1 Cells and Genomes

The surface of our planet is populated by living things—curious, intricately organized chemical factories that take in matter from their surroundings and use these raw materials to generate copies of themselves. The living organisms appear extraordinarily diverse. What could be more different than a tiger and a piece of seaweed, or a bacterium and a tree? Yet our ancestors, knowing nothing of cells or DNA, saw that all these things had something in common. They called that something "life," marveled at it, struggled to define it, and despaired of explaining what it was or how it worked in terms that relate to nonliving matter.

The discoveries of the past century have not diminished the marvel—quite the contrary. But they have lifted away the mystery as to the nature of life. We can now see that all living things are made of cells, and that these units of living matter all share the same machinery for their most basic functions. Living things, though infinitely varied when viewed from the outside, are fundamentally similar inside. The whole of biology is a counterpoint between the two themes: astonishing variety in individual particulars; astonishing constancy in fundamental mechanisms. In this first chapter we begin by outlining the universal features common to all life on our planet. We then survey, briefly, the diversity of cells. And we see how, thanks to the common code in which the specifications for all living organisms are written, it is possible to read, measure, and decipher these specifications to achieve a coherent understanding of all the forms of life, from the smallest to the greatest.

In This Chapter

THE UNIVERSAL FEATURES OF CELLS ON EARTH

It is estimated that there are more than 10 million—perhaps 100 million—living species on Earth today. Each species is different, and each reproduces itself faithfully, yielding progeny that belong to the same species: the parent organism hands down information specifying, in extraordinary detail, the characteristics that the offspring shall have. This phenomenon of *heredity* is central to the definition of life: it distinguishes life from other processes, such as the growth of a crystal, or the burning of a candle, or the formation of waves on water, in which orderly structures are generated but without the same type of link between the peculiarities of parents and the peculiarities of offspring. Like the candle flame, the living organism consumes free energy to create and maintain its organization; but the free energy drives a hugely complex system of chemical processes that is specified by the hereditary information.

Most living organisms are single cells; others, such as ourselves, are vast multicellular cities in which groups of cells perform specialized functions and are linked by intricate systems of communication. But in all cases, whether we discuss the solitary bacterium or the aggregate of more than 10^{13} cells that form a human body, the whole organism has been generated by cell divisions from a single cell. The single cell, therefore, is the vehicle for the hereditary information that defines the species (**Figure 1–1**). And specified by this information, the cell includes the machinery to gather raw materials from the environment, and to construct out of them a new cell in its own image, complete with a new copy of the hereditary information. Nothing less than a cell has this capability.

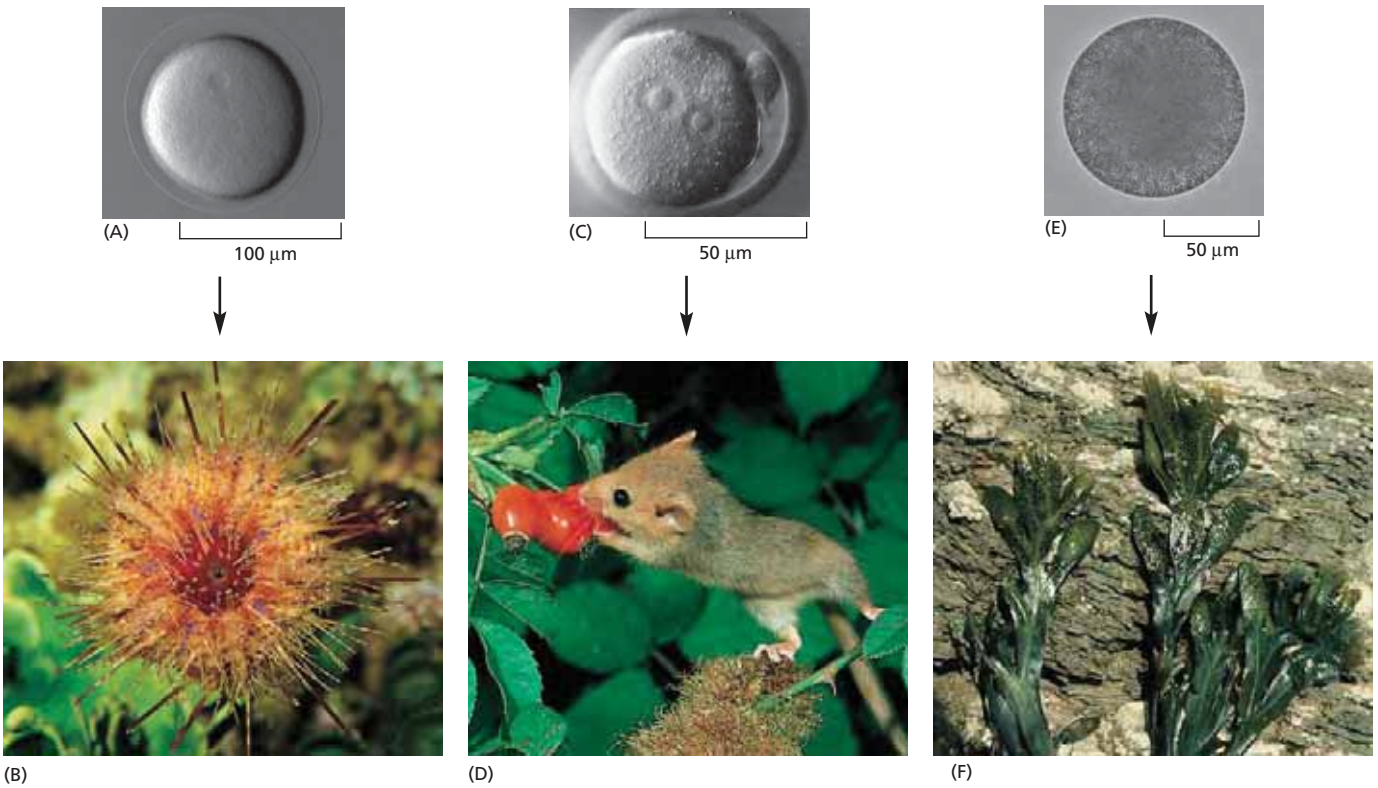


Figure 1–1 The hereditary information in the fertilized egg cell determines the nature of the whole multicellular organism. (A and B) A sea urchin egg gives rise to a sea urchin. (C and D) A mouse egg gives rise to a mouse. (E and F) An egg of the seaweed *Fucus* gives rise to a *Fucus* seaweed. (A, courtesy of David McClay; B, courtesy of M. Gibbs, Oxford Scientific Films; C, courtesy of Patricia Calarco, from G. Martin, *Science* 209:768–776, 1980. With permission from AAAS; D, courtesy of O. Newman, Oxford Scientific Films; E and F, courtesy of Colin Brownlee.)

All Cells Store Their Hereditary Information in the Same Linear Chemical Code (DNA)

Computers have made us familiar with the concept of information as a measurable quantity—a million bytes (to record a few hundred pages of text or an image from a digital camera), 600 million for the music on a CD, and so on. They have also made us well aware that the same information can be recorded in many different physical forms. As the computer world has evolved, the discs and tapes that we used 10 years ago for our electronic archives have become unreadable on present-day machines. Living cells, like computers, deal in information, and it is estimated that they have been evolving and diversifying for over 3.5 billion years. It is scarcely to be expected that they should all store their information in the same form, or that the archives of one type of cell should be readable by the information-handling machinery of another. And yet it is so. All living cells on Earth, without any known exception, store their hereditary information in the form of double-stranded molecules of DNA—long unbranched paired polymer chains, formed always of the same four types of monomers. These monomers have nicknames drawn from a four-letter alphabet—A, T, C, G—and they are strung together in a long linear sequence that encodes the genetic information, just as the sequence of 1s and 0s encodes the information in a computer file. We can take a piece of DNA from a human cell and insert it into a bacterium, or a piece of bacterial DNA and insert it into a human cell, and the information will be successfully read, interpreted, and copied. Using chemical methods, scientists can read out the complete sequence of monomers in any DNA molecule—extending for millions of nucleotides—and thereby decipher the hereditary information that each organism contains.

All Cells Replicate Their Hereditary Information by Templated Polymerization

The mechanisms that make life possible depend on the structure of the double-stranded DNA molecule. Each monomer in a single DNA strand—that is, each **nucleotide**—consists of two parts: a sugar (deoxyribose) with a phosphate group attached to it, and a *base*, which may be either adenine (A), guanine (G), cytosine (C) or thymine (T) (**Figure 1–2**). Each sugar is linked to the next via the phosphate group, creating a polymer chain composed of a repetitive sugar-phosphate backbone with a series of bases protruding from it. The DNA polymer is extended by adding monomers at one end. For a single isolated strand, these can, in principle, be added in any order, because each one links to the next in the same way, through the part of the molecule that is the same for all of them. In the living cell, however, DNA is not synthesized as a free strand in isolation, but on a template formed by a preexisting DNA strand. The bases protruding from the existing strand bind to bases of the strand being synthesized, according to a strict rule defined by the complementary structures of the bases: A binds to T, and C binds to G. This base-pairing holds fresh monomers in place and thereby controls the selection of which one of the four monomers shall be added to the growing strand next. In this way, a double-stranded structure is created, consisting of two exactly complementary sequences of As, Cs, Ts, and Gs. The two strands twist around each other, forming a double helix (Figure 1–2E).

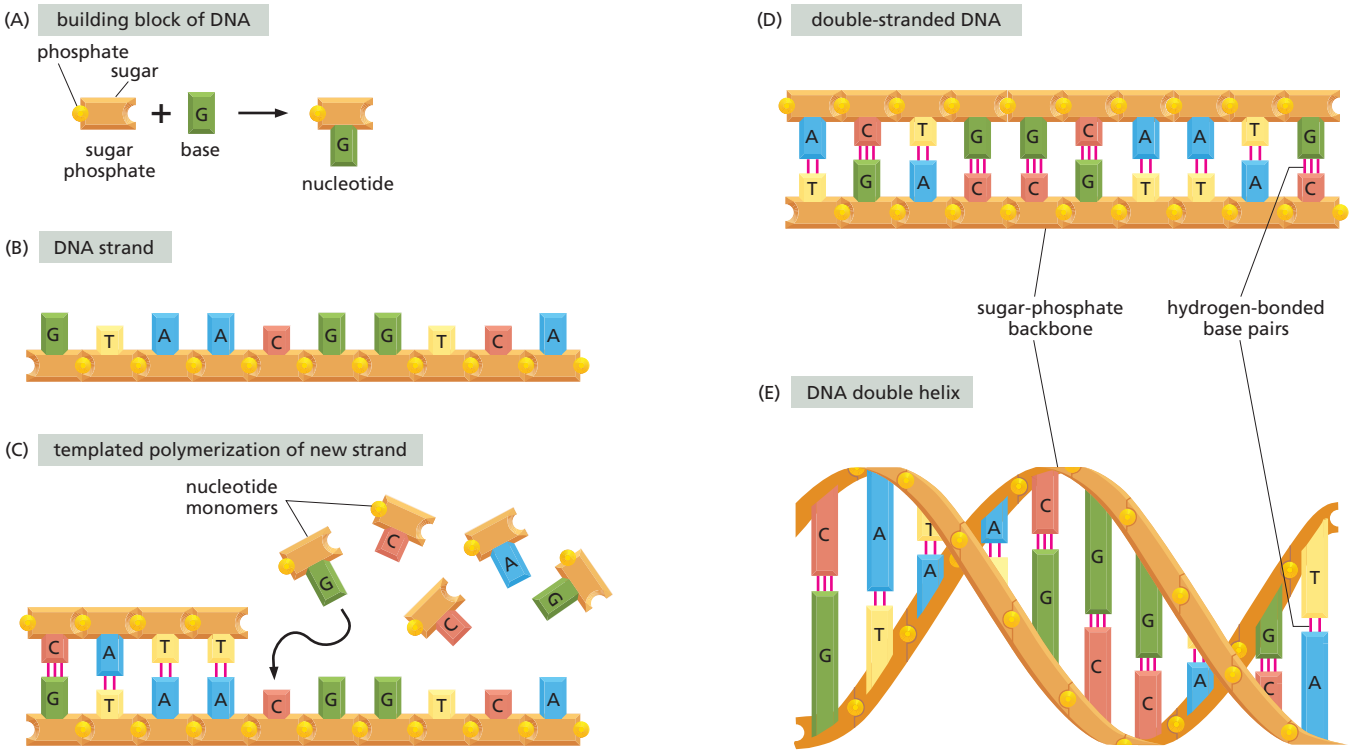


Figure 1–2 DNA and its building blocks. (A) DNA is made from simple subunits, called nucleotides, each consisting of a sugar-phosphate molecule with a nitrogen-containing sidegroup, or base, attached to it. The bases are of four types (adenine, guanine, cytosine, and thymine), corresponding to four distinct nucleotides, labeled A, G, C, and T. (B) A single strand of DNA consists of nucleotides joined together by sugar-phosphate linkages. Note that the individual sugar-phosphate units are asymmetric, giving the backbone of the strand a definite directionality, or polarity. This directionality guides the molecular processes by which the information in DNA is interpreted and copied in cells: the information is always "read" in a consistent order, just as written English text is read from left to right. (C) Through templated polymerization, the sequence of nucleotides in an existing DNA strand controls the sequence in which nucleotides are joined together in a new DNA strand; T in one strand pairs with A in the other, and G in one strand with C in the other. The new strand has a nucleotide sequence complementary to that of the old strand, and a backbone with opposite directionality: corresponding to the GTAA... of the original strand, it has ...TTAC. (D) A normal DNA molecule consists of two such complementary strands. The nucleotides within each strand are linked by strong (covalent) chemical bonds; the complementary nucleotides on opposite strands are held together more weakly, by hydrogen bonds. (E) The two strands twist around each other to form a double helix—a robust structure that can accommodate any sequence of nucleotides without altering its basic structure.

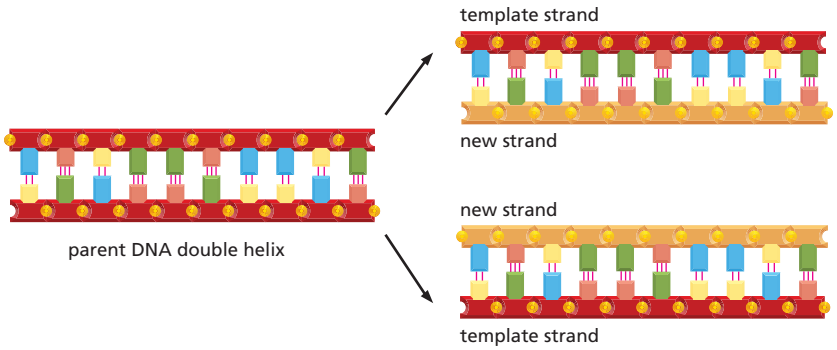


Figure 1–3 The copying of genetic information by DNA replication. In this process, the two strands of a DNA double helix are pulled apart, and each serves as a template for synthesis of a new complementary strand.

The bonds between the base pairs are weak compared with the sugar-phosphate links, and this allows the two DNA strands to be pulled apart without breakage of their backbones. Each strand then can serve as a template, in the way just described, for the synthesis of a fresh DNA strand complementary to itself—a fresh copy, that is, of the hereditary information (**Figure 1–3**). In different types of cells, this process of **DNA replication** occurs at different rates, with different controls to start it or stop it, and different auxiliary molecules to help it along. But the basics are universal: DNA is the information store, and *templated polymerization* is the way in which this information is copied throughout the living world.

All Cells Transcribe Portions of Their Hereditary Information into the Same Intermediary Form (RNA)

To carry out its information-bearing function, DNA must do more than copy itself. It must also *express* its information, by letting it guide the synthesis of other molecules in the cell. This also occurs by a mechanism that is the same in all living organisms, leading first and foremost to the production of two other key classes of polymers: RNAs and proteins. The process (discussed in detail in Chapters 6 and 7) begins with a templated polymerization called **transcription**, in which segments of the DNA sequence are used as templates for the synthesis of shorter molecules of the closely related polymer **ribonucleic acid**, or **RNA**. Later, in the more complex process of **translation**, many of these RNA molecules direct the synthesis of polymers of a radically different chemical class—the *proteins* (**Figure 1–4**).

In RNA, the backbone is formed of a slightly different sugar from that of DNA—ribose instead of deoxyribose—and one of the four bases is slightly different—uracil (U) in place of thymine (T); but the other three bases—A, C, and G—are the same, and all four bases pair with their complementary counterparts in DNA—the A, U, C, and G of RNA with the T, A, G, and C of DNA. During transcription, RNA monomers are lined up and selected for polymerization on a template strand of DNA, just as DNA monomers are selected during replication. The outcome is a polymer molecule whose sequence of nucleotides faithfully represents a part of the cell's genetic information, even though written in a slightly different alphabet, consisting of RNA monomers instead of DNA monomers.

The same segment of DNA can be used repeatedly to guide the synthesis of many identical RNA transcripts. Thus, whereas the cell's archive of genetic information in the form of DNA is fixed and sacrosanct, the RNA transcripts are mass-produced and disposable (**Figure 1–5**). As we shall see, these transcripts function as intermediates in the transfer of genetic information: they mainly serve as **messenger RNA** (**mRNA**) to guide the synthesis of proteins according to the genetic instructions stored in the DNA.

RNA molecules have distinctive structures that can also give them other specialized chemical capabilities. Being single-stranded, their backbone is flexible, so that the polymer chain can bend back on itself to allow one part of the

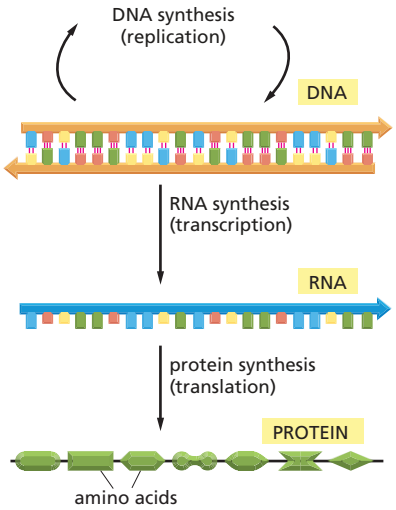


Figure 1–4 From DNA to protein. Genetic information is read out and put to use through a two-step process. First, in *transcription*, segments of the DNA sequence are used to guide the synthesis of molecules of RNA. Then, in *translation*, the RNA molecules are used to guide the synthesis of molecules of protein.

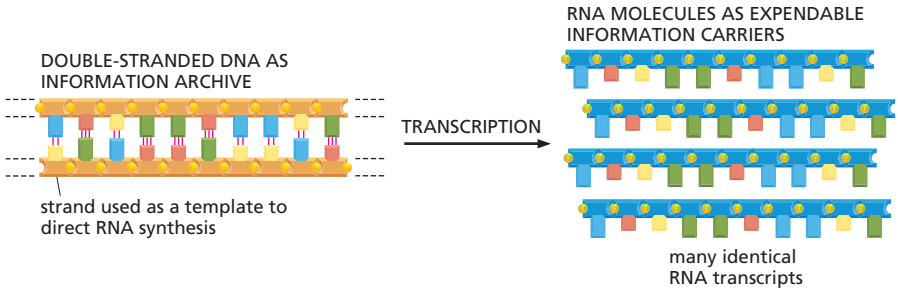


Figure 1–5 How genetic information is broadcast for use inside the cell. Each cell contains a fixed set of DNA molecules—its archive of genetic information. A given segment of this DNA guides the synthesis of many identical RNA transcripts, which serve as working copies of the information stored in the archive. Many different sets of RNA molecules can be made by transcribing selected parts of a long DNA sequence, allowing each cell to use its information store differently.

molecule to form weak bonds with another part of the same molecule. This occurs when segments of the sequence are locally complementary: a ...GGGG... segment, for example, will tend to associate with a ...CCCC... segment. These types of internal associations can cause an RNA chain to fold up into a specific shape that is dictated by its sequence (**Figure 1–6**). The shape of the RNA molecule, in turn, may enable it to recognize other molecules by binding to them selectively—and even, in certain cases, to catalyze chemical changes in the molecules that are bound. As we see in Chapter 6, a few chemical reactions catalyzed by RNA molecules are crucial for several of the most ancient and fundamental processes in living cells, and it has been suggested that more extensive catalysis by RNA played a central part in the early evolution of life.

All Cells Use Proteins as Catalysts

Protein molecules, like DNA and RNA molecules, are long unbranched polymer chains, formed by stringing together monomeric building blocks drawn from a standard repertoire that is the same for all living cells. Like DNA and RNA, they carry information in the form of a linear sequence of symbols, in the same way as a human message written in an alphabetic script. There are many different protein molecules in each cell, and—leaving out the water—they form most of the cell's mass.

The monomers of protein, the **amino acids**, are quite different from those of DNA and RNA, and there are 20 types, instead of 4. Each amino acid is built around the same core structure through which it can be linked in a standard way to any other amino acid in the set; attached to this core is a side group that gives each amino acid a distinctive chemical character. Each of the protein molecules, or **polypeptides**, created by joining amino acids in a particular sequence folds into a precise three-dimensional form with reactive sites on its surface (**Figure**

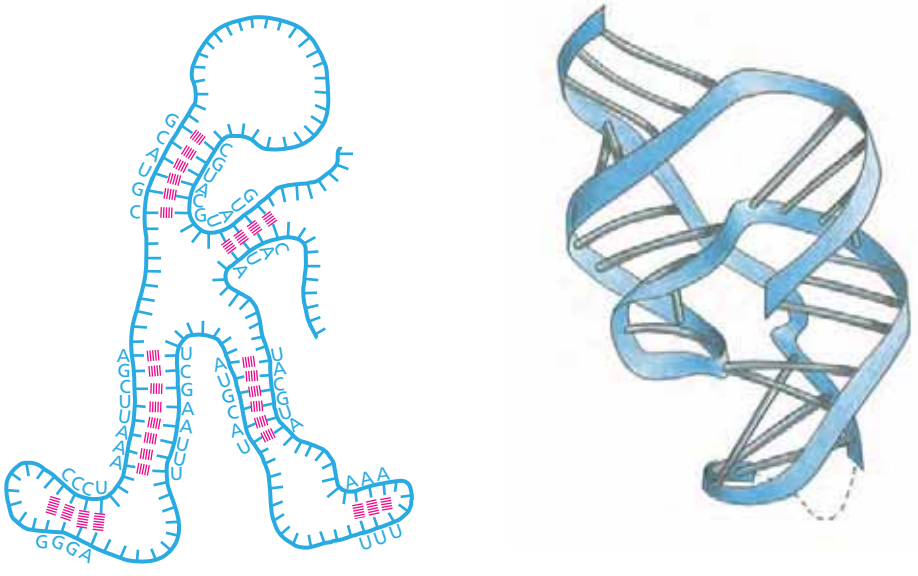


Figure 1–6 The conformation of an RNA molecule. (A) Nucleotide pairing between different regions of the same RNA polymer chain causes the molecule to adopt a distinctive shape. (B) The three-dimensional structure of an actual RNA molecule, from hepatitis delta virus, that catalyzes RNA strand cleavage. The *blue* ribbon represents the sugar-phosphate backbone; the bars represent base pairs. (B, based on A.R. Ferré D'Amaré, K. Zhou and J.A. Doudna, *Nature* 395:567–574, 1998. With permission from Macmillan Publishers Ltd.)

