Life

The Science of Biology

TWELFTH EDITION

David M. Hillis
University of Texas, Austin

H. Craig Heller
Stanford University

Sally D. Hacker
Oregon State University

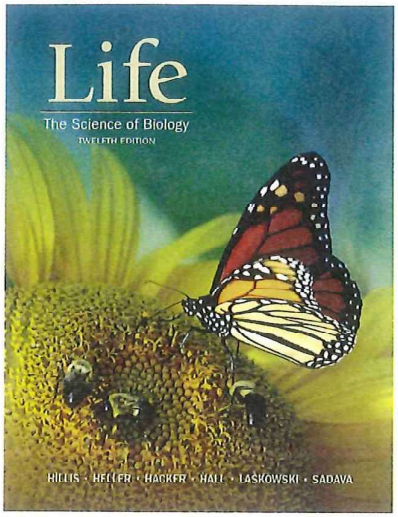
David W. Hall
University of Georgia

Marta J. Laskowski
Oberlin College

David Sadava
Emeritus, The Claremont Colleges

SINAUER ASSOCIATES

macmillan international HIGHER EDUCATION



© cbs/Alamy Stock Photo

Life: The Science of Biology, Twelfth Edition

Address editorial correspondence to:
Oxford University Press, 23 Plumtree Road, Sunderland, MA 01375 U.S.A.

Address orders to:
MPS/W. H. Freeman & Co., Order Department,
16365 James Madison Highway, U.S. Route 15,
Gordonsville, VA 22942 U.S.A.

Examination Copy Information: 1-800-446-8923

ISBN: 9781319315788

First Printing December 2019
Printed in the United States of America

Your body is more than 60% water by weight, excluding the minerals contained in bones. Water is the dominant component of virtually all living organisms, and most biochemical reactions take place in this watery, or aqueous, environment. Indeed, the search for life on other planets usually begins with the question of whether the planet has liquid water. Water is an unusual substance with unusual properties, many of them resulting from the polarity of the covalent bonds in the molecule and the ability to form hydrogen bonds with other molecules.

Water has a unique structure and special properties

The molecule H_2O has unusual physical and chemical properties.

ICE FLOATS In water's solid state (ice), individual water molecules are held in place by hydrogen bonds. Each molecule is bonded to four other molecules in a rigid, crystalline structure (Figure 2.15). Although the molecules are held firmly in place, they are farther apart from one another than they are in liquid water, where the molecules are moving about. In ice, there are cavities between individual water molecules. In other words, solid water is less dense than liquid water, which is why ice floats.

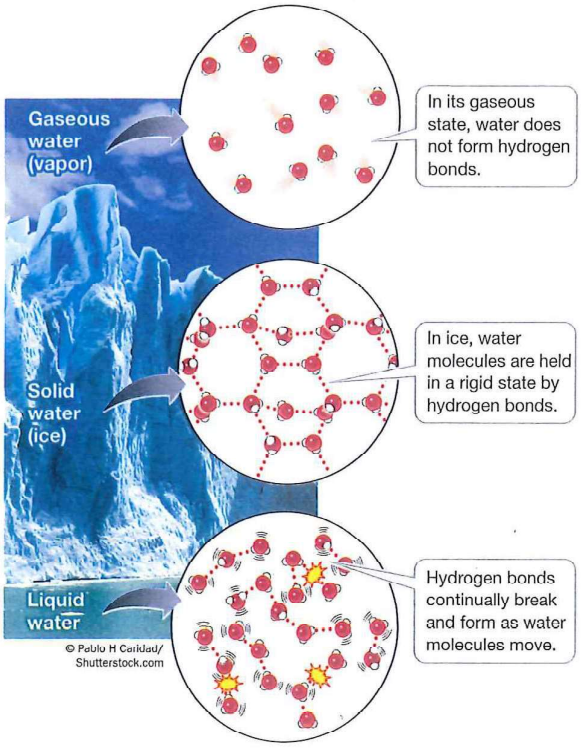


Figure 2.15 Hydrogen Bonding and the Properties of Water
Hydrogen bonding occurs between the molecules of water in both its liquid and solid states. Ice is more structured but less dense than liquid water, which is why ice floats. Water forms a gas when its hydrogen bonds are broken and the molecules move farther apart.

Think of the biological consequences if ice were to sink in water. A pond would freeze from the bottom up, becoming a solid block of ice in winter and killing most of the organisms living there. Once the whole pond was frozen, its temperature could drop well below the freezing point of water. But because ice floats, it forms an insulating layer on the top of the pond, and reduces heat flow to the cold air above. Thus fishes, plants, and other organisms in the pond are not subjected to temperatures lower than 0°C, which is the freezing point of pure water.

MELTING, FREEZING, AND HEAT CAPACITY Compared with many other substances that have molecules of similar size, ice requires a great deal of heat energy to melt. The amount of heat energy required to raise the temperature of 1 gram of a substance by 1 degree Celsius is called its **specific heat**. Water has a relatively high specific heat because so many hydrogen bonds connecting the water molecules in ice must be broken to change water from solid to liquid. In the opposite process—freezing—a great deal of energy is released to the environment. We say water has a high heat capacity. For example, water has twice the specific heat of ethyl alcohol and five times that of sand. This is why when you are on the beach as the sun sets, the sand cools much more quickly than the water.

Water also has a high **heat of vaporization**, defined as the heat a liquid absorbs in order to form a gas. A lot of heat (actually 580 calories per gram) is required to change 1 gram of water from liquid to gas. Once again, much of the heat energy is used to break the many hydrogen bonds between the water molecules. This heat must be absorbed from the environment in contact with the water. **Evaporation** (forming a vapor, or gas) thus has a cooling effect on the environment—whether a leaf, a forest, or an entire land mass. This effect explains why sweating cools the human body: as sweat evaporates from the skin, it uses up some of the adjacent body heat (Figure 2.16A).

On an environmental scale, a lot of solar energy that arrives at Earth is absorbed by the oceans. This heat absorption keeps ocean temperatures moderate. This not only allows survival of many marine organisms that otherwise might die in high temperatures, but also keeps the temperatures on coastal lands more moderate. If you have been to the San Francisco Bay area you know this: on a summer day, the temperature in the city is often 10 degrees Celsius cooler than that of inland cities, such as Walnut Creek. As the global climate gets warmer, heat absorption by bodies of water is becoming increasingly important.

Connect the Concepts Living systems use the evaporation of water, which disrupts hydrogen bonds, to dissipate excess heat that would otherwise cause problems. See Key Concept 37.1 for examples in plants, and Key Concepts 38.3–38.5 for examples in animals.

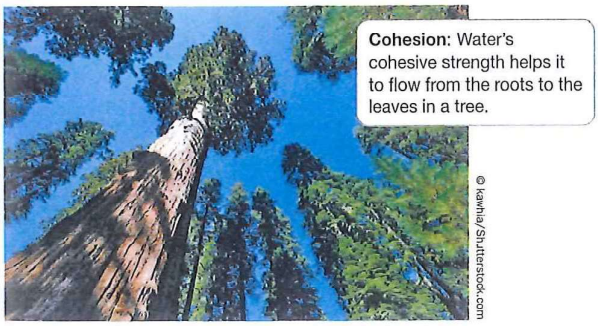
COHESION AND SURFACE TENSION In liquid water, individual molecules are able to move about. The hydrogen bonds between the molecules continually form and break (see Figure 2.15). Chemists estimate that this occurs about a trillion times a minute for a single water molecule!

At any given time, a water molecule forms on average 3.4 hydrogen bonds with other water molecules. These hydrogen bonds

explain the cohesive strength of liquid water. This cohesive strength, or **cohesion**, is defined as the capacity of water molecules to resist coming apart from one another when placed under tension. Water's cohesive strength permits narrow columns of liquid water to move from the roots to the leaves of tall trees. When water evaporates from the leaves, the entire column moves upward in response to the pull of the molecules at the top (Figure 2.16B). A related property is **adhesion**, the attraction of water molecules to other molecules of a different type. For example, when you put a straw into a cup with water, it "climbs" up the straw so the column is higher than the level in the cup. This adhesive behavior of water—sticking to the sides of the straw—reflects the adhesion of water forming the column.

(A)

(B)



(C)

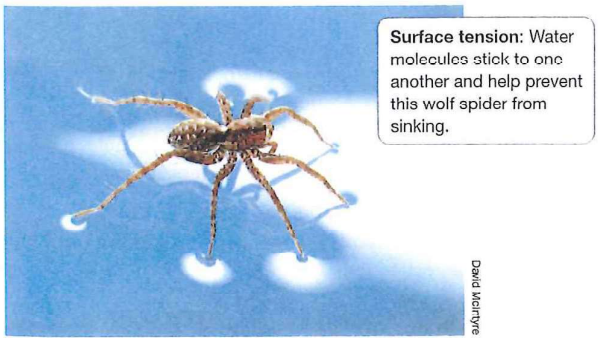


Figure 2.16 Water in Biology These three properties of water make it beneficial to organisms.

Connect the Concepts As described in Key Concept 33.2, the transpiration–cohesion–tension mechanism accounts for the movement of water from roots to leaves. Because of the cohesion between water molecules, water is pulled upward in continuous columns.

The surface of liquid water exposed to the air is difficult to puncture because the water molecules at the surface are hydrogen-bonded to other water molecules below them. This surface tension of water permits a container to be filled slightly above its rim without overflowing, and it permits spiders to walk on the surface of a pond (Figure 2.16C).

The reactions of life take place in aqueous solutions

A **solution** is produced when a substance (the **solute**) is dissolved in a liquid (the **solvent**). If the solvent is water, then the solution is called an aqueous solution. Water is polar, and because many important molecules in biological systems are polar, they readily dissolve in water. Being soluble doesn't mean that the molecules lose their identity and properties. They can still react, and indeed many important biochemical reactions occur in aqueous solutions.

Biologists who are interested in the biochemical reactions within cells identify the reactants and products and determine their amounts using two different types of analyses:

1. *Qualitative analyses* focus on identifying the substances involved in chemical reactions. For example, a qualitative analysis would be used to investigate the steps involved and the products formed when carbon-containing compounds are broken down to release energy in living tissues.
2. *Quantitative analyses* measure concentrations or amounts of substances. For example, a biochemist would use a quantitative analysis to measure how much of a certain product is formed in a chemical reaction. What follows is a brief introduction to some of the quantitative chemical terms you will see in this book.

A **mole** is the amount of a substance (in grams) that is numerically equal to its molecular weight. So a mole of hydrogen gas (H_2) weighs 2 g, a mole of sodium ion (Na^+) weighs 23 g, and a mole of table sugar ($C_{12}H_{22}O_{11}$) weighs about 342 g.

