A Computer Scientist's Guide to Cell Biology:

A Travelogue from a Stranger in a Strange Land

William W. Cohen
Machine Learning Department
Carnegie Mellon University

To Susan, Charlie, and Joshua.

Table of Contents

List of Figures

Introduction

For the past few months, I have been spending most of my time learning about biology. This is a major departure for me, as for the previous 25 years, I've spent most of my time learning about programming, computer science, text processing, artificial intelligence, and machine learning. Surprisingly, many of my long-time colleagues are doing something similar (albeit usually less intensively than I am). This document is written mainly for them—the many folks that are coming into biology from the perspective of computer science, especially from the areas of information retrieval and/or machine learning—and secondarily for me, so that I can organize and retain more of what I've learned.

I find it helpful to think of "biology" in three parts. One part of biology is **information about biological systems** (for instance, how yeast cells metabolize sugar). This is the focus of most introductory biological textbooks and overviews, and is the essence of what biologists actually study—what biologists are trying to determine from their experiments. However, it is not always what biologists spend most of their time talking about. If you pick up a typical biology paper, the *conclusions* are typically quite compact: often all the new information about biological systems in a paper appears in the title, and almost always it can be squeezed into the abstract. The bulk of the paper is about **experimental methods** and how they were used—this, I consider to be the second part of "biology." The third part of "biology" is the **language and nomenclature** used, which is rich, detailed, and highly impenetrable to mere laymen. To read and understand current literature in biology, it is necessary to have some background each of these three parts: core biology, experimental procedures, and the vocabulary.

I like to think of the last few months as something like a field trip to a new and exotic land. The inhabitants speak a strange and often incomprehensible language (the nomenclature of biology) and have equally strange and new customs and practices (the experimental methods used to explore biology). To further confuse things, the land is filled with many tribes, each with its own dialect, leaders, and scientific meetings. But all

the tribes share a single religion, with a single dogma—and all their customs, terms and rituals are organized around this religion. The highest goal of their religion is discover truth about living things—as much truth as possible, in as much detail as possible. This truth is "core" biology—information about living things. Knowing this "truth" is important, of course, but merely knowing the "truth" is not enough to understand a community of biologists, just as reading the Torah is not enough to understand a community of Jews.

In this document, I will provide a short introduction to "core" cell biology, mainly to introduce the most common terms and ideas. In doing so, I will occasionally oversimplify. This is deliberate. Computer scientists are used to analyzing complex systems by analyzing successively more complex abstractions, many of which are "real" (to the extent that anything computational is "real"): for instance, a push-down automaton is a generalization of a finite state machine, and both are useful for many real-world problems. One would like to operate in the same way in understanding biology, for instance, by first analyzing "finite-state" organisms, and then progressing to more complex ones. In biology, however, it is hardly ever the case that a clean and comprehensible abstract model perfectly models a real-life organism, so (almost) every simple general statement about how organisms function needs to be qualified—a tedious process in a document of this sort. I will also, by necessity, omit many interesting details, again deliberately. For a more comprehensive background on biology, there are many excellent textbooks, written by people far more qualified, some of which are mentioned in the final section of this paper).

After discussing "core" cell biology, I will then move on to discuss the most widely-used experimental procedures in biology. I will focus on what I perceive to be the high-level principles behind experimental procedures and mechanisms, and relate them to concepts well-understood in computer science whenever possible. Comments on nomenclature and background points will be made in side boxes.

How Cells Work

Prokaryotes: the simplest living things

One of the most fundamental distinctions between organisms is between the **prokaryotes** and the **eukaryotes**. **Eukaryotes** include all vertebrates (like humans) as well as many single-celled organisms, like yeast. The simpler **prokaryotes** are a distinct class of organisms, including various types of **bacteria** and **cyanobacteria** (blue-green algae). The best-studied prokaryote is *Escherichia coli***,** or *E. coli* to its friends, a bacterium normally found in the human intestine. Like more complex organisms, the life processes of *E. coli* are governed by the "**central dogma"** of biology: **DNA** acts as the long-term information storage; **proteins** are constructed using DNA as a template; and to construct a particular protein, a corresponding section of DNA called a **gene** is **transcribed** to a molecule called a **messenger RNA** and then **translated** into a protein by a giant molecular complex called a **ribosome**. After the protein is constructed, the gene is said to be **expressed.** To take a computer science analogy, DNA is a stored program, which is "executed" by transcription to RNA and expression as a protein. The "central dogma" is summarized in Figure 1.

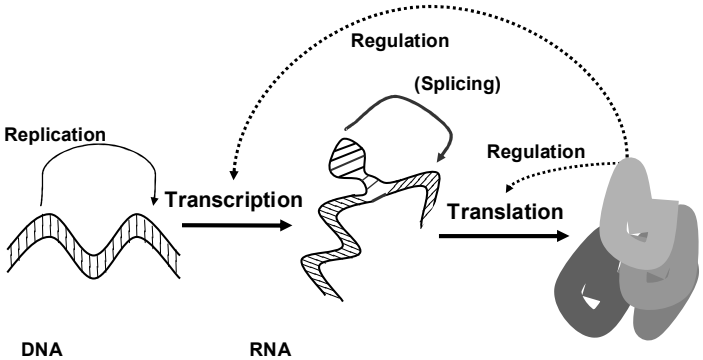
"Bacteria" can refer to all prokaryotes, but more commonly refers to **eubacteria**, a subclass.

DNA molecules are sequences of four different components, called **nucleotides**. Proteins are sequences of twenty different components called **amino acids**. Translation maps triplets of nucleotides called **codons** to single proteins: famously, nearly the same triplet-to-protein mapping is used by all living organisms.

This same process of DNA-to-mRNA-to-protein is carried out by all living things, with some variations. One variation, which occurs again in all organisms, is that some RNA molecules are used directly by the cell, rather than being used only indirectly, to make proteins. (For instance, key parts of ribosomes are made of **ribosomal RNA**, and mRNA translation also involves special molecules called **transfer RNAs.**)

Messenger RNA, ribosomal RNA, and transfer RNA are abbreviated as **mRNA**, **rRNA**, and **tRNA**, respectively. Another type of RNA, **small nuclear RNA** (**snRNA**), plays a role in splicing. A **gene product** is a generic term for a molecule (RNA or protein) that is coded for by a gene.

A second variation is that in the more complex **eukaryotic** organisms**,** mRNA is processed, before translation, by **splicing** out certain subsequences called **introns**. Surprisingly, the process of DNA-to-RNA-to-proteins is similar across all living organisms, not only in outline, but also in many details: scores of the genes that code for essential steps of the "central dogma processes" are highly similar in every living organism.



DNA
• bases A,T,C,G
• double-helical
• information storage for cell

RNA
• bases A,U,C,G
• varying shapes
• (usually) transfers info from DNA

Proteins
• long sequence of 20 different amino acids
• widely varying shapes
• carries out most functions of cells including translation and transcription
• regulates translation and transcription

The "central dogma" of biology: DNA is **transcribed** to RNA; mRNA is **translated** to proteins; proteins carry out most cellular activity, including control (**regulation***)* of transcription, translation, and **replication** of DNA.

(In more detail, RNA performs a number of functional roles in the cell besides acting as a "messenger" in mRNA.)

Figure 1. The "central dogma" of biology.

Prokaryotes are extremely diverse—they live in environments ranging from hot springs to ice-fields to deep-sea vents, and exploit energy sources ranging from light, to almost any organic material, to elemental sulphur. However, most prokaryotes are structurally quite simple: to a first approximation, they are simply bags of proteins. More specifically, a prokaryotic organism will consist of a single loop of DNA; an outer **plasma membrane** and (usually) a **cell wall;** and a complex mix of chemicals that the membrane encloses, many of which are **proteins**. Proteins are also embedded in the membranes of a cell.

Membranes are composed of two back-to-back layers of fatty molecules called **lipids,** hence biological membranes are often called **bilipid membranes.**

A **protein** is a linear sequence of twenty different building blocks called **amino acids**. Different amino-acid sequences will fold up into different shapes, and can have very different chemical properties. Proteins are typically hundreds or thousands of amino acids in length. The individual amino acids in a protein are connected with **covalent** bonds, which hold them together very tightly. However, when two proteins interact, they generally interact via a number of weaker inter-molecular forces; the same is true when a protein interacts with a molecule of DNA.