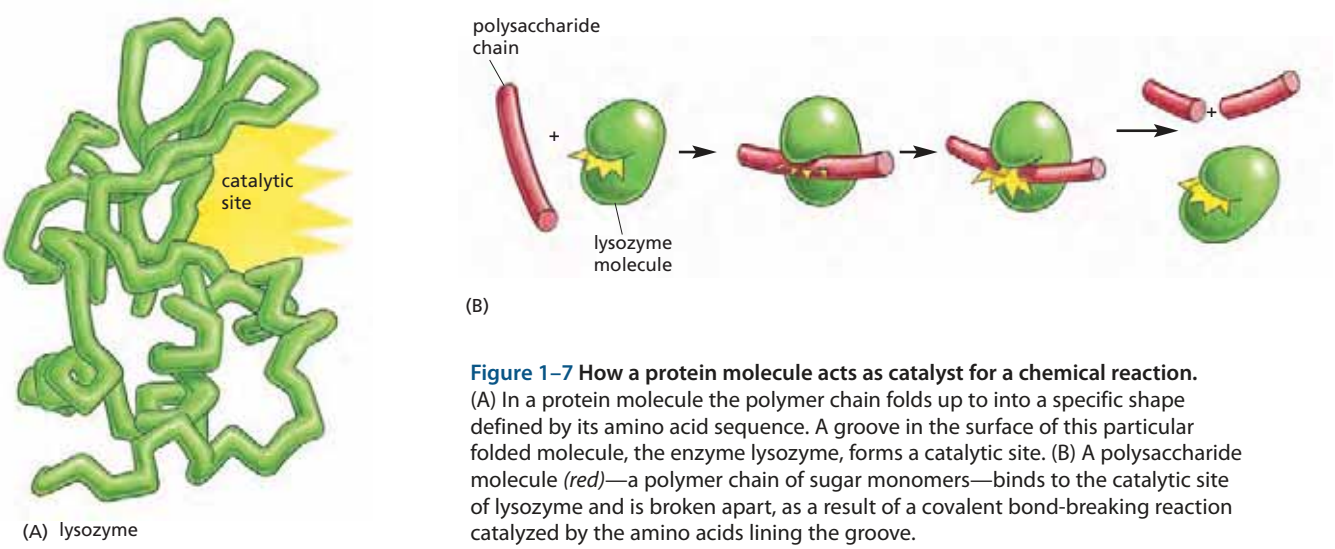


Figure 1–7 How a protein molecule acts as catalyst for a chemical reaction. (A) In a protein molecule the polymer chain folds up to into a specific shape defined by its amino acid sequence. A groove in the surface of this particular folded molecule, the enzyme lysozyme, forms a catalytic site. (B) A polysaccharide molecule *(red)*—a polymer chain of sugar monomers—binds to the catalytic site of lysozyme and is broken apart, as a result of a covalent bond-breaking reaction catalyzed by the amino acids lining the groove.

1–7A). These amino acid polymers thereby bind with high specificity to other molecules and act as **enzymes** to catalyze reactions that make or break covalent bonds. In this way they direct the vast majority of chemical processes in the cell (Figure 1–7B). Proteins have many other functions as well—maintaining structures, generating movements, sensing signals, and so on—each protein molecule performing a specific function according to its own genetically specified sequence of amino acids. Proteins, above all, are the molecules that put the cell's genetic information into action.

Thus, polynucleotides specify the amino acid sequences of proteins. Proteins, in turn, catalyze many chemical reactions, including those by which new DNA molecules are synthesized, and the genetic information in DNA is used to make both RNA and proteins. This feedback loop is the basis of the autocatalytic, self-reproducing behavior of living organisms (**Figure 1–8**).

All Cells Translate RNA into Protein in the Same Way

The translation of genetic information from the 4-letter alphabet of polynucleotides into the 20-letter alphabet of proteins is a complex process. The rules of this translation seem in some respects neat and rational, in other respects strangely arbitrary, given that they are (with minor exceptions) identical in all living things. These arbitrary features, it is thought, reflect frozen accidents in the early history of life—chance properties of the earliest organisms that were passed on by heredity and have become so deeply embedded in the constitution of all living cells that they cannot be changed without disastrous effects.

The information in the sequence of a messenger RNA molecule is read out in groups of three nucleotides at a time: each triplet of nucleotides, or *codon*, specifies (codes for) a single amino acid in a corresponding protein. Since there are 64 ($= 4 \times 4 \times 4$) possible codons, all of which occur in nature, but only 20 amino acids, there are necessarily many cases in which several codons correspond to the same amino acid. The code is read out by a special class of small RNA molecules, the **transfer RNAs** (**tRNAs**). Each type of tRNA becomes attached at one end to a specific amino acid, and displays at its other end a specific sequence of three nucleotides—an *anticodon*—that enables it to recognize, through base-pairing, a particular codon or subset of codons in mRNA (**Figure 1–9**).

For synthesis of protein, a succession of tRNA molecules charged with their appropriate amino acids have to be brought together with an mRNA molecule and matched up by base-pairing through their anticodons with each of its successive codons. The amino acids then have to be linked together to extend the growing protein chain, and the tRNAs, relieved of their burdens, have to be released. This whole complex of processes is carried out by a giant multimolecular machine, the ribosome, formed of two main chains of RNA, called **ribosomal RNAs**

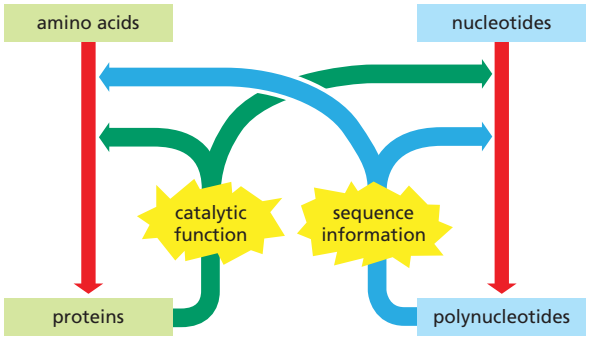


Figure 1–8 Life as an autocatalytic process. Polynucleotides (nucleotide polymers) and proteins (amino acid polymers) provide the sequence information and the catalytic functions that serve—through a complex set of chemical reactions—to bring about the synthesis of more polynucleotides and proteins of the same types.

(**rRNAs**), and more than 50 different proteins. This evolutionarily ancient molecular juggernaut latches onto the end of an mRNA molecule and then trundles along it, capturing loaded tRNA molecules and stitching together the amino acids they carry to form a new protein chain (**Figure 1–10**).

The Fragment of Genetic Information Corresponding to One Protein Is One Gene

DNA molecules as a rule are very large, containing the specifications for thousands of proteins. Individual segments of the entire DNA sequence are transcribed into separate mRNA molecules, with each segment coding for a different protein. Each such DNA segment represents one **gene**. A complication is that RNA molecules transcribed from the same DNA segment can often be processed in more than one way, so as to give rise to a set of alternative versions of a protein, especially in more complex cells such as those of plants and animals. A gene therefore is defined, more generally, as the segment of DNA sequence corresponding to a single protein or set of alternative protein variants (or to a single catalytic or structural RNA molecule for those genes that produce RNA but not protein).

In all cells, the *expression* of individual genes is regulated: instead of manufacturing its full repertoire of possible proteins at full tilt all the time, the cell adjusts the rate of transcription and translation of different genes independently, according to need. Stretches of *regulatory DNA* are interspersed among the segments

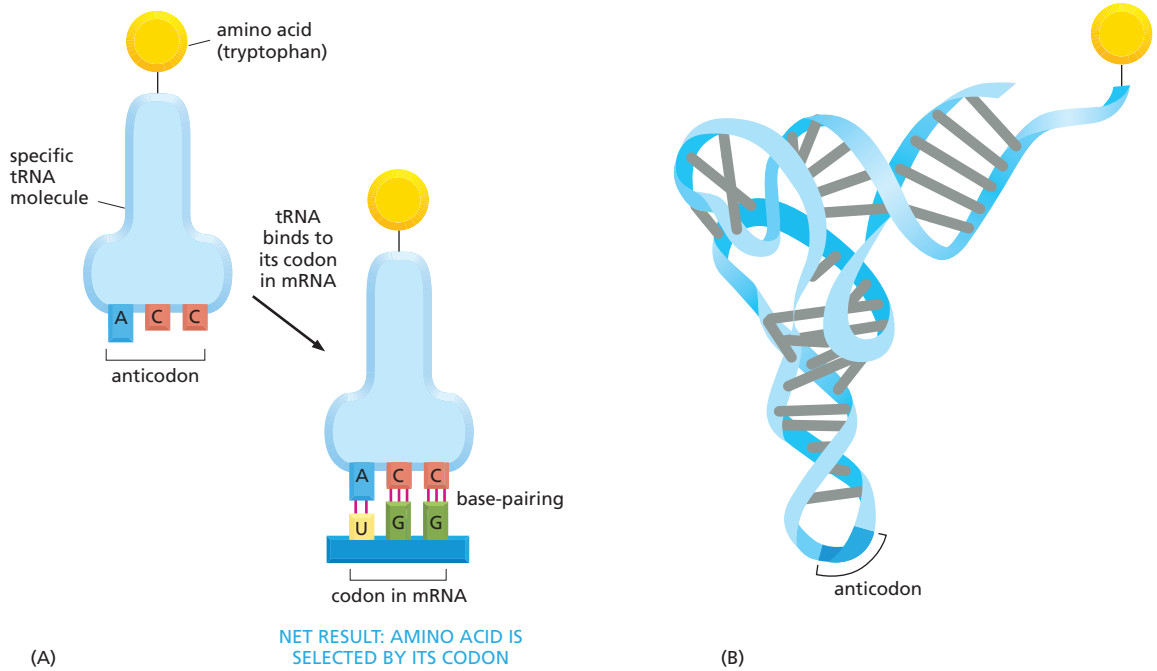


Figure 1–9 Transfer RNA. (A) A tRNA molecule specific for the amino acid tryptophan. One end of the tRNA molecule has tryptophan attached to it, while the other end displays the triplet nucleotide sequence CCA (its anticodon), which recognizes the tryptophan codon in messenger RNA molecules. (B) The three-dimensional structure of the tryptophan tRNA molecule. Note that the codon and the anticodon in (A) are in antiparallel orientations, like the two strands in a DNA double helix (see Figure 1–2), so that the sequence of the anticodon in the tRNA is read from right to left, while that of the codon in the mRNA is read from left to right.

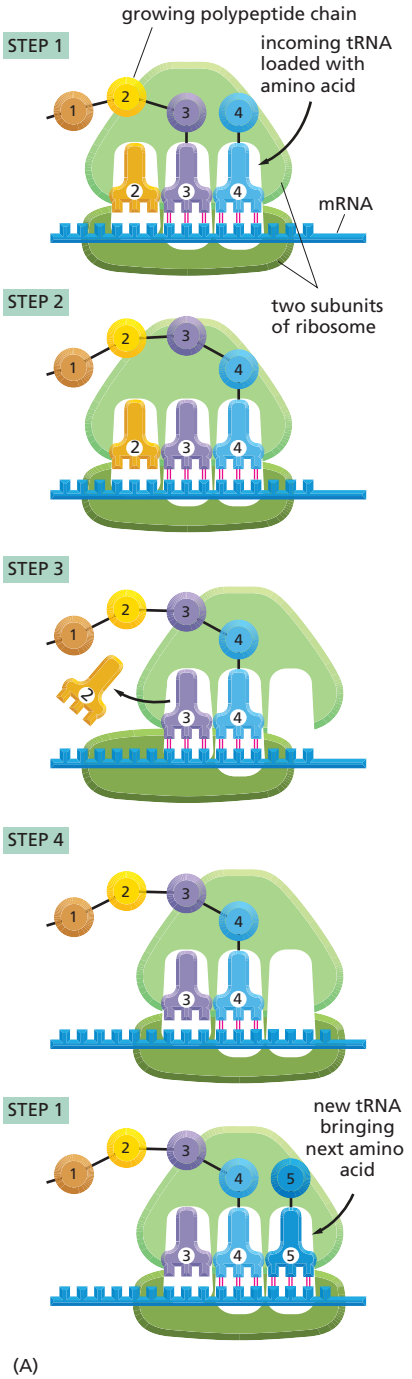


Figure 1–10 A ribosome at work. (A) The diagram shows how a ribosome moves along an mRNA molecule, capturing tRNA molecules that match the codons in the mRNA and using them to join amino acids into a protein chain. The mRNA specifies the sequence of amino acids. (B) The three-dimensional structure of a bacterial ribosome *(pale green* and *blue),* moving along an mRNA molecule *(orange* beads*),* with three tRNA molecules *(yellow*, *green*, and *pink)* at different stages in their process of capture and release. The ribosome is a giant assembly of more than 50 individual protein and RNA molecules. (B, courtesy of Joachim Frank, Yanhong Li and Rajendra Agarwal.)

that code for protein, and these noncoding regions bind to special protein molecules that control the local rate of transcription (**Figure 1–11**). Other noncoding DNA is also present, some of it serving, for example, as punctuation, defining where the information for an individual protein begins and ends. The quantity and organization of the regulatory and other noncoding DNA vary widely from one class of organisms to another, but the basic strategy is universal. In this way, the **genome** of the cell—that is, the total of its genetic information as embodied in its complete DNA sequence—dictates not only the nature of the cell's proteins, but also when and where they are to be made.

Life Requires Free Energy

A living cell is a dynamic chemical system, operating far from chemical equilibrium. For a cell to grow or to make a new cell in its own image, it must take in free energy from the environment, as well as raw materials, to drive the necessary synthetic reactions. This consumption of free energy is fundamental to life. When it stops, a cell decays towards chemical equilibrium and soon dies.

Genetic information is also fundamental to life. Is there any connection? The answer is yes: free energy is required for the propagation of information. For example, to specify one bit of information—that is, one yes/no choice between two equally probable alternatives—costs a defined amount of free energy that can be calculated. The quantitative relationship involves some deep reasoning and depends on a precise definition of the term "free energy," discussed in Chapter 2. The basic idea, however, is not difficult to understand intuitively.

Picture the molecules in a cell as a swarm of objects endowed with thermal energy, moving around violently at random, buffeted by collisions with one another. To specify genetic information—in the form of a DNA sequence, for example—molecules from this wild crowd must be captured, arranged in a specific order defined by some preexisting template, and linked together in a fixed relationship. The bonds that hold the molecules in their proper places on the template and join them together must be strong enough to resist the disordering effect of thermal motion. The process is driven forward by consumption of free energy, which is needed to ensure that the correct bonds are made, and made robustly. In the simplest case, the molecules can be compared with spring-loaded traps, ready to snap into a more stable, lower-energy attached state when they meet their proper partners; as they snap together into the bonded arrangement, their available stored energy—their free energy—like the energy of the spring in the trap, is released and dissipated as heat. In a cell, the chemical processes underlying information transfer are more complex, but the same basic principle applies: free energy has to be spent on the creation of order.

To replicate its genetic information faithfully, and indeed to make all its complex molecules according to the correct specifications, the cell therefore requires free energy, which has to be imported somehow from the surroundings.

All Cells Function as Biochemical Factories Dealing with the Same Basic Molecular Building Blocks

Because all cells make DNA, RNA, and protein, and these macromolecules are composed of the same set of subunits in every case, all cells have to contain and

manipulate a similar collection of small molecules, including simple sugars, nucleotides, and amino acids, as well as other substances that are universally required for their synthesis. All cells, for example, require the phosphorylated nucleotide *ATP (adenosine triphosphate)* as a building block for the synthesis of DNA and RNA; and all cells also make and consume this molecule as a carrier of free energy and phosphate groups to drive many other chemical reactions.

Although all cells function as biochemical factories of a broadly similar type, many of the details of their small-molecule transactions differ, and it is not as easy as it is for the informational macromolecules to point out the features that are strictly universal. Some organisms, such as plants, require only the simplest of nutrients and harness the energy of sunlight to make from these almost all their own small organic molecules; other organisms, such as animals, feed on living things and obtain many of their organic molecules ready-made. We return to this point below.

All Cells Are Enclosed in a Plasma Membrane Across Which Nutrients and Waste Materials Must Pass

There is, however, at least one other feature of cells that is universal: each one is enclosed by a membrane—the **plasma membrane**. This container acts as a selective barrier that enables the cell to concentrate nutrients gathered from its environment and retain the products it synthesizes for its own use, while excreting its waste products. Without a plasma membrane, the cell could not maintain its integrity as a coordinated chemical system.

The molecules forming this membrane have the simple physico-chemical property of being *amphiphilic*—that is, consisting of one part that is hydrophobic (water-insoluble) and another part that is hydrophilic (water-soluble). Such molecules placed in water aggregate spontaneously, arranging their hydrophobic portions to be as much in contact with one another as possible to hide them from the water, while keeping their hydrophilic portions exposed. Amphiphilic molecules of appropriate shape, such as the phospholipid molecules that comprise most of the plasma membrane, spontaneously aggregate in water to form a *bilayer* that creates small closed vesicles (**Figure 1–12**). The phenomenon can be demonstrated in a test tube by simply mixing phospholipids and water together; under appropriate conditions, small vesicles form whose aqueous contents are isolated from the external medium.

Although the chemical details vary, the hydrophobic tails of the predominant membrane molecules in all cells are hydrocarbon polymers ($–CH_2–CH_2–CH_2–$), and their spontaneous assembly into a bilayered vesicle is but one of many examples of an important general principle: cells produce molecules whose chemical properties cause them to *self-assemble* into the structures that a cell needs.

The cell boundary cannot be totally impermeable. If a cell is to grow and reproduce, it must be able to import raw materials and export waste across its plasma membrane. All cells therefore have specialized proteins embedded in their membrane that transport specific molecules from one side to the other (**Figure 1–13**). Some of these *membrane transport proteins,* like some of the proteins that catalyze the fundamental small-molecule reactions inside the cell,

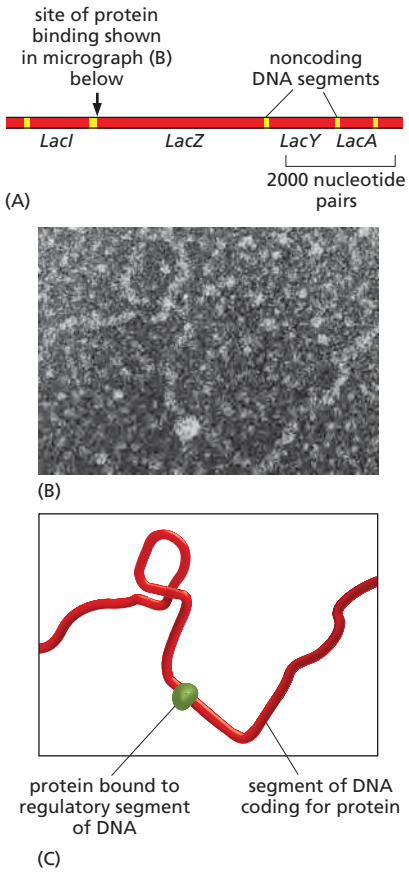


Figure 1–11 Gene regulation by protein binding to regulatory DNA. (A) A diagram of a small portion of the genome of the bacterium *Escherichia coli,* containing genes (called *LacI, LacZ, LacY,* and *LacA)* coding for four different proteins. The protein-coding DNA segments *(red)* have regulatory and other noncoding DNA segments *(yellow)* between them. (B) An electron micrograph of DNA from this region, with a protein molecule (encoded by the *LacI* gene) bound to the regulatory segment; this protein controls the rate of transcription of the *LacZ*, *LacY*, and *LacA* genes. (C) A drawing of the structures shown in (B). (B, courtesy of Jack Griffith.)

have been so well preserved over the course of evolution that we can recognize the family resemblances between them in comparisons of even the most distantly related groups of living organisms.

The transport proteins in the membrane largely determine which molecules enter the cell, and the catalytic proteins inside the cell determine the reactions that those molecules undergo. Thus, by specifying the proteins that the cell is to manufacture, the genetic information recorded in the DNA sequence dictates the entire chemistry of the cell; and not only its chemistry, but also its form and its behavior, for these too are chiefly constructed and controlled by the cell's proteins.

A Living Cell Can Exist with Fewer Than 500 Genes

The basic principles of biological information transfer are simple enough, but how complex are real living cells? In particular, what are the minimum requirements? We can get a rough indication by considering a species that has one of the smallest known genomes—the bacterium *Mycoplasma genitalium* (**Figure 1–14**). This organism lives as a parasite in mammals, and its environment provides it with many of its small molecules ready-made. Nevertheless, it still has to make all the large molecules—DNA, RNAs, and proteins—required for the basic processes of heredity. It has only about 480 genes in its genome of 580,070 nucleotide pairs, representing 145,018 bytes of information—about as much as it takes to record the text of one chapter of this book. Cell biology may be complicated, but it is not impossibly so.

The minimum number of genes for a viable cell in today's environments is probably not less than 200–300, although there are only about 60 genes in the core set shared by all living species without any known exception.

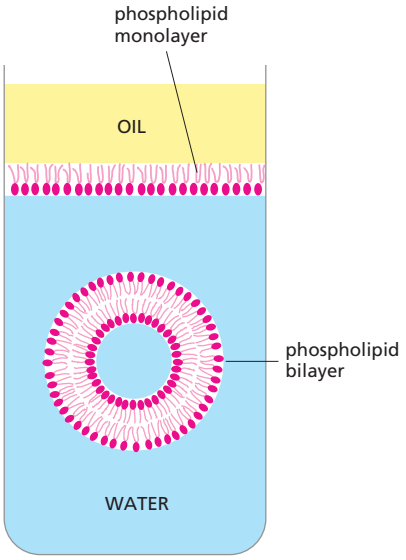


Figure 1–12 Formation of a membrane by amphiphilic phospholipid molecules. These have a hydrophilic (water-loving, phosphate) head group and a hydrophobic (water-avoiding, hydrocarbon) tail. At an interface between oil and water, they arrange themselves as a single sheet with their head groups facing the water and their tail groups facing the oil. When immersed in water, they aggregate to form bilayers enclosing aqueous compartments.

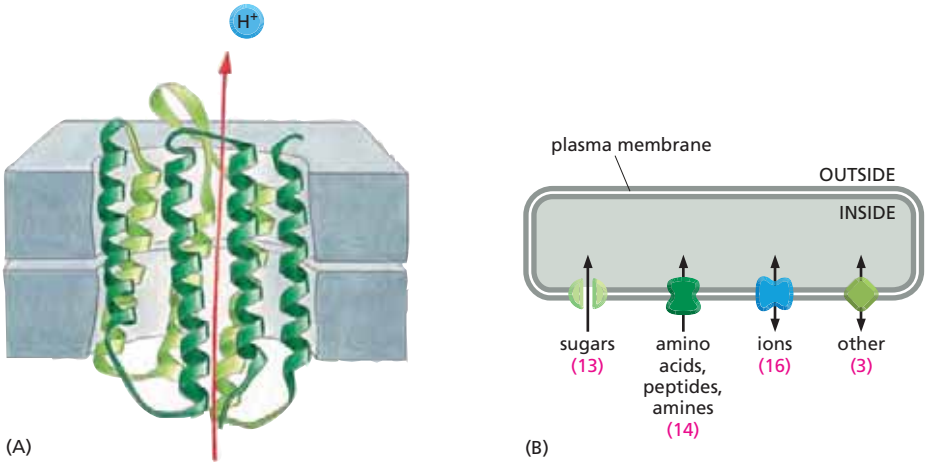


Figure 1–13 Membrane transport proteins. (A) Structure of a molecule of bacteriorhodopsin, from the archaeon (archaebacterium) *Halobacterium halobium*. This transport protein uses the energy of absorbed light to pump protons (H^+ ions) out of the cell. The polypeptide chain threads to and fro across the membrane; in several regions it is twisted into a helical conformation, and the helical segments are arranged to form the walls of a channel through which ions are transported. (B) Diagram of the set of transport proteins found in the membrane of the bacterium *Thermotoga maritima*. The numbers in parentheses refer to the number of different membrane transport proteins of each type. Most of the proteins within each class are evolutionarily related to one another and to their counterparts in other species.

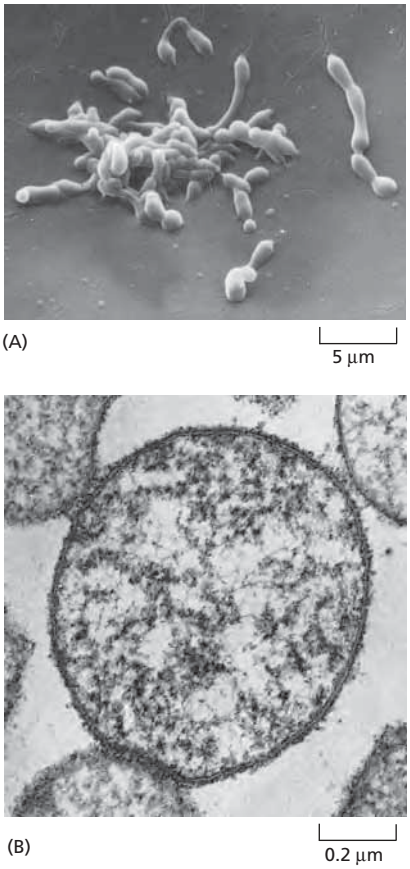


Figure 1–14 Mycoplasma genitalium. (A) Scanning electron micrograph showing the irregular shape of this small bacterium, reflecting the lack of any rigid wall. (B) Cross section (transmission electron micrograph) of a *Mycoplasma* cell. Of the 477 genes of *Mycoplasma genitalium*, 37 code for transfer, ribosomal, and other nonmessenger RNAs. Functions are known, or can be guessed, for 297 of the genes coding for protein: of these, 153 are involved in replication, transcription, translation, and related processes involving DNA, RNA, and protein; 29 in the membrane and surface structures of the cell; 33 in the transport of nutrients and other molecules across the membrane; 71 in energy conversion and the synthesis and degradation of small molecules; and 11 in the regulation of cell division and other processes. (A, from S. Razin et al., *Infect. Immun.* 30:538–546, 1980. With permission from the American Society for Microbiology; B, courtesy of Roger Cole, in Medical Microbiology, 4th ed. [S. Baron ed.]. Galveston: University of Texas Medical Branch, 1996.)

