaluate how well our device works and begin to consider the why of its design. We will make this attempt, even though it will be incomplete, and sometimes wrong.

Designing de novo

Engineers knows that they cannot create a general design for a general device—because there is no general material to embody it.[1,2] Engineers must proceed *from* the particular *to* the particular. So they start with a list of questions: Precisely what is this machine supposed to accomplish? How fast must it operate and over what dynamic range? How large can it be and how heavy? How much power can it use? What error rates can be tolerated, and which type of error is most worrisome—a false alarm or a failure to respond? The answers to these questions are design specifications.

Danger lurks in every vague expression: "very fast," "pretty small," "power-efficient," "error free." Generalities raise instant concern because one person's "very" is another's "barely." To a biologist, "brief" is a millisecond (10^{-3} s), but to an electronic engineer, "brief" is a nanosecond (10^{-9} s), and the difference is a millionfold. Engineers knows that no device can be truly instantaneous or error free—so they know to ask how high should we set the clock rate, how low should we hold the error rate, and at what costs?

The engineer realizes that every device operates in an environment and that this profoundly affects the design. A car for urban roads can be low slung with slender springs, two-wheel drive, and a transmission geared for highway speeds. But a pickup for rough rural roads needs a higher undercarriage, stouter springs, four-wheel drive, and a transmission geared for power at low speeds. The decision regarding which use is more likely (urban or rural) suffuses the whole design. Moreover an engineer always wants to quantify the particular environment to estimate the frequencies of key features and hazards.

One assumes, for example, that before building a million pickups, someone at Nissan bothered to measure the size distribution of rocks and potholes on rural roads. Then they could calculate what height of undercarriage would clear 99.99% of these obstructions and build to that standard.

Knowing the frequencies of various parameters allows rational consideration of safety factor and robustness: how much extra clearance should be allowed for the rare giant boulder; how much thicker should the springs be for the rare overload? Such considerations immediately raise the issue of expense—for a sturdier machine can always be built, but it will cost more and could be less competitive. So design and cost are inseparable.

Of course, environments change. Roads improve—and then deteriorate—so vehicle designs must take this into account. One strategy is to design a vehicle that is cheap and disposable and then bring out new models frequently. This allows adaptations to environmental changes to appear in the next model. Another strategy is to design a more expensive vehicle and invest it with intrinsically greater adaptive capacity—for example, adjustable suspension. Both designs would operate under the same basic principles; the main difference would lie in their strategies for adaptation to changes in demand. In biology the first strategy favors small animals with short lives; the second strategy, by conserving time and effort already invested, favors larger animals with longer lives. As we will see, these complementary strategies account for many differences between the brains of tiny worms, flies, and humans.

Design evolves in the context of *competition*. Most designs are not de novo but rather are based upon an already existing device. The new version tries to surpass the competition: lighter, faster, cheaper, more reliable—but each advance is generally modest. To totally scrap an older model and start fresh would cost too much, take too long, and so on. However, suppose a small part could be modified slightly to improve one factor—or simply make the model prettier? The advance might pay for itself because the device would compete better with others of the same class. A backpacker need not outrun the bear—just a companion—and the same is true for improvements in design. The revolutionary Model T Ford was not the best car ever built, but it was terrific for its time: cheaper and more reliable than its competitors.

How engineers design

An engineer takes account of the laws of physics, such as mechanics and thermodynamics. For example, a turbine works most efficiently when the pressure drop is greatest, so this is where to place the dam or hydro-tunnel. Similarly, power generation from steam is most efficient at high temperatures, which requires high pressures. But using pressure to do work is most efficient when the pressure change is infinitesimally small—which takes

infinitely long. There is no "right" answer here, but the laws of physics govern the practicality of power generation and power consumption—and thus affect many industrial designs.

Similarly, a designer is aware of unalterable physical properties. Certain particles move rapidly: photons in a vacuum (3×10^8 m in a second). In contrast, other particles move slowly: an amino acid diffusing in water (~1 μm in a millisecond)—a difference of 10^{14}. So for a communications engineer to choose photons to send a message would seem like a "no brainer"—except that actual brains rely extensively on diffusion! This point will be developed in chapters 5 and 6.

Designers pay particular attention to the interfaces where energy is transferred from one medium to another. For example, an automobile designed for a V-8 engine needs wide tires to powerfully grip the road. This is the final interface, tire-to-road, through which the engine's power is delivered; so to use narrow, lightly treaded tires would be worse than pointless—it would be lethal. More generally it is efficient to match components—for their operating capacities, robustness, reliability, and so on. Efficient designs will match the capacities of all parts so that none are too large or too small.

Matching may be achieved straightforwardly where the properties of the input are predictable, such as a power transformer driven by the line voltage, or a transistor switch in a digital circuit. But the engineer knows that the real world is more variable and allows for this in the design—by providing greater tolerances, or adjusting the matches with feedback. And to estimate what tolerances or what sorts of feedback are needed, the engineer—once again—must analyze the statistics of the environment. Chapters 8–12 will do this for vision.

What components?

Having identified a specific task, its context and constraints, a designer starts to sketch a device. The process draws on deep knowledge of the available components—their intrinsic properties (both advantageous and problematic), their functional relationships, robustness, modifiability, and cost. A mechanical engineer draws from a vast inventory of standard bolts, gears, and bearings and exploits the malleability and versatility of plastics and metal alloys to tailor new parts to particular functions. For example, Henry Ford, in designing his 1908 Model T, solved the mechanical problem of axles cracking on roads built for horses by choosing a tougher, lighter steel alloyed with vanadium.[3] An electrical engineer solves

an electronic problem by drawing on a parts catalog or else drawing on known properties and costs to design a new chip. Consequently, as models advance, the number of parts grows explosively: the Boeing 747 comprises 6 million parts.

In these respects the genome is a parts catalog—a list of DNA sequences that can be transcribed into RNA sequences ("messengers") that in turn can be translated into amino acid sequences—to create proteins that serve signaling in innumerable ways. This extensive genetic parts list is not the end, but rather a start—for there are vast opportunities for further innovation and tailoring (see chapter 5). An existing gene can be duplicated and then modified slightly, just as an engineer would hope, to produce an alternative function. For example, the protein (*opsin*) that is tuned to capture light at middle wavelengths (550 nm) has been duplicated and retuned in evolution by changing only a few amino acids out of several hundred to capture longer wavelengths (570 nm). This seemingly minor difference supports our ability to distinguish red from green.

At the next level, a single DNA sequence can be transcribed to produce shorter sequences of messenger RNA that can be spliced in alternative patterns to produce subtle but critical variants. For example, alternative splicing produces large families of receptor proteins with subtly different binding affinities—which give different time constants. Other variants desensitize at different rates. How these variations are exploited in neural design will be discussed, as will the capacity to further innovate and tailor the actual proteins by binding small ions and covalently adding small chemical groups (posttranslational modification). In short, with 20% of our genome devoted to coding neural signaling molecules, plus the additional variation allowed by duplication, alternative splicing, and posttranslational modification, the brain draws from a large inventory of adaptable parts. The versatility of these components, as explained further in chapter 5, is a major key to the brain's success.

At a still higher level biological design builds on preexisting structures and processes. Where a need arises from an animal's opportunity to exploit an abundant resource, natural selection can fashion a new organ from an old one that served a different purpose. Famously, for example, the panda's "thumb" evolved not from the first digit that humans inherited from earlier primates but from a small bone in the hand of its ancestors that served a different purpose (Gould, 1992). Thus, efficient designs can be reached via natural selection from various directions and various developmental sequences. This was recognized a century ago by a key founder of neuroscience, Santiago Ramón y Cajal (1909):

We certainly acknowledge that developmental conditions contribute to morphological features. Nevertheless, although cellular development may reveal how a particular feature assumes its mature form, it cannot clarify the utilitarian or teleological forces that led developmental mechanisms to incorporate this new anatomical feature. (edited for brevity)

Neatening up

A designer's list of needed functions might also reveal several that could be accomplished with a single, well-placed component. This has been termed "neatening up," which Ford certainly did for the Model T. For example, rather than manufacture separate cylinders and bolt them together (the standard method), he cast the engine as a solid block with holes for the cylinders. Moreover, rather than use a separate belt to drive the magneto (which provides spark to initiate combustion), he built magnets into the engine's flywheel—thereby reducing parts and weight. The sum of his efforts to improve components (vanadium steel) and design (flexible suspension) and neaten up (engine block, magneto) produced a model that was 25% lighter and delivered 25% more horsepower per pound than the competition, such as the Buick Tourabout.

Brain design reflects this process of neatening up. For example, one synapse can simultaneously serve two different pathways: fast and slow; ON and OFF. One neuron can alternately serve two different circuits: one during daylight and another during starlight (chapter 11). But this strategy must not compromise functionality.

Complicate but do not duplicate

Scientists are constantly lashed with the strop from Occam's razor. That is, we are forcefully encouraged to keep our explanatory models and theories *simple*. So the following design principle seems, not merely surprising, but actually counterintuitive: *if one design is simple and another complicated, choose the complicated* (Glegg, 1969; Pahl et al., 2007). Here is the reasoning: when one part is forced to do two jobs, it can do neither well. An example is the two-stroke engine.

The operating cycle of a four-stroke automobile engine involves four sweeps of the piston through the cylinder. One draws in the fuel, the next compresses it, the third delivers power as combustion drives the piston outward, and the fourth sweeps out the exhaust. The two-stroke engine discharges the exhaust with the same stroke that draws in fuel at the bottom

of the combustion stroke and the beginning of the compression stroke. This serves well for a lawn mower or a power saw because the simpler design avoids the need for a valve gear to separately port fuel and exhaust, giving a better ratio of power to weight. However, the four-stroke engine's more complicated design delivers much more power per liter of fuel and runs smoother and quieter. Moreover, its more efficient combustion discharges fewer pollutants (French, 1994).

There are general advantages to providing a separate part for each task. First, each part can be independently tuned for speed, sensitivity, and so forth—without compromise. Second, each part can be regulated independently. Third, more parts provide more opportunities for further refinement, innovation, and improvement—simply because there are more starting points.

Complicate! is such an important principle for neural design that it seems justified to give one example. The vertebrate retina might have used only one type of photoreceptor but instead it uses two: rod and cone. The rod photopigment is more stable but slower to regenerate, so it serves best in dim light. The cone photopigment is less stable but faster to regenerate, so it serves best in bright light. Having complicated the retina's cellular architecture with two cell types, each type has developed its own molecular refinements—specialized versions of the transduction molecules and of *their* regulatory molecules—all tuned for different light intensities. To take full advantage of these refinements, the two cell types have developed different circuits within the retina. However, just before the retinal output, there is a neatening up: rod and cone circuits merge to share a set of excitatory synapses onto a set of common output cells (ganglion cells). Further explanation is to be found in chapters 8 and 11.

There is another way to complicate a design: include several parts that appear to serve the same function. For example, a neuron may express several enzymes that produce the same product. And neighboring cells may express different versions of a similar protein; for example, axons and astrocytes (glial cells) in optic nerve both express a sodium/potassium pump but with subtly different properties. Also one region may connect to another via multiple pathways: dorsal spinocerebellar tract, ventral spinocerebellar tract, spino–reticulo–cerebellar tract, spino–olivo–cerebellar tract, and so on. These parallel features might once have been regarded as "redundant"—to increase reliability and protect against failure. But now most biologists appreciate that multiple pathways generally serve different roles and thus are not truly redundant.

Indeed, engineers try avoid redundancy, and for good reason. A part waiting for a function occupies space, adds weight, and costs extra. So, this consideration raises suspicion that the multiplicity of intracellular phosphatases, sodium/potassium pumps, spinocerebellar tracts, and so on represent complexity of the good kind.

Choosing materials

Engineers can choose from diverse materials. However, they must try to select the least costly material that is appropriate for the task. For a sailboat mast, wood was traditional, but it is heavy. Graphite can be equally stiff for less weight, but it is brittle; titanium gives the best physical performance, but it is costly. So the choice depends on whether the boat is a dinghy for weekend sailing or a 12 meter yacht for the America's Cup.

Brain design is forced to select from a far narrower set of materials. For example, biological membranes are composed of lipids and proteins. Although mechanisms for regulating the passage of substances and ions *across* the membrane in either direction are myriad (ion channels, pumps, cotransporters, antiporters, flippases, etc.), the intrinsic properties of the membrane itself are relatively constant. In particular, the membrane's specific capacitance is fixed at around 1 $\mu F\ cm^{-2}$. Neurons generate electrical signals by opening and closing channels in the membrane that allow ions to move down their electrochemical gradient and carry charge in and out of the cell.

The time constant of this electrical response is the product of the membrane resistance and capacitance, but capacitance is fixed. Therefore, to speed up an electrical process, a neuron of given surface area must reduce its membrane resistance by opening more channels, thus allowing more ions to cross the membrane. The cost of restoring these ions so as to maintain the electrochemical gradient is high—in fact, it is the human brain's major energy cost: more than 60% goes for pumping ions, making this a key constraint upon design. Thus, for the brain, as for 12 meter yachts and automobiles, speed comes at a premium—and the brain is forced to use it sparingly. This theme will recur.

Integrating across systems

Engineers look for trade-offs among individual components to improve overall performance. For example, because a truck's suspension reduces shock, investment in better springs and shock absorbers can be traded for

weight and strength in the axles. So designers evaluate the whole system to discover where investment in one component is more than compensated by savings in others. This integrated approach to design extends the principle of matching components to include cost.

An example relevant to neural design is the mobile telephone (Mackenzie, 2005). Like many animals, it is small and roams on limited power. Models compete fiercely, and success depends upon performance, beauty, and energy efficiency. One notable innovation provides the phone's "brain," its tiny internal computer, with a *turbo code* that extracts wireless signals from environmental noise. This code employs an algorithm for *belief propagation* that is computationally expensive. But the investment pays because the code eliminates noise so effectively that the efficiency of wireless communication approaches the theoretical limit defined by Shannon's equation (chapter 5).

Optimizing efficiency allows the phone to reduce the amplitude of its output signals. These consume the highest proportion of the phone's power because more energy is needed to transmit radio signals long distances in all directions than to send electrical pulses along short connections in a tiny computer. Consequently, the energy invested in the phone's brain for turbo coding produces much larger savings in the heavy work of signal transmission. By analogy, an animal's small brain saves energy by efficiently directing the activities of large, power-hungry muscles.

To understand the design of an integrated system requires teamwork. When no single person can grasp the details of every component and process, designers team up. Specialists integrate their detailed knowledge of each particular into an efficient whole. It was a team of specialists that reverse engineered the B-29. They needed to combine expertise in aerodynamics, structural engineering, materials science, fluid mechanics, control systems, and so on. Neuroscientists are reaching the same conclusion and forming teams that integrate specialized knowledge to reverse engineer their systems. Brains are integrated systems because they evolved to integrate, so how else can we understand them?

How to proceed and a caution

To consider brain design as a problem of reverse engineering, we must begin with an overview of its main tasks, establish some basic measures of performance, and then see how these relate to the investment of resources in particular mechanisms (chapters 2 and 3). Having established some basic principles, we select one important system—vision—and treat each stage of

processing in the framework of design. We present the environmental context, then the circuit structure and some "hows" of its functioning. Then for each stage we will point out some "whys" of the design and note where other neural systems use similar principles.

The design principles evinced here do not explain everything. In fact, principles cannot explain how anything works—not the B-29, not the Model T, and certainly not the brain. What, then, is their use? Design principles deepen our understanding of why things work the way they do, and armed with this deeper understanding, we can reverse engineer more efficiently. Of course, applying inappropriate or misguided principles would slow us down. Thus, principles derived theoretically, without real objects and mechanisms to illustrate them, are not yet of much use. So we attempt to balance the insights that come from principled explanations against the doubts that come from overdoing them.

2 Why an Animal Needs a Brain

Essentially only one thing in life interests us: our psychical constitution. The considerations which I have placed before you employ a scientific method in the study of these highest manifestations in the dog, man's best friend.

—Ivan Pavlov, Nobel lecture, 1904 (edited for brevity)

Brain books generally begin at the lowest levels—neurons, axons, synapses, and ion channels. But that approach ill suits our goal of reverse engineering. One cannot explain a B-29 by starting with the nuts and bolts. So we postpone the parts lists and detailed schematics to consider first a larger question: why do we *need* a brain?

One's first thought, of course, is that we need it for the magical activities and feelings it confers: art, music, love . . . consciousness. But although these features arouse intense curiosity—as Pavlov emphasized—we shall see that they are merely baroque decorations on the brain's fundamental purpose and should not be mistaken for the purpose itself. What we identify here as the brain's purpose, especially because we are seeking principles, should apply not only to humans but as well to the nematode worm, *C. elegans*, and to flies. The deep purpose of the nematode's brain of 302 neurons, the fruit fly's brain of 10^5 neurons, and our own brain of 10^{11} neurons (Azevedo et al., 2009) must be the same. By identifying the basic purpose, we set a context for later considering the "decorations." We expect that research on the mammalian cerebral cortex will not reveal many new principles—rather it will elaborate the core ones. In general, it should be easier to discover them in simpler brains.

The brain's purposes reduce to regulating the internal milieu and helping the organism to survive and reproduce. All complex behavior and mental experience—work and play, music and art, politics and prayer—are but strategies to accomplish these functions. Sharing these fundamental tasks, the brains of worms, flies, and vertebrates show significant

similarities—which will be discussed. But first, consider that a tiny bacterium, *E. coli*, and a much larger single-celled protozoan, *Paramecium,* manage these two tasks quite well without a brain. How?

Lives of the Brainless

A bacterium foraging

E. coli is miniscule (1 × 3 μm) and thrives in a nutritive soup—adrift in the intestinal digests of a large animal (figure 2.1; Alberts et al., 2008). The microbe is equipped with "taste" receptors, a battery of proteins each of which specifically binds an attractant (such as an amino acid or sugar) or a repellant. These receptor proteins cluster on the surface membrane and form signaling complexes within which they cooperate to increase sensitivity and response speed. The largest cluster is at the forward end ready to taste what comes as the bacterium ploughs through the soup. Although each cluster comprises thousands of molecules—to increase the chance of catching a taste—there are only five types of receptor molecule, each responding to a range of related compounds.

The first function of these receptors is to evaluate the *soup du jour*. Each potential nutrient (amino acid, sugar, etc.) requires its own specific transporter (*permease*) for uptake into the bacterium, plus a particular enzyme or even a whole set of enzymes to process it for energy and materials for growth. It would be uneconomical to maintain high levels of all possible transporters and processing enzymes when only a subset is needed at a given moment. Therefore, a cell refrains from synthesizing proteins for uptake and digestion until a taste receptor binds the target molecule. A receptor's binding affinity determines the concentration at which protein synthesis becomes economical.

For its default fuel *E. coli* uses glucose. But when glucose is off the menu, it can use lactose. This requires lactose detectors to call for two proteins: a permease to admit lactose and an enzyme, galactosidase, to split it. The genes coding these proteins are adjacent in *E. coli*'s DNA, comprising an *operon* (genes that work together). Their expression is blocked by a repressor protein that binds to this stretch of DNA and blocks the entry of RNA polymerase, the molecular machine that transcribes DNA to RNA (*RNA polymerase*) to initiate protein synthesis (figure 2.2). The repressor *is* the lactose detector which, upon binding allolactose (an isomer that always accompanies lactose) changes shape and releases from the DNA. This allows RNA polymerase to move off and transcribe the operon (figure 2.2; Phillips et al., 2009).

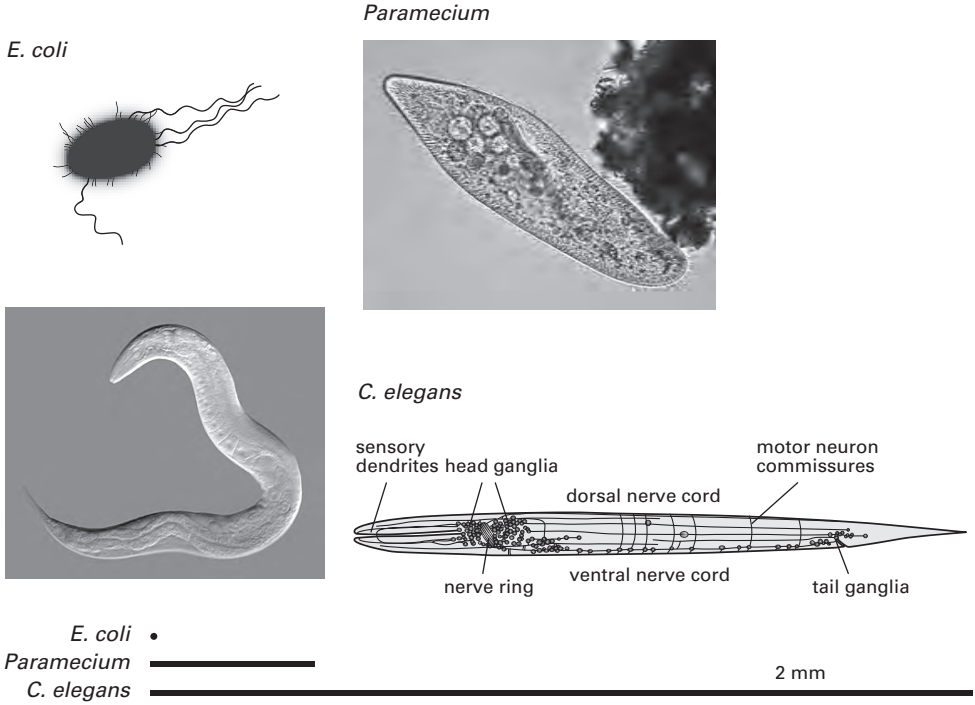


Figure 2.1

Three organisms of increasing size: bacterium, protozoan, and a nematode worm. Note the different scales: micrometers to millimeters. Body lengths are drawn to the same scale at the bottom of the diagram. *Paramecium caudatum* and *C. elegans* photos are light micrographs of live specimens. Diagram of worm indicates the positions of neurons that form the brain. Light micrographs from Wiki commons. *C. elegans* from Wikimedia Commons, CC BY-SA 3.0 / Bob Goldstein, UNC Chapel Hill, http://bio.unc.edu/people/faculty/goldstein/. *Paramecium* by Alfred Kahl, public domain, from Wikimedia Commons.

In effect, the lactose receptor *predicts* for the organism what it will need to exploit this new resource. By encoding the permease and the digestive enzyme together, one sensory signal can evoke all necessary components in the correct ratios. Thus, a given level of lactose in the soup calls for the proper amount of permease which is matched by the proper amount of galactosidase. This design principle—matching capacities within a coupled system—is a key to the organization of multicellular animals where it is called "symmorphosis" (Weibel, 2000). We see here that symmorphosis begins in the single cell.

The lac operon

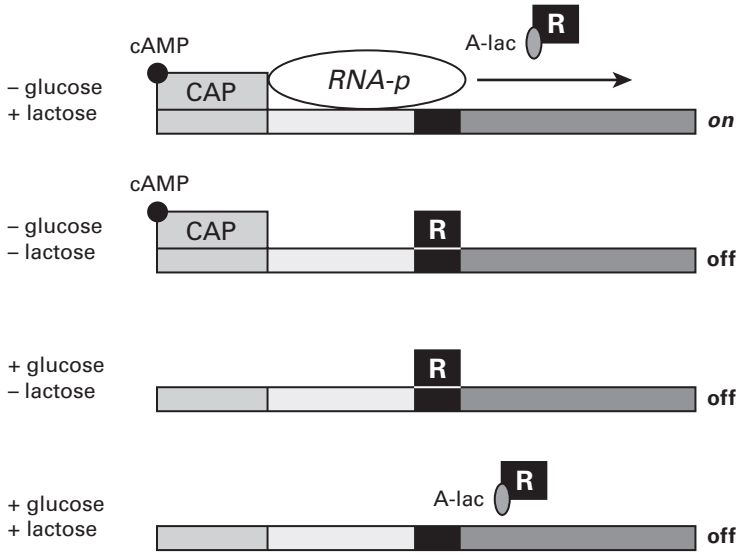


Figure 2.2

The lac operon: a molecular mechanism that discriminates between patterns of input and determines action. To transcribe the lac operon's genes, RNA polymerase (*RNA-P*) must bind to its site and move into the operon's DNA. Its movement is blocked by the repressor R, but R cannot bind and block when holding a molecule of allolactose (A-lac). To start moving, *RNA-p* must be activated by the protein CAP. This activator protein only binds to its site on the DNA when it is binding cAMP, and cAMP is eliminated in the presence of glucose. Thus, *RNA-p* only transcribes the lac operon when glucose is absent and lactose is present.

On occasions, such as when its host has eaten an ice cream, *E. coli* is presented with both lactose *and* glucose. Now the bacterium need not metabolize lactose and so need not build machinery to process it. To block this futile activity, there is a second molecular switch. RNA polymerase, to step along the DNA transcribing the lac operon, must be activated by the protein CAP, and CAP must be binding a small signaling molecule, cAMP. Biochemical pathways couple the production of cAMP to the concentration of glucose. As glucose rises, cAMP falls; this turns off the RNA polymerase (figure 2.2), and *E. coli* stops producing unneeded machinery.

Thus, a molecular control system combines information from two inputs to compute the correct conditions for processing lactose: IF lactose AND NO glucose, then GO; IF lactose AND glucose, then NO GO. The chemical

network controlling the lac operon enables a single cell to detect specific patterns of events and to mount concerted patterns of response that promote survival and reproduction. Of course, this is what a brain does on a larger scale, and in doing so it builds upon the capacities for executing logic that reside in the molecular control systems of single cells (Bray, 2009).

E. coli does more than just taste the soup and reprogram its digestive enzymes. The taste receptors also direct the cell to forage, that is, to discover and migrate to regions of higher nutrient concentration. To execute this process, *chemotaxis*, the bacterium propels itself with flagella, which are helical screws that rotate at 6,000 rpm. Their beating sends it tumbling off in random directions for brief periods, each followed by a short, straightish run. A surface receptor, sensing the instantaneous concentration of a nutrient, compares it to the past concentration—"past" lasting 1 s. If the new concentration is higher, the motor apparatus holds the forward course for a bit longer.

This search strategy (*biased random walk*, figure 2.3) resembles the party game where an object is hidden and a searcher is simply told "warmer . . . cooler . . . warmer, warmer. . ." The mechanism can sum signals from several attractants—maintaining the direction of motion for a longer time. Or, it can sum antagonistic signals (attractant + repellent) and change direction sooner. Thus, with a sensor, plus a "working memory" that controls a propeller, a microbe's wandering eventually delivers it to a greener pasture (Berg, 1993).

A microbe's memory

E. coli's working memory is simple: it is imprinted on the receptor protein by means of a negative feedback loop. The activated taste receptor causes an enzyme to attach methyl groups to the receptor complex, decreasing its sensitivity. The number of methyl groups on a receptor indicates how strongly it has been activated, and because the feedback loop is sluggish, the record stretches back into the bacterium's frantic past—1 s. The mechanism, by using the past to set receptor sensitivity, determines the bacterium's response in the present—a reasonable definition of memory. Thus, a single cell can store information cheaply through chemistry—by covalently modifying a signaling molecule.

In accomplishing the basics (preserve internal milieu and reproduce), this single cell uses mechanisms that are either optimal or highly economical: just the right number and distribution of taste receptors, just the right ratios of transporters and digestive enzymes, just the right levels of protein expression to match costs versus resources, plus the smallest signaling

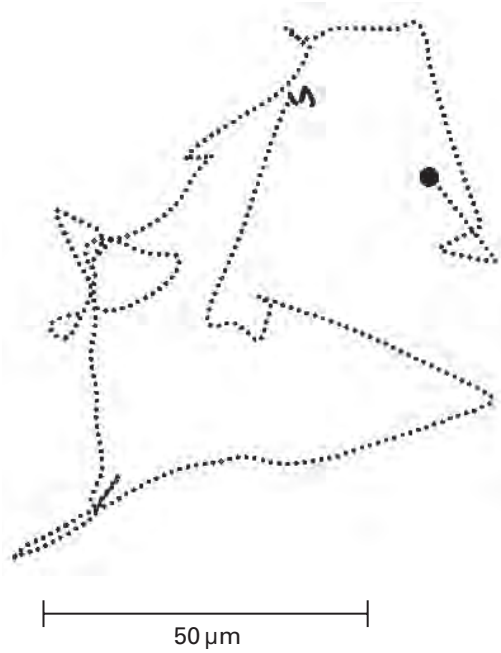


Figure 2.3
***E. coli*'s biased random walk**. By moving forward more and turning less, as the concentration of attractant increases, *E. coli* approaches the attractant's source. Tracing shows 26 runs over about 30 s with a mean speed of 21.2 µm/s. Reprinted with permission from Berg and Brown (1972). For videos of *E. coli* swimming see http://www.rowland.harvard.edu/labs/bacteria/index_movies.html/.

network for chemotaxis that could provide sufficiently robust performance. Moreover, its working memory suffices to steer the motor toward food and mates. Although a memory lasting only 1 s may not seem impressive, realize that to store a long history of lactose concentrations would be pointless—because they are themselves evanescent. Given its lifestyle, the bacterium's memory is just about as long as it *should* be.

This microbe easily lives like a Zen master—in the moment. Feed the cell, and in an hour it is gone, divided among its progeny. But once an organism becomes large enough for a brain, the Zen injunction—"Live in the moment"—itself becomes a Zen koan. A brain provides the organism with a more significant individual past and a more extended future with which to exploit it. But so equipped, staying in the moment becomes as unimaginable as the sound of one hand clapping.

Limitations to life as a microbe

Given that bacteria accomplish the basics so well, one must consider the limitations. First, their ability to respond to environmental challenge resides largely in genetic memory. A *population* thrives by reproducing rapidly and exchanging genetic material—so that when the environment changes, at least one individual in the population will contain a gene to deal with it. Thus, a population can "learn" to exploit new resources—such as potentially delicious industrial waste. However, an individual microbe, suddenly losing glucose in a lactose-rich medium, can respond only if its genome already contains the lac operon.

Second, an individual microbe cannot actively move very far. It can neither return to the site of its last meal nor deliberately transfer to a new host. This confines each species of microbe to the restricted environment for which it has specialized: a termite's gut or the skin of a human inner elbow (Grice et al., 2009)—where the bacterial genome is prepared for what it will likely encounter, and where surprises are relatively few. But this leaves a wider world unexplored and thus unexploited.

To explore would certainly increase the chances of encountering a more favorable medium—but there is a limiting challenge: size. For such a miniscule object, water is tremendously viscous. Top speed for *E. coli* is 30 μm per second, and when its effort ceases, there is insufficient inertia to carry it forward, so it abruptly stops within 0.01 nm (chapter 5; Purcell, 1977; Nelson, 2008). For a human it would be like swimming in thick molasses—agonizingly slow and energetically expensive. Consequently, to move over long distances, bacteria have evolved other methods, for example, by being sticky and hitching rides on animals.

In short, a bacterium inhabits a tiny universe—barely a few centimeters—where the critical factors are beyond its control. When transportation relies on random, energetically expensive self-propulsion or the kindness of strangers, life is precarious. A cell that could propel itself more rapidly and cheaply could forage more widely, but to overcome the effects of Brownian buffeting and high viscosity it must enlarge. And it need not get very large before motor coordination becomes an issue—as we now explain.

Protozoa: bigger and faster but still brainless

Paramecium, the familiar single-celled protozoan, measures up to 350 μm × 50 μm. Being 300,000-fold larger than *E. coli*, it is less subject to viscous forces. *Paramecium* propels itself with cilia that cover its surface and coordinate their beating to send synchronous waves from head to tail. Cruising

speed can reach roughly 1,400 μm per second, 50-fold faster than *E. coli* and with lower relative energy cost. In human terms this is the difference between exploring on foot at 4 mph and racing a car at 200 mph. Consequently, *Paramecium* can explore relatively enormous volumes of pond water and harvest bacteria by sweeping them into its "mouth." This microshark is guided by a variety of taste receptors to approach sites where bacteria proliferate, for example, clumps of rotting vegetation. It also has nociceptors to detect toxic sites, such as overripe sludge contaminated with hydrogen sulfide.

In its cluttered environment *Paramecium* inevitably encounters immovable obstacles, and to avoid the futility of continual ramming, *Paramecium* has evolved a useful response (figure 2.4; Jennings, 1904; Eckert, 1972). At the first bump it throws its cilia into reverse and backs off by a few millimeters. Then it does a quick twiddle, switches to forward, and sets off in a new direction. This avoidance response is fast—completed within a fraction of a second—and it has to be. Futile activity wastes time and energy; moreover, the immovable object might be a predator!

E. coli's chemical signaling systems could not trigger and coordinate this rapid response. Diffusion suffices for *E. coli* because the distance is short—a small intracellular messenger molecule diffuses throughout the bacterium in about 4 ms. But diffusion time increases as the distance squared (Nelson, 2008), so for a *Paramecium* that is 100-fold longer than *E. coli*, diffusion from "head" to "tail" would be 10,000-fold slower, about 40 s. Obviously, this is far too slow for receptors at the head to call "*Reverse!*" to the tail cilia. Electrical signals spread much faster: a change in membrane voltage initiated at the head reaches the tail in milliseconds.

Electrical signaling for this avoidance response requires several new components. First, a mechanoreceptor is needed to detect the bump. This involves a specialized cation channel inserted into the cell membrane. Stretch on the membrane deforms the channel, opening it to sodium ions that rapidly depolarize the membrane (<100 μs). Depolarization opens voltage-sensitive calcium channels that admit a rush of calcium ions—further depolarizing the membrane, opening still more calcium channels, and so on. This positive feedback produces a robust response that recruits calcium channels across the entire membrane (figure 2.4). They open briefly, then close and inactivate. Thus, the two components—stretch-gated sodium channel plus voltage-gated calcium channel—cooperate to deliver a synchronous pulse of calcium over the cell's entire surface.

The reason to spread the electrical signal via a calcium channel, rather than a voltage-gated sodium channel (such as used by nerve and muscle), is

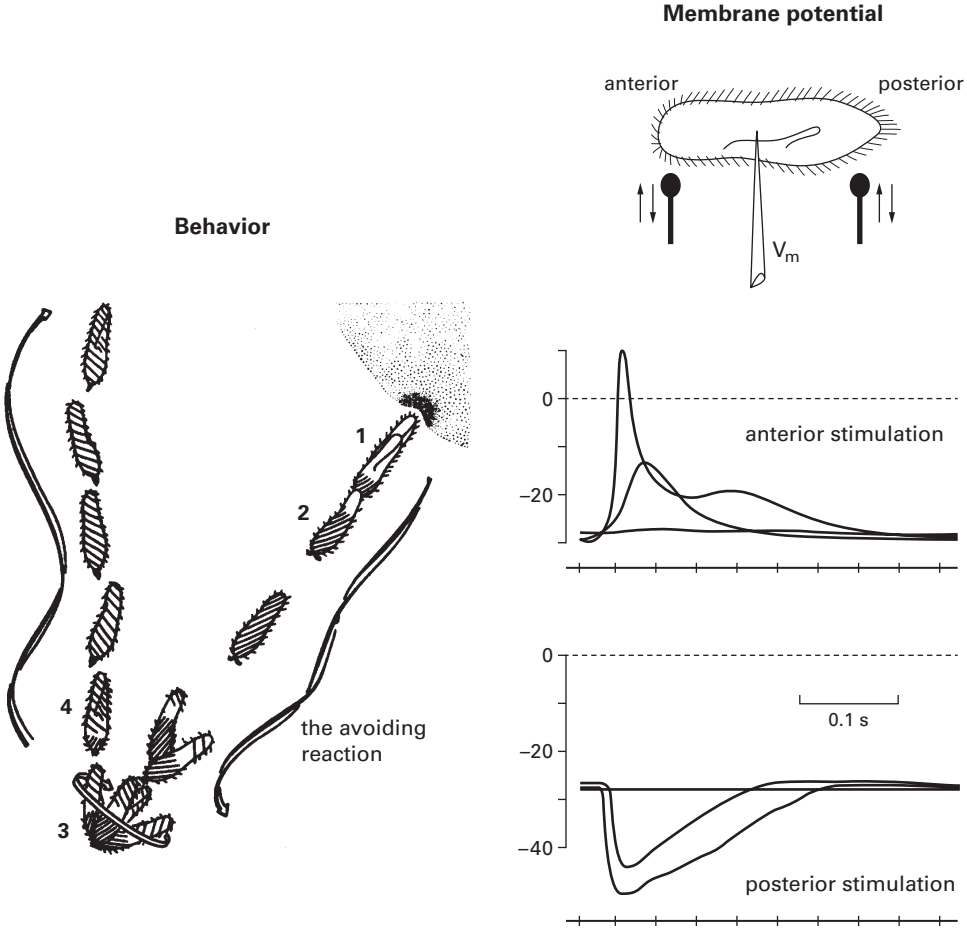


Figure 2.4

***Paramecium*'s avoidance response: behavior and electrical mechanism**. **Left**: The four stages of behavior. (1) Bumps up against immovable object, (2) backs off by reversing cilia, (3) gyrates while cilia switch from reverse to forward, and (4) sets off in a new direction. **Upper right**: Measuring electrical response to mechanical stimuli. Intracellular microelectrode records membrane potential and probes prod the membrane. **Middle right**: Membrane potential recorded following stimulation with anterior probe. A weak prod depolarizes membrane for 300 ms (lower trace). A strong prod generates a short calcium action potential followed by longer depolarization (upper trace). **Lower right**: Posterior prod hyperpolarizes. The response to the weaker prod is smaller and has a longer latency. Adapted from Eckert (1972), with permission.

that a calcium ion can also serve intracellularly as a chemical messenger. In this case the chemical message arrives synchronously at the base of all cilia, saying "*Reverse beat*," and their simultaneity adds power to the reversal. As *Paramecium* backs up, calcium pumps in the membrane vigorously reduce the calcium level, allowing patches of cilia to slip back into "forward"—explaining the indecisive twiddle. Once most of the calcium has been extruded and all cilia again beat forward, *Paramecium* heads off in a new direction (figure 2.4).

The system is polarized. The stretch channels are at the head, ensuring that the calcium pulse that reverses the cilia will also reverse the animal. The decision to reverse is structured as a simple threshold: when a bump is sharp enough, stretch channels open sufficiently to depolarize the membrane smartly enough to kick the calcium channels into their regenerative cycle. The numbers and sensitivities of stretch channels are adjusted to discriminate a truly immoveable obstacle from a yielding one. Conceivably, they are even tuned by experience via the attachment of some chemical group as with *E. coli*'s working memory.

Finally, the twiddle that sets *Paramecium* off in a new direction occurs because some patches of cilia enter forward gear before others, perhaps by the molecular noise in calcium pumps (chapter 6). Whatever the exact mechanism, the twiddle generates a random direction—which is good. Lacking distance receptors, *Paramecium* cannot predict which search direction is most likely to be best, so random behavior is optimal (Reynolds & Rhodes, 2009). Also, random motion prevents a predator from predicting *Paramecium*'s next move, thus making it harder to catch.

Where brains emerge

Despite the advantage of its fast control system for locomotion, *Paramecium*'s behavioral repertoire is limited. One impediment to richer behavior is that there is only one cell membrane and thus only one line for fast (electrical) communication. But more deeply, the cell is still so small that locomotion must be slow, and the environment remains so evanescent that richer behavior and longer memory offer no advantage. *Paramecium*'s exploitable world remains sufficiently restricted that one communication channel is plenty. Multicellularity can pay—but only when an animal becomes slightly larger and lives slightly longer in an environment where clues to food and danger persist.

The crossover—where multicellular animals arise and dominate (eat the unicellular)—occurs at a size of around 1 mm and a lifetime of days.[1] Then

cells specialize and associate to form tissues, tissues form systems, and systems cooperate to form a more versatile organism. Thus, multicellularity follows the engineering principle *complicate* (Glegg, 1969/2009a). The many tasks performed by a single cell are now divided among many specialized components. Naturally, coordination is required at each level (cell, tissue, organ, system, and organism) and across levels.

Coordination demands some mechanism with an overview that enables it to weigh alternatives, set priorities, and then exert ultimate authority to execute. Fortunately, the multicellular design that demands such integration also provides a special class of cells to accomplish it. These cells—neurons—now do what *Paramecium* could not: provide multiple fast lines for communication. In short, for a multicellular organism a brain becomes necessary, possible, and profitable.

Worm with tiny brain

The nematode worm, *C. elegans*, measures about 1×0.1 mm (figure 2.1) and in its predominant hermaphroditic form comprises exactly 959 somatic cells (Herman, 2006). It lives close to the soil surface and feeds on bacteria in rotting vegetable matter. Unlike *Paramecium*'s pond water chemicals in soil and humus are not swept away by convective currents—they move by diffusion and capillarity through a matrix, so traces persist (Félix & Braendle, 2010). The matrix and surface film provide firmer substrates for locomotion, and these allow the worm's sinuous crawl to open up whole new continents for exploitation.

The worm's enlarged territory and its locomotion through a labyrinthine matrix with persistent chemical traces warrant an upgrade. The worm improves the chemotaxis system and adds diverse sensors (of current state, opportunity, and danger), plus a larger repertoire of behavioral responses and a longer memory (de Bono & Maricq, 2005). Because bacteria-rich patches are oases where many species compete, the worm's success requires that it move smartly across a patch to efficiently find and exploit the productive regions, meet, mate, and lay eggs.

Improved foraging must be matched by more efficient systems for digestion, absorption, metabolic storage, and elimination. And as the behavioral repertoire expands, there is more need to evaluate and prioritize. For example, upon encountering a good hunting ground, how much heat or acidity should it tolerate? Upon encountering two chemical traces, which should it follow? When to search and when to graze? When to mate and when to be

stilled by "satisfaction"? In short many of the choices posed for humans by Ecclesiastes arise even for this apparently simple worm—which decides with its tiny brain.

The worm's brain may be small, but its 302 neurons plus 56 glial and support cells comprise nearly 40% of its body's entire complement. The figure in humans is close to 1%. So we first consider some behavioral advantages that justify its immense investment. Then we consider the brain's design, noting the features shared with larger brains that suggest they are governed by principles of neural design.

Locomotion

Grazers must keep on the move. The worm moves forward by bending just behind the head and then propagating the bend toward the tail. Driven by this sinusoidal wave, it threads its way through soil and rotting vegetable matter, swims through pools of fluid, and crawls across moist surfaces (e.g., decaying fruits, agar plates in laboratories). A worm travels fastest when rigid objects are regularly spaced at 0.5 mm (figure 2.5), and if this spacing is changed by just 10%, their forward speed halves. A worm seems designed to cope best with the average particle size in its preferred habitat, like a pickup truck designed for rough roads (Park et al., 2008).

But *C. elegans* is both truck and driver, continually adapting its propulsion to cope with changing conditions. When the worm goes from swimming in a pool to crawling across a wet surface, the surface tension increases viscous forces 10,000-fold, and the worm adjusts its undulations accordingly (figure 2.5). Frequency falls tenfold, wavelength shortens threefold, and more muscular power is transferred to the viscous medium. The worm continuously adjusts its drive train over a wide range of conditions, maintaining the wave's angle of attack at an efficient value, close to 45° (figure 2.5). To understand how, we must examine the integrated locomotor system: brain, muscles, body, and substrate.

A sequence of muscular contractions produces the moving wave (Sengupta & Samuel, 2009). Muscle cells on the upper side of the body contract to bow out the lower side, and when the upper cells relax, the body springs back, driven by an internal hydrostatic pressure of 0.5 atmospheres. The wave is propagated by sending two opposite bends along the body, one after the other (figure 2.6), and this sequence repeats at the frequency of undulation. When the head leads the tail, the wave moves down the worm, pushing it forward, and when the tail leads, the worm moves backward. The head also wags from side to side, and when the worm decides to

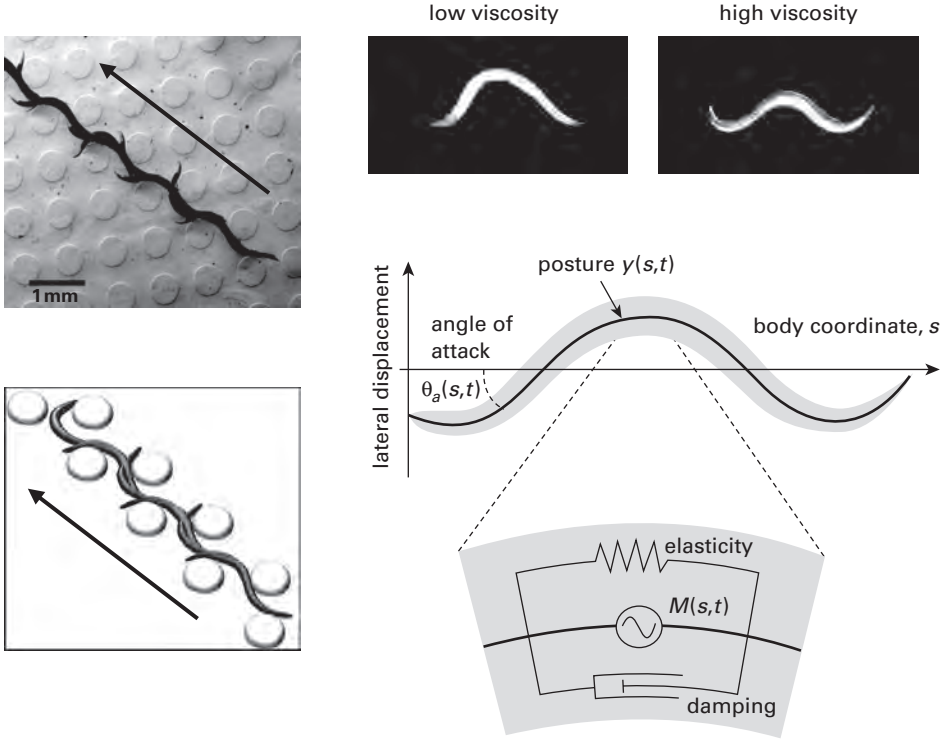


Figure 2.5

***C. elegans* locomotion matches the terrain and adapts to viscosity**. Spacing of soil particles affects forward speed, as shown when worm crawls through a regular array of agar posts of given spacing. **Upper left**: Superposition of 10 photos taken at 200-ms intervals as a worm traversed the array in which it moved forwards at maximum speed. **Lower left**: Tracings of five of the above photos, taken at 400-ms intervals, show why speed is maximum: body wavelength matches post spacing to distribute thrust efficiently. **Upper right**: The wavelength of undulation is longer in a low-viscosity medium and shorter in high viscosity. **Middle right**: Body posture is described by $y(s,t)$, the lateral displacement, y, changing with position along body, s, and time, t. The angle of attack at a given position and time, $\theta_a(s,t)$, is critical for determining thrust against the substrate. **Lower right**: The factors determining body posture and its dependence on viscosity. These vary with position along the body, s, and change with time t. In a simple biomechanical model the muscle force $M(s, t)$ interacts with body elasticity and viscous damping by the medium, to determine lateral displacement $y(s, t)$ and the angle of attack $\theta_a(s, t)$. Left reprinted with permission from Park et al. (2008). Right reprinted with permission from Fang-Yen et al. (2010).

suddenly change direction, it bends the whole body and then springs back—a good tactic for evasion and escape.

These four distinct patterns (forward, reverse, wag, and turn) are produced by 75 motoneurons that control 95 muscle cells. Each muscle cell receives input from one excitatory and one inhibitory neuron which are activated in strict alternation (Bullock, Orkand, & Grinnell, 1977). To bend the head, an excitatory motor neuron on one side of the body activates a muscle, and an inhibitory motor neuron suppresses the corresponding muscle on the other side. To propagate the bend as a wave, motor neurons activate sequentially along the body. Their output frequency determines the frequency of the undulation, and their phase determines its waveform. Excitatory motoneurons on one side activate with inhibitory motoneurons on the opposite side and alternate with excitatory motor neurons on that side (figure 2.6). Where should one look for the oscillators that produce these cycles of motor neuron activity?

Search for the oscillators

Early studies of animal locomotion were fraught with bitter argument about the origins of cyclical activity—such as stepping. Oscillations might be produced within the nervous system by local circuits (*central pattern generators*). Or they might be produced outside the nervous system by cycling sensory feedback (Marder & Bucher, 2001; Goulding, 2009). The feedback mechanism was proposed early for vertebrate stepping. One set of motor neurons excites muscles that extend the limb. This activates sensors that inhibit the extensor motor neurons and excite the flexor motor neurons, thus retracting the limb. Flexion activates sensors that inhibit the flexor motor neurons and excite the extensor neurons, and so on.

Many animals combine the two mechanisms. A central pattern generator sends cyclical commands to the motor neurons, and sensory feedback adjusts their phase, frequency, and amplitude to match changes in external load (Burrows, 1996). But the worm's circuitry seems not to use a central pattern generator. No intrinsically oscillating neurons have been found, nor does the brain's wiring diagram (see below) show the typical oscillatory circuit—a small group of neurons that send signals around a closed loop. Worms are capable of making central pattern generators—some of their cells use internal biochemical oscillators to control the rhythmical movements of ingestion, defecation, and copulation. That the worm can make central pattern generators but does not do so for locomotion suggests that it might have found a better way. Rather than relying on a pattern generator in its brain, the worm exploits its body.

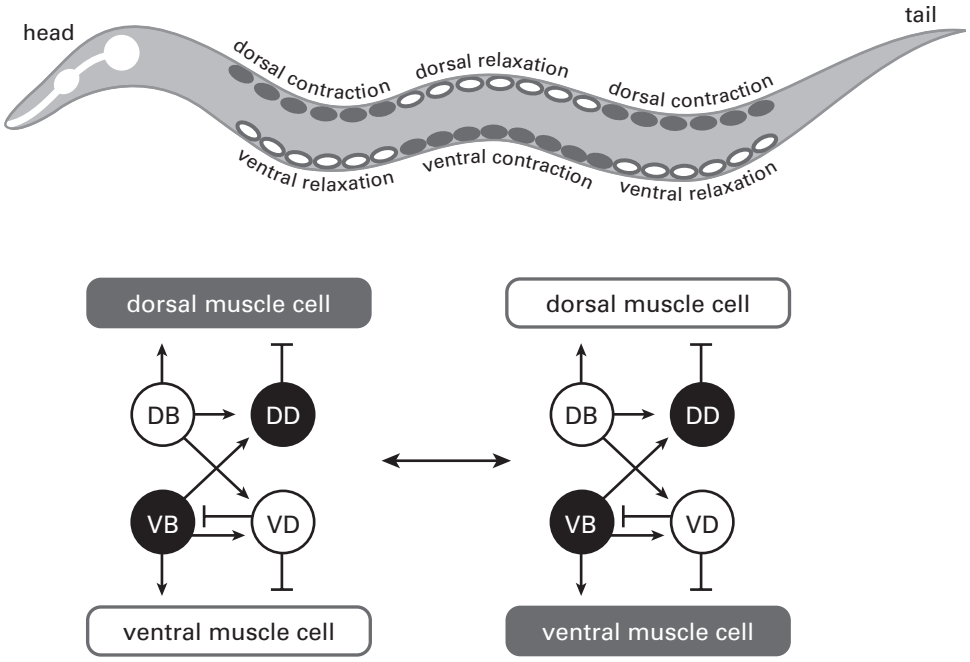


Figure 2.6

Neural circuit that bends the worm. Excitatory motor neurons (DB, VB) alternately cause dorsal and ventral muscles to contract, whereas inhibitory motor neurons (DD, VD) alternately cause them to relax. The excitatory motor neuron on one side drives the inhibitory neuron on the other side so that the body bows downward (DB and VD active), or upward (VB and DD active). This cross-inhibitory circuit repeats along the worm to promote a traveling wave. Modified from Sengupta & Samuel (2009), with permission.

Cycling with the body

The worm builds its oscillator by combining feedback with body mechanics. A burst of activity in motor neurons drives the muscles on one side. Their contraction bends the body and tensions the body's intrinsic spring—internal hydrostatic pressure. Sensors excited by these forces feed back to inhibit motor neurons, whereupon the muscles relax and the body springs back. This terminates the negative feedback, allowing the motor neurons to reactivate and start a new cycle (figure 2.6). Because the spring is damped by viscous forces (figure 2.5), the oscillation is well behaved. Also, it automatically adjusts to changes in viscous load, smoothly shifting the worm's gait to match operating conditions.

So by using its biomechanics the worm can dispense with a central pattern generator, thus freeing up brain space. Here, then, is a useful design principle for motor systems: lighten the brain's load by using the body. Engineers call this *embodied computation* (also embodied intelligence or cognition; Pfeifer & Bongard, 2006).

In the early days of robots, crawling and stepping movements were generated by an all-powerful central computer—an omniscient central pattern generator. This artificial intelligence collected sensory information and fed it into a complicated program that, by modeling the robot's mechanics, worked out the necessary commands and sent them to slavish limbs. To implement this top-down design required the robot to drag around a heavy computer, which, in turn, meant thicker limbs and stronger actuators—the result, a power-hungry behemoth. It was eventually realized that the robot and its limbs *are* a computer, an analogue computer that runs its mechanics in real time (Brooks, 1990). This analogue computer comes for free and can be set up to process information for control by, for example, being part of an oscillator. This insight inspired a new generation of small, efficient, and adroit stepping machines that blew away the behemoths. Thus, the worm exemplifies embodied computation with a neuromechanical system that matches and integrates a few basic components to meet specifications efficiently.

Neural circuits coordinate patterns of movement

Despite the contribution of body mechanics to the oscillator, neural circuits are still essential—they close the loop inside the worm. The neural circuits must be correctly configured and tuned to work with the biomechanics. Sensors must give the right feedback to motor neurons, and motor neurons must send the right signals to the right muscles with the right timing. Circuits are constructed to make this happen by ensuring that as muscles on one side of the body contract, the antagonistic muscles on the other side relax: motor neurons on one side inhibit the excitatory motor neurons for the antagonists and also excite their inhibitory motor neurons (figure 2.6). Here, then, is a circuit motif, *reciprocal inhibition* (Sherrington, 1906), that is widely employed in brains because it simply and effectively solves a common problem.

Changing direction

The brain produces motor rhythms for "forward" and "backward" using two separate sets of motor neurons. Each set has its own circuit: one works with the biomechanics to send the undulatory wave head-to-tail and the

other works to send the wave tail-to-head. This is not a popular design. Most animals use a single set of motor neurons as the final common pathway for all commands to muscle. Using two independent sets, each with a full complement of connections and synapses to muscles, seems wasteful, so why does the worm do this? We speculate that for a small brain with neuromechanical oscillation, two sets of motor neurons are cheaper than a complicated central pattern generator.

Directing action

Like *E. coli* and *Paramecium*, the worm acts to improve its chances of completing its production on the ecological stage. Equipped to move further and faster, its costs are higher and the risks greater, but so are the opportunities and rewards. So the acts must be directed appropriately (de Bono & Maricq, 2005; Lockery, 2011).