A **covalent bond** between two atoms means that the atoms share a pair of electrons. Weaker, inter-molecular forces include **ionic bonds** (between oppositely-charged atoms), and **hydrogen bonds** (in which a hydrogen atom is shared).

One attractive force that is often important between proteins is the **van der Waals force**, a weak, short range electrostatic attraction between atoms. Although the attraction between individual atoms is weak, van der Waals forces can strongly attract large molecules that fit very tightly together. Another strong "attractive force" is **hydrophobicity:** two surfaces that are **hydrophobic,** or repelled by water, will tend to stick together in a watery solution, especially if they fit together tightly enough to exclude water molecules. Proteins, like the amino acids from which they are formed, vary greatly in the degree to which they are attracted to or repelled by water.

The importance of all this is that the interactions between proteins in a cell are

A **bacteriophage**, or **phage,** is a virus that infects bacteria.

often highly **specific:** a protein P may interact with only a small number of other proteins—proteins to which some part of P "fits tightly." The chemistry of a cell is largely driven by these sorts of **protein-protein interactions.** Proteins also may interact strongly with certain very specific patterns of DNA (for instance, a protein might bind only to DNA containing the sequence "TATA") or with certain chemicals: many of the proteins in the plasma membrane of a bacteria, for instance, are **receptor proteins** that sense chemicals found in the environment.

Even simpler "living" things: viruses and plasmids

There are constructs simpler than prokaryotes that are lifelike, but not considered alive. **Viruses** contain information in nucleotides (DNA or RNA), but do not have the complete machinery needed to replicate themselves. Instead, they infect some other organism, and use its machinery to reproduce—just as an email virus uses existing programs on an infected machine to propagate. One well-studied virus is the **lambda phage,** which consists of a protein **coat** that encloses some DNA. The protein coat has the property that when it encounters the outer membrane of a cell, it will attach to the membrane, and insert the DNA into the cell. This DNA molecule has ends that attract each other, so it will soon form a loop—a loop similar to, but smaller than, the double-stranded loop of DNA that contains the genes in the host cell.

Most of the DNA in a cell is contained in **chromosomes.** In prokaryotes, a chromosome is generally a single long loop of DNA. Eukaryotic chromosomes have a more complex structure, and typical eukaryotes have several chromosomes.

Even though this DNA loop is not in the expected place for DNA—that is, it is not part of any **chromosome** of the cell—the machinery for transcription and translation that naturally exists inside the cell will recognize the viral DNA, and produce any proteins that are coded by it. The DNA from the lambda phage produces a protein called **lambda integrase**, which has the effect of inserting the viral lambda DNA into the host's chromosomal DNA. The cell is now a carrier of the lambda virus, and all its descendents will inherit the new viral DNA as well as the original host DNA. Eventually, some external event will make the virus become active: using the host's translation and replication machinery, it will excise its DNA out of the host's, create the materials (DNA and coat proteins) for many new viruses, assemble them, and finally destroy the

cell's plasma membranes, releasing new lambda phage viruses to the unsuspecting outside world.

If DNA is the source code for a cell, then a lambda phage produces a sort of self-modifying program: not only is the central-dogma machinery of the cell appropriated to make new viruses, but the DNA that defines the cell itself is changed. This sort of self-modifying code is actually quite common, especially in eukaryotes, and the basic unit of such a change is called a **transposon.** There are many types of transposons—sections of DNA that use lambda-phage-like methods to move or copy themselves around the **genome**—and a large fraction of the human DNA consists of mutated, broken copies of transposons.

The **genome** is the "main" component of the genetic material for an organism—e.g., the chromosomal DNA for a eukaryote, or the nuclear DNA for a bacterium.

Even simpler than a virus is a **plasmid**, which is simply a loop of double-stranded DNA, much like the DNA inserted by a virus. Biologists have determined that there is nothing special about viral DNA that encourages the cell to use it: in particular, the machinery for DNA replication that naturally exists inside the cell will recognize a plasmid and duplicate it as well, as long as it contains, somewhere on the loop, the correct "instructions" for the replication machinery: for instance, one specific sequence of nucleotides called the **origin of replication** indicates where replication will start**.** Furthermore, the plasmid's DNA will also be transcribed to RNA and expressed, as long as it contains the proper **promoters**. In short, the DNA "program" in a plasmid will be "executed" by a cell, and the plasmid will be copied and inherited by children of a cell—just like the normal host DNA.

Promoters are DNA sequences that bind to the machinery that initiates the transcription of a gene. Without a valid promoter, a gene will not be expressed.

Plasmids are found naturally—they are especially common in prokaryotes. Like viruses, plasmids also occasionally migrate from cell to cell, allowing genetic material to pass from one bacterium to another. (This is one way in which resistance to antibiotics can be propagated from one species of bacteria to another, for instance.) There are also other plasmid-like structures that replicate in cells, but do not migrate from cell to cell

easily—for instance, some yeast cells contain a loop of RNA that apparently encodes just the proteins needed for it to replicate.

All complex living things are eukaryotes

The class of **eukaryotes** includes all multi-celled organisms, as well as many single-celled organisms, like amoebas, paramecia, and yeast. Every plant or animal that you have ever seen without a microscope is a eukaryote. Surprisingly, in spite of their diversity, eukaryotes are quite similar at the biochemical level—there are more biochemical similarities between different eukaryotes than between different prokaryotes, for example.

Eukaryotes are much larger and more complex than prokaryotes. The well-studied *E. coli,* for instance, is about 2 μm long, but a typical mammalian cell is 10-30 μm long, roughly 10-20 times the length of *E. coli;* this is about the same size ratio as an average-size man to a 60-foot sperm whale, or a hamster to a human. The figure below indicates the relative scale of some of the objects we have discussed so far.

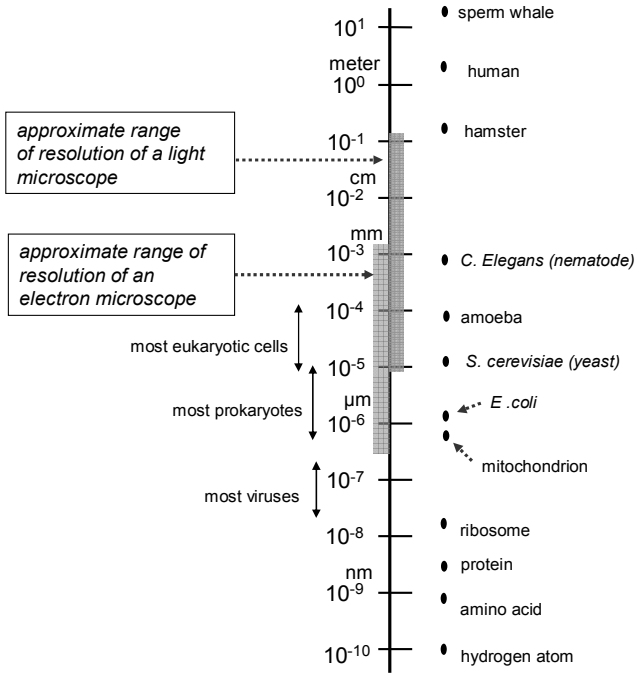


Figure 2. Relative sizes of various biological objects.

Unlike prokaryotes, eukaryotes have a complex internal organization, with many smaller subcompartments called **organelles.** For instance, the DNA is held in an internal **nucleus**, specialized compartments called **mitochondria** generate energy, the **endoplasmic reticulum** synthesizes most proteins, and long protein complexes called **microtubules** and **microfilaments** give shape and structure to the cell. Figure 3 illustrates some of the main components of a eukaryotic animal cell.