Summary

Living organisms reproduce themselves by transmitting genetic information to their progeny. The individual cell is the minimal self-reproducing unit, and is the vehicle for transmission of the genetic information in all living species. Every cell on our planet stores its genetic information in the same chemical form—as double-stranded DNA. The cell replicates its information by separating the paired DNA strands and using each as a template for polymerization to make a new DNA strand with a complementary sequence of nucleotides. The same strategy of templated polymerization is used to transcribe portions of the information from DNA into molecules of the closely related polymer, RNA. These in turn guide the synthesis of protein molecules by the more complex machinery of translation, involving a large multimolecular machine, the ribosome, which is itself composed of RNA and protein. Proteins are the principal catalysts for almost all the chemical reactions in the cell; their other functions include the selective import and export of small molecules across the plasma membrane that forms the cell's boundary. The specific function of each protein depends on its amino acid sequence, which is specified by the nucleotide sequence of a corresponding segment of the DNA—the gene that codes for that protein. In this way, the genome of the cell determines its chemistry; and the chemistry of every living cell is fundamentally similar, because it must provide for the synthesis of DNA, RNA, and protein. The simplest known cells have just under 500 genes.

THE DIVERSITY OF GENOMES AND THE TREE OF LIFE

The success of living organisms based on DNA, RNA, and protein, out of the infinitude of other chemical forms that we might conceive of, has been spectacular. They have populated the oceans, covered the land, infiltrated the Earth's crust, and molded the surface of our planet. Our oxygen-rich atmosphere, the deposits of coal and oil, the layers of iron ores, the cliffs of chalk and limestone and marble—all these are products, directly or indirectly, of past biological activity on Earth.

Living things are not confined to the familiar temperate realm of land, water, and sunlight inhabited by plants and plant-eating animals. They can be found in the darkest depths of the ocean, in hot volcanic mud, in pools beneath the frozen surface of the Antarctic, and buried kilometers deep in the Earth's crust. The creatures that live in these extreme environments are generally unfamiliar, not only because they are inaccessible, but also because they are mostly microscopic. In more homely habitats, too, most organisms are too small for us to see without special equipment: they tend to go unnoticed, unless they cause a disease or rot the timbers of our houses. Yet microorganisms make up most of the

total mass of living matter on our planet. Only recently, through new methods of molecular analysis and specifically through the analysis of DNA sequences, have we begun to get a picture of life on Earth that is not grossly distorted by our biased perspective as large animals living on dry land.

In this section we consider the diversity of organisms and the relationships among them. Because the genetic information for every organism is written in the universal language of DNA sequences, and the DNA sequence of any given organism can be obtained by standard biochemical techniques, it is now possible to characterize, catalogue, and compare any set of living organisms with reference to these sequences. From such comparisons we can estimate the place of each organism in the family tree of living species—the 'tree of life'. But before describing what this approach reveals, we need first to consider the routes by which cells in different environments obtain the matter and energy they require to survive and proliferate, and the ways in which some classes of organisms depend on others for their basic chemical needs.

Cells Can Be Powered by a Variety of Free Energy Sources

Living organisms obtain their free energy in different ways. Some, such as animals, fungi, and the bacteria that live in the human gut, get it by feeding on other living things or the organic chemicals they produce; such organisms are called *organotrophic* (from the Greek word *trophe*, meaning "food"). Others derive their energy directly from the nonliving world. These fall into two classes: those that harvest the energy of sunlight, and those that capture their energy from energy-rich systems of inorganic chemicals in the environment (chemical systems that are far from chemical equilibrium). Organisms of the former class are called *phototrophic* (feeding on sunlight); those of the latter are called *lithotrophic* (feeding on rock). Organotrophic organisms could not exist without these primary energy converters, which are the most plentiful form of life.

Phototrophic organisms include many types of bacteria, as well as algae and plants, on which we—and virtually all the living things that we ordinarily see around us—depend. Phototrophic organisms have changed the whole chemistry of our environment: the oxygen in the Earth's atmosphere is a by-product of their biosynthetic activities.

Lithotrophic organisms are not such an obvious feature of our world, because they are microscopic and mostly live in habitats that humans do not frequent—deep in the ocean, buried in the Earth's crust, or in various other inhospitable environments. But they are a major part of the living world, and are especially important in any consideration of the history of life on Earth.

Some lithotrophs get energy from *aerobic* reactions, which use molecular oxygen from the environment; since atmospheric O_2 is ultimately the product of living organisms, these aerobic lithotrophs are, in a sense, feeding on the products of past life. There are, however, other lithotrophs that live anaerobically, in places where little or no molecular oxygen is present, in circumstances similar to those that must have existed in the early days of life on Earth, before oxygen had accumulated.

The most dramatic of these sites are the hot *hydrothermal vents* found deep down on the floor of the Pacific and Atlantic Oceans, in regions where the ocean floor is spreading as new portions of the Earth's crust form by a gradual upwelling of material from the Earth's interior (**Figure 1–15**). Downward-percolating seawater is heated and driven back upward as a submarine geyser, carrying with it a current of chemicals from the hot rocks below. A typical cocktail might include H_2S, H_2, CO, Mn^{2+}, Fe^{2+}, Ni^{2+}, CH_2, NH_4^+, and phosphorus-containing compounds. A dense population of microbes lives in the neighborhood of the vent, thriving on this austere diet and harvesting free energy from reactions between the available chemicals. Other organisms—clams, mussels, and giant marine worms—in turn live off the microbes at the vent, forming an entire ecosystem analogous to the system of plants and animals that we belong to, but powered by geochemical energy instead of light (**Figure 1–16**).

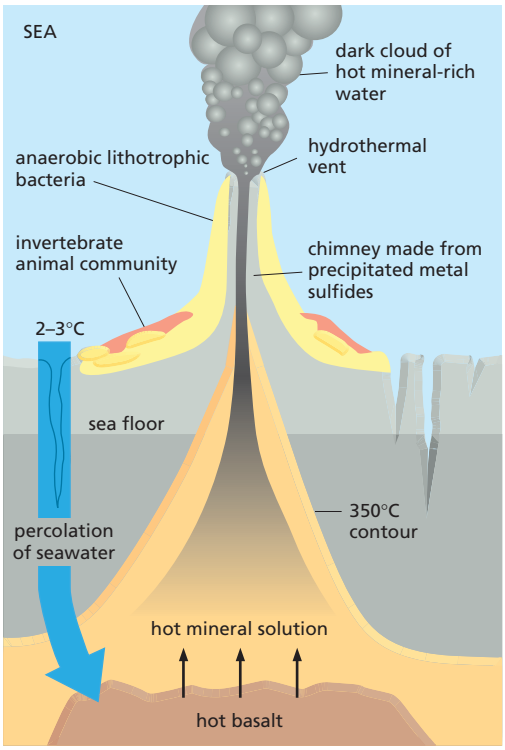


Figure 1–15 The geology of a hot hydrothermal vent in the ocean floor. Water percolates down toward the hot molten rock upwelling from the Earth's interior and is heated and driven back upward, carrying minerals leached from the hot rock. A temperature gradient is set up, from more than 350°C near the core of the vent, down to 2–3°C in the surrounding ocean. Minerals precipitate from the water as it cools, forming a chimney. Different classes of organisms, thriving at different temperatures, live in different neighborhoods of the chimney. A typical chimney might be a few meters tall, with a flow rate of 1–2 m/sec.

Some Cells Fix Nitrogen and Carbon Dioxide for Others

To make a living cell requires matter, as well as free energy. DNA, RNA, and protein are composed of just six elements: hydrogen, carbon, nitrogen, oxygen, sulfur, and phosphorus. These are all plentiful in the nonliving environment, in the Earth's rocks, water, and atmosphere, but not in chemical forms that allow easy incorporation into biological molecules. Atmospheric N_2 and CO_2, in particular, are extremely unreactive, and a large amount of free energy is required to drive the reactions that use these inorganic molecules to make the organic compounds needed for further biosynthesis—that is, to *fix* nitrogen and carbon dioxide, so as to make N and C available to living organisms. Many types of living cells lack the biochemical machinery to achieve this fixation, and rely on other classes of cells to do the job for them. We animals depend on plants for our supplies of

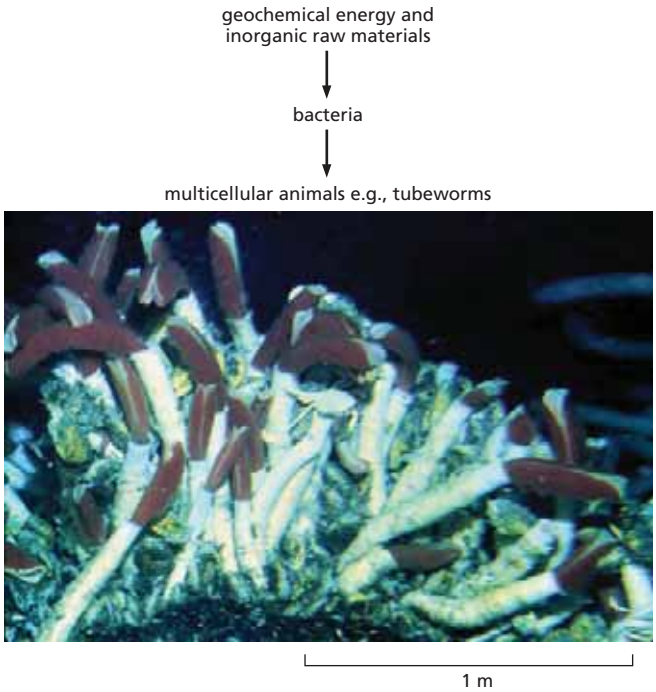


Figure 1–16 Living organisms at a hot hydrothermal vent. Close to the vent, at temperatures up to about 120°C, various lithotrophic species of bacteria and archaea (archaebacteria) live, directly fuelled by geochemical energy. A little further away, where the temperature is lower, various invertebrate animals live by feeding on these microorganisms. Most remarkable are the giant (2-meter) tube worms, which, rather than feed on the lithotrophic cells, live in symbiosis with them: specialized organs in the worms harbor huge numbers of symbiotic sulfur-oxidizing bacteria. These bacteria harness geochemical energy and supply nourishment to their hosts, which have no mouth, gut, or anus. The dependence of the tube worms on the bacteria for the harnessing of geothermal energy is analogous to the dependence of plants on chloroplasts for the harnessing of solar energy, discussed later in this chapter. The tube worms, however, are thought to have evolved from more conventional animals, and to have become secondarily adapted to life at hydrothermal vents. (Courtesy of Dudley Foster, Woods Hole Oceanographic Institution.)

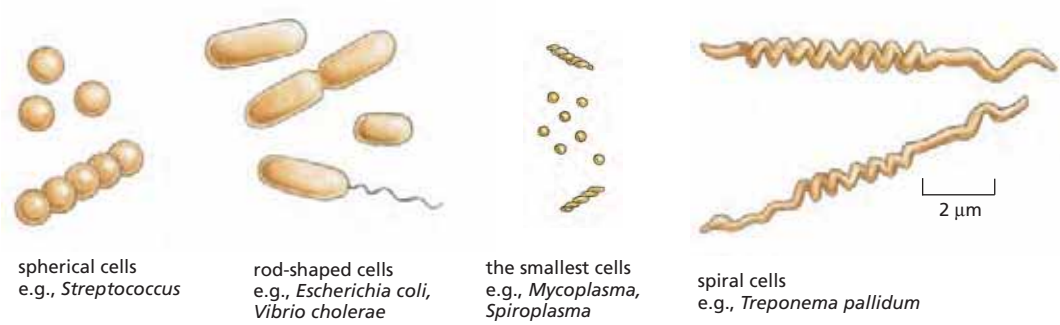


Figure 1–17 Shapes and sizes of some bacteria. Although most are small, as shown, measuring a few micrometers in linear dimension, there are also some giant species. An extreme example (not shown) is the cigar-shaped bacterium *Epulopiscium fishelsoni*, which lives in the gut of a surgeonfish and can be up to 600 μm long.

organic carbon and nitrogen compounds. Plants in turn, although they can fix carbon dioxide from the atmosphere, lack the ability to fix atmospheric nitrogen, and they depend in part on nitrogen-fixing bacteria to supply their need for nitrogen compounds. Plants of the pea family, for example, harbor symbiotic nitrogen-fixing bacteria in nodules in their roots.

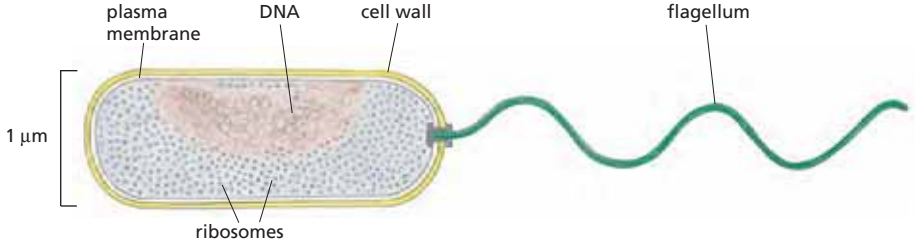
Living cells therefore differ widely in some of the most basic aspects of their biochemistry. Not surprisingly, cells with complementary needs and capabilities have developed close associations. Some of these associations, as we see below, have evolved to the point where the partners have lost their separate identities altogether: they have joined forces to form a single composite cell.

The Greatest Biochemical Diversity Exists Among Procaryotic Cells

From simple microscopy, it has long been clear that living organisms can be classified on the basis of cell structure into two groups: the **eucaryotes** and the **procaryotes**. Eucaryotes keep their DNA in a distinct membrane-enclosed intracellular compartment called the nucleus. (The name is from the Greek, meaning "truly nucleated," from the words *eu*, "well" or "truly," and *karyon*, "kernel" or "nucleus".) Procaryotes have no distinct nuclear compartment to house their DNA. Plants, fungi, and animals are eucaryotes; bacteria are procaryotes, as are archaea—a separate class of procaryotic cells, discussed below.

Most procaryotic cells are small and simple in outward appearance (**Figure 1–17**), and they live mostly as independent individuals or in loosely organized communities, rather than as multicellular organisms. They are typically spherical or rod-shaped and measure a few micrometers in linear dimension. They often have a tough protective coat, called a *cell wall,* beneath which a plasma membrane encloses a single cytoplasmic compartment containing DNA, RNA, proteins, and the many small molecules needed for life. In the electron microscope, this cell interior appears as a matrix of varying texture without any discernible organized internal structure (**Figure 1–18**).

Figure 1–18 The structure of a bacterium. (A) The bacterium *Vibrio cholerae,* showing its simple internal organization. Like many other species, *Vibrio* has a helical appendage at one end—a flagellum—that rotates as a propeller to drive the cell forward. (B) An electron micrograph of a longitudinal section through the widely studied bacterium *Escherichia coli (E. coli)*. This is related to *Vibrio* but has many flagella (not visible in this section) distributed over its surface. The cell's DNA is concentrated in the lightly stained region. (B, courtesy of E. Kellenberger.)



(A)

(B)

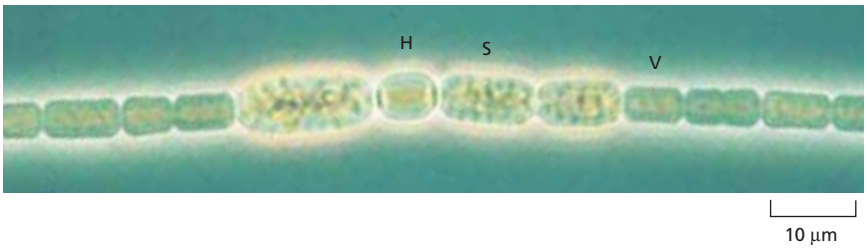


Figure 1–19 The phototrophic bacterium *Anabaena cylindrica* viewed in the light microscope. The cells of this species form long, multicellular filaments. Most of the cells (labeled *V)* perform photosynthesis, while others become specialized for nitrogen fixation (labeled *H),* or develop into resistant spores (labeled *S).* (Courtesy of Dave G. Adams.)

Procaryotic cells live in an enormous variety of ecological niches, and they are astonishingly varied in their biochemical capabilities—far more so than eucaryotic cells. Organotrophic species can utilize virtually any type of organic molecule as food, from sugars and amino acids to hydrocarbons and methane gas. Phototrophic species (**Figure 1–19**) harvest light energy in a variety of ways, some of them generating oxygen as a byproduct, others not. Lithotrophic species can feed on a plain diet of inorganic nutrients, getting their carbon from CO_2, and relying on H_2S to fuel their energy needs (**Figure 1–20**)—or on H_2, or Fe^{2+}, or elemental sulfur, or any of a host of other chemicals that occur in the environment.

Many parts of this world of microscopic organisms are virtually unexplored. Traditional methods of bacteriology have given us an acquaintance with those species that can be isolated and cultured in the laboratory. But DNA sequence analysis of the populations of bacteria in samples from natural habitats—such as soil or ocean water, or even the human mouth—has opened our eyes to the fact that most species cannot be cultured by standard laboratory techniques. According to one estimate, at least 99% of procaryotic species remain to be characterized.

The Tree of Life Has Three Primary Branches: Bacteria, Archaea, and Eucaryotes

The classification of living things has traditionally depended on comparisons of their outward appearances: we can see that a fish has eyes, jaws, backbone, brain, and so on, just as we do, and that a worm does not; that a rosebush is cousin to an apple tree, but less similar to a grass. As Darwin showed, we can readily interpret such close family resemblances in terms of evolution from common ancestors, and we can find the remains of many of these ancestors preserved in the fossil record. In this way, it has been possible to begin to draw a family tree of living organisms, showing the various lines of descent, as well as branch points in the history, where the ancestors of one group of species became different from those of another.

When the disparities between organisms become very great, however, these methods begin to fail. How do we decide whether a fungus is closer kin to a plant or to an animal? When it comes to procaryotes, the task becomes harder still: one microscopic rod or sphere looks much like another. Microbiologists have therefore sought to classify procaryotes in terms of their biochemistry and nutritional requirements. But this approach also has its pitfalls. Amid the bewildering variety of biochemical behaviors, it is difficult to know which differences truly reflect differences of evolutionary history.

Genome analysis has given us a simpler, more direct, and more powerful way to determine evolutionary relationships. The complete DNA sequence of an organism defines its nature with almost perfect precision and in exhaustive detail. Moreover, this specification is in a digital form—a string of letters—that can be entered straightforwardly into a computer and compared with the corresponding information for any other living thing. Because DNA is subject to random changes that accumulate over long periods of time (as we shall see shortly), the number of differences between the DNA sequences of two organisms can provide a direct, objective, quantitative indication of the evolutionary distance between them.

This approach has shown that the organisms that were traditionally classed together as "bacteria" can be as widely divergent in their evolutionary origins as

Figure 1–20 A lithotrophic bacterium. *Beggiatoa*, which lives in sulfurous environments, gets its energy by oxidizing H_2S and can fix carbon even in the dark. Note the yellow deposits of sulfur inside the cells. (Courtesy of Ralph W. Wolfe.)

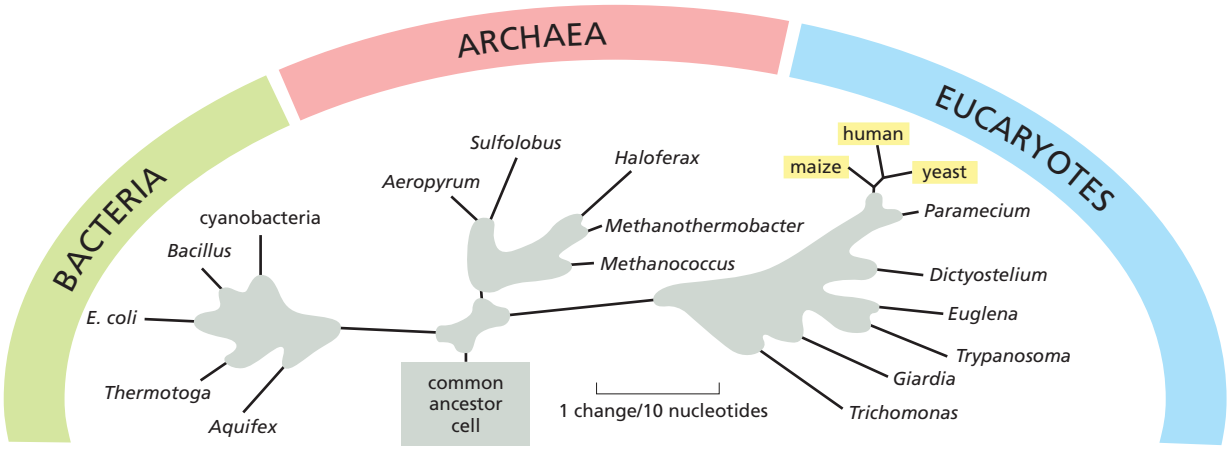


Figure 1–21 The three major divisions (domains) of the living world. Note that traditionally the word *bacteria* has been used to refer to procaryotes in general, but more recently has been redefined to refer to eubacteria specifically. The tree shown here is based on comparisons of the nucleotide sequence of a ribosomal RNA subunit in the different species, and the distances in the diagram represent estimates of the numbers of evolutionary changes that have occurred in this molecule in each lineage (see Figure 1–22). The parts of the tree shrouded in *gray cloud* represent uncertainties about details of the true pattern of species divergence in the course of evolution: comparisons of nucleotide or amino acid sequences of molecules other than rRNA, as well as other arguments, lead to somewhat different trees. There is general agreement, however, as to the early divergence of the three most basic domains—the bacteria, the archaea, and the eucaryotes.

is any procaryote from any eucaryote. It now appears that the procaryotes comprise two distinct groups that diverged early in the history of life on Earth, either before the ancestors of the eucaryotes diverged as a separate group or at about the same time. The two groups of procaryotes are called the **bacteria** (or eubacteria) and the **archaea** (or archaebacteria). The living world therefore has three major divisions or *domains*: bacteria, archaea, and eucaryotes (**Figure 1–21**).

Archaea are often found inhabiting environments that we humans avoid, such as bogs, sewage treatment plants, ocean depths, salt brines, and hot acid springs, although they are also widespread in less extreme and more homely environments, from soils and lakes to the stomachs of cattle. In outward appearance they are not easily distinguished from bacteria. At a molecular level, archaea seem to resemble eucaryotes more closely in their machinery for handling genetic information (replication, transcription, and translation), but bacteria more closely in their apparatus for metabolism and energy conversion. We discuss below how this might be explained.

Some Genes Evolve Rapidly; Others Are Highly Conserved

Both in the storage and in the copying of genetic information, random accidents and errors occur, altering the nucleotide sequence—that is, creating **mutations**. Therefore, when a cell divides, its two daughters are often not quite identical to one another or to their parent. On rare occasions, the error may represent a change for the better; more probably, it will cause no significant difference in the cell's prospects; and in many cases, the error will cause serious damage—for example, by disrupting the coding sequence for a key protein. Changes due to mistakes of the first type will tend to be perpetuated, because the altered cell has an increased likelihood of reproducing itself. Changes due to mistakes of the second type—*selectively neutral* changes—may be perpetuated or not: in the competition for limited resources, it is a matter of chance whether the altered cell or its cousins will succeed. But changes that cause serious damage lead nowhere: the cell that suffers them dies, leaving no progeny. Through endless repetition of this cycle of error and trial—of *mutation* and *natural selection*—

organisms evolve: their genetic specifications change, giving them new ways to exploit the environment more effectively, to survive in competition with others, and to reproduce successfully.

Clearly, some parts of the genome change more easily than others in the course of evolution. A segment of DNA that does not code for protein and has no significant regulatory role is free to change at a rate limited only by the frequency of random errors. In contrast, a gene that codes for a highly optimized essential protein or RNA molecule cannot alter so easily: when mistakes occur, the faulty cells are almost always eliminated. Genes of this latter sort are therefore *highly conserved*. Through 3.5 billion years or more of evolutionary history, many features of the genome have changed beyond all recognition; but the most highly conserved genes remain perfectly recognizable in all living species.