The simplest acts are aversive responses, similar in purpose and effect to *Paramecium*'s avoidance response. Tap the worm's head, and it immediately wriggles backward; tap its tail, and it wriggles forward. Two simple circuits generate this behavior (figure 2.7). Mechanosensory neurons at the front drive interneurons that activate the "backward" set of motor neurons, and mechanosensory neurons at the rear drive interneurons that activate the "forward" set. The two sets have cross connections to prevent their working in opposition.

Just as the purpose of *E. coli*'s actions is laid out in chemical circuits in a single cell, so the purpose of the worm's behavior is laid out in the connections between neurons. Naturally a brain with many neurons can generate richer behavior because, by forming connections between cells, it makes more circuits. How has the worm's brain harnessed this potential and moved its behavior beyond the simple reactions of *E. coli* and *Paramecium*?

Brain and behavior

Like the single-celled organisms the worm retreats from noxious chemicals, but its decision is more finely judged. A single sensor, the neuron labeled ASH in the brain's wiring diagram, controls this behavior by driving a "retreat" command interneuron, AVA, which shuts down the "forward" motor neurons and activates the "backward" motor neurons (figure 2.8). The sensor ASH expresses molecular receptors and detectors for a variety of potential threats, such as heavy metals, detergents, acids, or high temperature. Each input contributes to ASH activity, and when their sum suffices to trigger the command neuron, the worm backs off. Thus, a single neuron ASH serves as lawyer, jury, judge, and enforcer. It defines what constitutes

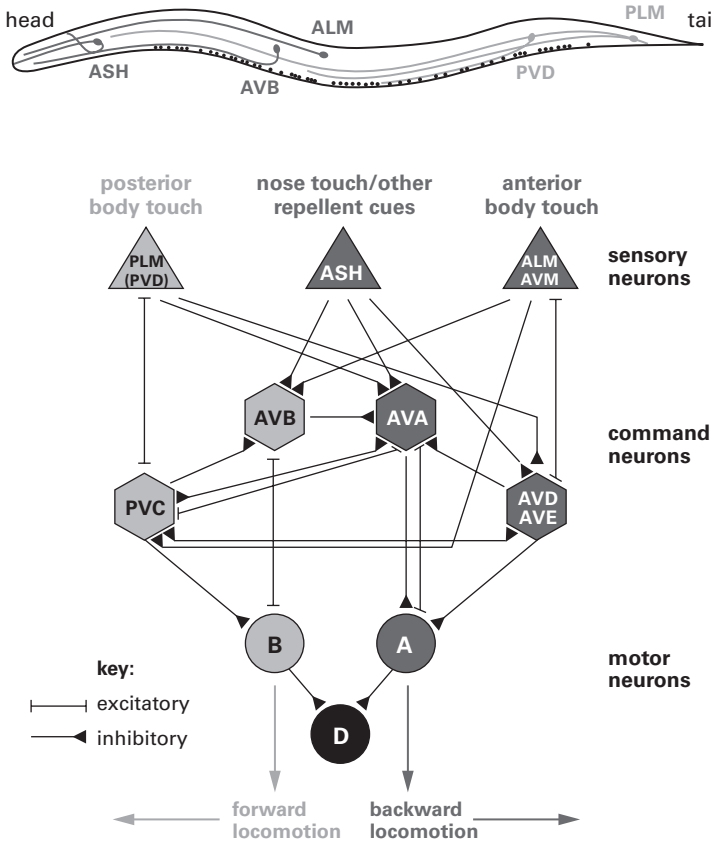


Figure 2.7
The circuit for aversive behavior. Mechanosensory neurons in the nose and in other anterior parts of the body drive command neurons for "backward" motor neurons. Mechanosensory neurons at the posterior end drive command neurons for "forward" motor neurons. These two pathways cross inhibit at the levels of command neurons and motor neurons. Adapted from de Bono & Maricq (2005), with permission.

evidence by selecting which receptors to express on its surface, collects the evidence, weighs it, judges if it warrants escape, and mandates the decision. The worm has several such sensory neurons, collecting other lines evidence for other actions.

Finding warmth, food, and mates

The worm seeks congenial places to feed, grow, and mate. *C. elegans* thrives and reproduces in a fairly narrow range of conditions: dim light,

temperature 13°–25° C, oxygen concentration 7%–14%, moderate pH, ample bacteria, and so on. To find these conditions, the worm needs a signal to warn it of imminent departure from the range—"bacteria depleted," "temperature dropping," and so forth. This search signal activates forward crawl. Foraging now for bacteria by taste and smell, the first whiff activates gradient ascent. Upon reaching favorable conditions, the worm needs a stop signal to announce "satisfaction"—what was sought is found. This signal activates a sequence of turns that places the worm in graze mode. But the worm remains vigilant. If at any moment sensors for noxious conditions are activated, they suppress the forward movement and turning, and they activate reverse.

The worm retains *E. coli*'s basic strategy for moving up or down a gradient, the biased random walk. As conditions improve, the worm turns less and runs ahead more; as conditions worsen, it turns more and runs ahead less. The mechanism is also similar: molecular receptors that drive the forward run adapt, and the decay of their output signals allows a turn. Stronger signals decay more slowly, prolonging the run.

However, with multicellularity comes an advance: ascending the gradient with paired sensors. For salt, a sensor on the right side of the head is excited by *increasing* salt, and a sensor on the left side is excited by *decreasing* salt. The right sensor excites the "forward" circuit and inhibits turning. Once the worm finds the peak concentration, this cell falls silent. If the worm moves off the peak, the left cell, excited by decreasing salt, reduces forward motion and excites turning. This search pattern, brief forward motion followed by turning, continues until the concentration starts to rise again.

The worm uses head wagging to expose both sensors to new territory and combines this action with forward thrust. This exemplifies a motor output modulated by sensing. This system also provides a case where two communication channels collect *identical* information, by sensing the same gradient, but extract different patterns and use them to drive opposite motor responses. Here is something else that a brain offers—new forms of pattern recognition that improve foraging.

Improved sensing and control are needed because *C. elegans* is to *E. coli* as a supertanker is to a rowboat. To steer a whole organism in random directions with gradual correction works on a small scale, but on a larger scale it becomes wasteful. Better for the worm to be more discriminating, to search with its *head* and inform the body once a course can be plotted. In still larger animals the sensors themselves are motorized—an insect antenna, a cat external ear, a human eye (chapter 4).

Because most worms use the same foraging circuits, they accumulate at the same sites—like undergraduates at a good café. And the subtext is similar: a place to feed is also a place to find mates. Moreover, the worms, unlike most undergraduates, are commonly hermaphroditic, so doubling their chances of a satisfying encounter. Even so, many worms enhance their attractiveness by releasing a pheromone to which intrinsically social worms are attracted. Movement toward the pheromone is controlled by a single neuron, RMG, a network hub that collects and integrates inputs from a suite of sensors and pheromones and drives the appropriate command interneurons (figure 2.8). A worm's degree of sociality is adjusted by a particular peptide released within the brain in response to changing conditions. The peptide, one member of a class of *neuromodulators*, binds to receptor proteins on specific neurons to change their activity—and hence behavior (Bargmann, 2012).

Stick and carrot

When local conditions begin to deteriorate, some definite signal is needed for the worm to move on. One such signal is the neuromodulator, octopamine. When food reserves fall, certain neurons release octopamine, which binds to receptors on particular target neurons, modifying their excitability and changing their synapses. This inhibits turning and activates the forward motor pattern. Thus, a single agent, released in response to a change in conditions, acts on specific neurons to alter circuits and switch the worm's program from "graze" to "roam."

When food is found, roaming stops and grazing resumes. This involves a second neuromodulator, dopamine. In mammalian brain, dopamine signifies (among other things) that a reward has exceeded its expected value. In worm, dopamine is released by the presence of food when, for example, mechanosensors touch particles the size of bacteria. Dopamine binds to receptors on target neurons, turning off the octopamine receptors and restoring the circuit to its previous configuration. This switches the worm from roaming to collecting its food reward. Thus, two neuromodulators, octopamine and dopamine, provide this tiny brain with a primordial stick and a primordial carrot to mediate, as they do in larger brains, "anxious" searching and "pleasurable" repetition (de Bono & Maricq, 2005).

Imminent starvation is not the only stress. Others include low oxygen, high CO_2, acidity and overcrowding. All suggest an exhausted patch—time to move on. As with humans, stress increases urgency. A comfortable worm

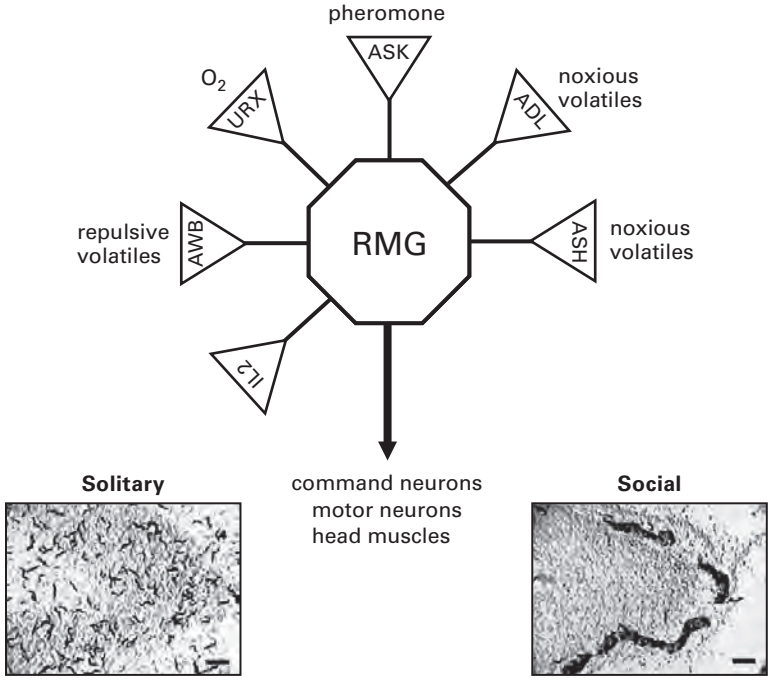


Figure 2.8

C. elegans. **Spoke and hub circuit controls solitary versus social behavior (dispersed vs. huddled). Upper**: The neuron RMG integrates social cues sensed by particular sensory neurons, ASK etc., and drives neurons that implement behavior. **Lower**: Social behavior. Solitary worms disperse and keep apart. Social worms huddle in groups. Each worm appears as a dark speck. Diagram adapted from Sokolowski (2010). Solitary and social worms from de Bono & Bargmann (1999), with permission.

moves leisurely up a promising chemical gradient, but a worm subjected to low oxygen for several hours ascends quickly. To change from stroll to rush, neuromodulators reconfigure the circuit for gradient ascent (Bargmann, 2012). For example, the sensors ADF and ASG respond to low oxygen by releasing another neuromodulator, serotonin.

Just as "carrot and stick" oversimplify human motivation, so it is for the worm. Competing for limited resources requires many factors to be weighed in deciding whether to roam or graze. A rich suite of neuromodulators allows the worm's brain of 302 neurons to evaluate contextual factors, such as nutritional status, food availability, crowding, and social signals, and then reconfigure accordingly.

Associative learning and memory

When life is good, the worm completes its life cycle (egg to egg) in 3.5 days and lives for several weeks. With a life span extending beyond the next mitotic cycle, allowing a past and a future, it now pays to recall what was good and what was bad. Far from living in the moment like *E. coli*, the worm uses its brain to associate events over time and thus draw on its experience (Ardiel & Rankin, 2010).

A worm remembers the temperature at which it was well fed and later seeks this temperature by moving up or down a thermal gradient. Finding the preferred temperature, it hangs there, searching along the isotherm. But dopamine decays promptly, so if the cupboard is bare, preference turns to aversion and the worm crawls off. Upon finding food and thus earning another shot of dopamine, the worm resets its temperature preference.

The mechanism for this learning resides within the thermal sensor that drives oriented crawling. This neuron senses changes of 0.003° C. Its response is minimal at the preferred temperature and rises on either side. The temperature for this minimum is reset by adjustments to the neuron's internal signaling; this requires protein synthesis and takes several hours. This learning process—chemical reprogramming within a single neuron—changes protein molecules but not synaptic connections.

Chemical preferences can also become associated with particular signals. For example, NaCl (salt) normally attracts worms, but when a worm has been starved in the presence of salt for only 10 minutes, it later avoids salt. A particular neuron downstream from the salt sensor releases another neuromodulator (insulin) that feeds back to an insulin receptor on the salt sensor to activate an internal signaling pathway (involving PIP3-kinase) to suppress attraction. Again, reprogramming a signaling pathway *within* a neuron allows experience to change the balance between attraction and repulsion. This mechanism also serves odorants. *C. elegans* even learns to avoid odorants from a particular pathogenic strain of bacteria that has made it sick.

These memory traces promote survival by extending the time over which an animal can identify and use patterns. The number of trials needed to establish an association is modest, five to ten repeats over 20 minutes. This makes sense in an environment where conditions are sufficiently shifty that to be useful, an association must establish rapidly and decay rapidly. In short, the worm's behavior demonstrates its reliance on information from three distinct sources: outside, inside, and the past. Its brain integrates these streams to select behaviors that, reflecting a wider context, improve the worm's vitality and reproductive success.

Some design aspects of this tiny brain

C. elegans' brain may be small, but it is not simple. To achieve its panoply of behaviors, the worm draws on a large catalog of molecular parts. This includes diverse proteins for intracellular chemical and electrical signaling, plus numerous parts for processing information at synapses. For example, signaling proteins occupy 20% of the worm's genome, and its 300+ synaptic parts amount to one third the number for mammals (Emes et al., 2008). In fact the worm brain uses many of the same components present in larger brains. Since parts are shared, one might expect some design rules to be shared as well. If some rules were not shared, that would also be instructive, for it might suggest costs and benefits of scaling up.

Here then are some design features gleaned from considering the worm's brain and what they might imply for bigger brains.

Computes as much as possible within a single cell

This feature is exemplified by the worm's *receptors* and their *sensors*. We distinguish these terms: "receptor" refers to an individual *protein molecule* that responds to a specific event—like stretch, temperature, protons, or chemical binding; "sensor" refers to an individual *neuron* that expresses one or more types of receptor. Although neuroscientists understand this difference perfectly well, for historical reasons they often use "receptor" for both the molecule and the neuron. We use different terms to reduce confusion for readers unfamiliar with the jargon, and also because they raise two design problems.

First, a single receptor molecule is subject to stochastic fluctuations, such as thermal noise. Therefore a neuron might need to improve the signal-to-noise ratio of signals conveyed by one receptor by averaging over a population of the same type. This raises the following design question: How many receptors of the same type should be expressed by each sensor? The answer will be given in chapter 6.

Second, receptors are more diverse than the sensor neurons that express them. Therefore, how should diverse receptor types be apportioned among sensors? For this problem *C. elegans* has a rule. If a set of receptors all lead to the same final action, they share a common sensor. For example, the sensor ASH collects signals from various types of receptor for noxious stimuli that require an aversive response; ASH couples its output to a single neuron that executes a command: *Scram!*

This rule explains receptor grouping generally. The worm uses more than 1,700 different types of receptor molecule for chemoreception (taste

and olfaction). This considerably exceeds the 800 or so used in mammals, but unlike mammals where each receptor type is typically assigned its own sensor, the worm provides only about 30 separate sensor neurons. Like sensors of noxious stimuli, each chemosensor sends its signal to a specific command neuron. So the signals from 1,700 different input channels (receptors for taste and olfaction) are assembled for action, not by circuits higher in the brain, but by a few dozen sensory neurons.

Computing *within* a cell economizes on neuron numbers. The worm meets all basic requirements for behavior (sensory pattern recognition, sensorimotor integration, and motor control) with small numbers of neurons. Thirty-eight sensors connect to 82 interneurons (whose processes are confined within the brain) that contact 119 motor neurons (cells whose processes leave the brain to contact the worm's 100 muscle cells). This reserves about 70 neurons for internal regulation and mating.

Yet there is a downside to performing several operations in a single cell. A cell's capacity to handle information is limited by factors such as internal noise, dynamic range, and energy supply. So a sensor that processes inputs from several types of receptor compromises its ability to handle the information from any one receptor type. A dedicated sensor can devote more receptors to its particular modality and thus improve sensitivity and signal-to-noise. This is the engineer's principle from chapter 1: to prevent one component from doing two tasks suboptimally, complicate.

Complication goes up the line. Better sensors warrant better sense organs: eyes for vision, ears for hearing, and so on. To benefit from these more accurate and discriminating sense organs, specialized sensory systems evolve in larger brains, each devoted to processing a single modality. The conclusion is obvious: as brains scale up to improve behavior, neurons specialize. Chapter 3 will suggest how and why, but now we consider a related question, how does a worm's tiny neuron manage to compute efficiently?

Uses chemistry wherever possible

Many worm neurons use internal molecular circuits to perform functions that in larger brains use a circuit of several neurons. For example, a single sensory neuron, AFD, determines the worm's temperature preference by adding new proteins to its intracellular signaling network. Another neuron, AWC^{ON}, changes a behavioral response to suit the situation. When an odorant is present *without* food, AWC^{ON}'s molecular receptors adapt and chemotaxis declines. However, when the same odorant is present *with* food, its receptors are sensitized, and chemotaxis increases (de Bono & Maricq, 2005). These competing responses are controlled by an intracellular

mechanism that switches the connection between sensor and behavioral output to reverse the control of chemotactic turning behavior (Pereira & van der Kooy, 2012).

These examples show that chemical computing by circuits *within* a neuron can manage behavior. Moreover, this can be very efficient because chemical signals are orders of magnitude cheaper than electrical signals (chapters 5 and 6). Chemical diffusion is slow for long distances, but the worm *is* small and slow. Thus, the worm's size and speed well suit its reliance on cheap chemical signaling. In addition, chemical signals can be broadcast to specific targets, which brings us to another design feature.

Uses neuromodulators to switch behaviors

Three neuromodulators were mentioned (octopamine, serotonin, and dopamine) that switch the worm's behavior in response to stress or the prospect of reward. But this is just page one from the parts catalog since the worm expresses 250 small peptides with known neuromodulatory functions. Their diversity and ubiquity is understandable because neuromodulation is so ingenious (Harris-Warrick & Marder, 1991). A neuromodulator can be broadcast widely yet still act locally and specifically, affecting only neurons that express an appropriate receptor. The receptors often couple to a protein that modulates intracellular signaling, so in effect a neuromodulator uses *trans*cellular chemistry to modulate *intra*cellular chemistry.

A neuromodulator's reach is further enhanced because its receptor diversifies into multiple subtypes that couple to different intracellular signaling networks. Consequently, one small molecule can retune and reconfigure a whole neural circuit without altering the anatomical connections. This allows every circuit to always be doing something and then to be recruited for something else as required. Thus, neuromodulators allow the brain to use components to their fullest.

Conserves synapses

The worm brain makes only about 6,400 chemical synapses. This is roughly the number that in a mammal contact a single retinal ganglion cell or a single cortical pyramidal cell. How can a worm operate with so few synapses? The neurons are far smaller and therefore can be driven by fewer synapses. But since a single synapse is unreliable, how can so few synapses signal reliably?

One answer is: *slowly*—a neuron can improve reliability by averaging over time. This can be tolerated because, compared to many animals, the worm lives in the slow lane. For example, its olfactory sensor uses a

chemical amplifier, a G protein signaling cascade that integrates for more than 20 s (chapter 5, figure 5.6). This sensor drives a synapse that integrates over several minutes. By comparison, a fly's olfactory system acts in less than 1 s. Locomotor waves descend the worm's body at 1 Hz, but an insect moves its legs faster than 10 Hz. So the worm can prosper with few synapses because it is slow. This suggests another feature: *send information as slowly as possible* because this uses fewer synapses, smaller cells, and less energy. Later chapters explain more.

Uses stereotyped components

Efficient design gives every component a definite task. Once all components are optimized for their tasks and optimally fitted together, it is efficient to repeat them across individuals. Similarly, every neuron in *C. elegans* has a definite role optimized by natural selection to meet a specified level of performance. Correspondingly every neuron is "identified," meaning that it exhibits a stereotyped morphology, chemistry, and location in every animal (White et al., 1986). The circuits are also identified, meaning that the synaptic connections are essentially identical across animals. This was established by reconstructing the entire nervous system from thousands of electron micrographs of serial sections—to produce the worm's *connectome* (figure 2.9). Identified neurons and circuits are consistently found in small brains: worm and water flea, leech and lobster, and so on.

Minimizes wiring costs

The layout of *C. elegans*' neural wiring suggests that all 302 neurons are located as near as possible to the sites where they are needed (Varshney et al., 2011). Chemical and thermal sensors concentrate at the head; tactile sensors that guide locomotion distribute along the body axis; motor neurons that propel the worm forward distribute along the rear half of the body, and motor neurons for reverse locomotion distribute along the front half (figure 2.6). But does the layout approach the optimum sought by chip designers—the unique set of placements that minimizes the total length of connections in the brain?

Designers of silicon chips have developed algorithms to optimize component placement. Their rule: place the most densely interconnected components close together and the more sparsely connected components further apart (figure 2.9). This algorithm applied to the worm's brain shows that 90% of neurons are optimally positioned (Cherniak, 1995; Chen et al., 2006; Pérez-Escudero & de Polavieja, 2007). The 10% of neurons not in their optimal position suggests competing needs. For example, neurons

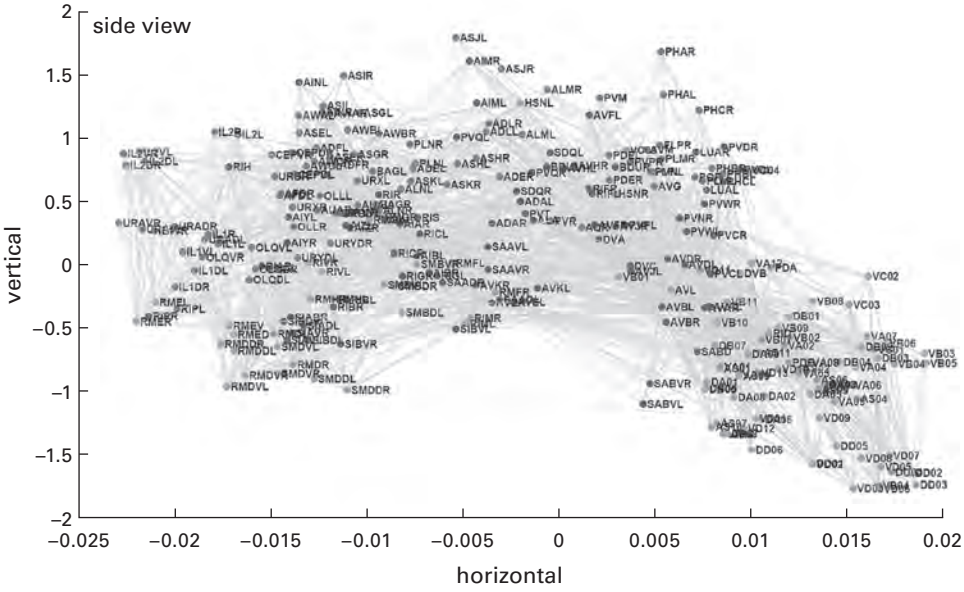


Figure 2.9

***C. elegans* connectome reconstructed from serial sections photographed in the electron microscope.** Each neuron is identified, and its synaptic connections are shown in gray. At the time of writing this is one of the most complete wiring diagrams established for any part of any brain (the other is the fly lamina cartridge, figures 9.2 and 9.3). Careful estimates suggest that this worm connectome is 93% accurate. Such are the technical difficulties of tracing neurons' thin connections that, after two decades of work on 302 neurons, 7% of connections are "missing." Reprinted with permission from Varshney et al. (2011).

that communicate most frequently with each other may be placed closer together to save energy and reduce conduction delays between them. Although layouts in larger brains certainly reflect this, conduction delay may be less relevant for *C. elegans* because the distances are so short, and the worm is so slow. "Short and slow" suggests another design feature.

Favors analogue over pulsatile

Because electrical signals in the worm travel less than a millimeter, neurons can conduct passively, as graded (analogue) changes in electrical potential. The brief, sharp, energy-intensive action potentials that dominate long-distance signaling in larger brains are unneeded, so the worm can rely solely on analogue computations, which are direct and energy efficient

(Sarpeshkar, 1998). Even its motor neurons operate in analogue mode. Over these short distances, analogue signaling transmits more information per neuron and at lower cost (chapter 7). So firmly does *C. elegans* hold to this feature that it has abandoned the gene that encodes the voltage-gated sodium channel used by larger, faster species to produce spikes.

Conclusions

Three organisms of ascending size, *E. coli*, *Paramecium*, and *C. elegans*, show why an animal needs a brain to process information on a larger scale. It is to increase opportunities for survival and reproduction in a competitive and variable environment.

The small single cell, *E. coli*, survives with surface receptors that relay information to the internal chemical signaling networks that determine metabolism, growth, reproduction, and movement. However, *E. coli* is a mere speck in space and time with most opportunities beyond its reach. A larger cell, *Paramecium*, moving more briskly travels farther, expanding opportunities, but is ultimately limited by its chemical signaling networks—diffusion and internal communication by intracellular motors are both too slow. Voltage-gated ion channels added to the cell membrane allow fast electrical signaling, but trapped in a viscous world, a single cell can only do so much.

The multicellular worm, *C. elegans*, overcomes viscosity by enlarging, and it moves faster and farther by specializing cells. This leads it to more opportunities and dangers—richer sources of information to be gathered and processed that finally need a brain. The key innovation is the neuron, a cell type specialized to collect, process, and communicate. Each neuron links its rich web of internal chemical communication to the electrical network at the surface membrane and thence to other neurons via synapses. Neuromodulators retune selected neurons to reconfigure whole circuits. Thus, a brain of only 302 neurons extends the worm's horizon by providing a behavioral repertoire that adapts to changing contexts.

The worm accomplishes the same tasks as a bacterium or protozoan—finds growth conditions and mates while avoiding unproductive or toxic sites. And it does so with similar behaviors, such as gradient ascent by biased random walk and avoidance. But with its brain *C. elegans* can cover more territory, and with its longer lifespan (weeks instead of minutes), it can adapt to nasty surprises as an *individual* rather than as a miniscule part of an adapting *population*.

Here emerges another design principle. Life span and lifestyle are related to the appearance of particular types of memory and particular decay times. Nothing should be remembered that is unlikely to enhance survival and reproduction. Nor should memories exceed the typical time constants of useful correlations—because when correlations decay, memory ceases to predict anything useful. But it *is* useful to establish the memory trace rapidly before it is outdated—and that seems to occur—few trials, closely spaced. This suggests that the longest and deepest human memories are not mere decoration but serve to shape character over a lifetime, promoting survival in our complex social fabric (chapter 14).

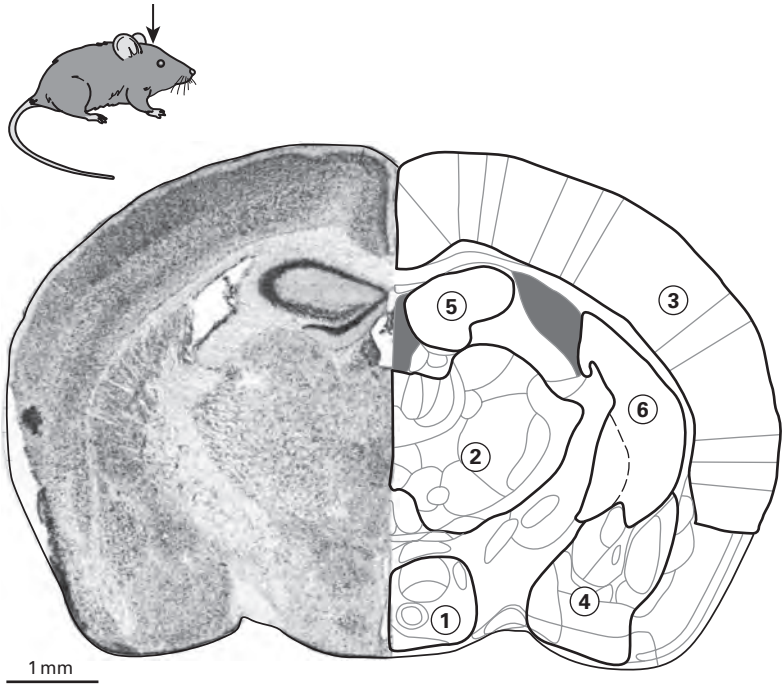
Finally, given that *C. elegans* does so well with only 302 neurons, one might look critically at an assumed truth—that it is better to have a bigger brain. So why *have* animals evolved still bigger brains?

3 Why a Bigger Brain?

This chapter will explain why, despite the worm's success with 302 neurons, brains expand. The mouse cerebral cortex contains about 10^7 neurons. This seems like a lot until you consider that the cortex of the macaque monkey, a key experimental model, is larger by 100-fold, and that human cortex is 10-fold larger still (Herculano-Houzel, 2011). Despite this huge range of scales, one feels comfortable generalizing about the "mammalian brain"—because every part identified in mouse can also be identified in macaque and human (figure 3.1; Kaas, 2005).

Consider also the fly brain. It has 500-fold fewer neurons than the mouse brain, but 500-fold more neurons than the worm brain, plus a rich structure—so warranting a slot in the "large brain" category. Insect and mammal brains share many similarities. For example, both gather their neurons into clusters and their axons into cables (*tracts*). Both employ special structures to accomplish the same broad tasks: store high-level input patterns, generate low-level output patterns, and retrieve patterns using reduced instructions. Of course, there are differences, given the differences in body design and behavior. Yet, despite half a billion years of evolutionary opportunity to diverge, brain designs in insect and mammal seem to have followed the same rules.

For designs to have persisted across this immensity of time and spatial scale implies that they are neither arbitrary nor accidental. Rather, they must have emerged as responses to some broad constraint. That is what elevates the shared responses to the status of *principles*. This chapter will identify the key constraint and indicate how it leads to three principles that govern the organization of larger brains.



1. Generate patterns for wireless signaling and appetitive behaviors.
2. "Preprocessing" to shape signals for higher processing.
3. High-level processing: assemble larger patterns, choose behaviors.
4. "Tag" high-level patterns for emotional significance.
5. Store and recall.
6. Evaluate reward predictions.

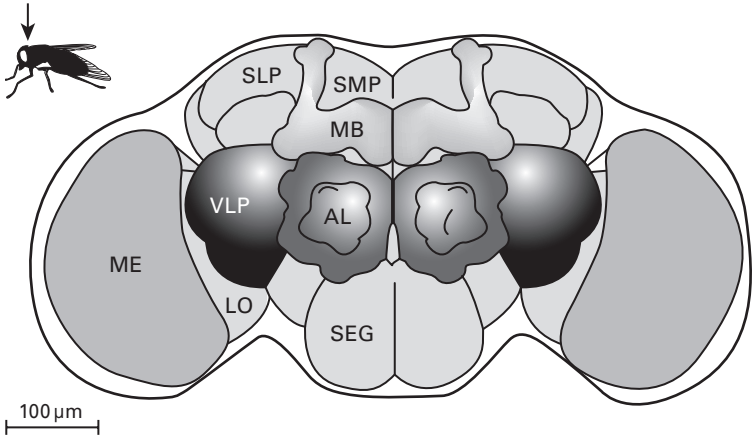


Figure 3.1
Mammalian and insect brains share many broad aspects of design. **Upper**: Cross section through mouse brain; inset indicates plane of section. **Left**: Fine dots are neurons; dark regions are neuron clusters; bright regions are myelinated tracts (chapter 4). **Right**: Numbered regions dedicated to core tasks: (1) *hypothalamus*; (2) *thalamus*; (3) *cerebral cortex*; (4) *amygdaloid complex*; (5) *hippocampus*; (6) *striatum*. Reprinted with modifications and permission from Franklin and Paxinos (1996). **Lower**: Cross section through fly brain; inset indicates plane of section. Brain is built of more than fifty clusters, each specialized for particular tasks. Depicted here are ME, medulla—detect and map local visual patterns; LO, lobula—assemble local visual patterns into larger patterns; AL, antennal lobe—preprocess olfactory signals for pattern recognition; VLP, ventrolateral protocerebrum; SLP, superior lateral protocerebrum; SMP, superior medial protocerebrum—all involved in high-level integration; MB, mushroom body—store and recall; SEG, subesophageal ganglion—integrate information for wired and wireless output to body.

A brain's core tasks

As animals emerge from the soil to a wider, less viscous world, the possibilities for foraging expand immensely. A worm explores mainly in two dimensions over an area of 0.01 m^2 whereas a honeybee typically covers an area of nearly 10^7 m^2, and a fly somewhat less. So foraging area expands by 10^9 (1 billionfold). Add the third dimension, and the volume to be explored becomes astronomical. Larger animals, such as fish, birds, and mammals, may migrate and thus forage over thousands of kilometers—thus millions of square kilometers.

Such gigantic territories contain immense resources and, of course, harbor innumerable dangers. For an animal to find the one and avoid the other requires it to rapidly gather vast amounts of information from the environment. To calibrate "vast" with one example, the eye sends the brain about 10 megabits per second, roughly the rate of an Ethernet connection (Koch et al., 2006). All sense data reach the brain in the form of tiny patterns—evanescent pieces of a dynamic jigsaw puzzle—and to be of any use, they require assembly to reveal a larger pattern. So if gathering information is to be at all rewarding, the brain must commit resources to assembling larger patterns on spatial and temporal scales that are relevant to behavior.

Yet, even a larger pattern might be useless until it is compared to a library of stored patterns where it can be identified: *edible/toxic*, *friend/foe*, or *search item not found*. Either outcome provides a basis for behavioral choice. A

match allows confident choice: eat or decline, approach or flee. A non-match suggests caution and need to gather more data. Thus, the brain requires "pattern comparators," and these must couple to mechanisms that select behaviors: *feed*, *fight*, *copulate*, *investigate*. These, in turn, couple to mechanisms for detailed motor patterns to drive muscles for moving limbs or wings.

Any given motor behavior *might* match exactly the action that was ordered: the arrow might strike the exact point at which it was aimed. But often there are errors due to environmental or neural perturbations, and these need to be identified, so that performance can progressively improve. Thus, a brain needs mechanisms to evaluate the mismatch between the orders it gave and the actual motor performance. So, in addition to sensing and processing patterns to discover "what's important out there," the brain also devotes considerable resources to sensing and processing its own motor errors, and other errors of internal "intentional" signaling in order to improve the accuracy and efficiency of the next round. This is "motor learning."

Behaviors are subject to another important class of errors. Every action has both costs and consequences. The costs are partly energetic: how much energy was spent? But also there are "opportunity costs": could the return have been greater and the risk less for some different action? Every behavior, even when perfectly executed, needs to be evaluated from this perspective: wise or foolish? repeat or not? These evaluations of *reward prediction*, like those for motor errors, are used to update stored knowledge in order to improve the next round of predictions. The nematode worm already shows this type of evaluation to some degree, but animals in the wider world allot it major neural resources.

In sum, to succeed in the wider world, an animal must exchange larger amounts of information with its external environment and also evaluate the costs and consequences of its actions. The seven core tasks that every brain must accomplish are summarized in figure 3.2. What the brain does for the external environment it also does for the internal environment which has also expanded and complexified. Moreover, the mechanisms for managing the internal and external environments need to couple closely in order to serve each other (figure 3.2).

Why the internal milieu needs a brain

To support richer external behaviors, an animal requires specialized internal tissues and organs. Some digest the bounty foraged from the outer

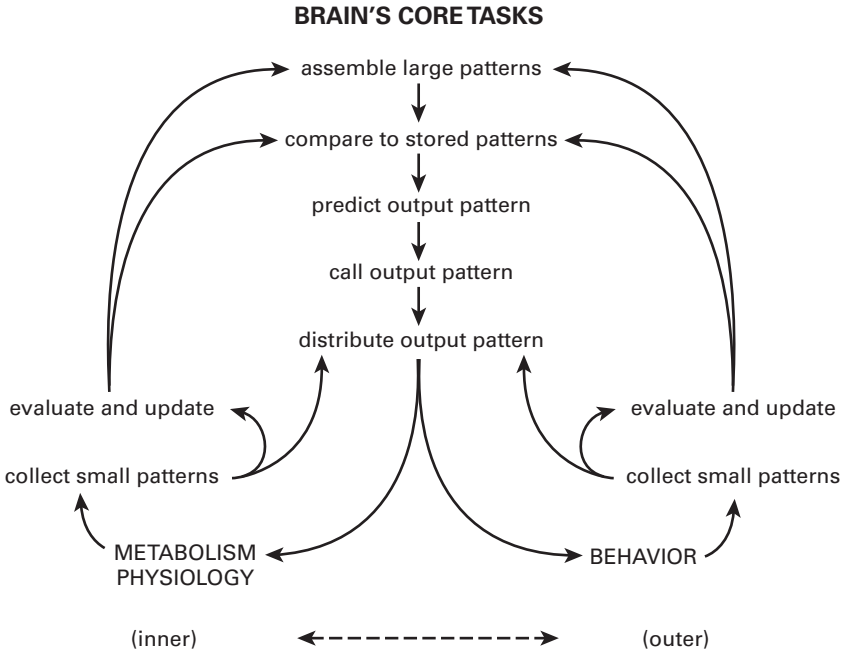


Figure 3.2
Large brains accomplish the same broad tasks. Note that inner and outer tasks couple to serve *each other* (↔).

world; others store metabolites and energy-rich compounds for release upon demand. Still others regulate ionic balance and cleanse the internal milieu, or distribute oxygen and metabolites to hungry tissues. Specialized organs of immunity protect against infectious agents and parasites. Organs couple to form systems, and systems cross-couple to optimize overall function.

The standard idea is that the internal systems more or less take care of themselves. Each parameter is supposed to have a set point, like a thermostat, from which deviations trigger feedback to correct the mismatch (*homeostasis*). Internal regulation also employs *autonomic nerves*—so termed because they are in some sense independent of voluntary control—thus, autonomous. We cannot "will" our heart to beat faster or our blood pressure to decrease. However, we can accomplish these shifts by recalling or imagining the appropriate scene. This implies the existence of neural pathways from pattern stores to pattern generators for autonomic circuits. Thus, although the autonomic nerves are generally supposed to serve

emergencies ("fight or flight"), they actually serve continuous regulation—not just for panic, but for efficiency.

Efficient regulation anticipates

In fact, all internal regulation, even the mildest sort, is far from autonomous. As the external environment presents opportunity or cause for concern, internal processes must predict what the external environment is about to deliver and must prepare particular responses that will probably be needed in support. For internal processes the goal is not to correct mismatches but to prevent them.

Such predictive regulation was demonstrated for feeding and digestion by Ivan Pavlov more than a century ago: the brain processes small patterns from the outside (sight or smell of some substance) and matches them to a stored pattern that identifies a particular food. Then the brain triggers secretions all along the digestive system to prepare for what's coming, starting in the mouth (if bread, then amylase; if fat, then lipase), then on to the stomach (if meat, then acid plus protease), the intestine (if fat, then bile), and finally the circulation (if glucose, then insulin). All of these secretions occur *before* and *during* the meal, triggered *predictively*—anticipating what will be coming down the gastrointestinal tract—thus preparing systems for absorption and uptake in order to prevent deviations that would need correction by negative feedback (Fu et al., 2011).

Modern work extends this point: as the stomach releases its contents to the next stage, it also signals the brain to prepare for the next bout of foraging. The brain responds by tuning up sensitivity of the olfactory receptors and by increasing the rate of sniffing (Julliard et al., 2007; Tong et al., 2011). Thus, the stomach warns the brain "Prepare to forage again"—well before the body has begun to deplete its reserves. Moreover, as fat reaches the small intestine, the gut can predict confidently the approach of satiety. Therefore, the gut warns the brain "cease feeding and proceed to the next activity"[1] (Fry et al., 2007).

Each "next activity" requires the brain to predict continuously, and in timely fashion, the need for a particular blood pressure. Consider the record of mean arterial pressure over 24 hours (figure 3.3). In early afternoon, as the subject attends a lecture, his brain anticipates reduced demand and allows him to doze: pressure falls. Startled awake by the jab of a pin, the brain predicts danger: pressure spikes; then, identifying a prank, the brain directs the nap to resume: pressure falls. At midnight the subject has sexual intercourse: pressure spikes, but then falls profoundly and stays low during sleep. Come morning, the brain predicting a busy day, restores the pressure.

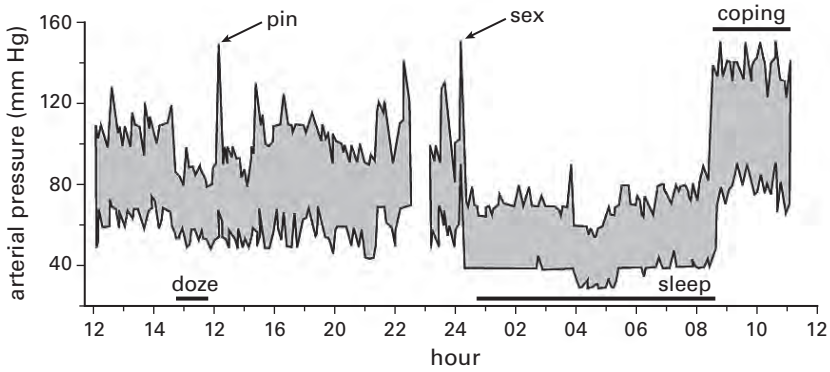


Figure 3.3

Internal systems match behavior. Arterial pressure fluctuates with demand. Each shift in pressure is accompanied by parallel shifts in hormonal and neural signaling that follow the broad catabolic and anabolic patterns. Redrawn from Bevan et al. (1969) and reprinted from Sterling (2004b).

Such anticipatory tuning requires coordinated action of multiple organs and organ systems. To raise pressure, the heart accelerates and vessels constrict. Also the kidney expands blood volume by pumping more salt water into the circulation. The kidney also signals the brain that the body will soon need more supplies of salt and water. Thus, like the gastrointestinal tract, the kidney alerts the brain well in advance of an upcoming need to resupply. Each contribution operates on a different timescale: faster for heart and vessels, slower for kidney's pumping, and still slower for the brain's rise of salt appetite and thirst. These contributions to internal regulation are all initiated simultaneously—and largely by the same signals.

In short, every move we make is matched by a corresponding cardiovascular and renal pattern. Of this we are generally unaware. Yet if the motor command ("Arise!") slightly precedes the internal command ("Tighten vessels!"), blood flow to the head drops, and we faint. That this experience, *postural hypotension*, occurs rarely attests to the rigorous coupling between the cardiovascular pattern and muscular patterns on a 100-ms timescale. On a slower timescale "Arise!" increases by eightfold a signal to the kidney to save water.[2]

Note that matching blood pressure to environmental context requires all of the brain's broad tasks as diagramed in figure 3.2—the collecting and assembling of patterns, the comparison to stores, and so forth. How else to decide if the jab is from a friend or enemy? Moreover, every high-level call

to external action is delivered simultaneously to multiple internal organs. Thus, collecting patterns and distributing patterns are both thoroughly coupled between inner and outer worlds. Where and how the brain effects this coupling will be treated in chapter 4.

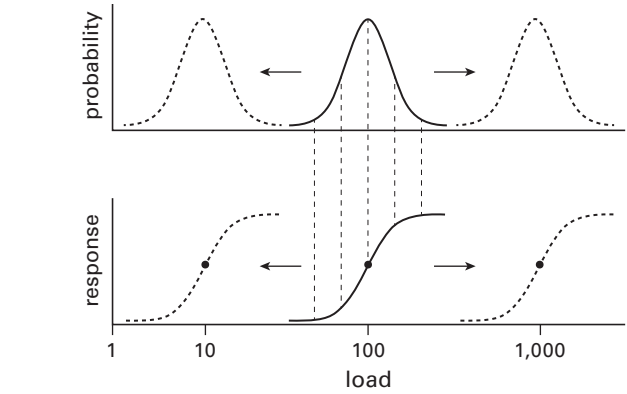
Adapt, match, trade

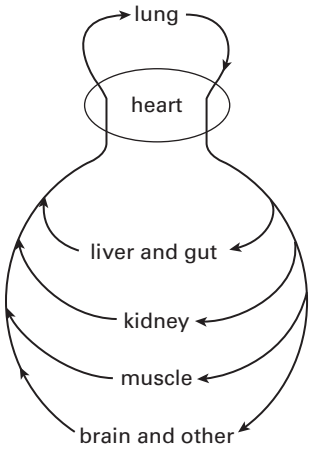
Although this book concerns efficient neural design, we must keep in mind that the brain comprises only 2% of the body's mass and 20% of its energy. So the body also needs to operate efficiently. Each organ should match its capacity to the anticipated need of the organ downstream. Too little and the system will fail; too much and capacity is wasted. So each organ needs constant tuning to anticipate the next demand (figure 3.4). But what happens when a need exceeds the capacity to supply? This problem is solved by arranging various short-term "trade-offs." Such cooperation enhances the range of performance while greatly reducing average excess capacity (figure 3.4).

For example, the "resting" heart pumps 6 L of blood per minute through the respiratory system and then out to the general circulation. Resting skeletal muscle uses about 20% of the oxygenated blood—matched to its modest need for maintaining posture. During peak exercise, muscle must increase its supply by nearly 20-fold, but the pulmonary and systemic circulation can increase their outputs only fourfold. Therefore, the body must either reduce its peak capacity for exercise or increase its peak pulmonary and cardiovascular capacity by fivefold—imagine the chest! Or it can borrow.

Indeed, during peak exercise the splanchnic circulation (gut and liver) and the renal circulation (kidney) both reduce their shares by four- to fivefold, enough to pay part of muscle's bill for exercise. During digestion, when the splanchnic circulation needs more blood, it borrows from muscle and skin—unless skin needs blood for cooling. The brain neither makes loans nor allows overdrafts that might cause it to overheat. Anyone who has eaten and then exercised in the sun will recall how these conflicting demands from muscle, gut, and skin are resolved: by corrective motor commands to internal systems ("Vomit!") and to external systems ("Lie down!"). Moreover, the experience receives a strongly negative evaluation that updates the knowledge store ("Do not repeat!").

This example illustrates three key rules for efficient regulation: (1) adapt response capacity to changes in input level, (2) match response capacities across the system, and (3) trade between systems. Regulatory responses begin promptly—as soon as there is sufficient statistical evidence to predict





Blood Flow (liters per minute)

	Rest	Exercise	Change
lung, heart	6.0	25.0	+ 325%
liver, gut	1.5	0.3	− 80%
kidney	1.1	0.3	− 80%
muscle	1.2	22.0	+ 1800%
brain, other	2.2	2.5	+ 10%

Figure 3.4

Adapt, match, trade. Upper: Adapt response capacity to load. Every system confronts some distribution of probable loads (bold). As conditions shift, so does the distribution (dashed). The response curve (bold) is typically sigmoid with its most sensitive region (steep part) matched to the most probable loads. As a sensor detects a statistically reliable change in the distribution, it prepares the effectors by shifting their response curves to match the new distribution (dashed). Each sensor also adapts its own sensitivity. Reprinted from Sterling (2004b). **Lower:** Organs and organ systems couple efficiently by matching loads to capacities. Trade-offs allow better performance while reducing unused capacity and enhancing "portability." Blood flow pattern changes with exercise: total flow quadruples, but that is insufficient for muscle. To meet the full need, blood is routed from liver, gut, and kidney, temporarily reducing their performance but eventually benefiting from what the muscular effort has accomplished. Data from Weibel (2000).

a new target level. By comparison, self-regulation by feedback to a set point would be hopelessly inefficient. But to execute these principles of predictive regulation requires an organ with knowledge of the outside, knowledge of the inside, and knowledge of the past to anticipate what the whole animal will need over various timescales—the whole brain (Sterling, 2012).

Bigger brains

We seem to have answered "Why a bigger brain?" In a wider world, a more effective brain expands the possibilities for behavior. Control of appendages such as fins, wings, and legs lends speed and scope to exploration, so that vastly more small patterns are encountered which then require selection and assembly. More large patterns require more comparisons, requiring a larger library; more comparisons also require more decisions, and these require more evaluation. Naturally, more neurons are needed, and since neuronal components are irreducibly small (chapter 7), a brain must enlarge.[3]

The larger brain, to be effective, must operate in real time. One need not watch a sloth for very long to realize the limits to life in slow motion. The larger, faster brain must still remain portable and also metabolically affordable. So a brain needs to be both functionally effective and cost-effective. These demands for speed, portability, and affordability all interact; therefore, individually and together they raise questions of brain design. We turn now to the fundamental constraint on any brain design that leads to the first three design principles. Then, in the context of these few principles, we discuss some actual designs (mammal and insect).

Design constraints

The fundamental constraint on brain design emerges from a law of physics. This law governs the costs of capturing, sending, and storing *information*. This law, embodied in a family of equations developed by Claude Shannon, applies equally to a telephone line and a neural cable, equally to a silicon circuit and a neural circuit. This law constrains neural design at all scales and cannot be avoided any more than a B-29 bomber can avoid the law of gravity. But, though the brain is fundamentally an organ that manipulates information, few neuroscientists are familiar with this law or aware of its value for understanding brain organization. We explain it briefly here and give more detail in chapters 5 and 6.

What *is* "information"?

Information is *the reduction of uncertainty about some situation X associated with observing any variable Y that is causally correlated with X*. Uncertainty defines the standard measure: one *bit* is the information needed to decide between two equally likely alternatives. Information depends on causality because, to reduce uncertainty, a message must be reliably relatable to its source, the event that caused it. Any factor that reduces the reliability of this connection, such as noise, increases uncertainty and destroys information.

Reduction of uncertainty succinctly describes the brain's purpose. A spike in an ON ganglion cell reduces the brain's uncertainty that a brighter than average object is located in a particular region of the visual field (chapter 11). And when the brain matches the sensory pattern coded by a patch of ganglion cells to a stored pattern, it reduces a key uncertainty: "Friend or foe?" The answer helps to select the next behavior and implement it. To this end, a motor neuron spike decreases the uncertainty that its target muscle fibers will contract and help the animal move in the appropriate direction. In short, to achieve its core purpose, the brain uses physical devices (neurons and circuits) that represent and manipulate information. So now we must ask: how much information can a neuron represent, and what constrains its capacity?

A neuron's information capacity

To convey information, a neuron must represent the state of its input as a distinct output (input and output must be causally related). It follows that a neuron's capacity to convey information is limited by the number of distinctly different outputs that it can generate. The number of different outputs a spiking neuron can generate in a given time is the number of distinctly different spike trains that it can produce in that time. This depends on two factors, mean firing rate (R spikes per second) and the precision of spike timing (Δt seconds). The upper bound on firing rate is set by spike duration plus the period following a spike when a neuron is refractory (cannot spike). Certain neurons reach this limit during brief bursts, but most neurons operate far below this limit. Precision is limited by channel noise and membrane time constant. Here biophysics limits information capacity.

What is the relation between spike rate, timing precision, and the number of different spike trains a neuron can produce? When a neuron transmits for 1 s, it produces R spikes with a timing precision of Δt (Rieke et al.,

1997). The number of different spike trains, M, is the number of ways the neuron can place its R spikes in $T = 1/\Delta t$ intervals (figure 3.5). Deriving M is a standard exercise in calculating combinations that is often set to students in quaint terms, such as placing peas in pots. The solution is

$$M = T!/(R!(T - R)!), \tag{3.1}$$

where ! denotes factorial and $(T - R)$ is the number of empty (spikeless) intervals.

The number of different messages, M, that a neuron can generate in 1 s converts to information rate. According to Shannon, the information, H, is given by

$$H = \log_2(M). \tag{3.2}$$

Substituting for M using (3.1) gives

$$H = \log_2(T!/(R!(T - R(!)) = \log_2(T!)—\log_2(R!)—\log_2((T - R)!). \tag{3.3}$$

Because Shannon used a logarithmic scale, a message lasting twice as long conveys twice as much information. And, because he used log base 2, information is in bits. Thus, H, the information that a neuron can transmit with messages 1 s long, is its information capacity in bits per second (figure 3.5).

With this expression we can "follow the money." That is, using a standard currency (bits) we can ask like good engineers: how fast does a neuron send information (bits per second) and how efficiently (bits per spike)? And at what cost in space (bits per cubic millimeter) and energy (bits per molecule of adenosine tri-phosphate)? This molecule, abbreviated *ATP*, is the standard intracellular molecule for transferring energy.

Information costs energy and space

Information rate increases with spike rate and with spike timing precision, that is, reduction in Δt. However, for any given precision, information rate increases sublinearly with spike rate (figure 3.5). Consequently, as spike rate rises, bits per spike should fall, and this theoretical decline in bits per spike is observed experimentally (figure 3.5).

There is another way to explain why more frequent spikes carry less information. A symbol that occurs less frequently is more surprising and so more informative (chapter 4, equation 4.2). This effect, which Shannon called *surprisal*, makes a code with fewer spikes more efficient. For example, a code that distributes spikes sparsely among a population of neurons conveys more bits per spike (chapter 12; Levy & Baxter, 1996).