Quantitative analyses do not yield counts of molecules. Because the amount of a substance in 1 mole is directly related to its molecular weight, it follows that the number of molecules in 1 mole is constant for all substances. So 1 mole of salt contains the same *number of molecules* as 1 mole of table sugar. This constant number of molecules in a mole is called **Avogadro's number**, and it is 6.02×10^{23} molecules per mole. Chemists work with moles of substances (which can be weighed in the laboratory) instead of actual molecules, which are too numerous to be counted. Consider 34.2 g (just over 1 ounce) of table sugar, $C_{12}H_{22}O_{11}$. This is one-tenth of a mole, or one-tenth of Avogadro's number: 6.02×10^{22} molecules.

A chemist can dissolve 1 mole of table sugar (342 g) in water to make 1 liter of solution, knowing that the mole contains 6.02×10^{23} individual sugar molecules. This solution—1 mole of a substance dissolved in water to make 1 liter—is called a 1 molar (1 M) solution. When a physician injects a certain volume and molar concentration of a drug into the bloodstream of a patient, a rough calculation can

KEY CONCEPT
8.1 Chemical Transformations Involve Energy and Energy Transfers

Learning Objectives

- **8.1.1** Apply the second law of thermodynamics to biological systems.
- **8.1.2** Differentiate between exergonic and endergonic reactions.

A **chemical reaction** occurs when atoms have sufficient energy to combine or change their bonding partners. Consider the hydrolysis of the disaccharide sucrose to its component monomers, glucose and fructose (see Figure 3.18):

$$\underset{(C_{12}H_{22}O_{11})}{\text{Sucrose}} + H_2O \rightarrow \underset{(C_6H_{12}O_6)}{\text{glucose}} + \underset{(C_6H_{12}O_6)}{\text{fructose}}$$

In this equation, sucrose and water are the **reactants**, and glucose and fructose are the **products**. During the reaction, some of the bonds in sucrose and water are broken and new bonds are formed, resulting in products with chemical properties that are very different from those of the reactants. The sum total of all the chemical reactions occurring in a biological system at a given time is called **metabolism**. Metabolic reactions involve energy changes; for example, the energy contained in the chemical bonds of sucrose (reactants) is greater than the energy in the bonds of the two products, glucose and fructose.

Physicists define **energy** as the capacity to do work, which occurs when a force operates on an object over a distance. In biochemistry, it is more useful to think of energy as *the capacity for change*. In biochemical reactions, energy changes are usually associated with changes in the chemical compositions and properties of molecules.

There are two basic types of energy

Energy comes in many forms: chemical, electrical, heat, light, and mechanical (Table 8.1). But all forms of energy can be considered as one of two basic types (Figure 8.1):

1. **Potential energy** is the energy of state or position—that is, stored energy. It can be stored in many forms: in covalent bonds, as a concentration gradient, or even as an electric charge imbalance.
2. **Kinetic energy** is the energy of movement—that is, the type of energy that does work, that makes things change. For example, heat causes molecular motions and can even break chemical bonds.

To produce a change, energy must be transformed, either by conversion from one form to another or from one place to another. Potential energy can be converted into kinetic energy and vice versa, and the form that the energy takes can also be converted. Think of reading this book: light energy is converted to chemical energy in your eyes, and then is converted to electrical energy in the nerve cells that carry messages to your brain. When you decide to turn a page, the electrical and chemical energy of nerves and muscles are converted to kinetic energy for movement of your hand and arm.

TABLE 8.1 | Energy in Biology

Form of energy	Example in biology
Chemical-bond: Stored in bonds	Chemical energy stored in covalent bonds is released during the hydrolysis of polymers
Electrical: Separation of charges	Electrical gradients across cell membranes help drive the movement of ions through channels
Heat: Transfer due to temperature difference	Heat can be released by chemical reactions, and this can alter the internal temperature of an organism
Light: Electromagnetic radiation stored as photons	Light energy is captured by pigments in the eye and by plant pigments in photosynthesis
Mechanical: Energy of motion	Mechanical energy is used in muscle movements and movements within cells

Figure 8.1 Energy Conversions and Work A leaping cat illustrates both the conversion between potential and kinetic energy and the conversion of energy from one form (chemical-bond) to another (mechanical).

Q: What are the energy conversions that occur when a woman dives from a board into a pool, splashes, and begins swimming?

There are two basic types of metabolism

Energy changes in living systems usually occur mostly as chemical changes, in which energy is stored in, or released from, chemical bonds.

Anabolic reactions (collectively anabolism) typically link smaller molecules to form larger, more complex molecules. Anabolic reactions require an input of energy. Energy is captured in the chemical bonds that are formed (for example, the glycosidic bond between the two monosaccharides forming sucrose). This captured energy is stored in the chemical bonds as potential energy:

$$\text{Glucose} + \text{fructose} + \text{energy} \rightarrow \text{Sucrose}$$

Catabolic reactions (collectively catabolism) break down larger, more complex molecules into smaller ones and often release the energy stored in the chemical bonds. For example, when sucrose is hydrolyzed, energy is released. In a biological system the released energy may be recaptured in new chemical bonds, or it may be used as kinetic energy—moving atoms, molecules, cells, or the whole organism:

$$\text{Sucrose} + H_2O \rightarrow \text{Glucose} + \text{fructose} + \text{energy}$$

Catabolic and anabolic reactions are often linked. The energy released in catabolic reactions is often used to drive anabolic reactions—that is, to do biological work. For example, the energy released by the breakdown of glucose (catabolism) is used to drive anabolic reactions such as the synthesis of triglycerides. This is why you can accumulate fat if you eat food in excess of your energy requirements.

The **laws of thermodynamics** (*thermo*, "energy," + *dynamics*, "change") were derived from studies of the fundamental physical properties of energy and the ways it interacts with matter. The laws apply to all matter and all energy transformations in the universe. Their application to living systems helps us understand how organisms and cells harvest and transform energy to sustain life.

The first law of thermodynamics: Energy is neither created nor destroyed

The first law of thermodynamics states that in any energy conversion, energy is neither created nor destroyed. In other words, during any conversion of energy—whether from one form to another or from one location to another—the total energy in the system under consideration before and after the conversion is the same (Figure 8.2A). As you will see in the next two chapters, the potential energy present in the chemical bonds of carbohydrates and lipids can be converted to potential energy in the form of adenosine triphosphate (ATP). This can then be converted into kinetic energy to do mechanical work (such as muscle contractions) or biochemical work (such as anabolism).

The second law of thermodynamics: Disorder tends to increase

Although energy cannot be created or destroyed, the second law of thermodynamics states that when energy is converted from one form to another, some of that energy becomes unavailable for doing work (Figure 8.2B). In other words, no physical process or

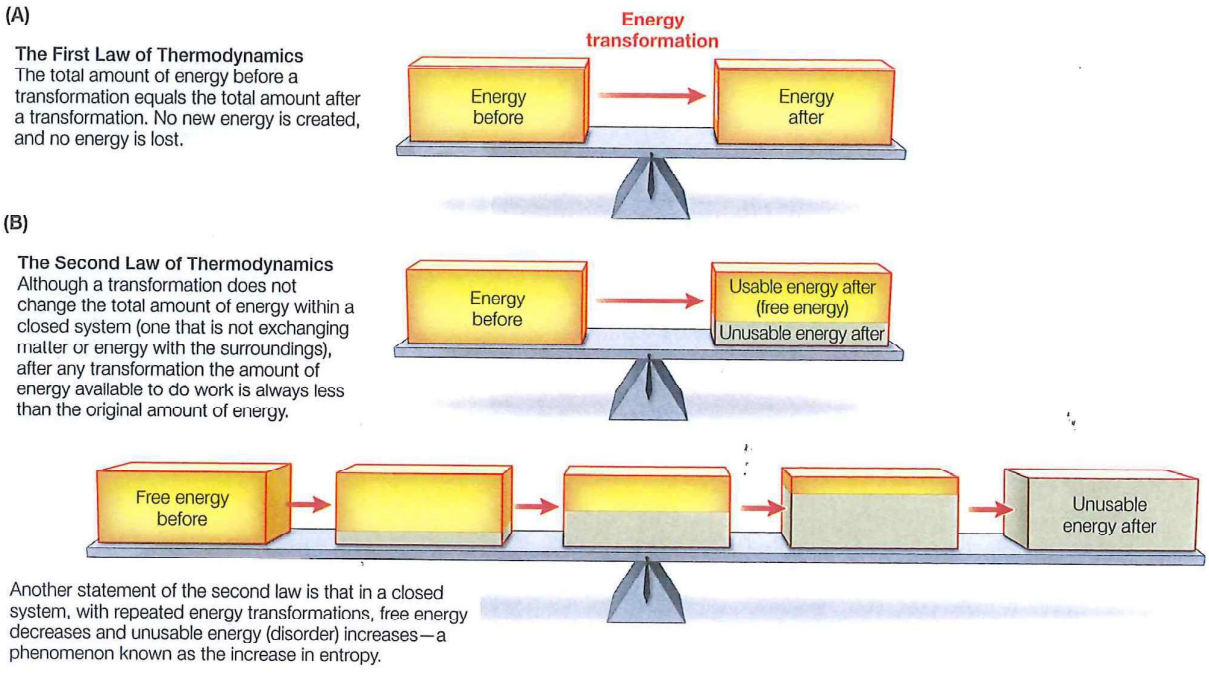


Figure 8.2 The Laws of Thermodynamics (A) The first law states that energy cannot be created or destroyed. (B) The second law states that after energy transformations, some energy becomes unavailable to do work.

chemical reaction is 100% efficient; some of the released energy is lost to a form associated with disorder. Think of disorder as a kind of randomness that is due to the thermal motion of particles; this energy is of such a low value and so spread out that it is unusable. **Entropy** is a measure of the disorder in a system.

It takes energy to impose order on a system. Unless energy is applied to a system, it will be randomly arranged or disordered. The second law applies to all energy transformations, but we will focus here on transformations in living systems. The second law applies on the micro scale, in chemical reactions and pathways in the cell. But it also applies on the macro scale, in communities of organisms interacting in the environment.

Connect the Concepts See Key Concept 56.3, which discusses the flow of energy in food chains, where there is a decline in usable energy in each step.

NOT ALL ENERGY CAN BE USED In any system, the total energy includes the usable energy that can do work and the unusable energy that is lost to disorder:

$$\text{Total energy} = \text{usable energy} + \text{unusable energy}$$

In biological systems, the total energy is called **enthalpy** (*H*). The usable energy that can do work is called **free energy** (G). Free energy is what cells require for all the chemical reactions involved in growth, cell division, and maintenance. The unusable energy is represented by entropy (*S*) multiplied by the absolute temperature (*T*). Thus we can rewrite the word equation above more precisely as:

$$H = G + TS \tag{8.1}$$

Because we are interested in usable energy, we rearrange Equation 8.1:

$$G = H - TS \tag{8.2}$$

Although we cannot measure *G*, *H*, or *S* absolutely, we can determine the *change* in each at a constant temperature. Such energy changes are measured in calories (cal) or joules (J).[1] A change is represented by the Greek letter delta (Δ). The change in free energy (Δ*G*) of any chemical reaction is equal to the difference in free energy between the products and the reactants:

$$\Delta G_{\text{reaction}} = G_{\text{products}} - G_{\text{reactants}} \tag{8.3}$$

A change in free energy can be either positive or negative; that is, the free energy of the products can be more or less than the free energy of the reactants. If the products have more free energy than the reactants (Δ*G* is positive), then there must have been some input of energy into the reaction. (Remember that energy cannot be created, so some energy must have been added from an external source.)