These latter genes are the ones we must examine if we wish to trace family relationships between the most distantly related organisms in the tree of life. The studies that led to the classification of the living world into the three domains of bacteria, archaea, and eucaryotes were based chiefly on analysis of one of the two main RNA components of the ribosome—the so-called small-subunit ribosomal RNA. Because translation is fundamental to all living cells, this component of the ribosome has been well conserved since early in the history of life on Earth (**Figure 1–22**).

Most Bacteria and Archaea Have 1000–6000 Genes

Natural selection has generally favored those procaryotic cells that can reproduce the fastest by taking up raw materials from their environment and replicating themselves most efficiently, at the maximal rate permitted by the available food supplies. Small size implies a large ratio of surface area to volume, thereby helping to maximize the uptake of nutrients across the plasma membrane and boosting a cell's reproductive rate.

Presumably for these reasons, most procaryotic cells carry very little superfluous baggage; their genomes are small, with genes packed closely together and minimal quantities of regulatory DNA between them. The small genome size makes it relatively easy to determine the complete DNA sequence. We now have this information for many species of bacteria and archaea, and a few species of eucaryotes. As shown in **Table 1–1**, most bacterial and archaeal genomes contain between 10^6 and 10^7 nucleotide pairs, encoding 1000–6000 genes.

A complete DNA sequence reveals both the genes an organism possesses and the genes it lacks. When we compare the three domains of the living world, we can begin to see which genes are common to all of them and must therefore have been present in the cell that was ancestral to all present-day living things, and which genes are peculiar to a single branch in the tree of life. To explain the findings, however, we need to consider a little more closely how new genes arise and genomes evolve.

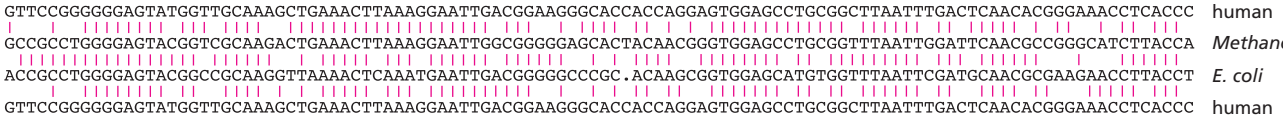


Figure 1–22 Genetic information conserved since the days of the last common ancestor of all living things. A part of the gene for the smaller of the two main RNA components of the ribosome is shown. (The complete molecule is about 1500–1900 nucleotides long, depending on species.) Corresponding segments of nucleotide sequence from an archaean *(Methanococcus jannaschii),* a bacterium *(Escherichia coli)* and a eucaryote *(Homo sapiens)* are aligned. Sites where the nucleotides are identical between species are indicated by a vertical line; the human sequence is repeated at the bottom of the alignment so that all three two-way comparisons can be seen. A dot halfway along the *E. coli* sequence denotes a site where a nucleotide has been either deleted from the bacterial lineage in the course of evolution, or inserted in the other two lineages. Note that the sequences from these three organisms, representative of the three domains of the living world, all differ from one another to a roughly similar degree, while still retaining unmistakable similarities.

Table 1–1 Some Genomes That Have Been Completely Sequenced

SPECIES	SPECIAL FEATURES	HABITAT	GENOME SIZE (1000s OF NUCLEOTIDE PAIRS PER HAPLOID GENOME)	ESTIMATED NUMBER OF GENES CODING FOR PROTEINS
BACTERIA				
Mycoplasma genitalium	has one of the smallest of all known cell genomes	human genital tract	580	468
Synechocystis sp.	photosynthetic, oxygen-generating (cyanobacterium)	lakes and streams	3573	3168
Escherichia coli	laboratory favorite	human gut	4639	4289
Helicobacter pylori	causes stomach ulcers and predisposes to stomach cancer	human stomach	1667	1590
Bacillus anthracis	causes anthrax	soil	5227	5634
Aquifex aeolicus	lithotrophic; lives at high temperatures	hydrothermal vents	1551	1544
Streptomyces coelicolor	source of antibiotics; giant genome	soil	8667	7825
Treponema pallidum	spirochete; causes syphilis	human tissues	1138	1041
Rickettsia prowazekii	bacterium most closely related to mitochondria; causes typhus	lice and humans (intracellular parasite)	1111	834
Thermotoga maritima	organotrophic; lives at very high temperatures	hydrothermal vents	1860	1877
ARCHAEA				
Methanococcus jannaschii	lithotrophic, anaerobic, methane-producing	hydrothermal vents	1664	1750
Archaeoglobus fulgidus	lithotrophic or organotrophic, anaerobic, sulfate-reducing	hydrothermal vents	2178	2493
Nanoarchaeum equitans	smallest known archaean; anaerobic; parasitic on another, larger archaean	hydrothermal and volcanic hot vents	491	552
EUCARYOTES				
Saccharomyces cerevisiae (budding yeast)	minimal model eucaryote	grape skins, beer	12,069	~6300
Arabidopsis thaliana (Thale cress)	model organism for flowering plants	soil and air	~142,000	~26,000
Caenorhabditis elegans (nematode worm)	simple animal with perfectly predictable development	soil	~97,000	~20,000
Drosophila melanogaster (fruit fly)	key to the genetics of animal development	rotting fruit	~137,000	~14,000
Homo sapiens (human)	most intensively studied mammal	houses	~3,200,000	~24,000

Genome size and gene number vary between strains of a single species, especially for bacteria and archaea. The table shows data for particular strains that have been sequenced. For eucaryotes, many genes can give rise to several alternative variant proteins, so that the total number of proteins specified by the genome is substantially greater than the number of genes.

New Genes Are Generated from Preexisting Genes

The raw material of evolution is the DNA sequence that already exists: there is no natural mechanism for making long stretches of new random sequence. In this sense, no gene is ever entirely new. Innovation can, however, occur in several ways (**Figure 1–23**):

1. *Intragenic mutation*: an existing gene can be modified by changes in its DNA sequence, through various types of error that occur mainly in the process of DNA replication.
2. *Gene duplication*: an existing gene can be duplicated so as to create a pair of initially identical genes within a single cell; these two genes may then diverge in the course of evolution.

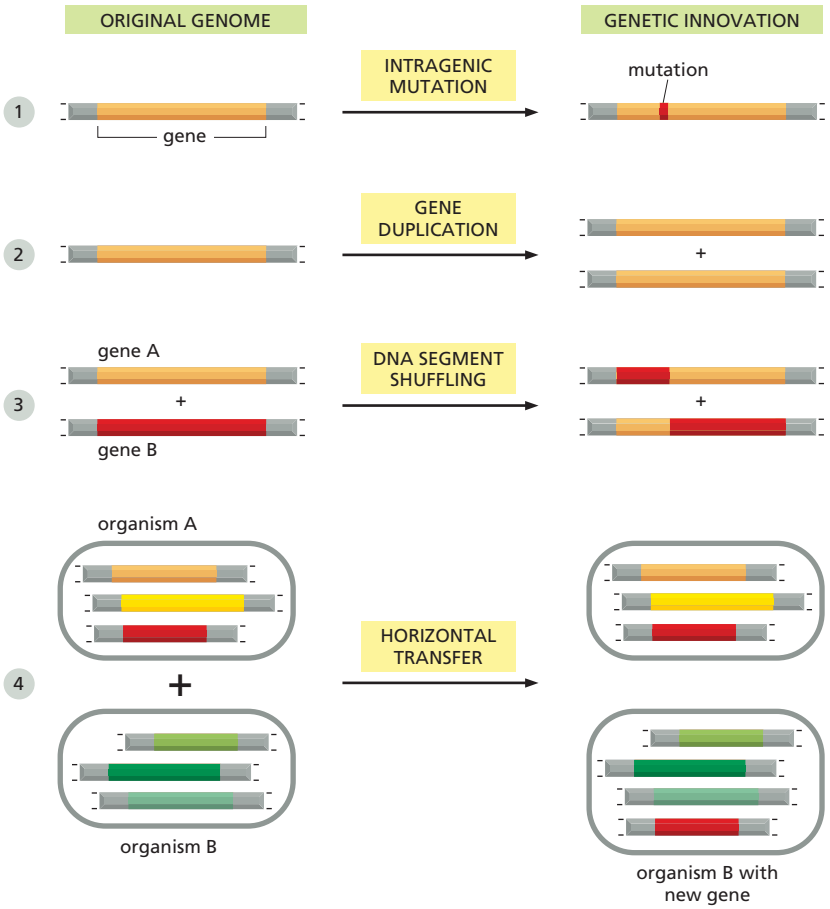


Figure 1–23 Four modes of genetic innovation and their effects on the DNA sequence of an organism. A special form of horizontal transfer occurs when two different types of cells enter into a permanent symbiotic association. Genes from one of the cells then may be transferred to the genome of the other, as we shall see below when we discuss mitochondria and chloroplasts.

3. *Segment shuffling*: two or more existing genes can be broken and rejoined to make a hybrid gene consisting of DNA segments that originally belonged to separate genes.
4. *Horizontal (intercellular) transfer*: a piece of DNA can be transferred from the genome of one cell to that of another—even to that of another species. This process is in contrast with the usual *vertical transfer* of genetic information from parent to progeny.

Each of these types of change leaves a characteristic trace in the DNA sequence of the organism, providing clear evidence that all four processes have occurred. In later chapters we discuss the underlying mechanisms, but for the present we focus on the consequences.

Gene Duplications Give Rise to Families of Related Genes Within a Single Cell

A cell duplicates its entire genome each time it divides into two daughter cells. However, accidents occasionally result in the inappropriate duplication of just part of the genome, with retention of original and duplicate segments in a single cell. Once a gene has been duplicated in this way, one of the two gene copies is free to mutate and become specialized to perform a different function within the same cell. Repeated rounds of this process of duplication and divergence, over many millions of years, have enabled one gene to give rise to a family of genes that may all be found within a single genome. Analysis of the DNA sequence of procaryotic genomes reveals many examples of such gene families: in *Bacillus subtilis*, for example, 47% of the genes have one or more obvious relatives (**Figure 1–24**).

When genes duplicate and diverge in this way, the individuals of one species become endowed with multiple variants of a primordial gene. This evolutionary

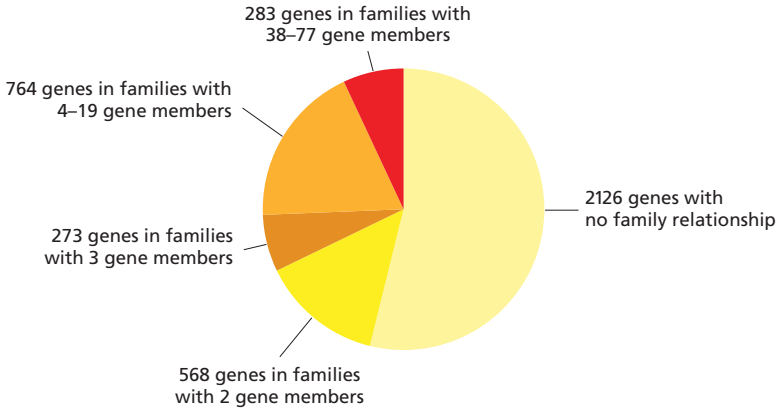


Figure 1–24 Families of evolutionarily related genes in the genome of *Bacillus subtilis*. The biggest family consists of 77 genes coding for varieties of ABC transporters—a class of membrane transport proteins found in all three domains of the living world. (Adapted from F. Kunst et al., *Nature* 390:249–256, 1997. With permission from Macmillan Publishers Ltd.)

process has to be distinguished from the genetic divergence that occurs when one species of organism splits into two separate lines of descent at a branch point in the family tree—when the human line of descent became separate from that of chimpanzees, for example. There, the genes gradually become different in the course of evolution, but they are likely to continue to have corresponding functions in the two sister species. Genes that are related by descent in this way—that is, genes in two separate species that derive from the same ancestral gene in the last common ancestor of those two species—are called **orthologs**. Related genes that have resulted from a gene duplication event within a single genome—and are likely to have diverged in their function—are called **paralogs**. Genes that are related by descent in either way are called **homologs**, a general term used to cover both types of relationship (**Figure 1–25**).

The family relationships between genes can become quite complex (**Figure 1–26**). For example, an organism that possesses a family of paralogous genes (for example, the seven hemoglobin genes α, β, γ, δ, ε, ζ, and θ) may evolve into two separate species (such as humans and chimpanzees) each possessing the entire set of paralogs. All 14 genes are homologs, with the human hemoglobin α orthologous to the chimpanzee hemoglobin α, but paralogous to the human or chimpanzee hemoglobin β, and so on. Moreover, the vertebrate hemoglobins (the oxygen-binding proteins of blood) are homologous to the vertebrate myoglobins (the oxygen-binding proteins of muscle), as well as to more distant

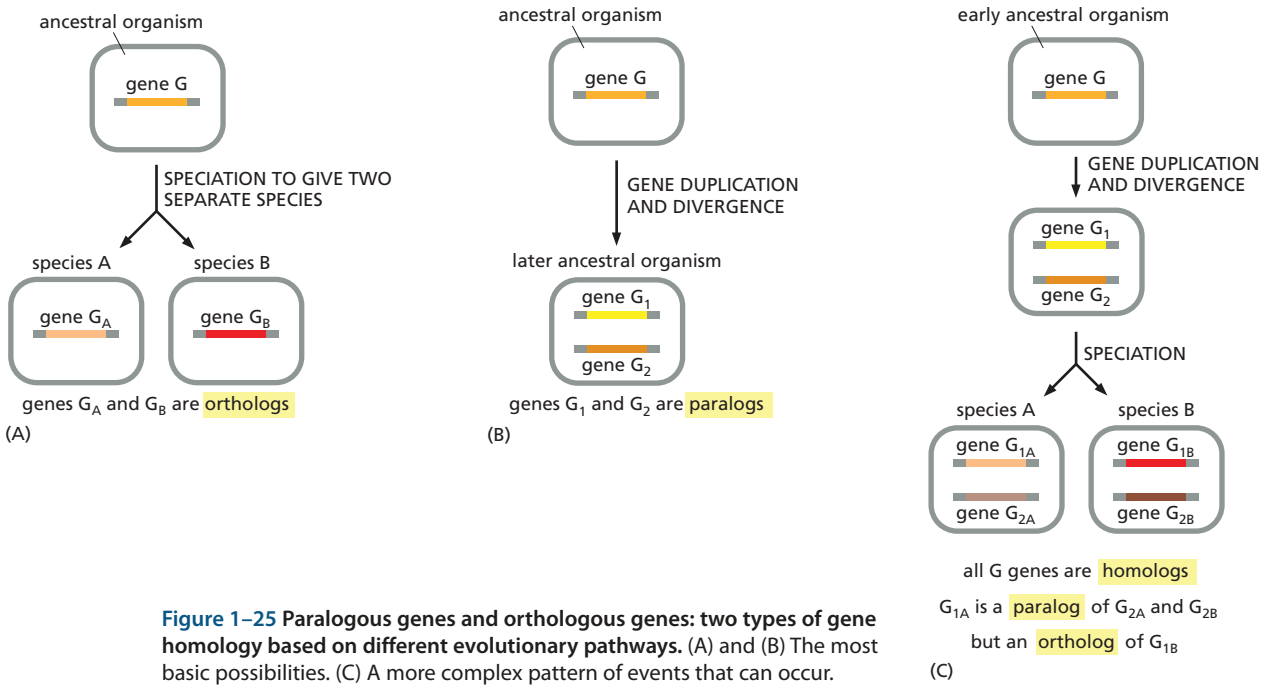


Figure 1–25 Paralogous genes and orthologous genes: two types of gene homology based on different evolutionary pathways. (A) and (B) The most basic possibilities. (C) A more complex pattern of events that can occur.

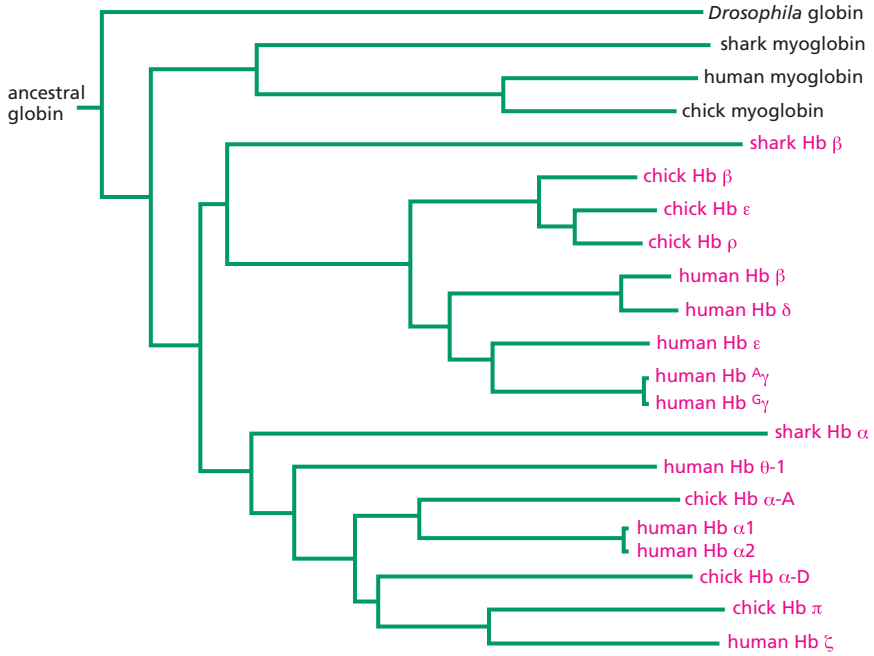


Figure 1–26 A complex family of homologous genes. This diagram shows the pedigree of the hemoglobin (Hb), myoglobin, and globin genes of human, chick, shark, and *Drosophila*. The lengths of the horizontal lines represent the amount of divergence in amino acid sequence.

genes that code for oxygen-binding proteins in invertebrates, plants, fungi, and bacteria. From the DNA sequences, it is usually easy to recognize that two genes in different species are homologous; it is much more difficult to decide, without other information, whether they stand in the precise evolutionary relationship of orthologs.

Genes Can Be Transferred Between Organisms, Both in the Laboratory and in Nature

Procaryotes also provide examples of the horizontal transfer of genes from one species of cell to another. The most obvious tell-tale signs are sequences recognizable as being derived from bacterial viruses, also called *bacteriophages* (**Figure 1–27**). **Viruses** are not themselves living cells but can act as vectors for gene transfer: they are small packets of genetic material that have evolved as parasites on the reproductive and biosynthetic machinery of host cells. They replicate in one cell, emerge from it with a protective wrapping, and then enter and infect another cell, which may be of the same or a different species. Often, the infected cell will be killed by the massive proliferation of virus particles inside it; but sometimes, the viral DNA, instead of directly generating these particles, may persist in its host for many cell generations as a relatively innocuous passenger, either as a separate intracellular fragment of DNA, known as a *plasmid,* or as a sequence inserted into the cell's regular genome. In their travels, viruses can accidentally pick up fragments of DNA from the genome of one host cell and ferry them into another cell. Such transfers of genetic material frequently occur in procaryotes, and they can also occur between eucaryotic cells of the same species.

Horizontal transfers of genes between eucaryotic cells of different species are very rare, and they do not seem to have played a significant part in eucaryote evolution (although massive transfers from bacterial to eucaryotic genomes have occurred in the evolution of mitochondria and chloroplasts, as we discuss below). In contrast, horizontal gene transfers occur much more frequently between different species of procaryotes. Many procaryotes have a remarkable capacity to take up even nonviral DNA molecules from their surroundings and thereby capture the genetic information these molecules carry. By this route, or by virus-mediated transfer, bacteria and archaea in the wild can acquire genes from neighboring cells relatively easily. Genes that confer resistance to an

antibiotic or an ability to produce a toxin, for example, can be transferred from species to species and provide the recipient bacterium with a selective advantage. In this way, new and sometimes dangerous strains of bacteria have been observed to evolve in the bacterial ecosystems that inhabit hospitals or the various niches in the human body. For example, horizontal gene transfer is responsible for the spread, over the past 40 years, of penicillin-resistant strains of *Neisseria gonorrheae,* the bacterium that causes gonorrhea. On a longer time scale, the results can be even more profound; it has been estimated that at least 18% of all of the genes in the present-day genome of *E. coli* have been acquired by horizontal transfer from another species within the past 100 million years.

Sex Results in Horizontal Exchanges of Genetic Information Within a Species

Horizontal exchanges of genetic information are important in bacterial and archaeal evolution in today's world, and they may have occurred even more frequently and promiscuously in the early days of life on Earth. Such early horizontal exchanges could explain the otherwise puzzling observation that the eucaryotes seem more similar to archaea in their genes for the basic information-handling processes of DNA replication, transcription, and translation, but more similar to bacteria in their genes for metabolic processes. In any case, whether horizontal gene transfer occurred most freely in the early days of life on Earth, or has continued at a steady low rate throughout evolutionary history, it has the effect of complicating the whole concept of cell ancestry, by making each cell's genome a composite of parts derived from separate sources.

Horizontal gene transfer among procaryotes may seem a surprising process, but it has a parallel in a phenomenon familiar to us all: sex. In addition to the usual vertical transfer of genetic material from parent to offspring, sexual reproduction causes a large-scale horizontal transfer of genetic information between two initially separate cell lineages—those of the father and the mother. A key feature of sex, of course, is that the genetic exchange normally occurs only between individuals of the same species. But no matter whether they occur within a species or between species, horizontal gene transfers leave a characteristic imprint: they result in individuals who are related more closely to one set of relatives with respect to some genes, and more closely to another set of relatives with respect to others. By comparing the DNA sequences of individual human genomes, an intelligent visitor from outer space could deduce that humans reproduce sexually, even if it knew nothing about human behavior.

Sexual reproduction is widespread (although not universal), especially among eucaryotes. Even bacteria indulge from time to time in controlled sexual exchanges of DNA with other members of their own species. Natural selection has clearly favored organisms that can reproduce sexually, although evolutionary theorists dispute precisely what the selective advantage of sex is.

The Function of a Gene Can Often Be Deduced from Its Sequence

Family relationships among genes are important not just for their historical interest, but because they simplify the task of deciphering gene functions. Once the sequence of a newly discovered gene has been determined, a scientist can tap a few keys on a computer to search the entire database of known gene sequences for genes related to it. In many cases, the function of one or more of these homologs will have been already determined experimentally, and thus, since gene sequence determines gene function, one can frequently make a good guess at the function of the new gene: it is likely to be similar to that of the already-known homologs.

In this way, it is possible to decipher a great deal of the biology of an organism simply by analyzing the DNA sequence of its genome and using the information we already have about the functions of genes in other organisms that have been more intensively studied.

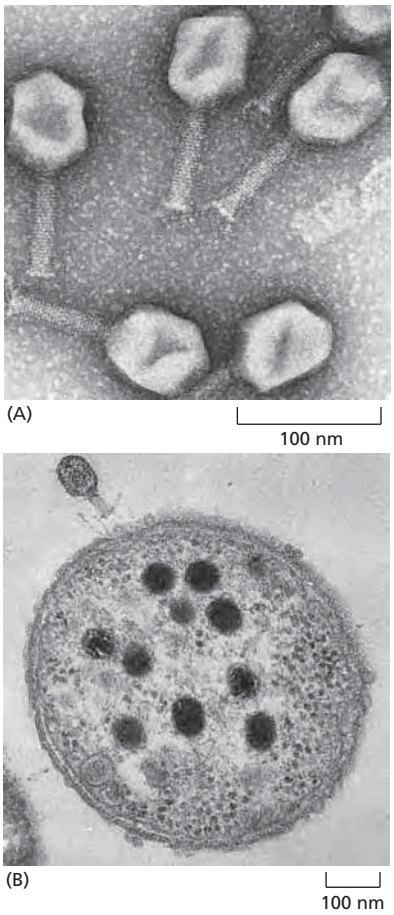


Figure 1–27 The viral transfer of DNA from one cell to another. (A) An electron micrograph of particles of a bacterial virus, the T4 bacteriophage. The head of this virus contains the viral DNA; the tail contains the apparatus for injecting the DNA into a host bacterium. (B) A cross section of a bacterium with a T4 bacteriophage latched onto its surface. The large dark objects inside the bacterium are the heads of new T4 particles in course of assembly. When they are mature, the bacterium will burst open to release them. (A, courtesy of James Paulson; B, courtesy of Jonathan King and Erika Hartwig from G. Karp, Cell and Molecular Biology, 2nd ed. New York: John Wiley & Sons, 1999. With permission from John Wiley & Sons.)