This simple law—infrequent spikes carry more bits—profoundly influences neural design because, following the money, one finds that spikes are

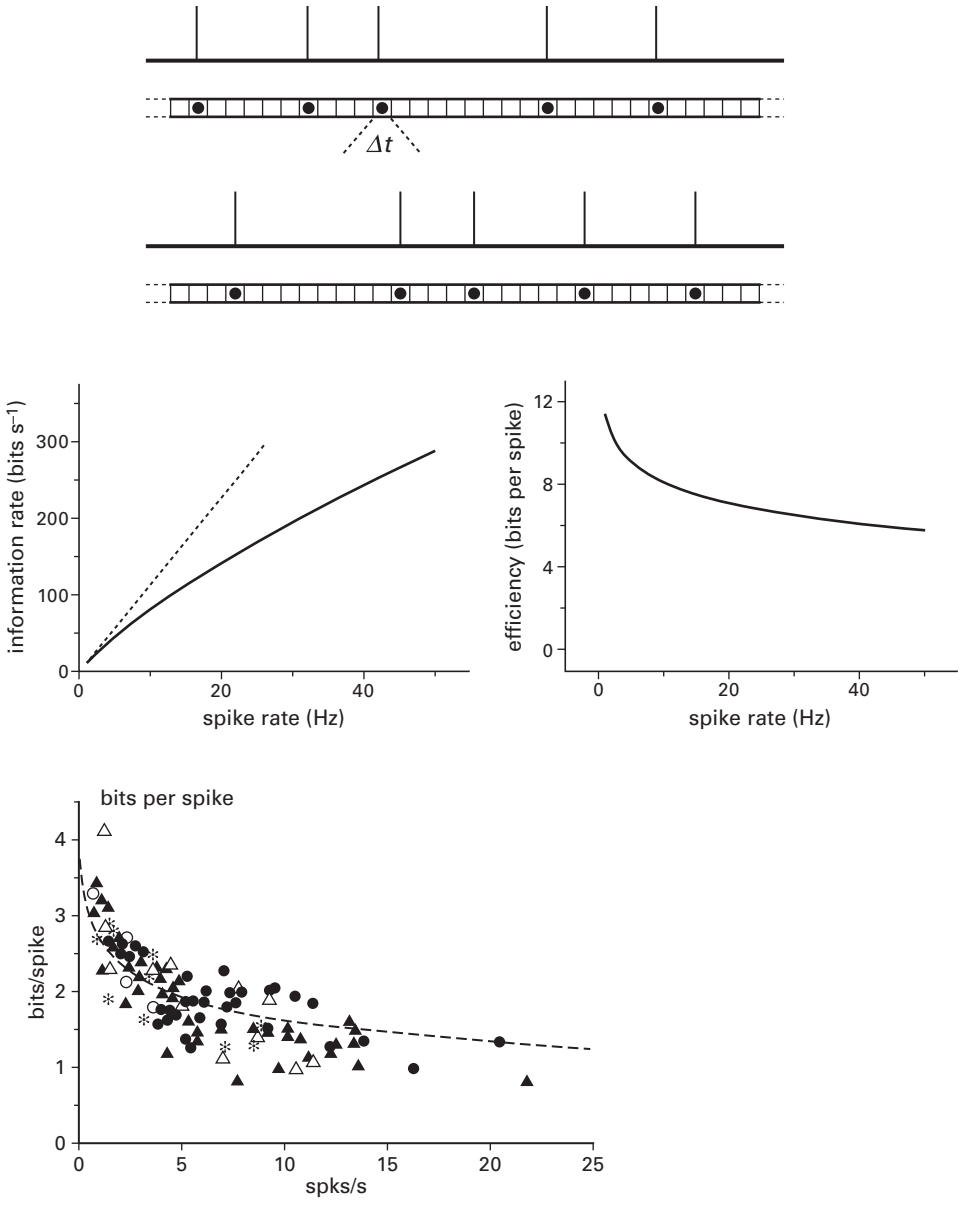


Figure 3.5

Mathematics and biophysics govern the representational capacity of signal trains. **Upper**: Distinct sequences of spikes in time intervals Δt represent different inputs. **Middle left**: Theory predicts information rate to increase sublinearly with spike rate, with the consequence shown at **middle right**: Increasing spike rate reduces the information transmitted per spike. These theoretical curves were calculated using the standard approximation for signal entropy at low spike rates (Rieke et al., 1997, equation 3.22). In general neurons do not achieve their theoretical capacity because of noise and redundancy; consequently, measured values of bits/spike are lower (figure 11.25). **Lower**: Measured bits per spike falls as mean spike rate increases. Data pooled from several classes of guinea pig retinal ganglion cell. Reprinted with permission from Balasubramanian & Sterling (2009).

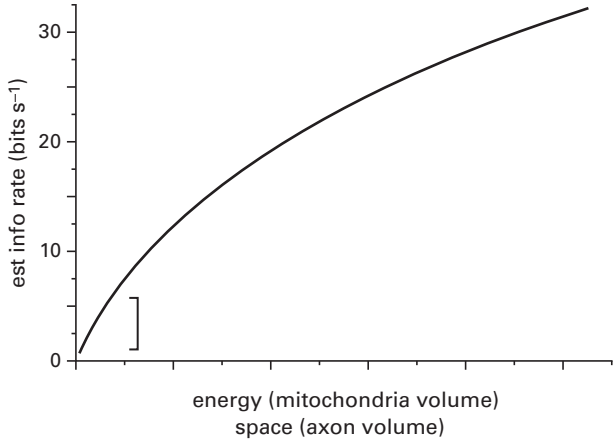


Figure 3.6
Law of diminishing returns. Doubling information rate of retinal ganglion cells more than doubles space and energy costs. Consequently, neural designs try to stay on the steep region of this empirically measured curve. Modified from Balasubramanian & Sterling (2009) and reprinted with permission.

expensive. They use about 20% of the brain's energy (Attwell & Laughlin, 2001; Sengupta et al., 2010). A spike charges a neuron's membrane capacitance by about 100 mV, and the membrane area is substantial due to a neuron's local branching. Higher mean spike rates require a larger cell body with greater membrane area; this increases energy cost per spike and adds to the cost of transmitting bits at high rates. Consequently, where spikes are sent sporadically and at low mean rates, more information can be sent for the same energy—more bits per ATP. This saving in energy by low rates is compounded by a saving in space.

Higher spike rates also require thicker axons.[4] Because axon diameter, d, increases directly with firing rate, axon volume rises as d^2; therefore, doubling the firing rate quadruples axon volume. The concentration of mitochondria, an indicator of energy cost, tends to be constant with axon diameter; therefore, as volume quadruples, so does the energy supply (Perge et al., 2009, 2012). In summary, there is a *law of diminishing returns*: cost per bit, both in energy and space, rises steeply with bit rate (figure 3.6).

Three principles of neural design

The inescapable cost of sending any information and the disproportionate cost of sending at higher rates lead to three design principles: *send only what is needed*; *send at the lowest acceptable rate*; *minimize wire, that is, length and*

diameter of all neural processes. This last principle seems obvious, but it actually reflects a subtle point that arises from the constraint on rate.

Designs should reduce wire, of course, because wire uses space and energy. But wires also use *time* for transmission, and that is time lost to processing and action (Howarth et al., 2012). The constraint is particularly onerous for neural wires because they transmit more slowly than copper wire. Neural conduction velocity is 100 millionfold lower and, for biophysical reasons, faster conduction requires thicker wires (chapter 7). Thus saving time by sending at higher information rates (bits per second) and higher conduction velocities (meters per second) requires thicker axons, which, as noted, involves disproportionate costs in energy and space (Wen & Chklovskii, 2005). Thus, the only economical way to save time is to rigorously shorten wires. This principle shapes brain design across all scales, from an axon's branching and the microscopic design of local circuits, to the overall layout (chapter 13).

With these few principles we can now consider how the mammalian and fly brains are organized on a scale of about 1 mm and why. This macroorganization cannot explain the actual computations because those occur mostly on a finer scale. Nor do we claim that every feature represents the best of all possible designs. Others might work just as well—but they have not been tested. All we can say is that these three principles illuminate the layout of real brains—across a millionfold range of scale and half a billion years of evolution.

4 How Bigger Brains Are Organized

I sensed the earth's slow turning into the dark. The shadow of night is drawn like a black veil across the earth, and since almost all creatures, from one meridian to the next, lie down after the sun has set, one might in following the setting sun, see on our globe nothing but prone bodies, row upon row, as if leveled by the scythe of Saturn.

—W. G. Sebald, paraphrasing Sir Thomas Browne (edited for brevity)

The preceding chapter established that for the brain to send information requires energy and space. Moreover, higher rates (more bits per second) require disproportionately more energy and space because they need thicker axons—for which both space and energy rise as the diameter *squared*. Consequently, the most efficient designs will send only information that is essential and will send it at the lowest rate allowable to serve a given purpose. If information can be sent without any wire at all, that is best. If wires are absolutely needed, they should be as short and as thin as possible. These principles allow substantial insight into how bigger brains are organized.

One design decision is so ubiquitous as to require immediate mention. Brains segregate the wires that interconnect local circuits with each other and with distant circuits. The reason is simple and fundamental: to mingle the wires with the circuits increases total wire length and thickness—violating the principle minimize wire (chapter 13). In mammals axons segregate if they travel beyond a few millimeters. The reason is that increasing distance requires increasing conduction speed to avoid computing delays, and this requires thicker axons. When axon diameter exceeds about 0.5 μm, the axon becomes wrapped in *myelin*, which increases conduction speed by about 6 mm ms^{-1} for every 1-μm increase in diameter (chapter 7). Because myelin in the living brain glistens white, extended sheets of myelinated axons are termed *white matter*.

Saturn's scythe sets brain design

The most profound condition for all life on Earth, the one that uniquely shapes every cell in every organism, is the daily rotation of our planet about its axis. This motion shifts the intensity of arriving solar radiation over the course of 24 hours by a factor of 10^{10}. The impact of this motion is so profound that for many cultures it opens the story of Creation. One familiar example waits only until line 4: "*. . . and God divided the light from the darkness . . . and there was evening and there was morning, one day*" (Genesis 1:4–5).

Animals can certainly survive without light (e.g., in caves), but those with access to light generally choose a particular time of day to forage and thus a particular range of light intensities. The basic choices are diurnal, nocturnal, and crepuscular (dawn and dusk).[1] This decides their investment in sensors: fine spatial vision with color versus acute hearing, possibly with echolocation, versus olfaction plus whiskers. Foraging period also decides their strategies to deal with predators occupying the same slot: camouflage, evasive flight, or skulking behavior.

During its active period the body expends chemical energy to support external behaviors, such as foraging, and internal activities, such as digestion and absorption. Some needs rely on both internal and external actions, for example, thermoregulation. Thus, the active phase involves a broad metabolic pattern, *catabolism*: (1) disassemble large polymeric molecules (proteins, fats, carbohydrates, nucleic acids) into their monomeric building blocks (amino acids, fatty acids, sugars, nucleotides); (2) distribute monomers to metabolically active tissues; (3) convert monomers into energy-bearing molecules, such as ATP, that drive cellular processes; and (4) use an aerobic (oxygen requiring) pathway to produce ATP because it is sixteenfold more efficient (ATP per glucose monomer) than the anaerobic pathway.

During its *in*active period, the body shifts to a broad pattern of renewal, *anabolism*: (1) assemble new polymers for growth, repair, remodeling, and immunity, and (2) replenish reserves by storing residual monomers as resynthesized polymers. Thus, liver converts spare glucose to the storage polymer glycogen; fat cells convert excess glucose to monomeric fatty acids which are then used to build the storage polymer, fat. Because catabolism and anabolism involve opposing sets of biochemical reactions, it would be inefficient to run them simultaneously. Thus, natural selection has separated internal processes into complementary patterns for different segments of the daily cycle.

The brain itself participates in the catabolic/anabolic cycle. During wakefulness it collects, processes, and distributes immense amounts of

information. During sleep, the brain switches over to anabolism via a specific regulatory enzyme and uses this phase to store recently acquired information (Dworak et al., 2010). This involves remodeling local circuits by retracting certain synapses and adding new ones and, in some cases, generating new neurons (chapter 14).

The obligatory alternation between catabolism and anabolism involves throttling down one set of biochemical pathways and revving up another—both of which take time. Consequently, each pattern needs to anticipate the environmental shift—in order to optimally match the key time windows for sleep and foraging. Thus, the pattern seen in figure 3.3, where blood pressure falls with sleep and rises with waking, is completely general: all processes in body and brain move through this cycle. So it is efficient for them to share the same broad signals, and although some processes cease during darkness and others during light, all must follow Saturn's scythe.

Brain clock

Many somatic cells contain an intrinsic clock, established by oscillations of interacting proteins, with a period of approximately 24 hours (*circadian*). But without a mechanism to trim them up, these clocks would soon drift out of phase. So a master clock is needed to track the day, including its continual shift, due to Earth's axial tilt, during its annual revolution about the sun. The master clock comprises a discrete cluster of neurons (about 8,600 in human), the suprachiasmatic nucleus (*SCN*).[2] One subgroup of SCN neurons contains a circadian clock that resets daily based on signals from the retina that track the slow shifts of light intensity across the day and season (figure 4.1).

The master clock requires neither color, nor spatial, nor fine temporal information—only slow intensity changes. Therefore, following two design principles, the retina sends as little as needed and sends as slowly as possible. It uses just a small fraction of retinal output neurons (0.2%), types that cover the retina sparsely and fire at very low rates, a few Hertz averaged over the day (Crook et al., 2013; Wong, 2012). SCN neurons themselves fire between about 8 Hz (day) and about 1 Hz (night; Häusser et al., 2004). To follow another principle, minimize wire, the SCN locates exactly where the optic tracts join the brain (see figure 4.1). But how does the master clock govern patterns across the entire body and the brain as well?

The SCN's relatively few neurons, about 10^4 in rat, could not conceivably contact all other cells directly (Güldner, 1983). Nor should they because their job is not to micromanage every cell but mainly to keep the time. Except for time, the SCN is fairly ignorant—largely unaware of internal

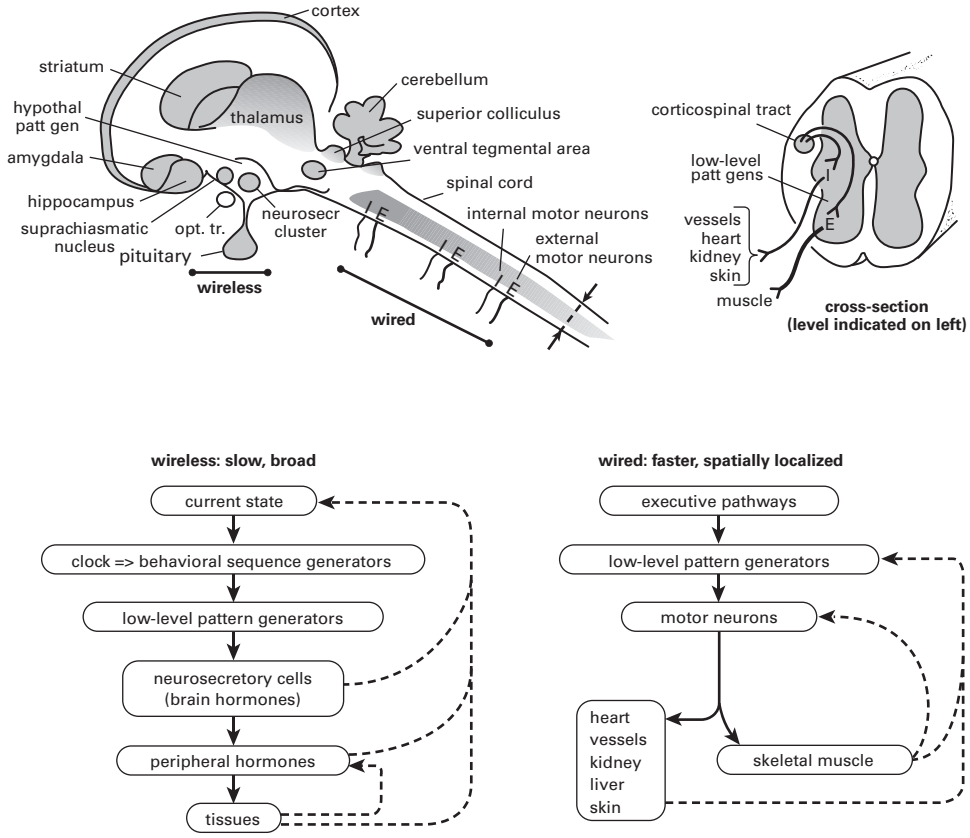


Figure 4.1

Brain's master clock (suprachiasmatic nucleus) informs a network of high-level pattern generators in hypothalamus. These coordinate internal physiology and external behavior. Their network selects a behavior, plus the endocrine and autonomic patterns needed to support it, and communicates the orders to low-level pattern generators, both wireless and wired. **Upper left**: Longitudinal section (diagrammatic) shows the spatial layout of this hierarchy. Opt. tr., optic tract; hypothal patt gen, hypothalamic pattern generator. **Upper right**: Cross section through spinal cord indicates that corticospinal tract tells local pattern generators to match internal physiology to external behavior. **Lower:** Schemes for wireless and wired control.

physiology and external behavior. Therefore, it could not responsibly tell either the body or the brain when to shift the broad pattern. For example, a rat normally forages at night, but what if food becomes sparse at night and plentiful at noon? Were the SCN to directly instruct a command center for foraging, it might send the rat to sleep without its supper.

Coupling clock to behavior: A hypothalamic network

Instead, the SCN couples to an adjacent region, the hypothalamus, that for its comparatively small extent is extremely well informed (figure 4.1; Saper et al., 2005; Thompson and Swanson, 2003). This region monitors myriad internal parameters, including temperature, blood levels of salt, and metabolites, hormonal signals for satiety, hunger, thirst, pain, fear, and sexual state. Some of its neuron clusters express their own endogenous oscillators, and at least one of these responds to changes in food availability (Guilding et al., 2009). This territory also monitors stored patterns—such as best places and times to forage and past dangers. And it monitors the external environment using every sense. Integrating all these data, plus SCN clock time, this region calculates which needs are urgent. Then, balancing urgency against opportunity and danger, it tells the rat whether to forage, mate, fight, or sleep. To execute, it does not micromanage but instead calls the appropriate pattern of behavior (Saper et al., 2005; Thompson & Swanson, 2003).

Hypothalamic circuits, designed to anticipate impending needs, generate signals that elicit various "motivated behaviors," that is, foraging for food, or drink, or sex in response to these integrated signals. As these motivating signals are broadcast to other brain regions, there arises a subjective component that we (among other animals) experience as desire. If one area can be considered as the wellspring of unconscious desires, this is it. It seems amazing that such a small region could access and integrate so much information and evoke such a variety of core behaviors. How could there be sufficient space for hypothalamic neurons to do so much?

Part of the answer is that this well-informed region dictates *sequences* of low-level patterns. For example, feeding behavior requires the sequence: sniffing → biting → chewing → swallowing. These components are programmed in detail by dedicated pattern generators located down in the brain stem near their effector muscles. The local pattern generators manage the exact timings of muscle contraction required for coordinated behavior. The broad sequence that smoothly calls each component into play can be dictated to local pattern generators with a reduced instruction set—something like a music conductor following a score to call forth a Beethoven

symphony from 80 low-level players with nothing but a slender baton. The analogy does not explain the magic in either case, but it does emphasize the design principle: send simple instructions and compute the complex details locally (Büschges et al., 2011).

This economical design allows the hypothalamic region to accommodate a dedicated circuit for each behavioral pattern. These are sufficiently compact that a fine electrode can stimulate them separately, revealing that each circuit evokes a full behavioral pattern, plus the appropriately matched visceral pattern (Hess, 1949; Bard & Mountcastle, 1947). For example, a cat with an electrode placed to evoke "angry attack" arches its back, hisses, and strikes with bared claws and teeth (somatic pattern). Simultaneously it dilates pupils, raises hackles, and increases cardiovascular activity (visceral pattern; Büschges et al., 2011; Hess, 1949).[3] Moving the electrode by a few millimeters can activate circuits for other behaviors: feeding or drinking or copulating or curling up to sleep. In short, many circuits fit in a small space because their output messages are simple.

Each behavior circuit is demonstrably guided by a rich set of input signals. For example, a cat electrically stimulated to feed will attack a ball of cotton that mimics a mouse, but only briefly, whereas it persistently attacks a real mouse until the current stops. If the mouse is replaced by a substantial rat, the cat retreats to its home corner. Evidently the feeding circuit is modulated by inputs that identify prey, distinguish true prey from false, and recognize dangerous prey—all based on comparison to stored patterns. Moreover, each behavior is imbued with a motivational component—apparent when an animal stimulated to feed will seek hidden food and work to obtain it (press a lever).

How does this small region, the hypothalamus, access the brain's core systems for perception, spatial memory, danger, economic value, and urgency? Again, it relies on details computed elsewhere and delivered only as conclusions: time from the SCN; integrated physiological data from myriad sources that define internal state; selected memories of location and danger from hippocampus and amygdala; recent history of reward value from the striatal system; high-level analysis of choices from prefrontal cortex. Because these inputs to the hypothalamic region all send summaries, they can use low information rates and thus fine fibers, thereby greatly conserving space (figure 4.1). Energy is also conserved, allowing this crucial region to have among the lowest metabolic rates (Sokoloff, 1977).

This strategy allows a major organ for memory, the hippocampus, to access key aspects of an animal's life history but send only modest clips to guide a particular behavior. This might explain why its output tract (*fornix*)

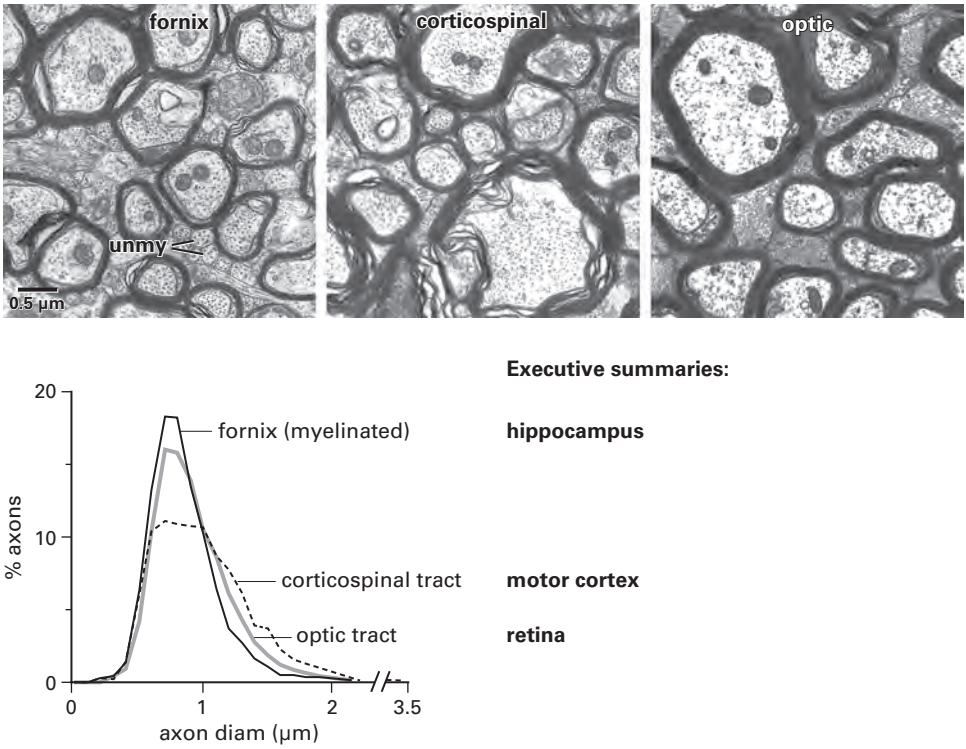


Figure 4.2
Fiber tracts that transmit summaries share an economical design. Their axon diameters distribute log-normally, with many thin axons and fewer thick ones. Halving the diameter reduces space and energy costs by fourfold. These "summary tracts" use low mean firing rates (see figure 4.6). Reprinted with permission from Perge et al., 2012).

can manage with mostly fine fibers, resembling the optic nerve, which itself sends strongly edited summaries from the retina (chapter 11). An apparently similar strategy allows sensorimotor areas of the cerebral cortex to lend speed and agility to motor behaviors via an output tract (*corticospinal tract*) of similar fine structure (figure 4.2; Quallo et al., 2012). In short, the hypothalamic network is designed to receive executive summaries as input and deliver broad memoranda as output (Perge et al., 2012).

Resurrection

To be awakened from a deep sleep feels horrible. And no wonder: every cell in the body and brain struggles to function according to its catabolic phase—against all central instructions to remain in anabolic phase. But

when anabolism has gone to completion—when the body has replenished stores, healed wounds, rebuilt muscles and immune systems, and when the brain's sorting mechanisms have punched "delete" or "save"—then all the cells and tissues finally wake up more or less simultaneously.

The SCN signals "dawn" to the hypothalamic network—which then decides, based on many factors, whether it is auspicious to awaken.[4] If so, the network signals a nearby cluster of neurons (comparable in size to SCN) to secrete the peptide transmitter *orexin*. The orexin neurons project widely over the brain to activate a cascade of systems that regulate arousal (Sakurai, 2007). Because orexin neurons couple the clock to the brain's arousal system, an animal lacking orexin tends to collapse unpredictably into sleep.

The orexin cluster specifically awakens olfactory sensors, enhancing their sensitivity, and it awakens motor mechanisms for foraging (Julliard et al., 2007). Informed by the master clock, the orexin cluster uses the hypothalamic pattern generator network to coordinate alertness, olfactory sensitivity, and the sense of hunger—all to initiate foraging at the proper time. Now it is time for brain signals to reinstate the broad catabolic pattern: mobilize energy stores from liver and oxygen carriers (red blood cells) from spleen and bone marrow; re-expand the vascular reservoir with salt water from the kidney. And it is time to *de*mobilize anabolic processes for growth, repair, and immunity.

In summary, the hypothalamic network manages the whole brain and all of its functions—without micromanaging. But now, what about micromanaging? A conductor is all well and good, but someone must play the bassoon. So how are the processes that do involve micromanaging governed by the design principles considered here?

Distributing output patterns

Wireless signaling

Design principles dictate that the slowest processes should be governed by the slowest effectors and the least wire. Where signals can be sent with zero wire, that is best. Consequently, the effectors for micromanaging the broad catabolic and anabolic patterns are endocrine glands. For example, the adrenal gland secretes a steroid hormone that enhances the kidney's uptake of sodium and a different one that enhances catabolism, mobilizing energy and suppressing growth and repair. Testis secretes anabolic steroids that enhance muscle, and liver secretes a hormone that stimulates red blood cell production. What coordinates these low-level effectors? Higher-level endocrine signals from the pituitary gland, which is in turn governed by

hormones from the brain. Wireless regulation of two particular functions, blood pressure and muscle contraction, is summarized in figure 4.3.

Brain hormones are secreted directly into the circulation by neurosecretory neurons whose clusters lie adjacent to the hypothalamic network of pattern generators. The pattern generators deliver their well-informed but simple orders via very fine, very short wires (figure 4.1). Each node in the hypothalamic network can call a particular pattern of brain hormones for release into the blood just upstream of the pituitary, thus stimulating it to release its own hormones into the general circulation. The whole endocrine network reaches every cell in the body within seconds. Not blazingly fast, but on the other hand, the messages are broadcast without any wire at all and with zero energy cost above what the heart is already doing.

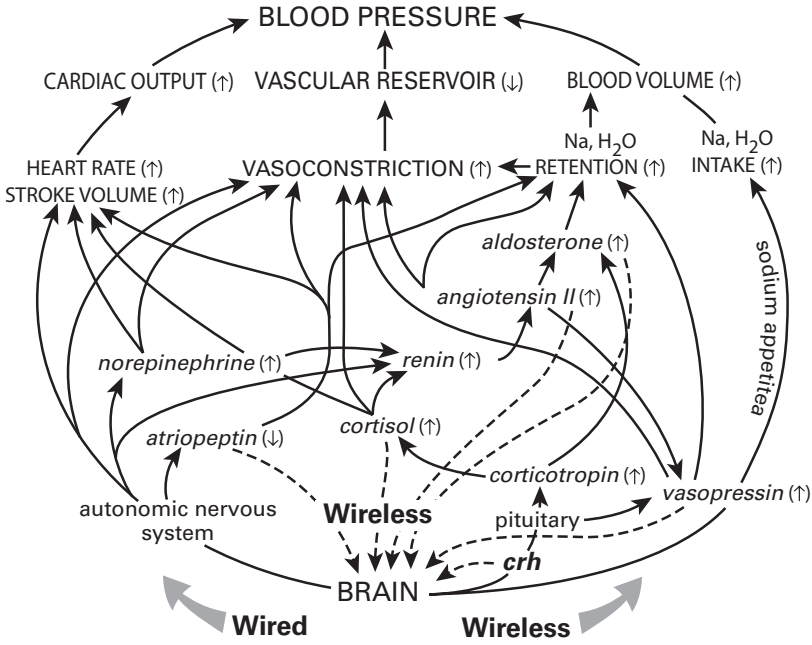
The genius of this wireless system lies partly with the receivers. Although all somatic cells are exposed to all hormones, only certain cell types download a given message. To do so, they produce a specific molecular receptor that binds a particular hormone and triggers a particular intracellular response. Thus, information broadcast diffusely to the whole body can be read out by a restricted number of cell types—whose responses to the signal are thereby coordinated. The molecular mechanism and reasons why it is so economical are described in chapter 5.

Another clever feature is that receiver cells can express different subtypes of the molecular receptor. Each subtype can couple within the cell to a particular *second messenger* with its own stereotyped action. For example, one messenger can greatly amplify the hormonal signal and use it to either activate or suppress some intracellular process. Thus, a single message broadcast wirelessly can evoke complex response patterns among different tissues that include negative as well as positive correlations.

For example, skeletal muscle acts rapidly on the outer world via fast signals over thick wires. Yet, it is also a tissue within the body and is thus regulated wirelessly by various hormones, including anabolic steroids, insulin, growth hormone, and thyroxin (figure 4.3, lower panel). Thus, wireless signaling helps the brain to efficiently couple inner and outer worlds.

Wireless collecting

The brain also uses wireless receivers, a small set of *circumventricular organs* that locate at specialized interfaces between brain and blood vessels. There the normal barrier between blood and brain parts, thus exposing neurons to circulating chemicals. These neurons select just what they need by expressing the appropriate molecular receptors. For example, the *subfornical organ* locates near the hypothalamic pattern generators that regulate


BLOOD PRESSURE
CARDIAC OUTPUT (↑)
VASCULAR RESERVOIR (↓)
BLOOD VOLUME (↑)
HEART RATE (↑)
STROKE VOLUME (↑)
VASOCONSTRICTION (↑)
Na, H2O RETENTION (↑)
Na, H2O INTAKE (↑)
aldosterone (↑)
angiotensin II (↑)
sodium appetite
norepinephrine (↑)
renin (↑)
cortisol (↑)
atriopeptin (↓)
corticotropin (↑)
vasopressin (↑)
autonomic nervous system
Wireless
pituitary
crh
BRAIN
Wired
Wireless

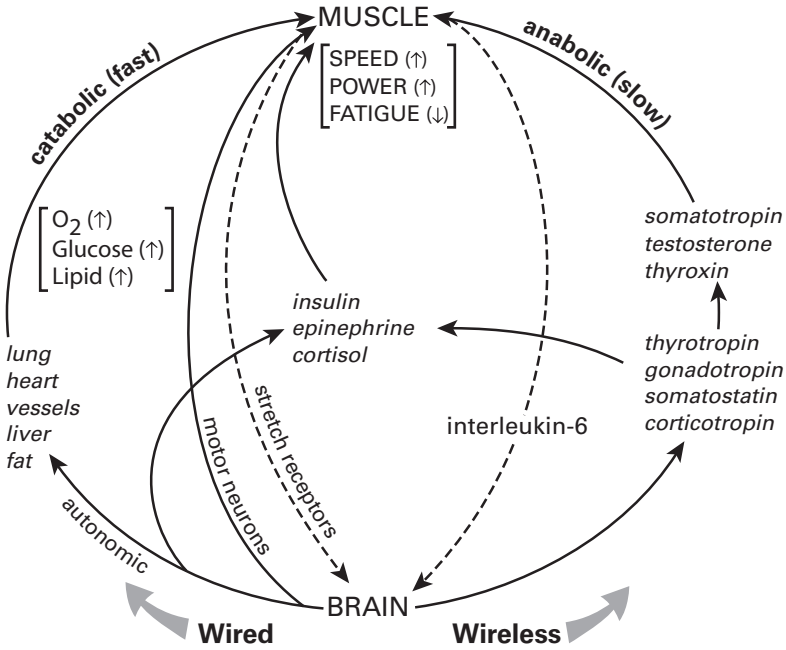

MUSCLE
SPEED (↑)
POWER (↑)
FATIGUE (↓)
catabolic (fast)
anabolic (slow)
O2 (↑)
Glucose (↑)
Lipid (↑)
somatotropin
testosterone
thyroxin
insulin
epinephrine
cortisol
thyrotropin
gonadotropin
somatostatin
corticotropin
lung
heart
vessels
liver
fat
stretch receptors
motor neurons
interleukin-6
autonomic
BRAIN
Wired
Wireless

Figure 4.3
Wireless regulation broadcasts slow signals to efficiently couple inner and outer worlds. Upper: To adjust blood pressure rapidly and locally, the brain uses wires (autonomic nerves). But to shift pressure slowly and broadly, it uses wireless signals (hormones) (*italicized*). Dashed lines indicate wireless feedbacks to brain. Feedbacks by wire are used by certain sensors, such as for oxygen and pressure, but are not shown. *CRH, corticotropin releasing hormone*. **Lower**: Catabolism in muscle activates rapidly to support contraction; so to rapidly activate catabolism, the brain uses wires. But anabolism in muscle is slower, so the brain activates those processes with wireless signals (*italicized*).

appetite for salt and water (figure 4.4). The neurons sense the blood's sodium level, plus levels of hormones (*angiotensin II* and *aldosterone*) that tell the kidney to conserve sodium (figure 4.3, upper panel). Thus, this wireless receiver closes the loop for anticipatory regulation: the brain sends instructions to kidney regarding salt and water, and the brain's subfornical organ wirelessly receives information about the current state[5] of sodium balance.

Need for wires: Faster, spatially directed signaling

Neurosecretions spread slowly (over seconds) and modulate target cells slowly because the packets of hormone molecules released into the voluminous vascular system become greatly diluted (to concentrations ~10^{-9} M). Therefore, molecular receptors need high affinity to capture the hormone, and thus their *un*binding is slow (chapter 6). However, this delay is inconsequential because the intracellular processes that they are regulating typically span minutes or hours. Thus, the slow rhythms of wireless signaling match their targets, physiological processes that rise and fall slowly.

Where faster responses are needed, the hormone is released into a *portal vessel* leading directly to a target downstream. Because the hormone is less diluted, it can be captured by lower affinity receptors, which unbind faster, and operate on the steep limb of the binding/response curve. For example, the brain hormone corticotropin-releasing hormone is secreted into portal vessels leading to the pituitary; the adrenal cortex secretes steroid hormones into portal vessels leading to the adrenal medulla. Yet certain internal process must proceed still more smartly, and that needs wire.

For example, for the brain to initiate a change in body posture, it must alter the pattern of muscle contraction. This will require a change in the distribution of oxygen and thus an altered vasomotor pattern to redistribute blood. Furthermore, active muscle will need to take up glucose, and that

will require triggering insulin secretion from pancreatic cells. These vascular and endocrine adjustments need to be initiated along with the muscle activity, and these faster, spatially localized signals demand wires.

This need is served by autonomic neurons whose axons contact every internal organ and blood vessel. Their mean firing rates are less than 1 Hz, and thus in Shannon's sense they transmit at low information rates. This seems intuitive, since a message—"Secrete some insulin" or "Constrict this vessel"—goes somewhat beyond "yes" or "no" (one bit), but not by much, and thus it can be accomplished with few spikes. Signals that transfer at rates below 1 Hz use the finest, cheapest axons.

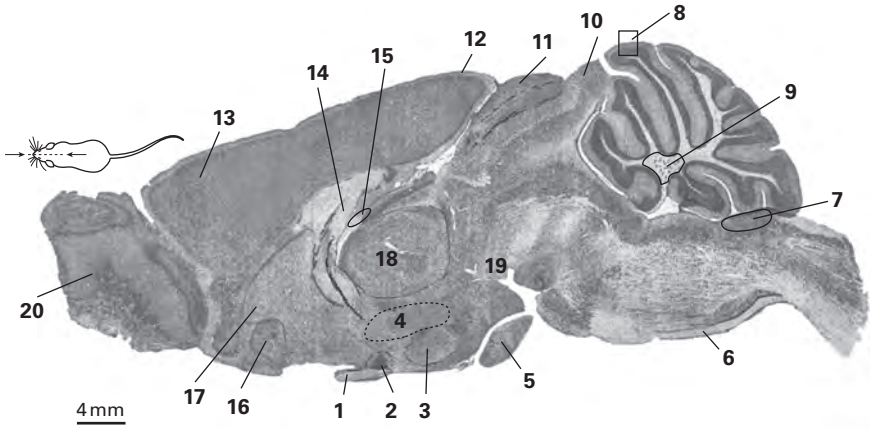
What manages these autonomic effectors? *Answer*: low-level pattern generators located in the brain stem and spinal cord near the output clusters (figure 4.1, right). The latter form two subsystems (*sympathetic* and *parasympathetic*), which employ different transmitters. Each transmitter couples to several receptor types, which in turn couple to different second messengers. Consequently, the autonomic effectors can generate rich internal patterns. They are the orchestral players—ready and waiting for the conductor to select the next pattern and tempo.

What manages the muscles that change the body's posture? Again the answer is low-level pattern generators located near the motor neuron clusters. These pattern generators must increase force from certain muscles and decrease it from others—in just the right amounts and at just the right instants. Sharp timing requires large currents, rapid integration (short time constants), and high mean firing rates (chapter 7). Therefore, these pattern generators need large neurons with thick dendrites and thick axons.[6] To reduce costs, they locate near their effectors. This lengthens the descending pathways that supervise them, but as noted, those are cheaper (figure 4.2).[7]

Motor control requires rapid feedback. The fastest signals from skin and joint receptors travel at about 50 m s^{-1}, and those from muscle receptors travel at about 100 m s^{-1}. These velocities require very thick, myelinated axons, 8–17 μm in diameter.[8] These fibers are 10-fold thicker than for the descending tracts and thus 100-fold greater in volume. Were pattern generators located higher in the brain, for example, nearer to the hypothalamic pattern generators, feedback would be delayed, even though these axons are huge. Thus, the combined needs for fast output and fast feedback constrain the low-level generators of motor patterns to locate near their effectors, the motor neurons (figure 4.5).

Arrangement of effector clusters

The neurosecretory clusters locate adjacent to the hypothalamic network, which can thus modulate them with very little wire (figure 4.1). But the



1. optic nerve
2. suprachiasmatic nucleus (clock)
3. hypothalamic neuroscretory cluster (brain hormones)
4. hypothalamic pattern generators (high-level)
5. pituitary gland (wireless signals → periphery)
6. corticospinal tract (summaries from motor cortex to low-level pattern generators)
7. area postrema (monitor blood chemistry)
8. cerebellar cortex (correct errors of intention)
9. cerebellar output clusters (integrates cerebellar output)
10. inferior colliculus (early auditory processing)
11. superior colliculus (orient head and eyes toward key information sources)
12. primary visual cortex (far from long-term storage sites)
13. frontal cortex (near long-term storage sites)
14. fornix (summaries from hippocampus to hypothalamic pattern generators)
15. subfornical organ (monitor blood sodium and related hormones)
16. amygdala (tag high-level patterns for storage)
17. striatum (evaluate predictions of reward)
18. thalamus (process signals for economical transfer to cerebral cortex)
19. ventral tegmental area (dopamine neurons → frontal cortex + striatum)
20. olfactory bulb

Figure 4.4

Longitudinal section through rat brain. This section shows relative size and location of various structures discussed in this chapter. From http://brainmaps.org/ajax-viewer.php?datid=62&sname=086&vX=-47.5&vY=-22.0545&vT=1 © The Regents of the University of California, Davis campus, 2014.

autonomic and somatic motor neuron clusters lie far from the hypothalamic network, distributing from the midbrain down through the spinal cord. This extended distribution allows space for their low-level pattern generators. The total volume of the autonomic effectors and their pattern generators, summed over the length of the spinal cord, is about 100-fold greater than that of the hypothalamic network.[9] This need for space easily justifies extending the brain tailward and helps explain why this design has been conserved. Moreover, the extension allows additional efficiencies.

Neurons that share input from the local-pattern-generator should cluster close together. Thus, the autonomic effector neurons that regulate internal organs and endocrine cells align in a column, allowing them to share input from the columnar low-level generator of autonomic patterns. Somatic motor neurons also align in columns—parallel to the autonomic column and near it; therefore, circuits for internal physiology and external behavior can be coordinated locally via short wires (figure 4.1).

Because low-level pattern generators for internal physiology and behavior locate together, descending tracts can regulate them together with no extra wire. For example, the corticospinal tract sends a reduced instruction set from motor cortex to low-level pattern generators for muscle (Yakovenko et al., 2011) and also to adjacent autonomic pattern generators for kidney (see figure 4.1). Thus, the descending message, "Arise!" can be sent efficiently to both effectors (Levinthal & Strick, 2012).

Somatic motor neurons extend this design for efficient component placement to a still finer level (figure 4.5). Motor neurons for a given muscle often fire together, implying shared inputs, so they cluster. Motor neurons for muscles that act synergistically across a joint also often fire together, also implying shared inputs, so their clusters stay close. Motor neurons for muscles that cooperate across multiple joints also fire together, but less often, so their clusters are further apart, distributing longitudinally with separations roughly corresponding to their frequencies of coactivation. Finally, motor neurons for antagonistic muscles tend to fire reciprocally, flexors excited/extensors inhibited. This reciprocity depends on a shared circuit (cross-inhibition, like the worm), so the clusters of antagonistic motor neurons also stay close—in parallel columns that run down the spinal cord (Sterling & Kuypers, 1967; figure 4.5).

In short, somatic motor neurons distribute according to a broad design rule: *neurons that fire together should locate together*.[10] This rule also governs sensory maps and all the brain's orderly topographic connections (chapters 12 and 13).

Design for an integrated movement

The placement of motor neurons in longitudinal columns allows a pattern generator to economically evoke an integrated limb movement (Bizzi & Cheung, 2013). The task is to excite contractile units in dozens of muscles across several joints and suppress their antagonists (Sherrington, 1910; Creed & Sherrington, 1926). The key is for motor neurons to send their dendrites longitudinally within a column for long distances (about 1 mm) so that dendrites of synergists overlap. Then, an input axon can coactivate synergists simply by branching as a T within the column and distributing synapses at regular intervals. Strong synergists will greatly overlap their dendrites and thus share more input than weaker synergists that overlap less (figure 4.5, lower). All inputs to the motor neuron columns follow this rule, including axons from sensory receptors, axons from local pattern generators, and axons from cortex (figure 4.5, lower). This design uses less wire than any other conceivable geometry, and thus it is optimal (chapter 13).

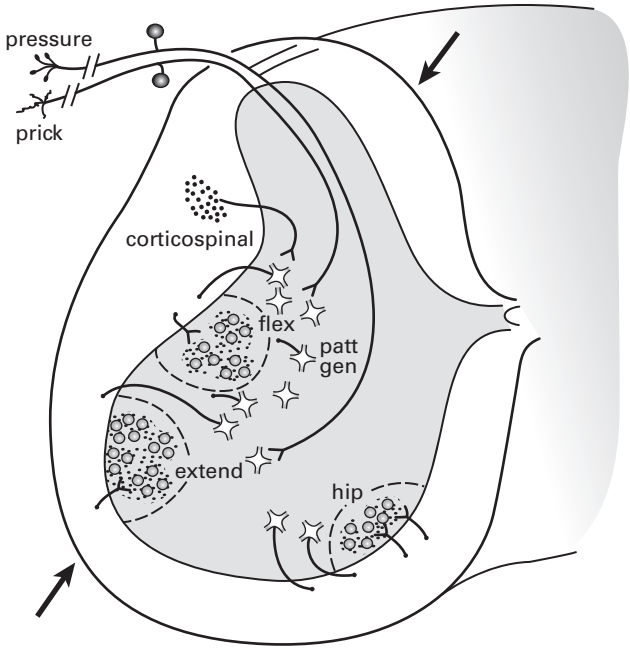
Pattern-generator neurons use thick, myelinated axons to synchronously activate motor neurons at different levels of the motor neuron column. To do this while least disturbing the synaptic circuitry, the axons are routed into the white matter where upon reaching the appropriate levels, they reenter the motor column and connect (figure 4.5).

One benefit of this architecture is that different sensory receptors from the same location can efficiently evoke opposite responses. Here, pressure receptors from the foot connect to the extensor pattern generator, so as weight shifts to that foot, all the extensors are excited to support the limb. Pain receptors connect to the flexor pattern generator, so as weight shifts to that foot, all the flexors are excited (and extensors inhibited) to withdraw the limb. These alternative decisions are accomplished at the lowest level, thereby avoiding the costs in time, space, and energy of consulting higher levels. The corticospinal tract delivers "executive summaries" from motor cortex to the pattern generators. So a corticospinal axon can simply say "Flex!" and local circuits do the rest (Bizzi & Cheung, 2013).

Collecting input patterns

Different senses, different costs

The wider world that makes a larger brain such a good investment contains a seeming infinity of patterns carried by diverse forms of energy: electromagnetic (light), heat, mechanical vibration of air (sound), direct


pressure
prick
corticospinal
flex
patt
gen
extend
hip

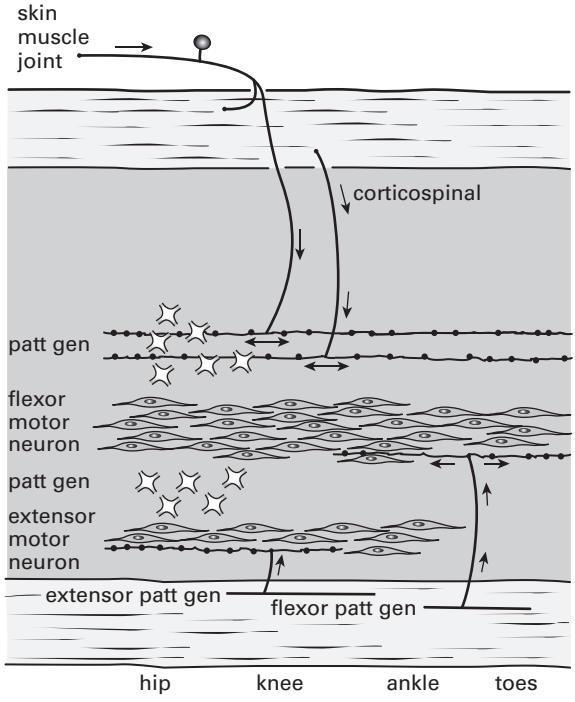

skin
muscle
joint
corticospinal
patt gen
flexor
motor
neuron
patt gen
extensor
motor
neuron
extensor patt gen
flexor patt gen
hip
knee
ankle
toes

Figure 4.5
Efficient wiring for integrated movement. Upper: Cross section through the spinal cord. Flexor and extensor motor neurons for the leg form separate clusters, which locate near each other and also near to the pattern-generator neurons that reciprocally excite and inhibit them. The flexor and extensor clusters form parallel columns extending over several segments of spinal cord. Each column is structured as a motor map: motor neurons for thigh muscles locate at higher spinal levels, then in descending order: knee, ankle, and toes. Within a column, the motor neuron dendrites extend longitudinally for about 1 mm in both directions; consequently motor neuron dendrites for synergistic muscles overlap. Their overlap allows a pattern-generator axon to excite motor neurons for synergistic muscles simply by spreading its axon arbor longitudinally within the dendritic plexus. This uses the least possible wire to excite motor neurons for several muscles. The longitudinal dendrites appear in this plane as dots scattered within the motor neuron clusters. Motorneuron clusters for hip muscles locate separately, near the midline. Patt gen, pattern generator. **Lower:** Longitudinal section through spinal cord in the plane indicated by arrows in upper diagram. This plane reveals the motor neurons' longitudinal dendritic plexus that spans the motor map from hip to toe. This plane shows the pattern generator axons leaving the white matter to enter a flexor or extensor dendritic plexus where they encounter overlapping dendrites of synergistic motor neurons. The pattern generator neurons do not orient longitudinally and thus do not overlap. Consequently, a sensory axon or a corticospinal axon, coursing longitudinally within the pattern-generator columns, can efficiently access a discrete subset of pattern-generator neurons and thus a subset of motor neurons for a particular integrated limb movement.

mechanical contact, volatile molecules (odorants), molecules in solution, electrical patterns, magnetic fields, and gravity. Animals evolve mechanisms to collect information carried by all these forms—and use them to find food and mates, to avoid predators, and to orient in space and time. The challenge is to decide which forms to invest in and how much. Some are intrinsically cheap whereas others are intrinsically costly. Yet for certain lifestyles, cheap won't work and expensive is well rewarded. So an animal selects from the universe of patterns according to how it makes a living and during what phase of the planet's daily rotation.

Animals that forage by day invest heavily in photoreceptors sensitive to wavelengths between 300–700 nm. Animals that forage by night invest heavily in other receptors. Snakes that hunt mice use temperature receptors to extend their range to the infrared[11] (about 800 nm). Moths and frugivorous bats invest heavily in olfactory receptors, but certain bats prefer moths over fruit and so invest heavily in sonar systems that produce, detect, and process ultrasound (frequencies up to 180 kHz).

Fish that inhabit clear water invest in photoreceptors and, because the spectral content shifts with depth toward blue, those that inhabit deeper waters shift their peak photosensitivity correspondingly. Fish residing in caves *dis*invest in photoreceptors and are essentially blind. Certain fish inhabiting rich, but turbid tropical rivers invest in electrosensory systems that interrogate their surroundings by emitting brief electrical pulses or sinusoidal waves up to 2 kHz, and measuring the electrical field with electroreceptors.

Sensors differ greatly in cost. Olfactory sensors are slow and relay information at low mean rates, so their axons are extremely fine, approaching the limit set by channel noise (chapter 7). Vision is faster, so retinal ganglion cell axons (optic nerve) fire at higher mean rates and are somewhat thicker; and hearing is still faster, so auditory axons are far thicker (figure 4.6). This progression of axon calibers corresponds to a linear progression of firing rates (figure 4.6). However, since space and energy costs rise steeply with diameter and firing rate, the thickest auditory axon costs 100-fold more than an olfactory axon (Perge et al., 2012).

Systems for sensing at the skin follow similar design rules. Mechanosensors employ various mechanisms to transduce and filter pressure and touch. Some sense high frequencies (vibration) and transmit via thick axons (figure 10.3); other mechanosensors sense lower frequencies and transmit via finer axons. Sensors for pain and temperature send at the lowest spike rates and use the finest axons. Centrally, the fast and slow systems are processed in parallel and to a large degree arrive at their thalamic relay over separate tracts (Willis & Coggeshall, 1991; Maksimovic et al., 2013; Boyd & Davey, 1968).

Of course, these costs of collecting primary patterns are merely down payments. Auditory patterns arriving at high rates must be *processed* at high rates—so their initial central circuits use thick wires and fast (expensive) synapses (Carr & Soares, 2002). The most expensive parts in a mammalian brain are those devoted to early auditory processing, for example, the *superior olivary nucleus and inferior colliculus* (see figure 4.4; Mogensen et al., 1983; Borowsky & Collins, 1989). Thus, the ultrasonic imaging system of an insectivorous bat is intrinsically more expensive than the olfactory system of a frugivorous bat.

For fish that use electrical signaling, the cost is tremendous. One set of neurons needs to produce high-frequency pulses; another needs to detect them and signal the brain. Then, as for the insectivorous bat, processing is expensive. The computations required by this system are executed by cerebellar circuits, so the cerebellum greatly expands (figure 4.7). Consequently,

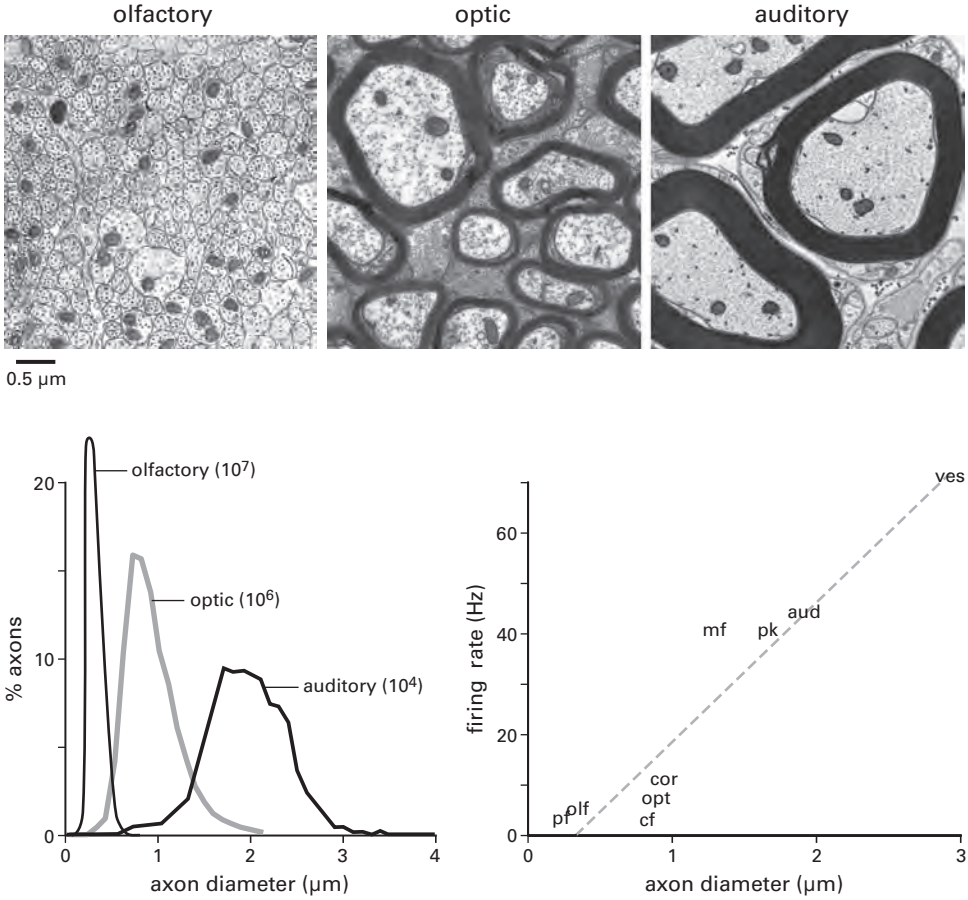


Figure 4.6

Unit cost of sending information differs greatly across senses. **Upper row**: Electron micrographs of cross sections through the olfactory, optic, and cochlear nerves shown at the same magnification. **Lower left**: Distributions of axon diameters. The auditory axons are nearly sevenfold thicker than the olfactory axons, so their unit volume and energy cost are nearly 50-fold greater. In parentheses are the number of axons serving that sense. The relation is reciprocal: low unit cost allows a many-unit design (olfactory) whereas high unit cost restricts the design to fewer units (auditory). **Lower right**: Higher mean firing rates require thicker axons. Vestibular axon unit cost is 100-fold greater than that unit cost of an olfactory axon. Reprinted with modifications and permission from Perge et al., 2012.

the brain of a mormyrid fish that uses electrical signaling is huge compared to a trout of comparable body size (figure 4.7) and requires 60% of the resting animal's energy budget! This emphasizes that the purpose of brain design is not necessarily to operate on the cheap—for that would limit functionality. Rather, it is to ensure that the brain's investment pays off.

Design and usage of sensor arrays

In the mammalian ear each auditory hair cell is tuned to a particular range of frequencies—with the cells mapped along the cochlea's basilar membrane from lowest frequency (20 Hz in human) at the apex to highest (20,000 Hz) at the base.[12] The axons serving the highest frequencies fire at higher mean rates and are roughly threefold thicker than those for the lowest frequencies. Consequently, they use nearly 10-fold more volume and energy (figure 4.8). For humans the most critical frequencies are those for speech—which peak below 500 Hz and decline gradually out to 3500 Hz (figure 4.8); From the perspective of brain economy it is fortunate that natural selection has placed human speech at the lower end of the auditory nerve's frequency range, which is the most economical (figure 4.8). This design decision also saves costs downstream for central processing.

It turns out that music uses the same frequencies as human speech. The most frequent intervals in music correspond to the greatest concentrations of power in the normalized spectrum of human speech. Moreover, the structure of musical scales, the preferred subsets of chromatic scale intervals, and the ordering of consonance versus dissonance can all be predicted from the distribution of amplitude–frequency pairings in speech (Schwartz et al., 2003). Thus, music's tonal characteristics match those of human vocalization, which are the predominant natural source of tonal stimuli. This match seems understandable given that music serves to express and communicate emotions. It seems that the blues evoke sadness because those are the sounds that ancient humans uttered in communicating *their* sadness (Bowling et al., 2012; Han et al., 2010).