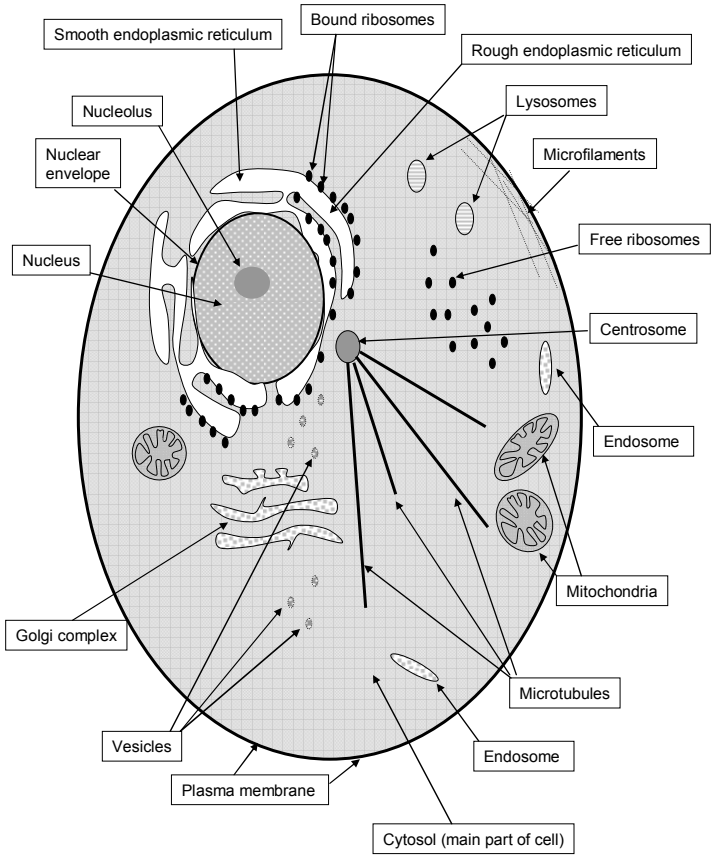


Figure 3. Internal organization of a eukaryotic animal cell.

Eukaryotes also use a more intricate scheme for storing their DNA "program." In prokaryotes, DNA is stored in what is essentially a single long loop. In eukaryotes, DNA is stored in complexes called **chromosomes,** wrapped around protein complexes called **nucleosomes.** The wrapping scheme that is used makes it possible to store DNA extremely compactly: for instance, if the DNA in a chromosome were about 1.5 cm long, the chromosome itself would be only about 2 μm long—four orders of magnitude shorter. Perhaps because of this ability to

compact DNA, eukaryotes tend to have much larger genomes than prokaryotes.

In addition to containing much more DNA than prokaryotes, eukaryotes also postprocess mRNA by a process called **splicing.** In splicing, some subsections of mRNA are removed before it is exported from the nucleus. Importantly, there can be multiple ways to splice the mRNA for a gene, so a single gene can produce many different proteins. This further increases the diversity of eukaryotes. Eukaryotes also have an additional set of mechanisms for regulating the expression of genes, because depending on its position relative to the nucleosomes, the DNA of a gene may or may not be accessible to the cell's transcription machinery.

The parts of a gene that are "spliced out" are called **introns.** The parts that are retained are called **exons.**

It is believed that some of the organelles inside eukaryotes evolved from smaller, independent organisms that began living inside the early proto-eukaryotes in a symbiotic relationship. For instance, mitochondria might have once been free-living bacteria. One strong piece of evidence for this theory is that mitochondria (and also **chloroplasts**, an organelle found in plants) have their own vestigial DNA, which uses a different code for translating DNA triplets into amino acids than the scheme used by any modern organism.

This theory of evolution is called **endosymbiosis.** A variety of modern endosymbionts exist, e.g., types of blue-green algae that live inside larger organisms. Some endosymbionts even contain a vestigial nucleus.

Cells cooperate

Humans, elephants, mushrooms, trout and oak trees are all eukaryotes. Interestingly, at the molecular level, the cells in multi-celled eukaryotes are in many ways very similar to single-celled organisms. The various cells that make up a multi-celled organism will share the same DNA, but are **differentiated**, meaning that they express a different set of genes: for instance, a kidney cell will express a different set of genes than a muscle cell.

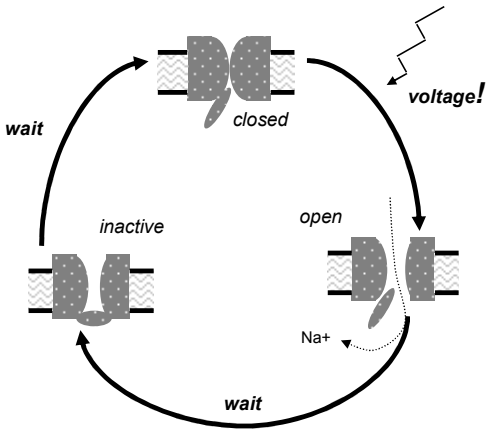
Cells in a multi-cellular organism also communicate, using a complex set of chemicals (mostly proteins) that are exchanged as **signals,** and received by **receptor sites** on the plasma membrane. Cells have many different

ways of sending, receiving and propagating signals. The most common types of receptors are **ion channels**, which allow small charged particles to pass through a membrane, and **G-protein coupled receptors** (which are discussed more below).

Neurons make use of ion channels to send messages from cell to cell, and also to propagate messages along a cell. Neurons have many branch-like protrusions called **dendrites** that receive signals. Outgoing signals pass through another protrusion called an **axon,** which can be several feet in length. To send a signal down an axon, a chain of **voltage-gated ion channels** are used—channels that open in response to a voltage signal. Opening an ion channel means that ions rush into the cell (since the ions are normally in a higher concentration outside the cell than inside it), which causes another voltage spike—a spike strong enough to cause nearby ion channels to open…which causes those channels to generate voltage spikes, and stimulate their neighboring channels, and so on. The process is somewhat like a "wave" at a football game, as is illustrated in Figure 5.

Of course, in order for the neuron to be ready to transmit the next signal, it is also necessary that the channels close again after the "wave" has passed by. One scheme for handling this is shown in Figure 4: shortly after a channel opens, it closes, and immediately after closing, the channel is **inactive**—i.e., unable to respond to voltage signals. The inactive phase keeps the wave moving in a single direction, but also requires ion-channel protein complexes to have some sort of short-term memory. Thus, ion channels are not simple holes in a membrane—they are quite complex molecular machines. Their shapes are also highly optimized to allow only certain ions through—the most common ones for signaling between cells being sodium (Na) and potassium (K).

After responding to a voltage signal of this sort, a neuron has absorbed many sodium ions. These are rapidly removed by special molecular complexes that "pump" unwanted ions out. The high concentration of ions outside the neuron that is produced by the pumps provides the energy needed to propagate the voltage signal.



A voltage-gated ion channel with three states: **closed**, which opens in response to voltage; **open**, which allows ions to pass through; and **inactive,** which blocks ions, and does not respond to voltage. The open and inactive states are temporary.

Figure 4. Voltage-gated ion channels in neurons.

Another type of ion channel is opened by the presence of a chemical called a **transmitter** rather than by voltage. **Transmitter-gated ion channels** are used to send signals from one neuron to another, as is shown Figure 6. Transmitter-gated ion channels are also common parts of the membranes inside cells: for instance, there are many channels that release calcium (Ca) ions from inside the endoplasmic reticulum—where it is found in abundance—into the cytoplasm. As in the re-uptake process of Figure 6, calcium-based signals require a means of removing "old" signaling material; hence, calcium-based signaling is often associated with the protein **calmodulin**, which binds readily to calcium.