More Than 200 Gene Families Are Common to All Three Primary Branches of the Tree of Life

Given the complete genome sequences of representative organisms from all three domains—archaea, bacteria, and eucaryotes—we can search systematically for homologies that span this enormous evolutionary divide. In this way we can begin to take stock of the common inheritance of all living things. There are considerable difficulties in this enterprise. For example, individual species have often lost some of the ancestral genes; other genes have almost certainly been acquired by horizontal transfer from another species and therefore are not truly ancestral, even though shared. In fact, genome comparisons strongly suggest that both lineage-specific gene loss and horizontal gene transfer, in some cases between evolutionarily distant species, have been major factors of evolution, at least among procaryotes. Finally, in the course of 2 or 3 billion years, some genes that were initially shared will have changed beyond recognition by current methods.

Because of all these vagaries of the evolutionary process, it seems that only a small proportion of ancestral gene families have been universally retained in a recognizable form. Thus, out of 4873 protein-coding gene families defined by comparing the genomes of 50 species of bacteria, 13 archaea, and 3 unicellular eucaryotes, only 63 are truly ubiquitous (that is, represented in all the genomes analyzed). The great majority of these universal families include components of the translation and transcription systems. This is not likely to be a realistic approximation of an ancestral gene set. A better—though still crude—idea of the latter can be obtained by tallying the gene families that have representatives in multiple, but not necessarily all, species from all three major domains. Such an analysis reveals 264 ancient conserved families. Each family can be assigned a function (at least in terms of general biochemical activity, but usually with more precision), with the largest number of shared gene families being involved in translation and in amino acid metabolism and transport (**Table 1–2**). This set of highly conserved gene families represents only a very rough sketch of the common inheritance of all modern life; a more precise reconstruction of the gene complement of the last universal common ancestor might be feasible with further genome sequencing and more careful comparative analysis.

Mutations Reveal the Functions of Genes

Without additional information, no amount of gazing at genome sequences will reveal the functions of genes. We may recognize that gene B is like gene A, but how do we discover the function of gene A in the first place? And even if we know the function of gene A, how do we test whether the function of gene B is truly the same as the sequence similarity suggests? How do we connect the world of abstract genetic information with the world of real living organisms?

The analysis of gene functions depends on two complementary approaches: genetics and biochemistry. Genetics starts with the study of mutants: we either find or make an organism in which a gene is altered, and examine the effects on the organism's structure and performance (**Figure 1–28**). Biochemistry examines the functions of molecules: we extract molecules from an organism and then study their chemical activities. By combining genetics and biochemistry and examining the chemical abnormalities in a mutant organism, it is possible to find those molecules whose production depends on a given gene. At the same time, studies of the performance of the mutant organism show us what role those molecules have in the operation of the organism as a whole. Thus, genetics and biochemistry together provide a way to relate genes, molecules, and the structure and function of the organism.

In recent years, DNA sequence information and the powerful tools of molecular biology have allowed rapid progress. From sequence comparisons, we can often identify particular subregions within a gene that have been preserved nearly unchanged over the course of evolution. These conserved subregions are likely to be the most important parts of the gene in terms of function. We can test their individual contributions to the activity of the gene product by creating in

Table 1–2 The Numbers of Gene Families, Classified by Function, That Are Common to All Three Domains of the Living World

GENE FAMILY FUNCTION	NUMBER OF "UNIVERSAL" FAMILIES
Information processing	
Translation	63
Transcription	7
Replication, recombination, and repair	13
Cellular processes and signaling	
Cell cycle control, mitosis, and meiosis	2
Defense mechanisms	3
Signal transduction mechanisms	1
Cell wall/membrane biogenesis	2
Intracellular trafficking and secretion	4
Post-translational modification, protein turnover, chaperones	8
Metabolism	
Energy production and conversion	19
Carbohydrate transport and metabolism	16
Amino acid transport and metabolism	43
Nucleotide transport and metabolism	15
Coenzyme transport and metabolism	22
Lipid transport and metabolism	9
Inorganic ion transport and metabolism	8
Secondary metabolite biosynthesis, transport, and catabolism	5
Poorly characterized	
General biochemical function predicted; specific biological role unknown	24

For the purpose of this analysis, gene families are defined as "universal" if they are represented in the genomes of at least two diverse archaea *(Archaeoglobus fulgidus* and *Aeropyrum pernix),* two evolutionarily distant bacteria *(Escherichia coli* and *Bacillus subtilis)* and one eucaryote (yeast, *Saccharomyces cerevisiae).* (Data from R.L. Tatusov, E.V. Koonin and D.J. Lipman, *Science* 278:631–637, 1997, with permission from AAAS; R.L. Tatusov et al., *BMC Bioinformatics* 4:41, 2003, with permission from BioMed Central; and the COGs database at the US National Library of Medicine.)

the laboratory mutations of specific sites within the gene, or by constructing artificial hybrid genes that combine part of one gene with part of another. Organisms can be engineered to make either the RNA or the protein specified by the gene in large quantities to facilitate biochemical analysis. Specialists in molecular structure can determine the three-dimensional conformation of the gene product, revealing the exact position of every atom in it. Biochemists can determine how each of the parts of the genetically specified molecule contributes to its chemical behavior. Cell biologists can analyze the behavior of cells that are engineered to express a mutant version of the gene.

There is, however, no one simple recipe for discovering a gene's function, and no simple standard universal format for describing it. We may discover, for example, that the product of a given gene catalyzes a certain chemical reaction, and yet have no idea how or why that reaction is important to the organism. The functional characterization of each new family of gene products, unlike the description of the gene sequences, presents a fresh challenge to the biologist's ingenuity. Moreover, we never fully understand the function of a gene until we learn its role in the life of the organism as a whole. To make ultimate sense of gene functions, therefore, we have to study whole organisms, not just molecules or cells.

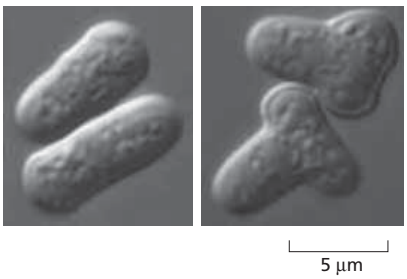


Figure 1–28 A mutant phenotype reflecting the function of a gene. A normal yeast (of the species *Schizosaccharomyces pombe)* is compared with a mutant in which a change in a single gene has converted the cell from a cigar shape *(left)* to a T shape *(right).* The mutant gene therefore has a function in the control of cell shape. But how, in molecular terms, does the gene product perform that function? That is a harder question, and needs biochemical analysis to answer it. (Courtesy of Kenneth Sawin and Paul Nurse.)

Molecular Biologists Have Focused a Spotlight on *E. coli*

Because living organisms are so complex, the more we learn about any particular species, the more attractive it becomes as an object for further study. Each

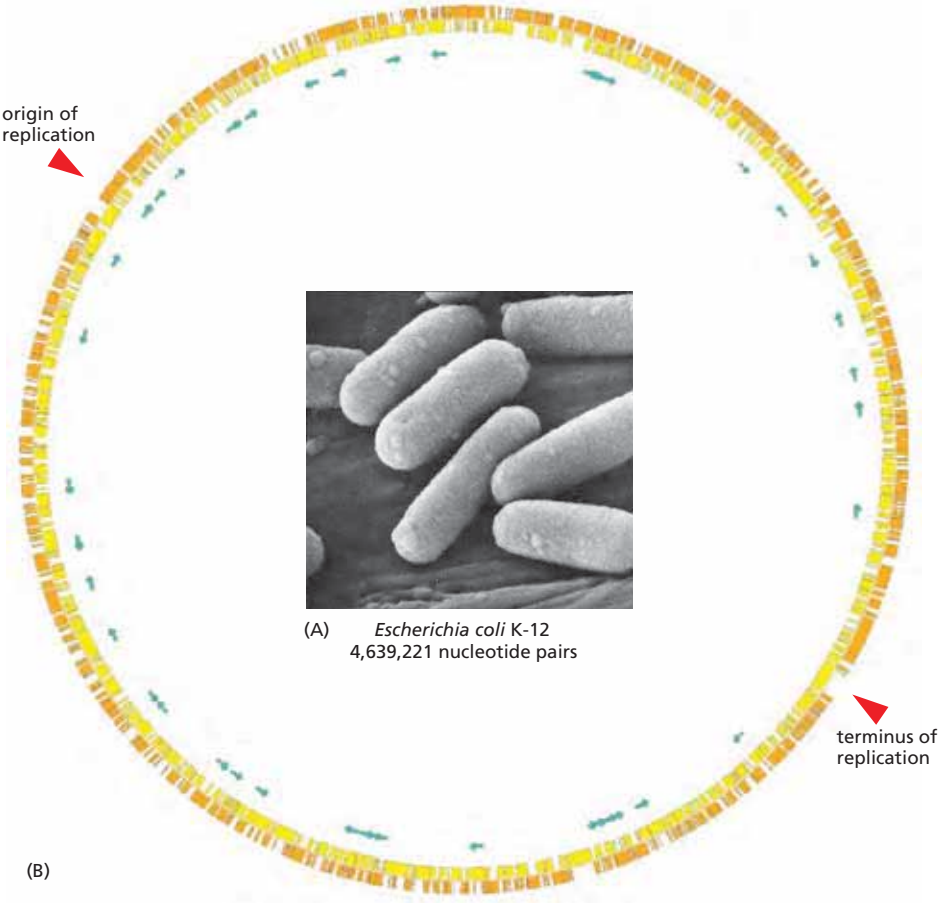


Figure 1–29 The genome of *E. coli*. (A) A cluster of *E. coli* cells. (B) A diagram of the genome of *E. coli* strain K-12. The diagram is circular because the DNA of *E. coli*, like that of other procaryotes, forms a single, closed loop. Protein-coding genes are shown as *yellow* or *orange bars*, depending on the DNA strand from which they are transcribed; genes encoding only RNA molecules are indicated by *green arrows*. Some genes are transcribed from one strand of the DNA double helix (in a clockwise direction in this diagram), others from the other strand (counterclockwise). (A, courtesy of Dr. Tony Brain and David Parker/Photo Researchers; B, adapted from F.R. Blattner et al., *Science* 277:1453–1462, 1997. With permission from AAAS.)

discovery raises new questions and provides new tools with which to tackle general questions in the context of the chosen organism. For this reason, large communities of biologists have become dedicated to studying different aspects of the same **model organism**.

In the enormously varied world of bacteria, the spotlight of molecular biology has for a long time focused intensely on just one species: *Escherichia coli*, or *E. coli* (see Figures 1–17 and 1–18). This small, rod-shaped bacterial cell normally lives in the gut of humans and other vertebrates, but it can be grown easily in a simple nutrient broth in a culture bottle. It adapts to variable chemical conditions and reproduces rapidly, and it can evolve by mutation and selection at a remarkable speed. As with other bacteria, different strains of *E. coli*, though classified as members of a single species, differ genetically to a much greater degree than do different varieties of a sexually reproducing organism such as a plant or animal. One *E. coli* strain may possess many hundreds of genes that are absent from another, and the two strains could have as little as 50% of their genes in common. The standard laboratory strain *E. coli* K-12 has a genome of approximately 4.6 million nucleotide pairs, contained in a single circular molecule of DNA, coding for about 4300 different kinds of proteins (**Figure 1–29**).

In molecular terms, we know more about *E. coli* than about any other living organism. Most of our understanding of the fundamental mechanisms of life—for example, how cells replicate their DNA, or how they decode the instructions represented in the DNA to direct the synthesis of specific proteins—has come from studies of *E. coli*. The basic genetic mechanisms have turned out to be highly conserved throughout evolution: these mechanisms are therefore essentially the same in our own cells as in *E. coli*.

Summary

Procaryotes (cells without a distinct nucleus) are biochemically the most diverse organisms and include species that can obtain all their energy and nutrients from inorganic chemical sources, such as the reactive mixtures of minerals released at hydrothermal vents on the ocean floor—the sort of diet that may have nourished the first living cells 3.5 billion years ago. DNA sequence comparisons reveal the family relationships of living organisms and show that the procaryotes fall into two groups that diverged early in the course of evolution: the bacteria (or eubacteria) and the archaea. Together with the eucaryotes (cells with a membrane-enclosed nucleus), these constitute the three primary branches of the tree of life. Most bacteria and archaea are small unicellular organisms with compact genomes comprising 1000–6000 genes. Many of the genes within a single organism show strong family resemblances in their DNA sequences, implying that they originated from the same ancestral gene through gene duplication and divergence. Family resemblances (homologies) are also clear when gene sequences are compared between different species, and more than 200 gene families have been so highly conserved that they can be recognized as common to most species from all three domains of the living world. Thus, given the DNA sequence of a newly discovered gene, it is often possible to deduce the gene's function from the known function of a homologous gene in an intensively studied model organism, such as the bacterium E. coli.

GENETIC INFORMATION IN EUCARYOTES

Eucaryotic cells, in general, are bigger and more elaborate than procaryotic cells, and their genomes are bigger and more elaborate, too. The greater size is accompanied by radical differences in cell structure and function. Moreover, many classes of eucaryotic cells form multicellular organisms that attain levels of complexity unmatched by any procaryote.

Because they are so complex, eucaryotes confront molecular biologists with a special set of challenges, which will concern us in the rest of this book. Increasingly, biologists meet these challenges through the analysis and manipulation of the genetic information within cells and organisms. It is therefore important at the outset to know something of the special features of the eucaryotic genome. We begin by briefly discussing how eucaryotic cells are organized, how this reflects their way of life, and how their genomes differ from those of procaryotes. This leads us to an outline of the strategy by which molecular biologists, by exploiting genetic information, are attempting to discover how eucaryotic organisms work.

Eucaryotic Cells May Have Originated as Predators

By definition, eucaryotic cells keep their DNA in an internal compartment called the nucleus. The *nuclear envelope,* a double layer of membrane, surrounds the nucleus and separates the DNA from the cytoplasm. Eucaryotes also have other features that set them apart from procaryotes (**Figure 1–30**). Their cells are, typically, 10 times bigger in linear dimension, and 1000 times larger in volume. They have a *cytoskeleton*—a system of protein filaments crisscrossing the cytoplasm and forming, together with the many proteins that attach to them, a system of girders, ropes, and motors that gives the cell mechanical strength, controls its shape, and drives and guides its movements. <GTTA> <ATGG> <TCGC> The nuclear envelope is only one part of a set of *internal membranes*, each structurally similar to the plasma membrane and enclosing different types of spaces inside the cell, many of them involved in digestion and secretion. Lacking the tough cell wall of most bacteria, animal cells and the free-living eucaryotic cells called *protozoa* can change their shape rapidly and engulf other cells and small objects by *phagocytosis* (**Figure 1–31**).

It is still a mystery how all these properties evolved, and in what sequence. One plausible view, however, is that they are all reflections of the way of life of a

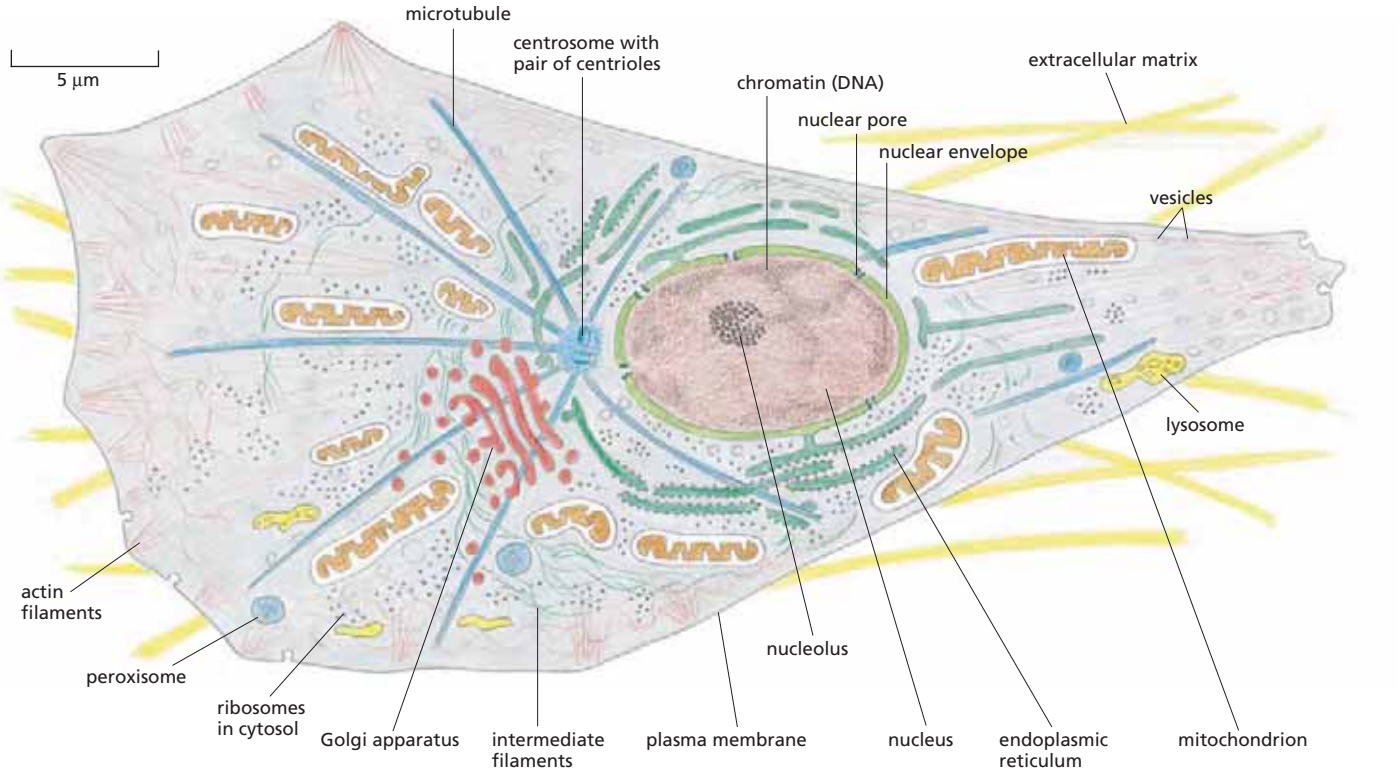


Figure 1–30 The major features of eucaryotic cells. The drawing depicts a typical animal cell, but almost all the same components are found in plants and fungi and in single-celled eucaryotes such as yeasts and protozoa. Plant cells contain chloroplasts in addition to the components shown here, and their plasma membrane is surrounded by a tough external wall formed of cellulose.

primordial eucaryotic cell that was a predator, living by capturing other cells and eating them (**Figure 1–32**). Such a way of life requires a large cell with a flexible plasma membrane, as well as an elaborate cytoskeleton to support and move this membrane. It may also require that the cell's long, fragile DNA molecules be sequestered in a separate nuclear compartment, to protect the genome from damage by the movements of the cytoskeleton.

Modern Eucaryotic Cells Evolved from a Symbiosis

A predatory way of life helps to explain another feature of eucaryotic cells. Almost all such cells contain *mitochondria* (**Figure 1–33**). These small bodies in the cytoplasm, enclosed by a double layer of membrane, take up oxygen and harness energy from the oxidation of food molecules—such as sugars—to produce most of the ATP that powers the cell's activities. Mitochondria are similar in size to small bacteria, and, like bacteria, they have their own genome in the form of a circular DNA molecule, their own ribosomes that differ from those elsewhere in the eucaryotic cell, and their own transfer RNAs. It is now generally accepted that mitochondria originated from free-living oxygen-metabolizing *(aerobic)* bacteria that were engulfed by an ancestral eucaryotic cell that could otherwise make no such use of oxygen (that is, was *anaerobic)*. Escaping digestion, these bacteria evolved in symbiosis with the engulfing cell and its progeny,

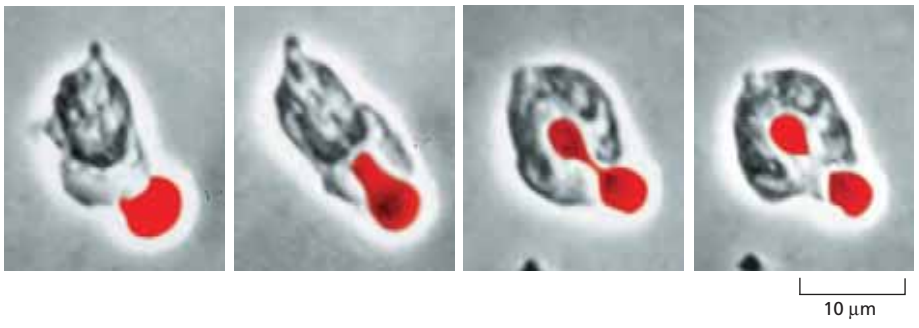


Figure 1–31 Phagocytosis. This series of stills from a movie shows a human white blood cell (a neutrophil) engulfing a red blood cell (artificially colored *red)* that has been treated with antibody. (Courtesy of Stephen E. Malawista and Anne de Boisfleury Chevance.)

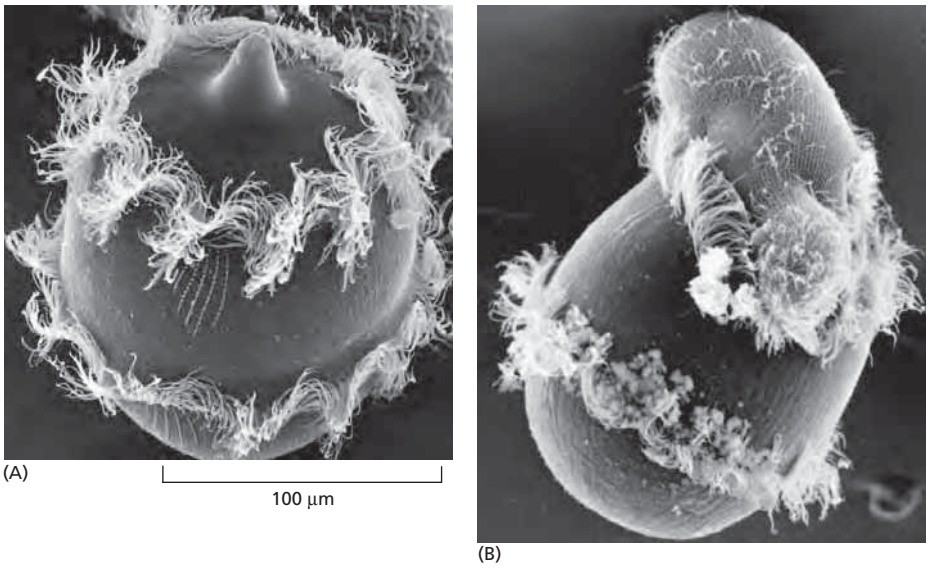


Figure 1–32 A single-celled eucaryote that eats other cells. (A) *Didinium* is a carnivorous protozoan, belonging to the group known as *ciliates*. It has a globular body, about 150 μm in diameter, encircled by two fringes of cilia—sinuous, whiplike appendages that beat continually; its front end is flattened except for a single protrusion, rather like a snout. (B) *Didinium* normally swims around in the water at high speed by means of the synchronous beating of its cilia. When it encounters a suitable prey, usually another type of protozoan, it releases numerous small paralyzing darts from its snout region. Then, the *Didinium* attaches to and devours the other cell by phagocytosis, inverting like a hollow ball to engulf its victim, which is almost as large as itself. (Courtesy of D. Barlow.)

receiving shelter and nourishment in return for the power generation they performed for their hosts (**Figure 1–34**). This partnership between a primitive anaerobic eucaryotic predator cell and an aerobic bacterial cell is thought to have been established about 1.5 billion years ago, when the Earth's atmosphere first became rich in oxygen.

