Life and Its Molecules
A Brief Introduction

Lawrence Hunter

■ To appreciate bioinformatics, it is necessary to understand some of the basic concepts and terminology of molecular biology. This article is a brief introduction to the extraordinarily complex phenomenon of life and to its molecular basis. We begin with the amazing diversity of life forms and the equally amazing unity in the molecules underlying life's processes. The challenge of accounting for both the variety and the commonalities among organisms is met by evolutionary theory; despite controversies, all scientific approaches to understanding life build on a shared core that can briefly be stated. One of the great insights of the last generation of biologists was the chemical instantiation of these evolutionary theories, whose discovery has driven biology toward the study of the structure and function of biological molecules. After an introduction to some of these key molecules and to the central dogma of molecular biology, we can begin to see the outlines of how such molecules can accomplish the tasks required of simple and then more complex life forms. The introduction concludes with a brief account of some of the new instruments and model systems that are now so rapidly advancing scientific understanding of life.

Life is an extraordinarily complex phenomenon. Although the study of living things dates at least as far back as Aristotle (ca. 300 BCE), the advent of tools that allow the interrogation of living systems in molecular detail and genomic breadth makes this a particularly exciting era in the history of biology. The purpose of this introduction is to help you begin to understand and appreciate our growing understanding of what living things are, what they do, and how they do it.

It is perhaps the holistic nature of the subject matter that makes creating an accessible introduction to biology so difficult. Understanding any aspect of living things can seem to require understanding of dozens of other aspects. There is no easy place to begin, no simple set of problems that can be grasped in isolation as a prelude to deeper understanding. The study of life is really many studies: evolution, biochemistry, genetics, pathology,[1] and ecosystems, just to name a few. The purpose of a brief introduction such as this one is to impart enough knowledge about enough different aspects of life to provide a foundation for more detailed understanding of the particulars relevant to the other articles in this issue.

A useful metaphor to keep in mind is that learning biology is akin to learning a foreign language. First, there is an extensive specialized vocabulary that biologists use to characterize living systems and their properties. To understand the biological literature, one must learn these terms and how they are used. This article introduces many such terms, using italics to set them off. As might become clear in this introduction, language is also a useful metaphor for understanding the structure and function of biological systems at the molecular level. The "book of life" is an apt and useful idea. Learning a foreign language involves more than just learning words. Languages are an intimate part of cultures; so, for example, learning French generally involves learning something about French culture as well as French words—likewise in biology. Biologists approach scientific problems somewhat differently than physicists, chemists, and other colleagues.

One of the major differences between biology and other physical sciences is the central role of detailed descriptions of the phenomena under study compared to general theoretical constructs. The English physicist Ernest Rutherford once dismissed biology as mere "stamp collecting," poking fun at this aspect of the science. Although it is true that a central aspect of biological science aims to create detailed (and accurate) descriptions of living things and their activities, a better metaphor than stamps might

be collecting biographies. Although stamps do vary, their variations are quite constrained, and most variations are not particularly tied to function. However, variation in living things, like variation in human life stories, is extraordinarily broad and so central to what it is to be alive that it is itself a phenomenon worthy of study. The many details of the complex story of an organism play a synergistic role in understanding it; reducing these details to a simpler characterization runs the risk of caricature. Let us therefore turn first to examining the diversity of life stories that make up the subject matter of biology.

Diversity

One of the most clearly distinguishing features of life as a whole is its diversity. Organisms exhibit an overwhelming collection of differences. Most people are familiar with only a tiny fraction of the kinds of life on earth, and even that small sample includes enormous variation. There are more than a million known species, and estimates of the number yet to be characterized range from 10 to 100 million additional species.

One of the most clearly distinguishing features of life as a whole is its diversity.

Consider the differences among just a few organisms, say, mayflies, grizzly bears, tortoises, dinosaurs, earthworms, guppies, and eagles. Some are huge, others tiny. They exhibit tremendous differences in how they feed, how they reproduce, what environments they can survive in, how long they live, what their sensory and motor abilities are, and so on. There are organisms whose home environments are so remote from our intuitions about what is hospitable to life that they are called *extremophiles,* for example, creatures that live in volcanic vents on the deep ocean floor or in acids so strong they can dissolve most familiar materials immediately.

Whole species differ greatly from one another, but there are also large variations among individuals within a single species. Even within a single individual, there can exist an amazing diversity of organs, tissues, and other components. This diversity in the activities and constituents of living things continues all the way down to the molecular level, where even in the simplest organisms many thousands of molecules interact with each other in as yet uncountable ways. As one becomes more familiar with the details of species, individuals, tissues, and molecules, life's diversity becomes even more striking.

An important part of understanding biology is developing at least a moderately detailed appreciation for the many species of living things and their relationships with each other; this study is called *taxonomy*. In many respects, microscopic life is considerably more varied than the life forms we can see. Many of these life forms consist of a single *cell,* which is the fundamental unit of life. (An adult human being contains more than a trillion cells.) The diversity of microscopic species in a cubic meter of seawater can rival that of macroscopic species in a cubic kilometer of rainforest. Because most microbes do not grow well in a laboratory, it has only been with the advent of molecular technology that the extent of microbial diversity has become apparent.

It is also worth noting that human beings have been present for only a vanishingly small portion of the history of life, at most the past million years or so. In comparison, dinosaurs roamed the earth for more than 160 million years. Peering deep into fossil history shows us many organisms that thrived for millions of years but are like nothing alive today, such as the five-eyed, vacuum-nose Opabinia from the Cambrian era.

One of the major challenges of biology as a science is to account for this diversity. How did it arise? How is it maintained? Why is it this particular set of diverse entities and processes and not some other?

Unity

Given the extraordinary diversity, one of the most surprising discoveries in the history of biology is near universality of the molecular detail underlying all living things. The instruments and experimental approaches necessary to even perceive anything molecular at all are a relatively recent phenomenon, and it is on the basis of that newfound ability to investigate life at a molecular level that so much of the recent progress and excitement has come.

All living things ever encountered depend crucially on the activities of the unusual and complex family of molecules called *proteins*. There are hundreds of thousands of different kinds of proteins, and they work together in large groups to carry out almost every biological function. Two rather extreme examples of proteins include *hemoglobin*, which carries oxygen in the blood, and *anthrax toxin,* a lethal poison created by a microbe. As described in more detail later, proteins are the entities responsible for the near miracles of chemistry required to turn food into bodies and offspring. The proteins that accomplish a particular function in one organism are generally quite similar to the proteins that do similar functions in many other organisms. The unity among organisms is not

merely that proteins generally do most of the biochemical work required for life but that very similar sets of proteins doing very similar kinds of things are found in extraordinarily diverse organisms. Many of the proteins in human beings are remarkably similar in structure and function to those found in, say, brewer's yeast!

The ubiquity of proteins is not the only remarkable unity among organisms. All living things make important use of another unusual and complex family of molecules, the nucleic acids. There are two distinct kinds of nucleic acids—(1) deoxyribonucleic acid (DNA) and (2) ribonucleic acid (RNA)—which play somewhat different but related roles as the information carriers of life.