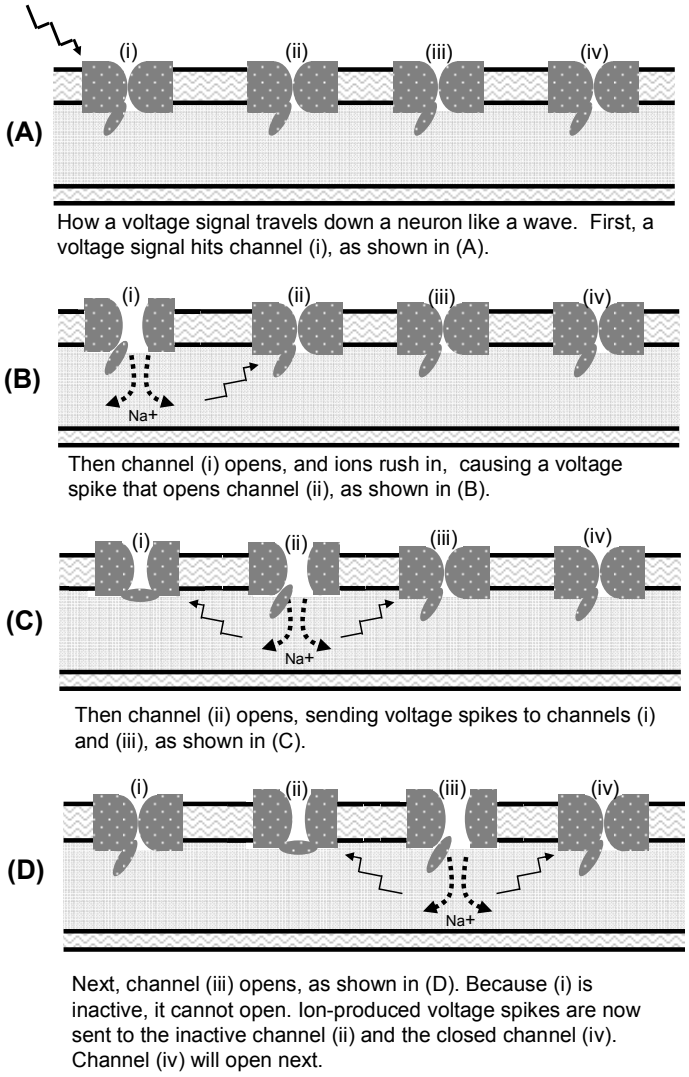
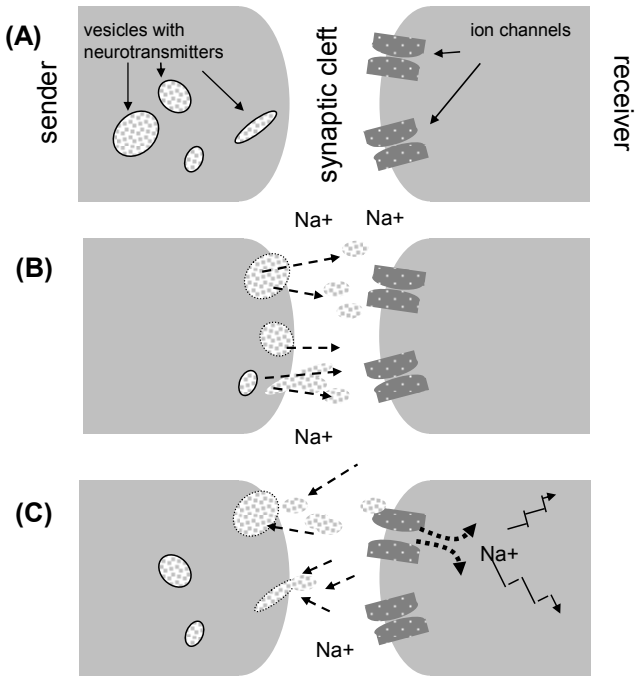


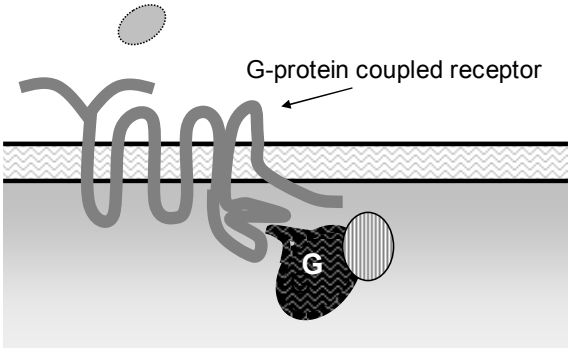
Figure 5. How signals propagate along a neuron.



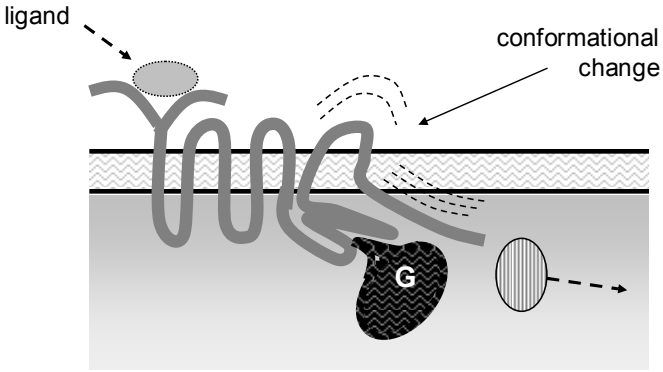
An example of a **transmitter-gated ion channel.** (A) shows the initial state. A substance used for signaling (for neurons, this is called a **neurotransmitter**) is held in vesicles by the sender cell. (B) In response to some internal change, the neurotransmitter is released. (C) Some of the neurotransmitter binds to ion channels on the receiver cell, and causes the channels to open. Most of the remainder of the neurotransmitter is re-absorbed by the sender cell, in a process called **re-uptake.**

A common neurotransmitter is **serotonin** (which is chemically related to the amino acid tryptophan). Many widely-used **antidepressants** (Prozac, Zoloft, and others) inhibit the reuptake step for serotonin, and are thus called **selective serotonin re-uptake inhibitors (SSRIs).** They cause serotonin to accumulate in the synaptic cleft, making it more likely that signals will propagate from cell to cell.

Figure 6. A transmitter-gated ion channel.



(A) A G-protein complex is bound to the G-protein coupled receptor on the inside of the cell. (There are many different types of G-proteins, and many types of receptors.)



(B) When the receptor binds to the **ligand** molecule, then the entire receptor changes shape. As a consequence, the G-protein complex is altered: part of it is released, to propagate the signal elsewhere in the cell.

Figure 7. A G-protein coupled receptor protein

Unlike ion channels, **G-protein coupled receptor proteins (GPCRs)** do not actually pass substances through a membrane. Instead, these receptors extend through the membrane on both sides. After the outside end of a **GPCR** binds to its target **ligand**, it changes **conformation** (i.e., shape) in such a way that a partner protein *inside* the membrane is affected. Typically, the partner **G protein** is actually a small collection of proteins bound together, some of which are released after the receptor detects the ligand. This process is shown in Figure 7.

A **ligand** is a molecule that binds to specific place on another molecule. The shape of a protein is called its **conformation.**

One important and well-studied example of such a receptor protein is **rhodopsin,** a protein found in our retina. Rhodopsin is somewhat atypical in that it responds to light, rather than a chemical stimulus.

Receptor proteins (and signaling pathways in general) are extremely important clinically, because they provide the easiest way for drugs to affect an organism. In general, cells make it difficult for outsiders to move chemicals across the plasma membrane; if you want to make them behave, it is often easiest to exploit the cell's "existing API" of signaling responses.

Cells divide and multiply

Cells also interact in another important way: by reproducing. The simplest way that cells reproduce is by division. In this process a cell will duplicate its DNA, separate the two copies of DNA, and then finally divide into two "daughter" cells, each with a copy of the parent cell's **genome.** In prokaryotes, this process is relatively simple: the DNA divides, each new strand attaches to a different place on the cell wall, and then the cell divides.

Cell division in eukaryotes is called **mitosis.**

Perhaps because the genetic material is organized into chromosomes, each of which must be duplicated and divided among the daughter cells, the process of division in eukaryotes is quite complex. Eukaryotic cells progress through a regular cycle of growth and division called the **cell cycle**, consisting of four phases: **S phase,** during which DNA is synthesized; **M phase,** during which the actual cell division (mitosis) occurs; and two gap phases**, G1** and **G2**, which fall between M&S and S&M respectively. The M phase consists of a number of

subphases: **prophase, prometaphase, metaphase, anaphase, telophase,** and **cytokinesis**, during each of which specific changes take place. (For instance, in metaphase, pairs of duplicate chromosomes are moved to the center of the nucleus.)