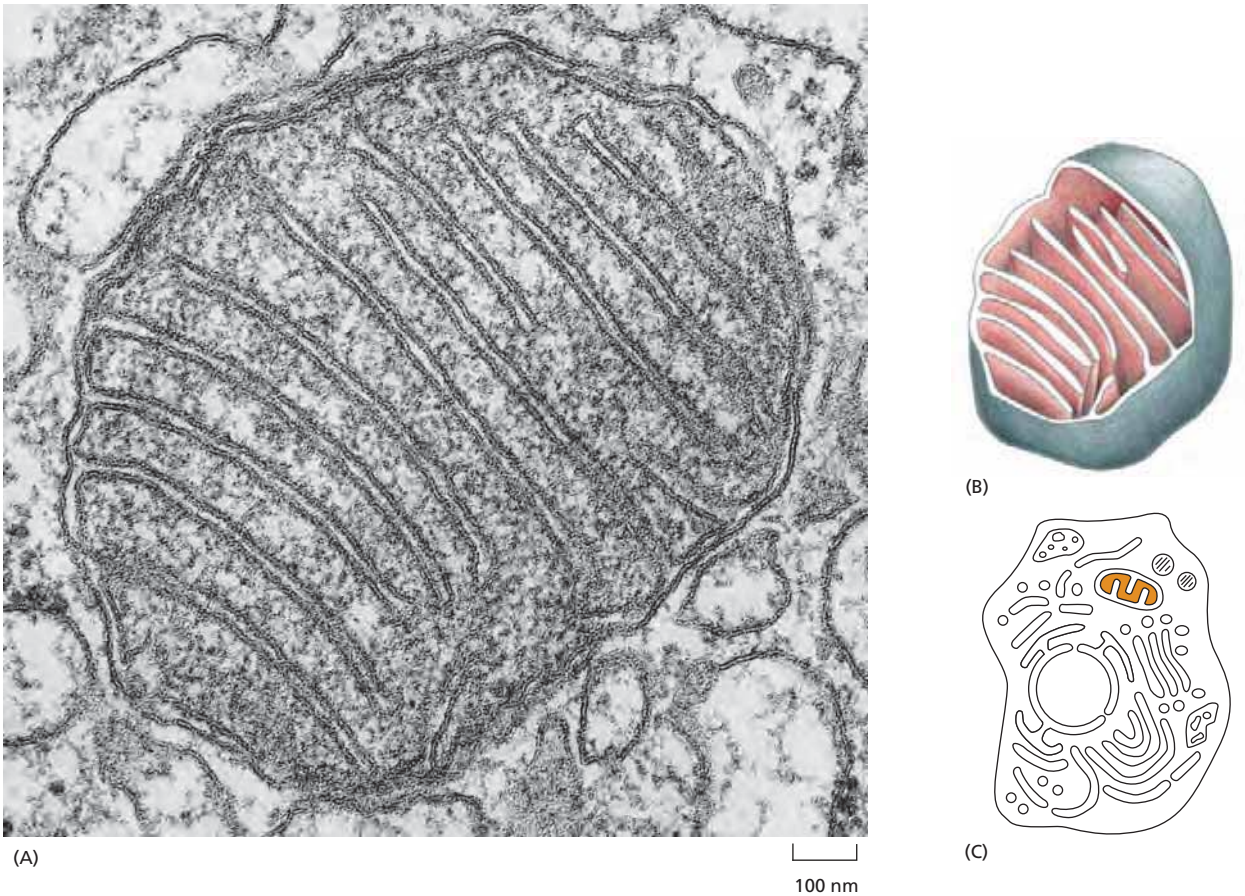


Figure 1–33 A mitochondrion. (A) A cross section, as seen in the electron microscope. (B) A drawing of a mitochondrion with part of it cut away to show the three-dimensional structure. (C) A schematic eucaryotic cell, with the interior space of a mitochondrion, containing the mitochondrial DNA and ribosomes, colored. Note the smooth outer membrane and the convoluted inner membrane, which houses the proteins that generate ATP from the oxidation of food molecules. (A, courtesy of Daniel S. Friend.)

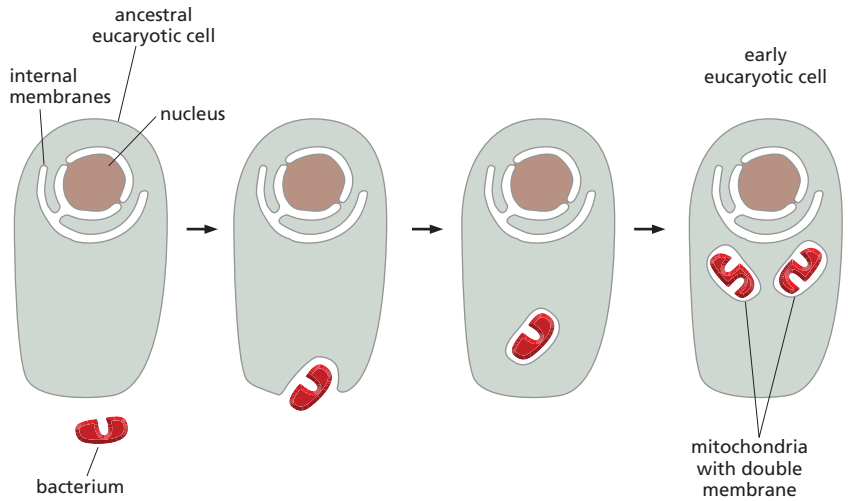


Figure 1–34 The origin of mitochondria. An ancestral eucaryotic cell is thought to have engulfed the bacterial ancestor of mitochondria, initiating a symbiotic relationship.

Many eucaryotic cells—specifically, those of plants and algae—also contain another class of small membrane-enclosed organelles somewhat similar to mitochondria—the *chloroplasts* (**Figure 1–35**). Chloroplasts perform photosynthesis, using the energy of sunlight to synthesize carbohydrates from atmospheric carbon dioxide and water, and deliver the products to the host cell as food. Like mitochondria, chloroplasts have their own genome and almost certainly originated as symbiotic photosynthetic bacteria, acquired by cells that already possessed mitochondria (**Figure 1–36**).

A eucaryotic cell equipped with chloroplasts has no need to chase after other cells as prey; it is nourished by the captive chloroplasts it has inherited from its ancestors. Correspondingly, plant cells, although they possess the cytoskeletal equipment for movement, have lost the ability to change shape rapidly and to engulf other cells by phagocytosis. Instead, they create around themselves a tough, protective cell wall. If the ancestral eucaryote was indeed a predator on other organisms, we can view plant cells as eucaryotes that have made the transition from hunting to farming.

Fungi represent yet another eucaryotic way of life. Fungal cells, like animal cells, possess mitochondria but not chloroplasts; but in contrast with animal cells and protozoa, they have a tough outer wall that limits their ability to move

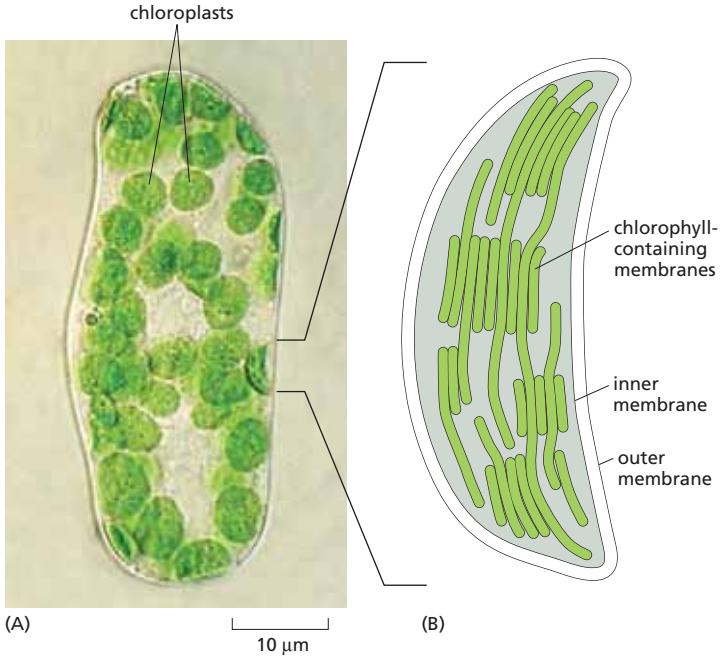


Figure 1–35 Chloroplasts. These organelles capture the energy of sunlight in plant cells and some single-celled eucaryotes. (A) A single cell isolated from a leaf of a flowering plant, seen in the light microscope, showing the green chloroplasts. (B) A drawing of one of the chloroplasts, showing the highly folded system of internal membranes containing the chlorophyll molecules by which light is absorbed. (A, courtesy of Preeti Dahiya.)

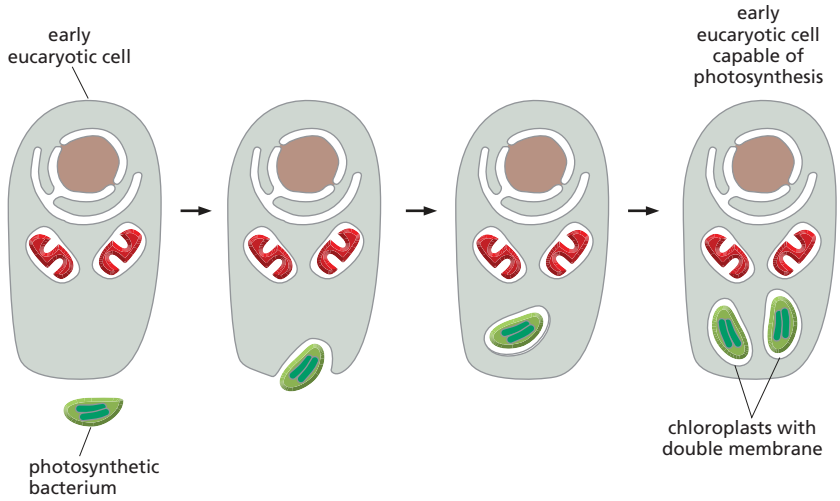


Figure 1–36 The origin of chloroplasts. An early eucaryotic cell, already possessing mitochondria, engulfed a photosynthetic bacterium (a cyanobacterium) and retained it in symbiosis. All present-day chloroplasts are thought to trace their ancestry back to a single species of cyanobacterium that was adopted as an internal symbiont (an endosymbiont) over a billion years ago.

rapidly or to swallow up other cells. Fungi, it seems, have turned from hunters into scavengers: other cells secrete nutrient molecules or release them upon death, and fungi feed on these leavings—performing whatever digestion is necessary extracellularly, by secreting digestive enzymes to the exterior.

Eucaryotes Have Hybrid Genomes

The genetic information of eucaryotic cells has a hybrid origin—from the ancestral anaerobic eucaryote, and from the bacteria that it adopted as symbionts. Most of this information is stored in the nucleus, but a small amount remains inside the mitochondria and, for plant and algal cells, in the chloroplasts. The mitochondrial DNA and the chloroplast DNA can be separated from the nuclear DNA and individually analyzed and sequenced. The mitochondrial and chloroplast genomes are found to be degenerate, cut-down versions of the corresponding bacterial genomes, lacking genes for many essential functions. In a human cell, for example, the mitochondrial genome consists of only 16,569 nucleotide pairs, and codes for only 13 proteins, two ribosomal RNA components, and 22 transfer RNAs.

The genes that are missing from the mitochondria and chloroplasts have not all been lost; instead, many of them have been somehow moved from the symbiont genome into the DNA of the host cell nucleus. The nuclear DNA of humans contains many genes coding for proteins that serve essential functions inside the mitochondria; in plants, the nuclear DNA also contains many genes specifying proteins required in chloroplasts.

Eucaryotic Genomes Are Big

Natural selection has evidently favored mitochondria with small genomes, just as it has favored bacteria with small genomes. By contrast, the nuclear genomes of most eucaryotes seem to have been free to enlarge. Perhaps the eucaryotic way of life has made large size an advantage: predators typically need to be bigger than their prey, and cell size generally increases in proportion to genome size. Perhaps enlargement of the genome has been driven by the accumulation of parasitic transposable elements (discussed in Chapter 5)—"selfish" segments of DNA that can insert copies of themselves at multiple sites in the genome. Whatever the explanation, the genomes of most eucaryotes are orders of magnitude larger than those of bacteria and archaea (**Figure 1–37**). And the freedom to be extravagant with DNA has had profound implications.

Eucaryotes not only have more genes than procaryotes; they also have vastly more DNA that does not code for protein or for any other functional product molecule. The human genome contains 1000 times as many nucleotide pairs as the genome of a typical bacterium, 20 times as many genes, and about 10,000

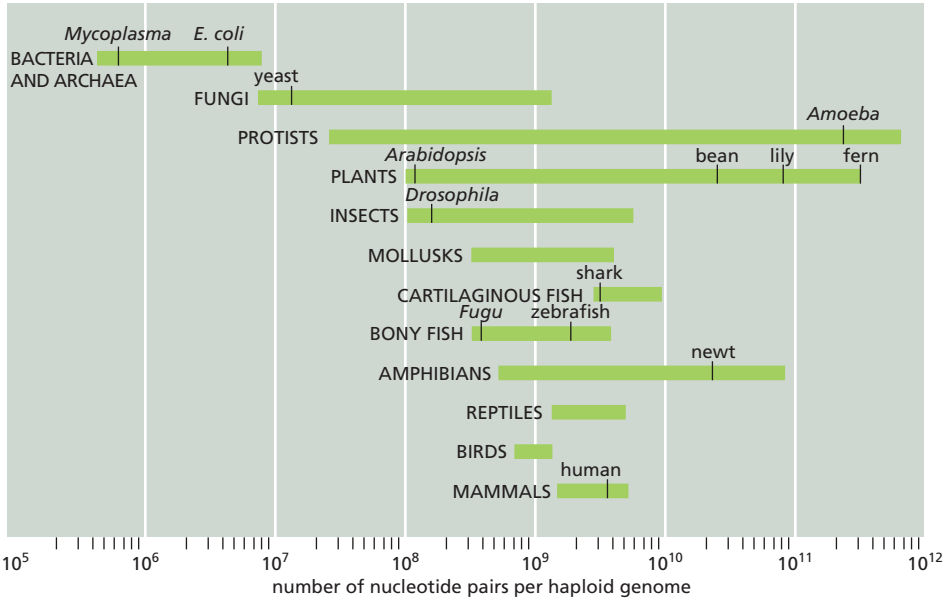


Figure 1–37 Genome sizes compared. Genome size is measured in nucleotide pairs of DNA per haploid genome, that is, per single copy of the genome. (The cells of sexually reproducing organisms such as ourselves are generally diploid: they contain two copies of the genome, one inherited from the mother, the other from the father.) Closely related organisms can vary widely in the quantity of DNA in their genomes, even though they contain similar numbers of functionally distinct genes. (Data from W.H. Li, Molecular Evolution, pp. 380–383. Sunderland, MA: Sinauer, 1997.)

times as much noncoding DNA (~98.5% of the genome for a human is noncoding, as opposed to 11% of the genome for the bacterium *E. coli*).

Eucaryotic Genomes Are Rich in Regulatory DNA

Much of our noncoding DNA is almost certainly dispensable junk, retained like a mass of old papers because, when there is little pressure to keep an archive small, it is easier to retain everything than to sort out the valuable information and discard the rest. Certain exceptional eucaryotic species, such as the puffer fish (**Figure 1–38**), bear witness to the profligacy of their relatives; they have somehow managed to rid themselves of large quantities of noncoding DNA. Yet they appear similar in structure, behavior, and fitness to related species that have vastly more such DNA.

Even in compact eucaryotic genomes such as that of puffer fish, there is more noncoding DNA than coding DNA, and at least some of the noncoding DNA certainly has important functions. In particular, it regulates the expression of adjacent genes. With this regulatory DNA, eucaryotes have evolved distinctive ways of controlling when and where a gene is brought into play. This sophisticated gene regulation is crucial for the formation of complex multicellular organisms.

The Genome Defines the Program of Multicellular Development

The cells in an individual animal or plant are extraordinarily varied. Fat cells, skin cells, bone cells, nerve cells—they seem as dissimilar as any cells could be. Yet all these cell types are the descendants of a single fertilized egg cell, and all (with minor exceptions) contain identical copies of the genome of the species.

The differences result from the way in which the cells make selective use of their genetic instructions according to the cues they get from their surroundings in the developing embryo. The DNA is not just a shopping list specifying the molecules that every cell must have, and the cell is not an assembly of all the items on the list. Rather, the cell behaves as a multipurpose machine, with sensors to receive environmental signals and with highly developed abilities to call different sets of genes into action according to the sequences of signals to which the cell has been exposed. The genome in each cell is big enough to accommodate the information that specifies an entire multicellular organism, but in any individual cell only part of that information is used.

A large fraction of the genes in the eucaryotic genome code for proteins that regulate the activities of other genes. Most of these *gene regulatory proteins* act by

Figure 1–38 The puffer fish *(Fugu rubripes).* This organism has a genome size of 400 million nucleotide pairs—about one-quarter as much as a zebrafish, for example, even though the two species of fish have similar numbers of genes. (From a woodcut by Hiroshige, courtesy of Arts and Designs of Japan.)

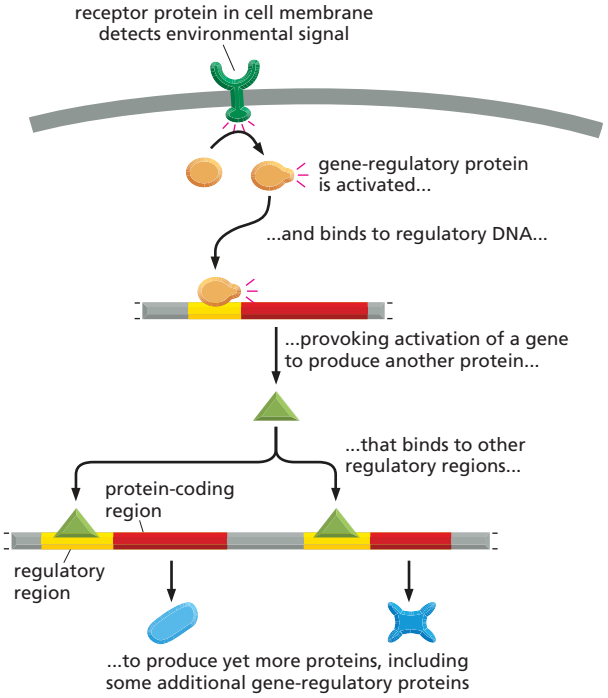


Figure 1–39 Controlling gene readout by environmental signals. Regulatory DNA allows gene expression to be controlled by regulatory proteins, which are in turn the products of other genes. This diagram shows how a cell's gene expression is adjusted according to a signal from the cell's environment. The initial effect of the signal is to activate a regulatory protein already present in the cell; the signal may, for example, trigger the attachment of a phosphate group to the regulatory protein, altering its chemical properties.

binding, directly or indirectly, to the regulatory DNA adjacent to the genes that are to be controlled (**Figure 1–39**), or by interfering with the abilities of other proteins to do so. The expanded genome of eucaryotes therefore not only specifies the hardware of the cell, but also stores the software that controls how that hardware is used (**Figure 1–40**).

Cells do not just passively receive signals; rather, they actively exchange signals with their neighbors. Thus, in a developing multicellular organism, the same control system governs each cell, but with different consequences depending on the messages exchanged. The outcome, astonishingly, is a precisely patterned array of cells in different states, each displaying a character appropriate to its position in the multicellular structure.

Many Eucaryotes Live as Solitary Cells: the Protists

Many species of eucaryotic cells lead a solitary life—some as hunters (the *protozoa*), some as photosynthesizers (the unicellular *algae*), some as scavengers (the unicellular fungi, or *yeasts*). **Figure 1–41** conveys something of the variety of forms of these single-celled eucaryotes, or *protists*. The anatomy of protozoa,

Figure 1–40 Genetic control of the program of multicellular development. The role of a regulatory gene is demonstrated in the snapdragon *Antirrhinum*. In this example, a mutation in a single gene coding for a regulatory protein causes leafy shoots to develop in place of flowers: because a regulatory protein has been changed, the cells adopt characters that would be appropriate to a different location in the normal plant. The mutant is on the left, the normal plant on the right. (Courtesy of Enrico Coen and Rosemary Carpenter.)

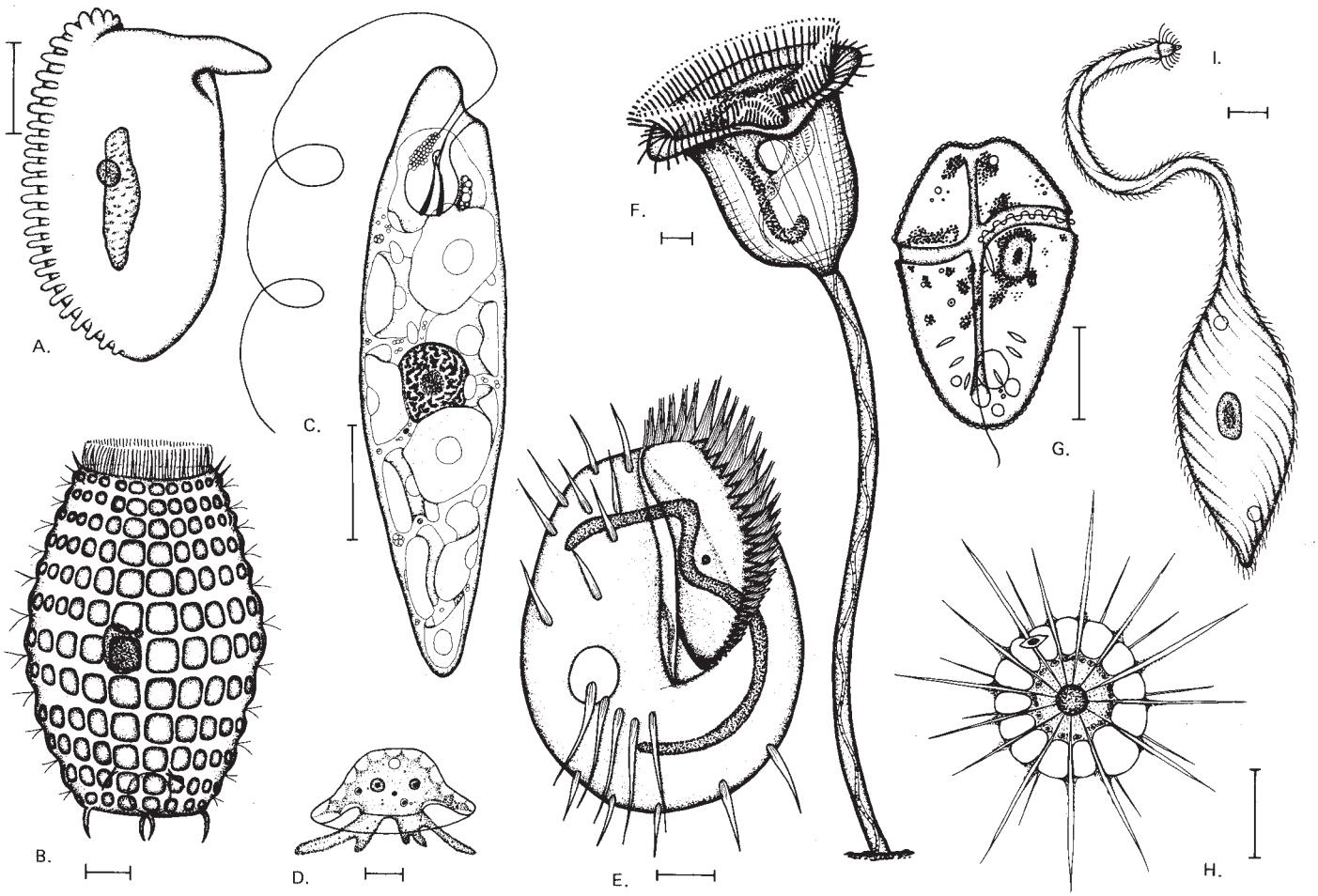


Figure 1–41 An assortment of protists: a small sample of an extremely diverse class of organisms. The drawings are done to different scales, but in each case the scale bar represents 10 μm. The organisms in (A), (B), (E), (F), and (I) are ciliates; (C) is a euglenoid; (D) is an amoeba; (G) is a dinoflagellate; (H) is a heliozoan. (From M.A. Sleigh, Biology of Protozoa. Cambridge, UK: Cambridge University Press, 1973.)

especially, is often elaborate and includes such structures as sensory bristles, photoreceptors, sinuously beating cilia, leglike appendages, mouth parts, stinging darts, and musclelike contractile bundles. Although they are single cells, protozoa can be as intricate, as versatile, and as complex in their behavior as many multicellular organisms (see Figure 1–32). <ATGG> <TCGC>

In terms of their ancestry and DNA sequences, protists are far more diverse than the multicellular animals, plants, and fungi, which arose as three comparatively late branches of the eucaryotic pedigree (see Figure 1–21). As with procaryotes, humans have tended to neglect the protists because they are microscopic. Only now, with the help of genome analysis, are we beginning to understand their positions in the tree of life, and to put into context the glimpses these strange creatures offer us of our distant evolutionary past.