At a constant temperature Δ*G* is defined in terms of the change in total energy (Δ*H*) and the change in entropy (Δ*S*):

$$\Delta G = \Delta H - T\Delta S \tag{8.4}$$

[1] A calorie is the amount of heat energy needed to raise the temperature of 1 gram of pure water from 14.5°C to 15.5°C. In the SI system, energy is measured in joules. 1 J = 0.239 cal; conversely, 1 cal = 4.184 J. Thus, for example, 486 cal = 2,033 J, or 2.033 kJ. Although they are defined here in terms of heat, the calorie and the joule are measures of mechanical, electrical, or chemical energy. When you compare data on energy, always compare joules with joules and calories with calories.

Equation 8.4 tells us whether free energy is released or required by a chemical reaction:

- If Δ*G* is negative (Δ*G* < 0), free energy is released.
- If Δ*G* is positive (Δ*G* > 0), free energy is required.

If the necessary free energy is not available, the reaction does not occur. The sign and magnitude of Δ*G* depend on the two factors on the right side of the equation:

- Δ*H*: In a chemical reaction, Δ*H* is the total amount of energy added to the system (Δ*H* > 0) or released (Δ*H* < 0).
- Δ*S*: Depending on the sign and magnitude of Δ*S*, the entire term, *T*Δ*S*, may be negative or positive, large or small. In the chemistry lab, temperature can affect Δ*G*. But living systems don't vary their temperature that much, so the magnitude and sign of Δ*G* can depend on changes in entropy.

If a reaction increases entropy, its products are more disordered or random than its reactants. For example, when a protein is hydrolyzed to its amino acids, the products have considerable freedom to move around. The disorder in a solution of amino acids will be large compared with that in the protein, in which peptide bonds and other forces prevent free movement. So in hydrolysis, the change in entropy (Δ*S*) will be positive. Conversely, if there are fewer products and they are more restrained in their movements than the reactants (as for amino acids being joined in a protein), Δ*S* will be negative. This means that according to Equation 8.4, Δ*G* is positive and energy input is required.

DISORDER TENDS TO INCREASE Consider the human body, with its highly organized tissues and organs composed of large, complex molecules. You might think that this order and complexity are in conflict with the second law—after all, entropy, or disorder, should be maximized—but they are not, for two reasons:

1. *Getting ordered is coupled to the generation of disorder.* Making 1 kg of a human body (soft tissues, not bones) requires the catabolism of about 10 kg of highly ordered biological materials (our food), which are converted into CO_2, H_2O, and other simple molecules. So this process creates far more disorder (more energy is lost to entropy in the small molecules) than the amount of order (total energy; enthalpy) stored in large molecules in 1 kg of a person.
2. *Life requires a constant input of energy to maintain order.* Without this energy, the complex structures of living systems would break down. Because energy is used to generate and maintain order, there is no conflict with the second law of thermodynamics.

Having seen that the laws of thermodynamics apply to organisms, let's see how these laws apply to biochemical reactions inside the cell.

Chemical reactions release or consume energy

As you saw earlier, anabolic reactions link simple molecules to form more complex molecules, so they tend to increase complexity (order) in the cell. By contrast, catabolic reactions break down

complex molecules into simpler ones, so they tend to decrease complexity (generate disorder).

- Catabolic reactions may break down an ordered reactant into smaller, more randomly distributed products. Reactions that release free energy ($-\Delta G$) are called **exergonic** reactions (Figure 8.3A). For example:

  Complex molecules → free energy + small molecules

- Anabolic reactions may make a single product (a highly ordered substance) out of many smaller reactants (less ordered). Reactions that require or consume free energy ($+\Delta G$) are called **endergonic** reactions (Figure 8.3B). For example:

  Free energy + small molecules → complex molecules

In principle, chemical reactions are reversible and can run both forward and backward. For example, if compound A can be converted into compound B (A → B), then B, in principle, can be converted into A (B → A), although *the concentrations of A and B determine which of these directions will be favored.* You can think of the overall reaction as resulting from competition between the forward and reverse reactions (A ⇌ B). According to the law of mass action, increasing the concentration of A makes the forward reaction (A → B) happen more often than the reverse reaction, just as increasing the concentration of B favors the reverse reaction (B → A).

There are concentrations of A and B at which the forward and reverse reactions take place at the same rate. At these concentrations, no further *net change* in the system is observable, although individual molecules are still forming and breaking apart. This balance between forward and reverse reactions is known as **chemical equilibrium**. Chemical equilibrium is a state of no net change, and a state in which $\Delta G = 0$.

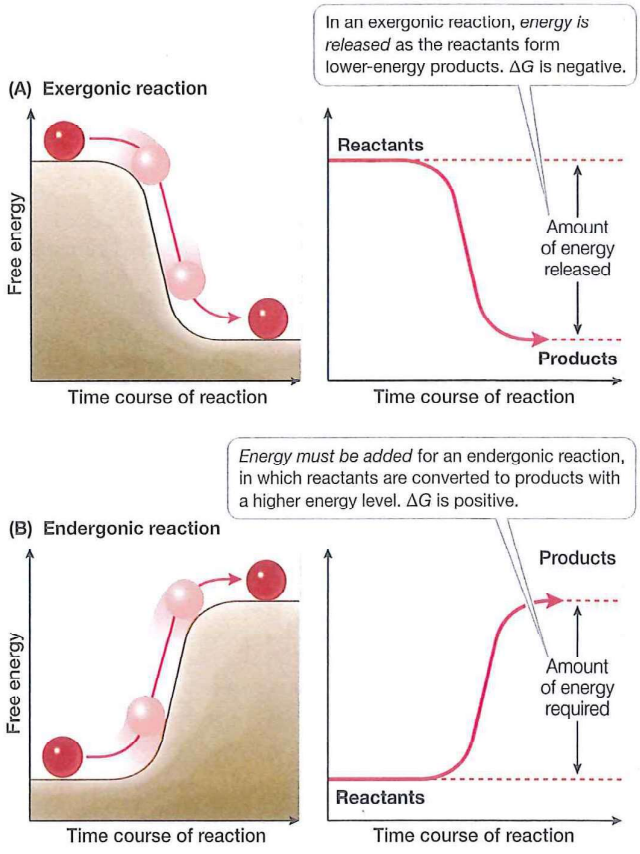


Figure 8.3 Exergonic and Endergonic Reactions (A) In an exergonic reaction, the reactants behave like a ball rolling down a hill, and energy is released. (B) A ball will not roll uphill by itself. Driving an endergonic reaction, like moving a ball uphill, requires the addition of free energy.

Chemical equilibrium and free energy are related

Chemical reactions do not necessarily proceed to completion, with all reactants converted into products. Each reaction has a specific equilibrium point, which is related to the relative free energy contents of the reactants and products. To understand the principle of equilibrium, consider the interconversion of glucose 1-phosphate and glucose 6-phosphate, a rearrangement of a phosphate group from one position on the ring of carbon atoms to another:

Glucose 1-phosphate ⇌ glucose 6-phosphate

Unraveling what goes on chemically in the cell is made possible by that fact that in many cases, after a cell's membrane has been ruptured to release its contents, the **cell-free system** with the cell contents (cytoplasm) in a test tube, usually in a buffered solution, has the same chemical properties as the intact cell. So the reaction can be studied in a test tube. Suppose we start out with an aqueous solution of glucose 1-phosphate that has a concentration of 0.02 M. (M stands for molar concentration; see Key Concept 2.4.) The solution is maintained under constant environmental conditions (25°C and pH 7). As the reaction proceeds to equilibrium, the concentration of the product, glucose 6-phosphate, rises from 0 M to 0.019 M, while the concentration of the reactant, glucose 1-phosphate, falls from 0.02 M to 0.001 M, at which point equilibrium is reached (Figure 8.4). *At equilibrium, the reverse reaction, from glucose 6-phosphate to glucose 1-phosphate, progresses at the same rate as the forward reaction.*

At equilibrium, then, this reaction has a product-to-reactant ratio of 19:1 (0.019/0.001), so the forward reaction has gone 95% of the way to completion ("to the right," as written above). This result is obtained every time the experiment is run under the same conditions.

The change in free energy (ΔG) for any reaction is related directly to its point of equilibrium. The further toward completion the point of equilibrium lies, the more free energy is released. In an exergonic reaction, ΔG is a negative number. The total value of ΔG also depends on the beginning concentrations of the reactants and products and other conditions such as temperature, pressure, and pH of the solution. Biochemists

KEY CONCEPT

14.1 Genes Code for Proteins

Learning Objectives

- **14.1.1** Describe the attributes of model organisms used for genetic studies.
- **14.1.2** Determine the order of intermediates in a biosynthetic pathway, given growth data for various mutant strains on these intermediates.
- **14.1.3** Determine the locations of mutant enzymes in a biosynthetic pathway, given growth data for the mutant strains.

In Chapter 4 we introduced DNA and its role in gene expression. Then in Chapter 13 we presented evidence that DNA is the carrier of genetic information and described how DNA is replicated prior to cell division. Here we focus on the evidence that proteins are the major products of gene expression, and we describe how a gene is expressed as protein.

Scientists had a molecular understanding of phenotypes before it was known that DNA was the genetic material; they had studied the chemical differences between individuals carrying wild-type and mutant alleles in organisms as diverse as humans and bread molds and knew that the major phenotypic differences resulted from differences in specific proteins. But just *how* different proteins arise in different individuals was not understood.

Observations in humans led to the proposal that genes determine enzymes

The identification of a gene product as a protein began with a mutation. In the early twentieth century, English physician Archibald Garrod saw several children with a rare disease. One symptom of the disease was that the urine turned dark brown or black when exposed to air, which was especially noticeable on the infants' diapers. The disease was given the descriptive name alkaptonuria ("black urine") (Figure 14.1A).

Garrod noticed that the disease was most common in children whose parents were first cousins. Mendelian genetics had just been "rediscovered," and Garrod realized that because first cousins share on average one-eighth of their alleles, the children of first cousins might inherit a rare mutant allele from both parents. He proposed that alkaptonuria was a recessive phenotype caused by a mutant allele that causes the disease phenotype.

Garrod took the analysis further by identifying the biochemical abnormality in the affected children. He isolated from them an unusual substance, homogentisic acid, which accumulated in blood, joints (where it crystallized and caused severe pain), and urine (where it turned black). The chemical structure of homogentisic acid is similar to that of the amino acid tyrosine (Figure 14.1B).