The cell cycle is orchestrated by a set of proteins called **cyclins** and **cyclin dependent kinases** (**Cdks**). The many actual movements that take place in mitosis are produced by "molecular motor" proteins that interact with the cell's microtubules**.**

A **kinase** is a protein that modifies another protein by adding a phosphate group. This process is called **phosphorylation**.

Like many things, this whole process becomes even more complicated when sex is involved. Organisms that reproduce sexually have two types of cells: **diploid** cells, which contain two copies of each chromosome, and **haploid** cells, which contain only one copy. Haploid cells are produced by a different type of cell division (called **meiosis)** which is illustrated below in Figure 8.

Only a single pair of chromosomes is shown in Figure 8, which simplifies the drawing. Unfortunately, considering a single pair of chromosomes also *overly* simplifies the process in an important way. Consider a diploid cell with N chromosome pairs: for convenience, call these pairs $(m_1, f_1),\ldots(m_N, f_N)$. Meiosis will produce four haploid cells, each of which contains either m_1 or f_1, either m_2 or f_2, and so on; thus there are 2^N possible haploid daughter cells. The huge number of possible ways in which chromosomes can be divvied up during meiosis is reason why eukaryotic species, like ourselves, can be genetically diverse.

In fact, the number of possible haploids is much larger than this, due to **genetic recombination**, a process in which segments of DNA are "swapped" between chromosomes. As shown in Figure 8D, this typically occurs when bivalents are formed. These swaps, or **crossover events**, happen on average 2-3 times on each pair of human chromosomes.

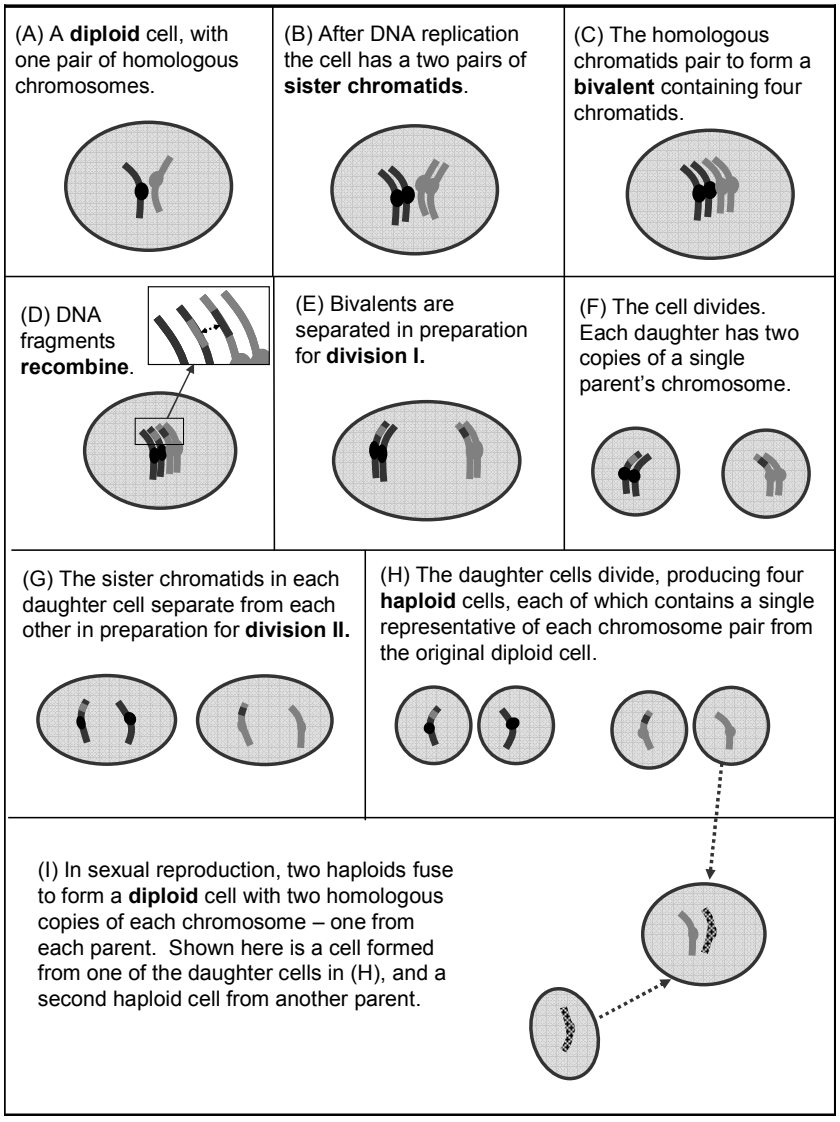


Figure 8. Meiosis produces haploid cells.

Diploid cells are more complex to study, if your goal is to understand which genes cause which effects, because the two copies of each gene need not be exact copies: instead, there can be slightly different DNA sequences that produce similar gene products. The variant sequences are said to be different **alleles** of the gene. Often, only one of the alleles (the **dominant allele**) will be expressed, and the other **recessive allele** will be "hidden" (in the sense that its effects are masked.)

An organism with two copies of the same allele for a gene is **homozygous** for that gene. An organism with two different alleles for a gene is **heterozygous** for the gene.

In humans, there are only two types of haploid cells: egg cells and sperm cells. All other cells are diploid. A popular organism for genetic studies is **yeast**, a single-celled eukaryote that can grow and reproduce as a haploid, but can also reproduce sexually. There are no male or female yeast: instead the "sexes" for yeast are called **type a,** and **type α.** When yeast cells "want" to mate, they release a chemical called a **mating factor** (which, by the way, is detected by a type of G-protein coupled receptor). Yeast cells are not always receptive to mating signals—for instance, when there is plenty of food in the environment, they often "prefer" to eat. Sometimes, however, when a "Greek" type-α yeast cell detects a mating factor from a "Roman" type-a cell, it will start building a protuberance called a "schmoo tip"—a name derived from the classic "Lil' Abner" cartoons by Al Capp. Eventually the "schmoo tips" of the parent cells grow together and the cells can fuse and mate, producing a diploid child.

Prokaryotes do not undergo meiosis, but they can exchange genetic material via plasmids. One special type of plasmid, called a **fertility plasmid** or **F-plasmid**, contains genes that enable an *E. coli* to initiate a process called **conjugation**. Bacteria containing the F-plasmid are called "male," and have the ability to construct a long tubular organelle called a sex pilus, which is used (you'll be relieved to read) as a sort of a grappling hook to grab another *E. coli* and bring it in close. The organisms then form a "conjugate bridge" and exchange genetic material—including the F-plasmid itself. Mating usually involves groups of 5-10 bacteria, and in the kinky world of the *E. coli,* all of them become "male" after conjugation, by virtue of their newly-received F-plasmid.

The Complexity of Living Things

Complexes and pathways

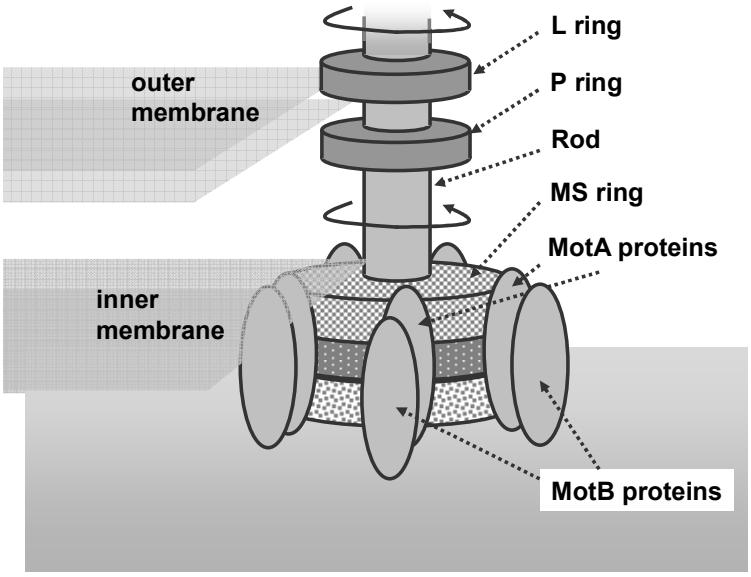
Although the basic mechanisms that underlie cellular biology are surprisingly few, there are many instances and many variations on these mechanisms, leading to an ocean of detail concerning (for instance) how the process of microtubule attachment to a centrosome differs across different species. Cellular-level systems, because they are so small, are also difficult to observe directly, which means that obtaining this detail experimentally is a long and arduous process, often involving tying together many pieces of indirect evidence. Most importantly, cellular biology is hard to understand because living things are extremely complex—in several different respects.