A Yeast Serves as a Minimal Model Eucaryote

The molecular and genetic complexity of eucaryotes is daunting. Even more than for procaryotes, biologists need to concentrate their limited resources on a few selected model organisms to fathom this complexity.

To analyze the internal workings of the eucaryotic cell, without the additional problems of multicellular development, it makes sense to use a species that is unicellular and as simple as possible. The popular choice for this role of minimal model eucaryote has been the yeast *Saccharomyces cerevisiae* (**Figure 1–42**)—the same species that is used by brewers of beer and bakers of bread.

S. cerevisiae is a small, single-celled member of the kingdom of fungi and thus, according to modern views, at least as closely related to animals as it is to plants. It is robust and easy to grow in a simple nutrient medium. Like other fungi, it has a tough cell wall, is relatively immobile, and possesses mitochondria but not chloroplasts. When nutrients are plentiful, it grows and divides almost as

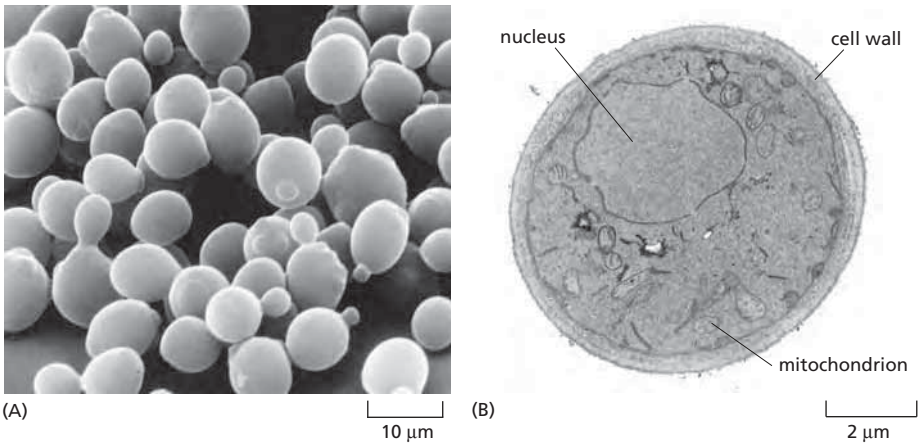


Figure 1–42 The yeast *Saccharomyces cerevisiae*. (A) A scanning electron micrograph of a cluster of the cells. This species is also known as budding yeast; it proliferates by forming a protrusion or bud that enlarges and then separates from the rest of the original cell. Many cells with buds are visible in this micrograph. (B) A transmission electron micrograph of a cross section of a yeast cell, showing its nucleus, mitochondrion, and thick cell wall. (A, courtesy of Ira Herskowitz and Eric Schabatach.)

rapidly as a bacterium. It can reproduce either vegetatively (that is, by simple cell division), or sexually: two yeast cells that are *haploid* (possessing a single copy of the genome) can fuse to create a cell that is *diploid* (containing a double genome); and the diploid cell can undergo *meiosis* (a reduction division) to produce cells that are once again haploid (**Figure 1–43**). In contrast with higher plants and animals, the yeast can divide indefinitely in either the haploid or the diploid state, and the process leading from the one state to the other can be induced at will by changing the growth conditions.

In addition to these features, the yeast has a further property that makes it a convenient organism for genetic studies: its genome, by eucaryotic standards, is exceptionally small. Nevertheless, it suffices for all the basic tasks that every eucaryotic cell must perform. As we shall see later in this book, studies on yeasts (using both *S. cerevisiae* and other species) have provided a key to many crucial processes, including the eucaryotic cell-division cycle—the critical chain of events by which the nucleus and all the other components of a cell are duplicated and parceled out to create two daughter cells from one. The control system that governs this process has been so well conserved over the course of evolution that many of its components can function interchangeably in yeast and human cells: if a mutant yeast lacking an essential yeast cell-division-cycle gene is supplied with a copy of the homologous cell-division-cycle gene from a human, the yeast is cured of its defect and becomes able to divide normally.

The Expression Levels of All The Genes of An Organism Can Be Monitored Simultaneously

The complete genome sequence of *S. cerevisiae*, determined in 1997, consists of approximately 13,117,000 nucleotide pairs, including the small contribution (78,520 nucleotide pairs) of the mitochondrial DNA. This total is only about 2.5 times as much DNA as there is in *E. coli*, and it codes for only 1.5 times as many distinct proteins (about 6300 in all). The way of life of *S. cerevisiae* is similar in many ways to that of a bacterium, and it seems that this yeast has likewise been subject to selection pressures that have kept its genome compact.

Knowledge of the complete genome sequence of any organism—be it a yeast or a human—opens up new perspectives on the workings of the cell: things that once seemed impossibly complex now seem within our grasp. Using techniques

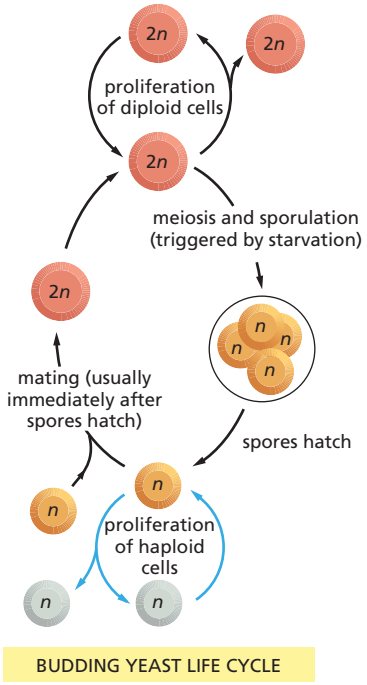


Figure 1–43 The reproductive cycles of the yeast *S. cerevisiae*. Depending on environmental conditions and on details of the genotype, cells of this species can exist in either a diploid (2*n*) state, with a double chromosome set, or a haploid (*n*) state, with a single chromosome set. The diploid form can either proliferate by ordinary cell-division cycles or undergo meiosis to produce haploid cells. The haploid form can either proliferate by ordinary cell-division cycles or undergo sexual fusion with another haploid cell to become diploid. Meiosis is triggered by starvation and gives rise to spores—haploid cells in a dormant state, resistant to harsh environmental conditions.

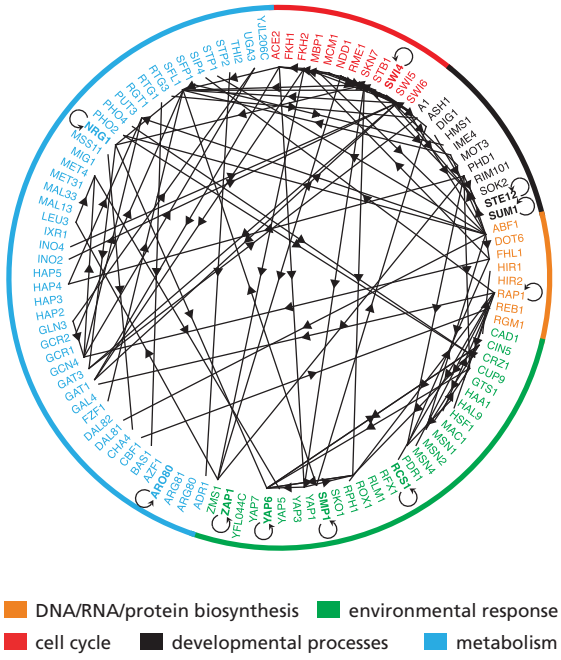


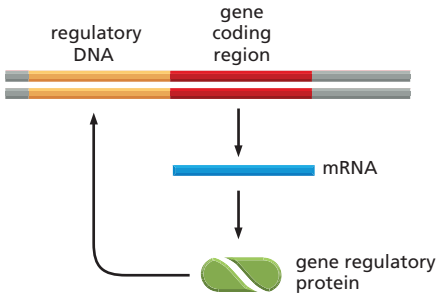
Figure 1–44 The network of interactions between gene regulatory proteins and the genes that code for them in a yeast cell. Results are shown for 106 out of the total of 141 gene regulatory proteins in *Saccharomyces cerevisiae*. Each protein in the set was tested for its ability to bind to the regulatory DNA of each of the genes coding for this set of proteins. In the diagram, the genes are arranged in a circle, and an arrow pointing from gene A to gene B means that the protein encoded by A binds to the regulatory DNA of B, and therefore presumably regulates the expression of B. Small circles with arrowheads indicate genes whose products directly regulate their own expression. Genes governing different aspects of cell behavior are shown in different colors. For a multicellular plant or animal, the number of gene regulatory proteins is about 10 times greater, and the amount of regulatory DNA perhaps 100 times greater, so that the corresponding diagram would be vastly more complex. (From T.I. Lee et al., *Science* 298:799–804, 2002. With permission from AAAS.)

to be described in Chapter 8, it is now possible, for example, to monitor, simultaneously, the amount of mRNA transcript that is produced from every gene in the yeast genome under any chosen conditions, and to see how this whole pattern of gene activity changes when conditions change. The analysis can be repeated with mRNA prepared from mutants lacking a chosen gene—any gene that we care to test. In principle, this approach provides a way to reveal the entire system of control relationships that govern gene expression—not only in yeast cells, but in any organism whose genome sequence is known.

To Make Sense of Cells, We Need Mathematics, Computers, and Quantitative Information

Through methods such as these, exploiting our knowledge of complete genome sequences, we can list the genes and proteins in a cell and begin to depict the web of interactions between them (**Figure 1–44**). But how are we to turn all this information into an understanding of how cells work? Even for a single cell type belonging to a single species of organism, the current deluge of data seems overwhelming. The sort of informal reasoning on which biologists usually rely seems totally inadequate in the face of such complexity. In fact, the difficulty is more than just a matter of information overload. Biological systems are, for example, full of feedback loops, and the behavior of even the simplest of systems with feedback is remarkably difficult to predict by intuition alone (**Figure 1–45**); small

Figure 1–45 A very simple gene regulatory circuit—a single gene regulating its own expression by the binding of its protein product to its own regulatory DNA. Simple schematic diagrams such as this are often used to summarize what we know (as in Figure 1–44), but they leave many questions unanswered. When the protein binds, does it inhibit or stimulate transcription? How steeply does the transcription rate depend on the protein concentration? How long, on average, does a molecule of the protein remain bound to the DNA? How long does it take to make each molecule of mRNA or protein, and how quickly does each type of molecule get degraded? Mathematical modeling shows that we need quantitative answers to all these and other questions before we can predict the behavior of even this single-gene system. For different parameter values, the system may settle to a unique steady state; or it may behave as a switch, capable of existing in one or other of a set of alternative states; or it may oscillate; or it may show large random fluctuations.



changes in parameters can cause radical changes in outcome. To go from a circuit diagram to a prediction of the behavior of the system, we need detailed quantitative information, and to draw deductions from that information we need mathematics and computers.

These tools for quantitative reasoning are essential, but they are not all-powerful. You might think that, knowing how each protein influences each other protein, and how the expression of each gene is regulated by the products of others, we should soon be able to calculate how the cell as a whole will behave, just as an astronomer can calculate the orbits of the planets, or a chemical engineer can calculate the flows through a chemical plant. But any attempt to perform this feat for an entire living cell rapidly reveals the limits of our present state of knowledge. The information we have, plentiful as it is, is full of gaps and uncertainties. Moreover, it is largely qualitative rather than quantitative. Most often, cell biologists studying the cell's control systems sum up their knowledge in simple schematic diagrams—this book is full of them—rather than in numbers, graphs, and differential equations. To progress from qualitative descriptions and intuitive reasoning to quantitative descriptions and mathematical deduction is one of the biggest challenges for contemporary cell biology. So far, the challenge has been met only for a few very simple fragments of the machinery of living cells—subsystems involving a handful of different proteins, or two or three cross-regulatory genes, where theory and experiment can go closely hand in hand. We shall discuss some of these examples later in the book.

Arabidopsis Has Been Chosen Out of 300,000 Species As a Model Plant

The large multicellular organisms that we see around us—the flowers and trees and animals—seem fantastically varied, but they are much closer to one another in their evolutionary origins, and more similar in their basic cell biology, than the great host of microscopic single-celled organisms. Thus, while bacteria and eucaryotes are separated by more than 3000 million years of divergent evolution, vertebrates and insects are separated by about 700 million years, fish and mammals by about 450 million years, and the different species of flowering plants by only about 150 million years.

Because of the close evolutionary relationship between all flowering plants, we can, once again, get insight into the cell and molecular biology of this whole class of organisms by focusing on just one or a few species for detailed analysis. Out of the several hundred thousand species of flowering plants on Earth today, molecular biologists have chosen to concentrate their efforts on a small weed, the common Thale cress *Arabidopsis thaliana* (**Figure 1–46**), which can be grown indoors in large numbers, and produces thousands of offspring per plant after 8–10 weeks. *Arabidopsis* has a genome of approximately 140 million nucleotide pairs, about 11 times as much as yeast, and its complete sequence is known.

Figure 1–46 *Arabidopsis thaliana,* the plant chosen as the primary model for studying plant molecular genetics. (Courtesy of Toni Hayden and the John Innes Foundation.)

The World of Animal Cells Is Represented By a Worm, a Fly, a Mouse, and a Human

Multicellular animals account for the majority of all named species of living organisms, and for the largest part of the biological research effort. Four species have emerged as the foremost model organisms for molecular genetic studies. In order of increasing size, they are the nematode worm *Caenorhabditis elegans*, the fly *Drosophila melanogaster*, the mouse *Mus musculus*, and the human, *Homo sapiens*. Each of these has had its genome sequenced.

Caenorhabditis elegans (**Figure 1–47**) is a small, harmless relative of the eelworm that attacks crops. With a life cycle of only a few days, an ability to survive in a freezer indefinitely in a state of suspended animation, a simple body plan, and an unusual life cycle that is well suited for genetic studies (described in Chapter 23), it is an ideal model organism. *C. elegans* develops with clockwork precision from a fertilized egg cell into an adult worm with exactly 959 body cells

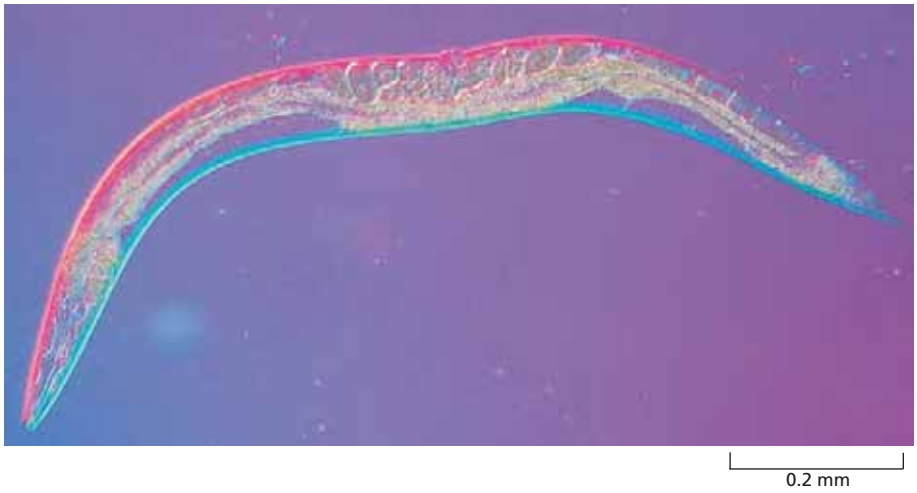


Figure 1–47 ***Caenorhabditis elegans,*** **the first multicellular organism to have its complete genome sequence determined.** This small nematode, about 1 mm long, lives in the soil. Most individuals are hermaphrodites, producing both eggs and sperm. The animal is viewed here using interference contrast optics, showing up the boundaries of the tissues in bright colors; the animal itself is not colored when viewed with ordinary lighting. (Courtesy of Ian Hope.)

(plus a variable number of egg and sperm cells)—an unusual degree of regularity for an animal. We now have a minutely detailed description of the sequence of events by which this occurs, as the cells divide, move, and change their characters according to strict and predictable rules. The genome of 97 million nucleotide pairs codes for about 19,000 proteins, and many mutants and other tools are available for the testing of gene functions. Although the worm has a body plan very different from our own, the conservation of biological mechanisms has been sufficient for the worm to be a model for many of the developmental and cell-biological processes that occur in the human body. Studies of the worm help us to understand, for example, the programs of cell division and cell death that determine the numbers of cells in the body—a topic of great importance in developmental biology and cancer research.

Studies in *Drosophila* Provide a Key to Vertebrate Development

The fruitfly *Drosophila melanogaster* (**Figure 1–48**) has been used as a model genetic organism for longer than any other; in fact, the foundations of classical genetics were built to a large extent on studies of this insect. Over 80 years ago, it provided, for example, definitive proof that genes—the abstract units of hereditary information—are carried on chromosomes, concrete physical objects whose behavior had been closely followed in the eucaryotic cell with the light microscope, but whose function was at first unknown. The proof depended on one of the many features that make *Drosophila* peculiarly convenient for genetics—the

Figure 1–48 ***Drosophila melanogaster.*** Molecular genetic studies on this fly have provided the main key to understanding how all animals develop from a fertilized egg into an adult. (From E.B. Lewis, *Science* 221:cover, 1983. With permission from AAAS.)

giant chromosomes, with characteristic banded appearance, that are visible in some of its cells (**Figure 1–49**). Specific changes in the hereditary information, manifest in families of mutant flies, were found to correlate exactly with the loss or alteration of specific giant-chromosome bands.

In more recent times, *Drosophila*, more than any other organism, has shown us how to trace the chain of cause and effect from the genetic instructions encoded in the chromosomal DNA to the structure of the adult multicellular body. *Drosophila* mutants with body parts strangely misplaced or mispatterned provided the key to the identification and characterization of the genes required to make a properly structured body, with gut, limbs, eyes, and all the other parts in their correct places. Once these *Drosophila* genes were sequenced, the genomes of vertebrates could be scanned for homologs. These were found, and their functions in vertebrates were then tested by analyzing mice in which the genes had been mutated. The results, as we see later in the book, reveal an astonishing degree of similarity in the molecular mechanisms of insect and vertebrate development.

The majority of all named species of living organisms are insects. Even if *Drosophila* had nothing in common with vertebrates, but only with insects, it would still be an important model organism. But if understanding the molecular genetics of vertebrates is the goal, why not simply tackle the problem head-on? Why sidle up to it obliquely, through studies in *Drosophila*?

Drosophila requires only 9 days to progress from a fertilized egg to an adult; it is vastly easier and cheaper to breed than any vertebrate, and its genome is much smaller—about 170 million nucleotide pairs, compared with 3200 million for a human. This genome codes for about 14,000 proteins, and mutants can now be obtained for essentially any gene. But there is also another, deeper reason why genetic mechanisms that are hard to discover in a vertebrate are often readily revealed in the fly. This relates, as we now explain, to the frequency of gene duplication, which is substantially greater in vertebrate genomes than in the fly genome and has probably been crucial in making vertebrates the complex and subtle creatures that they are.

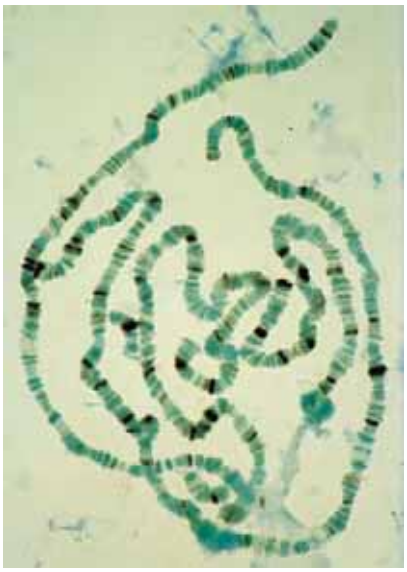


Figure 1–49 Giant chromosomes from salivary gland cells of *Drosophila*. Because many rounds of DNA replication have occurred without an intervening cell division, each of the chromosomes in these unusual cells contains over 1000 identical DNA molecules, all aligned in register. This makes them easy to see in the light microscope, where they display a characteristic and reproducible banding pattern. Specific bands can be identified as the locations of specific genes: a mutant fly with a region of the banding pattern missing shows a phenotype reflecting loss of the genes in that region. Genes that are being transcribed at a high rate correspond to bands with a "puffed" appearance. The bands stained dark brown in the micrograph are sites where a particular regulatory protein is bound to the DNA. (Courtesy of B. Zink and R. Paro, from R. Paro, *Trends Genet.* 6:416–421, 1990. With permission from Elsevier.)

The Vertebrate Genome Is a Product of Repeated Duplication

Almost every gene in the vertebrate genome has paralogs—other genes in the same genome that are unmistakably related and must have arisen by gene duplication. In many cases, a whole cluster of genes is closely related to similar clusters present elsewhere in the genome, suggesting that genes have been duplicated in linked groups rather than as isolated individuals. According to one hypothesis, at an early stage in the evolution of the vertebrates, the entire genome underwent duplication twice in succession, giving rise to four copies of every gene. In some groups of vertebrates, such as fish of the salmon and carp families (including the zebrafish, a popular research animal), it has been suggested that there was yet another duplication, creating an eightfold multiplicity of genes.

The precise course of vertebrate genome evolution remains uncertain, because many further evolutionary changes have occurred since these ancient events. Genes that were once identical have diverged; many of the gene copies have been lost through disruptive mutations; some have undergone further rounds of local duplication; and the genome, in each branch of the vertebrate family tree, has suffered repeated rearrangements, breaking up most of the original gene orderings. Comparison of the gene order in two related organisms, such as the human and the mouse, reveals that—on the time scale of vertebrate evolution—chromosomes frequently fuse and fragment to move large blocks of DNA sequence around. Indeed, it is possible, as we shall discuss in Chapter 7, that the present state of affairs is the result of many separate duplications of fragments of the genome, rather than duplications of the genome as a whole.

There is, however, no doubt that such whole-genome duplications do occur from time to time in evolution, for we can see recent instances in which duplicated chromosome sets are still clearly identifiable as such. The frog

Figure 1–50 Two species of the frog genus *Xenopus*. *X. tropicalis*, above, has an ordinary diploid genome; *X. laevis*, below, has twice as much DNA per cell. From the banding patterns of their chromosomes and the arrangement of genes along them, as well as from comparisons of gene sequences, it is clear that the large-genome species have evolved through duplications of the whole genome. These duplications are thought to have occurred in the aftermath of matings between frogs of slightly divergent *Xenopus* species. (Courtesy of E. Amaya, M. Offield and R. Grainger, *Trends Genet.* 14:253–255, 1998. With permission from Elsevier.)

genus *Xenopus*, for example, comprises a set of closely similar species related to one another by repeated duplications or triplications of the whole genome. Among these frogs are *X. tropicalis*, with an ordinary diploid genome; the common laboratory species *X. laevis*, with a duplicated genome and twice as much DNA per cell; and *X. ruwenzoriensis*, with a sixfold reduplication of the original genome and six times as much DNA per cell (108 chromosomes, compared with 36 in *X. laevis*, for example). These species are estimated to have diverged from one another within the past 120 million years (**Figure 1–50**).

Genetic Redundancy Is a Problem for Geneticists, But It Creates Opportunities for Evolving Organisms

Whatever the details of the evolutionary history, it is clear that most genes in the vertebrate genome exist in several versions that were once identical. The related genes often remain functionally interchangeable for many purposes. This phenomenon is called **genetic redundancy**. For the scientist struggling to discover all the genes involved in some particular process, it complicates the task. If gene A is mutated and no effect is seen, it cannot be concluded that gene A is functionally irrelevant—it may simply be that this gene normally works in parallel with its relatives, and these suffice for near-normal function even when gene A is defective. In the less repetitive genome of *Drosophila*, where gene duplication is less common, the analysis is more straightforward: single gene functions are revealed directly by the consequences of single-gene mutations (the single-engined plane stops flying when the engine fails).

Genome duplication has clearly allowed the development of more complex life forms; it provides an organism with a cornucopia of spare gene copies, which are free to mutate to serve divergent purposes. While one copy becomes optimized for use in the liver, say, another can become optimized for use in the brain or adapted for a novel purpose. In this way, the additional genes allow for increased complexity and sophistication. As the genes take on divergent functions, they cease to be redundant. Often, however, while the genes acquire individually specialized roles, they also continue to perform some aspects of their original core function in parallel, redundantly. Mutation of a single gene then causes a relatively minor abnormality that reveals only a part of the gene's function (**Figure 1–51**). Families of genes with divergent but partly overlapping functions are a pervasive feature of vertebrate molecular biology, and they are encountered repeatedly in this book.