The function of enzymes as biological catalysts had just been discovered. Garrod proposed that homogentisic acid was a breakdown product of tyrosine. He suggested that while homogentisic acid ordinarily would be converted to a harmless product, it was not being converted in children with alkaptonuria. Garrod further hypothesized that the enzyme required for the breakdown of homogentisic acid was not being produced in these children. He suggested that a normal human allele (for the wild-type phenotype) was required for the synthesis of an enzyme that catalyzed this

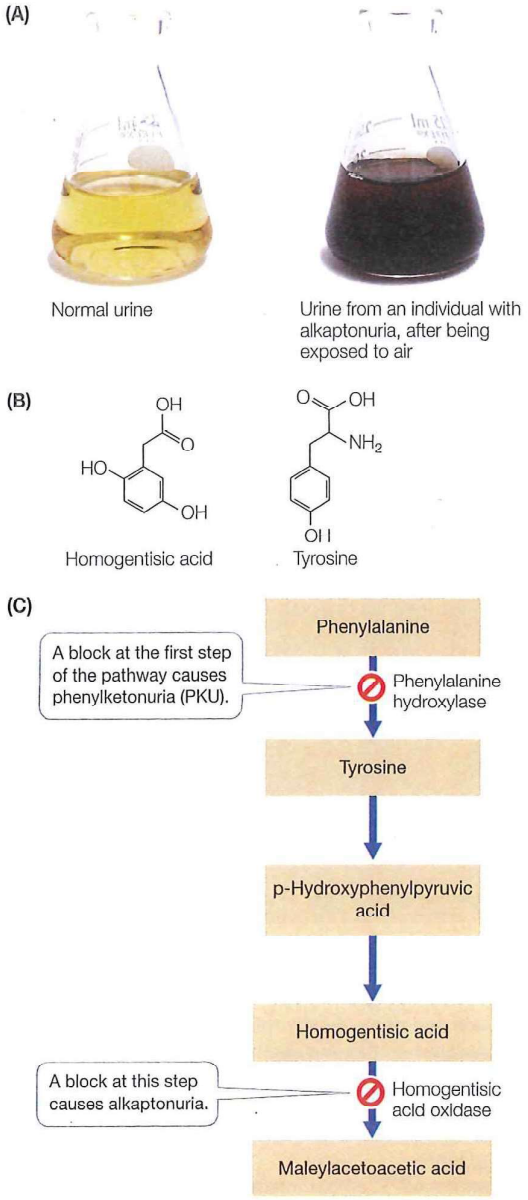


Figure 14.1 Alkaptonuria and Phenylketonuria (A) The physician Archibald Garrod showed that the accumulation of homogentisic acid in the urine of patients with alkaptonuria causes their urine to turn black when exposed to air. (B) Because the chemical structure of homogentisic acid is similar to that of the amino acid tyrosine, Garrod proposed that it was a breakdown product of the amino acid. (C) Both alkaptonuria and phenylketonuria are caused by mutations in specific enzymes in the pathway that breaks down phenylalanine.

conversion. If the allele were mutated, the enzyme would be inactive and homogentisic acid would accumulate instead.

Garrod concluded that there must be *one gene to one enzyme* and coined the term "inborn error of metabolism" to describe this genetically determined biochemical disease. But his hypothesis was not confirmed until the enzyme and pathway were identified (Figure 14.1C). The enzyme, homogentisic acid oxidase, was identified as active in healthy people and inactive in alkaptonuria patients in 1958. The specific DNA mutation that caused the enzyme to be nonfunctional was not described until 1996. An enzyme that catalyzes another step in this pathway is nonfunctional in individuals who have phenylketonuria (PKU). From the pathway, you can see that PKU causes elevated levels of phenylalanine, and this leads to significant intellectual disability if untreated. Fortunately it is now routine to test for high phenylalanine levels in the blood of newborn infants, and if an affected child consumes a diet low in phenylalanine, intellectual disability is avoided.

To relate genes and enzymes more generally, biologists turned to simpler organisms that could be manipulated in the laboratory.

Experiments on bread mold established that genes determine enzymes

As they work to explain the principles that govern life, biologists often turn to organisms that are easy to manipulate experimentally. Such **model organisms** have certain characteristics that make them attractive experimental subjects. For example, model organisms (1) are easy to grow in the laboratory or greenhouse, (2) have short generation times, (3) often produce large numbers of progeny, and (4) are easy to manipulate genetically, by crossing or by other methods. Biologists use model organisms to discover principles that can then be applied to other organisms. You have encountered some model organisms in previous chapters, including the pea plants (*Pisa sativum*) used by Mendel (see Key Concept 12.1), the fruit flies (*Drosophila*) used by Morgan (see Key Concept 12.4), and the *E. coli* bacteria used by Meselson and Stahl (see Key Concept 13.3).

To this list we now add the bread mold *Neurospora*, a type of sac fungus (see Chapter 28). In addition to having the traits shared by other model organisms, this mold is haploid for most of its life cycle, so there are no issues with dominance of alleles: all alleles are expressed phenotypically and not masked by a heterozygous condition. Biologists at Stanford University led by George Beadle and Edward Tatum undertook studies to biochemically define the phenotypes in *Neurospora*.

Like Garrod, Beadle and Tatum hypothesized that the expression of a specific gene results in the activity of a specific enzyme. They set out to test this hypothesis directly. They grew *Neurospora* on a nutritional medium containing sucrose, minerals, and biotin, which is the only vitamin that wild-type *Neurospora* cannot synthesize itself. Using this minimal medium, the enzymes of wild-type *Neurospora* could catalyze all the metabolic reactions needed for growth.

The scientists then treated the wild-type *Neurospora* with X-rays, which function as a mutagen. A **mutagen** is something that damages DNA, causing **mutations**: heritable alterations in the DNA sequence. After the X-ray treatment, some *Neurospora* strains could no longer grow on the minimal medium. These mutant strains grew only if they were supplied with specific additional nutrients, such as particular vitamins. Beadle and Tatum hypothesized that these genetic strains had mutations in the genes that code for production of enzymes needed to synthesize the additional nutrients. For each mutant strain, the scientists were able to find a single compound that, when added to the minimal medium, supported the growth of that strain. These results suggested that mutations have simple effects, and that each mutation causes a defect in only one enzyme in a metabolic pathway. These conclusions confirmed Garrod's **one-gene, one-enzyme hypothesis.**

Mutations provide a powerful way to determine cause and effect in biology. Nowhere has this been more evident than in the elucidation of biochemical pathways. Such pathways consist of sequential steps (chemical reactions) in which each event is dependent on the occurrence of the preceding event. If each reaction is catalyzed by an enzyme, and each enzyme is encoded by a gene, then it should be possible to block a pathway at any step by knocking out (rendering nonfunctional) the gene encoding that enzyme by mutation. Strains with mutated genes will thus not be able to synthesize the final product of the pathway and will exhibit a mutant phenotype. Addition of a compound in the pathway that occurs before the step with the mutated enzyme will not allow the mutant strain to produce the final product of the pathway. In contrast, addition of a compound occurring after the step with the mutated enzyme will allow the mutant strain to produce the final product of the pathway.

Two colleagues of Beadle and Tatum, Adrian Srb and Norman Horowitz, used this reasoning to determine characteristics of the biosynthetic pathway for arginine. They isolated *Neurospora* mutants that could not survive without the amino acid arginine in their growth medium. By adding particular compounds hypothesized to be in the arginine biosynthetic pathway to the medium, Srb and Horowitz were able to identify a series of steps in the biochemical pathway leading to the synthesis of arginine (Figure 14.2).

One gene determines one polypeptide

The one-gene, one-enzyme relationship has undergone several modifications in light of our current knowledge of molecular biology. Many proteins, including many enzymes, are composed of more than one polypeptide chain, or subunit (recall quaternary protein structure from Key Concept 3.2). So it is more correct to speak of a **one-gene, one-polypeptide relationship.**

Connect the Concepts Hemoglobin, illustrated in Figure 3.12 is an example of a protein composed of more than one polypeptide chain. Hemoglobin has four polypeptides—two α and two β subunits, and each kind of subunit is encoded by a separate gene.

So far we have seen that in terms of protein synthesis, the *function of a gene is to prescribe the production of a single, specific polypeptide.* But not all genes code for polypeptides. As we will see later in this chapter and in Chapter 16, many DNA sequences are transcribed to RNA molecules that are not translated into polypeptides, but instead have other functions.

Experiment

Figure 14.2A One Gene, One Enzyme

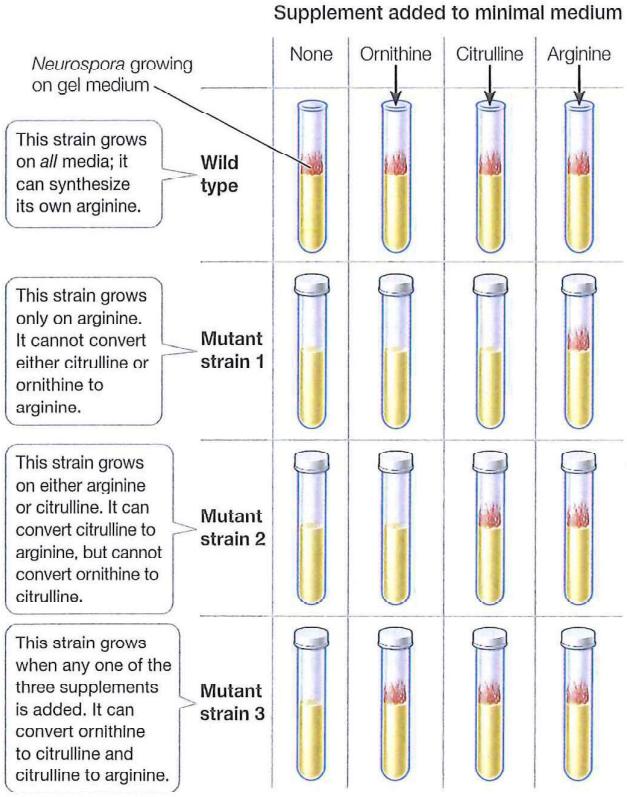
Original Paper: A. M. Srb and N. H. Horowitz. 1944. The ornithine cycle in *Neurospora* and its genetic control. *J Biol Chem* 154: 129–139.

Srb and Horowitz isolated several mutant strains of *Neurospora* that could not make arginine (arg). Several compounds, including ornithine and citrulline, are intermediates in the pathway for arginine synthesis. By systematically adding each of these molecules to the growth media for the mutant strains, the researchers deduced that each mutant strain was deficient in one enzyme along the arginine biosynthetic pathway.

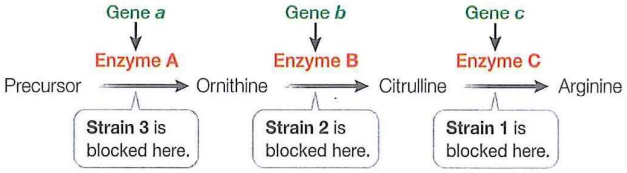
HYPOTHESIS▶ Each gene determines an enzyme in a biochemical pathway.