Music is processed by auditory areas in the right hemisphere, the side specialized for perceiving and expressing emotion; language is processed by corresponding areas on the left. It might seem redundant to analyze sounds with the same frequencies and structure in both hemispheres, but the computations are quite different, so it is economical to separate the circuits. What is the payoff for investing such substantial neural resources? Human survival and reproduction requires social cooperation—which depends upon communicating emotionally as well as cognitively. In short, music

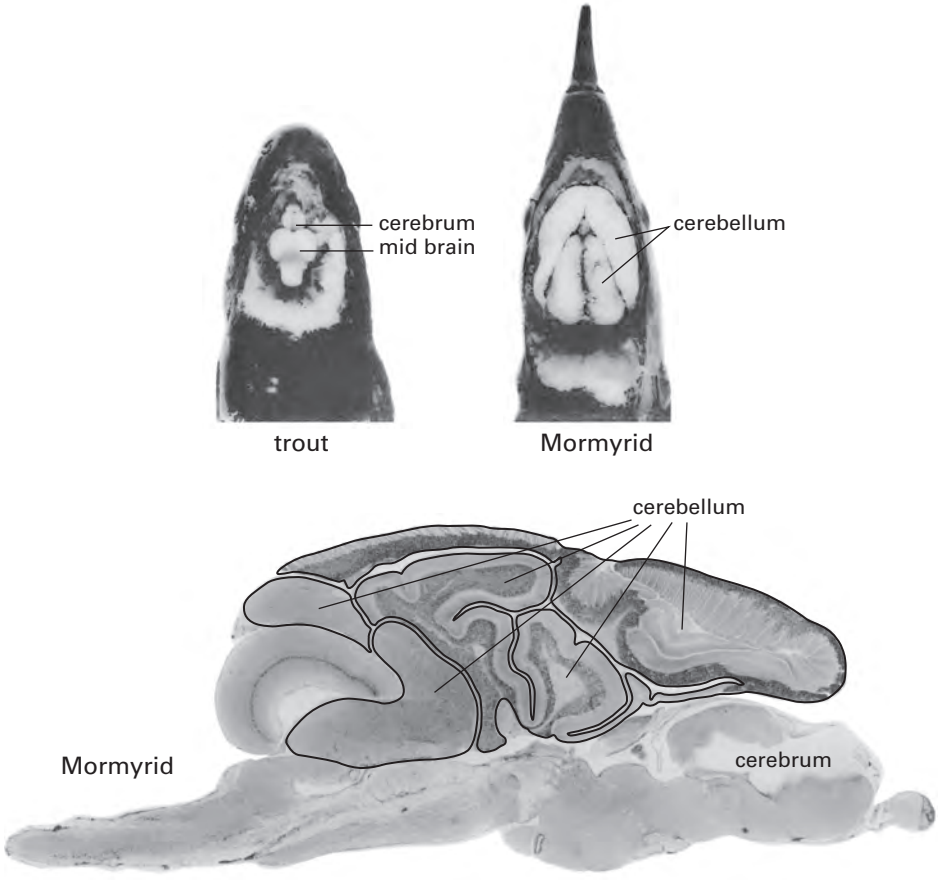


Figure 4.7

Mormyrid brain greatly expands cerebellar structures. Upper: Electrosignaling Mormyrid from turbid waters resembles trout in body size but requires a far larger brain, most of which is a highly elaborated cerebellum. **Lower:** Longitudinal section shows that the cerebellum (outlined) occupies most of the brain, completely obscuring the cerebrum. Central processors of high temporal frequencies often use a cerebellar-like design, including, in mammals, the dorsal cochlear nucleus (Oertel & Young, 2004; Bell et al., 2008). Reprinted from Nieuwenhuys & Nicholson (1969).

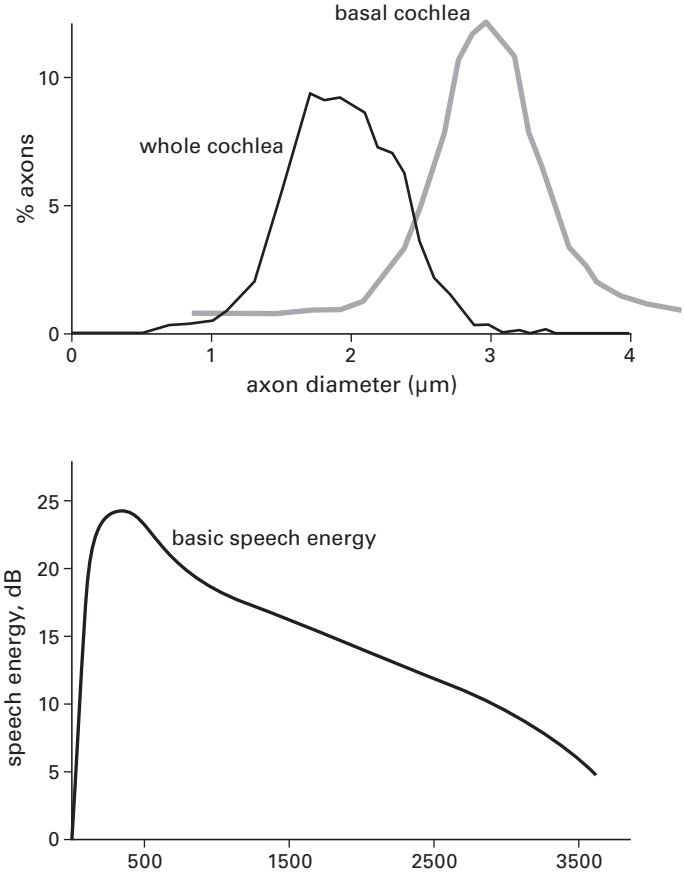


Figure 4.8

Speech uses lower frequencies and thus finer axons. Upper: Axons from the high-frequency end of cochlea (basal) are thicker and cost more space and energy than axons from the low-frequency end. **Lower**: Human speech occupies mostly frequencies below 500 Hz—the cheaper end. Upper, reprinted with permission from Perge et al., 2012); lower, after Freeman (1999).

helps communal life, made difficult by large brains, to be at least tolerable and occasionally joyous (Chanda & Levitin, 2013).

A sensor array must be fine enough to resolve the details that are critical to its task. For example, human vision resolves a spatial pattern of 60 cycles per degree, and this requires 120 cones per degree (Nyquist's rule). In two dimensions this amounts to 200,000 cones mm^{-2} (Packer et al., 1989). Again, this is just the down payment—for to *preserve* this spatial resolution, the communication line from each cone must remain separate all the way up to the visual cortex. All design must foresee the subsequent costs.

The general solution is to sample densely with a small part of the array and more sparsely with the rest. Therefore, our retina packs half of all its cones densely in a tiny patch (*fovea*), which occupies only 1% of the retinal surface. In this design the visual cortex devotes half of its volume to processing what the fovea delivers—thus allowing a fine analysis without unacceptably expanding the cortex.

For this strategy to work, it is often necessary to make the sampling array mobile—so that it can be trained on any feature of potential importance. Therefore, a fovea requires a system of muscles to move the eye, plus a control system to direct its constant exploration, and a higher-level system to select an object to be tracked. The effect is to stabilize the object on the fovea, allowing it to be sampled at high spatial resolution.[13] Stabilization confers an additional economy: it reduces the range of temporal frequencies on the fovea, allowing foveal neurons (and their subsequent processors) to operate at lower information rates, that is, on the steep segment of the rate-versus-cost curve for space and energy (figure 3.6).

This strategy also works for the tactile sense—dense distributions of sensors to fingertips, lips, and tongue—and explains the distorted *homunculus* in maps of human cortex, also the *barrel fields* in mouse cortex that represent the whiskers (Pammer et al., 2013) and the bizarre countenance of the star-nosed mole (figure 4.9).

Motorizing the sensors

The strategic choice of a fine, mobile sampler raises two other design issues: first, how to point the sensor where it is needed and, second, how to tell the brain that the sensor is *being* pointed. Both design issues require a dedicated part, the *superior colliculus* (figure 4.4).

The mechanism that chooses where to point the sensor needs visual input. When a retinal region outside the fovea senses a moving object, retinal signals drive a motor mechanism to smartly move the fovea onto that object and track it. The superior colliculus does this efficiently by placing a


1cm

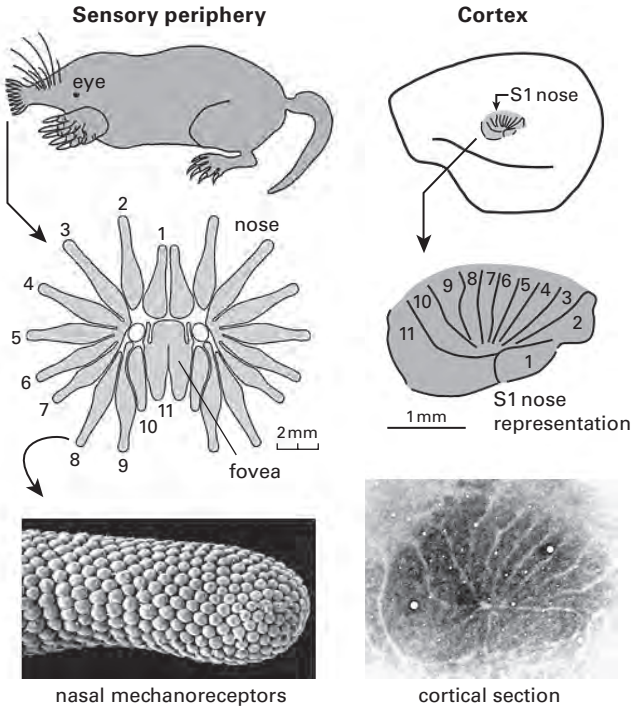

Sensory periphery
Cortex
eye
nose
fovea
2mm
S1 nose
1mm
S1 nose
representation
nasal mechanoreceptors
cortical section
Behavior
saccade
prey item
first contact
after saccade to nose fovea

Figure 4.9
Design of sampling arrays. Fine sampling required for spatial acuity requires large areas of cortex. Shown here is the mechanosensory system of the star-nosed mole. **Upper:** Frontal view shows tip of nose surrounded by 22 fleshy appendages. **Middle left**: Each nostril surrounded by 11 appendages, all covered by mechanoreceptors. No. 11 bears the densest distribution of receptors and thus serves as a mechanosensory fovea. **Middle right**: Each appendage is represented separately in somatosensory cortex (S1), with no. 11 occupying the greatest area. **Lower:** When a lateral appendage contacts an object of interest, the nose shifts to touch it with no. 11, the foveal appendage. Reprinted with permission from Sachdev and Catania (2002).

retinal map in register with a motor map so that each retinal point, using extremely short axons (~0.1 mm) can excite the corresponding point in the motor map and drive the eyes toward that location. Other senses also couple to the same motor map so that any of them—a flash, a bang, a slap—can announce which region of space needs the brain's immediate attention.[14]

Of course, we also attend to milder stimuli that match some stored pattern, especially when aroused by an internal signal of desire (food, sex). So the collicular mechanism for orienting the sensors needs to be informed by many issues. Decisions regarding where to look are made at the cortical level, which requires the cerebral cortex to communicate with the superior colliculus.

The upper collicular layers receive visual patterns and relay upward for further processing by cortical areas concerned with motion, and they receive signals from the same areas (Berman & Wurtz, 2010). The deeper collicular layers collect signals from the highest executive levels—frontal and parietal cortex—which convey a highly informed decision regarding where to look. The computations needed to reach that decision are extensive, involving much of the brain. But the decision can be relayed to the superior colliculus via a rather modest tract that requires only 6% of the corticocollicular pathway (Collins et al., 2005).

In short, the deeper layers of the colliculus know *where* to direct the eyes—that circuit is hardwired between the motor map and the low-level pattern generators that coordinate eye muscles. The deeper layers learn *whether* to move the eyes and *when*, by integrating raw-ish sensory inputs[15] with processed signals descending from cortex. The integrated output delivers instructions regarding vector and timing to pattern generators in brainstem that micromanage eye movements, and to those in the upper spinal cord that micromanage head movements. Thus, the descending collicular tract resembles various other tracts, such as fornix, hypothalamic, optic, and corticospinal, in being organized to send minimal instructions.

The motor stratum of the superior colliculus represents an intermediate-level pattern generator. It is tweaked by succinct executive decisions from above and delivers succinct instructions to low-level pattern generators. But it must also fulfill one more responsibility—to inform higher levels that its order: "Look!" has been sent.[16] This signal, termed *corollary discharge*, informs frontal cortex that the sensor is being repositioned. Why is this signal needed?

Corollary Discharge

When the retina is swept passively across a scene, the scene appears to move. The reader can confirm this by closing one eye and jiggling the other with a forefinger (gently!). However, when the superior colliculus *orders* the eye to sweep actively, the scene appears stable. The trick to stabilizing the scene when the brain moves the eye is to relay the order: "move!" to brain regions where the smaller patterns have finally been assembled into large, coherent patterns—corresponding to integrated perceptions. These areas, lying anteriorly in the parietal and prefrontal cortex (*frontal eye field*), know where the eye is looking—but they also need to know where the eye is *about to look*, so that they can compensate in advance before the motion occurs. This prediction, by allowing compensation, stablilizes perception—when we move our eyes, the world appears to remain stationary, as it should (Sommer & Wurtz, 2008; Wurtz et al., 2011).

The anterior frontal cortex is as far away from the superior colliculus as it is possible to be, so one might wonder why spend so much wire? One reason is that large patterns are assembled step-wise by cortical areas that press ever forward (chapter 12). By the stage where behaviorally relevant patterns have been assembled, compared to stores, and readied for use in selecting an action, the anterior frontal lobe is pretty much the last bit of available real estate. Moreover, because this cortical region decides where to look, it is precisely the site that needs corollary discharge to compensate for self-motion.

Another reason to control eye movements from the anterior frontal lobe is that, beyond their aid to sensing, eye movements also serve social communication. When someone looks us in the eye (or fails to), we notice. Even a dog notices and becomes aggressive when stared down by an unfamiliar human. Thus, as the cortical areas for social communication expand in the frontal and temporal lobes (chapter 12), they require a mechanism for sending executive summaries down to the superior colliculus. So design again economizes by using long pathways to send modest messages: "Look here!" or "*Don't* look here!"—skipping long, expensive explanations.

In summary, to build efficient sensors, the brain makes them mobile. It also compensates for self-induced motion, targeting the highest levels where choices and actions are being selected. These high-level mechanisms can then efficiently direct the low-level circuits that generate stereotyped patterns of movement. This *motif* drives orienting movements: the primate's eyes, the cat's external ear, and the rodent's whiskers and sniffing. These circuits use modest tracts to govern low-level pattern generators located near the relevant motor neuron clusters. This is the same motif that regulates internal systems and behavior.

Processing and storage of input patterns

Patterned inputs encounter the same constraints as patterned outputs, and to economize, they follow the same principles. First, the inputs deliver what can be computed locally; second, they relay upward only what is needed to assemble larger patterns. Each successive stage of processing sheds unneeded information. These principles also apply to storage: save only what is needed, for as long as it is needed, and in the most compact form.

Compute locally

Economy begins with sensory transduction. Because sending information at high rates costs more (figure 3.6), sensors use separate lines for different rates. For example, certain mechanosensors in the skin are wrapped in an onion-like capsule that filters out slow changes and delivers the fast ones to a mechanosensitive cation channel in the nerve terminal at the onion's core (figure 10.3). Other types with different capsules locate at different depths within the skin to help filter out the fast changes and capture slower ones. Skin sensors of temperature, noxious pressure, and noxious chemicals operate still more slowly—which allows still lower spike rates and finer axons. Consequently, the distribution of fiber diameters from sensory nerves resembles that of central tracts: many fine fibers and fewer thick ones.

Exemplifying the rule, *compute locally,* are two types of pressure receptor located on the foot. Each demands a prompt behavioral response without waiting 200 ms and expending more wire to consult higher processors. The responses are opposite: one extends the limb to support the body; the other flexes the limb to remove it from contact with the ground.

For example, pressing your bare foot on a smooth surface activates an array of low-frequency pressure receptors that excites the pattern generator

for limb extension to support your weight. But pressing your foot on a sharp point activates higher frequency pressure receptors that excite the pattern generator for limb flexion to withdraw your weight and for limb extension on the opposite side to support your weight. This occurs faster than you can *feel* "Ouch!" because the higher frequency pressure responses travel over thick, fast-conducting wires that couple directly to the local pattern generator (figure 4.1).

Such direct functional connections between specific sensory inputs and specific motor outputs were historically termed *reflexes* (Sherrington, 1906). By now the design is seen as coupling each receptor type to the appropriate pattern generator. This design saves time, wire . . . and grief.

Relay to cortex

The small pattern carried by a single sensory axon resembles a piece of jigsaw puzzle to be assembled with other pieces into a larger pattern of sufficient quality for comparison to stored patterns. Assembly is a task for the cerebral cortex, but to reach that level, input arrays require serial "preprocessing" to reduce firing rates by stripping away redundancy and unneeded information. This requires that slow and fast signal components that were transduced separately maintain their separation via *parallel pathways* all the way to cortex. Thus, skin sensors signaling pain and temperature with low mean rates are processed by one set of circuits near their entry points (spinal cord and lower brainstem) whereas sensors signaling joint angle, muscle length, and whisker deflection with high mean rates are processed by different circuits[17] in lower brainstem.

For most sensors the spike rates are still too high for direct relay to cortex, so a central integrator (*thalamus*) is interposed to concentrate the message, that is, more bits per spike (figure 3.5C). This allows a two- to fourfold reduction in mean spike rate on the path to cortex. The thalamus is also used by other brain regions, such as cerebellum, *striatum*, and superior colliculus, for the same function (Bartlett & Wang, 2011; Sommer & Wurtz, 2004).[18] The computational strategy and synaptic mechanisms to achieve this function are described in chapter 12. The exceptions to this design are the olfactory sensors which signal at such low rates that, following a single stage of preprocessing in the *olfactory bulb*, they are allowed to skip the thalamic relay and ascend directly to cortex (Friedrich & Laurent, 2001).

Cortex finds larger patterns

The task of sensory cortex is to rapidly capture correlations of higher order from the array of local correlations relayed from thalamus. This proceeds by

stages, first across layers of each primary area (*V1, S1, A1*) and then across successive areas, until single neurons eventually report patterns of clear behavioral relevance that identify an object by sight, touch, or sound (figure 12.11). Such patterns emerge in specialized patches where most neurons respond only to that pattern and not to the fragments that comprise it, thus an area for faces, objects, scenes, and so on (chapter 12).

A reader might worry that the world's infinity of categories would require a corresponding infinity of cortical areas, but actually, the number only needs to match categories that matter most deeply to the animal. Smaller brains operate with fewer categories, so the whole mouse cortex divides into about 20 areas, whereas human cortex has about 200 (Kaas, 2008). As areas attain higher levels of abstraction, each contains less information and thus requires less space. So the early cortical areas, which first process thalamic input, are large, whereas later areas for high-order patterns are small (figure 12.11)

This design—many small areas operating in parallel—continues the principles of economy. Resources can be assigned according to what matters most to the animal. Processing can proceed at lowest acceptable rates and at lowest acceptable spatial resolution. For example, an *object-grasp area* that needs only coarse patterns can download them at an earlier stage than an *object identification area* that needs more detail (Srivastava et al., 2009; Fattori et al., 2012). Wire is saved by locating areas that assemble the patterns near to the areas that use them (chapters 12 and 13). For example, face areas locate anteriorly in the temporal lobe on the path toward areas that evaluate facial expression. An object-grasp area locates posteriorly in the parietal lobe—on the path toward motor cortex that guides grasping. Thus, the overall processing scheme for cortex reflects the three design principles seen at lower levels: send only what's needed; send slowly as possible; minimize wire.

Storing Patterns

To store small, evanescent patterns encoded by an array of thalamic neurons, would be costly. If patterns were all stored at this level, high-level images could in principle be reconstructed. However, with the optic nerve delivering 10 Mbit s^{-1} to the thalamus, storage needs would soon exceed any conceivable capacity. Moreover, if data were stored raw, it could only be filed by order of arrival—so to retrieve images from stored fragments would be a computational nightmare and impractically slow. So an animal should store high-level patterns and only *particular* ones that can improve future behaviors.

Each species stores patterns critical for its economic strategies. For example, a nutcracker jay living at high altitude caches nuts at numerous sites in autumn and descends to a valley for winter. Returning in spring, it recalls myriad cache locations to sustain itself until the summer brings fresh groceries. For humans, what matters most is our ability to rapidly recall a face, along with any historical significance that we can attach to the face we are facing. This allows the best chances for selecting an appropriate behavior.

Yet we must not store every face encountered on a stroll through the park—only ones likely to prove significant. So a potentially important face needs to be tagged—cognitively and affectively—and then filed. Upon reencounter, the original image is retrieved and held in "working memory" for comparison to the current image. These various processes require cooperation between several neural structures. The main cortical face area connects with the amygdala, which "stamps" the image from its catalog of innate emotional expressions. To further annotate the image, the striatal system for reward prediction connects to the face area via a long loop and to the amygdala (Middleton & Strick, 1996). Then, they all connect to sites for working memory and behavioral choice in prefrontal cortex.

These organs for pattern recognition, storage, evaluation, and behavioral choice interconnect strongly; therefore, by locating near each other, wire is reduced. Their location anteriorly in temporal and frontal lobes is no mystery: the posterior regions are already occupied by areas concerned with pattern assembly. Thus, in mammals where higher degrees of sociality require the brain to enlarge, the expansion occurs disproportionately in anterior regions for cognition and emotional expression (Dunbar & Shultz, 2007). Thus, although human and macaque collect similar amounts of sensory information (e.g., their retinas are nearly identical),[19] humans greatly expand the number and size of cortical areas for assembling the higher order patterns. This occurs especially in forward regions that include amygdala, prefrontal cortex, and hippocampus.

Correcting errors

Evaluating behavior: Two kinds of prediction error

The parts of the motor system that directly generate and distribute final output patterns (behavior) require only a small fraction of total brain volume. However, the adjective "final" is slightly misleading. Each motor act is also a beginning: it is a provisional answer to some predicted need. Since needs recur, output patterns might be improved if their effectiveness could

be evaluated. Therefore, the brain invests heavily in several systems for evaluation and error correction.

One system asks, "How precisely did the actual output pattern match the intended pattern?" This system computes the difference between the intended pattern and the actual pattern; then it feeds the error back to command structures that gradually improve performance. This serves *motor learning*—what is gained from practicing the piano or the golf swing. Mindful repetition improves speed and accuracy—and also efficiency—since a motion that begins awkwardly eventually gains grace and saves energy (Huang et al., 2012). This system also serves cognitive and affective processes: it compares intended cognitive and emotional patterns to what actually occur and then feeds back to improve subsequent performance. Thus, motor learning is subset of *intention learning*.

Another system asks, "Was the act, however well performed, worth the energy and the risk?" This system compares the expected payoff from a particular act to what was actually gained. The neural mechanism rewards a better outcome by releasing a pulse of dopamine at key brain sites and punishes a poorer outcome by reducing dopamine and enhancing other chemical signals. This is *reward-prediction learning*, and one can easily imagine its myriad ramifications. Reward-prediction learning evaluates every choice and thus charts the course of our lives: cereal or toast; law or medicine; choice of mate, friends, and retirement fund (chapter 14).

Intention learning and reward-prediction learning employ different brain structures, and both are large (Doya, 2000). The organ for intention learning is the cerebellum, and the organ for reward-prediction learning is the striatum (figure 4.10). Neither structure directly modulates the final output: they do not send wires to the low-level pattern generators. Rather, they return error signals to particular high-level organizers of behavior. For example, the cerebellar region that serves motor learning (*anterior lobe*) returns its updating signal to motor cortex. Cerebellar regions that serve perceptual, cognitive, and affective learning return their updates to cortical areas for pattern recognition in temporal and parietal cortex and to areas for behavioral choice, such as prefrontal cortex (Strick et al., 2009; Schmahmann & Pandya, 2008).

Cerebellar and striatal output tracts both use high spike rates that require thick axons. In fact, the striatum is named for striations due to bundles of thick, myelinated axons (figure 4.10). High spike rates should be reduced before the messages are broadcast. Both circuits do this, as noted, via a thalamic relay. Cerebellar and striatal design will be considered further in chapter 13.

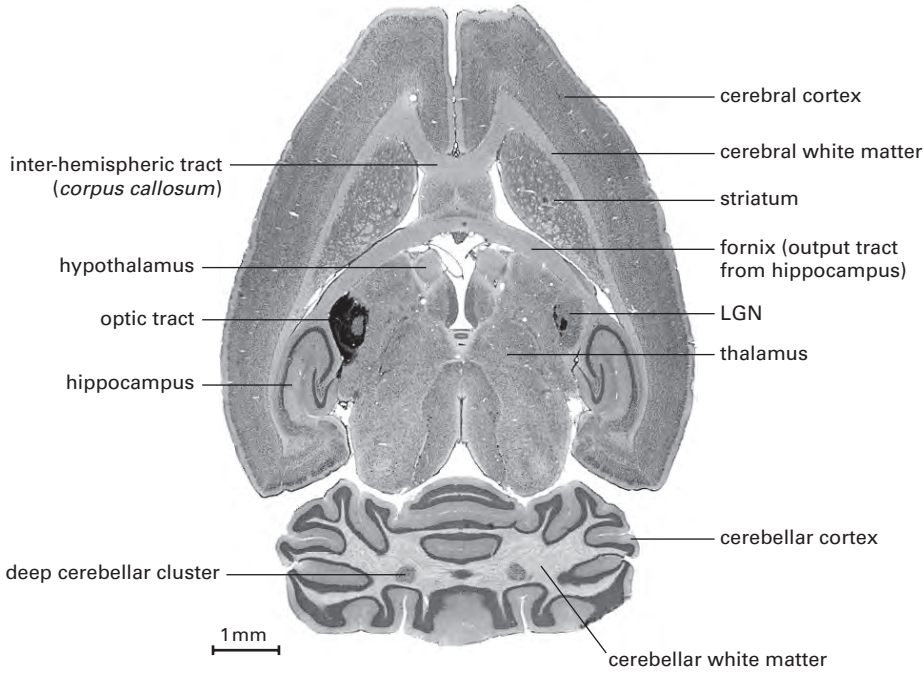


Figure 4.10
Rat brain in horizontal section. Note that striatum lies nearest to the anterior cerebral cortex. Striatum contains dense bundles of myelinated axons (pale) whose large caliber reflects their high spike rates. Note also the deep cerebellar clusters which reduce the number of high-rate axons before projecting to thalamus where rates are reduced before relay to cerebral cortex. Left optic tract is dark because a protein tracer injected into the eye was taken up by ganglion cells and transported inside their axons to the brain. Tracer is visualized here by a specific chemical reaction. Image courtesy of H. J. Karten and reprinted with permission; © The Regents of the University of California, Davis campus, 2014.

Conclusions regarding organization of mammal brain

This chapter has sketched how three principles (*send only what is needed*; *send at the lowest acceptable rate*; *minimize wire*) shape brain design to accomplish its seven broad tasks (see figure 3.2). The layout explained here extends upward from a scale of millimeters. It does not explain design of local circuits that analyze and integrate input patterns or generate output patterns. Those compute on a scale of nanometers to micrometers and are topics for chapters 7–11. This chapter also does not explain the brain's

striking structural diversity on the scale of micrometers to millimeters, such as the different structures for cerebellar versus cerebral cortex and the specialized substructures of cerebral cortex. These are treated in chapter 13.

Insect Brain

We consider now the insect brain, emphasizing *Drosophila*, because of its importance for genetic analysis—like mouse. But we also include other insects, such as locust, wasp, cricket, and bee that share various broad features of somatic and neural design and are profoundly specialized for particular lifestyles and habitats (Burrows, 1996; Strausfeld, 2012). Just as we referred to "mammalian" brain in preceding sections, we will refer to "insect" brain in this section.[20]

The first point is that the insect brain needs to accomplish the same basic tasks as the mammalian brain (figure 3.2). Second, it is governed by the identical constraints: the law of diminishing returns (figure 3.6), plus the need to minimize wire (figure 4.1). Third, the insect brain also predictively regulates the internal environment and efficiently couples the internal organs (see figure 3.4). Finally, the insect brain couples the inner and outer worlds (figures 3.2 and 4.3) and, following the scythe of Saturn, encounters the same types of information, which it must analyze and integrate to satisfy similar behavioral demands. So we should expect similarities of macro-organization. Indeed, they are numerous and striking (figure 4.11).

The insect brain, like the mammalian, is organized into defined neural clusters with locally dense connections plus distinct tracts for more distant connections (Chiang et al., 2011). Brain outputs include a rich system for wireless signaling, starting with two neurosecretory bodies at the back of the brain (*corpora cardiaca* and *corpora allata*) whose neurons secrete neuromodulators and hormones into the circulation (analogous to the hypothalamic neurosecretory clusters). These neuromodulators and hormones, which include over 50 neuropeptides, govern the insect's internal milieu by acting on energy metabolism, salt and water balance, growth/molting, and reproduction. Autonomic neurons cooperate with these hormones to coordinate visceral function with behavior (Cognigni et al., 2011). For example, gut neurons interact with hormones to increase intestinal throughput to fuel egg production and also to control appetite. These concerted actions of wireless and slow wire processes in insects resemble those of the vertebrate hypothalamo-pituitary and autonomic systems, and there appears to be a common evolutionary origin (Arendt, 2008).

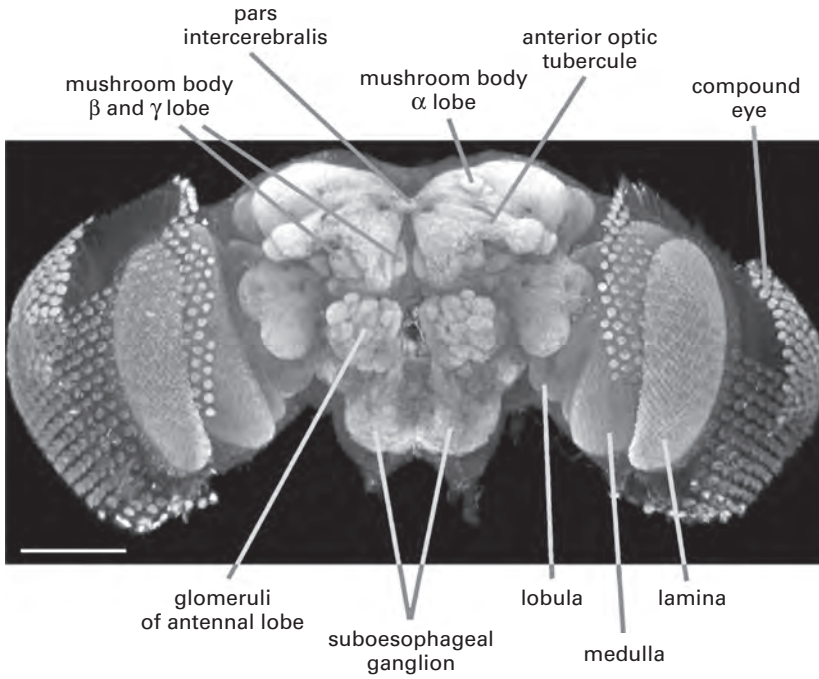


Figure 4.11

Frontal view of fly brain shows prominent areas devoted to specific functions. *Vision:* Compound eye, hexagonal array of optical sampling units passes information sequentially to lamina—collect and sort inputs, medulla—detect local patterns, lobula (and lobula plate not seen in this view)—assemble small patterns into larger patterns, anterior optic tubercule—associate larger patterns. *Olfaction:* Glomeruli in antennal lobe—collect and sort inputs and project to mushroom bodies, which identify patterns. *Learning*: mushroom bodies—integrate diverse information, learn patterns and associate with punishment and reward. *Integration:* Pars intercerebralis connects two sides of brain. *Distribution:* Suboesophageal ganglion—integrate information for wired and wireless output to body. View of a three-dimensional reconstruction of a *Drosophila* brain stained with antibody for synapses to show areas where processing takes place. Image courtesy of Ian Meinertzhagen. Reconstruction can be rotated and viewed from different angles at http://flybrain.neurobio.arizona.edu/Flybrain/html/contrib/1997/sun97a/.

Insect brain uses these systems to coordinate visceral, behavioral, and immune responses to stress, but instead of the vertebrate's adrenalin (*epinephrine*), insects use octopamine (Verlinden et al., 2010). Thus, during emergencies, "fight or flight," octopaminergic neurons raise octopamine concentration in hemolymph (like adrenalin in vertebrate blood), which acts broadly on endocrine cells and fat body (similarities with vertebrate liver) to mobilize energy reserves, on muscle to increase power, and on sensory receptors and circuits to increase sensitivity and response speed. Octopamine neurons also directly contact endocrine glands, heart, muscle, and certain brain regions for specific purposes. For example, in locust 40 identified neurons (*DUM*) innervate flight muscles to regulate fuel supply (Burrows, 1996). At rest the neuron fires steadily, maintaining the supply of "fast burning" sugars needed for takeoff. During steady flight the DUM is silenced, and energy supply switches to the larger reserves of slower burning fats. The DUM's low mean firing rate, 0.5–1 Hz, resembles mammalian autonomic nerves.

Insect brains also have clocks set by light—indeed the molecular mechanism of animal clocks was first determined in *Drosophila* (Weiner, 1999). *Drosophila*'s roughly 150 clock neurons form a distributed system that governs catabolic/anabolic phases, including a sleep phase for the consolidation of neural processing (Allada & Chung, 2010; Crocker & Sehgal, 2010). Some clock neurons form small clusters, mini-SCNs, that collect specific entraining inputs from the compound eye, the simple eyes (*ocelli*), and a pair of photoreceptor cells within the brain. Other clock neurons express their own photopigment and so can collect photons through translucent cuticle. Thus, the fly's clocks locate anywhere they are needed. We speculate that this distributed design saves wire in a small brain.

Collecting patterns

Investment in sensors to collect patterns is strongly tuned to social and economic strategies. *Drosophila*'s compound eye is relatively small, and the photoreceptors gather information at a low rate—good enough for hovering over decaying fruit. However, *Coenosia*, a close relative with similar body size, is an aerial predator and, to resolve and track its prey, requires a threefold larger eye and photoreceptors with fourfold higher bit rates (Gonzalez-Bellido et al., 2011). In accordance with the law of diminishing returns for photoreceptors, *Coenosia*'s high-rate eye costs more space and energy per bit (chapter 8).

To identify rotting fruit and detect *pheromones* (secreted chemical factors that trigger social responses), *Drosophila* invests in about 50 types of

olfactory receptor. These are more than are used by the louse that parasitizes humans (10) but fewer than are used by the honeybee (160) and fire ant (400) for their extensive foraging and chemical communication. Certain insects also invest in mechanical apparatus to improve their efficiency at pheromone detection. For example, male moths commonly use broad antennae as molecular sieves, which they push through the air to trap molecules of female attractant.

Meanwhile, *Drosophila*'s antennae specialize to register not the aroma of courtship but its music. Both sexes sing to each other. The vibrations, reaching 500 Hz, are received via the antenna and transmitted to its base to activate about 500 mechanosensors (*Johnston's organ*). These are equipped with mechanical feedback to boost the gain, like hair cells in the mammalian cochlea, to operate near the sensitivity limit set by Brownian noise (Immonen & Ritchie, 2011).

Moths are hunted by bats using echolocation. So the moth invests in a pair of simple ears, each with only one or two sensors, and couples their outputs to a simple pattern generator for evasive flight. When the sensors detect a bat's ultrasonic chirp, evasive flight is engaged, and the moth dives to the ground (Roeder, 1967). This system provides a cheap answer to the bat's high-tech, super-expensive sonar.

Insect sensor arrays, like mammalian sensor arrays, are subject to the sampling theorem (Nyquist's rule). To achieve high resolution at acceptable cost, they too combine broad, coarse sampling with local, fine sampling—both in space and time. For example, a male housefly pursuing an evasive female at high angular velocities is aided by his visual *lovespot*. The forward-facing photoreceptors pack especially densely to improve spatial resolution, and they produce especially fast electrical responses to improve temporal resolution—both needed to track the speedy female (Burton & Laughlin, 2003). But the lovespot, like a mammalian fovea, must not be too broad, because it is expensive, so the fly uses the same solution: motorize the sensor. During pursuit, a dedicated tracking system controls head and body movements to keep the lovespot centered on the target.

In short, insects invest in sensors according to need and locate the sensors where they will be most useful: olfactory and auditory sensors on the antennae that project into the air stream; auditory sensors on crickets' forelegs to space them widely (thereby improving sound localization), taste sensors on the landing gear (feet), mechanosensors on the wing. Each sensory system is used to inform the others; for example, an odor that attracts *Drosophila* increases the accuracy with which its visual system guides its flight (Chow et al., 2011)—cross-modal interactions that are also used by mammals (Burge et al., 2010).

Processing and storage

Insect sensory processing resembles mammalian processing in that small patterns collected by sensors are filtered and then assembled into larger patterns. To assemble visual patterns, the fly identifies spatial and temporal correlations via successive neural layers (figure 4.12). First, the lamina sums correlated inputs and removes redundancy associated with the level of illumination. Then the medulla identifies local features which the next layers (*lobula* and *lobula plate*) use to detect larger and more complicated patterns. Then their outputs distribute to various smaller regions (*optic glomeruli*) where they are processed before projecting forward to integrative centers in the *protocerebrum*. Each optic glomerulus collects inputs from a particular ensemble of neurons in the lobula, suggesting that higher order patterns are being segregated.

The architecture of the fly visual system resembles in several respects that of mammal. The fly preserves spatial continuity of the retinal image by mapping the output from one layer, point by point, onto the next layer—across the many stages of processing. However, at the final stage, the optic glomeruli abandon retinotopic organization, thus shedding "where" information while sorting out "what," reminiscent of the *ventral stream* of the mammalian cortical pathway (chapter 12).

The layers and maps of vision's earlier stages are computationally efficient because all parts of an object represented in the retinal image are spatially and temporally continuous. These properties of the input allow local features (local motion, local edges) to be extracted and mapped at the lower levels and then assembled at higher levels to define objects and scenes. Extracting all local features first, as with insect medulla and mammalian visual cortex, provides a communal data set to be shared by various higher order mechanisms, and this conserves space and energy. Local processing, mapping, and the orderly projections from each layer to the next also save wire, as do the orderly maps of different modalities within a tract (Niu et al., 2013; chapter 13).

Despite an efficient architecture, visual processing for form, motion, and color is computationally demanding. The visual system uses 70% of the fly's neurons, of which most are in the medulla, which extracts local features using about 150 different types of identified neuron. Thirty-five types, replicated in each of the medulla's 800 retinotopic columns, interrogate the image for local features. In this respect the fly's medulla is analogous to the mammal's primary visual cortex, also the largest visual area (chapter 12).

The olfactory system is structured differently (figure 4.12). Whereas vision assembles patterns stepwise across four layers, olfaction uses just two (Masse et al., 2009). The first layer (*antennal lobe*) collects input from 45

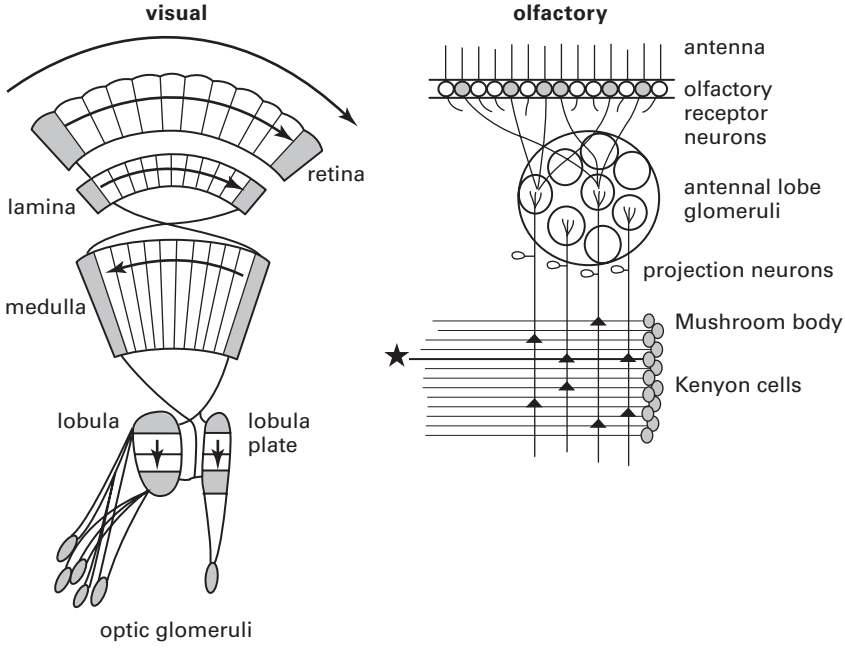


Figure 4.12

The visual system is deep and maps spatial position. The olfactory system is shallow and processes globally, without reference to spatial position. Left: Fly visual system processes retinal image in four successive layers. Lamina assembles and sums correlated inputs and reduces redundancy (chapter 9); medulla extracts local features; lobula and lobula plate assemble larger patterns (lobula—color, form and motion; lobula plate—motion). The first three layers map retinal image (arrow) across columns of neurons. The last layer, optic glomeruli, does not map, it generalizes. Each glomerulus collects from all neurons coding the same pattern, irrespective of spatial position. **Right:** Fly olfactory system processes information in just two layers. First the antennal lobe assembles and sums correlated inputs from receptor neurons. A glomerulus collects from neurons with same olfactory receptor and filters to reduce redundancy. Then 2500 Kenyon cells in mushroom body extract from all 45 glomeruli the patterns that define odors. Each Kenyon cell associates synaptic input (triangles) from small subset of 10 glomeruli to form an efficient sparse code (2 associated synapses shown on starred Kenyon cell). Diagrams simplified and not to scale. Visual based on Strausfeld (2012); olfactory based on Masse et al. (2009).

types of olfactory receptor on the antenna, collects each type in a separate synaptic glomerulus, sums these correlated inputs to reduce noise and filters to reduce redundancy. The results are relayed to the second processing layer, residing in the mushroom body, which is the insect's seat of learning (see below). The second-stage neurons compare all 45 olfactory inputs and learn by association the unique patterns of glomerular input that define particular odors. The mammalian olfactory system employs a remarkably similar structure (Wilson & Mainen, 2006). It uses an olfactory bulb with glomeruli, one for each receptor type, and after filtering, it projects straight to cortex for association and learning.

Two-stage processing works for olfaction because, unlike vision, there are no local features. The molecule or mixture that characterizes an odor arrives in a volume of air for a certain time, but there are no higher order spatial correlations to help identify it. The correlations that identify an odor are distributed across receptors: each type binds a spectrum of molecular species, each with different affinity. Thus, an odorant, whether from a single molecular species or a mixture, activates several receptor types to different degrees, to produce a correlated pattern of receptor activations—and that defines an odor.

The pattern from the array of glomeruli transfers to the mushroom body (Laurent, 2002; figure 4.12). Because each odorant stimulates several receptors, and each receptor contributes to the coding of many odorants, the mushroom body's task is to find correlations across receptor inputs—the pattern that defines a particular odor. When a new and significant odor is encountered, the new pattern is learned. To optimize the number of different patterns that can be represented by the mushroom body's 2500 Kenyon cells, the information is coded sparsely with few spikes (Jortner et al., 2007).

In short, there are profound differences across sensing systems within an animal, and profound similarities for a given sensing system across animals (insect vs. mammal). Olfactory and visual designs differ because the small patterns that they collect present different statistics and thus require different processing. Olfactory designs are similar because the input statistics for insect and mammal are the same and thus require similar processing. The same goes for visual designs.

Nonetheless, insect and mammalian designs are not identical, probably because they are differently constrained. For example, the fly visual system lacks a thalamus, which the mammal needs to reduce spike rates. Many fly visual neurons connect centrally over distances less than 0.5 mm, which means that signals can travel passively in graded (analogue) form. This saves space and energy in two ways: analogue can transmit high

information rates cheaply (chapter 5), and can avoid costly analogue → pulsatile and pulsatile → analogue conversions. Thus the insect brain uses a more efficient design that cannot be implemented in a larger brain.

Assembling patterns and choosing an action

A fly assesses its current state from sensory patterns, compares this state to stored patterns to learn how its state is changing, and adjusts behavior accordingly. For example, it may steer flight to maintain a constant bearing with respect to the sun or change course to approach a rewarding object or avoid an aversive one. The *central complex*, a compact modular structure strategically placed deep in the brain, plays a pivotal role in these processes of assessment, decision, and direction (Strausfeld, 2012; Strauss et al., 2011).

The central complex links sensory patterns to motor commands within a framework of body orientation (figure 4.13). Its three largest structures, the *protocerebral bridge*, the *fan-shaped body*, and the *ellipsoid body* are linear arrays of neural modules that map the angle of azimuth (compass bearings on a horizontal plane) around the fly. The 16 modules of the protocerebral bridge map 16 sectors, eight on the fly's left and eight on its right (figure 4.13), and project to eight modules in the fan-shaped body. Each fan-shaped body module accepts input from a protocerebral bridge module on the left side, and from its opposite number on the right side. This convergence establishes eight horizontal axes that pass through the center of the fly.

The eight fan-shaped body modules then connect straight to the eight modules of the ellipsoid body which, in turn, connect to the lateral accessory lobes. Here the outputs from the central complex contact the descending neurons that drive motor pattern generators in the segmental ganglia. In short, by explicitly linking signals to azimuthal bearings (horizontal lines of sight from the fly's cockpit), the central body relates the position of a sensory pattern to the body's orientation and direction of movement.

Information on sensory patterns and stored patterns project across the directional modules via *horizontal neurons*. Some horizontal neurons establish memory traces, and this allows generalization. Information gathered from a pattern observed in one direction is distributed so that an object learned in one location can be recalled in another. The horizontal projections stratify the fan-shaped body, and two of its layers have been linked to specific components of visual patterns: one layer to the orientation of visual contours and the other to the elevation of an object above the horizon.

Some patterns processed by the central complex are used for navigation. The central complex serves as a sky compass that enables locusts and monarch butterflies to fly on a constant bearing by maintaining the body at a

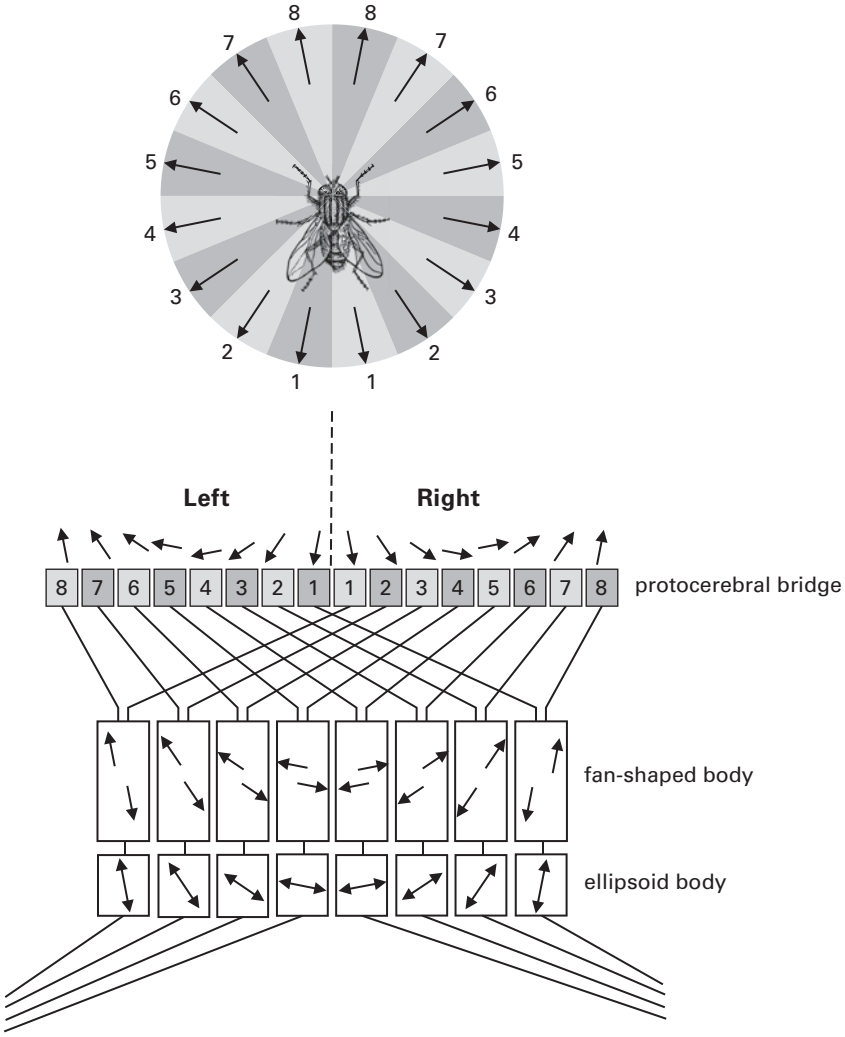


Figure 4.13

Central complex maps horizontal lines of sight. Protocerebral bridge's 16 modules map 16 sectors viewed from head, 8 on insect's left and 8 on its right. Projection to fan-shaped body's 8 modules connects opposite sectors (e.g. Left 1 and Right 8) to establish and map axes that pass through centre of head. This map is projected to ellipsoid body's 8 modules, for output to neurons that select and control motor patterns. The central complex then sends information about position of stimuli with respect to the head to neurons that control body orientation and the direction of locomotion. Figure based on Strausfeld (2012) and Strauss et al. (2011). Fly image from http://openclipart.org/image/800px/svg_to_png/120457/HouseFly2_.png.

given angle to the solar azimuth. When the sun is obscured, the pattern of polarized light in blue sky is used instead. For bees returning to the hive, and especially for monarch butterflies migrating 3,000 miles from Canada to Mexico, it is important to maintain the same true bearing (e.g., 185 degrees South Southwest) throughout the day. For this, the sky compass mechanism uses clock information to correct for the sun's movement, and neurons in the central complex are involved (Heinze & Reppert, 2011).

In short, the central complex is aptly named because it is both centrally located and central to the brain's broad tasks that were indicated in figure 3.2 (assemble larger patterns, compare to stored patterns, predict a promising output pattern, and call an integrated output). Thus in many ways the central complex is homologous to the mammal's basal ganglia (Strausfeld & Hirth, 2013). It seems remarkable that the central complex achieves all this with less than 600 neurons (592 at the last count). But how are output patterns implemented?

Distributing motor patterns

The insect brain places its motor neurons in the body segments where they are needed and drives their detailed firing sequences with pattern generators located at the same site (like the mammalian spinal cord). These final pattern generators are coordinated across segments (e.g., three pairs of legs) via fibers that connect to pattern generators in other segments. These are organized into complex behaviors which the brain can call or restrain via descending neurons. Most famously, for a male mantis to copulate, he needs only to shed a descending restraint—which occurs when an obliging female . . . bites off his head.

Significantly, though perhaps anticlimactically, the distribution of fiber diameters in the connecting tracts resembles the mammal: many fine axons and fewer thick ones (figure 4.14).

The activation or disinhibition of some rhythmic and stereotyped behaviors—for singing, mating, fighting, and so on—is controlled by small numbers of command neurons that activate dedicated networks (Hedwig, 2000). To an observer, these behaviors appear quite complex and plastic—for example, Google "drosophila aggression" and watch a YouTube film that resembles a professional boxing match. Complex behaviors can be evoked from larger insect brains by electrical stimulation of single command neurons—recalling the complex behaviors evoked with fine electrodes from the mammalian hypothalamus.

The insect brain, like the mammal, needs to distinguish activity created by its own motor commands from activity originating in the environment,

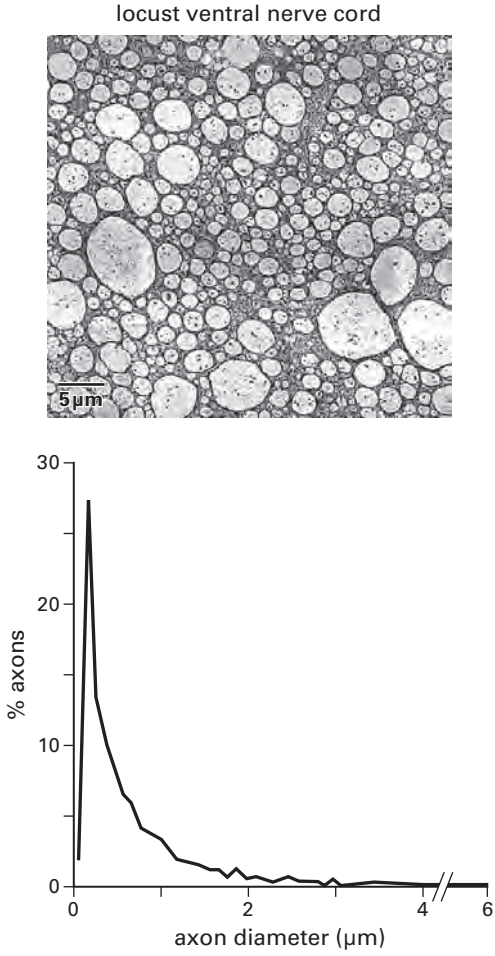


Figure 4.14

Distribution of fiber diameters in insect nerve cord. This distribution resembles many long pathways in mammalian brain. Reprinted with permission from Perge et al. (2012).

that is, it needs mechanisms for corollary discharge. For example, a cricket producing loud chirps risks desensitizing its own auditory system, which would prevent it from detecting softer external sounds (Poulet & Hedwig, 2007). To avoid desensitization, the small motor circuit that generates the chirp drives a single neuron that directly blocks inputs from the two ears (figure 4.15). This simple circuit shuts down auditory inputs for precisely the duration of a chirp, leaving the cricket free to listen for responses between chirps. This precise blanking-out of disruptive input resembles the suppression of visual inputs during a saccadic eye movement. The point here is that for most tasks that a mammalian brain needs to accomplish, so too must an insect brain. Moreover, the insect brain often uses similar strategies—but benefits from the smaller scale: fewer neurons and shorter distances (Chittka & Niven, 2009).

Correcting errors: Motor learning

The prominence of the cerebellum in mammalian brain might predict an obvious insect analogue, but there is no structure totally dedicated to motor learning. The suggestions are that motor learning is one of many tasks assigned to the mushroom bodies and to the central complex (Farris, 2011; Strauss et al., 2011). Indeed, with fewer body segments to coordinate, stiffer mechanics, and a body that is not continually growing, an insect arguably has less need for motor learning.

Nonetheless, some motor learning is essential. For example, flies improve their motor performance with practice (Wolf et al., 1992). Normally when a fly (or any animal) turns in one direction, the visual scene moves in the opposite direction. If this relationship between action and consequence is reversed by placing the fly in a flight simulator, the fly adjusts within 24 hours. Now when it wants to approach a promising target, it turns *away* from the target, and *voilá,* the target enters its field of view. This resembles Kohler's famous experiment: after students wore inverting spectacles for a day or two, the world appeared to be right-way up, but when they removed the spectacles, it appeared upside down. Why do flies need this motor learning? Motor learning is built into their flight control system to cope with changes of body mass (feeding, defecating, growth, and laying eggs) and damage to the wings.