Together, the nucleic acids and the proteins are called biological *macromolecules,* based on their large size compared to most inorganic molecules. If one could stretch it out, a single DNA molecule can be more than a meter long (although only a few hundred angstroms wide)! Both proteins and nucleic acids are linear *polymers,* which are molecules made up of long strings of just a few basic components. The components of proteins are called *amino acids,* and the components of nucleic acids are *nucleotides*. It is the particular relationships among components that give an individual macromolecule its distinguishing characteristics. The specific order of components is called the *sequence* of the macromolecule. Macromolecular sequences can be thought of as strings of "letters" that form the words, sentences, chapters, and books of living things.

In contrast to the macromolecules, all the other many molecules in the world relevant to living things (such as water, sugars, fats, and drugs) are often called *small molecules*. The study of the actions of biomolecules large and small is generally termed *biochemistry.* Both scientific understanding of the structure and function of macromolecules and instrumentation engineering advances in the ability to investigate the details of particular members of these families are a crucial driving force in the expanding understanding of life, hence the term *molecular* biology (figure 1).

Evolution

There is another unity among all forms of life that is even more important than the molecular one: *evolution*. Evolution is, without a doubt, the most important concept in biology, and it was discovered long before biological molecules were even conceived. Although biological evolution is itself a complex topic, the basic idea is again simple: All organisms are part of a continuous line of ancestors and descendants. This is the only statement in biology to which there is no exception.

There are some very important consequences to this statement. Every creature that ever existed on earth is related (however distantly) to every other creature. If you go back far enough, every pair of organisms shares a common ancestor. Not only are humans related to (that is, share a common ancestor with) chimpanzees, we are relatives to dinosaurs and even bacteria! There is, in fact, a "universal ancestor" that is the great-great-great ... great-grandparent of every organism on the planet.

The existence of common ancestors is an important part of the explanation of the similarities we see within families of organisms. For example, because the use of nucleic acids to code for proteins is universal throughout life, evolution suggests that the most recent universal ancestor must have done the same thing. Other similarities among smaller sets of organisms, such as bilateral symmetry in body shapes or the presence of oxygen-carrying hemoglobin in circulating blood, are usually shared by virtue of their inheritance from a common ancestor.

Evolutionary relatedness leads to a straightforward explanation of similarities among organisms, analogous to the observation that offspring are similar to their parents. The more difficult challenge is to balance an explanation of our similarities with an explanation of the diversity. Its success in this challenge is what has made evolution so central to our understanding of life.

Evolution is a controversial topic with several competing theories, but the overall structure of all the competitors involves three basic phenomena. First, evolution requires *self-replication*. That is, entities that evolve must make copies of themselves. The entity that does the replicating is the *parent,* and the resulting new entity is the *offspring*. There is a lot of subtlety hidden in the word *copy* in that definition. Simply stated, offspring must share at least some of the characteristics of the parent; this sharing is called *inheritance*. Inheritable characteristics are called *traits*. Inheritance is one of the forces that drives life toward unity.

The second requirement for evolution is a source of *variation*—if offspring were all perfect replicates of parents, there would be no evolution. There are many sources of inheritable variation in biology. Some of these sources of variation are random, such as mutation; others are systematic, such as the mix of inheritance from multiple parents in sexual reproduction. Variation is one of the forces that drives life toward diversity.

Evolution is, without a doubt, the most important concept in biology, and it was discovered long before biological molecules were even conceived.

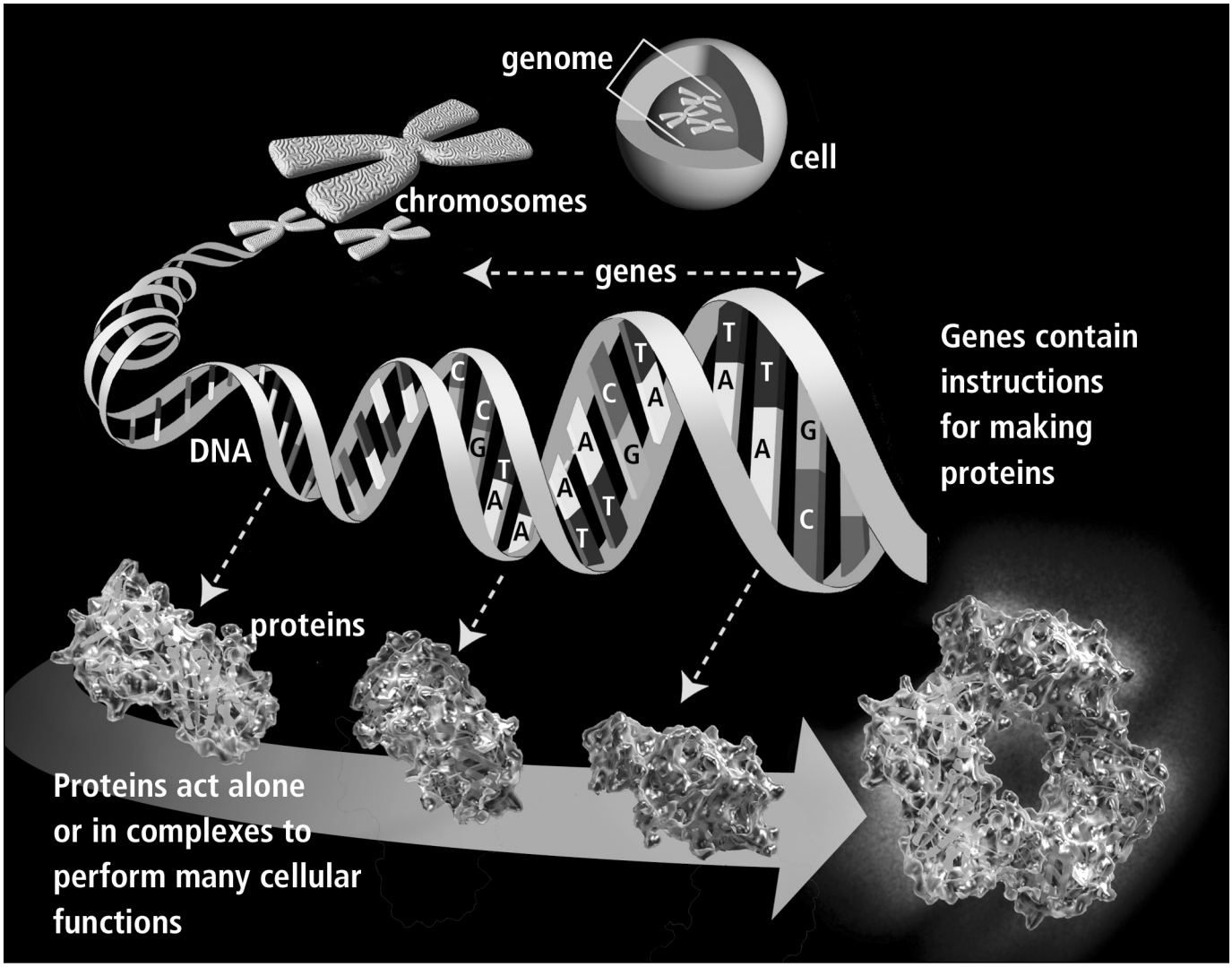

Figure 1. From the Cell to Protein Machines.