METHOD▶ Place each *arg* mutant strain on a minimal nutritional medium with and without supplements.

RESULTS▶



INTERPRETATION▶



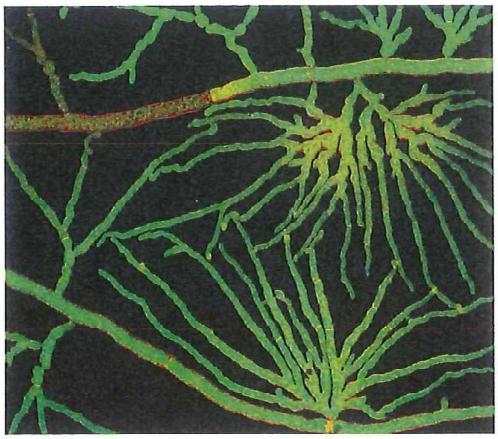
If an organism cannot convert one particular compound to another, it presumably lacks an enzyme required for conversion, and the mutation is in the gene that codes for that enzyme.

CONCLUSION▶ Each gene specifies a particular enzyme.

Work with the Data

Figure 14.2B One Gene, One Enzyme

Original Paper: A. M. Srb and N. H. Horowitz. 1944. The ornithine cycle in *Neurospora* and its genetic control. *J Biol Chem* 154: 129–139.



Courtesy of Eric Kalkman, University of Glasgow

Beadle and Tatum used X rays to cause mutations in *Neurospora*. They isolated mutant strains that were unable to grow on minimal medium but were able to grow if the medium was supplemented with particular compounds. Their colleagues Adrian Srb and Norman Horowitz analyzed 15 mutant strains (the *arg* mutants) that could not synthesize arginine but could grow on medium supplemented with arginine. The scientists tested various compounds and found two, ornithine and citrulline, that could be used instead of arginine to support the growth of some of the mutant strains (as seen in Figure 14.2A). The results for three of the strains are shown in the table, with growth expressed as dry weight in milligrams of fungal material after growth for 5 days.

Strain	No addition	Ornithine added	Citrulline added	Arginine added
34105	1.1	25.5	30.0	33.2
33442	2.3	2.5	42.7	43.8
36703	0.0	0.0	0.0	20.4

QUESTIONS▶

1. Based on the biochemical pathway for arginine synthesis shown in Figure 14.2A, which gene (*a*, *b*, or *c*) was mutated in each strain?
2. Why was there some growth in strains 34105 and 33442 even when there were no additions to the growth medium?
3. Nineteen other amino acids were tested as substitutes for arginine in the three strains. In all cases, there was no growth. Explain these results.

Go to **Achieve** for a companion **Data in Depth** exercise.

KEY CONCEPT
14.1 Recap and Assess

Studies of mutations in humans and bread molds led to our understanding of the one-gene, one-polypeptide relationship. In most cases, the function of a gene is to code for a specific polypeptide.

1. What is a model organism, and why is *Neurospora* a good model for studying biochemical genetics?
2. How were the experiments on mutant strains of *Neurospora* set up to determine the order of steps in a biochemical pathway?
3. In bacteria, the biosynthesis of the amino acid tryptophan (T) from the precursor chorismate (C) involves four intermediate chemical compounds, which we will call D, E, F, and G. Here are the phenotypes of various mutant strains. Each strain has a mutation in a gene for a different enzyme; + means growth with the indicated compound added to the medium, and 0 means no growth. Based on these hypothetical data, order the compounds (C, D, E, F, G, and T) and enzymes (1, 2, 3, 4, and 5) in the tryptophan biosynthetic pathway.

	Addition to medium					
Mutant strain	**C**	**D**	**E**	**F**	**G**	**T**
1	0	0	0	0	+	+
2	0	+	+	0	+	+
3	0	+	0	0	+	+
4	0	+	+	+	+	+
5	0	0	0	0	0	+

Now that you have seen the evidence for the one-gene, one-polypeptide relationship, how does it work? That is, how is the information encoded in DNA used to produce a particular polypeptide?

KEY CONCEPT
14.2 Information Flows from Genes to Proteins

Learning Objective

14.2.1 Describe the central dogma of molecular biology and give examples of exceptions.

As we discussed in Chapter 13 and in Key Concept 14.1, DNA is the hereditary material and codes for RNAs and proteins. In the remainder of this chapter we focus on the processes that occur when a protein-coding gene is expressed. We briefly outlined gene expression in Key Concept 4.1. To review, this process occurs in two major steps:

1. During **transcription**, the information in a DNA sequence (a gene) is copied into a complementary RNA sequence.
2. During **translation**, this RNA sequence is used to create the amino acid sequence of a polypeptide.

The "**central dogma** of molecular biology" describes the primary ways in which information flows in a cell: from DNA to DNA, from DNA to RNA, and from RNA to protein (Figure 14.3). We have already discussed replication (see Key Concept 13.3) and now turn to transcription and translation.

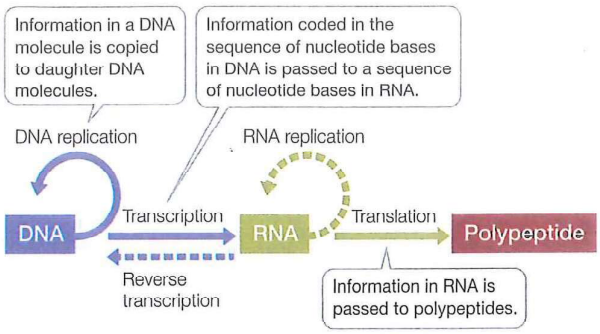


Figure 14.3 The Central Dogma of Molecular Biology
The central dogma describes the primary ways in which information flows in the cell: from DNA to DNA, from DNA to RNA and from RNA to protein (solid lines). Information is stored in DNA and copied during replication. Certain DNA sequences are transcribed into RNA, and some RNA sequences are translated into amino acid sequences. Substantially less common, RNA replication (in some RNA viruses) and synthesis of DNA from RNA (reverse transcription) can occur (dashed lines). In no case is amino acid information used to synthesize RNA.

Q: Have you seen any examples in previous chapters where information flow does not follow the central dogma (dashed lines)?

Three types of RNA have roles in the information flow from DNA to protein

There are numerous types of RNA. Three of them have vital roles in gene expression and represent the most common RNAs in a typical cell.

MESSENGER RNA When a protein-coding gene is expressed, one of the two DNA strands in the gene is transcribed to produce a complementary RNA strand, which in eukaryotic cells is then processed to produce **messenger RNA (mRNA)**. In eukaryotic cells, the mRNA travels from the nucleus to the cytoplasm, where it is translated into a polypeptide (Figure 14.4). The nucleotide sequence of the mRNA determines the ordered sequence of amino acids in the polypeptide chain, which is synthesized by a ribosome. mRNAs represent about 5% of the RNA in a cell.

RIBOSOMAL RNA The **ribosome** is essentially a protein synthesis factory composed of multiple proteins and several **ribosomal RNAs (rRNAs)**. One of the rRNAs catalyzes peptide bond formation between amino acids, to form a polypeptide. rRNAs represent about 80% of the RNA in a cell.

TRANSFER RNA Another type of RNA called **transfer RNA (tRNA)** can both bind a specific amino acid and recognize specific sequences of nucleotides in mRNA. It is the tRNA that recognizes which amino acid should be added next to a growing polypeptide chain. tRNAs represent about 15% of the RNA in a cell.

Media Clip 14.1 Protein Synthesis: An Epic on a Cellular Level
Life12e.com/mc14.1

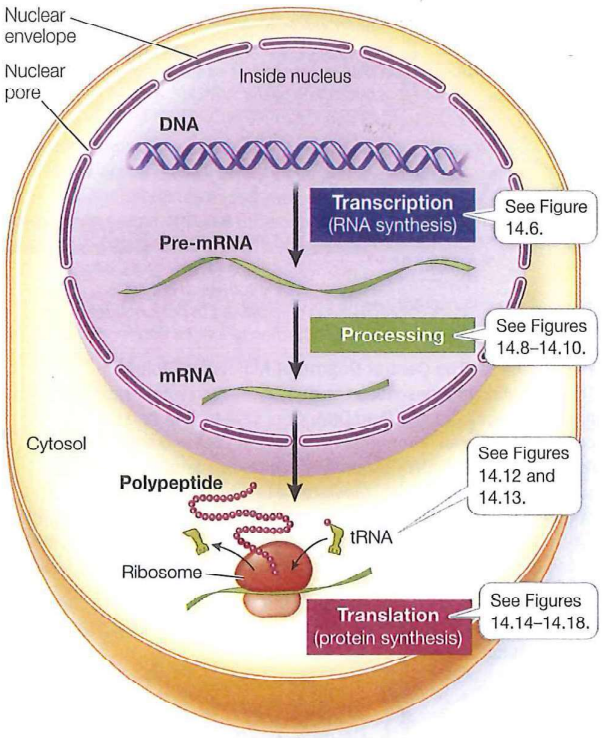


Figure 14.4 From Gene to Protein This diagram summarizes the processes of gene expression in eukaryotes. Note that the nucleus is typically about one-fourth the size shown here.

Q: In general, how would this diagram be different for a prokaryotic cell?

View in Achieve
Activity 14.1 Eukaryotic Gene Expression

In some cases, RNA determines the sequence of RNA or DNA

While we have said that DNA is the genetic material, some viruses present exceptions to this aspect of the central dogma. As we saw in Key Concept 13.1, a virus is a non-cellular infectious particle that reproduces inside cells. Many viruses, such as the ones that cause influenza and polio, have RNA rather than DNA as their genetic material. That is, the nucleotide sequence of viral RNA acts as an information carrier and can be expressed as a protein. Because RNA is usually single-stranded, the question arises: how do these viruses replicate? More specifically, how do they duplicate their genetic material? Most viruses replicate by transcribing from RNA to RNA, making an RNA strand that is complementary to their genomes. This "opposite" strand is then used to make multiple copies of the viral genome.

Not all viruses whose genomes consist of RNA replicate by transcribing from RNA to RNA. Some, such as human immunodeficiency viruses (HIVs) and certain rare tumor viruses, make a DNA copy of their genome after infecting a host cell. This DNA copy is then incorporated into the host's genome. Synthesis of DNA from RNA is called **reverse transcription**, and viruses that employ this kind of transcription are called **retroviruses**. Retroviruses rely on the host cell's transcription machinery to make more RNA. This RNA can either be translated to produce viral proteins, or incorporated as the viral genome into new viral particles. You might also recall that telomerase extends telomeres by synthesizing DNA from an RNA template and is thus another example of reverse transcription.

KEY CONCEPT

14.2 Recap and Assess

The expression of protein-coding genes can be broken down into two fundamental steps: transcription and translation. In transcription one strand of DNA is used to produce a complementary mRNA, whose sequence determines the mature mRNA transcript. The mRNA sequence determines the order of amino acids in a polypeptide. Translation is the process by which this information in mRNA is converted into a polypeptide chain.