Reward-prediction error

Insect brains are wired for associative learning and employ a system for computing reward-prediction error that follows the same basic learning rules as in mammals. The internal reward system uses dopamine and

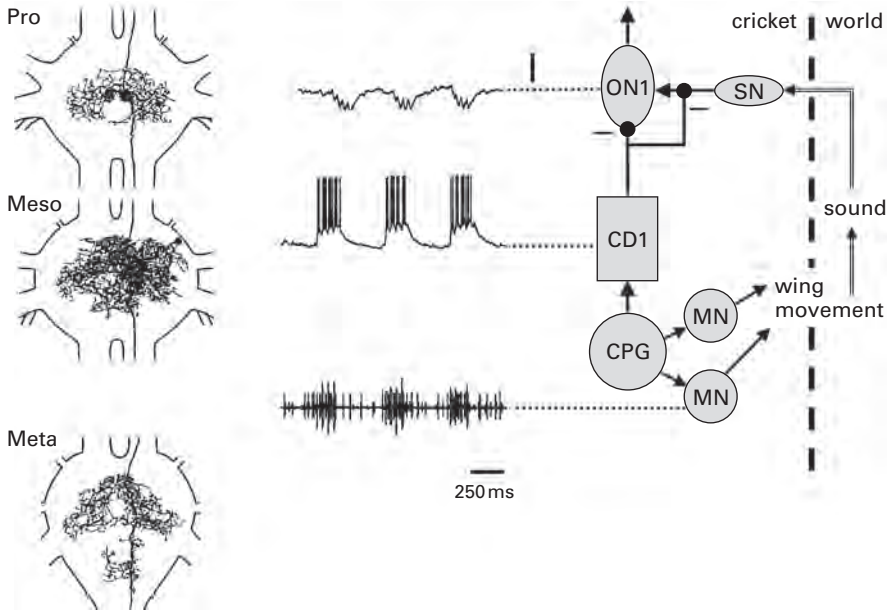


Figure 4.15

CD1, the neuron that prevents a cricket being deafened by its own chirp. Right: CD1's circuit. Central pattern generator (CPG) drives motononeurons (MN) rhythmically to produce wing movements that generate chirps. Each chirp excites sensory neurons (SN) in cricket's ear. SN output synapses excite omega neuron (ON), which conveys auditory information to brain. CPG also drives CD1, which inhibits ON1 and output synapses of SN, thereby blocking signal to brain while chirp is being produced. **Middle**: Recordings of signals within circuit. Bottom trace: Extracellular recording of spikes in MN, driven rhythmically by CPG. Middle trace: Intracellular recording from CD1. Excitatory synapses from CPG depolarize CD1 to produce bursts of spikes that follow CPG rhythm. Top trace: Intracellular recording from ON. Inhibitory synapses made by CD1 produce rhythmical bursts of IPSPs that block ON1 output during chirps. **Left:** Morphology of CD1 revealed by intracellular dye injection. Axon connects dendritic arbors in the three thoracic ganglia, meta-, meso-, and pro-. Mesothoracic dendrites receive excitatory synapses from CPG. Prothoracic dendrites make inhibitory synapses onto omega neuron, ON1, and all sensory neurons, SN. Vertical scale bar: 20 mV for CD1; 5 mV for ON1. Reproduced from Poulet and Hedwig (2006) with permission.

octopamine. The systems for computing reward-prediction error and for storing the lessons both reside in the mushroom body (figure 14.11).

The mushroom body, like mammalian cerebral cortex, participates in olfactory learning, associative learning, spatial learning, visual pattern recognition, attention, and sensory integration. The mushroom body, like cortex, shapes its circuit architecture to view multiple inputs, looking for coincidences to associate with reward or punishment. This suggests a multipurpose cross-correlator that can be wired to evaluate a variety of associations and store the lessons.

As with other computing devices, new models allow new opportunities. Primitive parasitic wasps (early model) use elaborate mushroom bodies to find and store the locations of grubs hidden at particular sites within a plant (Farris & Schulmeister, 2011). Social wasps (later model) use this capacity to recognize each colony member by its distinctive face and body markings and to store this information along with knowledge of its position in the dominance hierarchy (Sheehan & Tibbetts, 2011). Thus, the later model supports a complex social behavior that confers the benefits of communal foraging and the division of labor. Social insects, like social primates, build upon the low-level sensors, adding brain parts that enable social behavior. The parts that expand are those that recognize patterns, store them, and evaluate them via the system of reward prediction.

What a honeybee can do with a brain of 10^6 neurons seems prodigious. A bee learns to break camouflage, to navigate a maze via symbolic cues (blue, turn left; yellow, turn right), and to associate a flower with the time of day during which that particular species produces nectar. Bees can also perform delayed match-to-sample and symbolic match-to-sample tasks[21] that were thought, until recently, to be confined to monkeys, human, dolphin, and pigeon (Srinivasan, 2010; Menzel, 2012). In short, absolute numbers of neurons seem not to be everything. What seems most important is that design takes full advantage of small size.

Efficiencies of small size

Because an insect is small, it can use an external skeleton. Small body and exoskeleton both allow a smaller brain, which is intrinsically more efficient. A small brain uses disproportionately less wire than a larger brain, so it can locate cell bodies at the brain's margins, out of the way of wires and tracts (chapter 13). As well as saving space, this also saves energy because a distant cell body reduces load on a neuron's electrical circuit (chapter 7). Shorter wires allow more analogue signaling (e.g., worm; chapter 2), and what spikes are needed can travel at lower velocities on thinner axons.

Furthermore, a small brain allows a compact neuron to coordinate the activities of an entire system (figure 4.15) or to spread its dendrites broadly enough to extract a pattern from an entire sensory field.

An insect brain economizes by relaxing the specifications for workaday behavior. A low-mass insect, clad in tough exoskeleton, sustains less damage in a collision or a stumble, so it can tolerate accident rates that would for humans be criminally negligent. The exoskeleton also lessens the burden of motor control. Shorter limbs with stiffer joints and viscous damping are easier to manage, and the ability to place sensors in the exoskeleton to measure the most informative forces reduces the need to compute at higher levels. Mammals use the same strategy (see above), but an exoskeleton provides insects with more opportunities for sensor construction and placement. Insects do require some high-performance control systems; it would be impossible for a fly to fly without one, but in many respects the insect body is less demanding and more adaptable.

The exoskeleton provides opportunities to reduce demands on the brain through embodied computation. For example, to beat its wings at 200 Hz, *Drosophila* builds an oscillator from its flexible exoskeleton and muscles that, when excited, contract in response to stretch (Dickinson & Tu, 1997). To kick start, a dedicated neural circuit excites an auxiliary muscle to contract sharply and stretch the muscles that elevate the wings. As the elevators contract, they stretch the muscles that lower the wings. Coupled by the resonant exoskeleton, the antagonists pull back and forth, beating the wings. To keep the muscles excited, the brain need only deliver spikes at less than 10 Hz. Thus, an intermittent, low-rate input from the brain produces a high-rate, patterned output from the body, significantly reducing computational load. The kick-start muscle also yanks the legs straight, thrusting the fly upward as the wings start to beat, a case of "neatening up" (chapter 1).

The brain can further reduce its computational load by taking shortcuts. Challenging problems are solved with simpler solutions that, while inexact, work well enough, and many animals, including humans, use these efficient *heuristics* (Gigerenzer, 2008). Insects often use them to judge the sizes of much larger objects (Wehner, 1987)—an egg to be parasitized by a tiny wasp, a chamber in which to build a whole ants' nest, a target of given angular diameter—is it small and close by or big and far off?

A big problem for any animal is how to find its way in the world and return. The honeybee uses the sun as a compass to set its bearings from the hive to a productive clump of flowers. When the sun is obscured, the bee infers the sun's position from the pattern of polarized light in patches of

blue sky. To relate a fragment of this polarization pattern to its stored map seems difficult, but the bee employs a shortcut: it reduces the two-dimensional sky map to a one-dimensional map of *e-vector* versus bearing to the sun, ignoring the sun's arc as it travels across the sky (Rossel & Wehner, 1982). As expected, this extreme simplification produces serious errors (up to 30 degrees depending on the time of day), but these are of little consequence because the bees all use the same map. Thus, when a scout returns to the hive on a bearing that, according to her faulty map is 50 degrees from the sun, she communicates this bearing to food gatherers. When these foraging bees set off, they head in the right direction because they use the same faulty map to set a bearing of 50 degrees.

In short, what insect designs demonstrate to an astonishing degree is the advantage of specialization. If a task is specified for a modest range of conditions, then it can be done with a highly specialized design. This is the significance of J.B.S. Haldane's famous remark that God seems to have had an "inordinate fondness for beetles." Their primordial design apparently allowed them to specialize enormously—so each could do with great efficiency what its niche required. A brain comprising small, specialized areas will, like an ecosystem of interacting specialists, be complicated.

Conclusions

Mammalian and insect brains accomplish the same core tasks and are subject to the same physical constraints, so both are designed to *send at the lowest acceptable rate* and *minimize wire*. Both brains regulate the body's internal milieu via slow, wireless (endocrine) signals, plus thin wires with extremely low firing rates (autonomic). Both send long-distance signals via tracts with mostly thin axons. Both arrange their sensors and brain regions in similar positions and use similar structures to perform similar computations. These designs operate at or above the level of the single neuron. But lower levels—molecules and intracellular networks—are subject to similar constraints and therefore follow similar principles, as described next in chapter 5.

5 Information Processing: From Molecules to Molecular Circuits

Chapter 3 explained that information is transmitted when a signal reduces uncertainty about the state of a source. It further explained that in transmitting information by pulses, the information rate (bits/s) depends on the pulse rate and timing precision. That chapter noted a law of diminishing returns: as pulse rate rises, there is less information per pulse (figure 3.6). Moreover, higher information rates (i.e., higher pulse rates and greater timing precision) use disproportionately more space and energy, both of which are limiting resources. These resource constraints directly suggested three principles for efficiency in transmitting information: *send only what is needed*; *send at the lowest acceptable rate*; *minimize wire*. Chapter 4 showed that these principles shape many aspects of brain design on a spatial scale of centimeters down to micrometers.

Yet, as pulses transfer information over distance, they are mainly reporting results. The actual processing of information occurs mostly on a 1,000-fold finer spatial scale, the scale of molecules. There information is processed by chemical reactions: molecules diffuse, bind, exchange energy, change conformation, and so on. The key actors at this level are single protein molecules (~6 nm). They are targets for diverse inputs, such as small "messenger" molecules that, upon binding to a receiver protein, reduce its uncertainty about a source. Protein molecules also provide diverse outputs that, for example, alter the energy or concentration of other molecules, thereby reducing their uncertainty.

These processes not only operate at different scale, they often use a different format. Rather than being pulsatile, molecular signals are often graded, that is, analogue. Despite the change in format, the task remains the same: to reduce uncertainty. Therefore, the same principles for communicating information still apply. Chapter 5 explains how information is processed by single molecules. It identifies constraints on the information

capacity of a single protein molecule, and the irreducible cost of registering one bit. A logical place to begin is where information from an electrical pulse is forced to change format to a chemical concentration.

When one neuron sends a pulse to another neuron, there is a problem. The source wire that delivers it is separated physically from the receiver neuron by a gap of 20 nm. When a signal manages to cross that gap, there is another formidable barrier, a double layer of hydrophobic membrane about 5 nm thick. How to cross both barriers and finally deliver information to the receiver? The membrane is equally a problem for wireless signals (chapter 4): how can a hormone outside the cell deliver its information to the inside? The solution in both cases is for the message to change format. This presents boundless opportunities to process information and also opportunities to lose it.

Information from a pulse crosses the gap as a puff of small molecules—appropriately termed *transmitter*. Information finally enters a receiver neuron when one or more transmitter molecules bind to a protein molecule that spans the cell membrane. Binding triggers the protein to change conformation, and that carries information into the cell. A wireless messenger (hormone) works the same way—binds to a transmembrane protein to change its conformation.[1] *Thus, most transfer of information from a source neuron to a receiver neuron occurs via chemistry (concentrations, binding reactions) and physics (changes in molecular structure).*

Information can enter a cell in myriad ways. The change in protein conformation may open a channel through the membrane to admit ions that carry electrical current. Or it may cause a protein's cytoplasmic tail to release a small molecule that binds and alters other proteins. An altered protein may search out targets by random walk (diffusion). To save time its search may be reduced from three dimensions to two by allowing the altered protein to skate with little feet along the membrane's inner surface.

Such mechanisms accomplish much of the brain's information processing. They amplify, perform logical operations, store and recall, and so on. Although these mechanisms may be triggered by an all-or-none pulse, they themselves are generally graded: small molecules vary in concentration, activated proteins vary in number, ionic currents vary in amplitude, and so on. The information content of these analogue signals, as for the pulse code, can be usefully analyzed by Shannon's formulas. A very few equations, all intuitive, can explain fundamentally: (1) what constrains information processing by signals; (2) what reduces their information; and (3) why higher information rates are more expensive.

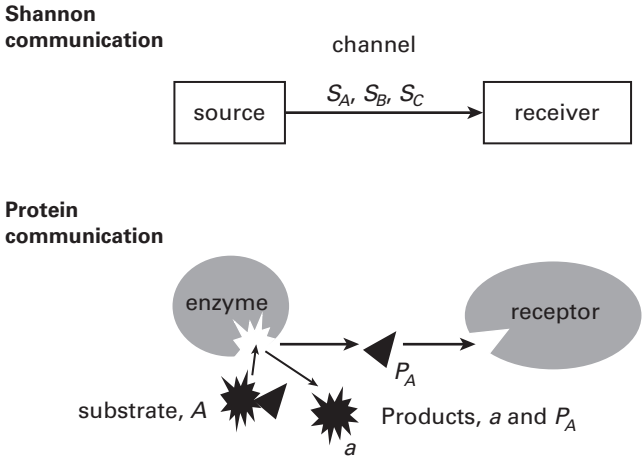


Figure 5.1
Shannon's general communication system maps onto communication between two protein molecules. When Shannon's source is in states *A*, *B*, or *C*, it transmits signals S_A, S_B, or S_C, so eliminating the receiver's uncertainty about the state of the source. The protein source is an enzyme that, upon encountering substrate *A*, produces two products P_A and *a*. The protein receiver is a receptor that specifically binds P_A. P_A's presence or absence at the receptor's binding site establishes the state at the source, namely, that *A* is present or absent.

The reward is the same as for pulses (equation 3.3): with these formulas one can "follow the money" and thereby discover how constrained neural resources are spent. Moreover, following the money at this nanoscale leads to all the remaining principles of neural design. So now we explain how Shannon calculated the amount of information needed to specify a source and how much information a signal can carry (figure 5.1; Shannon & Weaver, 1949).

How much information is needed to specify a source?

The information needed to specify a source increases with the number of states that the source might occupy. Where there is only one state, there is no uncertainty, so no information is required and signals indicating this known state are *redundant*. Efficient designs will reduce redundancy to satisfy the principle *send only what is needed*.

If there are two equally likely states, *A* and *B*, then by definition 1 bit of information eliminates uncertainty by identifying *A* or *B* (e.g., *A* = 0; *B* = 1).

Increase the source's degrees of freedom from two to four states, *A*, *B*, *C*, *D*, and the probabilities are lower, for example,

$p(A) = p(B) = p(C) = p(D) = 0.25.$

Now the situation is more uncertain, and to decide requires two bits. The first bit decides between two equally likely pairs, for example, (*A, B*) versus (*C, D*), and the second bit decides between members of the pair. These two bits constitute a 2-bit code for states, such as

$A = 00$; $B = 01$; $C = 10$; $D = 11$.

The fact that 1 bit specifies two states and 2 bits specifies four states illustrates a general relationship. When a source can be in any one of *U* equally likely states, to identify the state of the source a receiver must obtain at least

$$I = \log_2(U) \text{ bits.} \tag{5.1}$$

Note that, as expected, the quantity of information needed to define the state of a source increases with the complexity of the situation—here the number of possibilities, *U*.

Most sources in nature have states whose likelihoods differ, and this affects the quantity of information needed to specify a state. For example, when we change the probability distribution of the four states, *A*, *B*, *C*, *D* to

$p(A) = 0.125$; $p(B) = 0.5$; $p(C) = 0.25$; $p(D) = 0.125$,

all four states can be identified by a 2-bit code: ($A = 00$; $B = 01$; $C = 10$; $D = 11$), but a 3-bit code is more efficient (figure 5.2). The first bit decides if the state is *B,* the second if it is *C,* and the third if it is *D* or *A*. Note that each choice is binary and equiprobable—1 bit. When used repeatedly, this 3-bit code is, on average, more efficient than the 2-bit code. On 50% of occasions the 3-bit code needs just 1 bit to identify the correct state, $p(B) = 0.5$. On 25% of the occasions, it needs 2 bits to identify the correct state, $p(C) = 0.25$), and on 25% it needs three bits to identify the correct state, $p(A) + p(D) = 0.25$. With usage so distributed, the average number of bits per determination of state is

$0.5 \times 1 \text{ bit} + 0.25 \times 2 \text{ bits} + 0.25 \times 3 \text{ bits} = 1.75 \text{ bits}.$

Thus, the 3-bit code is 12.5% more efficient than a 2-bit code. This illustrates one of Shannon's discoveries: it is efficient to match a coding scheme to the statistical distribution of the states being coded. The brain got there first (chapter 9).

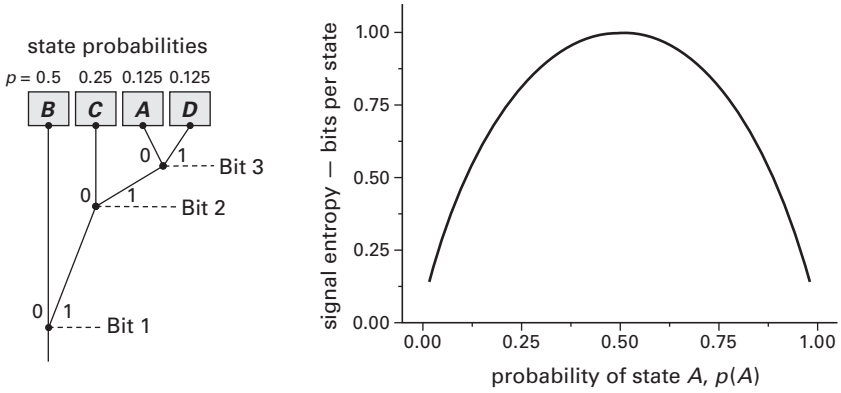


Figure 5.2

Two ways to improve efficiency with which signal states represent information. Left: This decision tree implements a 3-bit code to represent four states that have different probabilities. An alternative would be to assign 2 bits to every state, but 3-bit code is more efficient because half the signals transmitted (those for state *C*) use only 1 bit, and this more than compensates for giving the least frequent states (*A*, *D*) 3 bits. **Right**: A limited number of signal states is used most efficiently when all states are used equally often. In this two-state system the condition $p(A) = p(B) = 0.5$ maximizes information capacity at 1 bit per state. Left reprinted from Laughlin (2011).

The 3-bit code also reminds us that the number of bits needed to specify a state increases with the state's uncertainty. One bit specifies the most likely state, $p(B) = 0.5$; two bits the next most likely, $p(C) = 0.25$; and three bits the least likely, $p(A) = p(D) = 0.125$. In general, when the probability of encountering state x is $p(x)$, the information required to specify x is

$$I_x = \log_2(1/p(x)) = -\log_2(p(x)) \text{ bits.} \tag{5.2}$$

This relationship is consistent with equation 5.1: when there are U equally likely states, $p(x) = 1/U$.

The four-state source explains the basics, but how does information theory apply to the riotous possibilities of the real world? For practical applications, such as the design of his employer's telephone network, Shannon derived a general equation. The number of bits needed to specify the state of *any* source is

$$H(x) = -\sum_{1}^{x} p(x)\log_2(p(x)). \tag{5.3}$$

This quantity, $H(x)$, takes the information per state, as defined by its probability $p(x)$ in equation 5.2, multiplies it by the proportion of time the state is used, $p(x)$, and sums this quantity across all states.

Shannon named this quantity, $H(x)$, *entropy* because its equation (5.3) has the same form as Boltzmann's equation for the entropy of a thermodynamic system. Indeed, the two entropies derive similar quantities. Boltzmann's entropy quantifies a system's total disorder. Shannon's entropy quantifies a system's total uncertainty, and it enabled him to answer our next question.

How much information can a signal carry?

The number of bits carried by a signal is given by the entropy equation, here the entropy of signal states. We start with a signal's ability to specify a source. When every source state is allotted its own signal state (a 1:1 mapping of source onto signal), the signal can carry all of the information needed to specify the source because it can always represent each and every state of the source, and from equation 5.3 this information is the *source entropy*. This equality suggests a general method to calculate the information carried by a signal. Identify the signal's states and use them to calculate the signal's entropy in bits. The calculation obviously holds when source states map 1:1 onto signal states, but is it valid when the source and signal states greatly differ? For example, is it valid when analogue signals from a microphone are transferred to the digital format of a CD or when analogue synaptic potentials trigger trains of action potentials?

Shannon proved mathematically that entropies equate across formats. Thus, it is always possible to devise a mapping whereby a signal with entropy H bits specifies the states of a source with an entropy H bits. Thus the information from a meandering source with many rare states, such as sounds in a telephone conversation, can be compressed into snappier codes that use fewer states more often, such as high frequency radio signals or bits in a digital network.[2] In short, to quantify how much information a signal can carry, just calculate its Shannon entropy using equation 5.3. Having done so, one can consider design issues for the signals that couple a neural source to a neural receiver.

Entropy sets the upper bound to a system's information capacity, but communication systems generally and neural systems in particular are unable to fill that capacity. The first constraint is noise because, when noise enters a system, information is lost. Thus, we must consider how noise

affects the design of neural circuits. The second constraint is redundancy because, repeating a signal reduces a system's capacity to send *new* information. However, when noise is present, repetition can enhance the system's ability to specify the source. Consequently, noise and redundancy in every real communication system are complementary.

How noise destroys information

Noise (random fluctuation that does not correlate with changes in signal state) destroys information by introducing uncertainty. In a noise-free system, the receiver can associate a given signal state with a source state with total confidence; however, when noise is present, is a change sensed by the receiver signal, or is it noise? The quantity of information destroyed by noise depends on the uncertainty introduced by noise and, because bits resolve uncertainty, this is also the number of bits required to describe the noise—its Shannon entropy (equation 5.3). It follows that the information carried by a signal in the presence of noise is the signal entropy minus the noise entropy. Because entropy tends to increase logarithmically with the number of states (equation 5.1), and subtracting logarithms is equivalent to division, information increases as the logarithm of the ratio between signal and noise; $\log_2(S/N)$.

Redundancy

Redundancy (signal state that represents something already known) carries no information. Redundancy comes in two forms. The first is a less extreme form of repetition—states are no longer completely correlated; they are partially correlated. When state *A* correlates with state *B*, receiving *A* increases the probability of receiving *B*, thus reducing the uncertainty associated with *B*, and hence *B*'s information content. Circuits commonly use lateral and self-inhibition to remove this form of redundancy in order to *send only what is needed*, information (chapters 9 and 11).

In the second form of redundancy, the signal states are carrying less information than they might because they are used too frequently or too rarely. Consider a binary signal with two states, *A* and *B*. The information carried by these two states depends upon the signal entropy,

$$H = -p(A)\log_2(p(A)) - p(B)\log_2(p(B)) , \tag{5.4}$$

and H peaks at 1 bit per state when $p(A)$ is equal to $p(B)$ (figure 5.2). This optimum coding strategy, use states equally often, generalizes to systems with many states, and is widely employed in systems where the number of available signal states is severely limited by power restrictions and noise

(e.g., satellites, mobile phones). Retinal neurons face similar limitations and use this same strategy as do many other cell signaling systems (chapters 9, 11; Bialek, 2012).

Now we know: (1) how much information is needed to describe a set of events (equation 5.3); (2) how much information a signal conveys about these events (also equation 5.3); and (3) how these quantities depend on redundancy and noise. This leads to another question: when events change rapidly, how can the brain keep up? To answer this, we derive an expression for information rates in bits per second.

Calculating the information rates of continuously changing signals

The rate of information transfer depends upon the amount of information conveyed by each signal state and the rate at which these states evolve over time. Chapter 3 gave the information rate for action potentials by calculating their entropy. This quantity depends on a physical property of the signal: discrete pulses timed with a given precision, a property that makes low rates cheaper. Other formats have different properties, and these impose different constraints on relationships between signal quality, bit rate, and efficiency.

Much of the brain's information is represented by analogue signals that, by definition, change continuously. These include changes in concentration of messenger molecules, changes in the number of receptor proteins activated by a ligand, and changes in the electrical potentials generated across neural membrane by ion channels. As an analogue signal varies, it runs through a series of signal states (figure 5.3). These states deliver information at a rate that is the number of bits conveyed per state multiplied by the rate at which states change. The number of discriminable states is the range of response covered by signal and noise, $(S + N)$, divided by the noise (figure 5.3). Thus, from equation 5.2, each state delivers $\log_2(1 + S/N)$ bits. The analogue signal can change level in time Δt (figure 5.3). Thus, states are delivered at a rate $R = 1/\Delta t$, and when successive signal states are uncorrelated (i.e., no redundancy in the input), the information rate is

$$I = R.\log_2(1 + S/N) \text{ bits s}^{-1}. \tag{5.5}$$

In many practical systems, calculating rate is more complicated. Equation 5.5 assumes that redundancy is zero, that is, there is no correlation between signal states. To achieve this, signal states must change randomly. To be truly random, the signal must be able to jump from any one state to

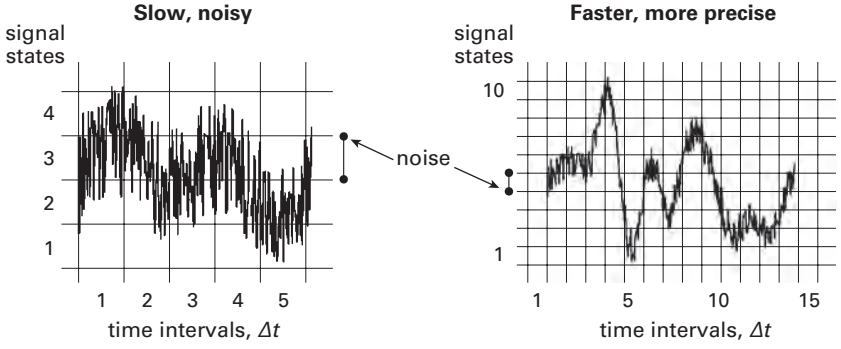


Figure 5.3

Signal range, noise, and response dynamics determine the information rates of analogue signals. Noise divides a waveform's signal range into discriminable states, and states can change at time intervals Δt. The faster, more reliable waveform obviously conveys more details of the signal. From equation 5.5, it also has a higher information rate because it has a higher S/N and, with shorter Δt, changes level at a higher rate.

any other, but this ability is constrained by the time needed to make the jump. For example, an enzyme generates a product at a finite rate, so it requires time to change the product's concentration in a compartment of given volume; similarly an electrical current supplied through a resistor requires time to charge a capacitor. Thus, the number of different states to which a signal can jump in one time interval, Δt, is limited by the rate at which the signal can change, but given sufficient time, it can move to any state. This time dependency complicates the calculation of information rates.

Shannon solved this problem by using the Fourier transform to convert the continuous analogue signal and noise into their frequency components. Each frequency component is independent, in the sense that changing the amplitude or phase of one frequency component has no effect on any other frequency; consequently, every frequency carries its own information. It follows that the total information carried by the signal is the sum of the information carried by each of its component frequencies.

$$I = \int_0^{co} \log_2[1 + S(f)/N(f)] \cdot df \ , \tag{5.6}$$

where I is bits per second, $S(f)$ and $N(f)$ are the power spectra[3] of signal and noise, and co, the signal's cutoff frequency, defines its bandwidth.

There are two provisos to this derivation of information rate (equation 5.6). The system must be linear, and both the signal and the noise must vary randomly with Gaussian distributions so that the frequencies being transmitted are uncorrelated. These conditions are reasonably well met when systems are driven with low-amplitude Gaussian inputs (e.g., Rieke et al., 1997). Note that when the spectrum of $S(f)/N(f)$ is flat, the sum across frequencies reduces to equation 5.5, with a bandwidth of $1/2\Delta t$ replacing the rate R,

$$I = (bandwidth)\log_2(1 + S/N). \tag{5.7}$$

This relationship between *bandwidth*, S/N, and information rate, I, affects neural design because to transmit information at higher rates, a neuron needs a wider bandwidth (faster responses) plus higher S/N, and these require extra materials and energy. Thus, we have a trade-off between resources and performance that, as we will see, profoundly influences neural design.

Information in any real system must be embodied physically or chemically. The brain uses *signaling proteins* to process information, so we now examine their physics and chemistry.

How protein molecules transmit and process information

A protein acquires its specific function by folding to reduce its free energy

A protein molecule is formed from a linear chain of amino acids linked in a genetically specified sequence (Alberts et al., 2008). The linear sequence becomes a useful molecule as follows. The chain is flexible, so it bends and folds to reduce its free energy by minimizing potential energy and maximizing entropy (Williamson, 2011; Dror et al., 2012). The charged amino side groups attempt to form pairs of attractive opposites (+ with –) and to avoid repellant likes (+ with +) or (– with –). To increase entropy, the hydrophobic side chains avoid polar groups and coalesce into oily cliques. All of this jostling for position must be achieved within packing constraints.

Buffeted by thermal energy, yanked up and down potential gradients, exchanging order for disorder, the protein molecule constantly changes its three-dimensional structure (*conformation*) until it falls into a local minimum free energy that is deep enough to resist thermal motion. The protein molecule has reached a stable conformation (figure 5.4).

This stable conformation determines the protein molecule's physical and chemical properties (Williamson, 2011). A typical protein, with several hundred amino acids, folds into a 5- to 10-nm structure to adopt a form

that supports its function: long fiber to make a hair, globular block with attachment knobs to build the cytoskeleton, part of a stepping leg to move materials, and so on. A protein may locate a subset of amino acids where they can bind and interact with a specific molecule. Such a binding site enables the protein to collect and send information.

Binding specificity allows information transfer

Recall that information transfers when a change at the receiver can be associated with the state of the source. Chemical binding satisfies this requirement. For example, when an enzyme molecule reacts with its substrate to produce its product, the enzyme only binds the substrate, and the receiver, a receptor protein, only binds the product (figure 5.1). Thus, the source tells the receiver "substrate present" by using a diffusible messenger, the product. If the enzyme and/or the receptor were to relax their binding specificities, other molecules in the cytoplasm would also bind. Such cross talk would reduce the probability that the receiver is responding to the presence of one particular substrate at the source. Thus, binding specificity enables information transfer.

Once a protein's binding site receives information, how can it be further processed? By *allostery*. This is a protein's ability to respond to a specific input, such as binding a messenger, by switching to a new stable conformation.[4]

How allostery works

Consider the protein molecule continuously changing conformation as it descends to its lowest available free energy level. This progression is, in effect, a voyage across an energy landscape (figure 5.4) in which the map coordinates represent the protein's conformation and the altitude represents its free energy.

The descent follows gradients in the energy landscape, and thermal jiggles push it over bumps. Thus, the protein explores a locale and finds a path to lower regions. When the protein enters a valley too deep for thermal forces to boost it out, it is trapped, and the conformation becomes confined to a small region (figure 5.4). Here the protein may shuttle between a small set of functionally distinct conformations, or it may remain centered on one stable conformation (figure 5.4). Thus confined, the molecule assumes a role dictated by its conformation.

Consider now what happens when an external factor alters the energy landscape. An external input could be a change in pH or electrical potential, it could be binding or releasing a specific molecule, or it could be an

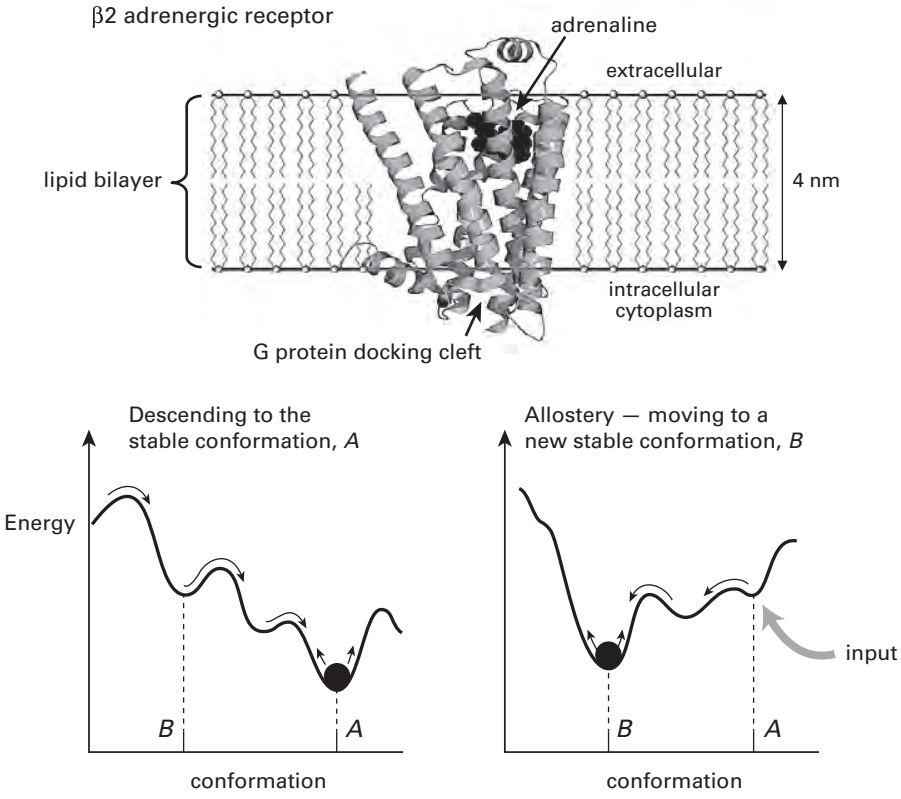


Figure 5.4

Protein structure, conformational state, energy landscape, and allostery. Upper: The β2 adrenergic receptor protein spans the cell membrane's lipid bilayer. Here it is shown in the conformation where binding an adrenaline molecule at a site on the outside has opened a cleft for binding a G protein molecule on the inside. Note the prominent helices crossing the membrane. **Lower left**: Section through a protein molecule's energy landscape. During folding, the protein descends the energy landscape and adopts the stable conformation A. **Lower right**: An external input changes the energy landscape and the protein moves to conformation B. This is allostery. Upper adapted from http://en.wikipedia.org/wiki/Beta-2_adrenergic_receptor#mediaviewer/File:2RH1.png.

injection of energy via the attachment of a high-energy phosphate group to an amino side group. Such inputs alter the protein's energy landscape, depressing some regions and elevating others (figure 5.4). The protein responds by moving, within microseconds to milliseconds (Williamson, 2011; Dror et al., 2012), to a new stable conformation. The new conformation differs physically and chemically from the previous one, so the molecule reacts differently to chemical and physical inputs. This change enables it process information.

How a protein uses allostery to process information

A finite-state machine[5] processes information by running through a well-defined sequence of state changes (transitions), each triggered by a particular condition, such as the presence or absence of an input, or a conjunction of inputs. This is allostery. As a protein molecule runs through a sequence of state changes, each conditional upon a particular input, it produces an output conditional upon those inputs (Huber & Sakmar, 2011). Thus, allostery enables a single protein molecule to compute (Bray, 1995). For example, a single molecule is easily programmed to perform the Boolean operation, AND (figure 5.5).

The rest of this chapter treats one particular finite-state machine that comprises a pair of interacting proteins. The receptor protein accepts the wireless signal, adrenalin, a hormone that prepares an organism to fight or flee, then relays the information ("Adrenalin present!") across the cell membrane. There it transmits to receiver proteins on the membrane's inner face that amplify and broadcast the information within the cell. Both proteins then reset for the next signal. The receptor protein is the β2 adrenergic receptor, and the receiver protein is a G protein.

We choose this example for several reasons. First, the β2 adrenergic receptor and its G protein represent a broad, ubiquitous class of finite-state machines (chapters 2, 7, and 8). The human genome specifies more than 800 different receptor proteins that couple to a G protein and more than 100 different G proteins. Second, this example indicates the spatial scale used by most neural computations. Third, it exemplifies computation by amplifying, and in doing so illustrates molecular solutions to a broad design problem, overcoming noise. Fourth, it clarifies the reason to compute at this spatial scale: high efficiency in space and energy. The energy cost of 1 bit in this system, as will be explained, approaches the theoretical lower limit to within a factor of about 30.

The final reason to choose this example over other possibilities is that the sequence of conformational changes, triggered by adrenalin's

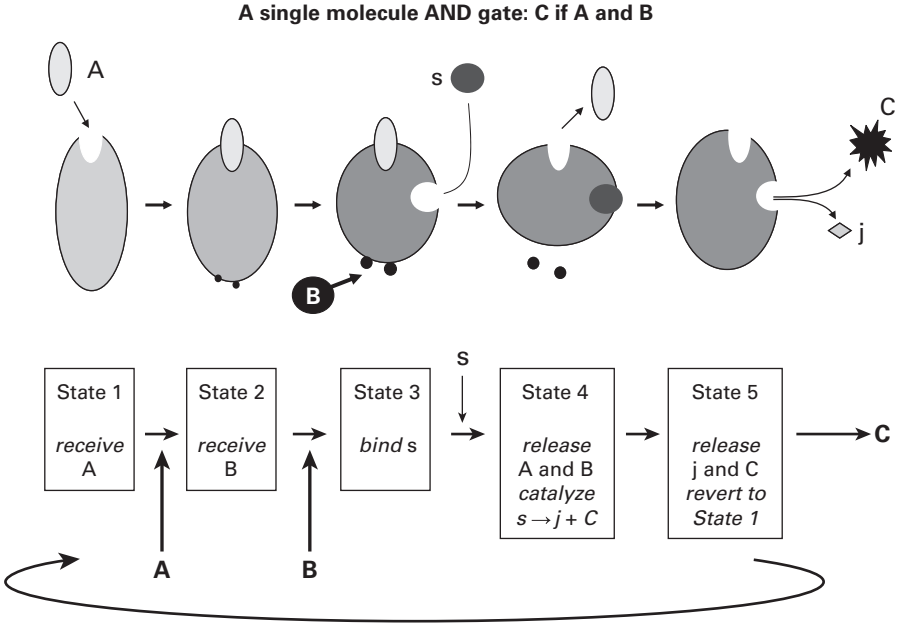


Figure 5.5
The allosteric protein as a finite-state machine. How a sequence of stimulus-evoked changes in allosteric state could enable a single protein molecule to perform a simple computation, here a logical AND on the two inputs A and B. Ligand A binds to the protein, exposing two sites to be phosphorylated by kinase B. The pair of attached phosphates alters the protein's conformation, exposing a catalytic site that digests the substrate s to produce products j and C. Bottom row gives the corresponding program of state transitions.

binding to the receptor and completed by the release of activated G proteins, has been documented at the atomic scale, by x-ray diffraction (Rasmussen et al., 2011; Chung et al., 2011; summarized in Schwartz & Sakmar, 2011).

Allostery in action

The system is ready to receive when the receptor's conformation exposes its adrenaline binding site on the cell membrane's outer face and masks the G protein's binding site on the inner face (figure 5.6). G proteins diffuse on the inner face, colliding with receptors, but encounter no signal. When adrenalin binds to the receptor, the protein changes conformational state

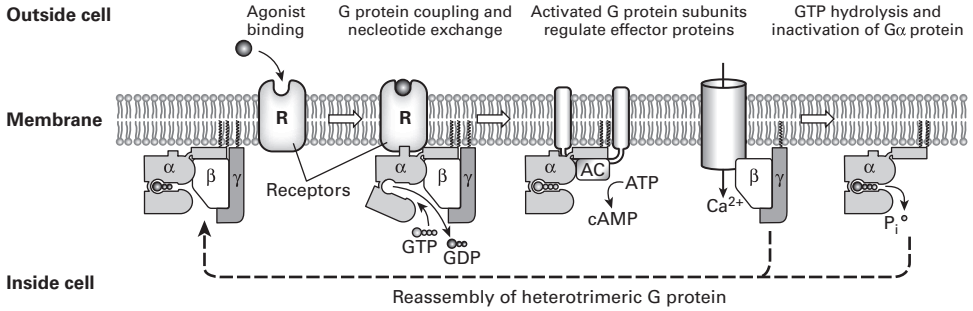


Figure 5.6

β2 adrenergic receptor and its G protein use allostery to operate as a finite-state machine. Receptor receives a wireless signal outside the cell and, by changing conformation, relays it across the membrane to G protein. G protein dissociates and α subunit broadcasts signal to effector proteins by diffusing on inner surface of the membrane. α subunit hydrolyses bound GTP and reverts to conformation that the binds the other subunits. G protein is reconstituted, ready to signal again. Further details in text. Figure adapted from summary diagram from the definitive study of structural changes that pass information through these two molecules (Rasmussen et al., 2011), with permission.

(figures 5.4 and 5.6). One of the seven helical coils that span the membrane (coil number 6) moves 1.4 nm and others move shorter distances. Together they open a cleft in the receptor molecule at the inner face to expose the G protein's binding site. At the next collision, a G protein engages this site with a special knob and docks securely (figure 5.6).

This coupling changes the energy landscape of both molecules. The G protein embarks on a sequence of conformational changes (figure 5.6). Two of its three subunits, β and γ, detach and diffuse into the cytoplasm. The α subunit responds to the loss of its partners by swinging apart two large sections at their hinge. This motion, spanning more than 110° and requiring several hundreds of microseconds, reveals, like an oyster showing its pearl, a small molecule, guanosine diphosphate (*GDP*), bound deep within the protein. The exposed GDP promptly exchanges with a molecule from the cytoplasm, guanosine triphosphate (*GTP*), whose additional phosphate gives it higher energy.

GTP's binding transfers energy to the α-subunit, again changing the landscape. The hinged gates swing closed, retaining the high-energy GTP that is fueling the sequence of state changes. The knob retracts, thereby uncoupling the α subunit from the receptor and freeing it to diffuse on the membrane's inner face. Now another binding site on the α subunit is

exposed for other proteins to bind and change *their* conformation in response to the signal "Adrenaline!" (figure 5.6). In short, an orderly sequence of conformational state changes has carried information, "Adrenaline!," across the cell membrane, and by releasing an activated GTP-α subunit, it has started the process of broadcasting this information wirelessly within the cell.

How allostery amplifies

This form of allostery easily amplifies. When one GTP-α uncouples from the activated receptor protein, another docks in its place, is activated, then is released, and so on. The rates vary from 10–500 per second, depending mainly on the density of G proteins on the membrane—for this sets their frequency of encountering a receptor protein. The number of G proteins activated and released by a receptor increases with time as the cleft stays open. The amplification (*gain*) varies across systems, ranging from 4 in a system with short time constant, such as a fast fly photoreceptor (chapter 8) to 100 in systems with long time constant, such as a slow-acting hormone.

Amplification is a form of redundancy since each copy simply repeats a message without adding new information. Thus, multiple G proteins activated by the β2 receptor simply repeat, "Adrenalin!," "Adrenalin!" . . . Yet this redundancy is essential for two reasons. To produce a concerted response to adrenalin, the signal must reach many parts of the cell in good time, hence the activation of several G proteins. Second, the system must guard against noise. Because a thermal bump occasionally activates a single G protein molecule, the receptor must activate several molecules to generate a reliable message. Thus, when amplification protects information from noise, it also introduces inefficiency in the form of redundancy. An efficient design will strike an appropriate balance by matching the gain of amplification to the level of noise (chapter 6).

Although the β2 receptor and its G protein have worked together to amplify and broadcast the signal "Adrenaline!," the process is incomplete. This finite-state machine, which turned on in order to signal danger, must turn off when the warning has been sent. Then the machine must reset to be ready once again to deliver the message.

How allostery terminates the message and resets the system

Turnoff and reset are accomplished by continuing to move the receptor and the G proteins it activated through their sequences of conformational states. As for the all preceding steps of activation, each transition for

deactivation serves a specific purpose. To inactivate the β2 receptor, an enzyme (*kinase*) accepts a high-energy phosphate group from an ATP molecule and attaches it covalently to a particular site on the β2 receptor. The phosphorylation of several such sites raises the receptor's energy level sufficiently to change its conformation, now exposing a binding site for a different protein molecule, *arrestin*. When arrestin binds, it blocks access to the G protein's docking cleft, thus preventing transmission.

Something is needed to protect unoccupied β2 receptors from being inactivated while they are in the receptive conformation, waiting for adrenalin. The receptor is engineered so that the receptive conformation hides the phosphorylation sites, and they become exposed only in the conformation triggered by binding adrenalin. Something is also needed to give time for an activated receptor to amplify, that is, to activate and release several G proteins. To achieve this, the kinases that attach high-energy phosphates are designed to work slowly. Moreover, by modulating this rate of phosphorylation, both the gain and time constant of amplification are adjusted for no extra space and little extra energy.

Once arrestin blocks transmission to the G protein, the β2 molecule resets—by continuing its journey through conformational states. The adrenalin molecule, whose initial binding to the receptor opened a cleft for docking the G protein, eventually *un*binds adrenalin, and this closes the docking cleft. This allows a *phosphatase* enzyme to remove the added phosphates, releasing arrestin, and restoring the receptor to its initial state. Its *signaling cycle* is complete: it has received, transmitted, and reset.

But what prevents the activated α subunit from continuing its diffusive search for partners? This subunit is also an enzyme that removes the high-energy phosphate from its own bound GTP (figure 5.6), and this provides an automatic cutout. Withdrawing the high-energy phosphate from the α subunit triggers its final sequence of conformational state changes. It rebinds the βγ units and once more protrudes its docking knob. Now the G protein has reset to the inactive αβγ-GDP form and is again ready to dock with an adrenalin-bound receptor.

In summary, this molecular finite-state machine uses two parts, receptor and G protein. It exploits three properties of a protein molecule—binding specificity, allostery, and diffusion—to execute a program of state changes. The program receives a signal at the cell surface and transmits it *mechanically* across the cell membrane. The program then amplifies the signal, broadcasts it within the cell, and resets. This computational device, the G-protein-coupled receptor (*GPCR*), being ubiquitous, will be discussed

further (chapters 6–8). But here we explain another invaluable property of signaling proteins—how their energy efficiency approaches the thermodynamic limit.

Energy efficiency of protein devices

Why must molecular devices consume energy to process information?

A protein's signaling cycle starts and finishes at the same point in the energy landscape. If every conformational state within the cycle had the same free energy, the cycle could be completed without consuming energy. However, the protein would then depend on random thermal fluctuations to change states. Moreover, if free energy were constant, each transition would be reversible—with equal probabilities of moving forward or backward. To complete the cycle would be theoretically possible: a signal could be delivered without expending energy. However, such a lossless system would be impractical because, relying on a chain of improbable and reversible events, the receiver would wait for long and indeterminate times (Bennett, 1982, 2000).

Energy eliminates this intolerable wait by driving the protein through the conformational state transitions in the intended direction. Moreover, the effect is progressive: adding more energy speeds the cycle. But what about the lower bound: what is the least energy that can deliver information usefully?

Lower bound to energy cost in signaling

Thermodynamics suggests a minimum, the energy required to register one bit of information (Landauer, 1996; Schneider, 2010),

$$\Delta E = k_B T \ln(2) \approx 0.7\ k_B T \text{ joules} \approx 3 \times 10^{-21} \text{ joules per bit}, \tag{5.8}$$

where k_B is Boltzmann's constant and T is temperature in degrees Kelvin. ΔE is tiny,[6] but single protein molecules are also tiny and so approach this thermodynamic limit to energy efficiency.

The signaling cycles of the β2 adrenergic receptor and its G protein can each register a bit by switching from OFF to ON and then resetting to OFF. Each protein draws energy from the cell's standard currency, the high-energy molecule, ATP. Hydrolysis of one ATP delivers 25 $k_B T$ joules, and the receptor uses at least three ATP molecules when it is phosphorylated (figure 5.5). This gives an efficiency of 75 $k_B T$ joules per bit, which is two orders of magnitude above the thermodynamic limit (equation 5.8). The G protein consumes the equivalent of 1 ATP when it hydrolyzes its GTP to GDP

(figure 5.6), giving an efficiency of *25* k_BT joules per bit, between one and two orders of magnitude above the thermodynamic limit.

Thus, both proteins process a bit of information for less than the cost of a covalent bond (~100 k_BT). This seems plausible because a protein is a soft device, more like a machine made from jelly than a rigid clockwork (Williamson, 2011). Indeed, the free energy to stabilize a protein (folded vs. unfolded) is less than a quarter of the free energy to form a covalent bond and is about equal to the energy delivered by ATP.

What prevents these two protein molecules from operating closer to the thermodynamic limit? Realize that the 0.7 k_BT limit is the cost of simply registering a bit as a change of state. It does not include transmitting the bit. To send a bit across the membrane, the β2 receptor moves its helix number 6 by 1.4 nm, and to relay the bit into the cytoplasm, the G protein opens its large hinged section by 110°. Both movements require work (Howard, 2001), and work consumes energy. Energy is also used to drive the cycle at a rate appropriate for the function—recall that the β2 receptor signals "Emergency!" Considering that the energy cost of transmission by the GPCR includes these extra tasks, protein signaling appears astonishingly close to the thermodynamic limit. An order of magnitude is a reasonable guess.

Energy and the design of efficient signaling molecules

The receptor and G protein turn on and off abruptly and reliably—like a mechanical switch. The latter avoids accidental tripping by using an energy barrier. Some of the energy needed to trip it is recycled so that once triggered, the change goes quickly. Where safety is critical, the energy barrier is high, but where it is less critical, the barrier can be lowered to save energy. Likewise, a protein's energy landscape seems engineered to require just the right energy input for each state transition. The design also involves tradeoffs between speed, reliability, and energy. For example, were viscous forces within a protein to increase with switching rate, the energy cost per transition would increase disproportionately, making lower rates more efficient. Thus, a design principle observed at the microscopic level for axons, *send at the lowest acceptable rate* (chapter 3), may also hold at the nanoscopic level for protein molecules, albeit for different reasons.

Summary

The signaling systems established by protein molecules receive and transmit information, as defined by Shannon, using different physical and

chemical processes from the ones that Shannon originally treated. Three physical and chemical properties of proteins support the transmission and processing of information. Binding makes specific connections between molecules, enzymatic activity provides a potent means of generating and amplifying signals, and allostery enables information to pass through single molecules. Allostery also equips a single protein molecule to compute by operating as a finite-state machine. By running through a well-defined program of state changes, triggered by specific inputs, the molecule completes a program only when it encounters a specific combination of inputs. These properties equip proteins to form circuits of molecules that compute.

Circuits built from proteins satisfy two design principles. First, if we rule out quantum computation, these circuits are irreducibly small, and this saves space and materials. Protein circuits also save energy because protein molecules operate near the thermodynamic limit of energy efficiency. Moreover, protein chemistry allows energy to be delivered efficiently in just the amounts needed to meet the circuit's need for speed and accuracy. Thus, the performance of components in protein circuits can be matched to their tasks to gain economies that come with sending at the lowest rate.

These advantages—compactness, energy efficiency, and ability to adapt and match—all suggest the principle *compute with chemistry*. It is cheaper. But to realize the savings, protein circuits must support the brain's core tasks. Chapter 6 now explains how proteins equip molecular circuits to meet a brain's requirements for information processing.

6 Information Processing in Protein Circuits

Chapter 5 explained that information is encoded whenever a source's change in state registers as a change in state at a receiver. The primary mechanism at the nanometer scale is a protein's ability to connect specific inputs to specific outputs by, for example, binding molecules, catalyzing reactions, and changing conformation. These reactions are employed universally in biology and have two advantages for brains—energy efficiency and compactness. As noted in the previous chapter, the energy used by a protein molecule to register 1 bit approaches the thermodynamic minimum. Also, for changing conformation, its unique task, a protein is irreducibly small. Smaller would be better since a moderate-sized protein molecule (100 kDa) spans about 6 nm and occupies about 100 nm^3. But although a smaller peptide can serve as a ligand, it lacks a protein's rich possibilities for stable folds, pockets, and allostery that are essential to its receiving and processing information.

Chapter 5 noted that a protein molecule can compute. For example, it can amplify (one adrenalin bound to one β2 receptor protein activates several G proteins), and it can do logic (e.g., compute the Boolean AND; figure 5.5). However, one logical operation doesn't make a brain. A brain needs to do a lot more math than that. For starters, it needs mechanisms on the nanometer scale to calculate the four linear arithmetical operations (+, -, ×, ÷) and various nonlinear operations such as $\log(x)$ and x^n. It also needs switches (where an input causes a step change in output), filters (to remove certain frequencies and attend to particular timescales), correlators (to associate events), and so on.

For such nanometer-scale computations, the genome serves as a parts catalog—listing the codes for thousands of protein structures, each specified for some particular input/output (I/O) function. But executing an orderly sequence of operations that computes something requires something more: a specific subset of I/O components that link correctly. A cell's

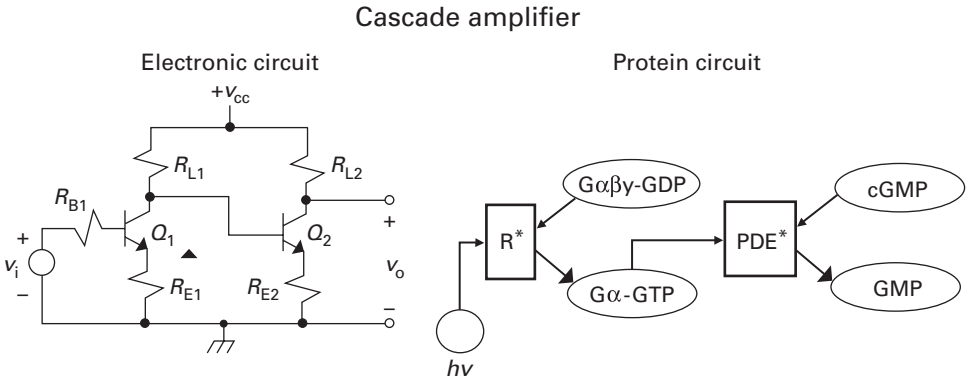


Figure 6.1

Circuit for cascade amplifier: silicon versus protein. In silicon, an input voltage, v_1, drives the first transistor Q_1, which amplifies the signal. Q_1' s output drives transistor Q_2, which amplifies the signal again and generates the output v_o. In protein, a photon ($h\nu$) activates one molecule of a receptor protein (R), changing its conformation to (R*). Like the β-adrenergic receptor (figure 5.6), R* amplifies by catalyzing 20 G proteins to change from Gαβγ-GDP to Gα-GTP. Each Gα-GTP activates a molecule of the enzyme phosphodiesterase (PDE), which again amplifies by catalyzing the hydrolysis of 100s of messenger molecules of cGMP to GMP. Both silicon and protein amplifiers multiply the input by the product of the gains of the two amplification stages. Electronic circuit from http://en.wikipedia.org/wiki/Cascade_amplifier. Protein circuit for phototransduction in rods is described in chapter 8. GDP, guanosine diphosphate; GTP, guanosine triphosphate; R*, the photosensitive molecule rhodopsin, activated by a photon.

internal mechanism ensures that this occurs—that the right proteins are delivered to the right places at the right times (Alberts et al., 2008). In both respects—using components with specific I/O functions and linking them correctly—protein circuits resemble electronic circuits (figure 6.1).