A **flagellum** is a whip-like appendage that certain bacteria have. It functions as a sort of propeller to help them move. An *E.coli* flagellum rotates at 100Hz, allowing the *E.coli* to cover 35 times its own diameter in a second.

One source of complexity is the sheer number of objects that exist in a cell. At the molecular level of detail, there are thousands of different proteins in even the simplest one-celled organisms. These individual proteins can themselves be quite large, and assemblies of multiple proteins (appropriately called **protein complexes**) can be extremely intricate. One notable example for bacteria is the "molecular motor" which spins the **flagellum**—an assembly of dozens of copies of some twenty distinct proteins that functions as a highly efficient rotary motor. (See Figure 9.) This motor is atypical in some ways—most protein complexes are less well-understood, and do not resemble familiar mechanical devices like turbines—but it is far from unrivaled in its size or in the number of protein components. (Ribosomes, for instance, are much larger.) Unraveling this type of complexity is part of the discipline of biochemistry.

A second type of complexity associated with living things are the complex ways in which proteins interact with each other, with the environment, and with the "central dogma" processes that lead to the production of other proteins. A *simplified* illustration of one of the best-studied such processes is shown in Figure 10, which illustrates how *E. coli* "turns on" the genes

that are necessary to import lactose when its preferred nutrient, glucose, is not present. Briefly, the gene *lacZ* is regulated by two proteins (called *CAP* and the *lac repressor protein*), which function by binding to the DNA near the site of the *lacZ* gene, and a feedback loop involving lactose and glucose affect the relative quantities of CAP and the lac repressor protein; however, as the figure shows, the details of this feedback process are nontrivial.



Structure of a bacterial flagellum (simplified). About 40 different proteins form this complex. The MS ring is made up of about 30 FliG subunits, and about 11 MotA/MotB protein pairs surround the MS ring. It is believed that these pairs, together with FliG, form an ion channel. As ions pass through the channel, conformational changes cause the MS ring to rotate, much like a waterwheel.

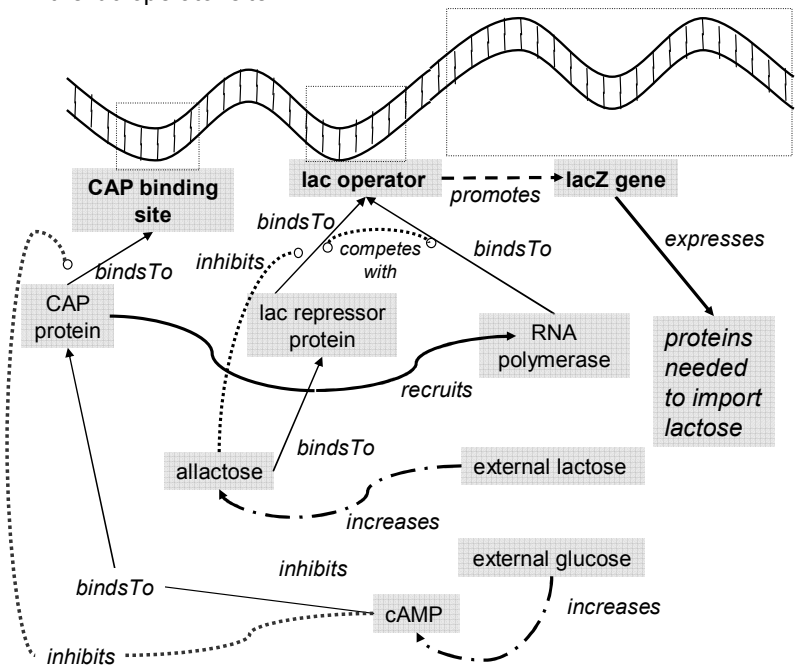
A similar "molecular motor" is used in ATP synthesis in a mitochondrion: rotation, driven by ions flowing through a channel, is the energy used to convert ADP to ATP. (See the section below, "Energy and Pathways").

Figure 9. The bacterial flagellum.

Many cell processes involve this sort of "interaction complexity," and often the interactions are far from being completely deciphered, let alone

understood. Like the molecular motor that drives the flagellum, the chemical interactions in a cell have been optimized over billions of years of evolution, and like any highly-optimized process, they are extremely difficult to comprehend.

The *lacZ* gene is transcribed only when CAP binds to the CAP binding site, and when the *lac* repressor protein does not bind to the *lac* operator site.



This network presents simplified view of why *E.coli* produces lactose-importing proteins only when lactose is present, and glucose is not.

Figure 10. How *E. coli* responds to nutrients

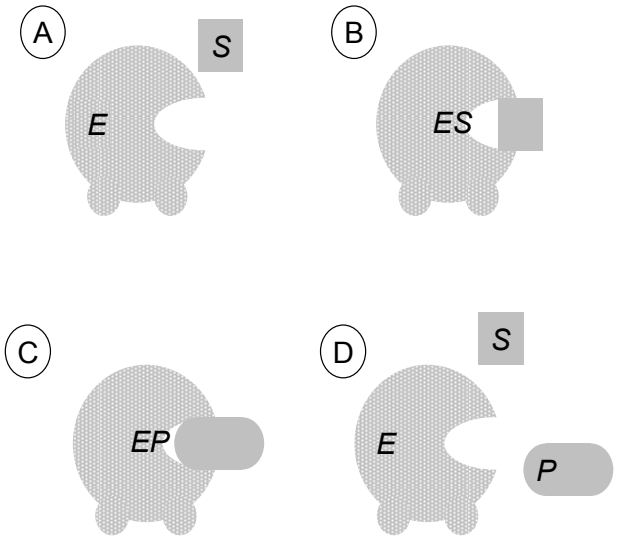
Individual interactions can be complicated

Networks of chemical interactions like the one shown in Figure 10 are also complex in a different respect: not only is there a complex network that defines the *qualitative* interactions that take place, the individual interactions can be *quantitatively* complex. To take an example, increases in glucose *might* increase the quantity of cAMP linearly—but often there will be complex non-linear relationships between the parts of a biological chemical pathway.

The reason for this is that most biological reactions are mediated by **enzymes**—proteins that encourage a chemical change, without participating in that change. Figure 11 gives a "cartoon" illustrating how an enzyme might encourage or **catalyze** a simple change, in which molecule *S* is modified to form a new molecule *P*. It is also common for enzymes to catalyze reactions in which two molecules *S* and *T* combine to form a new product.

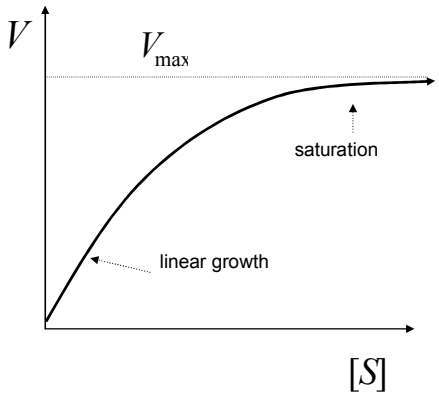
Enzymes can accelerate the rate of a chemical reaction by up to three orders of magnitude, so it is not a bad approximation to assume that a change (like S → P above) can only occur when an enzyme E is present. This means that if you assume a fixed amount of enzyme E and plot the rate of the chemical reaction (let's call this "velocity," V) against the amount of the substrate S (and like chemists, let's write the amount of S as [S]), the result will be the curve shown below. Velocity V will increase until the enzyme molecules are all being used at maximum speed, and then flatten out, as shown in Figure 12.

This model is due to Michaelis and Menten and is called "saturation kinetics." In fact, the shape of the curve shown is quite easy to derive from basic probability and a few additional assumptions—the ambitious reader can look at the mathematics in Figure 13 and Figure 14 to see this.