1. What is the central dogma of molecular biology?
2. Do retroviruses violate the central dogma? Explain.

We will revisit viral genetics in later chapters (in Key Concepts 16.3 and 24.4). In the rest of this chapter we focus on gene expression in prokaryotes and eukaryotes. Understanding this process is essential for understanding how organisms function at the molecular level and is key to the application of biology to human welfare, in areas such as agriculture and medicine. We'll begin by describing how the information in DNA is transcribed to produce RNA.

KEY CONCEPT 14.3 DNA Is Transcribed to Produce RNA

Learning Objectives

14.3.1 Describe characteristics of RNA polymerases.
14.3.2 Describe the function of the promoter and terminator sequences in transcription.

The synthesis of RNA is directed by DNA. The base sequence of one strand of DNA is used as a template for RNA synthesis, so that the RNA made is complementary in sequence to that DNA strand, with the exception that in RNA there is uracil (U) instead of thymine (T) and in RNA the sugar is ribose instead of deoxyribose. It is important to realize that although the RNA made is a mirror image of its DNA template, the RNA has the same sequence as the other, non-template strand of DNA. So the information content of DNA is indeed preserved in RNA. Since the sequence of nucleotides in the non-template strand is the same as in the protein-coding RNA (except with Ts instead of Us) it is usually referred to as the coding strand.

Transcription—the formation of a specific RNA sequence from a specific DNA sequence—requires several components:

- A DNA template for complementary base pairing—one of the two strands of DNA

TABLE 14.1 | Some RNAs in Eukaryotic Cells

RNA type	Location of activity	Role
Ribosomal RNA (rRNA)	Cytoplasm (ribosome)	Component of ribosome, catalysis of peptide bond formation
Messenger RNA (mRNA)	Cytoplasm	Carrier of protein-coding sequence
Transfer RNA (tRNA)	Cytoplasm	Intermediary between mRNA and protein sequences
MicroRNA (miRNA)	Nucleus and cytoplasm	Regulates mRNA stability and translation
Small interfering RNA (siRNA)	Nucleus and cytoplasm	Regulates stability of other RNAs
Small nuclear RNA (snRNA)	Nucleus	Mediates mRNA processing

- The four ribonucleoside triphosphates ATP, GTP, CTP, and UTP, to act as substrates
- An RNA polymerase enzyme
- Salts and a pH buffer to create an appropriate chemical environment for RNA polymerase (if transcription is performed in a test tube)

Several kinds of RNA are made from DNA templates. The most important from a genetic point of view is mRNA. But transcription also produces tRNA and rRNA, whose roles in protein synthesis are described in Key Concept 14.6. Like polypeptides, these last two RNAs are encoded by specific genes. Eukaryotes also make many kinds of small RNAs, including small nuclear RNA (snRNA), microRNA (miRNA), and small interfering RNA (siRNA), which are also transcribed. Table 14.1 summarizes some of the RNAs found in eukaryotic cells. We will discuss the roles of miRNA and siRNA in Chapter 16.

RNA polymerases share common features

RNA polymerases from both prokaryotes and eukaryotes catalyze the synthesis of RNA from the DNA template. There is only one kind of RNA polymerase in bacteria, whereas there are several kinds in eukaryotes; however, they all share a common structure (Figure 14.5). Like DNA polymerases, RNA polymerases catalyze the addition of nucleotides in a 5′-to-3′ direction, but unlike DNA polymerases (see Figure 13.12), RNA polymerases *do not require a primer.*

Transcription occurs in three steps

Transcription can be divided into three distinct processes: (1) initiation, (2) elongation, and (3) termination. Follow these processes in Figure 14.6.

INITIATION Transcription begins when RNA polymerase binds to a special sequence of DNA called a **promoter** (see Figure 14.6A). Eukaryotic genes generally have one promoter each, whereas in prokaryotes and viruses, several genes often share one promoter. Promoters are important control sequences that "tell" the RNA polymerase two things:

1. Where to start transcription
2. Which strand of DNA to transcribe

A promoter has directionality, so the RNA polymerase is oriented to the appropriate strand to use as a template. The promoter also determines the **initiation site**, where transcription begins. Groups of nucleotides lying "upstream" from the initiation site (5′ on the coding strand and 3′ on the template strand) help the RNA polymerase bind. Other proteins, which can bind to specific DNA sequences and to RNA polymerase, help direct the polymerase onto the promoter. These proteins, which include **sigma factors** in prokaryotes and **transcription factors** in eukaryotes, help determine the specific genes that are expressed at a particular time in the cell.

Although every gene has a promoter, not all promoters are identical. Some are more effective at transcription initiation than others. Furthermore, there are differences between transcription initiation in prokaryotes and in eukaryotes. We will discuss promoters and their roles in regulating gene expression in Chapter 16.

ELONGATION After RNA polymerase has bound to the promoter, it begins the process of **elongation** (see Figure 14.6B). DNA unwinds about ten base pairs at a time, and RNA polymerase

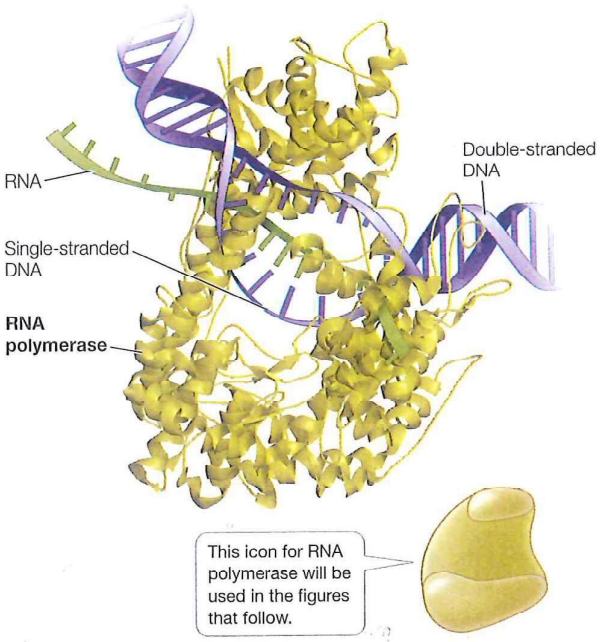


Figure 14.5 RNA Polymerase Interacting with DNA The RNA polymerase enzyme depicted (in yellow-brown) is from bacteriophage T7, but it is representative of most other RNA polymerases. The inset indicates how this enzyme will be represented in the figures that follow. (RNA polymerase by David McIntyre based on data from PDB ID: 1MSW. Y. W. Yin and T. A. Steitz. 2002. *Science* 298: 1387–1395.)

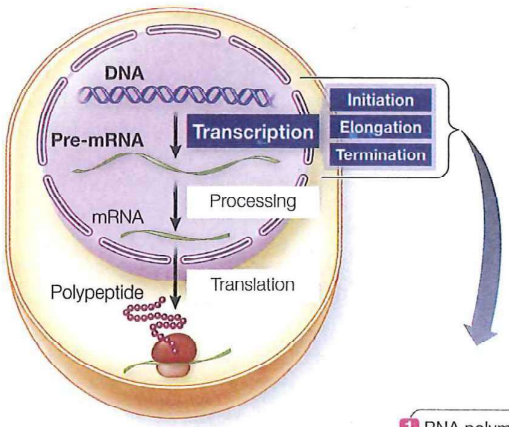
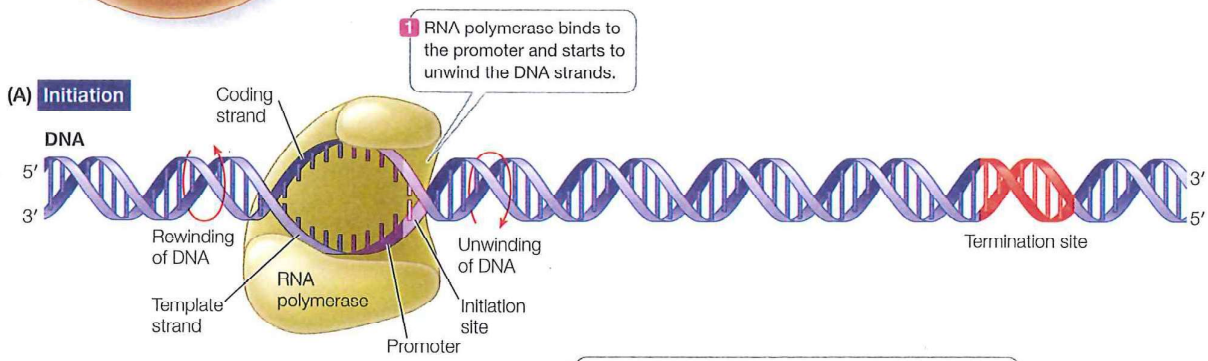
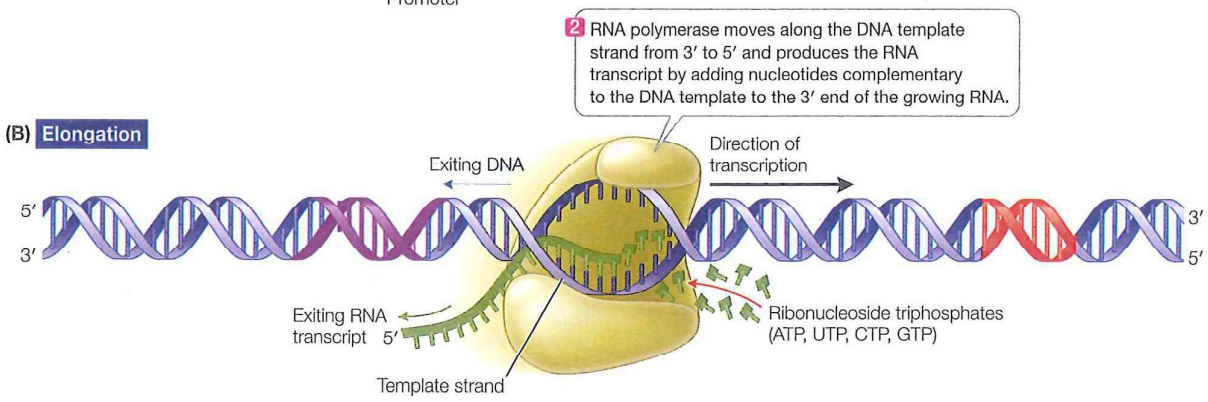


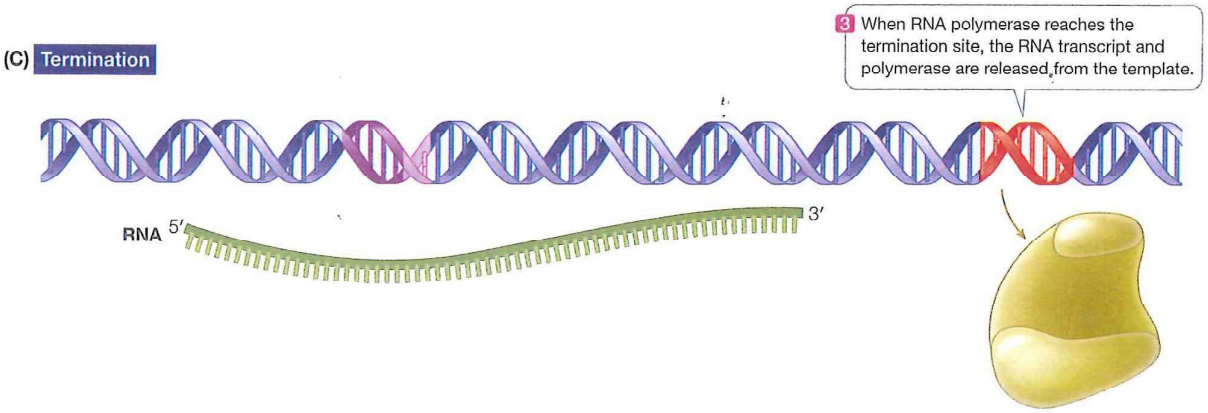
Figure 14.6 **DNA Is Transcribed to Form RNA** DNA is locally unwound by RNA polymerase, and one strand then serves as a template for RNA synthesis. The RNA transcript is formed and then peels away, allowing the DNA that has already been transcribed to rewind into a double helix. Three distinct processes—initiation (A), elongation (B), and termination (C)—constitute transcription. RNA polymerase is larger in reality than indicated here (see Figure 14.5).