To understand neural computing at the nanometer scale, one must consider what shapes a protein's I/O function. What determines, for example, whether it will take a sum or a logarithm, whether it will switch or filter? These functions emerge from a protein's three-dimensional structure, through its ability to react chemically, mechanically, and electrically, and to change state in response to these inputs—allosterically.

One must also consider how a sequence of I/O functions should couple to make a useful circuit. For example, should a protein couple directly to its target, should it diffuse, should it anchor and send a small messenger, or should it communicate electrically via the cell membrane? Here the broad

answers are simple: diffusion slows as the square of molecular weight, and proteins are heavy, so the best choice for coupling depends on the required distance and allowable time. Diffusion time increases as distance squared and concentration decays exponentially. Thus, molecular size and concentration, plus the laws of diffusion, shape protein circuit design. Consequently, when distances are large and time is short, circuits use electrical signals. This chapter will explain further with some simple examples, starting with ligand binding. The concepts and principles introduced here will be exemplified more thoroughly in all subsequent chapters.

I/O functions emerge from the kinetics of chemical binding

I/O functions from a single binding site

A ligand diffuses under thermal bombardment to a specific site on a protein and binds. That is, it sticks for a time, and then comes off. While the ligand is bound, the protein adopts an active conformation in which it produces its *output*, for example, it is able to bind a downstream protein or catalyze a chemical reaction. Thus, the protein's *output* is proportional to the fraction of time it binds the ligand, and this is determined (Phillips et al. 2009, chapter 6, "Entropy Rules!"; Bialek, 2012) by the ligand concentration [ligand] and rate constants for unbinding (k_{OFF}) and binding (k_{ON}):

$$output/output_{max} = [\text{ligand}]/(k_{OFF}/k_{ON} + [\text{ligand}]). \quad (6.1)$$

This I/O function is *hyperbolic*; it rises steeply at first, and then tapers off as the binding site approaches saturation, $output_{max}$ (figure 6.2). The ratio k_{OFF}/k_{ON} is the dissociation constant k_D, and equals the ligand concentration required to produce a half maximal *output*. The same binding kinetics apply to protein–protein binding, so what is here explained for ligand–protein binding applies also to protein–protein binding.

The hyperbolic I/O function computes. It can perform, depending on input, three analogue operations:

1. At lower inputs levels (those causing < 0.25 maximum output), the function is linear (figure 6.2), so small inputs add.
2. At medium input levels (those causing 0.25 to 0.75 maximum output), the function is approximately logarithmic (figure 6.2). This reduces the sensitivity of the output to the absolute level of the input and scales the inputs proportionally such that a constant fractional change in input, Δ[ligand]/[ligand,] causes a constant change in output, *Δoutput*. This type of scaling exists at the behavioral level for many categories of sensory

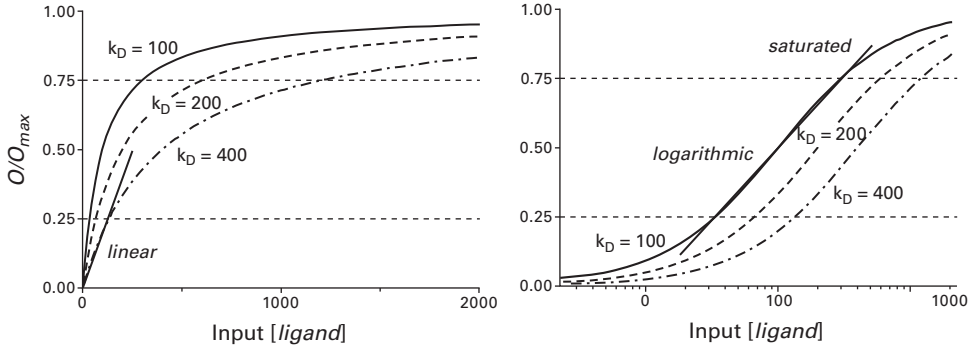


Figure 6.2

Input/output (I/O) function generated by binding kinetics performs the same computations across widely different input ranges by altering dissociation constant, and hence binding affinity. **Left:** Output (normalized to its maximum) is plotted against input, [ligand], for three dissociation constants, k_D. When the output is small (<0.25 max), the I/O function is linear and adds. **Right**: Output plotted against log([ligand]). When the output is medium (0.25–0.75 max) the function is logarithmic. In saturated regime (0.75 max – max) function's slope approaches zero.

discrimination (*Weber–Fechner law*). Thus, a computation that serves behavior starts with chemical kinetics at the nanometer scale.

3. When large, sudden increases in input drive the response from zero to maximum, the function is a step and thus can serve as an ON/OFF switch for Boolean operations.

Sensitivity depends on the protein's affinity for the ligand. Higher affinity (tighter binding) decreases the OFF rate, thus reducing the k_D. The effect is to reduce the concentration of ligand needed to cause a half-maximum output. By adjusting k_D, a given I/O function can execute the same set of computations across a wide range of mean ligand concentrations (figure 6.2). All that is needed is to tweak the protein's binding site to match its affinity to the level of ligand by changing the protein's conformation slightly. This can be executed stably in the genome, by changing the codons that specify influential amino acids, to produce a different *isoform* of the protein, or it can be done dynamically as the protein operates—for example, by using a kinase to add an energetic phosphate.

This capacity of a protein to implement its I/O function with altered binding affinity serves in innumerable ways. For example, at low affinity (high k_D) a protein can receive information from its ligand across a short distance at high concentration, in a brief time, for example,

neurotransmitter diffusing across a 20-nm synaptic cleft. At high affinity (low k_D) the protein can receive information from the same ligand at 1,000-fold lower concentration over a much longer time, for example, a circulating hormone. These capacities are implemented for adrenalin by adrenergic receptors, probably by different isoforms. Dynamic adjustments to affinity can be used for physiological adaptation—to match the I/O function to changes in mean concentration of ligand (figure 3.4).

Protein molecules with different binding affinities transmit different temporal frequencies. High-affinity receptors cannot transmit high frequencies because they do not release their ligand quickly. Consequently, they maintain the same level of output for some time after the input ligand concentration falls. Thus, a high-affinity receptor acts as a low-pass filter—for example, at retinal synapses (chapter 11). By comparison, low-affinity receptors release their ligands promptly, so they transmit high frequencies as well as low, and this gives them a wider bandwidth.

Temporal filtering by a single protein molecule can be modified by *desensitization*. This property curtails the output even while the input ligand remains bound, so allowing a protein with sufficient binding affinity for a low mean concentration of ligand to cut off its response faster than the ligand can unbind. Now, the protein is a high-pass filter. For example, upon binding synaptic transmitter, a protein receptor changes conformation to open an ion channel, but conformational change continues and closes the channel long before the ligand comes off. Speed of desensitization is designed into a protein as part of its energy landscape (Sun et al., 2002), and its use in temporal filtering will be exemplified in chapter 11.

Steeper I/O functions from cooperative binding

A protein's hyperbolic I/O function is steepened by adding more binding sites for the ligand and requiring that several bind to generate the output (Koshland et al., 1982). When n sites have to cooperate, the I/O function follows the nth power of the ligand concentration:

$$output/output_{max} = [\text{ligand}]^n/(k_D + [\text{ligand}]^n). \tag{6.2}$$

Now the I/O function's lower region (figure 6.3) approximates a power function: $output/output_{max} = [\text{ligand}]^n$, and its logarithmic midregion (figure 6.3) is n times steeper: $output/output_{max} = \log([\text{ligand}]^n) = n \log([\text{ligand}])$. By adjusting both binding affinity and *cooperativity*, an I/O function's position and slope can be matched to the distribution of its input levels (figure 3.4)—which in the fly visual system optimizes coding efficiency (figure 9.10; Laughlin, 1981; Nemenman, 2012).

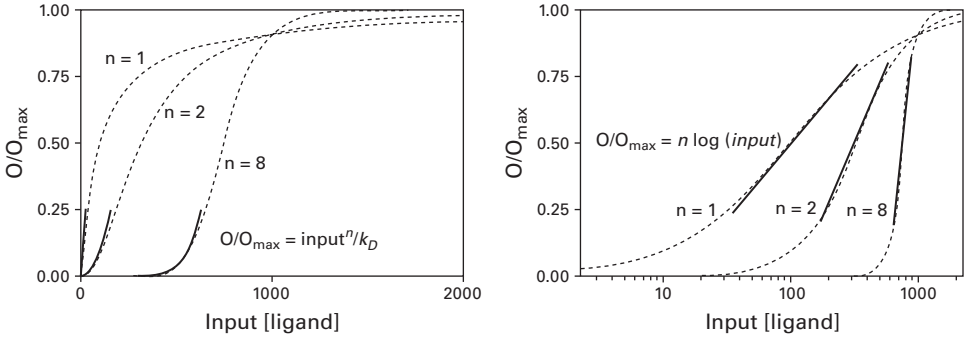


Figure 6.3

Cooperativity changes the input/output (I/O) function generated by binding kinetics to provide different computations. I/O functions are plotted with cooperativities $n = 2$ and $n = 8$ and, for comparison, without cooperativity ($n = 1$). k_D is constant. **Left**: Cooperativity implements the power function $output = input^n/k_d$ with small outputs (<0.25 max). It also shifts the I/O function to higher input values without losing sensitivity (the slope remains steep). In the extreme, for example, $n = 8$, cooperativity creates a switch. **Right**: Cooperativity implements the function $n\log(input)$ in the medium output range (0.25–0.75 max).

A high cooperativity provides a steeper I/O function for digital switching (figure 6.3, $n = 8$) which, by thresholding, can prevent input noise from passing further along a protein circuit. For example, in the protein circuit that releases a synaptic vesicle (chapter 7), a critical step is triggered by the protein synaptotagmin binding calcium ions at several sites. This cooperativity shifts the I/O function to higher concentrations (figure 6.3), so that noisy fluctuations in a cell's baseline calcium concentration rarely release a vesicle. Cooperativity also narrows the range of calcium concentrations that trigger release by increasing the I/O function's slope. Thus, when a voltage-gated calcium channel releases a puff of calcium, synaptotagmin responds promptly, and this increases the temporal precision of release.

Chemical circuitry supports analogue processing

In addition to the functions implemented by binding, proteins' chemical reactions support analogue processing with a rich repertoire of primitives. In brief, simple chemical circuits have equivalent electronic circuits (Sarpeshkar, 2010; figure 6.1) and are capable of implementing procedures used in analogue electronics, namely, amplify, oscillate (Tyson et al., 2003), differentiate, and integrate (Oishi & Klavins, 2011). As well as taking logs (figure 6.2) and raising to powers (figure 6.3), chemical circuits support the

arithmetic operations add, subtract, multiply, and divide (figure 6.4). Small chemical circuits also have the ability to perform more complicated functions—for example, take *n*th roots (Buisman et al., 2008), compute polynomials, and solve quadratic equations. Whether the brain explicitly implements this more advanced algebra in small chemical circuits[1] is an open question, but the point is made. Chemical circuits support Turing's Universal Computation (Hjelmfelt et al., 1991), which means that they can in principle be configured to compute any function.

Chemical circuits cover the time domain

Not only does chemistry compute, it equips the brain to compute over the range of timescales observed in animal behavior—from the microseconds of the electric sense and hearing to a century of memory. Binding and conformational change take microseconds to seconds. Sequences of reactions executed by protein circuits take from milliseconds (phototransduction, chapter 8) to days (the circadian clock, chapter 4). In chapter 14 we describe how memories that are first laid down by the modification of synaptic receptor proteins are then consolidated for years by the chemical synthesis of new proteins and the assembly of new structures.

What makes a protein circuit efficient?

Computation by circuits built from protein molecules is efficient for several reasons. It is efficient in energy because binding and conformational change approach the thermodynamic limit (chapter 5). It is efficient in space because a single molecule computes. Moreover, computation at this level proceeds directly—that is, by implementing "analogue primitives" (Sarpeshkar, 1998; 2014). Analogue computation typically needs fewer steps than digital to complete a basic operation. For example, analogue multiplies directly, but digital takes $PR^{1.585}$ steps, where PR is the numerical precision in bits (Moore & Mertens, 2011), so even with a low precision of 4 bits, eight steps are saved.

Transmission within a chemical circuit is wireless, so space for wires also reaches an absolute minimum and circuits share space seamlessly. Wireless transmission distributes signals with a minimum of equipment. Once a messenger molecule is broadcast, it can be received by any protein with the appropriate binding site. Thus, wireless transmission makes it easier to reconfigure circuits to change behavior—in the short term by sculpting circuits with neuromodulators (chapter 2) and in the long term by evolving new connections (Katz, 2011). Nor is additional energy needed for wireless

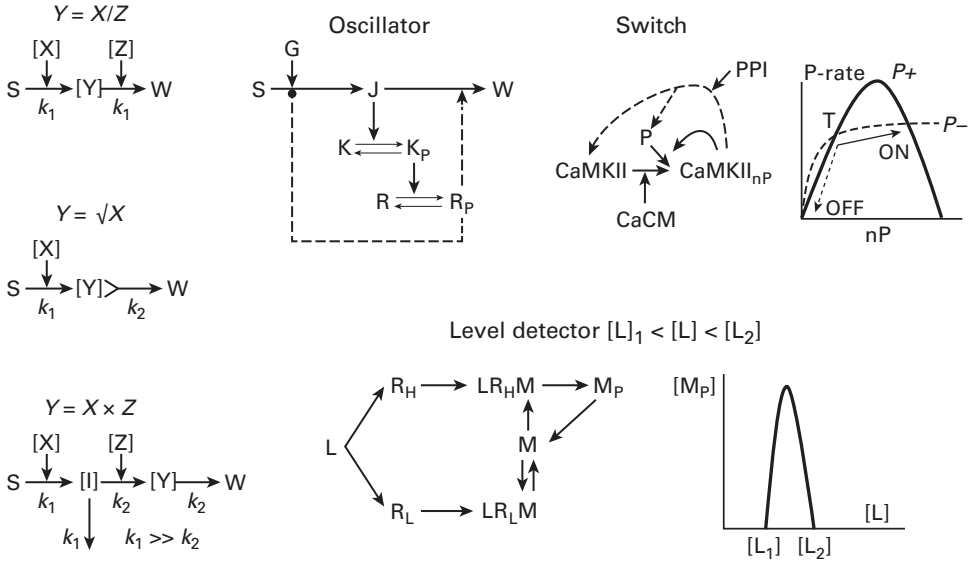


Figure 6.4

Computation by chemical circuits. **Left**: Circuits that divide, calculate square root, and multiply. The steady-state concentrations of enzymes [X] and [Z] determine the steady-state concentration [Y]. The substrate S is replenished to maintain its high concentration, and the waste product, W, is eliminated so that neither limit reaction rates. k_1 and k_2 are rate constants. In the square-root circuit, two molecules of Y react to form W. In the multiplication circuit, the enzyme X produces an intermediate I. Adapted from Buisman et al. (2008). **Upper middle**: Oscillates when enzyme G is activated. J builds up rapidly and also activates two delayed negative feedback loops (dashed line) by promoting the slower buildup of K_P and R_P. R_p depresses J by catalyzing its removal of J and blocking its production. As J falls, K_P and R_P convert back to K and R, negative feedback ceases, and the next cycle starts with the production of J. Adapted from Novák and Tyson (2008). **Upper right**: Autocatalytic switch implicated in synaptic memory storage (chapter 14). The switch protein, CAM Kinase II (CAMKII) has 12 phosphorylation sites. If two sites are phosphorylated by the input, the calcium binding protein CaCM, then CAM Kinase II becomes autocatalytic and attaches more phosphates to itself. Rate of phosphate attachment, *P*+, increases steeply with *nP,* the number of attached phosphates, but then declines at high *nP* as more phosphorylation sites are occupied. The rate of phosphate removal, *P*–, by the phosphatase PPI increases with *np* and saturates at a medium *nP*. Consequently, when CaCM is strong enough to drive CAM Kinase II phosphorylation to the trip point, *T*, where *P*+ > *P*–, autocatalysis drives *nP* to the ON position. Here *P*+ = *P*– and the switch can remain ON indefinitely. When CaCM fails to drive the system to *T*, PPI wins out and removes all phosphates—the switch remains OFF. Adapted from Miller et al. (2005). **Lower middle/right**: Level-detector circuit responds by generating M_p when concentration of [L] lies between $[L_1]$ and $[L_2]$. Two receptor types bind L, high-affinity R_H and low-affinity R_L. LR_H phosphorylates M to active M_P, but LR_L just binds M reversibly At low [L] only the high-affinity LR_H binds, and M_p production increases with L. At high [L] the low-affinity R_L also binds; it outcompetes LR_H for M, so M_p production falls. Adapted from Bray (1995).

transmission. Once the messenger is synthesized and concentrated, it diffuses down its gradient, agitated by thermal bombardment (Brownian motion).

The thermal bombardment that aids diffusion also randomizes movement, and this limits efficiency by introducing noise. Each messenger molecule that reaches a binding site has done so independently of all other messenger molecules; moreover, it has arrived *accidentally* by random walk (figure 2.3). It is the same for a protein designed to deliver information by skating on the membrane (chapters 5 and 8): it finds a receiver by random walk in two dimensions. Moreover, the processes that pass information *through* a protein—binding, allosteric state-transition, catalysis, and release—are also randomized by thermodynamic fluctuations. Therefore, chemical computation in molecular circuits has an associated degree of noise that, as noted in chapter 5, destroys information. Such thermodynamic noise cannot be eliminated, so it must be managed, as we now explain.

Managing noise in a protein circuit

Following the principle *send only what is needed*, a circuit should generally avoid sending noise.[2] Where noise is inevitable, it should be minimized before transmission, so most neural designs try to prevent noise or reduce it at early stages.

Where proteins remain tightly bound in small complexes, signals go directly, thereby avoiding Brownian noise. For more extensive circuits, molecules must move more freely. Now Brownian motion introduces uncertainty. This is reduced by placing proteins close to each other, on the membrane or attached to the cytoskeleton, and by confining diffusible messengers to small compartments. Small compartments also reduce costs—less messenger need be made to produce a signal of given concentration.

By reducing diffusion distances, complexes and compartments shorten delays and lower noise. This occurs where proteins are held together by a protein scaffold—for example, on both sides of a chemical synapse (chapter 7). A presynaptic complex of at least five different proteins (Eggermann et al., 2012) binds a synaptic vesicle and attaches it to the membrane, ready for release. When activated by a surge of calcium, the proteins run through their finite-state routines within 100 μs, to release the vesicle with a minimum of Brownian noise. Postsynaptically, a larger complex of protein species couple to each other and to the membrane. When the vesicle's transmitter molecules cross the 20-nm synaptic cleft and bind a receptor

protein, the change in state triggers a host of postsynaptic protein pathways. This complex occupies a 25- to 50-nm layer beneath the postsynaptic membrane (figure 7.3). Compartments and complexes are used in all chemical synapses, in dendrites (chapter 7), in photoreceptors (chapter 8), and indeed in all cells, to promote economy, to speed responses, and to reduce noise.

Some of the noise associated with changes in a protein's conformational state can be prevented by elevating the barriers on the molecule's energy landscape (chapter 5). Although this reduces reaction rates and hence bandwidth, these can be restored by injecting more energy to drive the process. Thus, there are trade-offs between energy consumption, response speed (bandwidth), and reliability (S/N). This sort of intramolecular noise can also be removed by thresholding with a molecular switch (figure 6.3), but there are three penalties: (1) the high energy cost of having a system full ON when only partial ON would do; (2) the low information capacity of a binary system; and (3) the loss of analogue's ability to process directly. But despite complexes, small compartments, and binary switches, some noise remains. What then?

Noise reducer of last resort

There is another way to reduce noise, or more precisely, to improve S/N. The trick is to replicate a noisy signal, then send the replicates in parallel through multiple components, and sum their outputs. The amplitude of the transmitted signal increases linearly with the number of components, but because their noise is uncorrelated, noise increases as the square root. Thus, with an array of M identical components generating noise independently, the output S/N increases as $\sqrt{M}$. Such a parallel array can increase its S/N to arbitrarily high levels by adding more components. However, the solution must be used as a last resort, and then judiciously, because it is expensive.

The dependence of S/N on $\sqrt{M}$ imposes a law of diminishing returns. Cost rises in proportion to M, but benefit rises as $\sqrt{M}$, so efficiency falls as $1/\sqrt{M}$. Here then is the downside of molecular processing. A single molecule can process near the thermodynamic limit to energy efficiency, but that molecule suffers thermodynamic fluctuations. This noise can be countered with a parallel array of the self-same molecules, but the additional resources consume some of what was saved by operating near thermodynamic limit. Therefore, the best a circuit can do is maximize the efficiency of its parallel array, and this it does by matching the size of the array (M) to the costs associated with the array, and to the S/N of the input.

Maximizing efficiency in a parallel array

To evaluate costs and benefits in the design of a parallel array, we use a general measure of performance, information capacity (Schreiber et al., 2002). An array's information capacity depends on S/N (chapter 5) and increases as $\log_2 (1 + S/N) = \log_2 (1 + \sqrt{M})$. However, the energy cost of passing the signal through the array increases as M. Thus, as M increases, the array's efficiency falls—unavoidably—because the array is redundant: all components try to transmit the same signal. Therefore, efficiency is maximum when $M = 1$. Unfortunately the signal generated by one protein molecule is usually too weak and noisy to be useful.

A more practical optimum emerges upon including the *fixed cost* of building and maintaining the circuit that contains the array. Then, as M increases, information per unit cost of signaling falls through redundancy, but information per unit fixed cost rises. An optimum occurs where these two competing tendencies balance. Consequently, a higher ratio of fixed cost to signaling cost gives a larger optimum array (figure 6.5, inset). The optimum array size also depends on the costs in other parts of the circuit. Where expensive components generate a high S/N and then couple to cheaper components, the cheaper array should enlarge beyond its optimum to retain the hard-won benefit. In general, good design distributes investment among components to maximize performance across the entire system (Alexander, 1996; Weibel, 2000).

A good design does not necessarily optimize an array's efficiency. Initially information capacity and efficiency both rise steeply with M (figure 6.5). But then the capacity curve starts to flatten, and an optimum is reached for given fixed cost where efficiency peaks (figure 6.5). As M rises above the optimum, capacity continues to increase, but efficiency declines, albeit more gradually than it rose. Consequently, an array should set M somewhat above the optimum to reduce the possibility of losing both efficiency and information when unexpected perturbations force it to operate below the optimum. Thus operating at the exact optimum may not be best. Robustness is important, too (Schreiber et al., 2002; Sterling & Freed, 2007).

But what *is* a protein circuit's fixed cost? Given that a circuit's viability requires the whole animal, must one count all vital functions? Although the far end to fixed costs looks hazy, the beginning is certainly clear: it is the cost of making a circuit's protein molecules. The average cost to synthesize an amino acid and insert it into a protein is approximately 5.2 molecules of ATP (chapter 5; Phillips et al., 2009), so to build a typical protein of 300 amino acids costs about 1,700 ATP molecules. Protein delivery and installation are extra. By comparison, the cost per signaling cycle is one to

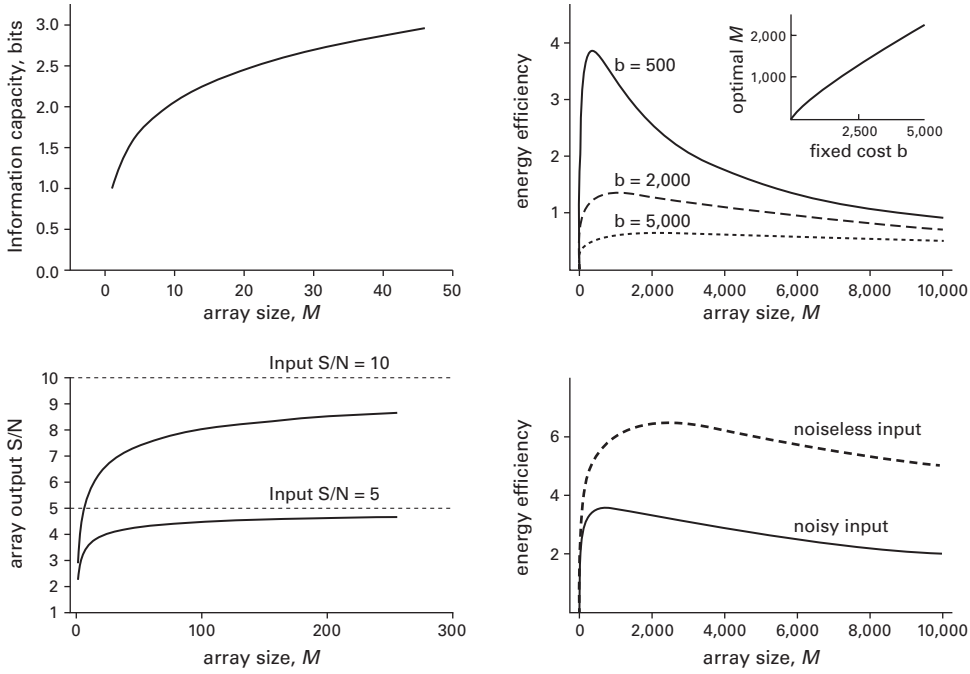


Figure 6.5

Optimizing the noise reducer of last resort—an array of *M* identical components. **Upper left**: Increasing an array's size increases its information capacity with diminishing returns. **Upper right**: Energy efficiency (information capacity/energy cost) is optimized at an array size, *M*, that depends on the fixed cost, *b*. Efficiency is in arbitrary units, *b* is in units of signaling cost. Inset shows how optimal array size increases with fixed cost. **Lower left**: With a noisy input the output S/N cannot exceed the input S/N (dashed lines). Lowering this ceiling reduces the advantage of larger arrays. **Lower right**: Reducing input *S/N* reduces the size of the optimum array. Upper and lower right redrawn from Schreiber et al. (2002). Upper and lower left calculated using their formulae.

five ATP molecules (chapter 5), and this suggests a rule of thumb: the cost of operating a protein molecule (signaling cost) will exceed its construction cost when the molecule has completed 500–1,000 signaling cycles.

Returning to efficiency, the S/N of an input profoundly affects the array's optimum size. The array cannot reduce input noise but can only let noise cancel by averaging. Consequently, input noise imposes a ceiling to be approached by the array's S/N. This reduces the efficacy of a large array at low input S/N (figure 6.5) and the size of the most efficient array (figure

6.5). In other words, because an input with low S/N contains less information, and a smaller array has a lower information capacity, the optimum array matches its capacity to its input.

The matching of array size to input S/N follows the principle of symmorphosis (Weibel, 2000), whereby capacities match within a system to avoid waste. What was illustrated for flow of oxygen through lungs, heart, vessels, and muscle (figure 3.4) applies equally to the flow of information through an array of protein molecules. We will see that symmorphosis also holds for parallel arrays of ion channels in a membrane (below), for photoreceptors in a retina (chapter 8), for synapses in a neural circuit (chapter 9), and for neurons in a pathway (chapter 11).

Summary: Pros and cons of computing with chemical circuits

A chemical circuit processes information efficiently on several counts. Operating near the thermodynamic limit it is energy efficient, and its molecules makes efficient use of space and materials. Chemical computation is direct (analogue), which uses fewer steps than digital. Chemistry is wireless, which reduces space and energy for transmission and, by making it easier to form new connections, facilitates behavioral plasticity and evolutionary innovation. A downside is noise, which is handled in four ways. Some Brownian noise is avoided by coupling proteins in complexes and small compartments; some thermodynamic noise is avoided by raising intramolecular energy barriers; and some noise is removed by molecular switches. Unavoidable noise can be mitigated by signaling with parallel, redundant components that add n signals linearly and noise as the square root.

The cost of signaling increases with the concentration of the messenger. Therefore, efficiency might seem to favor high-affinity receptors that bind at low concentration. Yet, there is a penalty and, hence, a trade-off. High-affinity receptors decrease signal bandwidth by slowing the rate at which a signal decays. Low-affinity receptors need higher concentrations, which cost more but, releasing the ligand faster, provide higher bandwidths (Attwell & Gibb, 2005). Thus, speed and bandwidth consume materials and energy, making it advisable to send at the lowest rate.

Despite the advantages of chemical computing, there remains the important proviso *compute with chemistry wherever possible*. Chemistry is fast at the nanometer scale, but because diffusion slows and dilutes signals, chemistry beyond a few microns is too slow to coordinate immediate behavior. Thus, as for *Paramecium* (chapter 2), the need for speed over distance forces a more expensive option—protein circuits that process information electrically.

Information processing by electrical circuits

How electrical circuits meet the need for speed over distance

Electrical current in a silicon device is carried by electrons, but in a biological device it is carried by ions. The cell membrane, comprising a bilayer of nonpolar lipid, is impermeant to ions, so it separates charge, sustains a voltage difference across it, and has a capacitance of about 1 μF cm^{-2}. Charging the membrane's capacitance constrains the speed of electrical signaling. The membrane's time constant, τ, is its resistance times its capacitance, RC, so τ can be shortened to speed up the signal by shrinking the membrane area and by reducing its resistance to the passage of ions.

An ion passes through the membrane via a channel (Hille, 2001); a large protein molecule assembled as a ring of subunits to form an aqueous pore in the membrane (figure 6.6). The pore is constructed to selectively pass particular ion species in single file by adjusting its width and strategically positioning charged amino acid side groups. A typical sodium ion channel is 10 times more permeable to sodium than to either calcium or potassium, and a potassium channel is more selective still—100-fold more permeable to potassium than to sodium, and almost totally impermeable to calcium.

The channel's energetically stable conformation sets it either closed or open. And thus it remains until a specific input, such as a ligand binding or a change in membrane potential, and/or thermal fluctuations cause the channel to open or close, allosterically. Any net transfer of charge through a channel changes the voltage across the membrane. This voltage signal transmits further and faster along the membrane than chemical diffusion allows, millimeters in milliseconds. But although allostery allows a cheap input, a channel's ionic current is an expensive output, as we now explain.

To charge the membrane quickly, ions must be driven through channels at high rates. The primary driving force is a concentration gradient maintained across the membrane by ion pumps (figure 6.7). Most important is the sodium–potassium pump, which maintains low sodium concentrations and high potassium concentrations inside the neuron. This pump is a molecular machine, a protein complex spanning the membrane which hydrolyzes one ATP molecule to export three sodium ions and import two potassium ions. This asymmetrical exchange generates an outward current of one positive charge per pump cycle and sets up the two concentration differences, $[K]_{in} > [K]_{out}$ and $[Na]_{in} < [Na]_{out}$. These two gradients power most of the brain's electrical circuits. Consequently, the sodium–potassium pump consumes 60% of the brain's energy (Attwell & Laughlin, 2001).

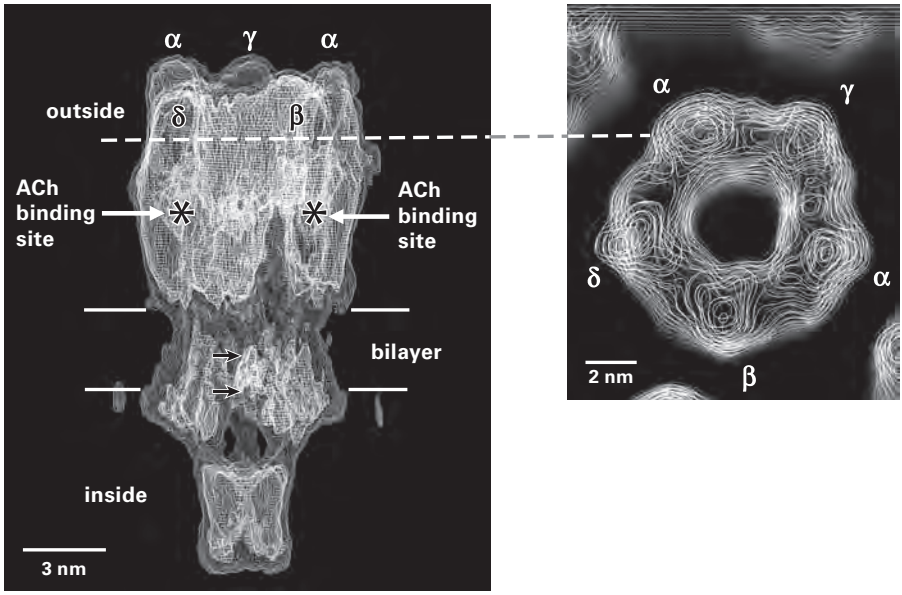


Figure 6.6

An ion channel is a large protein with a pore that conducts ions across the membrane. Ligand gated channel from the electric organ of a torpedo ray opens to admit sodium ions and potassium ions when it binds two molecules of the neurotransmitter acetylcholine, Ach. **Left:** Channel imaged side-on. The channel is formed by a ring of five protein subunits, two αs, β, γ, and δ. All contribute to the extracellular vestibule, the narrower pore that crosses the membrane's lipid bilayer, and the intracellular domain. Asterisks show binding sites for neurotransmitter acetylcholine on the two α subunits. When both bind the channel opens and passes sodium ions and potassium ions. Large intracellular domain has phosphorylation sites for modulating channel's sensitivity. **Right:** Cross section through channel at level indicated on left by dashed line. Three-dimensional structure of channel reconstructed from electron micrographs of crystalline channel arrays, with a resolution of 0.4 nm. Image courtesy of Nigel Unwin. Further details in Unwin (2013).

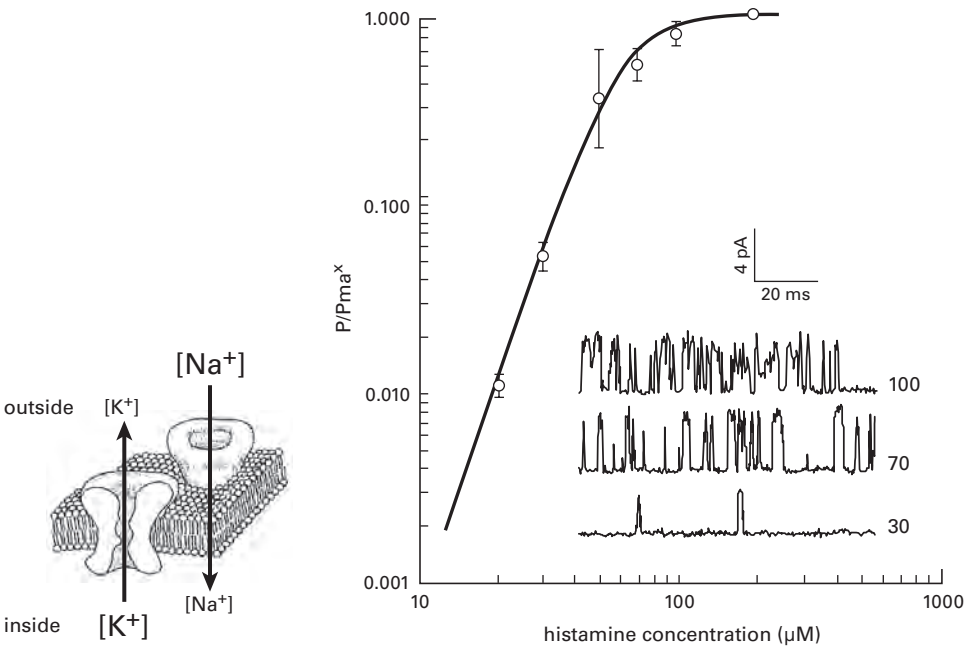


Figure 6.7

Concentration gradients drive ions through channels that open and close rapidly in response to a specific input. **Left**: Sodium and potassium ions cross the membrane through ions channels, driven by concentration gradients. **Right**: A chloride ion channel opens to pass ~4 pA of current when it binds the neurotransmitter histamine. Currents recorded from a single channel, by patch clamp, at three histamine concentrations: 30, 70, and 100 μM. The open probability increases with histamine concentration according to the binding equation, 6.2, with cooperativity $n = 3$. Channel recorded in membrane of a large monopolar cell from the fly lamina (chapter 9). Left, after Hille (2001). Right modified and reprinted with permission from Hardie (1989).

The concentration gradient is equivalent to a battery whose voltage drives ions through the channel at the same rate (figure 6.8). The battery's voltage is given by the Nernst equation, which converts the chemical potential of the concentration difference into an equivalent electrical potential. Thus, for ionic species, x, its battery's voltage is

$$E_x = RT/(zF)\ \ln([X]_o/[X]_i) = 2.303\ RT/(zF)\ \log([X]_o/[X]_i), \qquad (6.3)$$

where $[X]_o$ and $[X]_i$ are the concentrations of ion x outside and inside the cell, z is its charge, R is the universal gas constant, T is the temperature in Kelvin, and F is Faraday's constant.

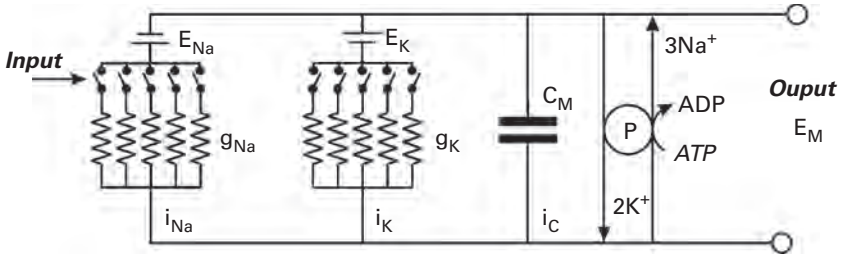


Figure 6.8

The simple resistor–capacitor (RC) circuit formed by ion channels in the neuronal membrane. The input opens sodium channels, and the output is the membrane potential, E_M. A bank of potassium channels, each with conductance g_K, passes outward current i_K, driven by the potassium ion battery E_K. Without input, the potassium channels maintain a *resting potential* of E_K. Input opens sodium channels, g_{Na}, which, driven by the sodium battery, E_{Na}, pass inward current, i_{Na}. To change the output, E_M, the membrane capacitance, C_M, is charged and discharged by the capacitative current, i_C. Sodium-potassium pump, P, keeps batteries charged using energy obtained from hydrolysis of one molecule of ATP to ADP to export 3 sodium ions and import 2 potassium ions, thereby generating an outward pump current.

The two ionic batteries that dominate electrical signaling, potassium with E_K ~ –85 mV and sodium with E_{Na} ~ +50 mV, provide a dynamic range of about135 mV. A neuron exploits this to the fullest when it generates its fastest signal, an action potential. Before the action potential the neuron is at rest. Mainly potassium channels are open, and the membrane potential sits close to E_K. Here a sodium ion experiences its maximum force, pulled inward by a membrane potential of –85 mV, and pushed inward by a concentration difference equivalent to +50 mV. So when a sodium channel opens to initiate an action potential, sodium ions surge in, driven by 135 mV, and their powerful current helps meet the need for speed.

Less than a millisecond later, when the action potential peaks close to E_{Na}, a potassium ion experiences its maximum force, so when a potassium channel opens to return the membrane to rest, potassium ions surge in, driven by 135 mV. Again, this helps meet the need for speed by increasing the power of the potassium current.

To improve power delivery, a channel's bore is designed to transmit rapidly: ions pass at rates up to 10^8 s^{-1} (Williamson, 2011). These are the highest output rates known for protein molecules (Hille, 2001). By comparison, the fastest chemical output by an enzyme (carbonic anhydrase) is 20-fold slower, and most enzymes are 100-fold slower (Williamson, 2011).

Chemical signaling by molecules, such as ligand-binding receptors and G proteins, operate slower than an ion channel by 4 to 7 orders of magnitude. With its exceptional output rate, a voltage-gated sodium channel opening for 1 ms admits 6,000 Na^+ ions. This 1 pA ionic current delivers 2.4×10^4 k_BT joules, giving a power rating of 200 fW.

Fast processing also requires molecules that switch quickly. Channels are structured to open or close in tens of microseconds (figure 6.7)—near the limits of allosteric state change (Chakrapani & Auerbach, 2005). The energy used to open a channel, ~25 k_BT joules (Chowdhury & Chanda, 2012), is 35 times the thermodynamic minimum for a bit (chapter 5), high enough above to be reliable, but low enough not to put too much of a brake on processing speed. With an input energy of 25 k_BT joules and an output of 2.4×10^4 k_BT joules, a sodium channel opening for 1 ms has a power gain ×1,000. Thus, a channel's combination of sensitivity, fast switching, and gain satisfies the need for speed. But as noted, it comes at a price.

The price is paid to keep ionic batteries fully charged. An ion passing through a channel drops its battery's voltage by reducing the concentration gradient (equation 6.3). The gradient is restored by pumping the ion back across the membrane, so when a sodium channel opens for 1 ms and admits 6,000 Na^+ ions, sodium-potassium pumps hydrolyze 2,000 ATP molecules to ADP to pump these ions back. The efficiency of the conversion of the chemical energy supplied by ATP to the electrical energy delivered by the channel is reasonably high, 50%.[3] Nevertheless, a channel's signaling cycle (open, admit ions for a millisecond, close, restore ions) uses 2,000 times more ATP than a G protein's cycle. This is the price paid for speed over distance.

In summary, an ion channel changes a neuron's membrane potential rapidly by operating as a power transistor that is irreducibly small and operates close to thermodynamic limits. Engineers seek similar efficiency savings by developing their version of a single molecule power transistor. Biology evolved this device over a billion years ago and solved the not inconsiderable problem of connecting its molecular "transistors" to form circuits.

How circuits built from ion channels operate electrically

Ion channels naturally form electrical circuits because they connect two lower resistances (extracellular space, cytoplasm) across an insulating membrane. Consider the simplest circuit, two types of ion channel working against each other to code an analogue input as an analogue output, namely, a change in membrane potential, E_M (figure 6.8).

The circuit's behavior is captured by an electrical model in which each channel is a switched resistor, connected to its battery (figure 6.8; Koch, 1999). The resistor represents the channel's conductance, g, (conductance = 1/resistance) and the switch opens the channel. For a channel that passes ions of species x, the current, i_x, is given by Ohm's law:

$$i_x = g_x\,(E_x - E_m), \tag{6.4}$$

where E_m is membrane potential, E_x is the electromotive force (EMF) of the ionic battery (equation 6.3), and g_x is the single-channel conductance for ion x. Note that when $E_M = E_x$, there is a tipping point where the direction of current reverses. This point is used to determine E_x experimentally, so it is often called the *reversal potential*.

For ion channels to change the membrane potential, they must charge and discharge the membrane's capacitance (~1 μF cm^{-2}), represented in the model by the capacitor, C_M. The fourth component, the sodium–potassium pump, P, hydrolyzes ATP to keeps the ionic batteries charged. Because the rate at which the pump exchanges three sodium ions for two potassium ions is effectively independent of membrane potential, it is treated as a constant current source.

This RC circuit model describes how the membrane potential changes when channels open and close. Applying Kirchoff's law,

$$i_{Na} + i_K + i_C + i_P = 0, \tag{6.5}$$

where i_C is the capacitative current and i_P is the pump current. Substituting for the currents flowing through the channels and the capacitor,

$$(E_{Na} - E_M)N_{Na}g_{Na} + (E_K - E_M)N_K g_K + C_M\, dE_M/dt + i_P = 0, \tag{6.6}$$

where N_{Na} and N_K are the numbers of open sodium channels and open potassium channels. Because the pump maintains the concentration gradients for sodium and potassium, $i_P = 0.5\ i_K$, giving

$$(E_{Na} - E_M)\ N_{Na}g_{Na} + 3/2(E_K - E_M)\ N_K g_K + C_M\, dE_M/dt = 0. \tag{6.7}$$

This current-balance equation captures the biophysics of electrical signaling across a neural membrane and easily extends to include other channels (including ones that depend on time and voltage), other pump currents, and currents generated by ion exchangers. Consequently, an equation of this form is the core of the many more complicated models of electrical interactions in neurons (Hodgkin & Huxley, 1952; Koch, 1999). One insight is that this irreducibly simple circuit is inherently *self-shunting*. That is, current driven through a channel pushes the membrane voltage toward the channel's reversal potential, thereby progressively diminishing the current

passed per channel as more channels of this type open. This nonlinear behavior shapes the circuit's I/O function and supports information processing.

I/O function of the basic circuit

To explain the circuit's I/O function we drive it with an input that opens sodium channels. Sodium ions enter, pushing E_M toward the positive potential of the sodium battery. This shift in voltage encodes the input intensity, I, as an output. To derive the relationship between input and output, assume that the input acts linearly, so the number of open sodium channels is

$$N_{Na} = aI, \tag{6.8}$$

where a is channel gain, in open channels per unit input. Thus, the sodium conductance is

$$G_{Na} = g_{Na}N_{Na} = g_{Na}aI. \tag{6.9}$$

The opposing potassium conductance is held constant, $G_K = g_K N_K$, where g_K is the conductance of a single potassium channel and N_K is the number of open potassium channels.

The circuit's I/O function now follows. Without input, $G_{Na} = 0$, and the circuit rests with $E_M = E_K$. A step rise in I opens aI sodium channels whose inward current charges the membrane capacitance to a new steady voltage with a time constant

$$\tau_M = C_M R_M, \tag{6.10}$$

where R_M, the membrane resistance, is $1/(G_{Na} + G_K)$. This steady state is reached long before pump currents change because they are slow (see below) whereas τ_M is typically milliseconds; consequently $i_C = i_P = 0$. Solving the circuit's current balance equation gives the new steady-state membrane potential

$$E_M = (G_{Na}E_{Na} + G_K E_K)/(G_{Na} + G_K). \tag{6.11}$$

Dividing through by G_K, we see that E_M depends on the conductance ratio, G_{Na}/G_K,

$$E_M = (E_{Na}G_{Na}/G_K + E_K)/(G_{Na}/G_K + 1). \tag{6.12}$$

This relationship is simplified by expressing the voltage output relative to a baseline of zero input so that $output = E_M - E_K$, then normalizing output to its maximum, $output_{max} = E_{Na} - E_K$. Note that the setting of E_K to zero simply

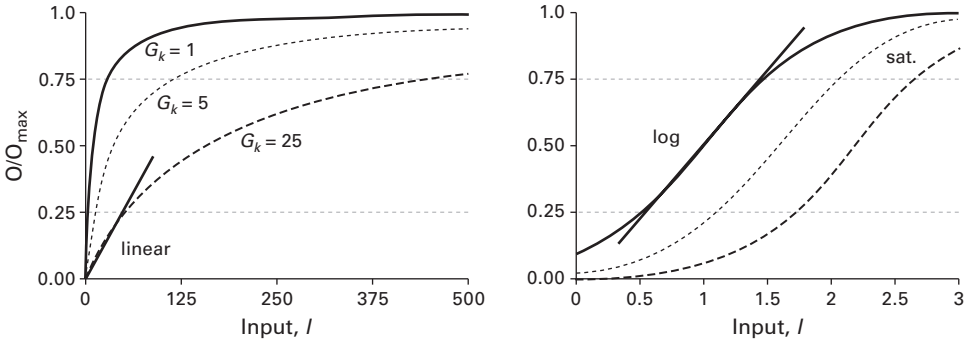


Figure 6.9

Input/output (I/O) function generated by the basic electrical circuit allows the same computations across different input ranges by changing the shunting conductance G_K. Normalized output, O/O_{max}, is plotted against input, I, for three different shunting conductances. **Left**: When output is small (<0.25 max), the I/O function adds. **Right:** When the output is medium (0.25–0.75 max), the function is logarithmic. In saturated regime (0.75 max – max) function's slope approaches zero. Note similarity with I/O function produced by chemical binding (figure 6.2).

shifts the voltage scale without altering the EMFs experienced by ions, so response amplitudes are unaffected. Now

$$output/output_{max} = (G_{Na}/G_K)/(G_{Na}/G_K + 1). \quad (6.13)$$

Substituting aIg_{Na} for G_{Na}, we obtain a simple form of the circuit's I/O function

$$output/output_{max} = kI/(kI + 1), \quad (6.14)$$

where the gain factor $k = ag_{Na}/G_K$. The electrical circuit's I/O function is hyperbolic (equation 6.14; figure 6.9), like the I/O function for chemical binding, because it too saturates. And like the chemical circuit, the electrical circuit's hyperbolic I/O provides operators for processing information (Koch, 1999; Silver, 2010).

An electrical circuit's hyperbolic I/O supports six operators

1. *Addition* (A + B) occurs when the circuit operates in the bottom quartile of the I/O function where it is approximately linear (figures 6.2 and 6.9), When inputs A and B open the same species of ion channel, they add.
2. *Subtraction* (A – B) also occurs in this linear region when input A opens an ion channel that carries current inward and B opens a channel that

carries a current outward. The changes in conductance and voltage must be small enough for the channels to approximate constant current sources driving a constant load.

3. The *log* transform occurs in the middle region of the I/O function, where *output* ~ *k*log *I* (figure 6.9B). As with chemical circuitry, this log transform is widely used in sensory circuits to scale responses to changes in input level, so that a constant $\Delta I/I$ produces equal changes in output throughout this logarithmic range.

4 & 5. *Multiplication* (×) *and division* (÷) are performed by changing the gain factor, *k*, in the I/O function (equation 6.14). This can be accomplished by altering the channel gain (*a*) and/or the potassium conductance (G_K). For example, increasing G_K shunts the input from G_{Na}. This mechanism is widely used for multiplicative gain control and divisive normalization (chapters 8 and 12), procedures that optimize coding and facilitate the extraction of patterns (Koch, 1999; Carandini & Heeger, 2012). Changing channel gain, *a*, does not, strictly speaking, multiply and divide within the circuit, but it has this effect on the I/O function. The important distinction for design is that increasing G_K increases both signal quality (S/N, bandwidth) and energy consumption by increasing the number of open channels, whereas reducing *a* reduces signal quality and energy consumption by reducing the number of open channels.

6. *Exp* (inverse of log) is implemented by installing cooperativity in ion channels—for example, by requiring that *n* binding sites be occupied to open a ligand-gated channel. As in chemical circuits, cooperativity raises the output to the *n*'th power of the input, so steepening the I/O curve and shifting it to higher input levels. Cooperativity is used at blowfly photoreceptor output synapses to match a neuron's coding function to the range of input levels (figure 3.4). The neurotransmitter, histamine, must occupy 3 binding sites to open a postsynaptic chloride channel. This steepens the I/O function (figure 6.7) to help achieve a match with the probability distribution of input signals (figure 9.10).

How electrical circuits support analogue processing

Ion channels implement the four elements of analogue electrical circuits, resistance, R; capacitance, *C*; inductance, *L*; and memristance, *M* (Chua, 1971). Resistance and capacitance are obvious (figures 6.7 and 6.8), but the uses of inductance and memristance need explanation. With an inductance the voltage is proportional to the rate of change of current. Thus, when the current is increasing more rapidly, the voltage is larger, and this

advances the phase of the response to a sinusoidal input. Voltage-gated potassium channels advance phase by means of delayed negative feedback (Koch, 1999).

A memristor changes its resistance in proportion to the quantity of charge it has conveyed and then holds this resistance when charge stops flowing (Strukov et al., 2008). This resistance with memory is provided by a channel that couples electrical signaling to chemical signaling. For example, take an ion channel that passes mostly sodium with a little calcium. This calcium provides a measure of the total charge flowing through the channel. Arrange that calcium binds to the mechanism that opens the channel, and alters its open probability. Now one has a memristor in which charge entry couples to the channel's effective conductance. Photoreceptors use this mechanism to control their gain (chapter 8).

How voltage-gated channels meet a need for speed over distance

A voltage-gated channel opens or closes allosterically, in response to membrane potential. Thus, a voltage-gated channel can be activated within milliseconds by channels opening millimeters away. In addition, a voltage-gated channel amplifies an electrical input. By virtue of these properties, voltage-gated channels can produce a larger signal that transmits more quickly and reliably than the signals generated by ligand-gated channels—most notably an action potential (figure 6.10).

A typical action potential, an approximately 100-mV pulse lasting about 1 ms (figure 6.10), is produced by a large and sudden influx of sodium ions followed by a similar efflux of potassium ions. These currents are produced by sodium channels and potassium channels (figure 6.10) that, gated by depolarization, generate the action potential and propagate it along the membrane at speeds of 0.3–80 mm ms^{-1} without loss of amplitude.