Cells are the fundamental working units of every living system. All the instructions needed to direct their activities are contained within the chemical deoxyribonucleic acid (DNA).

Although genes get a lot of attention, it's the proteins that perform most life functions and even make up the majority of cellular structures. Proteins are large, complex molecules made up of smaller subunits called *amino acids*. Chemical properties that distinguish the 20 different amino acids cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell. *Figure courtesy, U.S. Department of Energy Human Genome (www.ornl.gov/hgmis).*

However, the variation we see in life is clearly not wholly random. The final key in understanding evolution is the idea of *selection.* Not every organism gets to reproduce. Selection is the process by which some organisms have offspring, and others don't. There are many aspects to selection. Although some aspects of reproductive success are random, others are related to the traits of the organism. If we look at groups of closely related organisms (called *populations*) rather than individuals, it is possible to demonstrate that variation in traits that are positively related to reproductive success will tend to become common in the population over time, and those variations that negatively effect reproduction will tend to disappear.

The relationship between the particular set of traits of an organism and its reproductive success is termed *fitness*. Charles Darwin's famous dictum, "survival of the fittest," emphasizes the high stakes and inherent competition involved in differential reproductive success. Of course, fitness is a very complex function, which depends on many things, including the environment in which the organism lives and

even other organisms in that environment. Variations that improve fitness in a particular environment are called *adaptations* to that environment.

The combination of inheritance, variation, and selection is the essence of evolution, the force that created both the unity and the diversity of living things that we observe today. Biology's first understanding of the basic mechanisms of evolution arose from Gregor Mendel's study of inheritance of characteristics in sexually reproducing plants. Offspring had long been thought to be blends of the traits of the parents. However, if this were true, all variation would quickly disappear. Darwin noted this paradox but never resolved it. Mendel's answer was that traits were particulate; traits do not blend but are inherited (or not) as a unit. Evolutionarily speaking, a *gene* is the particle of inheritance, or the smallest inheritable unit. An organism can be said to have a *genotype,* that is, the complete set of genes that were inherited from its parents.

It is important to remember that not all the characteristics of an organism (that together are called its *phenotype*) are determined by inheritance. Organisms with precisely the same genotype can end up with quite different phenotypes, like the differences between twins reared apart. One of the basic concepts of biology is that genotype interacts with environment to determine phenotype.

The genotype of one individual can be different from that of another individual, even of the same species. Alternative forms of the same gene are called *alleles*. For example, the *color of a flower* might be determined by a single gene, and the phenotype *pink* might arise from a particular allele of the flower color gene. Organisms of a particular species all have the same genes but have different alleles.

Mendel proposed that each sexually reproducing organism has two alleles for each gene, one from each parent. If the two alleles are the same, the phenotype reflects it. These organisms are called *homozygotic* for that allele. If the two alleles are different, the organism is called heterozygotic, and the phenotype reflects the *dominant* allele. The allele that is not dominant is called *recessive*. Recessive alleles are reflected in the phenotype only when they are homozygotic.

Although nearly lost to history (Mendel worked before Darwin), this theory has held up through the transition from a purely evolutionary definition of a gene to the contemporary chemical one. Inheritance of genes in sexually reproducing organisms is determined by Mendel's laws. However, only characteristics that are *monogenic* (related to a single gene) display Mendelian inheritance at the phenotypic level. All continuously varying quantitative traits (such as height) must be *polygenic* (involving several genes), which can make inheritance appear to be a blending rather than an all or nothing phenomenon. Most medically important traits (such as proclivity to cancer or heart disease) are polygenic; that is, they involve multiple genes. That is why claims about having found **the** gene for breast cancer (or intelligence or any other complex phenomena) are generally journalistic oversimplifications.

The Central Dogma

Biological macromolecules have many remarkable properties, and their study is the essence of molecular biology. The most central of these, the one that was discovered by James Watson and Francis Crick, is that the DNA molecule is the carrier of the gene. The relationship between genes and phenotype is that nucleotide sequences in a DNA molecule code for the amino acid sequences of proteins. DNA coding for protein is the biochemical basis for the connection between genes and phenotype (a bit more biochemistry is necessary to fully appreciate that statement but keep reading!).

Biological macromolecules have many remarkable properties, and their study is the essence of molecular biology.

The specific relationship between nucleic acids and proteins is so important to modern biology that it is called the *central dogma.* The central dogma itself is relatively simple, although the chemical mechanisms underlying it are not. The dogma states, "DNA molecules contain information about how to create proteins; this information is *transcribed* into RNA molecules, which, in turn, direct chemical machinery which *translates* the nucleic acid message into a protein." One way to remember the difference between transcription and translation is to note that DNA and RNA are different mechanisms for information storage (so exchange among them is mere transcription) but that protein is a mechanism for action, so its production requires translating information into action. Although it is most often the case that a single DNA sequence specifies a single protein, sometimes the DNA sequence of a gene can specify multiple proteins (through alternative splicing) or even no protein (when the transcribed RNA plays a direct functional role).

The central dogma states that the flow of information is one way, from DNA to RNA to protein. This is a good time to make the observation that there is practically no statement in all biology that is universally true, without exception. It turns out that there are even excep-

tions to the central dogma! For example, the AIDS virus, and its relatives in the broader family of retroviruses, is able to translate RNA into DNA—hence the "retro" in retrovirus. However, these kinds of exceptions are quite rare, and the central dogma is about as lawlike as any statement in biology ever gets.

A profound implication of the central dogma is that nearly all the information necessary to construct and operate a living thing is contained in its DNA.[2] We call the complete complement of DNA (and therefore the collection of all the genes) in a particular species its *genome*. That is why genome sequencing projects, which determine the exact sequence of all the DNA in an organism, are so important.

Structure and Function

The components and activities of living things are studied in two distinct and complementary ways: (1) their structure and (2) their function.

The components and activities of living things are studied in two distinct and complementary ways: (1) their structure and (2) their function. *Structure,* whether of an entire organism or of a single biomolecular component, describes physical composition and physical relationships. *Function* describes the role that a component plays in the processes of life. Much research in molecular biology is done to relate a known function to the (unknown) structures that instantiate it, or relating a known structure to the (unknown) function that it supports.

The pressure of evolutionary selection ensures that the main function of all living things is to turn environmentally available matter and energy into offspring more successfully than competitors. The structures that support this function consist minimally of three components: (1) boundaries separating the organism from its environment, (2) the organism's inheritable characteristics, and (3) all the other materials necessary for survival and reproduction. Boundaries take the form of *membranes* and are made of a class of small molecules called *lipids*. The inheritable characteristics, that is, the genome, are physically embodied in structures called *chromosomes* that consist primarily of DNA. The other materials necessary for life form a complex and highly structured mixture loosely called the *cytoplasm*.