The Mouse Serves as a Model for Mammals

Mammals have typically three or four times as many genes as *Drosophila,* a genome that is 20 times larger, and millions or billions of times as many cells in their adult bodies. In terms of genome size and function, cell biology, and molecular mechanisms, mammals are nevertheless a highly uniform group of organisms. Even anatomically, the differences among mammals are chiefly a matter of size and proportions; it is hard to think of a human body part that does not have a counterpart in elephants and mice, and vice versa. Evolution plays freely with quantitative features, but it does not readily change the logic of the structure.

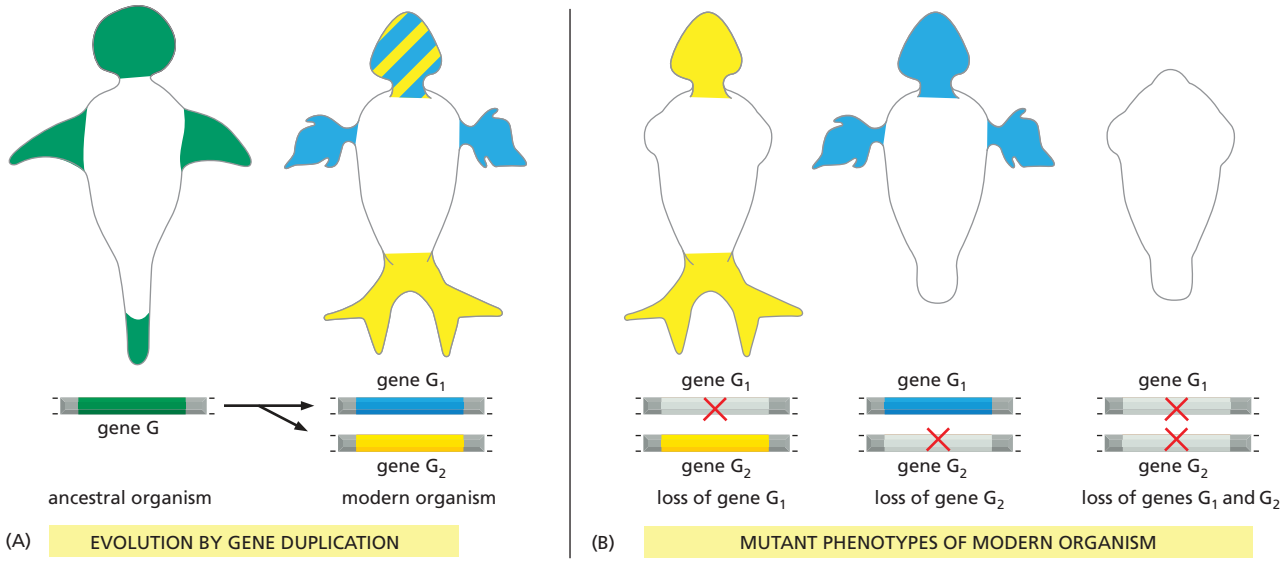


Figure 1–51 The consequences of gene duplication for mutational analyses of gene function. In this hypothetical example, an ancestral multicellular organism has a genome containing a single copy of gene G, which performs its function at several sites in the body, indicated in *green*. (A) Through gene duplication, a modern descendant of the ancestral organism has two copies of gene G, called G_1 and G_2. These have diverged somewhat in their patterns of expression and in their activities at the sites where they are expressed, but they still retain important similarities. At some sites, they are expressed together, and each independently performs the same old function as the ancestral gene G (alternating *green* and *yellow stripes);* at other sites, they are expressed alone and may serve new purposes. (B) Because of a functional overlap, the loss of one of the two genes by mutation *(red cross)* reveals only a part of its role; only the loss of both genes in the double mutant reveals the full range of processes for which these genes are responsible. Analogous principles apply to duplicated genes that operate in the same place (for example, in a single-celled organism) but are called into action together or individually in response to varying circumstances. Thus, gene duplications complicate genetic analyses in all organisms.

For a more exact measure of how closely mammalian species resemble one another genetically, we can compare the nucleotide sequences of corresponding (orthologous) genes, or the amino acid sequences of the proteins that these genes encode. The results for individual genes and proteins vary widely. But typically, if we line up the amino acid sequence of a human protein with that of the orthologous protein from, say, an elephant, about 85% of the amino acids are identical. A similar comparison between human and bird shows an amino acid identity of about 70%—twice as many differences, because the bird and the mammalian lineages have had twice as long to diverge as those of the elephant and the human (**Figure 1–52**).

The mouse, being small, hardy, and a rapid breeder, has become the foremost model organism for experimental studies of vertebrate molecular genetics. Many naturally occurring mutations are known, often mimicking the effects of corresponding mutations in humans (**Figure 1–53**). Methods have been developed, moreover, to test the function of any chosen mouse gene, or of any noncoding portion of the mouse genome, by artificially creating mutations in it, as we explain later in the book.

One made-to-order mutant mouse can provide a wealth of information for the cell biologist. It reveals the effects of the chosen mutation in a host of different contexts, simultaneously testing the action of the gene in all the different kinds of cells in the body that could in principle be affected.

Humans Report on Their Own Peculiarities

As humans, we have a special interest in the human genome. We want to know the full set of parts from which we are made, and to discover how they work. But even if you were a mouse, preoccupied with the molecular biology of mice, humans would be attractive as model genetic organisms, because of one special property: through medical examinations and self-reporting, we catalog our own genetic (and other) disorders. The human population is enormous, consisting today of some 6 billion individuals, and this self-documenting property means that a huge database of information exists on human mutations. The complete human genome sequence of more than 3 billion nucleotide pairs has now been determined, making it easier than ever before to identify at a molecular level the precise gene responsible for each human mutant characteristic.

By drawing together the insights from humans, mice, flies, worms, yeasts, plants, and bacteria—using gene sequence similarities to map out the correspondences between one model organism and another—we enrich our understanding of them all.

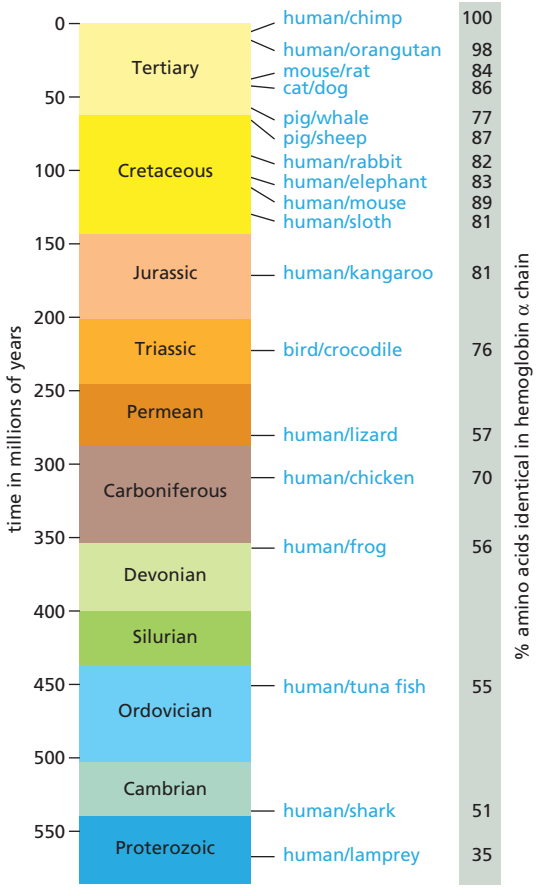


Figure 1–52 Times of divergence of different vertebrates. The scale on the left shows the estimated date and geological era of the last common ancestor of each specified pair of animals. Each time estimate is based on comparisons of the amino acid sequences of orthologous proteins; the longer a pair of animals have had to evolve independently, the smaller the percentage of amino acids that remain identical. Data from many different classes of proteins have been averaged to arrive at the final estimates, and the time scale has been calibrated to match the fossil evidence that the last common ancestor of mammals and birds lived 310 million years ago. The figures on the right give data on sequence divergence for one particular protein (chosen arbitrarily)—the α chain of hemoglobin. Note that although there is a clear general trend of increasing divergence with increasing time for this protein, there are also some irregularities. These reflect the randomness within the evolutionary process and, probably, the action of natural selection driving especially rapid changes of hemoglobin sequence in some organisms that experienced special physiological demands. On average, within any particular evolutionary lineage, hemoglobins accumulate changes at a rate of about 6 altered amino acids per 100 amino acids every 100 million years. Some proteins, subject to stricter functional constraints, evolve much more slowly than this, others as much as 5 times faster. All this gives rise to substantial uncertainties in estimates of divergence times, and some experts believe that the major groups of mammals diverged from one another as much as 60 million years more recently than shown here. (Adapted from S. Kumar and S.B. Hedges, *Nature* 392:917–920, 1998. With permission from Macmillan Publishers Ltd.)

We Are All Different in Detail

What precisely do we mean when we speak of *the* human genome? Whose genome? On average, any two people taken at random differ in about one or two in every 1000 nucleotide pairs in their DNA sequence. The Human Genome Project has arbitrarily selected DNA from a small number of anonymous individuals for sequencing. The human genome—the genome of the human species—is, properly speaking, a more complex thing, embracing the entire pool of variant genes that are found in the human population and continually exchanged and reassorted in the course of sexual reproduction. Ultimately, we can hope to document this variation too. Knowledge of it will help us understand, for example, why some people are prone to one disease, others to another; why some respond well to a drug, others badly. It will also provide new clues to our history—the population movements and minglings of our ancestors, the infections they suffered, the diets they ate. All these things leave traces in the variant forms of genes that have survived in human communities.

Figure 1–53 Human and mouse: similar genes and similar development. The human baby and the mouse shown here have similar white patches on their foreheads because both have mutations in the same gene (called *Kit)*, required for the development and maintenance of pigment cells. (Courtesy of R.A. Fleischman.)

Knowledge and understanding bring the power to intervene—with humans, to avoid or prevent disease; with plants, to create better crops; with bacteria, to turn them to our own uses. All these biological enterprises are linked, because the genetic information of all living organisms is written in the same language. The new-found ability of molecular biologists to read and decipher this language has already begun to transform our relationship to the living world. The account of cell biology in the subsequent chapters will, we hope, prepare you to understand, and possibly to contribute to, the great scientific adventure of the twenty-first century.

Summary

Eucaryotic cells, by definition, keep their DNA in a separate membrane-enclosed compartment, the nucleus. They have, in addition, a cytoskeleton for support and movement, elaborate intracellular compartments for digestion and secretion, the capacity (in many species) to engulf other cells, and a metabolism that depends on the oxidation of organic molecules by mitochondria. These properties suggest that eucaryotes may have originated as predators on other cells. Mitochondria—and, in plants, chloroplasts—contain their own genetic material, and evidently evolved from bacteria that were taken up into the cytoplasm of the eucaryotic cell and survived as symbionts. Eucaryotic cells have typically 3–30 times as many genes as procaryotes, and often thousands of times more noncoding DNA. The noncoding DNA allows for complex regulation of gene expression, as required for the construction of complex multicellular organisms. Many eucaryotes are, however, unicellular—among them the yeast Saccharomyces cerevisiae, *which serves as a simple model organism for eucaryotic cell biology, revealing the molecular basis of conserved fundamental processes such as the eucaryotic cell division cycle. A small number of other organisms have been chosen as primary models for multicellular plants and animals, and the sequencing of their entire genomes has opened the way to systematic and comprehensive analysis of gene functions, gene regulation, and genetic diversity. As a result of gene duplications during vertebrate evolution, vertebrate genomes contain multiple closely related homologs of most genes. This genetic redundancy has allowed diversification and specialization of genes for new purposes, but it also makes gene functions harder to decipher. There is less genetic redundancy in the nematode* Caenorhabditis elegans *and the fly* Drosophila melanogaster, *which have thus played a key part in revealing universal genetic mechanisms of animal development.*

PROBLEMS

Which statements are true? Explain why or why not.

1–1 The human hemoglobin genes, which are arranged in two clusters on two chromosomes, provide a good example of an orthologous set of genes.

1–2 Horizontal gene transfer is more prevalent in single-celled organisms than in multicellular organisms.

1–3 Most of the DNA sequences in a bacterial genome code for proteins, whereas most of the sequences in the human genome do not.

Discuss the following problems.

1–4 Since it was deciphered four decades ago, some have claimed that the genetic code must be a frozen accident, while others have argued that it was shaped by natural selection. A striking feature of the genetic code is its inherent resistance to the effects of mutation. For example, a change in the third position of a codon often specifies the same amino acid or one with similar chemical properties. The natural code resists mutation more effectively (is less susceptible to error) than most other possible versions, as illustrated in **Figure Q1–1**. Only one in a million computer-generated "random" codes is more error-resistant than the natural genetic code. Does the extraordinary mutation resistance of the genetic code argue in favor of its origin as a frozen accident or as a result of natural selection? Explain your reasoning.

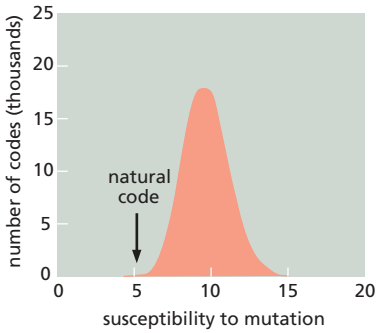


Figure Q1–1 Susceptibility of the natural code relative to millions of computer-generated codes (Problem 1–4). Susceptibility measures the average change in amino acid properties caused by random mutations. A small value indicates that mutations tend to cause minor changes. (Data courtesy of Steve Freeland.)

1–5 You have begun to characterize a sample obtained from the depths of the oceans on Europa, one of Jupiter's moons. Much to your surprise, the sample contains a life-form that grows well in a rich broth. Your preliminary analysis

shows that it is cellular and contains DNA, RNA, and protein. When you show your results to a colleague, she suggests that your sample was contaminated with an organism from Earth. What approaches might you try to distinguish between contamination and a novel cellular life-form based on DNA, RNA, and protein?

1–6 It is not so difficult to imagine what it means to feed on the organic molecules that living things produce. That is, after all, what we do. But what does it mean to "feed" on sunlight, as phototrophs do? Or, even stranger, to "feed" on rocks, as lithotrophs do? Where is the "food," for example, in the mixture of chemicals (H_2S, H_2, CO, Mn^+, Fe^{2+}, Ni^{2+}, CH_4, and NH_4^+) spewed forth from a hydrothermal vent?

1–7 How many possible different trees (branching patterns) can be drawn for eubacteria, archaea, and eucaryotes, assuming that they all arose from a common ancestor?

1–8 The genes for ribosomal RNA are highly conserved (relatively few sequence changes) in all organisms on Earth; thus, they have evolved very slowly over time. Were ribosomal RNA genes "born" perfect?

1–9 Genes participating in informational processes such as replication, transcription, and translation are transferred between species much less often than are genes involved in metabolism. The basis for this inequality is unclear at present, but one suggestion is that it relates to the underlying complexity. Informational processes tend to involve large aggregates of different gene products, whereas metabolic reactions are usually catalyzed by enzymes composed of a single protein. Why would the complexity of the underlying process—informational or metabolic—have any effect on the rate of horizontal gene transfer?

1–10 The process of gene transfer from the mitochondrial to the nuclear genome can be analyzed in plants. The respiratory gene *Cox2*, which encodes subunit 2 of cytochrome oxidase, was functionally transferred to the nucleus during flowering plant evolution. Extensive analyses of plant genera have pinpointed the time of appearance of the nuclear form of the gene and identified several likely intermediates in the ultimate loss from the mitochondrial genome. A summary of *Cox2* gene distributions between mitochondria and nuclei, along with data on their transcription, is shown in a phylogenetic context in **Figure Q1–2**.

A. Assuming that transfer of the mitochondrial gene to the nucleus occurred only once (an assumption supported by the structures of the nuclear genes), indicate the point in the phylogenetic tree where the transfer occurred.

B. Are there any examples of genera in which the transferred gene and the mitochondrial gene both appear functional? Indicate them.

C. What is the minimal number of times that the mitochondrial gene has been inactivated or lost? Indicate those events on the phylogenetic tree.

D. What is the minimal number of times that the nuclear gene has been inactivated or lost? Indicate those events on the phylogenetic tree.

E. Based on this information, propose a general scheme for transfer of mitochondrial genes to the nuclear genome.

1–11 When plant hemoglobin genes were first discovered in legumes, it was so surprising to find a gene typical of animal blood that it was hypothesized that the plant gene arose

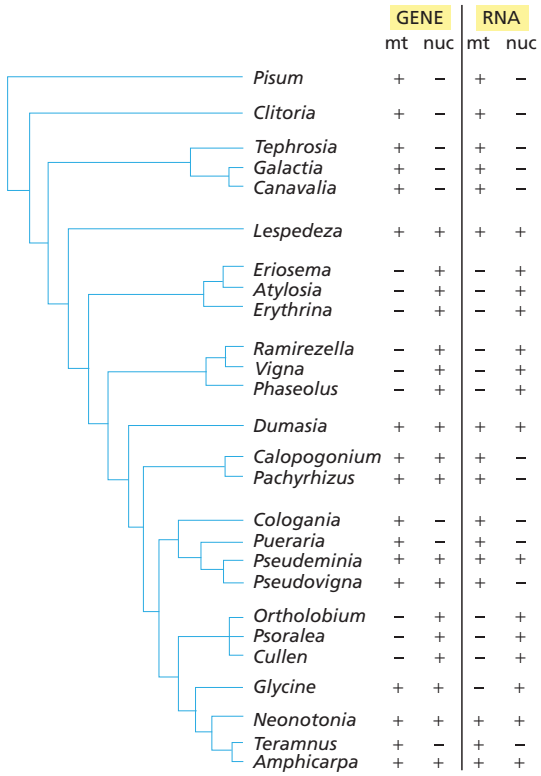


Figure Q1–2 Summary of *Cox2* gene distribution and transcript data in a phylogenetic context (Problem 1–10). The presence of the intact gene or a functional transcript is indicated by (+); the absence of the intact gene or a functional transcript is indicated by (–). mt, mitochondria; nuc, nuclei.

by horizontal transfer from an animal. Many more hemoglobin genes have now been sequenced, and a phylogenetic tree based on some of these sequences is shown in **Figure Q1–3**.

A. Does this tree support or refute the hypothesis that the plant hemoglobins arose by horizontal gene transfer?

B. Supposing that the plant hemoglobin genes were originally derived from a parasitic nematode, for example, what would you expect the phylogenetic tree to look like?

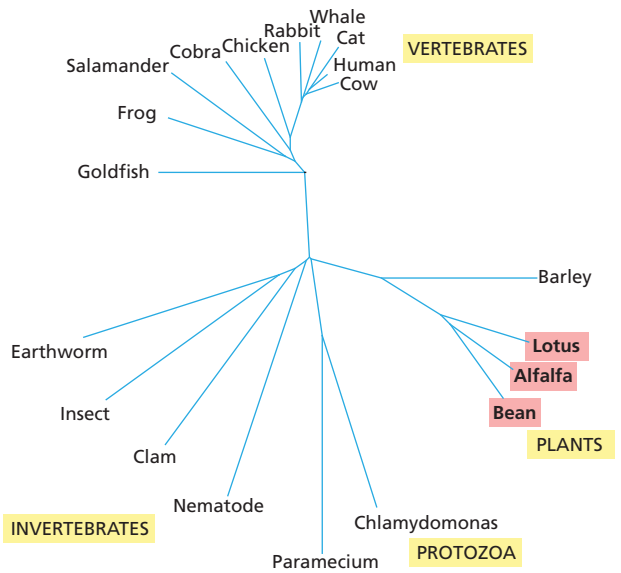


Figure Q1–3 Phylogenetic tree for hemoglobin genes from a variety of species (Problem 1–11). The legumes are highlighted in red.

1–12 Rates of evolution appear to vary in different lineages. For example, the rate of evolution in the rat lineage is significantly higher than in the human lineage. These rate differences are apparent whether one looks at changes in protein sequences that are subject to selective pressure or at changes in noncoding nucleotide sequences, which are not under obvious selection pressure. Can you offer one or more possible explanations for the slower rate of evolutionary change in the human lineage versus the rat lineage?

REFERENCES

General

Alberts B, Bray D, Hopkin K et al (2004) Essential Cell Biology, 2nd ed. New York: Garland Science.

Barton NH, Briggs DEG, Eisen JA et al (2007) Evolution. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Darwin C (1859) On the Origin of Species. London: Murray.

Graur D & Li W-H (1999) Fundamentals of Molecular Evolution, 2nd ed. Sunderland, MA: Sinauer Associates.

Madigan MT & Martinko JM (2005) Brock's Biology of Microorganisms, 11th ed. Englewood Cliffs, NJ: Prentice Hall.

Margulis L & Schwartz KV (1998) Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth, 3rd ed. New York: Freeman.

Watson JD, Baker TA, Bell SP et al (2007) Molecular Biology of the Gene, 6th ed. Menlo Park, CA: Benjamin-Cummings.

The Universal Features of Cells on Earth

Andersson SGE (2006) The bacterial world gets smaller. *Science* 314:259–260.

Brenner S, Jacob F & Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190:576–581.

Fraser CM, Gocayne JD, White O et al (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397–403.

Harris JK, Kelley ST, Spiegelman et al (2003) The genetic core of the universal ancestor. *Genome Res* 13:407–413.

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–338.

Watson JD & Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* 171:737–738.

Yusupov MM,Yusupova GZ, Baucom A et al (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science* 292:883–896.

The Diversity of Genomes and the Tree of Life

Blattner FR, Plunkett G, Bloch CA et al (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1474.

Boucher Y, Douady CJ, Papke RT et al (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* 37:283–328.

Cole ST, Brosch R, Parkhill J et al (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393:537–544.

Dixon B (1994) Power Unseen: How Microbes Rule the World. Oxford: Freeman.

Kerr RA (1997) Life goes to extremes in the deep earth—and elsewhere? *Science* 276:703–704.

Lee TI, Rinaldi NJ, Robert F et al (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298:799–804.

Olsen GJ & Woese CR (1997) Archaeal genomics: an overview. *Cell* 89:991–994.

Pace NR (1997) A molecular view of microbial diversity and the biosphere. *Science* 276:734–740.

Woese C (1998) The universal ancestor. *Proc Natl Acad Sci USA* 95:6854–6859.

Genetic Information in Eucaryotes

Adams MD, Celniker SE, Holt RA et al (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.

Andersson SG, Zomorodipour A, Andersson JO et al (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396:133–140.

The *Arabidopsis* Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.

Carroll SB, Grenier JK & Weatherbee SD (2005) From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design, 2nd ed. Maldon, MA: Blackwell Science.

de Duve C (2007) The origin of eukaryotes: a reappraisal. *Nature Rev Genet* 8:395-403.

Delsuc F, Brinkmann H & Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Rev Genet* 6:361–375.

DeRisi JL, Iyer VR & Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–686.

Gabriel SB, Schaffner SF, Nguyen H et al (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.

Goffeau A, Barrell BG, Bussey H et al (1996) Life with 6000 genes. *Science* 274:546–567.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Kellis M, Birren BW & Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617–624.

Lynch M & Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.

Mulley J & Holland P (2004) Comparative genomics: Small genome, big insights. *Nature* 431:916–917.

National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/

Owens K & King MC (1999) Genomic views of human history. *Science* 286:451–453.

Palmer JD & Delwiche CF (1996) Second-hand chloroplasts and the case of the disappearing nucleus. *Proc Natl Acad Sci USA* 93:7432–7435.

Pennisi E (2004) The birth of the nucleus. *Science* 305:766–768.

Plasterk RH (1999) The year of the worm. *BioEssays* 21:105–109.

Reed FA & Tishkoff SA (2006) African human diversity, origins and migrations. *Curr Opin Genet Dev* 16:597–605.

Rubin GM, Yandell MD, Wortman JR et al (2000) Comparative genomics of the eukaryotes. *Science* 287:2204–2215.

Stillman B & Stewart D (2003) The genome of *Homo sapiens*. (Cold Spring Harbor Symp. Quant. Biol. LXVIII). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–2018.

Tinsley RC & Kobel HR (eds) (1996) The Biology of *Xenopus*. Oxford: Clarendon Press.

Tyson JJ, Chen KC & Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* 15:221–231.

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. *Science* 291:1304–1351.