The voltage-gated channels generate the action potential as follows (figure 6.10). At resting potential, typically –70 mV to –60 mV, the voltage-gated channels for sodium and for potassium open with a low probability. When an analogue input depolarizes the membrane, the open probability increases and a small proportion of sodium channels opens immediately. Driven by their maximum force, sodium ions surge in and depolarize the membrane further, creating a positive feedback loop (figure 6.10). Almost all of the voltage-gated potassium channels remain closed because they respond to depolarization more slowly. A longer activation time constant is programmed into their finite state transitions to keep them closed while the sodium channels are starting to open. This delayed opening increases

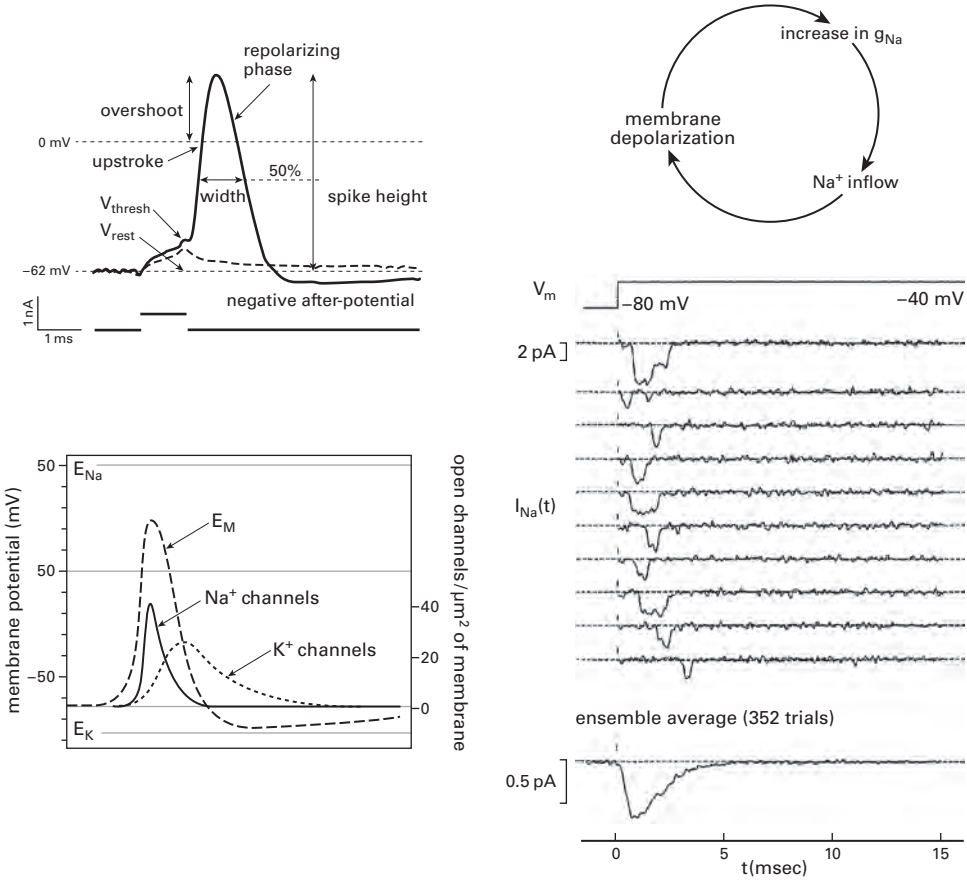


Figure 6.10

Voltage-gated sodium channels and voltage-gated potassium channels meet the need for speed by producing an action potential. **Upper left**: Action potential waveform. Spike initiated when suprathreshold current depolarizes membrane potential from resting potential, V_{rest} to threshold, V_{thresh}. Fast upstroke overshooting to peak height and repolarizing phase complete rapidly to produce spike with narrow width (measured at 50% spike height). Slower negative after-potential follows. **Upper right**: Positive feedback loop that accelerates spike upstroke and drives overshoot to maximum amplitude. Increase in voltage-gated sodium conductance, g_{Na}, increases inflow, depolarizes membrane and increases voltage-gated sodium conductance. **Lower left**: Time course of spike (E_M, left axis) and voltage-gated sodium and potassium conductance, plotted as density of open channels (right axis). The rapid increase in the number of open sodium channels that drives the upstroke is short-lived because sodium channels quickly inactivate. The voltage-gated potassium channels open more slowly to repolarize, and generate the negative after-potential. **Lower right**: Recordings of the activity of two voltage-gated sodium channels show that, following a step depolarization, each opens with a randomly varying latency for a randomly varying time. Averaging 352 individual responses demonstrates that a large array of channels averages out noise to produce a reliable sodium current. Upper left redrawn from Bean (2007). Upper right and lower left from Shepherd (1994) with permission. Lower left, data from J. B. Patlek, plotted after Hille (2001), with permission.

the efficiency with which sodium channels charge the membrane capacitance by preventing the charge being carried in by sodium from being negated by charge carried out by potassium. Blocking this futile cycle allows the action potential to develop and, by reducing the number of ions crossing the membrane, saves pump energy.

At a critical level of depolarization, the *threshold* potential (figure 6.10), sodium's positive feedback takes off. All available sodium channels open (figure 6.10), more sodium ions surge in, and, unopposed by the more sluggish potassium channels, their current depolarizes the membrane toward equilibrium potential (E_{Na} = 50 mV) in less than 1 ms. As the membrane potential approaches this peak, large numbers of voltage-gated potassium channels are starting to open (see figure 6.10). Potassium ions experience their maximum force and surge out, driving the membrane potential back down, toward rest. At the same time, the open sodium channels change conformation and lock shut. This *inactivation*, programmed into a sodium channel's state changes, stops incoming sodium ions from negating the charge being carried by outgoing potassium, thereby increasing efficiency. The voltage-gated potassium channels drive the membrane potential to resting potential within 0.5 ms and, being no longer depolarized, start to close. But because potassium channels change their state more slowly, many remain open; the membrane potential dips below rest and approaches E_K, creating a negative afterpotential (figure 6.10).

While potassium channels are repolarizing the membrane, the voltage-gated sodium channels remain inactive. To reset to its initial state (closed but responsive to depolarization), a sodium channel must experience the strong negativity of potentials close to rest. This state change is programmed to have a time constant of ~3 ms. The resulting delay, plus the residue of open potassium channels, makes it impossible to trigger another action potential during a *refractory period* of 2 ms.[4] Although being refractory places a ceiling on action potential frequency, it ensures that an action potential cannot trigger a resurgent sodium current during its repolarizing phase. This prevents a single action potential from starting a continuous train of spikes.

In summary, an action potential is the product of three electrical feedback loops, all formed by voltage-gated channels. Sodium's positive feedback loop depolarizes the membrane to the action potential's peak (figure 6.10), and potassium's delayed negative feedback repolarizes to rest. Speed and efficiency are enhanced by a third negative feedback loop, mediated allosterically by sodium channel inactivation. Because channels gate each other electrically, the action potential is brief. This increases timing

precision and hence the number of bits carried by an action potential (chapter 3). Being electrical, an action potential travels rapidly along a neuron's membrane at speeds up to 100 mm in a millisecond (chapter 7) yet retains its information because it is faithfully regenerated by feedback. But how can the information carried by such an electrical signal drive a chemical circuit? The answer is a voltage-gated channel with a chemical output.

How a voltage-gated calcium channel links electrical to chemical

A voltage-gated calcium channel admits an ion that readily binds a protein and changes its conformation. As noted in a chemical synapse, calcium entering via channels opened by presynaptic depolarization binds the protein synaptotagmin, which then changes conformation and triggers vesicle release. A calcium ion is especially effective at changing a protein's conformation because, being divalent, it pulls negatively charged parts of a protein closer together.

Calcium is especially effective as a chemical messenger because cells pump it out to keep the internal concentration low, 30–200 nM. This creates a steep concentration gradient, equivalent to a battery of 130 mV that, aided by the –70 mV resting potential, drives calcium in through a channel at a rate of ~10^7 ions per second. With so little internal calcium, the proteins within nanometers of the channel experience a 100-fold increase in calcium concentration within 100 μs. This nanodomain calcium signal has a wide bandwidth because it decays as rapidly as it rises. The puff of calcium injected by a channel vanishes within 500 μs by diffusing rapidly into a large sink, the well-buffered bulk of the cell's cytoplasm. Viewed from the channel's nanodomain, this rapid removal mechanism comes for free. The calcium puff is mopped up by buffering proteins, distant pumps, and exchangers.

In summary, the simplest electrical circuits demonstrate how the brain satisfies the need for speed over distance. Whereas chemical signaling can send information in a millisecond, but only over 1 μm, passive electrical signaling can send it a millimeter in the same time—1,000-fold faster. Active electrical signaling (action potentials) can send it still faster, by another 100-fold, over much longer distances. Electrical circuits can be constructed to use the same operators as chemical circuits (figure 6.2; cf. figure 6.9). But operating more rapidly over longer distances requires more power. An electrical circuit consumes orders of magnitude more energy than a chemical circuit and, because electrical signaling uses wires, costs more space.

Given the costs, one expects efficient design. Since ion channels are allosteric proteins and operate stochastically, they present the same issues of

S/N, bandwidth, and redundancy that were identified for chemical circuits. We should also expect the same need to match output to input—symmorphosis. Moreover, when short-range chemical circuits have filtered signals into parallel streams with different S/N and bandwidth and information content, the electrical circuits that relay this information rapidly over distance need to match these inputs with appropriate outputs. This requires a diversity of ion channels with subtly different sensitivities and speeds—that is, access to the large "part list" contained in the genome.[5]

Constraints on information processing by circuits of ion channels

Biophysical constraints

Three biophysical factors limit the performance of electrical circuits formed by ion channels: (1) the high electrical resistance of single channels, (2) membrane capacitance, and (3) channel noise from thermal fluctuations in single proteins.

First, channel resistance. Despite having a high transport number for a protein molecule, a single channel nevertheless has a high resistance, $R_{Ch} \sim 10^{11}\ \Omega$. The reason is that selectivity requires the ions to pass in single file. Driven by a typical range of voltages, 10 mV–100 mV, a channel passes 0.1–1 pA. In a neuron with typical input resistance, $10^8\ \Omega$, such currents are sufficient to change membrane voltage by 10 μV to 100 μV. That's not much. For example, the voltage change needed to reliably trigger an action potential is about 1 mV–10 mV, that is, 10 to 1,000-fold larger. Moreover, voltage decays exponentially with distance, so a single channel's signal soon disappears in the membrane voltage noise. A larger voltage signal will travel further and support a workable S/N at its destination, and this is easily achieved (equations 6.9 and 6.11; figure 6.9)—by opening more channels.

Second, membrane capacitance. As noted, capacitance limits a signal's rate of change. One channel, passing 0.5 pA, charges the membrane slowly, and this limits temporal frequency and bandwidth. For example, one channel charges the 314 μm^2 membrane of a spherical neuron, 10 μm in diameter, with a time constant of 88 ms, giving a bandwidth of 12 Hz. This limit too can be raised—by opening more channels.

Third, channel noise. Channels, like other proteins, change conformational state stochastically because they are subject to thermodynamic fluctuations. Therefore, a channel opens and closes stochastically with probabilities that depend upon its input (figure 6.7). This stochastic

opening adds noise. The ratio of signal to noise can be improved—by opening more channels.

Channels operating in an electrical parallel array, as in figure 6.8, obey the same rule as molecules in a chemical array (figure 6.5). The S/N of an array of M parallel channels increases as $\sqrt{M}$, and as M increases, efficiency falls. Consequently, an efficient electrical circuit will match its number of channels to three factors: fixed cost, costs of other signals in the circuit, and input S/N (figure 6.5). In summary, one adjustment, opening more channels, improves four measures of performance: signal amplitude, signal bandwidth, S/N, and information capacity (equation 5.6). So, what constrains the numbers of channels that a circuit can employ to improve its performance?

What limits the number of channels in a circuit?

A circuit could maximize its performance by maximizing the number of channels it uses. Some parts of protein circuits (e.g., ligand-gated channels on a postsynaptic membrane) achieve this locally by packing channels in the cell membrane as a crystalline array (~2.5×10^3 channels per μm^2). This produces tremendous local currents which charge the membrane with extreme rapidity, a design used by the electric eel to discharge its electric organ. However, such a power drain could not be sustained globally across an entire neuron.

The number of channels is limited by membrane space for pumps. A pump molecule has approximately the same footprint as a channel, but, operating at 200 cycles s^{-1}, it extrudes only 600 sodium ions s^{-1}. To match the throughput of one open sodium channel (6×10^6 sodium ions s^{-1}) requires 10,000 pump molecules, which occupy 4 μm^2 of membrane. Thus the density of *open* channels that a neuron can sustain is reduced to one channel per 4 μm^2, 10,000-fold less than their maximum packing density. This translates into a 10,000-fold lower bandwidth and a 100-fold reduction in S/N. Being proportional to bandwidth and $\log_2$ (S/N), the sustainable information rate is cut by almost five orders of magnitude. Placing the circuit's battery chargers (pumps) alongside the circuit's transistors (ion channels) limits a neuron's ability to process information, but cell biology offers few alternatives.[6]

Were a neuron to fully pack its membrane with channels and their obligatory pumps, could it power them? The essential ATP is generated within the neuron by mitochondria. These occupy space, so the maximum sustainable ATP production is proportional to cytoplasmic volume and mitochondrial density. Typically 4×10^5 ATP s^{-1} can be generated per μm^3 (based

on a specific metabolic rate for cortical neurons of 40 μmoles ATP/g/min; Attwell & Laughlin, 2001), which means that generating the power for one open sodium channel requires about 5 μm^3 of cytoplasm. Thus, when operating at the pump limit of one open channel per 4 μm^2 of membrane, 5 μm^3 of cytoplasm is required to provide the pumps' ATP, giving a surface area to volume ratio of 1:1.25. Therefore, a spherical neuron must be greater than 7.5 μm in diameter to operate at the pump limit, but a smaller sphere has a larger surface area:volume ratio, so it is limited by the ability of mitochondria to generate ATP. Many neuronal cell bodies have diameters greater than 7.5 μm, but to connect efficiently they branch (chapter 13), and this increases surface area:volume. Thus, a pyramidal neuron, with a surface area:volume ratio of about 3:1, cannot reach the pump limit to open channel density. Forced to operate with fewer open channels, it must reduce the rate, temporal precision, and accuracy of its electrical signals. Housing the system that burns fuel to supply energy also limits a neuron's processing power, but again, that's cell biology.

In short, the molecular power transistor (ion channel), its molecular battery charger (ion pump), and its intracellular power station (mitochondrion) prevent the brain from reaping a major benefit of irreducibly small molecular components, high-density computing. Thus, unlike conventional engineering design, neural design must maximize performance at low-power density. Given that opening more channels inevitably costs space—membrane area for pumps and cytoplasmic volume for mitochondria—it is all the more critical to open the minimum number of channels required to meet functional specifications. To paraphrase a now familiar principle, a low-energy-density brain should send information with the lowest rate of channel opening.

Providing speed and accuracy with low energy density circuits

Given that low energy density limits the minimum time constant and maximum S/N by limiting the number of open channels, how can a brain respond quickly and accurately? A solution adopted by most brains is to open many channels infrequently in concentrated groups—that is, use powerful signals that are sparsely distributed in space and time, as happens with action potentials and synapses.[7] This design leads to an apparent paradox. These concentrated electrical signals are costly and consume most of the brain's energy, so they are part of the problem, but, given the need to send accurate signals far and fast, they are also part of the solution.

Although concentrated bursts promote temporal precision by increasing S/N and reducing the membrane time constant, their spatial and temporal

sparsity enforces low mean rates. For example, the power density of cortical gray matter limits the mean firing rate, averaged across all classes of cortical neuron, to less than 10 Hz (Attwell & Laughlin, 2001; Lennie, 2003; Sengupta et al., 2010; Howarth et al., 2012). How the brain manages to process information effectively within this limit is a major theme in neural design.

Can arguments based on energy density be extended to establish an upper limit to a brain's processing power—as bits per volume per second? Possibly, but this would only consider the expensive electrical signals. Over short distances and longer times, chemical processing is orders of magnitude cheaper, thus the principle *compute with chemistry*. Chemical and electrical circuits can process information with similar operators, but at scales and costs that differ by orders of magnitude. Therefore the design task for a neuron is to integrate across these scales to achieve the best result in space, time, and energy. This is the subject of chapter 7.

7 Design of Neurons

Chapter 6 explained that much of the brain's computing occurs by chemistry at the scale of single protein molecules and protein circuits. Computing by chemistry offers good S/N at irreducibly low cost in space and energy. Moreover, where the reaction vessel shrinks, the principle of mass action can operate on high concentrations with small numbers of molecules. High concentrations allow low binding affinities to achieve useful signaling rates. Small volumes also shorten distances—over which diffusion is rapid. Also, because concentrations of diffusing molecules decay steeply in space and time, many computations can be accomplished wirelessly—simply by placing detectors at different distances from a source and letting Brownian motion do the math.

Computing with proteins allows a nearly infinite parts catalog—because a protein can be customized by changing a single amino acid—and that is effected simply by swapping a single base pair in the DNA. Thus, natural selection can shape every component precisely for a specific task—for example, to match a particular binding affinity and a particular cooperativity to a particular signal (figures 6.2 and 6.3). The ease of adjusting protein structure has generated immense diversity: overall, the mammalian brain transcribes 5,000 to 8,000 genes and uses alternative splicing to produce 50,000 to 80,000 distinct proteins.[1]

Chemical computing works brilliantly across a spatial scale of nanometers to micrometers and a temporal scale of 100 μs to seconds (e.g., rod phototransduction; chapter 8). Yet to serve behavior, computations must retain the same timescale but travel up to 1 millionfold farther. To achieve speed over distance requires recoding the chemical signals to electrical signals. Recoding begins with an allosteric trigger, such as ligand-binding or G protein activation, but allostery must eventually open an ion channel in the membrane to establish an electrical signal. This is one key task for a neuron: use allostery to send an electric signal somewhere fast.

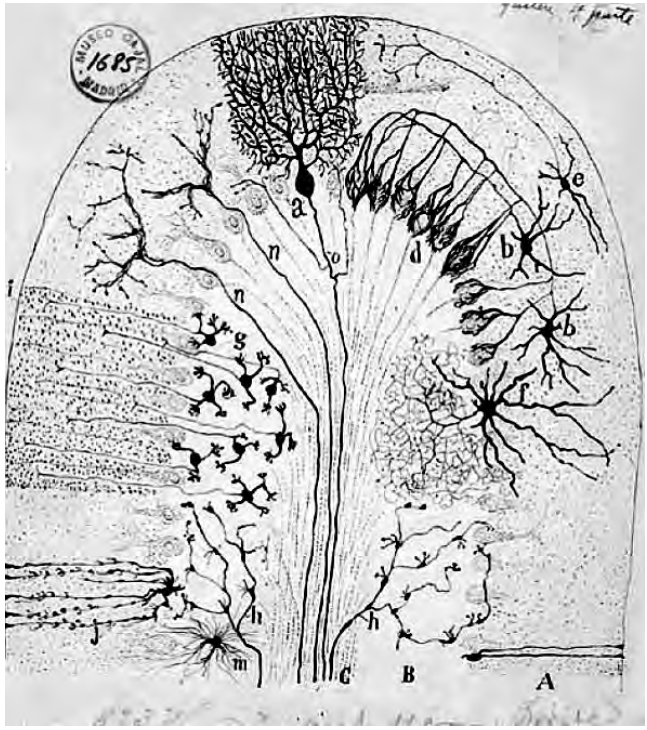


Figure 7.1
Neurons and glial cells of cerebellar cortex. Neuron types shown here (a, b, e, f, g) all express the standard polarized design: inputs to multiple dendrites converge to cell body and output to a single axon. Neuron types: a, Purkinje; b, basket; e, stellate; f, Golgi; g, granule. Input axons: h, mossy fiber; n, climbing fiber. Two types of glia (j, m) are shown at lower left. Each Bergman glial cell (j) wraps the dendritic arbor of a single Purkinje neuron (a). Drawing by S. Ramón y Cajal. Reprinted with permission from Sotelo (2003).

"Somewhere fast" has two parts. First, a chemical signal released by a *presynaptic* neuron targets a *postsynaptic* neuron on a short branch (*dendrite*). The chemical transmitter binding to a protein receptor allosterically opens its ion channel. This initiates an electrical signal that spreads passively along the dendrite toward a central locus (cell body or specialized cable segment) for integration with signals from other dendrites. Second, the integrated electrical signal recodes to an all-or-none pulse that spreads actively down a single cable (*axon*) toward presynaptic terminals that contact other neurons (figure 7.1).

Dendrites are 10–1,000 micrometers long, depending on neuron type, and over such distances "fast" means up to 50 micrometers per millisecond. Axons are 1–1,000 millimeters long, and over such distances, "fast" means at least 1 millimeter per millisecond. Thus, dendrites conduct passive electrical signals about 50-fold faster than chemical diffusion, and axons conduct active electrical signals at least 20-fold faster than dendrites.

A neuron steps up from the nanometer scale of protein circuits to the micrometer scale of a synapse (1,000-fold), then to the millimeter scale of a dendritic tree (1,000-fold), and then to the meter scale of the longest mammalian axons (1,000-fold), ultimately integrating processes that span a 10^9 range of spatial scale. This greatly increases the cost of space, materials, and energy. A protein molecule allosterically encoding 1 bit occupies about 50 nm^3; whereas the smallest neuron cell body encoding 1 bit occupies 10^9 greater volume; and the largest neuron cell body encoding 1 bit occupies 10^{12} greater volume and correspondingly more materials. The energy cost of encoding 1 bit rises from about 25 k_BT to about 10^9 k_BT.[2]

Such numbers explain why neurons need to be efficient. The microvessels that deliver oxygen and metabolic supplies distribute densely, forcing neurons to occupy their interstices (figure 7.2). Were neurons to be energetically less efficient, they would need more mitochondria to produce more ATP—and that would require a denser capillary network at the expense of efficient neuron layout (chapter 13). The same constraint applies to space and materials. For example, the diameter of a cerebellar Purkinje cell body is 10-fold greater than that of a cerebellar granule neuron, but its volume is greater by 1,000-fold (figure 7.1). Therefore, we must explain how the design of each neural component: synapse, dendrite, cell body, and axon match each other and conserve space and energy.

Synapse

Synapses enable neurons to process information in neural circuits by transferring and transforming signals at specific connections. The simplest synapses are electrical, made from proteins that form an array of channels that connect two neurons. Where it serves as a simple resistor, an electrical synapse is as inexpensive and noiseless as a connection can be.[3] This is why electrical synapses are widely used to weakly couple neurons, for example, to compute the mean signal over a patch of retina to reduce redundancy (chapter 11), to synchronize rhythmical activity among the cortical interneurons, and to synchronize motoneurons that drive the same muscle. But coupling with a resistor does not equip a circuit to compute much. More

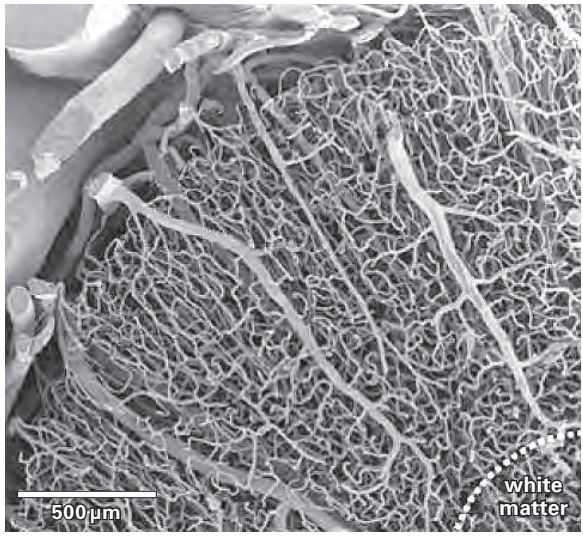


Figure 7.2
Blood vessels distribute densely in gray matter with even mesh. The 500-µm scale bar corresponds roughly to the dimension of largest local circuits. Therefore, neurons and glia must fit into the interstices of the capillary network. Were energy cost to rise, due either to lower neuronal efficiency or enhanced neuronal performance, vessel density would rise at the expense of efficient neuron layout. Cerebral cortex from superior temporal gyrus of monkey. Reprinted with permission from Weber et al. (2008).

transformations are required, and signals must be amplified to produce fast responses that are resistant to noise. These requirements are met by chemical synapses, so called because a presynaptic neuron transmits chemically by sending a pulse of neurotransmitter to receptors on a postsynaptic neuron.

Origin of graded chemical signal

A chemical pulse originates when a vesicle docked to a presynaptic *active zone* fuses with the plasma membrane and releases transmitter molecules through a *fusion pore* into the *synaptic cleft* (figure 7.3). The vesicle contains about 4,000 molecules of transmitter concentrated by a transporter protein in the vesicle membrane to roughly 100 mM. Discharge through the pore requires about 100 µs, during which molecules are diffusing away; yet their concentration at the postsynaptic receptor proteins clustered 20 nm across the cleft rises briefly to about 10 mM (figure 7.3). This suffices for

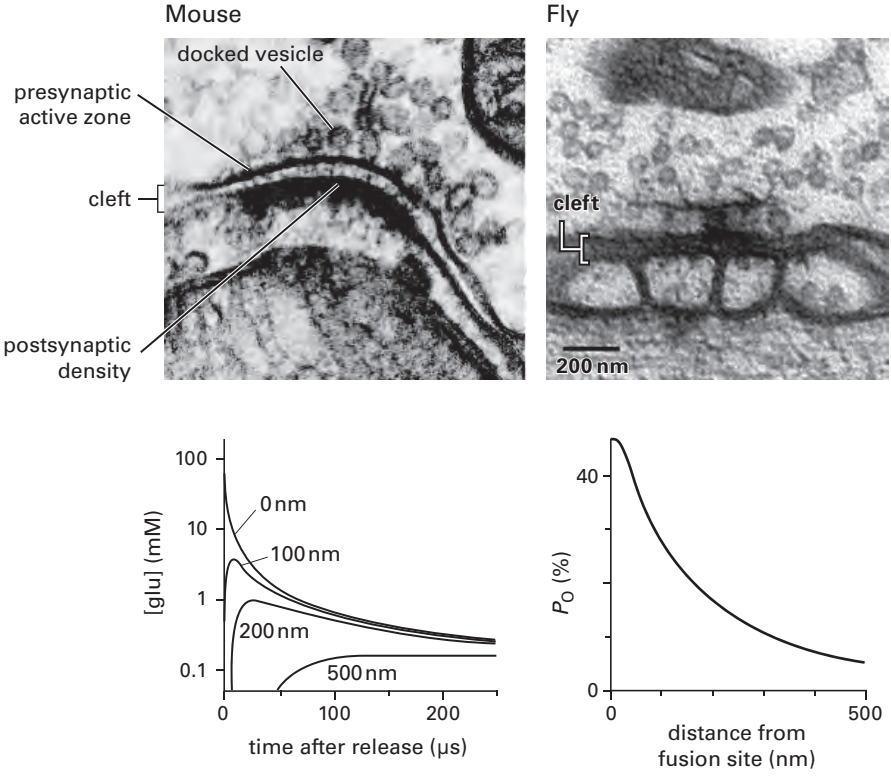


Figure 7.3

Fusion of one synaptic vesicle briefly raises the concentration of transmitter within the synaptic cleft to 10 mM. **Upper left**: Synaptic vesicles, about 40 nm, from mouse cerebellar cortex docked to presynaptic membrane across the synaptic cleft from the postsynaptic density. This density houses a complex of proteins that support, among other functions, long-term potentiation (see chapter 14). Courtesy of K. H. Harris. **Upper right**: Synaptic vesicles, about 30 nm, from *Drosophila* medulla. Note that cleft width and vesicle size are similar for mammal and fly. Fly vesicle contains similar number of transmitter molecules as mammal (Borycz et al., 2005). Courtesy of Zhiyuan Lu, Patricia Rivlin & Fly EM Team, Janelia, HHMI. **Lower left**: Concentration decays steeply in space and time. **Lower right**: Concentration at 20 nm from fusion site suffices to bind and open roughly half of the postsynaptic ion channels. Half-saturation by one vesicle allows multivesicular release to enhance the response. Decay to lower concentration with time and distance can be exploited by other types of receptor molecules with higher binding affinities (figure 11.10). P_o, open probability of postsynaptic ion channels. Graphs are modified and reprinted with permission from Xu-Friedman & Regehr (2004).

transmitter to bind cooperatively to low affinity receptors, thereby opening their channels and initiating the electrical signal, a miniature postsynaptic current (*MPSC*).

This design requires matching closely the number of molecules and their concentration within a vesicle to its emptying time, diffusion distance, and receptor binding constant. Were the vesicle to contain fewer molecules or a lower initial concentration, the final concentration at the postsynaptic receptor would be too low for its binding affinity. The same would occur if the vesicle emptied more slowly or if the diffusion distance across the cleft were greater. Any of these factors could be compensated for by a higher affinity at the receptor, but that would sacrifice bandwidth (chapter 6). These factors could also be compensated by a narrower cleft, but that would increase cleft electrical resistance and reduce postsynaptic current. Thus, cleft width appears to optimally balance transmitter concentration at the postsynaptic receptors and electrical resistance (Savtchenko and Rusakov, 2007; Graydon et al., 2014). So here is symmorphosis at the nanometer scale.

Molecular mechanism of vesicle fusion

For vesicle fusion (*exocytosis*) to work at all requires multiple allosteric processes. And for it to transfer the information encoded chemically to an electrical signal, while preserving temporal precision and S/N, these allosteric processes must couple efficiently as now explained.

To preserve temporal precision, vesicle fusion must occur promptly as a triggered event. This requires *docking* it in advance to a specialized *active* zone and then *priming* the vesicle with multiple *SNARE*s, each a complex of four protein molecules. A SNARE, upon binding the vesicle tightly to the presynaptic membrane, adopts a high free energy conformation that is metastable (figure 5.4). Consequently, a small signal can push a SNARE over the hump on its energy landscape and trigger fusion.

The trigger is a surge of calcium ions reaching the docked vesicle through voltage-gated channels clustered at the active zone. When channels open in response to a presynaptic depolarizing electrical signal, several hundred calcium ions enter to raise the local concentration by 50-fold in less than 500 μs. Several calcium ions are bound by the protein *synaptotagmin* attached to the vesicle, which then binds to the SNARE and pushes it over the energy hump (Südhof, 2013). As the SNARE plummets to a lower energy conformation, the freed energy causes violent tugs on the vesicle. The combined force from three SNAREs suffices to fuse the vesicle to the presynaptic membrane and wrench open a pore with consequence already noted. The

entire process, from presynaptic electrical signal to postsynaptic receptor activation, occurs fast enough to preserve temporal precision and to be completed within 600 μs.

Speed and temporal precision emerge from several design principles. The molecular components are irreducibly small and locate close together—within nanometers. This allows fast chemistry with irreducibly few molecules to achieve the high concentrations needed to transmit fast signals (high bandwidth). Chemistry achieves speed and gain by storing energy and then releasing it with concatenated switches: (1) voltage switch opens a calcium channel, releasing energy stored in calcium's electrochemical gradient (figure 7.4); (2) synaptotagmin binds calcium at low affinity, releasing energy stored within the SNAREs; (3) SNAREs fuse a vesicle, releasing energy stored by concentrating transmitter.

Timing is sharpened by cooperativity that steepens the response curves (figure 6.3): several voltage-gated calcium channels cooperate to establish a sufficient calcium concentration, several calcium ions cooperate to cause synaptotagmin to bind a SNARE, and several SNAREs cooperate to fuse a vesicle. Thus, via switches directing stored energy, the chemical signal recodes to electrical with good S/N and temporal precision.

In order to transmit high frequencies, a steeply rising chemical signal must also fall steeply. Thus, each stage must terminate quickly: (1) calcium channels close instantaneously as the membrane repolarizes; (2) calcium concentration collapses locally within tens of microseconds as calcium is bound rapidly by low-affinity buffering proteins; (3) synaptotagmin switches off sharply because of its steep dependence on calcium; (4) transmitter concentration decays within less than 1 ms by fast binding to transporter proteins on synaptic and glial membranes and by diffusing from the cleft.

In short, rapid release and rapid termination produce a chemical signal in the cleft that peaks within about 0.6 ms and lasts less than 1.5 ms, thereby transmitting information with irreducible delay and a bandwidth of about 1 kHz. This suffices to transmit most frequencies coded by the neuron's electrical signals because the membrane time constant is constrained by energy cost (chapter 6; Attwell & Gibb, 2005). Thus, the mechanism of chemical signaling at the synapse matches the bandwidth of the presynaptic neuron.

Vesicle release is stochastic

An action potential reaching a presynaptic terminal may cause a single vesicle to be released, or it may fail. Release is stochastic with a probability that

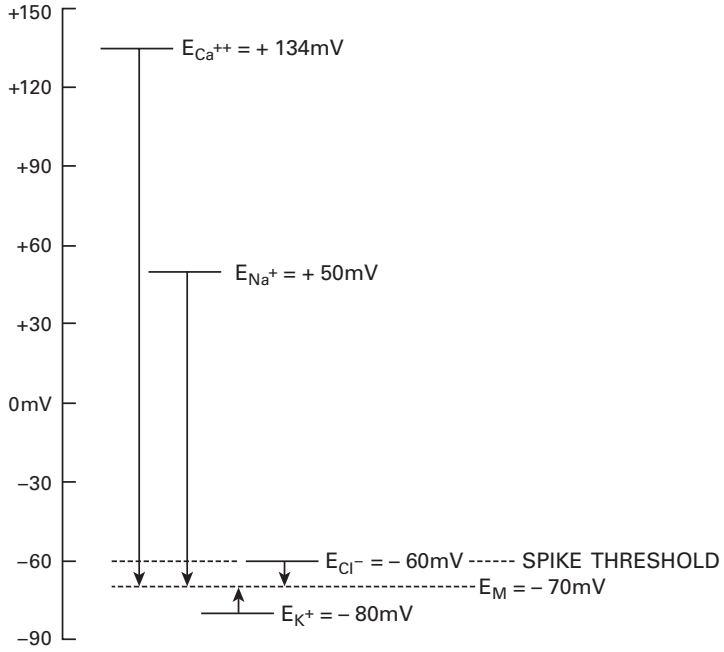


Figure 7.4

Driving forces on key ions at the neuron's resting potential. Calcium and sodium ions are driven strongly inward whereas the forces on potassium and chloride ions are weak until the membrane is depolarized. Then their driving forces increase with the degree of depolarization and tend to oppose it. Since a neuron must depolarize to release a graded chemical signal or to fire an all-or-none spike, inputs whose allosteric effects open calcium or sodium channels are excitatory, and inputs whose allosteric effects open chloride or potassium channels are inhibitory.

can vary between 0.1 and 0.9. The uncertainty of release introduces noise, thereby reducing information, but it can have advantages (Harris et al., 2012). For example, it reduces the likelihood that two vesicles will redundantly carry the same signal. It also offers a mechanism to adjust the effectiveness of a synapse—its *weight*—by tuning release probability. This provides mechanisms for homeostasis and plasticity (chapter 14). The costs and benefits of stochastic release are discussed further in later chapters.

Recovery and cost of presynaptic chemical signal

Vesicles fusing by exocytosis expand the presynaptic membrane, thus increasing its capacitance and time constant. To prevent this, the terminal retrieves each fused vesicle by *endocytosis*, folding inward the added patch

of membrane and pinching it off. This sounds simple, but, of course, it requires allosteric action by several types of proteins—to reverse what was accomplished by the SNAREs. Obviously, a synaptic terminal must maintain a strict balance between exo- and endocytosis.

Although a synaptic vesicle appears morphologically simple, it is really a complex molecular machine comprising about 800 protein molecules of about 40 protein species (Fernández-Chacón & Südhof, 1999). Beyond the structural membrane proteins and transporter proteins to fill it with transmitter, there are specific proteins for other tasks: to link the vesicle to neighbors near the presynaptic membrane for rapid recruitment to docking sites (Hallermann & Silver, 2013; Hallermann et al., 2010), to provide two of the proteins in each SNARE plus the synaptotagmins to trigger them, and to mark the vesicle for endocytotic retrieval. By specifically retrieving vesicle membrane, rather than nonvesicle membrane, the specific vesicle proteins are retrieved together, allowing the vesicle to be refilled and readied for rerelease within a minute.

Using allostery of chemically coupled proteins, the vesicle release mechanism is efficient in space, materials, and energy: to extrude the modest numbers of calcium ions, ~12,000 ATP; to energize the SNAREs, <100 ATP; to retrieve the vesicle, <500 ATP; and to fill the vesicle, ~11,000 ATP (Attwell & Laughlin, 2001).[4] Therefore, the total cost of a presynaptic chemical quantum is ~23,000 ATP. The postsynaptic electrical response to this signal costs roughly 10-fold more, as we now explain.

Postsynaptic electrical response

The transmitter molecules reaching receptor proteins across the cleft bind stochastically, and when two or more molecules bind cooperatively to the same protein, it changes conformation to open a channel (figures 6.6 and 6.7). The contents of one vesicle, a *quantum*, open about half of the available channels (figure 7.3). Therefore, enlarging the quantum by increasing transmitter concentration within a vesicle, or releasing several quanta (*multivesicular release*) can produce a graded increase in the fraction of open channels (see below, figure 7.17). Thus, the information packet presented to a postsynaptic receptor cluster is graded, as is the opening of channels that capture the amplitude and timing of the upstream event. Still in chemical mode, it is cheap.[5]

But now ions flow through the open channel with a direction and amplitude that depend on their electrochemical driving force (figure 7.4). When the transmitter is glutamate and the postsynaptic receptor is a ligand-gated cation channel (chapter 6), sodium ions, and in certain cases calcium ions,[6]

are strongly driven inward to depolarize the membrane. The channel is also permeable to potassium, but its weaker driving force moves fewer ions outward. The net ionic current converts the graded chemical signal to a graded electrical signal. The cost is 10-fold more energy, but this is essential to send over distance at acceptable time delays. Recall that, whereas a graded chemical signal diffuses cheaply, about 1 μm in a millisecond, a graded electrical signal needs batteries but travels 50-fold farther in the same time.

The greater energy cost demands measures to reduce noise. This follows the principle *send only what is needed*, and that means sending the least possible noise. The noise sources include: stochastic release of vesicles, timing of vesicle fusion, size of a transmitter quantum,[7] times of receptor binding/unbinding, and channels opening/closing (Ribrault et al., 2011). So the neuron takes various measures to mitigate them.

The number of molecules released in the quantal pulse from a vesicle varies across neuron types. The number depends strongly on vesicle diameter, d, since volume goes as d^3, and on the final intravesicle concentration, roughly 100 mM. Each neuron type selects a vesicle between 30 and 50 nm in diameter (figure 7.3). This allows a roughly fivefold range in number of molecules. Functional vesicles as small as 20 nm in diameter have been produced experimentally by genetic manipulation, but they contain fewer transmitter molecules than a 30-nm vesicle—insufficient to establish an effective postsynaptic concentration. Small vesicles with more transporter molecules might conceivably establish higher internal concentrations, but actually, overexpression of transporter proteins gives larger vesicles with similar mean concentration. Thus, the 30-nm vesicle seems to be a lower bound from fly to mammal (figure 7.3).

Increasing the number, M, of postsynaptic receptors improves postsynaptic S/N by $\sqrt{M}$ and reduces the time constant by charging the capacitance more briskly (chapter 6). A small synapse clusters about 20 receptors, improving S/N compared to one receptor by about 4.5-fold. A large synapse may expand the receptor cluster up to about 10-fold and thereby improve S/N by about 14-fold (figure 7.5).[8] But this threefold benefit comes with a 10-fold greater cost, so it is reserved for special purposes, such as auditory synapses that transmit with high temporal precision (chapter 10). In any case, the graded electrical signal now serves the neuron's next task: integrate and send an output.

Different protein receptors process on different timescales

A neuron must register events on different timescales and does so using postsynaptic receptors with different kinetics. Consider as an example one

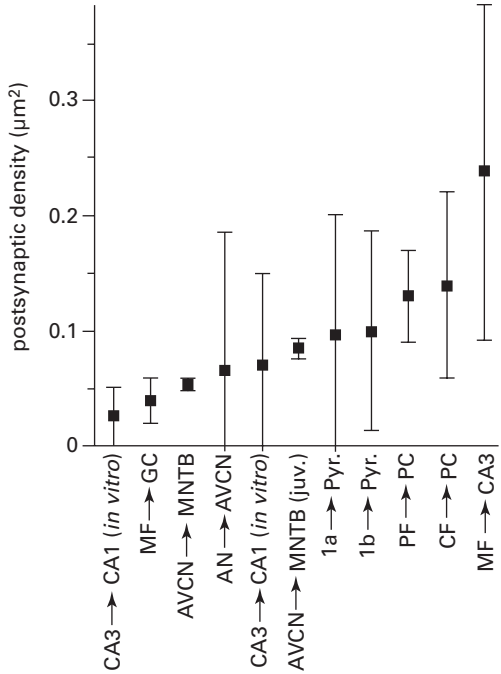


Figure 7.5

Postsynaptic receptor clusters from different synapse types span a 10-fold range of size. This reflects customized S/N for each type, according to needed benefit and cost. Area of postsynaptic density corresponds to size of the receptor cluster. CA3 → CA1= hippocampus; MF → GC = cerebellar mossy fiber to granule cell; AVCN → MNTB = anteroventral cochlear nucleus to medial nucleus of trapezoid body; AN → AVCN = auditory nerve to anteroventral cochlear nucleus; 1a, 1b → Pyr = pyriform cortex; CF → PC = climbing fiber to Purkinje cell; PF → PC = parallel fiber to Purkinje cell; MF → CA3 = hippocampal mossy fiber to hippocampal pyramidal cell. Reprinted with modification and permission from Xu-Friedman and Regehr (2004).

broad family of ligand-gated cation channels, the glutamate receptors (Attwell & Gibb, 2005). Two types, the NMDA receptor and the AMPA receptor,[9] bind glutamate with similar ON rates, but the NMDA receptor's OFF rate is 400 times slower. This slower OFF gives the NMDA receptor a more sustained response that covers a longer timescale (figure 7.6); it also causes a 400-fold higher sensitivity to glutamate: whereas an AMPA receptor requires glutamate concentrations of nearly 1 mM to cooperatively bind two glutamate molecules, an NMDA receptor is doubly bound at about 1 μM.

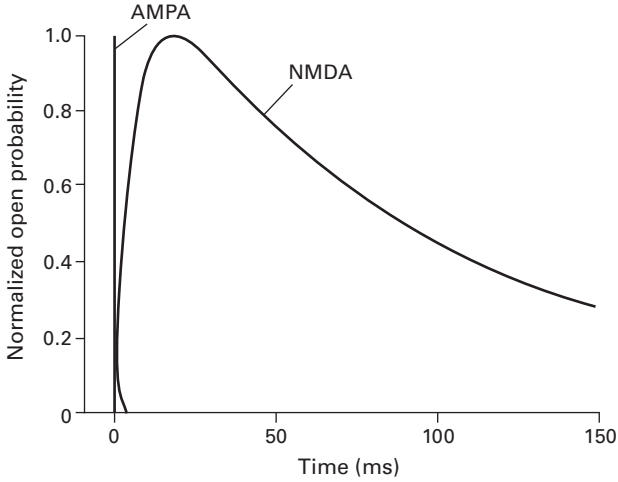


Figure 7.6

Neurons use different types of receptor to encode a range of temporal frequencies. Graph shows time course of activation for AMPA and NMDA receptors by a 0.3-ms pulse of 1 mM glutamate. To more finely optimize binding for a range of high frequencies, different AMPA types are employed along with different regulatory molecules. Reprinted with permission from Attwell & Gibb (2005).

The AMPA receptor covers a neuron's shortest timescale (figure 7.6). Its high rates for glutamate ON and OFF match fast rate constants for channel opening and closing, so an AMPA receptor completes its electrical response with a time constant of about 0.8 ms. This speed preserves the temporal information transmitted by fast vesicle release. It also matches the lower limit to a neuron's membrane time constant imposed by the energy cost of a low membrane resistance (time constant = capacitance × resistance). In other words, an AMPA receptor equips an excitatory synapse to transmit the full bandwidth of neuronal response.

For AMPA receptors to cover this fast timescale and wide bandwidth, glutamate at the postsynaptic receptor site must decline promptly. Glutamate diffuses about 1 μm per millisecond, so by restricting AMPA receptors to a small postsynaptic patch, less than 1 μm in diameter, its concentration falls by e-fold within a millisecond (figure 7.3). Glutamate accumulation during successive releases is prevented by active uptake. Transporter proteins on the surrounding neuron and glial membranes bind glutamate rapidly and move it into the cell, powered by the influx of three sodium ions. The transporter works slowly, requiring a full minute to retrieve the 4,000

glutamate molecules released by one vesicle. Therefore, to mop them up on the required millisecond timescale requires transporters to be concentrated around the synapse: to bind 4,000 glutamate molecules within a millisecond requires 10,000 transporters that bind with high affinity (Attwell & Gibb, 2005).

To efficiently match components, AMPA receptors locate near the vesicle release site, within about 20 nm, where they see a fast, high peak and steep decay of glutamate concentration (figure 7.3). Thus, the receptor's molecular structure is matched by vesicle content and distance from the release site (figure 7.3). NMDA receptors in certain cases locate farther from the release site where they see slower rise and fall of glutamate concentration, allowing them to integrate contributions from successive release events. Their sustained responses can also be integrated with fast AMPA receptors (Clark & Cull-Candy, 2002).

The NMDA receptor's sustained response allows it a role in detecting *coincidence detection*. This exploits the NMDA receptor's voltage sensitivity. A receptor, binding glutamate when the membrane is at resting potential (–65 mV), reacts weakly because its channel is partly blocked by a positively charged magnesium ion. When the membrane depolarizes—for example, due to AMPA-mediated currents from synapses on the same dendrite—magnesium is forced out, allowing sodium and calcium to enter. Thus, an NMDA receptor detects the coincidence of a presynaptic input (glutamate) AND a postsynaptic response (depolarization) within a time window of about 100 ms set by the slow unbinding of glutamate.

The NMDA receptor's ability to detect and signal coincidence equips a neuron for pattern recognition and learning (chapter 14). An active receptor emphasizes the coincidence by amplifying and extending a synapse's excitatory input; moreover, it marks the synapses whose signals coincide. Only synapses that recently delivered glutamate have NMDA receptors primed for action. When these receptors are unblocked by depolarization, they admit chemical messengers (calcium ions) that initiate structural change. Because an action potential also depolarizes synapses, the NMDA receptor enables a neuron to take a first step in learning; it can identify and modify those synapses whose inputs coincide with a definitive output.

The duration of an NMDA receptor's time window is critical for learning. Shorter would increase false negatives—the receptor would miss correlations between events that take longer to unfold. Longer would increase false positives—more unrelated events would occur in the same time window. The NMDA receptor's OFF rate creates the 100-ms time window that seems about right for many of life's more immediate events.

However, this mechanism creates a problem (Attwell & Gibb, 2005). The slow OFF rate retains some glutamate bound until nearly every last molecule has been removed from the synaptic cleft. This requires a transporter to harness the energy of *three* sodium ions: two sodiums can pull extracellular glutamate down to 180 nM, but this leaves 13% of NMDA receptors still bound. Thus, for NMDA's time window to close within 100 ms requires a transporter with appropriate stoichiometry—at a cost of 50% more energy.

Intervals still longer than the NMDA receptor's time window are covered by a *metabotropic* glutamate receptor, *mGluR*. This receptor belongs to the same molecular family as the β-adrenergic receptor (chapter 5), and like that receptor, it activates a G protein to deliver an amplified signal. Responses driven by mGluR can be tuned to cover a range of time intervals, from about 0.1 s to tens of seconds. Moreover, mGluR's second messengers can, like calcium from the NMDA receptor, institute longer lasting structural changes.

Processing multivesicular release

We have explained how receptors process on different timescales by varying their output to a singular input, a puff of glutamate. Receptors can also detect variations in input timescale, produced by different temporal patterns of vesicle release, by varying the kinetics of receptor activation, deactivation, and desensitization. For example an AMPA receptor desensitizes in response to prolonged glutamate, causing the response to a sustained burst of action potentials to decline over time. This fast desensitization favors signals that change on a short timescale, thereby tuning AMPA's bandwidth to higher frequencies and eliminating redundancy. Conversely, an mGluR activates slowly and does not desensitize, thereby favoring inputs that change on longer timescales.

In summary, the glutamate receptor families enable a neuron to process on different timescales by producing synaptic responses of different durations. Receptors are constructed from different parts, that is, from different combinations of a receptor's protein subunits, to provide the key differences in kinetics, sensitivity, and output: an AMPA receptor that unbinds glutamate at a high rate to create a narrow window and desensitizes to favor high frequencies; an NMDA receptor that is voltage sensitive and delivers calcium ions; and an mGluR that acts more slowly via second messengers. But simply engineering receptors is insufficient. A receptor's kinetics and sensitivity must be matched by the stoichiometry and affinity of transporter molecules, by their density around a synapse, and by the dimensions of the synapse itself (Attwell & Gibb, 2005). This conclusion, that to be

effective requires design in depth, is more amply illustrated by photoreceptors (chapter 8).

Efficiency from synaptic inhibition

Neurons employ various forms of synaptic inhibition. All increase efficiency by improving timing precision (Sengupta et al., 2013) and reducing redundancy. These effects all help concentrate information, so that the space and energy used to send expensive electrical signals are well spent. Inhibition also serves to delete information unneeded by a downstream user. These effects are discussed with specific examples in chapters 9 through 12. Here we explain the mechanisms and why they are cheap.

One type of synaptic inhibition is achieved when transmitter binding opens a membrane channel for chloride ions. The open chloride channel reduces the depolarization produced by an inward cationic current and hence reduces the probability of triggering an output—a vesicle release or spike. The inhibition is achieved in two ways. First, when E_{Cl} is negative to the membrane potential (figure 7.4), chloride enters and neutralizes the charge carried by entering cations. Second, irrespective of whether chloride flows in or out, the open chloride channel lowers membrane resistance, thus shunting the depolarization. This effect dominates when chloride currents are small, which is generally so because E_{Cl} is generally near E_M (figure 7.4). Small currents require less restorative ion pumping, which makes chloride's *shunting inhibition* energy efficient.

Shunting inhibition is also achieved by opening a potassium channel. As for chloride, the potassium current is small because E_K is near E_M (figure 7.4). The inhibitory potassium channels are not gated by chemical transmitter but rather by G proteins, membrane voltage, or calcium. Thus, they add substantially to the parts list for energy efficient inhibition.

The transmitters for ligand-gated chloride channels are GABA, glycine, and histamine (figure 6.7). The GABA receptors ($GABA_A$) comprise a diverse family of molecules. The receptor assembles as a pentamer from several classes of subunit (alpha, beta, gamma, etc.), each of which has subtypes. This permits customized properties, such as different binding constants, different speeds of opening, and different rates of desensitization. The $GABA_A$ receptor's ligand binding is modulated allosterically at several sites on the molecule by brain chemicals, such as steroids and by various exogenous chemicals, such as alcohol, barbiturates, and benzodiazepine "tranquilizers." This suggests, by analogy with endogenous modulators of other types of receptor (endogenous opiates and endocannabinoids), that there should be endobenzodiazepines. One such molecule has been reported, a

secreted protein, *diazepam binding inhibitor*, that potentiates the $GABA_A$ receptor (Christian et al., 2013).

Dendrites expand a neuron's information capacity

To gather more information, a neuron must supply more membrane for synaptic contacts while minimizing the length of wire devoted to connecting (chapter 13). The design solution is to grow dendrites, which also offer compact compartments for electrical and chemical computing and for integrating signals. A dendrite is structured like an electrical cable—with a conducting core (cytoplasm) and outer insulation (the membrane's lipid bilayer). Voltages decay exponentially on a cable (figure 7.7) with a length constant[10] that depends upon resistance per unit length of insulating membrane, r_M, and conducting core, r_{cyt}:

$$\textit{length constant} = \sqrt{(r_M/r_{cyt})}. \tag{7.1}$$

A dendrite does not transmit far: r_M is too low because potassium channels stay open to maintain resting potential, and r_{cyt} is too high because hydrated ions in cytoplasm conduct poorly. Therefore, a dendrite's length constant is generally less than 1 mm, and dendrites preserve signal amplitude by staying shorter than their length constant.

A dendrite may increase its length constant by growing thicker, thus reducing r_{cyt}, but the improvement goes only as the square root of diameter (Koch, 1999). With length constant increasing as $\sqrt{d}$ and volume increasing as $d^2 \times \textit{length}$, the total cost of space and materials increases as $(\textit{length})^5$. Such a steeply diminishing return requires designs that keep dendrites short (chapter 13). Temporal resolution (bandwidth) also requires short dendrites because longer ones increase capacitance. This delays signals and spreads them out, thus attenuating high frequencies (figure 9.9). In short, signals traveling passively on a dendrite longer than 1 mm would be too weak and too slow to carry much information. Yet these cable properties that limit dendritic length can be exploited to process information as it is gathered.

Dendrites process directly

Dendritic biophysics provides cheap and robust analogue processing (Koch, 1999). For example, by placing input synapses that carry less information distally on the dendrite, they can be given less weight in the final output, whereas signals that carry more information can be given more weight by placing them nearer the cell body (figures 11.15 and 11.16). More generally, inward currents from excitatory synapses can be combined with outward

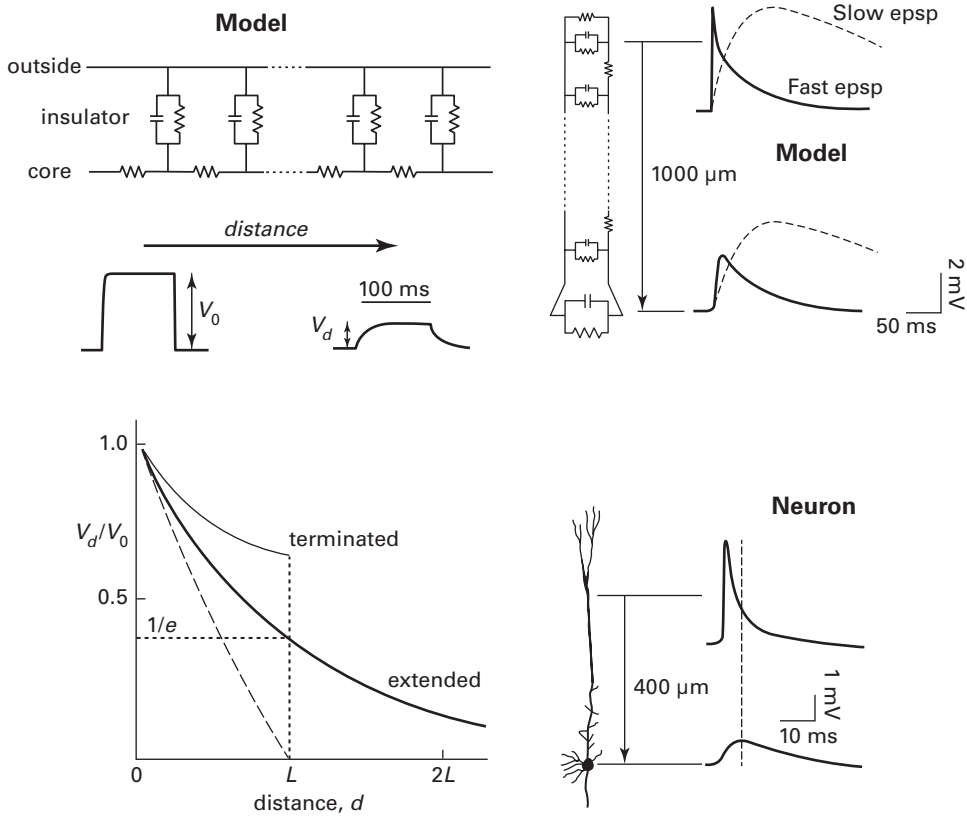


Figure 7.7

Passive transmission by dendrite. Upper left: Passive cable modeled as series of thin segments, each contributing to resistance and capacitance of insulator (dendrite's membrane), and to resistance of conducting core (dendrite's cytoplasm). **Middle left:** Cable smooths square pulse as it transmits over distance, *d*, and attenuates steady-state voltage amplitude from initial value, V_0, to V_d. **Lower left:** Amplitude transmitted, V_d/V_0, decreases exponentially when cable is extended many length constants, *L* (middle curve, steady-state response of *semi-infinite cable*). Transmission improves when cable is terminated and end is sealed (top curve). Transmission worsens when end is open and short circuits core's signal (bottom curve). **Upper right:** Model of passive dendritic cable transmitting to cell body (larger resistor and capacitor at base of cable). Fast EPSP severely attenuated, slow EPSP less so. **Lower right:** Neuron transmitting fast EPSP to cell body via dendrite. Fast EPSP is smoothed, thereby delaying time to peak (dashed vertical line), and attenuated. Cable transmission curves after Rall (1959). Modeled transmission of fast EPSP and slow EPSP replotted from Spruston et al. (1998). Cortical pyramidal cell morphology and recordings of EPSP adapted from Stuart & Spruston (1998) with permission.

currents from inhibitory synapses and potassium channels. Thus, a dendrite serves as an analogue electrical circuit that adds, subtracts, divides, multiplies, and takes logarithms (figure 6.2).