The main function that life must support (turning food into babies) is realized through a wildly complex set of chemical reactions. At its essence, bonds among the atoms in the matter that living things consume as food must be broken and remade into the molecules needed for life—all at the right times and in the right amounts. Some of the reactions that living things use to turn food into offspring are thermodynamically feasible but would happen very slowly or infrequently on their own. Other reactions that living things manage to exploit are not even thermodynamically feasible. What makes these reactions happen at the rates needed for life? This question is why chemistry is so central to understanding biology.

Proteins, functioning as *enzymes,* provide the activation energy necessary to catalyze thermodynamically feasible reactions. Although they play other roles as well, many proteins have an enzymatic function. Even more importantly, reactions that are not thermodynamically feasible at all can be made to happen by coupling them to other reactions that break down energy-rich compounds and provide compensating entropy. This is the essence of *metabolism* and why organisms need energy (either from sunlight or food) to live.

Metabolism involves *catabolism,* the transformation of external energy into forms that can be used by the organism, and *anabolism,* the synthesis of the material components necessary for maintenance and reproduction of life (such as particular lipids, proteins, and nucleic acids). The material being acted on by an enzyme is often called its *substrate,* and the result of the enzymatic transformation of the substrate is the *product*. Generally speaking, metabolism is realized by sets of linked chemical reactions, where the product of one reaction becomes the substrate for the next. Each reaction in this chain is catalyzed by a different protein; such a set of reactions is called a *metabolic pathway.* There are hundreds of such pathways even in the simplest organisms, and these pathways branch and loop various ways to form complex metabolic networks (figures 2, 3).

How is it that proteins can accomplish these amazing feats of chemistry? The details of enzymatic mechanisms depend on the quantum mechanics of electrons and bonds, but without going into that level of detail, it is still possible to gain a rough understanding of the structure and function of proteins.

The enzymatic function of a protein generally has three aspects: (1) activity, (2) specificity, and (3) regulation. The *activity* of an enzyme is what it does chemically; for example, it might break a particular kind of bond. There are about a dozen very broad classes of activities and many variations on these themes. *Specificity* is the ability of proteins to recognize and act on only particular substrates, often being able to discriminate between extremely subtle chemical differences. Finally, the activity of a protein can often be turned on or off or modulated more finely by other molecules, termed the *regulation* of the enzyme.

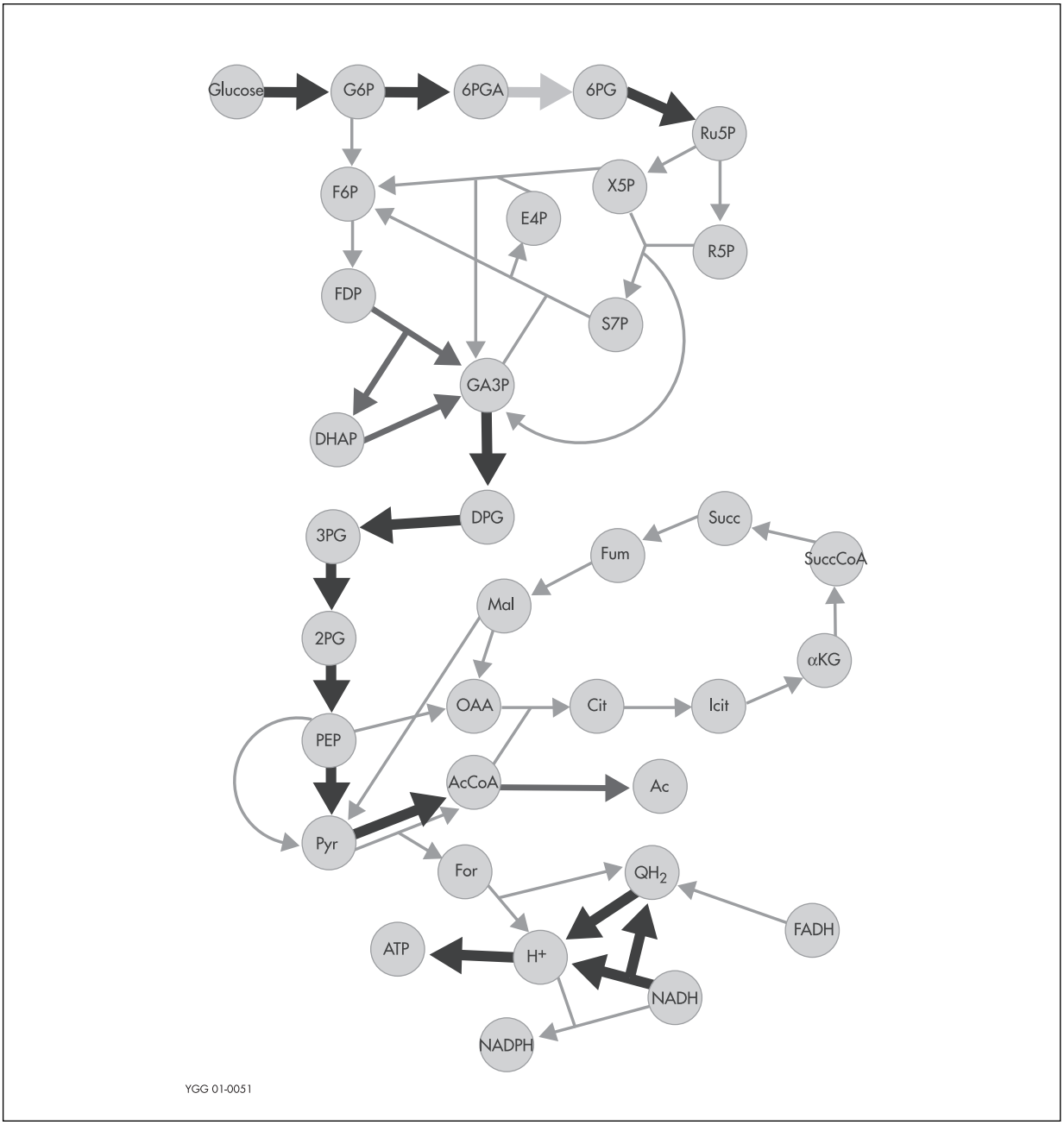

Figure 2. Metabolic Network Model for Escherichia coli.