A cartoon showing how an enzyme catalyzes a change from *S* to *P*. (A) Initially, the enzyme *E* and "substrate" *S* are separate. (B) They then collide, and bind to form a "complex" *ES*. (C) While bound to *E*, forces on the substrate *S* cause it to change to form the "product" *P*. (D). The product is released, and the enzyme is ready to interact with another substrate molecule *S*. A chemist would summarize this as: E+S→ ES → EP → E+P

Figure 11. How enzymes work.



Reaction velocity with a fixed quantity of an enzyme *E*, and varying amounts of substrate *S*. When little substrate is present, an enzyme *E* to catalyze the reaction is quickly found, so reaction velocity *V* grows linearly in substrate quantity [*S*]. For large amounts of substrate, availability of enzymes *E* becomes a bottleneck.

Figure 12. Saturation kinetics for enzymes.

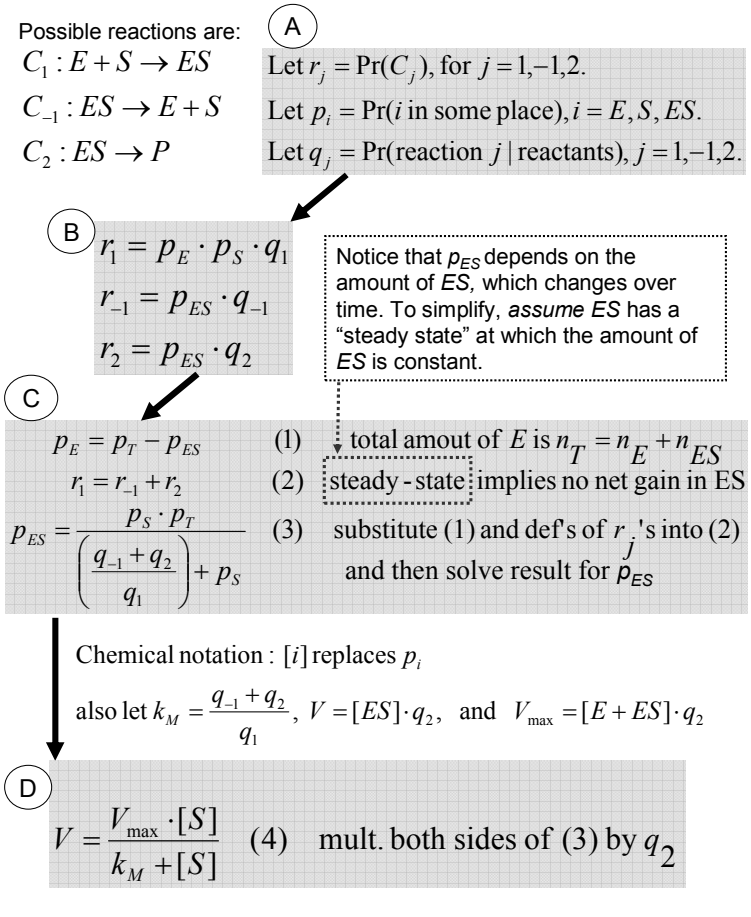
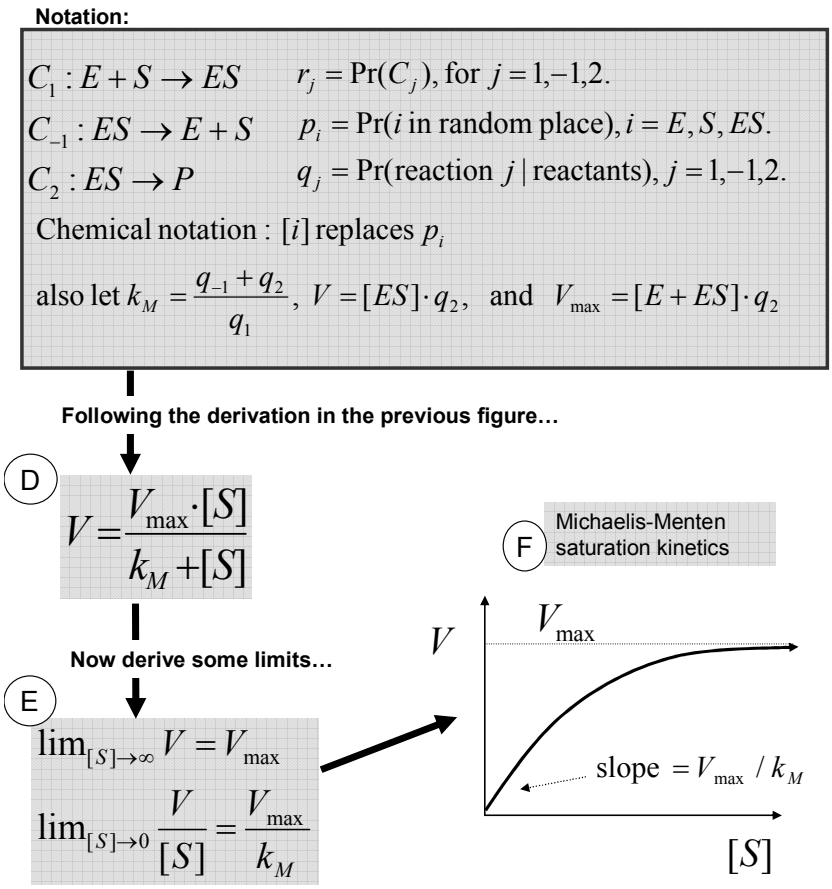


Figure 13. Derivation of Michaelis-Menten saturation kinetics.



The first limit shows that *V*, the velocity at which *P* is produced, will asymptote at V_{max}.

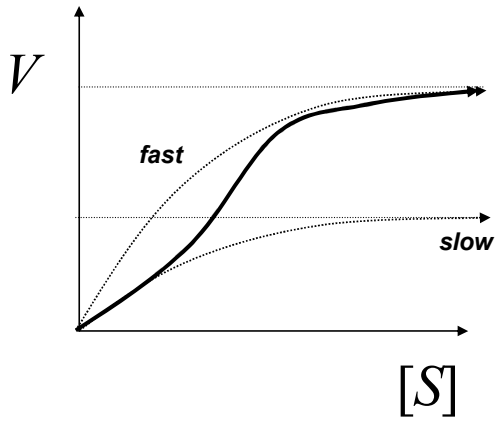
The second limit shows that for small concentrations of S, the velocity V will grow linearly with [*S*], at a rate of $V_{max}/k_{M.}$

Figure 14. Interpreting Michaelis-Menten saturation kinetics.

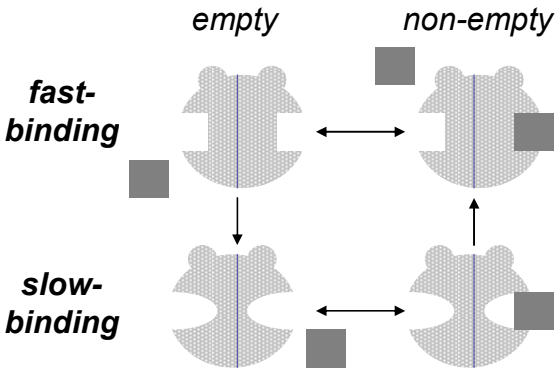
Enzymes with more complicated structures can lead to more complicated velocity-concentration curves, as shown in Figure 15. A typical example would be an enzyme with two parts, each of which has an **active site** (a location at which the substrate *S* can bind), and each of which has two possible **conformations** or shapes. One conformation is a fast-binding shape, which has a high maximum velocity $V_{maxFast}$, and the other is a slower-binding shape with maximum velocity $V_{maxSlow.}$ The lower part of the figure shows a simple state diagram, in which: (a) both parts of the enzyme change conformation at the same time, (b) shifts from the slow to fast conformation happen more frequently when the enzyme is binding the substrate, and (c) shifts from fast to slow tend to happen when the enzyme is "empty," i.e., not binding any substrate molecule. In this case, as substrate concentration increases, the enzymes in a solution will gradually shift conformation from slow-binding to fast-binding states, and the actual velocity-concentration plot will gradually shift from one saturation curve to another, producing a **sigmoid (**i.e., S-shaped) curve—shown in the top of the figure. A sigmoid is a smooth approximation of a step-function, which means that enzymes can act to switch activities on quite quickly.