Such operations implemented directly by an RC circuit combine many synaptic inputs to produce an output within milliseconds. This rapid many-to-one integration, which serves behavioral requirements for prompt decision, would be difficult to implement with a chemical circuit. The many-to-one ability is further exploited by joining several dendrites to the cell body or an integrating segment to form a final common output (figure 7.1).

To increase their processing abilities, dendrites *complicate their design* (Branco & Häusser, 2010). For one example, a dendritic twig at the distal tip of an elaborate dendritic tree makes a coincidence detector by strategically expressing voltage-sensitive sodium channels (Harnett et al., 2013). Glutamate released at excitatory synapses binds AMPA and NMDA receptors, but only when AMPA currents from several synapses coincide does the dendritic twig depolarize sufficiently to unblock the NMDA receptors. Through them, calcium and sodium ions enter to amplify the coincidence, and voltage-gated sodium channels register it by producing a robust pulse, an action potential.

This pulse does not propagate far because the sodium channels are confined to the twig. However, it generates sufficient current to drive a detectable signal into the larger dendritic tree. Here, strategically positioned potassium channels shunt responses from particular twigs and branches, thereby selectively blocking some inputs or controlling their gain. Thus, more complicated dendrites provide two layers of processing, local within a single dendrite and global within the larger dendritic tree.

A dendrite can add another layer of processing by receiving a synapse on the globular head of a dendritic spine, 0.5–1.5 μm in diameter (figure 7.8; Yuste, 2013; Sala & Segal, 2014). Evoked current and chemical messengers pass from spine head to dendrite through a thin neck, 0.05–0.25 μm in diameter and 0.5–2 μm long. The neck resists the flow of intracellular current and chemicals, thereby creating a computing compartment in the head that can be regulated by varying neck diameter and length. This design neatly resolves two conflicting demands.

A synapse must merge its current with many other currents in an extended RC network (dendrite and tree). In doing so, it loses individuality because the amplitude of its *EPSP (excitatory postsynaptic potential)* depends largely on the state of the network, as determined by the multitude of synaptic and voltage-gated conductances (Yuste, 2013). Yet, a synapse must

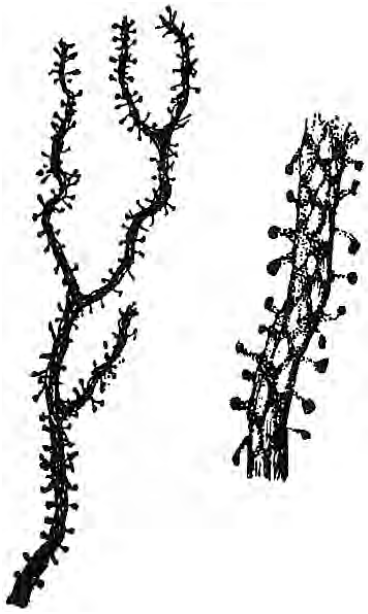


Figure 7.8
Purkinje cell dendrites studded with spines of varied morphology. Reprinted from Ramón y Cajal (1909).

also *adapt, match, learn, and forget*. For this, it must retain individuality to monitor and adjust its own EPSP (chapter 14). A specific example is explained for the retina's horizontal cell (chapter 11, figure 11.5).

The spine neck's variable resistance allows it to adjust the amplitude of the EPSP in the spine head.[11] This robust private measure of output can then drive electrical and chemical circuits in the head that modify the output to adapt, match, learn and forget (chapter 14). The head's small volume allows chemical processing to be fast, reliable and efficient. A spine also increases wiring efficiency by extending a dendrite's reach at minimum cost in space and materials (chapter 13, figure 13.3).

In short, dendrites collect and process analogue signals before final recoding to faster, regenerative pulses for transmission down the axon. But dendrites can also process in reverse. A dendritic tree may express voltage-gated sodium and calcium channels that allow spikes triggered in the axon to propagate back into the tree. This sends information about the neuron's output back to the dendrites where it serves various purposes, for example, to strengthen the synapses that generated the neuron's

output—for learning (chapter 14). Now we consider the design to minimize information loss in converting dendritic analogue to a pulse code at the axon, for faster transmission over distance.

The axon converts to a pulse code

Having gathered and processed analogue signals from multiple dendrites, the neuron needs to send the information over long distance to its synaptic terminals via a single specialized cable (figure 7.1). Passive conduction of a graded signal is too slow, and its exponential decay would lose information to noise (figure 7.7). Consequently, for distances greater than one length constant (~1 mm), an axon recodes to all-or-none pulses—action potentials (figure 6.10). These travel faster and avoid decay by regenerating as they go. Information encoded as spikes travels far with little loss.

However, the step of recoding from analogue to pulses loses a lot of information—as much as 90% (Sengupta et al., 2014). Whereas an analogue signal tracks changes continuously by allowing several response levels (figure 5.3), pulses allow the membrane potential to change intermittently between just two levels (figure 3.5). Thus, a neuron using analogue can transmit more than 2,000 bits s^{-1}, whereas using spikes, it can manage fewer than 500 bits s^{-1}. Furthermore, a 100-mV spike costs far more energy (chapter 6), so recoding to pulses massively reduces energy efficiency (bits per ATP).

A neuron tries to minimize these losses by converting to spikes at a specialized initiation site.[12] This *initial segment* locates at some distance from the cell body's large capacitance to shorten the membrane time constant (see below, figure 7.11). The initial segment also packs membrane channels extra densely to further reduce the time constant and increase S/N. With higher S/N and shorter time constant, the site triggers spikes faster and more reliably, both of which increase bits per spike, thereby reducing loss of information and improving efficiency.

To increase bits per spike right at the initial segment is critical because transmission down the axon is expensive. A transmitted spike charges the entire axon's membrane capacitance by about 100 mV by admitting sodium ions, and to pump them out costs 6.3×10^3 ATP per square micrometer of membrane (chapter 6). A pyramidal neuron in cerebral cortex uses an irreducibly thin axon (~0.3 μm in diameter) to send spikes through its intracortical circuit. Even so, the cost per spike is 6×10^6 ATP per millimeter, and over its length of 4 cm, the cost is 2.4×10^8 ATP (Attwell & Laughlin, 2001).

Considering the full population of pyramidal neurons, their spikes account for more than 20% of the energy used to process information in cortical circuits (Sengupta et al., 2010; Howarth et al., 2012).

In short, the conversion of synaptic and dendritic analogue signals to axonal spikes creates an expensive bottleneck. Bits are limited and costs increased, which makes it imperative to use each spike efficiently according to principles of neural design. In particular, an axon should *send only what is needed* by reducing redundancy and noise and by limiting transmission to what downstream neurons need to know (chapter 11). Thus, the carving out of essential features for transmission is a critical process in which the final step, the conversion of analogue to spikes by voltage-gated channels, plays a decisive role (Aguera y Arcas et al., 2003).

Supply and recycling

The neuron cell body must continually deliver fresh organelles, such as mitochondria, vesicles, and rafts of receptor proteins, outward to distant dendritic and axonal arbors. These nether regions must return recyclable materials and inform the cell nucleus of their needs.[13] Traffic in both directions needs to be faster than diffusion—at rates up to about 10 mm per hour. For efficiency's sake, bidirectional traffic goes along the same track, a protein monorail.

The *microtubule*, constructed from subunits as a cylindrical polymer, is irreducibly fine, about 30 nm in cross section. Two molecular motors, *kinesin* and *dynein*, step along the tubule, ferrying cargo, in opposite directions. The motors operate with lever arms that, jutting orthogonally from the tubule, require clearance; therefore, microtubule spacing can be no closer than about 50 nm. Attach some cargo, and the circumferential zone about the tubule needs to be about 30 nm. These molecular structures set a lower bound to the caliber of a neuronal process. The finest axons at about 100 nm are just thick enough to accommodate a single microtubule plus its molecular motors and cargo. Axons are encouraged to shrink because space and energy costs rise as d^2, but eventually they hit this lower bound because of their irreducible need for fast transport.

Transport along the microtubule monorail is relatively cheap. Each 8-nm step of kinesin down the tubule costs 1 ATP, so to move a cargo 1 mm costs about 10^5 ATP. Although this might seem expensive, the molecular stepping motor is slow, so the cost per second is only about 100 ATP, far lower than the cost of electrical signaling by an ion channel (chapter 6).

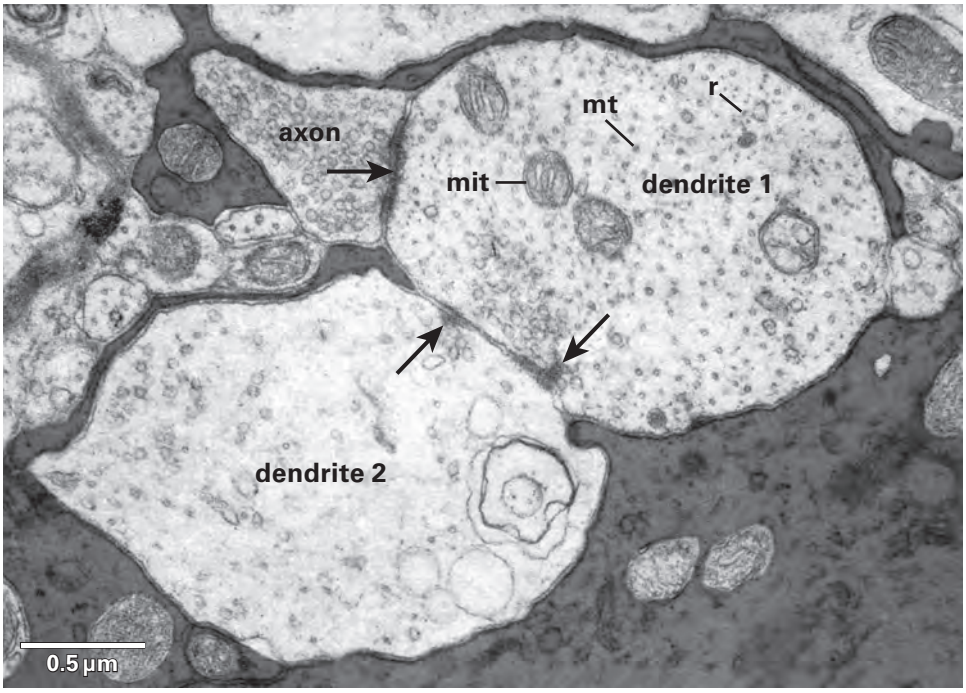


Figure 7.9
Local computing by chemical axodendrodendritic synapses. Axon terminal contacts dendrite 1 that contacts adjacent dendrite 2 that feeds back to dendrite 1. Note the array of evenly spaced microtubules (mt) in cross section and glial wrappings (shaded). Tiny dots are ribosomes (r) that support dendritic protein synthesis. Reprinted with permission from Famiglietti (1970).

Variations on the standard design

The standard polarized design requires every synaptic input voltage to travel a fair distance along a dendrite (up to a millimeter) over a significant time (2–20 ms) and then still longer distances along an axon, costing additional time and energy. Computations that avoid this long, slow loop would save considerable resources. Thus, neural designs include various features for computing locally, both at the synaptic and neuron levels.

Synapses designed for local computing

Certain designs allow direct computing between dendrites. A dendrite may form a chemical synapse onto a neighboring dendrite. The presynaptic

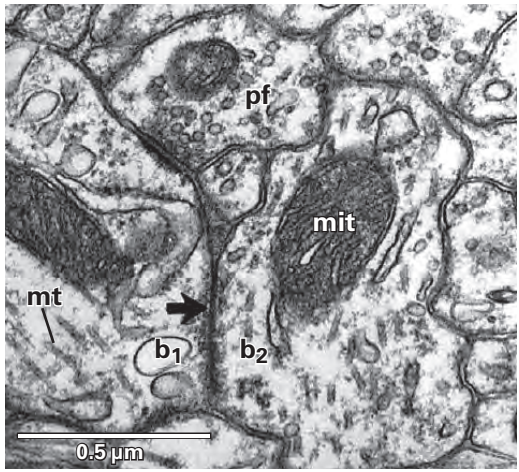


Figure 7.10

Local computing via electrical synapse between two basket cell dendrites. When chemical synapse from parallel fiber (pf) depolarizes b_1, current flows across the electrical synapse (arrow) to instantaneously and inexpensively depolarize b_2. Rat. Reprinted with permission from Sotelo & Llinás (1972).© 1972 Rockefeller University Press.

dendrite releases transmitter onto the postsynaptic dendrite, which may reciprocate with a return chemical synapse to its neighbor (figure 7.9). Such *dendrodendritic* chemical synapses are employed by circuits in many brain regions, including spinal cord, thalamus, olfactory bulb, superior colliculus, retina, and cerebellar cortex (Shepherd, 2004). They are efficient because (1) computations remain in analogue mode, which is direct and cheap; (2) expensive long-distance electrical signaling is avoided; and (3) different branches of the same neuron can compute independently, thereby expanding by up to 100-fold the computational possibilities of a single neuron (Grimes et al., 2010)

Dendrodendritic synapses can also be electrical, via a *gap junction*. This connection passes current directly between dendrites through a transcellular channel (*connexon*). It is fast—essentially instantaneous—because it dispenses with all the steps needed by a chemical synapse that take a millisecond or longer. Moreover, it does not amplify, so it is energetically cheap, and, as evident in figure 7.10, it requires no space at all. This type of synapse, illustrated here for cerebellar cortex, is ubiquitous. Specific computational functions will be treated in chapter 11.

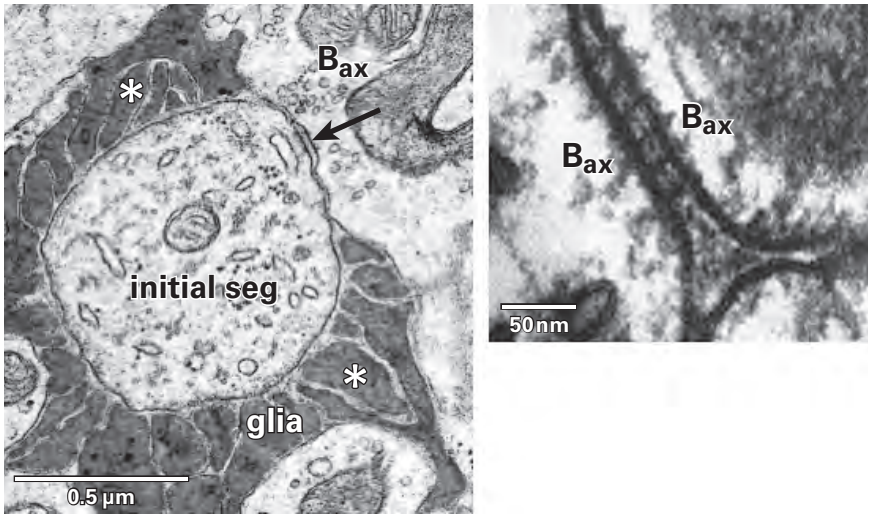


Figure 7.11
Two types of inhibitory synapse at the Purkinje cell axon initial segment. Left: Basket cell axon terminal (Bax) forms a chemical synapse (GABA) onto Purkinje cell initial segment. Glial fingers form a sheath around the axon. These fingers contain dense arrays of potassium transporters that rapidly bind potassium released during the action potential. Rat. Reprinted with permission from Palay and Chan-Palay (1974). **Right**: Basket cell axon tendrils form *septate-like junctions* with high resistance cross-bridges. Their capsule surrounds the glial fingers and apparently elevates the extracellular resistance. Basket cell spikes that synchronously invade the tendrils depolarize the extracellular space, thus reducing the intracellular depolarization to modulate spike timing. Reprinted with permission from Sotelo & Llinás (1972).

Local computing is also accomplished by axoaxonic synapses. One design delivers a chemical synapse to an axon's initial segment (figure 7.11). This site, as noted, critically regulates spike frequency and timing, so a chemical synapse can act powerfully without consulting the neuron's slower and costlier integrative apparatus. Another design uses a special form of electrical inhibition—fast and cheap. The cerebellar basket cell axon uses both designs, as now explained.

The basket cell axon, having richly enveloped the Purkinje cell body (figure 7.1), sends fine tendrils to surround the initial segment (see below, figure 7.15). The tendrils penetrate the glial wrapping and deliver a chemical synaptic contact that releases GABA (figure 7.11). Additionally, the tendrils join together using high resistance cross-bridges to form a capsule surrounding the glial fingers (figure 7.11). The capsule apparently elevates

extracellular resistance so that basket cell spikes, locally synchronized by the dendrodendritic electrical synapses (figure 7.10), invade the fine tendrils and depolarize the extracellular space. This instantaneously reduces the initial segment's intracellular depolarization and modulates spike timing (Korn & Axelrad, 1980). Polarizing the extracellular space to control a neuron's output is direct, economical, and relatively noise free—so it is used elsewhere (chapters 9 and 11).

Other types of local computing use an axoaxonic synapse, directed not to the initial segment, but rather to another synaptic terminal. Matched to function, these can be either electrical or chemical, and sometimes both are combined at the same junction. Electrical junctions between synaptic terminals of the same type improve S/N (chapter 11, figure 11.4). They also increase temporal precision because, as one synapse depolarizes, its coupled neighbor draws off some current, advancing its own depolarization and retarding the first (Pereda, 2014). This also increases synchrony between coupled terminals, a valuable property in many circuits achieved directly at negligible cost in space and energy.

A chemical axoaxonic synapse, depending on transmitter and receptor type, can be excitatory or inhibitory. For example, a retinal axon terminal releases glutamate onto AMPA receptors at an axon terminal of a thalamic interneuron (figure 12.4). Another synapse may release GABA onto a $GABA_A$ receptor, which gates a chloride channel. This connection is usually inhibitory because a particular protein pump (KCC2) sets E_{Cl} near the resting membrane potential (figure 7.4). However, certain terminals express a different pump (NKCC1) that sets E_{Cl} positive to the resting potential. In this case a synapse matching GABA to $GABA_A$ will *depolarize* the postsynaptic terminal and be excitatory. This allows a circuit to diametrically reverse its function, not by altering the anatomical structure, nor the transmitter, nor its receptor. Instead it simply swaps chloride pumps.

Neurons designed for local computation

Neurons differing from the standard polarized design are numerous, so here we note from retina two radical alternatives. One type radiates dendrites symmetrically from its cell body to collect chemical synaptic inputs. This *starburst* neuron lacks an axon but forms chemical outputs at the distal dendritic tips (figure 11.24). The design is such that a visual stimulus moving centripetally, from cell body toward the dendritic tips, releases GABA onto a ganglion cell and thereby blocks its spiking to that direction of motion (figure 11.24).

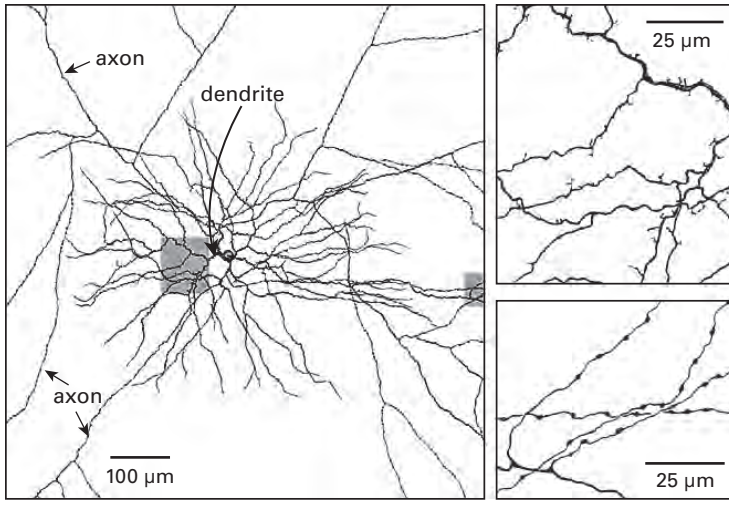


Figure 7.12
Polyaxonal amacrine neuron reverses the standard polarized design. **Left:** Instead of collecting input on dendrites and funneling to a single axon, it radiates multiple axons from distal dendritic tips. Gray boxes indicate regions shown at higher magnification. **Upper right**: Dendrites express spines for receiving inputs. **Lower right**: Axons express varicosities for sending outputs. Reprinted with permission from Davenport et al. (2007).

This neuron breaks another rule: whereas most neurons release only one chemical transmitter, the starburst neuron releases *acetylcholine* in addition to GABA. These transmitters are packaged by different transporters into different vesicles that cluster at different presynaptic sites and contact different dendrites. Thus, a starburst process, which computes locally and connects locally, can, by releasing different transmitters onto different neurons, evoke opposite responses to the same stimulus. One computation produces two outcomes for the same price.

Another radical design radiates dendrites symmetrically about the cell body to collect local information; then, each dendrite radiates an axon from its distal tip to broadcast the information over millimeters (figure 7.12). This *polyaxonal amacrine* neuron is polarized, but in reverse. The cell body, rather than converging information for a single axon, diverges via multiple axons in all directions (figure 7.12).

We conclude that the core rule for designing a neuron is to build it for a particular task. This achieves the needed performance for least cost (chapters 9, 12, and 13).

Glial cells in design of neurons

Glial cells comprise a substantial fraction of the brain's volume (Halassa & Haydon, 2010). In white matter (tracts) *astrocyte* cell bodies and processes use more than 30% of the space. Myelin sheaths occupy an additional 25%, and *oligodendrocyte* cell bodies that provide the myelin wrapping use an additional 13%. Thus, in tracts the space allotted to glia comes to about 65% (figure 13.21; Perge et al., 2009). In gray matter (circuits) the fraction for astrocyte processes varies locally by design, but overall is about 10%, plus some added allowance for cell bodies (Mischenko et al., 2010; Schüz & Palm, 1989). So on average, the total space for glia in gray matter is roughly 15%.

Glia expend considerable energy. For example, the mitochondrial volume fraction of astrocyte processes in white matter is more than 3%, more than twice that of the myelinated axons. In the optic nerve astrocytes contain more than 70% of the mitochondria (see below, figure 7.21; Perge et al., 2009). In gray matter less than 5% of the mitochondria are in glia, but gray matter processes information in dense neural circuits, so its overall metabolic rate per volume is threefold higher (Attwell & Laughlin, 2001; Harris & Attwell, 2012). Given glia's substantial costs in space and energy, what are the benefits to neural design?

White matter: Benefits of myelin and astrocytes

A naked axon conducts action potentials efficiently, but conduction velocity rises only as $\sqrt{d}$. Therefore, where speed is required, the naked axon must become inordinately thick. This cost is accepted for a command neuron that triggers the escape response of an invertebrate; most famously, the squid giant axon is about 1 mm in diameter. This works if there are only a few giant axons, but they could not be used routinely because they would take far too much brain space. Yet vertebrates move fast and need many fast axons.

When speed requires an axon thicker than about 0.5 μm, the design solution is for an oligodendrocyte process to wrap a segment of the axon in a multilayered, jelly roll of plasma membranes. This is myelin. Its multiple layers effectively reduce the axon's membrane capacitance and increase its resistance. This increases space constant and reduces time constant. These improvements allow the advancing foot of the action potential to spread further and faster. Thanks to myelin wrapping, action potential velocity increases in direct proportion to axon diameter at about 6,000 mm/s per micron diameter.

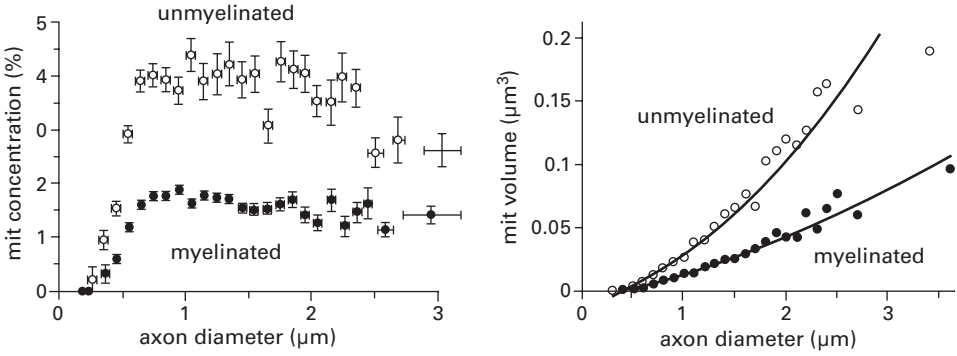


Figure 7.13

Myelination saves energy as well as conduction time. Left: In myelinated axons thicker than about 0.7 μm mitochondria occupy a constant proportion of cytoplasmic volume, about 1.5%. In unmyelinated segments of the same axons mitochondria occupy about 4% of cytoplasmic volume. **Right**: Mitochondrial volume per unit axon length rises linearly with diameter for fine axons ($d < 0.7$ μm) and quadratically for thicker ones. This figure compares ganglion cell axons within the retina, where they are unmyelinated, to their continuations in the optic nerve where they are myelinated. Thus, both plots represent the same neuron types. Reprinted with permission from Perge et al. (2009).

The myelin-wrapped segment extends for about 0.5–1 mm, so as the voltage pulse flashes passively across this distance, it soon encounters a naked spot of neural membrane (*node of Ranvier*) which concentrates sodium channels. These are of a specific type, Nav1.6, that open rapidly and synchronously to regenerate the action potential—which then continues its passive course to the next node. The sodium channels pack so densely at the node (up to 2,000/μm^2) that the potassium channels needed to repolarize are displaced laterally.

Nodal spacing increases directly with axon diameter. This works because thicker axons produce larger nodal currents and increase the number of myelin wraps, further increasing the space constant. In systems where spike arrival time is critical, nodal spacing can be tweaked to compensate for different conduction distances (Cheng & Carr, 2007; Carr & Boudreau, 1993). One might imagine that concentrating sodium channels at a few sites rather than distributing them over the whole axon would save energy, and this proves to be so (figure 7.13). This saving, though substantial, does not begin to explain the threefold difference in energy cost of white matter compared to gray matter and, thus, its far sparser supply of blood vessels. That is explained by the absence of synaptic currents (Harris & Attwell, 2012).

Astrocytes in white matter are critical to the design that generates an efficient action potential. Following a spike, the axon repolarizes by releasing a pulse of potassium into the nodal extracellular space. Because neighboring axons are all firing, the concentration gradient needed for the pulse to diffuse from the node is diminished. Therefore, potassium must be removed rapidly by sodium-potassium pumps (chapter 6). But where to place them?

Space at the nodal axon membrane is so fully occupied by sodium channels that the potassium channels are displaced laterally. Thus, the nodal membrane cannot accommodate large numbers of pump molecules. So the design places many fast-binding pumps on the astrocyte membranes and some slower ones along the axon (Ransom et al., 2000). Thus, astrocytes in white matter are key to rapid removal of extracellular potassium, and this may explain their large proportion of space and energy capacity.

Gray matter: astrocyte compartments for transmitter diffusion

Certain computations benefit when neighboring synapses operate independently of each other. Other computations benefit when neighboring synapses share their transmitter. The degree of independence versus sharing is set partly by the degree of astrocyte wrapping. Certain types of synapse are individually wrapped, allowing each pulse of transmitter to bind the postsynaptic receptors and then to be removed by transporter proteins on the astrocyte membranes (see below, figure 7.17). Other types are poorly wrapped, allowing longer persistence and spread of transmitter to neighboring synapses (see below, figure 7.16.

Still other types of synapse provide multiple synaptic contacts from closely spaced release sites that are all wrapped together by a glial capsule (figure 12.4). The glial membranes densely express transporter proteins for the particular transmitter released within the capsule (Josephson and Morest, 2003). This allows pulses from one release site to spill over to neighboring receptor patches with consequences to be discussed below. Such *glomerular synapses* are used in various locations, including spinal cord, cochlear nucleus, thalamus (figure 12.4), and cerebellum (see below, figure 7.16).

Other tasks for glia

Astrocytes display myriad other properties. For example, they respond to neuronal activity and neurotransmitters via G-protein-coupled receptors. Moreover, they release gliotransmitters, such as glutamate, D-serine and ATP, which act on neurons. Astrocyte-derived ATP can modulate synaptic

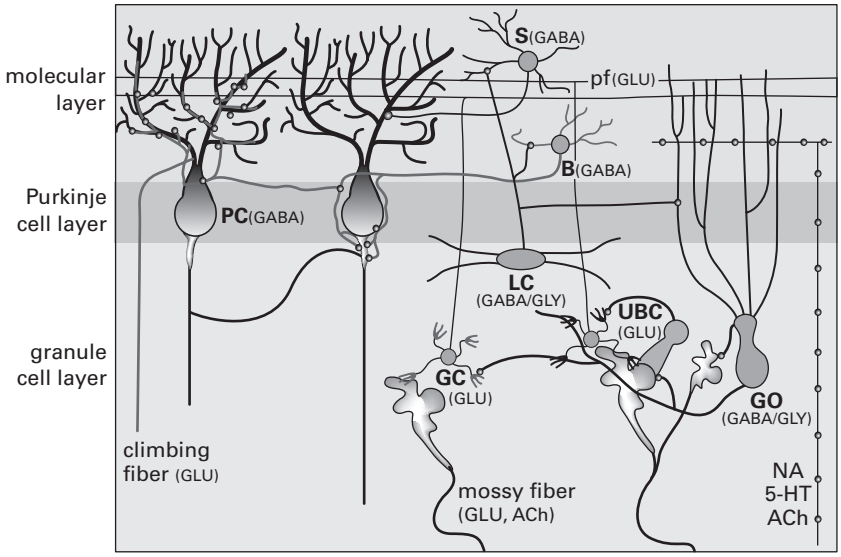


Figure 7.14

Wiring diagram of cerebellar cortex. Granule cell (GC) integrates excitatory contacts (from distant mossy fibers and local unipolar brush cells (UBC)) with inhibitory contacts (from Golgi neurons (GO) with excitatory contacts from parallel fibers). Purkinje cell (PC) integrates excitatory contacts from a single climbing fiber and 175,000 parallel fibers (pf) with inhibitory contacts from stellate (S) and basket (B) cells that receive inhibitory input from the Lugaro cell (LC). Purkinje cell sends recurrent GABAergic contacts to axon initial segment of neighbors. The basket cell extends fine processes to encapsulate the Purkinje axon initial segment, an arrangement associated with electrical inhibition. GLU, glutamate; Ach, acetylcholine; GLY, glycine; NA, noradrenaline; 5-HT, serotonin. Redrawn from Sotelo (2008).

transmission, either directly or through its metabolic product, adenosine (Schmitt et al., 2012). Astrocytes are also interposed between blood vessels and neurons and thus play important roles in regulating metabolic responses (Howarth, 2014). This does not exhaust the list of properties and contributions of glial cells to neural function. However, understanding remains too incomplete to fully grasp how the various features contribute to efficient design.

Each neuron's design serves a larger circuit

Cerebellum illustrates the extent to which neurons and glia are adapted for specific functions within a larger circuit. A schematic diagram of cerebellar

cortex identifies the circuit's four types of excitatory neuron, and four types of inhibitory neuron (figure 7.14). Although a complete account of their cooperative effort remains a goal, enough is already known to see how some features of their functional architecture serve efficient processing.

The cerebellar circuit performs two operations. First, it remaps information coded at high mean rates by a modest number of mossy fibers to much lower mean rates carried by a much larger number of granule cells. This occurs in the inner synaptic layer. Second, it formulates an output by a quite small number of Purkinje cells in the outer layer. The granule cells project their individually sparse representation into the outer layer via a massive array of axons, running parallel to save wire (chapter 13). Each Purkinje neuron integrates single synaptic inputs from 175,000 parallel fibers to send an output pattern via its axon. To efficiently implement these two operations—remap and send an output—requires different functional architectures.

Functional architecture of inner synaptic layer

The inner synaptic layer densely packs astronomical numbers of granule cells in small clusters (figure 7.15). The cell bodies are irreducibly small, 6–7 μm in diameter. The cell body is filled almost completely by the nucleus, leaving a mere crescent of cytoplasm essential for protein synthesis. The dendrites are limited to four short processes, about 12 μm long, terminating in specialized claws that collect all the synaptic input (figures 7.1 and 7.14).

The granule cell transmits to Purkinje cells with an irreducibly thin axon. Its diameter can be less than 0.2 μm, which allows just enough space for 1–2 microtubules plus their motors and cargo, and an internal resistance just low enough to prevent noise. Any narrower and the current entering through one sodium channel would see an internal resistance so high as to bring the membrane to threshold. This would allow a lone channel opened solely by thermal buffeting to generate a spontaneous spike, thereby introducing noise (Faisal et al., 2005). This design—thin axon—supports only a very low mean spike rate because, with a high surface area/volume ratio, it can contain relatively few mitochondria.

Granule cell design allows no input synapses, except at the four specialized claws. This provides high membrane resistance, which reduces the cost of maintaining resting potential. Even so, the bill for resting potential is considerable (see below, figure 7.21). As a neuron shrinks, the ratio of surface area to volume increases as 1/diameter, so the ratio for a granule cell is nearly 10-fold greater than for a Purkinje cell. Such a relatively large membrane area presents an expanse for leakage, and since the granule cell is the

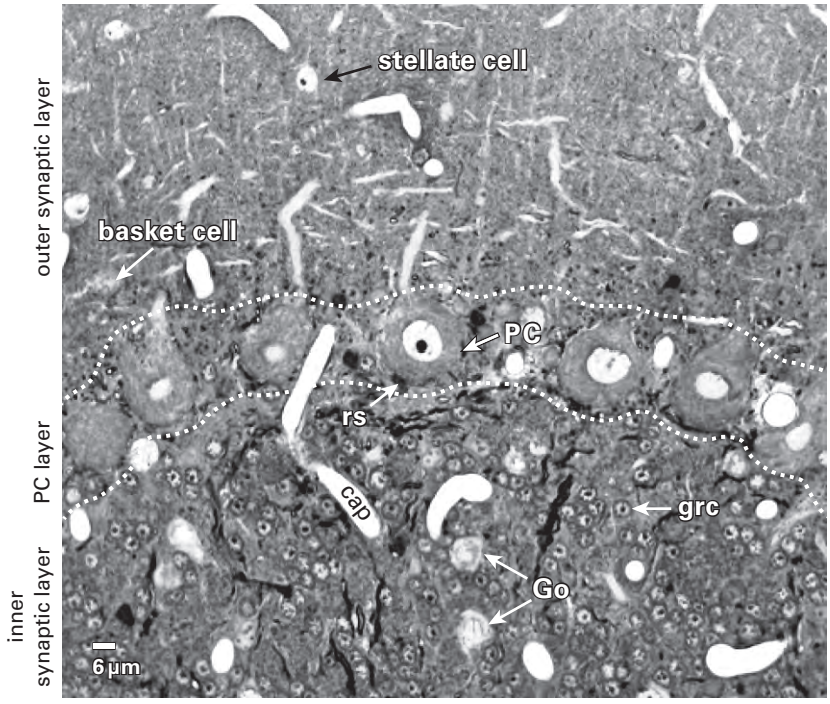


Figure 7.15

Largest cerebellar neuron occupies more than a 1,000-fold greater volume than smallest neuron. Thin section (~1 µm) through monkey cerebellar cortex. Purkinje cell body (PC) and nucleus are far larger than those of granule cell (grc). The latter cluster to leave space for mossy fiber terminals to form glomeruli with grc dendritic claws and space for Golgi cells (Go). Note rich network of capillaries (cap). Fine, scattered dots are mitochondria. Courtesy of E. Mugnaini.

brain's most numerous neuron, this small cost grows large (see also chapter 13).

Much of the inner synaptic layer is occupied by the large axon terminals of mossy fibers (figures 7.1 and 7.16). A terminal interlaces with multiple (~15) dendritic claws, each from a different but neighboring granule cell, and forms a complex knot (*glomerulus*), nearly as large as a granule cell body (figures 7.1 and 7.16). The mossy fiber axon fires at an unusually high mean rate (up to 200 Hz) and is therefore among the brain's thickest (figure 4.6).

To match the axon's high rate, a terminal expresses 150 active zones, 10 per postsynaptic granule cell (figure 7.16). These sites are capable of driving

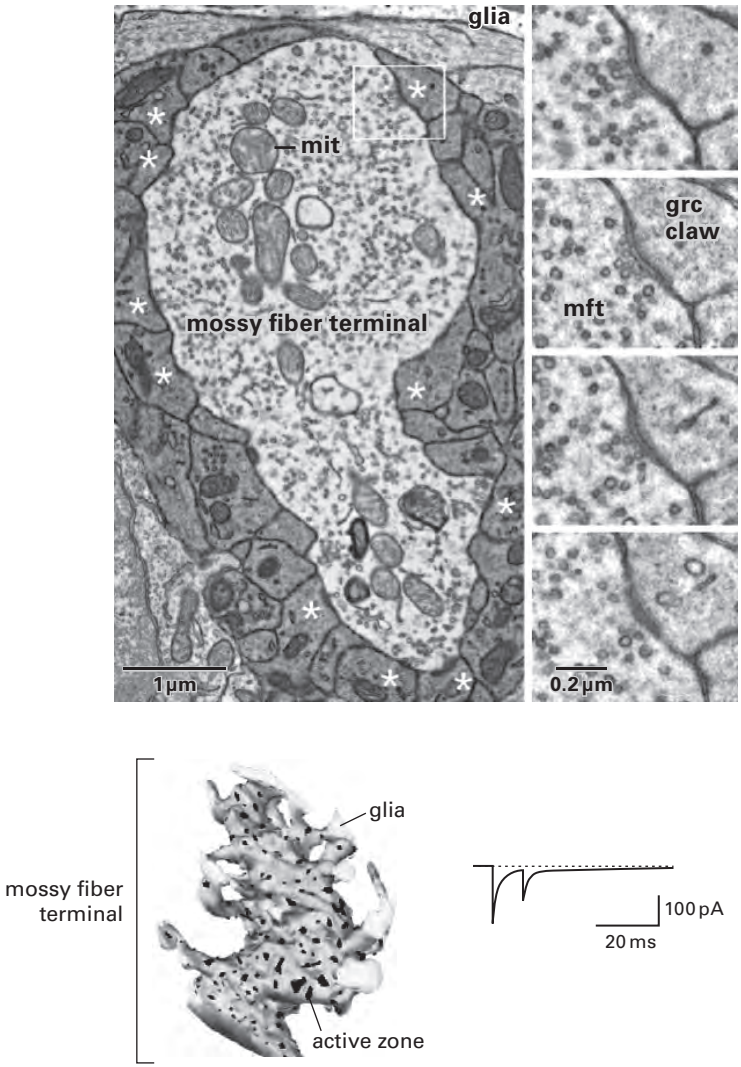


Figure 7.16

A large terminal may contact many small dendrites without intervening glia. This design pools transmitter from many synapses. It densensitizes postsynaptic receptors to reduce spike rate while improving temporal precision. **Upper left**: Central mossy fiber terminal contacts onto granule cell dendritic claws at release sites marked by *. Postsynaptic processes pack closely with no intervening glia. The whole structure (mft + grc claws) is partially encapsulated by glial membranes (above, shaded). **Upper right**: Serial sections through one release site, showing docked vesicles. **Lower left**: Three-dimensional reconstruction of mossy fiber terminal. Terminal provides hundreds of active zones. Each active zone has approximately seven neighbors closer than 1 µm. Glia cover about 20% of outer surface. **Lower right**: Granule cell strongly excited by first spike in mossy fiber but less strongly by next spike 10 ms later. Reprinted with permission from Xu-Friedman & Reghr (2003).

a granule cell synchronously, with one vesicle per presynaptic spike, up to frequencies of 700 Hz (Saviane & Silver, 2006). This direct mapping of spike input to vesicle output preserves the bandwidth and dynamic range of the mossy fiber axon. However, to sustain signaling, the terminal must replenish, dock, and prime vesicles at the axon's mean rate. The enlarged terminal provides the volume to accommodate a releasable pool of about 5×10^4 vesicles, about 300 per active zone. Given that synchronous release maintains bandwidth and signal range, how is the terminal designed to maintain the other determinant of information rate, S/N?

To maintain S/N, the enlarged terminal contacts many small dendrites without intervening glia (figure 7.16). This design allows transmitter released at one site to diffuse to neighboring sites. Thus, a postsynaptic receptor cluster on one granule cell dendrite receives a pulse of transmitter from its own release site, plus pulses from at least seven other sites less than 1 μm distant. This *spillover* of transmitter from several sites reduces noise produced by probabilistic release, and being adjacent, diffusion affects only slightly response duration. To allow spillover, the synaptic complex is largely devoid of glia expressing transporters. Consequently, transmitter released across the terminal at high rates tends to persist, and this too is put to good use.

The persistent spillover densensitizes postsynaptic receptor clusters, which acts as a negative feedback to reduce the amplitude of the granule cell's excitatory postsynaptic currents (EPSCs; figure 7.16). Thus, a granule cell can integrate numerous temporally correlated inputs to improve temporal precision, yet since each input delivers a small postsynaptic current, the mean spike rate is drastically reduced. In other words, a glomerulus with large terminal, multiple postsynaptic clusters, and scant glia, is well designed to remap information from densely coding mossy fiber axons to sparsely coding granule cells.

The roughly 50-fold step down of mean spike rate from mossy fiber to granule cell is efficient partly because bits/spike increases for lower rates (figure 3.5). Efficiency is further increased by improving a spike's temporal precision (chapters 5 and 6) to generate a *sparse code*, in which each granule cell is mostly silent and only fires brief, well-timed bursts at frequencies greater than 100 Hz (Ruigrok et al., 2011). Timing precision and brevity of the burst are enhanced by another contributor to the glomerulus: the Golgi neuron, whose cell bodies distribute sparsely within the inner synaptic layer and contribute GABAergic inhibitory contacts to the glomerulus (figure 7.14). This contribution to sparsifying the signal is relatively cheap because the individual cell is of modest size, low rate, and modest numbers. Moreover, inhibition is far cheaper than excitation (see below, figure 7.20).

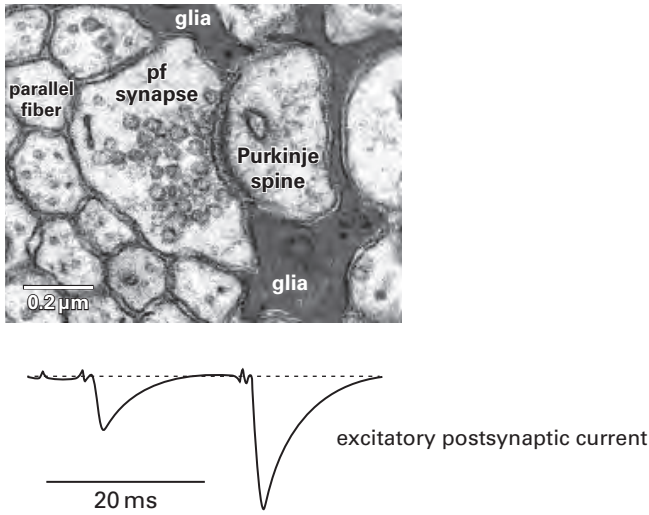


Figure 7.17
Parallel fiber synapse to Purkinje cell spine is ensheathed by glia and is facilitated at high frequencies. Upper: Glial wrapping is nearly 70%, which greatly reduces spillover between neighboring active zones. **Lower**: Excitatory postsynaptic response to brief burst of spikes at parallel fiber synapse (two spikes at 50 Hz). The second response is larger, probably because two vesicles were released simultaneously. The enhanced pulse of transmitter diffuses further from the release site to reach extrasynaptic NMDA receptors (Nahir & Jahr, 2013). Reprinted with permission from Xu-Friedman & Reghr (2001).

Now consider the circuit's second operation—integrate granule cell messages to form a Purkinje cell output. The granule cell axon ascends to the outer synaptic layer and branches as a T to run parallel with its neighbors and perpendicular to the fan-like Purkinje cell dendritic arbors (figures 7.1 and 7.14). Extending for 2 mm, it contacts one dendritic spine on many neurons, including Purkinje cell, stellate, and basket cells. The presynaptic active zone docks substantial numbers of vesicles (figure 7.17) so that the release of one vesicle does not deplete the ready pool. This design allows a second spike to admit sufficient calcium to release several vesicles simultaneously and produce a larger postsynaptic response (figure 7.17). The glutamate from single and multiple releases spills over to NMDA receptors just beyond the postsynaptic density, thereby extending the Purkinje neuron's time window for coincidence detection (figure 7.5).

The parallel fiber contacts upon spines tend to be well wrapped by glia, which reduces spillover between neighboring synapses (figure 7.17). Thus,

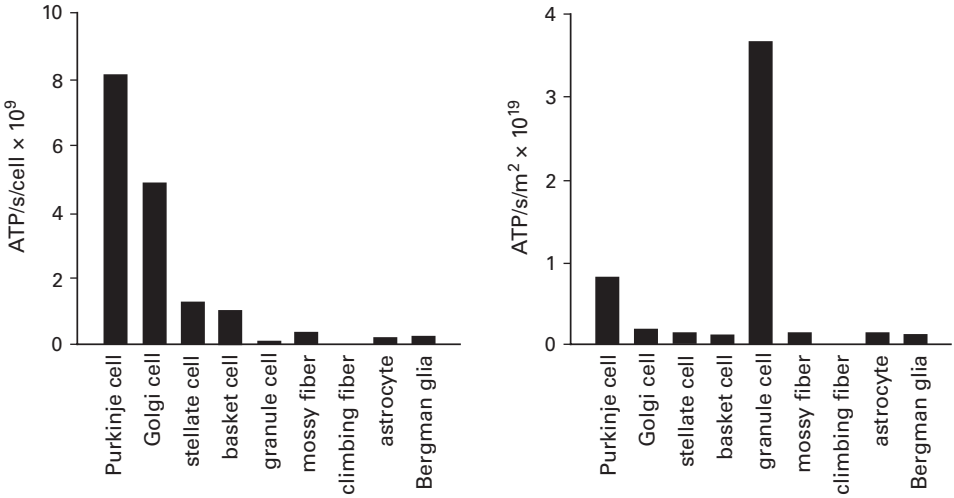


Figure 7.18
Energy costs by cell type. Left: Purkinje cell is the most expensive neuron, and granule cell is cheapest. **Right**: Granule cell array is the most expensive, and Purkinje cell array is far cheaper. Glial cells are cheap individually and as arrays. Reprinted with permission from Howarth et al. (2012).

a granule cell's output synapse is structured to reliably deliver a precisely timed message—privately (Nahir & Jahr, 2013). All 150,000 parallel fiber synapses onto an individual Purkinje cell tend to be wrapped by the same glial cell (*Bergman glia*), whose form mimics that of the Purkinje cell's extensive dendritic tree (figure 7.1).

The different tasks of inner and outer cerebellar layers and their consequent different designs illustrate why there can be no generic neuron. In the inner layer, high-rate synapses improve S/N by pooling excitatory responses and sharpen timing precision with feedback inhibition—to allow a burst of information-rich spikes (figure 7.16). In the second case, spikes deliver this information by a synaptic design that facilitates to a burst (figure 7.17). The first design reduces glial wrapping to enhance synaptic spillover; the second design does the opposite. Now we can ask: what are the costs of these two designs?

Costs of different neuron designs

Energy costs by cell type

When the various energy costs are totaled, the individual Purkinje cell proves to be the most expensive neuron, and granule cell proves to be the

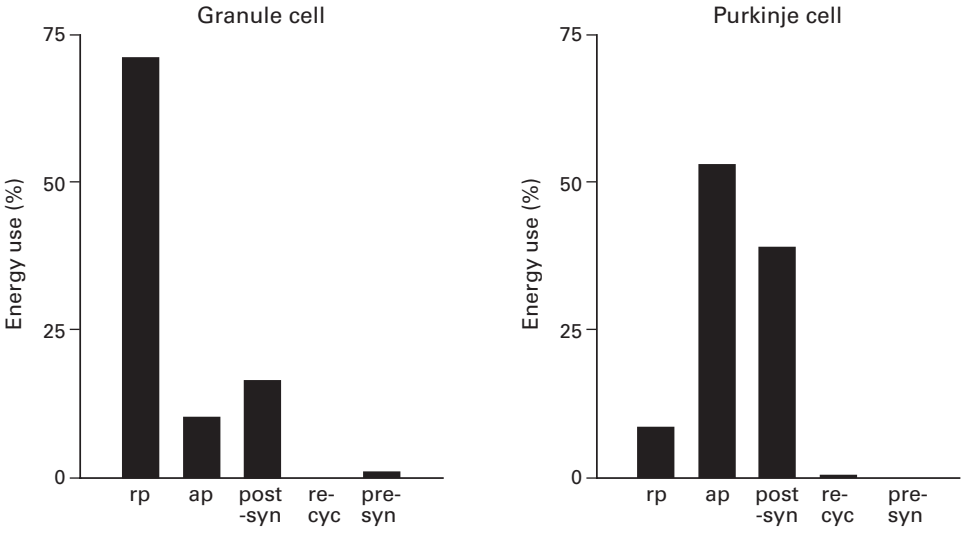


Figure 7.19
Energy costs by cell function: Granule cell versus Purkinje cell. rp, resting potential; ap, action potential; postsyn, postsynaptic receptor currents; recyc, transmitter recycling (ATP for uptake by glial transporters, metabolic processing, and vesicular transporters); presyn, presynaptic calcium entry and vesicle cycling. Reprinted with permission from Howarth et al. (2012).

cheapest (figure 7.18). This should be no surprise since the Purkinje neuron is far larger, receives many more synapses, and fires at a greater than 10-fold higher mean rate. On the other hand, it seems mildly surprising that, the granule cell array is approximately fourfold more expensive than the Purkinje cell array. This reflects the granule cells' outnumbering the Purkinje cells by 274 to 1.The local inhibitory neurons are cheap individually and as arrays, except for the Golgi neuron. The latter is costly as a single cell because it receives a high rate of excitation from mossy fibers and is considerably larger than the granule cell (figure 7.14). Its major cost (75%) goes for postsynaptic receptor currents (Howarth et al., 2012). However, it is cheap overall because the array is sparse. Cerebellar glial cells are cheap individually and as arrays.

Energy costs by cellular function and computational stage

Each part of a neuron has its own cost, and the proportions vary according to the cell's design (figure 7.19). Thus, the granule cell's thin axon (which, ascending to the outer synaptic layer, branches as a T to become the irreducibly fine parallel fiber) sends each action potential cheaply. And, because

of sparse coding, its mean spike rate is low. Therefore, less than 10% of a granule cell's energy goes for spikes (Howarth et al., 2012). On the other hand, because the fine axon has a high surface area/volume, much energy is needed to maintain the resting potential against leaks. Postsynaptic currents at the input are costly, but synaptic release along the parallel fiber is cheap.

The Purkinje cell reverses the pattern. Its thick axon sends each spike at greater expense; moreover, the cell fires at high mean rates (figure 4.6). Therefore, the cell uses most energy for action potentials. The Purkinje cell's second greatest cost is for postsynaptic excitatory currents due to its vast number ($>10^5$) of glutamatergic contacts from parallel fibers and the climbing fiber. Its costs for recycling vesicles and transmitter are negligible because, except for a very few recurrent contacts, a Purkinje cell's outputs are all outside the cerebellar cortex.[14]

Note that for both neuron types there are presynaptic costs. These include extruding accumulated calcium from synaptic terminals via a sodium/calcium exchanger, retrieving vesicles by endocytosis, refilling vesicles via a transporter, and retrieving and resynthesizing transmitter. For both granule and Purkinje neurons these presynaptic costs are negligible (figure 7.19), and the same is true for all the inhibitory interneurons (Howarth et al., 2012). Some of these processes are cheap because they use chemistry (neural exo- and endocytosis, metabolic processing of transmitters). Calcium extrusion is cheap because, although it uses energy for active transport, the calcium current to release a vesicle is miniscule compared to the current needed to open the calcium channel and compared to the postsynaptic current that the vesicle evokes. This reemphasizes the economy of chemical processes and the high cost of electrical amplification.

Cerebellar cortex contains more types of inhibitory neurons than excitatory ones (figure 7.14). However, the inhibitory neurons distribute sparsely and their synaptic currents are far cheaper. Therefore, excitation costs nearly four-fold more than inhibition (figure 7.20). Moreover, since the inhibitory processes serve to reduce redundancy and restrain expensive excitation, they seem a particularly good investment.

Cerebellar neurons parcel out their computations so neatly that the computational costs can be evaluated (figure 7.21). The inner synaptic layer is tasked to step down mossy fiber firing rates and concentrate information with a sparse code in granule cells. This costs slightly more than half of the total energy. The sparse code must then propagate to all the neurons in the outer synaptic layer, and this costs nearly one third of the total. Finally,

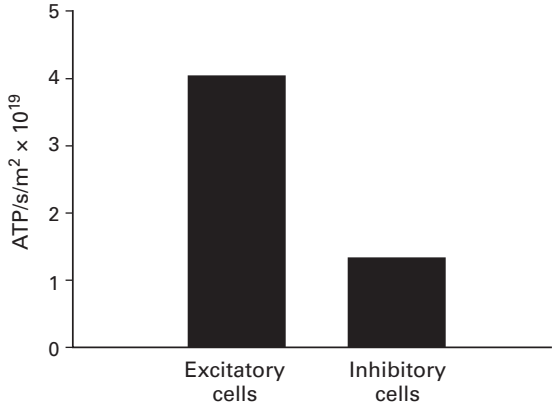


Figure 7.20
Excitatory neurons in cerebellar cortex cost nearly fourfold more than inhibitory neurons. Reprinted from Howarth et al. (2012).

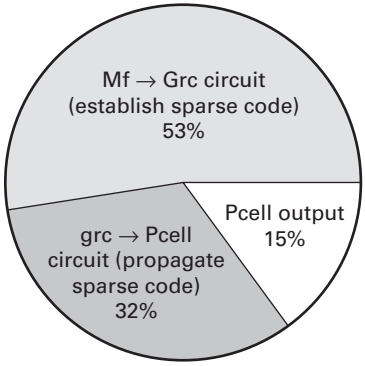


Figure 7.21
Input layer to cerebellar cortex consumes most energy. Intermediate layer consumes less, and output is cheapest. Mf, mossy fiber; Grc, granule cell; Pcell, Purkinje cell. Relabeled from Howarth et al. (2012).

the result of the outer-layer computation must be sent as Purkinje cell output which, due to the step down in neuron numbers, costs least (15%). These calculations correspond rather neatly to the distribution of *cytochrome oxidase*, which serves the final step of mitochondrial energy production: dense patches within the inner synaptic layer, corresponding to the synaptic glomeruli versus broad but weak distribution in the outer synaptic layer and strong in the Purkinje neurons (figure 13.20).

Conclusion

This chapter has focused on a few of the many ways that a neuron integrates information encoded at the input as chemical signals to transmit an output electrically at speed over distance. A core point is that "the neuron" is a shape-shifter. It can assume any form within limits ultimately set by physics, chemistry, and cell biology in order to function according to the basic principles of economy in neural design. The chapter did not explain how a neuron matches its internal chemical signals to achieve high efficiency, or how it couples them with equally high efficiency to its electrical outputs. That is the topic of chapter 8.