Metabolic maps provide a framework for studying the consequences of genotype changes and the relationships between genotypes and phenotypes. This metabolic network model for *Escherichia coli* incorporated data on 436 metabolic intermediates undergoing 720 possible enzyme-catalyzed reactions. In this diagram, the circles contain abbreviated names of the metabolic intermediates, and the arrows represent enzymes. The very heavy lines indicate links with high metabolic fluxes. Analyses were correct 90 percent of the time in predicting the ability of 36 mutants with single-gene deletions to grow on different media. (*Image courtesy U.S. Department of Energy Genomes to Life Program, doegenomestolife.org. Figure adapted from J. S. Edwards and B. O. Palsson, Proc. Nat. Acad. Sci. 97, 5528-33 [2000].)*

Each of these aspects of enzymatic function is realized by a corresponding aspect of the structure of the protein. Recall that proteins are linear polymers, made up of a particular sequence of amino acids. An average protein might contain a bit more than a hundred amino acids; very large ones can have thousands. There are 20 different naturally occurring amino acids that are assembled into proteins, which means that the total number of possible proteins is enormous. Each different amino acid has somewhat different chemical

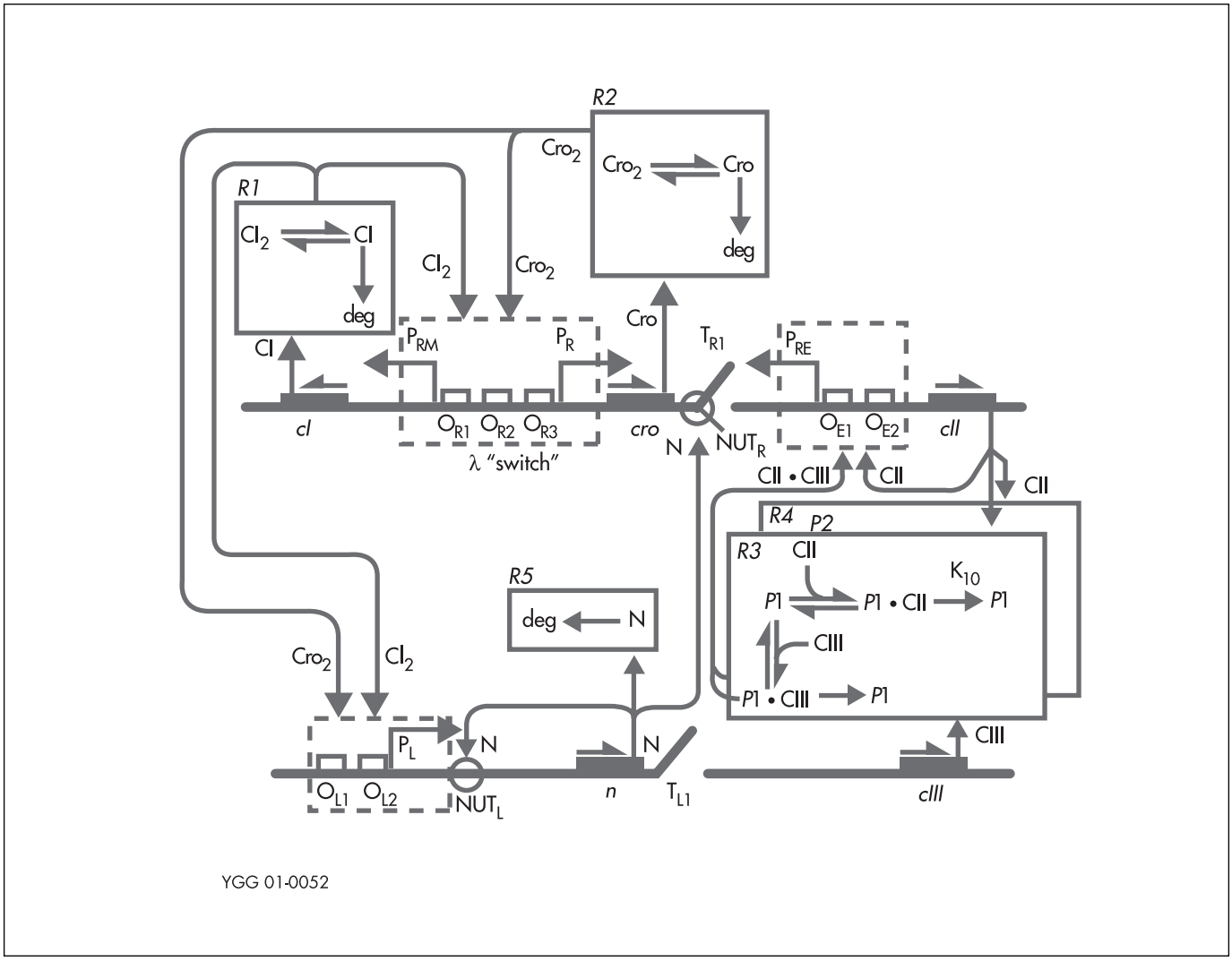

Figure 3. Pathway Kinetics.

The pathway kinetics model depicts the mechanisms of the "decision circuit" that commits a bacterial virus (lambda) to one of two alternate pathways in its life cycle. The lytic path sets the stage for immediate replication of the virus and destruction of its *Escherichia coli* host cell, and the lysogenic path selects for the incorporation of viral DNA into the host genome, allowing the virus to remain in a dormant state.

In the diagram, bold horizontal lines indicate stretches of double-stranded DNA, arrows over genes show the transcription direction, and dashed boxes enclose operator sites that make up a promoter control complex. The core of the decision circuit is the four-promoter, five-gene regulatory network; initiation of pathway actions involves other coupled genes not shown. (*Image courtesy U.S. Department of Energy Genomes to Life Program, doegenomestolife.org. Figure adapted from A. Arkin, J. Ross, and H. H. McAdams, Genetics 149, 163348 [1998].)*

properties; some are charged, some are heavy, some have aromatic rings,[3] some are hydrophobic,[4] and so on. The precise arrangement of amino acids determines the structure, activity, specificity, and mechanism of regulation of the protein.

Although proteins are linear polymers, the linear chain (also known as the backbone of the protein) self-assembles into quite complex three-dimensional shapes (or conformations). These three-dimensional shapes are crucial to the function of the protein. When dissolved in water, most proteins fold into either a single conformation or a small ensemble of similar conformations. However, there are an enormous number of physically possible conformations, and a complete understanding of this process or even enough of an understanding to predict the three-dimensional shape of a protein given its sequence of amino acids is still elusive, although much progress has been made recently. The correct folding of a protein is crucial for its function; the recently discovered prion diseases (such as mad cow disease) are caused by misfolded proteins. How particular confirmations impart particular activities,

specificities, and regulation mechanisms is also the subject of intense scrutiny. For example, such an understanding is often important in the development of new pharmaceuticals.

The Molecular Biology of the Gene

The central dogma connects the Mendelian idea of the function of a gene (that is, as the unit of genetic transmission) with a particular structure: Genes specify proteins. How does this work? DNA, which is the structure that embodies the genome, has to support two functions: First, it must be copied with high fidelity so that those instructions can be passed to offspring. Second, it must contain the specification of the proteins that ultimately determine (with the environment) the phenotype of an organism.

First, consider how inheritance works. DNA is also a linear polymer, this time of *nucleotides*. In some senses, the structure DNA is simpler than proteins because there are only four different kinds of nucleotides found in DNA, and no matter what its sequence, DNA forms pretty much the same three-dimensional structure, the famous double helix. However, the remarkable aspect of DNA's structure is how it supports replication. The nucleotide elements of DNA are adenosine, guanine, cytosine, and thymine, abbreviated *A, G, C*, and *T*, respectively. Each nucleotide forms chemical bonds with one of the other nucleotides, which is called *complementary: A* is complementary to *T* and *G* to *C*. Each nucleotide element of the polymer is always matched with its complement; the term for one of these units is a *base pair.* A sequence of nucleotides (called a *strand*) is also directional in that the "head" end can be distinguished from the "tail" end.[5] Thus, each DNA molecule actually embodies two sequences: (1) one going from head to tail and (2) the complementary strand going the other way:

head	*ACTGACTG*	tail
tail	*TGACTGAC*	head

The replicative machinery takes advantage of this complementarity. To copy a DNA molecule, it is unzipped (starting at either end), and a complementary nucleotide is bonded to each of the two unzipped segments. By the time the entire molecule is unzipped, new base pairs have been attached to each position in both of the original strands, which then form two copies of the original double-stranded DNA.