A molecule that is composed of two identical subunits is a **dimer**; three identical subunits compose a **trimer;** and *N* identical subunits compose a **polymer.** An enzyme in which binding sites do not behave independently is an **allosteric** enzyme; in the example here, the enzyme exhibits **cooperative binding.**

Sigmoid curves and network structures are also familiar in computer science, and especially in machine learning: they are commonly used to define **neural networks.** A neural network is simply a directed graph in which the "activation level" of each node is a sigmoid function of the sum of the activation levels of all its input (i.e., parent) nodes. It is well-known that neural networks are very expressive computationally: for instance, finite-depth neural networks can compute any continuous function, and also any Boolean function. Although I am not familiar with any formal results showing this, it seems quite likely that protein-protein interaction networks governed by enzymatic reactions are also computationally expressive—most likely Turing-complete, in the case of feedback loops. This is another source of complexity in the study of living things.



Allosteric enzymes switch from a slow-binding state to a fast-binding state, and tend to remain in the fast-binding state when the substrate *S* is common. Their kinetics follows a sigmoid curve.



A typical allosteric enzyme: when one half is being used, the whole molecule tends to shift to the fast-binding state.

Figure 15. An enzyme with a sigmoidal concentration-velocity curve.

Energy and pathways

More properly, ATP is combined with water to produce ADP plus inorganic phosphate, yielding energy: ATP+H_20 → ADP + Pi. This reaction is called **hydrolysis.**

Enzymes are important in another way. Running the machinery of the cell requires energy. Most of this energy is stored by pushing certain molecules into a high-energy state. The most common of these "fuel" molecules is **adenosine,** which can be found in two forms in the cell: **adenosine triphosphate (ATP),** the higher-energy form, and **adenosine diphosphate (ADP),** the lower-energy form. Enzymes are the means by which this energy is harnessed. Usually this is done by coupling some reaction P→Q that *requires* energy with a reaction like ATP→ADP, which releases energy. If you visualize the potential energy in a molecule as vertical position, you might think of this sort of enzyme as a sort of see-saw, in which one molecule's energy is increased, and another's is decreased, as in the figure below. (Dotted lines around a shape indicate a high-energy form of a molecule.)

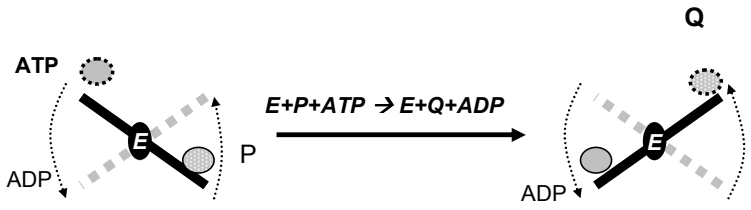
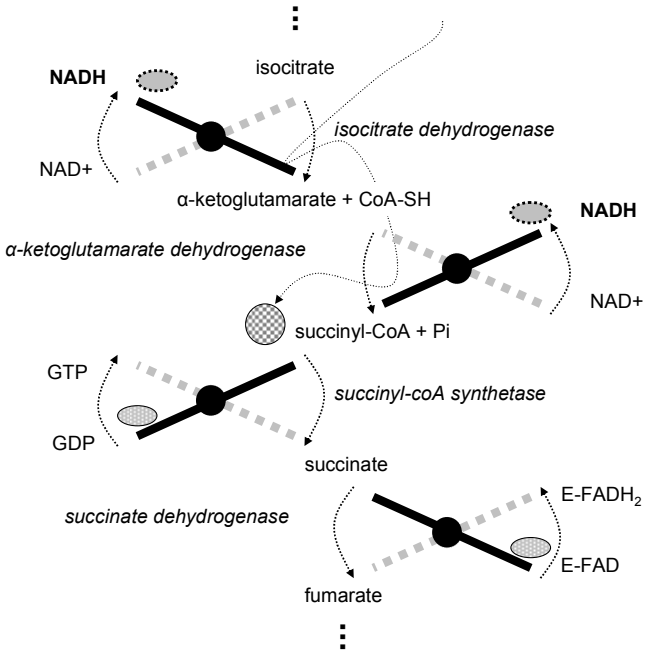


Figure 16. A coupled reaction.

Cellular operations that require or produce energy will often use an **enzymatic pathway**—a sequence of enzyme-catalyzed reactions, in which the output of one step becomes the input of the next. One well-known example of such a pathway is the TCA cycle, which is part of the machinery by which oxygen and sugar is converted into energy and carbon dioxide. A small part of this pathway is shown below in Figure 17. (Notice that this particular pathway produces energy, rather than consuming energy).



Part of the TCA cycle (also called the citric acid cycle or the Krebs cycle) in action. A high-energy molecule of isocitrate has been converted to a lower-energy molecule called α-ketoglutamarate and then to a still lower-energy molecule, succinyl-CoA (as shown by the path taken by the hashed circle). In the process two low-energy NAD+ molecules have been converted to high-energy NADH molecules. Each "see-saw" is an enzyme (named in italics) that couples the two reactions. The next steps in the cycle will convert the succinyl-CoA to succinate and then fumarate, producing two more high-energy molecules, GTP and $E\text{-}FADH_2$.

Figure 17. Part of an energy-producing pathway.

Since each intermediate chemical in the pathway (e.g., fumarate, succinate, etc) is different, each enzyme is also different: thus a pathway that either consumes or produces large amounts of energy will often involve many different enzymes, again contributing to complexity.

Amplification and pathways

Sometimes a pathway will act to amplify a weak initial signal. A good example of this is the pathway associated with **rhodopsin**. Rhodopsin is a G-linked protein receptor that detects light. Each rhodopsin protein cradles a "chromophore" molecule called **11-cis-retinal.** When a photon is absorbed by the 11-cis-retinal molecule, it changes shape, which causes rhodopsin to change shape and become "active." "Active" rhodopsin can then "activate" a second protein called **transducin.** Transducin, in turn, "activates" a third protein called **cGMP phosphodiesterase (PDE)**, an enzyme that hydrolyses a somewhat ATP-like molecule called **cyclic guanine monophosphate** (**cGMP**). In the **rod** and **cone** cells in the retina—the cells which sense light—cGMP acts somewhat like a chemical doorstop, propping open certain ion channels. When the concentration of cGMP is reduced, these ion channels close, changing the electrical charge of the cell and finally leading to a voltage signal. The process is thus something like this, where **R** is rhodopsin, **T** is transducin, and a ***** denotes the active form:

The "fuel" used in a cell is chemically related to the bases of DNA and RNA. There are four **nucleobases** (aka **bases**) that form DNA: **adenosine, thymine, cytosine, and guanine,** abbreviated A, T, C, and G. (In RNA **uracil** replaces **thymine.)** A **nucleoside** is a base attached to a sugar: either **ribose** (for RNA) or **deoxyribose** (for DNA). A **nucleotide** is a nucleoside attached to a phosphate group: either mono-, di-, or triphosphate. These are abbreviated with 3- and 4-letter codes: e.g., ATP is adenosine triphosphate, and cAMP is cyclic adenosine monophosphate.

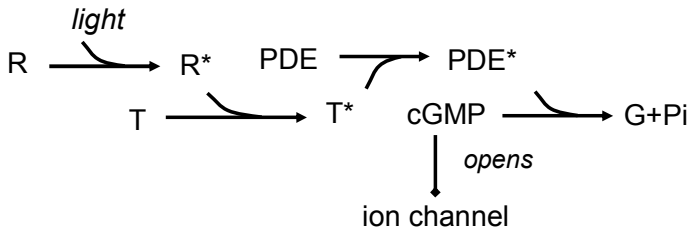


Figure 18. How light is detected by rhodopsin.

Acknowledgements

I would like to thank Susan Cohen, for indexing the book, and encouraging me to write it; Dan Kundin, for proofreading a late version of the book; Eric Xing, for comments on an earlier version; and the National Institutes of Health, for supporting this work under NIH Grant DA017357-01.

Index