Q: Comparing RNA synthesis with DNA replication, what are the common features?

View in Achieve
Animation 14.1 Transcription







reads the template strand in the 3′-to-5′ direction. The first nucleotide in the new RNA forms its 5′ end, and subsequent nucleotides complementary to the DNA template are added to its 3′ end. Thus the RNA transcript is antiparallel to the DNA template strand.

You may recall from Key Concept 13.3 that DNA polymerase uses dNTPs (deoxyribonucleoside triphosphates) as substrates, and forms covalent bonds between each incoming dNTP and the 3′ end of the growing polynucleotide chain (see Figure 13.10). Energy released by the removal of two phosphate groups from the dNTP is used to drive the reaction. Similarly, RNA polymerase uses (ribo)nucleoside triphosphates (NTPs) as substrates, removing two phosphate groups from each substrate molecule and using the released energy to drive the polymerization reaction.

Because RNA polymerases do not proofread, transcription errors occur at a rate of one for every 10^4 to 10^5 bases. Because many copies of RNA are made, however, and because they often have only a relatively short life span, these errors are not as potentially harmful as mutations in DNA.

TERMINATION Just as initiation sites in the DNA template strand specify the starting point for transcription, particular base sequences specify its **termination** (see Figure 14.6C). In prokaryotes there are two mechanisms for ending transcription. For some genes, the newly formed transcript folds back on itself and forms internal hydrogen bonds between bases. A loop forms, and this structure causes the transcript to fall away from the DNA template and the RNA polymerase. In other cases a protein binds to specific sequences on the transcript and causes the RNA to detach from the DNA template. Eukaryotes terminate transcription using methods that are similar to, but more complicated than, those found in prokaryotes.

KEY CONCEPT
14.3 Recap and Assess

Transcription, which is catalyzed by an RNA polymerase, proceeds in three steps: initiation, elongation, and termination.

1. Describe the actions of RNA polymerase during transcription.
2. Knocking out (rendering nonfunctional) a particular protein in a bacterium causes many, but not all, genes to produce longer mRNAs. Explain this result.
3. Errors in RNA transcription occur about 100,000 times more often than errors in DNA replication. Why can this higher rate be tolerated in RNA but not in DNA synthesis?

The general features of transcription that we have described were first elucidated in model prokaryotes, such as *E. coli*. Biologists then used the same methods to analyze this process in eukaryotes and found that there are some notable (and important) differences. We turn now to a more detailed description of eukaryotic gene expression.

KEY CONCEPT
14.4 Eukaryotic Pre-mRNA Transcripts Are Processed prior to Translation

Learning Objectives

14.4.1 Describe the mRNA processing steps in eukaryotes.
14.4.2 Predict the outcome of a defect in mRNA processing in a eukaryotic cell.
14.4.3 Compare the features of a eukaryotic gene with its mature mRNA transcript.

The process of gene expression from DNA to protein is essentially the same in eukaryotes as it is in prokaryotes. In the last key concept we described the basic features of transcription that are common to both. Here we focus on the differences, listed in Table 14.2, between prokaryotic and eukaryotic gene transcription.

In prokaryotes and eukaryotes, the sequence of an mRNA that reaches the ribosome is complementary to the sequence of a gene in the organism's DNA. One way to show this is by the technique of **nucleic acid hybridization**, shown in Figure 14.7A. This technique involves two steps:

1. A sample of chromosomal DNA containing the gene is denatured to break the hydrogen bonds between the base pairs and separate the two strands.
2. The single-stranded mRNA (called a **probe**) is incubated with the denatured DNA. If the probe has a base sequence complementary to the target DNA, a probe–target double helix forms by hydrogen bonding between the bases. Because the two strands are from different sources, the resulting double-stranded region is called a hybrid.

Hybridization experiments can be performed with various combinations of DNA and RNA (RNA as target and DNA as probe; DNA as both target and probe, etc.). In many hybridization experiments,

TABLE 14.2 | Differences between Prokaryotic and Eukaryotic Gene Expression

Characteristic	Prokaryotes	Eukaryotes
Transcription and translation occurrence	At the same time, in the cytoplasm	Transcription in the nucleus, then translation in the cytoplasm
Gene structure	DNA sequence aligns perfectly with mRNA sequence	Regions of DNA in gene do not align with mature RNA (introns), and RNA 3′ region does not align with DNA (poly A tail)
Modification of mRNA after initial transcription but before translation	Usually none	Introns spliced out; 5′ end cap and 3′ poly A tail added

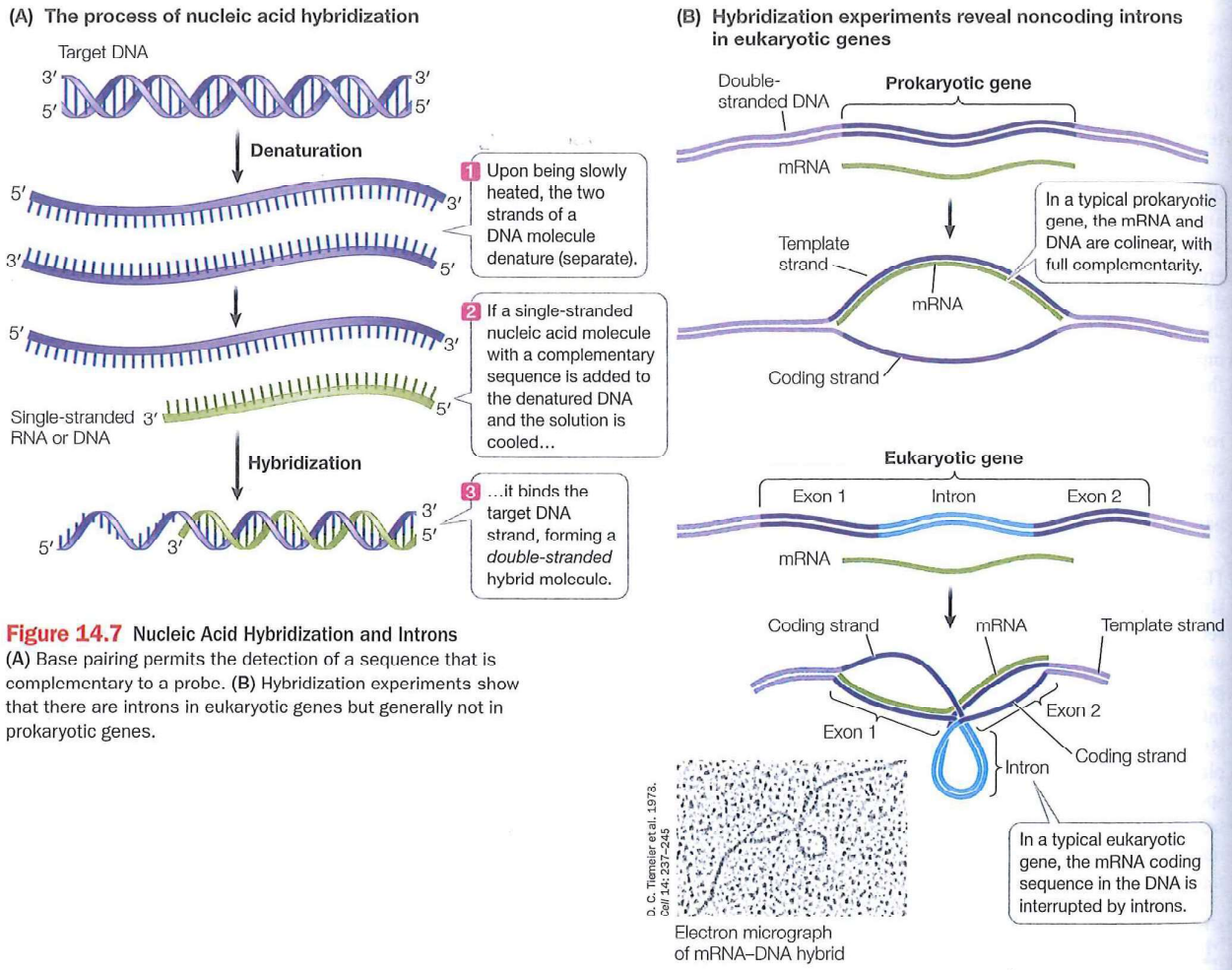


Figure 14.7 Nucleic Acid Hybridization and Introns
(A) Base pairing permits the detection of a sequence that is complementary to a probe. **(B)** Hybridization experiments show that there are introns in eukaryotic genes but generally not in prokaryotic genes.

the probe is labeled in some way so that its binding to a specific target sequence can be detected. The double-stranded hybrids can also be viewed by electron microscopy.

Noncoding sequences called introns often appear between genes in eukaryotic chromosomes

Differences between prokaryotic and eukaryotic transcription are revealed when mRNA probes from prokaryotes and eukaryotes are incubated with their respective chromosomal DNAs:

- In prokaryotes (Figure 14.7B, top) there is usually a 1:1 linear complementarity between the base sequence of the mRNA at the ribosome and that of the chromosomal DNA.
- In eukaryotes (Figure 14.7B, bottom) one or more non-hybridized DNA loops are often observed sticking out of the mRNA–DNA hybrid, indicating that there are stretches of DNA sequence that do not have a complementary sequence in the mRNA that is translated at the ribosome.