The second function that DNA must serve is to specify proteins. In fact, the chemical definition of a gene is "a sequence of nucleotides in DNA that codes for an amino acid" sequence. DNA is a linear polymer whose constituents are drawn from four nucleotides, and DNA sequences are often written down as strings of their abbreviations, for example, *ACCATAGGACTT*. The *genetic code* is the mapping by which nucleotide sequences are translated into amino acid sequences. Recall that there are 20 different amino acids but only 4 nucleic acids. The information required to specify 1 of 20 amino acids therefore requires at least 3 nucleotides (because there are only 16 possible combinations of 2 nucleotides). Nucleotide triplets are called *codons,* and they specify a particular amino acid. Because there are 64 possible codons and only 20 amino acids, there is some redundancy in the code; for example, both the *GGT* codon and *GGC* codon specify the amino acid glycine. There are also three *stop codons* that are used to indicate the end of a protein sequence and a commonly used *start codon* that also codes for the amino acid methionine.

The mapping from DNA sequences to proteins closes the loop in understanding how phenotype can be influenced by genotype. The phenotype of an organism is highly dependent on how its components are synthesized, that is, on its metabolism. Metabolism, in turn, is determined by the precise details of the organism's constituent proteins. Those details of protein structure and function can be traced to the protein's precise amino acid sequences, and from there to the organism's DNA, which exactly encodes the amino acid sequence of each protein an organism can make. That, briefly stated, is the connection between the structure of DNA and the function of Mendel's gene. Following the metaphor of the book of life, consider genomic DNA to be the text, proteins the words, metabolic processes the sentences and paragraphs, and the phenotype a movie based on the text.

Not all the DNA in an organism codes for proteins. The remainder, however, is not just junk. Some RNA molecules have roles other than transcriptional messengers and can be functional end products themselves. A key role of functional RNA molecules is as part of the *ribosome,* an assembly of proteins and RNA molecules that carries out the translation of messenger RNA into protein. Even DNA that doesn't code for protein or RNA can play an important role. One function of noncoding DNA is the regulation of gene expression, that is, the amount of each protein that is being made from a given gene at any particular time. The synthesis of proteins from the DNA code has a cost in matter and energy, and the organism doesn't need the same amount of each

protein at all times. Some proteins (say, those specialized for digesting unusual food sources or managing temperature stress) might not be needed at all in most circumstances. An important aspect of the way an organism controls how much protein is present is to regulate the synthesis of the protein (other mechanisms of control include the regulation of degradation and transport). The synthesis of a particular protein coded for by a particular gene is called the *expression* of that gene.

The expression of a gene begins with the transcription of genomic DNA into RNA. That process is itself controlled by the binding of particular proteins, called *transcription factors,* to the DNA molecule. Transcription factors recognize particular sequences of nucleotides that occur just upstream (before) from the start of the coding sequence of the gene. Transcriptional control is combinatorial, in that one transcription factor can influence the expression of many genes, and most genes are influenced by more than one transcription factor. Genes that are influenced by multiple transcription factors have multiple upstream DNA sequences that those transcription factors recognize. Regions that are required for transcription are called *promoters;* optional additional regions called *repressors* bind transcription factors that reduce the expression level, and others, called *enhancers,* bind transcription factors that increase it. Because transcription factors are themselves proteins, their activities are under the control of still other transcription factors, forming complex feedback loops that combine stability with responsiveness to environmental and other signals. These relationships among genes and transcription factors is the *genetic regulatory network*.

… prokaryotes are evolving organisms that are able to find novel and complex mechanisms to increase their fitness in the contemporary world….

Prokaryotes

Armed with an appreciation of evolution, chemistry, and the central dogma, one can begin to understand some of life's simplest forms: the bacteria. Bacteria, or to be more technically precise, the *prokaryotes*, are important in science (making up the vast majority of the biomass on the planet), economics (as tools in bioengineering) and human health (as pathogens). They are also utterly ubiquitous. The human gut is populated with the prokaryote *Escherceria coli*, which is necessary for people to properly digest their food. Small variations in that *E. coli* bacterium can turn it into a nasty pathogen, responsible for sometimes lethal cases of food poisoning. The shower curtain in your bathroom is probably home to billions of bacteria of thousands of different species, as is the dirt in your front yard, the sponge in your kitchen sink, and nearly every other place on the planet.

Prokaryotes are simple in a variety of respects. Prokaryotes are microscopic, single-celled organisms with minimal internal structure. Their fitness is largely determined by the speed at which they can reproduce. They contain fewer genes, organized in simpler regulatory patterns than other organisms. In fact, the study of certain bacteria was central in identifying the chemical components and processes that are absolutely necessary to life.

However, being as simple as possible still involves a fair degree of complexity. Prokaryotic molecular systems accomplish most of the key tasks in all living things: the capture, transport, and application of energy; the synthesis of all the molecules necessary for life from environmentally available materials; sensation, awareness, and response to the environment; the processing and even exchange of nucleic acid instructions; and, of course, reproduction, in the form of cell division called *mitosis*. All these cellular processes are carried out by complex networks of proteins and catalyzed reactions. Despite their relative simplicity and the decades of research done on *E. coli*, a complete mechanistic understanding of all the activities of even these organisms is still elusive. In addition, it is important to keep in mind that prokaryotes are evolving organisms that are able to find novel and complex mechanisms to increase their fitness in the contemporary world, for example, for resisting antibiotics, eating plastics, or otherwise living in human-dominated environments.

The Eukaryotes and the Eukaryotic Cell

Despite the ubiquity of bacteria, most people are more familiar with the broad class of organisms that include plants and animals; that is, the *eukaryotes*. Eukaryotes include a tremendous range of organisms, from tiny free-living single-celled organisms to human beings consisting of more than a trillion cells. All multicellular organisms (and therefore all that are visible to the naked eye) are eukaryotic. Even some single-celled eukaryotes are familiar, such as brewer's yeast or athlete's foot fungus. (Although not discussed further in this introduction, there is also a third main branch of life called *archaea*. These organisms are all single celled but are more like eukaryotes than prokaryotes in a variety of ways; many are extremophiles that live in environments that once had been thought to be incompatible with life.)

At the cellular level, eukaryotes are easily distinguishable from bacteria by their greatly elaborated internal structure. Eukaryotic cells have a variety of internal compartments separated by membranes and various specialized internal components called *organelles*. The origins of the eukaryotic cell are particularly interesting; they appear to be the result of symbiotic communities of prokaryotes losing their individuality and merging into a single entity. Organelles are remnants of this merging—one even has kept its own genome, which reproduces and is inherited separately from the rest of the organism's genome.

The most striking organelle of the eukaryotic cell is the *nucleus,* which contains all the genetic material of the cell. Under typical microscopic stains, the nucleus appears as a large, dark (or false-colored blue) central sphere in a light (or false-colored pink) background of cytoplasm. The nucleus is where gene expression is controlled and where DNA replication occurs. The initial transcription of genomic DNA into RNA also occurs in the nucleus, but that messenger RNA (mRNA) is then transported to another organelle in the cytoplasm, called a *ribosome,* for translation into protein. Unlike bacteria, the DNA that codes for protein in eukaryotes can be interrupted by noncoding regions of DNA called *introns*. Special proteins remove the noncoding regions from the mRNA before it is translated into protein at the ribosome. The parts of the gene that are not spliced out and that continue to be translated into proteins are called *exons*. One way to remember which is which is to think that *ex*ons are *ex*pressed, and *in*trons *in*terrupt. The functional role that introns play is not entirely clear. At least one function is to allow a single gene to code for multiple related proteins through the use of alternative splicing, or using different subsets of exons from a single gene to code for entire families of related proteins.

There are many other differences between prokaryotes and the eukaryotic cell. Another important difference is the mechanism by which the expression of genes is regulated. In prokaryotes, functionally related genes are adjacent to each other in the genome, and a single promoter can control the whole group of genes. Regulation of the expression of eukaryotic genes is much more complex.

Remarkably enough, the biochemistry of nearly all eukaryotic organisms is quite similar. The mechanisms of metabolism, protein coding, gene regulation, genome replication, and so on are quite similar in humans and, say, earthworms.

Multicellular Organisms

Every organism visible to the naked eye contains many cells and is called *multicellular.* All multicellular organisms are eukaryotes. What differentiates a multicellular organism from a colony of single-celled organisms is that a multicellular organism is composed of cells of many different types, each of which specializes to accomplish a particular task. *Cellular specialization* is the hallmark of multicellularity.

All the cells in a multicellular organism are descendants of a single fertilized egg cell and, therefore, have exactly the same DNA. However, the cells differentiate from one another into *lineages*. Most cells become committed once they are differentiated; that is, they and their descendants cannot change to become any other cell type. Differentiated cells specialize in particular functions; for example, muscle cells contract, nerve cells process signals, and so on. Cellular specialization is a clear example of the importance of differential gene expression. Different cell types in an organism all have exactly the same genome; it is differences among the expression levels of their genes that make a muscle cell different from a nerve cell.

The most basic division among cell types is between *germline* cells (that is, reproductive cells such as eggs and sperm) and *somatic* cells. Although the somatic cells sometimes divide, none of their genomes will make it directly into another generation of the entire organism. Any variation that arises in a somatic cell will be lost forever, although variations that arise in germline cells will be passed along to the next generation. From the somatic cell's perspective, this is a remarkable loss of function. The grand deal that made cellular specialization possible is a fascinating evolutionary story that also likely involves the origin of sexual reproduction. The breakdown of this deal, when somatic cells begin to reproduce without regard to their being part of an organism, is what we call *cancer.*

Most multicellular organisms reproduce sexually, although some can also reproduce by budding. Sexual reproduction involves a different mechanism than mitosis, the process by which other cells divide. This special process, called *meosis,* is the mechanism by which two parental genomes are transformed into a single offspring's genome, and it underlies the Mendelian nature of inheritance.

Reproduction is not the only place that cells in multicellular organisms have to cooperate. Nearly every organismal function depends on tight coordination among cells, and an elaborate mechanism for sending and receiving signal has developed in response to this need. This mechanism involves both the release and

Web Sites to Visit

Larry Hunter's Web Site

compbio.uchsc.edu/hunter/
My web site has pointers to many useful teaching and learning resources.

Human Genome Project (DoE) Education

www.ornl.gov/hgmis/education/education.html

National Human Genome Research Institute (NIH) Education Pages

www.genome.gov/page.cfm?pageID=10000002

The National Library of Medicine (NIH)

www.nlm.nih.gov

NCBI

www.ncbi.nlm.nih.gov
NCBI provides GenBank and many other databases.

Molecular Biology Textbooks

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

AcademicInfo.net Biology Research Pages

www.academicinfo.net/biology.html

AcademicInfo.net Biology Education Pages

www.academicinfo.net/biologyed.html

All Species Foundation

www.all-species.org/
The All Species Foundation is attempting to catalog all the species on earth.

The University of California Museum of Paleontology History of Life

www.ucmp.berkeley.edu/historyoflife/histoflife.html

The Hooper Virtual Paleontological Museum

hannover.park.org/Canada/Museum/lobby.html

Mass Extinctions

hannover.park.org/Canada/Museum/extinction/homepg.html
This museum has a very nice site on mass extinctions.

Anthropogenic Mass Extinction

www.well.com/user/davidu/extinction.html
This is an excellent web site collecting information about the possible current anthropogenic mass extinction event.

The Tree of Life Web Project

tolweb.org/tree/phylogeny.html
This is a good taxonomy web site.

AccessExcellence

www.accessexcellence.org/BF/bf02/lipps/
The origin of multicellular organisms is well described by Jere Lipps on this site, which also has great pictures.

Protein Databank

www.rcsb.org/pdb/

Protein Databank Molecule of the Month

www.rcsb.org/pdb/molecules/molecule_list.html

Protein Databank Ribosome

www.rcsb.org/pdb/molecules/pdb10_1.html

RNA Structure

www.rnabase.org
RNAbase is a database of RNA structures.

RNA Structure Primer

www.rnabase.org/primer
RNAbase has an excellent primer on RNA structure..

Visible Human Anatomy

www.uchsc.edu/sm/chs/open.html

reception of signaling molecules and the creation of an appropriate response to signals within a cell. Cell membranes are studded with an enormous variety of molecules called *receptors* that receive these signals. Receptors respond only to very specific molecular signals; the molecule that a particular receptor responds to is called its *ligand*. The process set into motion by the binding of a ligand to its receptor is called *signal transduction*. This process involves interacting cascades of reactions in which proteins called *secondary messengers* chemically modify one another, integrating multiple incoming signals over time and ultimately activating various transcription factors, causing changes in gene expression within the cell.

Signal transduction plays a key role in the function of many pharmaceuticals. Because the function of cell membranes is largely to keep undesired substances (such as drugs) out of the cell, it is much easier to find a chemical that interacts with a receptor than it is to find one that can get inside a cell. Receptors can trigger an enormous number of responses (even cell suicide, or *apoptosis*), and hence, drugs that bind

to them can mediate effects such as reducing blood pressure, suppressing immune responses, or even alleviating depression.

Tissues, Organs, and Development

Not only are cells in multicellular organisms specialized, they are precisely arrayed in particular spatial patterns. *Tissues* are collections of cells of a particular type in a particular spatial distribution. The cells in a tissue all arise from the same lineage. Tissues that work together to execute a particular biological function are called *organs,* such as a kidney or a leaf. Organs, in turn, are grouped into *organ systems*.

The four main human tissue types are (1) epithelium, (2) connective tissue, (3) muscle tissue, and (4) nerve tissue. *Epithelium* is the tissue covering all body surfaces and the lining of internal organs and glands. There are three subclasses of simple epithelium and then various combinations. *Connective tissue* is distinguished not only by the type of cells it contains but also by the extracellular material around the cells. Different types of connective tissue include supportive tissues such as bone and cartilage as well as fat, blood, and lymphatic (or immune system) tissue. There are three distinct kinds of muscle tissue: First, striated muscle is the kind under voluntary control, such as that in arm or leg muscles. Second, smooth muscle is found in places such as blood vessel walls and the intestines and is generally not under voluntary control. Third, the heart is made of a special kind of cardiac muscle. Nervous tissue is specialized for sending and receiving signals and makes up the constituents of the nerves, brain, and sense organs.

Specifying where one organ (or even organ system) ends and another begins is largely a matter of definition. Consider just a few of the dozen or so human organ systems. The *circulatory system* includes the heart, blood, and blood vessels. The *digestive system* includes the mouth, teeth, tongue, esophagus, stomach, and intestines. Some would also include the glands of digestion (the pancreas, the gallbladder, and the liver) in the digestive system, but others would include them in a glandular system.

The transformation from fertilized egg to mature adult is called the *process of development.* The study of development pursues three main questions: *(1) differentiation,* or how a single cell gives rise to all the many cell types found in adult organisms; (2) *morphogenesis,* or how tissues are organized spatially to make organs and how organs are arranged into a body plan; and (3) *growth,* or how proliferation (and cell death) is regulated and how cells know when to divide and when not to. Developmental mechanisms, although still not completely understood, evolve very slowly and tend to be widely shared across large numbers of organisms.

The structure of the tissues and organs of the body is the subject of the study of *anatomy,* and their function is the study of *physiology,* both of which are beyond the scope of this introduction.

Instrumentation and Experimental Systems

Biology involves not only its subject matter but also the methods used to study it. A great deal of the excitement in contemporary biology comes from the rapid pace of innovation in biological instrumentation, which for the first time is producing data about living systems that is both molecular in detail and genomewide in scope. Such instrumentation is generally referred to as *high throughput,* meaning that whatever is being assayed can be measured quickly enough to look at a large proportion of the biomolecules in an organism.

The first high-throughput instrumentation, and in some ways the most fundamental, provides the ability to determine the specific sequence of a molecule of DNA. Longer sequences are harder to obtain, so although there are thousands of viral genomes that have been determined, and hundreds of bacterial ones, there are fewer than two dozen completely sequenced eukaryotic genomes.

However, these complete sequenced organisms are not chosen at random. They represent organisms that are either of significant economic importance themselves (for example, rice) or are particularly amenable to experimentation and explanation. Such creatures are called *model organisms* and include mice, the fruit fly *Drosophila melangaster,* a simple multicellular worm called *Caenorhabditis elegans,* and the single-celled eukaryote brewer's yeast. Although the complete genomic sequence is available for only a tiny fraction of the world's organisms, the sequences of at least some genes of particular interest (for example, spider silk) have been determined for tens of thousands of organisms.

DNA sequencing provides information not only about the genome but also can be used to look at variations, or *polymorphisms,* among different individuals of the same species. Assaying a particular set of polymorphisms among different individuals, whether by DNA sequencing or other means, is called *genotyping*. The smallest possible difference is a *single nucleotide polymor-*

phism (SNP), and technology for high-throughput SNP genotyping is beginning to be used in molecular genetics laboratories.

Gene sequences generally contain all the information necessary to specify the three-dimensional structure, hence the chemical function, of a protein; however, there is as yet no practical method for mapping from an amino acid sequence to its folded structure. Instead, instrumentation involving X-ray crystallography or nuclear magnetic resonance can be used to empirically determine a protein's structure. Although still difficult, protein structure determination is rapidly increasing in speed. *Structural genomics* is the name for the effort now under way to determine a representative set of three-dimensional structures, including all medically important human proteins and proteins from important pathogens and model organisms.

Recall that the differences between various cell types and tissues in an organism are not determined directly by genotype (which is the same for all its cells) but instead by differences in expression levels among the various genes over time. Once the sequence of the genes of an organism is known, it is possible to fabricate a device, called an *expression array* or *gene chip,* that takes snapshots of the expression level of all genes simultaneously.

Even at a particular instant, snapshots of gene expression do not tell the whole story of what is going on in a cell. Recall also that signal transduction, which is very important for the invention of new drugs, involves the modification of existing proteins rather than the synthesis of new ones. *Mass spectrometry* can be used to assay proteins and modifications to them. Although the technology is still evolving, this approach is also becoming high throughput and forms the basis for *proteomics,* or the study of the complete set of proteins in a living system.

Despite these spectacular innovations, biology is still quite limited in its measurement abilities (and their associated costs). For example, although gene expression can vary quite significantly among individual cells in a tissue, current expression array technology requires RNA from a large number of cells combined to get a signal. Innovation in molecular instrumentation is proceeding very rapidly, and it is reasonable to expect that new methods will provide a great deal more valuable information, and the insights into the functions of living things that come from them, in coming years.

Conclusions

Learning molecular biology is a process that goes in a spiral; one repeatedly studies the same aspects of living systems but each time in more detail and with new perspective. This introduction is intended to bring the reader around the circuit once, providing only the coarsest generalities and few examples or details. Nevertheless, the careful reader should now be able to put into context many of the biological problems addressed by the computational methods described in this issue.

For those whose appetite has been whetted, an abundance of excellent textbooks and references are available as well, some online (see sidebar). The two most widely used college-level textbooks are *Molecular Cell Biology* (W. H. Freeman, 2003), by Harvey Lodish et al., and *Molecular Biology of the Cell* (Garland, 2002), by Bruce Alberts et al. For a somewhat more gentle introduction, I recommend either the wonderfully illustrated *The Way Life Works* (Three Rivers Press, 1998), by Mahlon Hoagland and Bert Dodson, or *Molecular Biology Made Simple and Fun* (Cache River, 2000), by David Clark and Lonnie Russell.

Notes

1. Mechanisms of disease.

2. There is also a role for the maternal proteins in the fertilized egg cell, but this is relatively minor compared to the contribution of the DNA.

3. Chemical structures when the atoms are bonded together to form a loop.

4. Water hating, such as oil.

5. One end is called 5′ (pronounced five-prime), and the other end is called 3′ (three prime).

Lawrence Hunter is one of the founders of bioinformatics. After graduating with a Ph.D. in computer science from Yale University in 1989, he spent more than a decade as a research scientist at the National Institutes of Health before joining the faculty of the University of Colorado School of Medicine, where he now directs the Center for Computational Pharmacology. He is writing a book-length introduction to molecular biology to be published by The MIT Press in 2004. His e-mail address is larry.hunter@uchsc.edu.