The discovery that some stretches of eukaryotic DNA sequence are not present in the mRNA that reaches the ribosome for translation initiated the question of whether this "extra" DNA actually gets transcribed. Does transcription somehow skip these sequences, or are they transcribed and then edited out of the mRNA transcript before it arrives at the ribosome? To find out, an experiment can be conducted in which the initial mRNA transcript in the cell nucleus—the **precursor mRNA**, or **pre-mRNA** (see Figure 14.4)—is hybridized with chromosomal DNA. This experiment shows full, linear, loop-free hybridization between the pre-mRNA and the template DNA, allowing us to conclude that the noncoding, intervening regions of DNA, known as **introns**, do get transcribed but are then spliced out of the pre-mRNA in the nucleus. The remaining sequences (**exons**) make up the mRNA that reaches the ribosome. The step of splicing out the introns is one of the steps in **RNA processing** (Figure 14.8).

Introns *interrupt, but do not scramble*, the DNA sequence of a gene. The base sequences of the exons in the template strand, if joined and taken in order, form a continuous sequence that is complementary to that of the mature mRNA. In some cases the separated exons often encode different functional regions, or **domains**, of the protein. For example, the globin polypeptides that make up

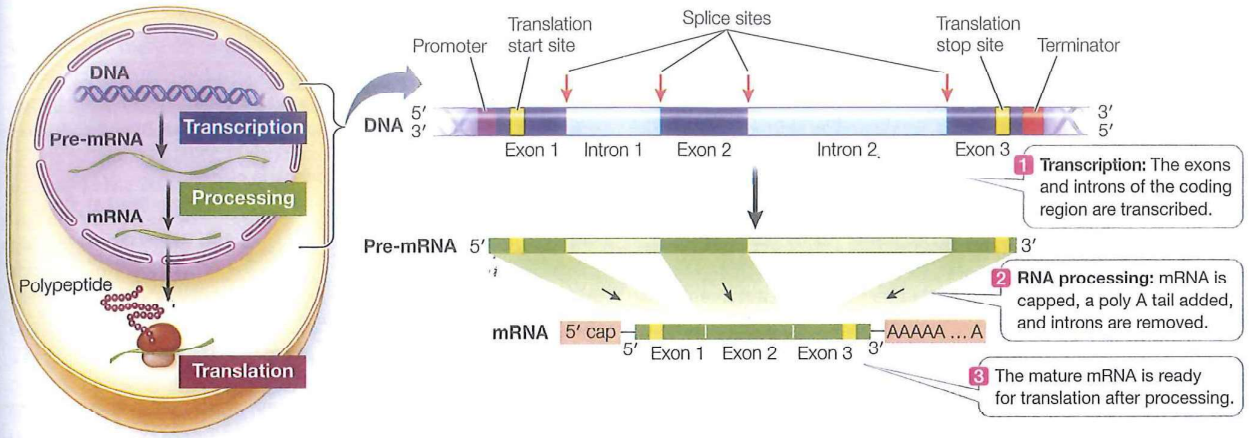


Figure 14.8 Transcription of a Eukaryotic Gene The β-globin gene diagrammed here is about 1,700 base pairs (bp) long. The three exons contain the sequence encoding 147 amino acids plus a signal (translation stop site) to terminate translation. The two introns—noncoding sequences of DNA containing almost 1,000 bp among them—are initially transcribed but are spliced out of the pre-mRNA transcript during processing.

hemoglobin each have two domains: one for binding to a nonprotein pigment called heme, and another for binding to the other globin subunits. These two domains are encoded by different exons in the globin genes. Most (but not all) eukaryotic genes contain introns, and in rare cases, introns are also found in prokaryotes. Introns can be numerous and large, though the average human gene has only eight, each with an average length of 3,300 bp. The largest human gene encodes a muscle protein called titin; it has 365 exons, which together code for 24,000–36,000 amino acids.

Pre-mRNA processing prepares the mRNA transcript for translation

The transcript of a eukaryotic gene is modified in several ways before it leaves the nucleus: both ends of the pre-mRNA are modified, and the introns are removed.

MODIFICATION AT BOTH ENDS Two steps in the processing of pre-mRNA take place in the nucleus, one at each end of the molecule (Figure 14.9):

1. A **5′ cap** is added to the 5′ end of the pre-mRNA as it is transcribed. The 5′ cap, which is a chemically modified molecule of guanosine triphosphate (GTP), facilitates the binding of mRNA to the ribosome for translation, and it protects the mRNA from being digested by ribonucleases that break down RNAs.
2. A **poly A tail** is added to the 3′ end of the pre-mRNA at the end of transcription. Transcription ends downstream of the protein-coding region in DNA. In eukaryotes there is usually a "polyadenylation" sequence (AAUAAA) near the 3′ end of the pre-mRNA, downstream of the protein-coding region. This sequence acts as a signal for an enzyme to cut the pre-mRNA. Immediately after this cleavage, another enzyme adds 100 to 300 adenine nucleotides (the poly A tail) to the 3′ end of the pre-mRNA. This tail helps in the export of mature mRNA from the nucleus and is important for mRNA recognition and stability in the cytosol.

SPLICING TO REMOVE INTRONS Within the nucleus, a process called **RNA splicing** removes the introns and splices the exons together. You can follow the steps of this important process in Figure 14.10. There are specific sequences that are commonly found in introns (consensus sequences) that allow the splicing machinery to identify and precisely remove introns from pre-mRNAs. The splicing machinery includes proteins and RNA–protein complexes called small nuclear ribonucleoproteins (snRNPs) that together recognize and cut out each intron.

Molecular studies of human genetic diseases have provided insights into RNA splicing. For example, people with the genetic

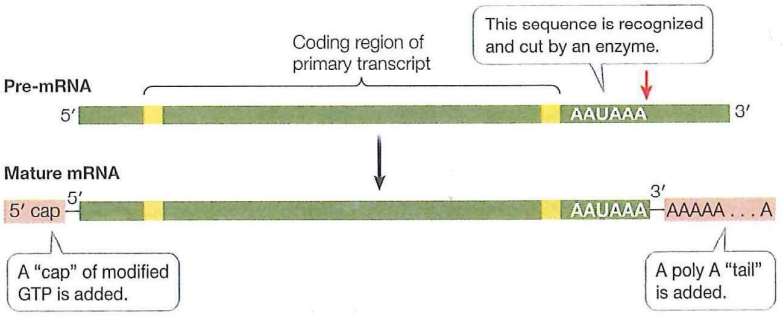


Figure 14.9 Processing the Ends of Eukaryotic Pre-mRNA Modifications at each end of the pre-mRNA transcript—the 5′ cap and the poly A tail—are important for mRNA function.

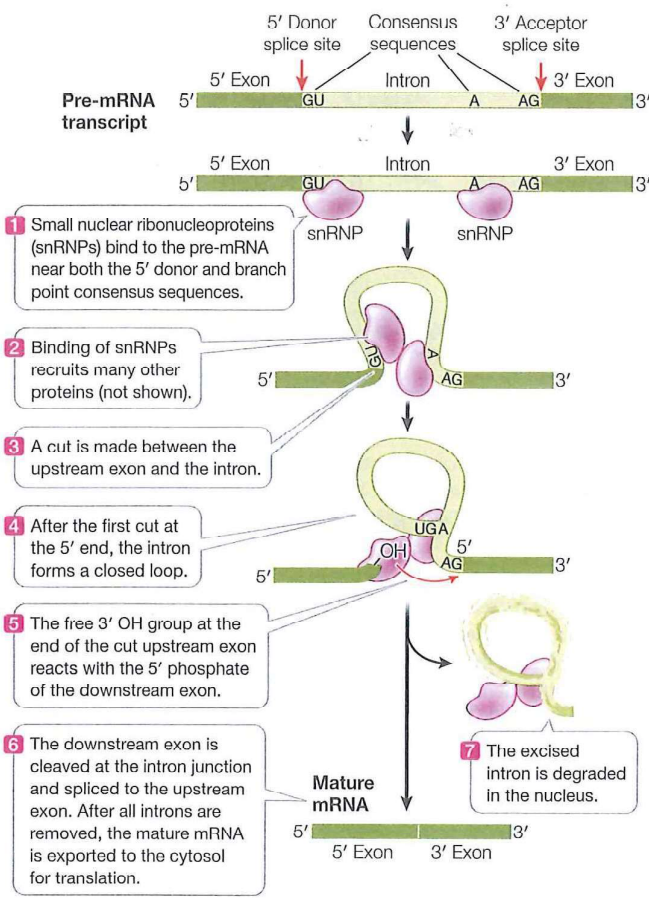


Figure 14.10 RNA Splicing The binding of snRNPs to consensus sequences bordering the introns on the pre-mRNA results in a series of proteins binding and cutting the pre-mRNA with great precision.

View in Achieve
Animation 14.2 RNA Splicing

disease beta thalassemia have a defect in the production of one of the hemoglobin subunits. These people suffer from severe anemia because they have an inadequate amount of hemoglobin in their red blood cells. In some cases, the genetic mutation that causes the disease occurs at an intron consensus sequence, where the splicing machinery binds to the RNA (see Figure 14.10, step 1) in the β-globin gene. Consequently, β-globin pre-mRNA is spliced incorrectly, and the resulting mRNA encodes a nonfunctional polypeptide.

After processing is completed in the nucleus, the mature mRNA has a polypeptide-coding region that will be translated by ribosomes, a 5′ untranslated region (upstream of the polypeptide-coding region), and a 3′ untranslated region (downstream of the polypeptide-coding region). The mature mRNA moves out into the cytosol through the nuclear pores. A protein bound to the 5′ nucleotide cap during processing is recognized by a receptor at the nuclear pore, stimulating export. Unprocessed or incompletely processed pre-mRNAs remain in the nucleus.

KEY CONCEPT

14.4 Recap and Assess

Most eukaryotic genes contain noncoding sequences called introns, which are removed from the pre-mRNA transcript in the nucleus. The mature mRNA transcript with 5′ cap and 3′ tail is then exported through a nuclear pore to ribosomes in the cytosol where translation takes place.

1. How and why is the pre-mRNA transcript modified at the 5′ and 3′ ends during mRNA processing?
2. Outline the steps involved in RNA splicing. What are the consequences if it does not happen correctly?
3. A eukaryotic gene that is 1,440 base pairs long codes for a polypeptide that is 192 amino acids long. However, only 576 base pairs encode the polypeptide. Discuss the discrepancy.

Transcription and posttranscriptional events produce an mRNA that is ready to be translated into a sequence of amino acids in a polypeptide. We turn now to the relationship between mRNA sequence and amino acid sequence.

KEY CONCEPT

14.5 The Genetic Code Determines the Protein Sequence Encoded by an mRNA

Learning Objectives

14.5.1 Use the genetic code table to determine the amino acid sequence encoded by an mRNA sequence.

14.5.2 Define redundancy in the genetic code.

14.5.3 Distinguish between sense and nonsense, or stop, codons.

The **genetic code** is the informational key by which a sequence of mRNA nucleotides corresponding to a gene is translated into the sequence of amino acids composing the protein expressed by that gene. That is, the genetic code specifies which amino acids will be used to build a protein. You can think of the code as consisting of a series of sequential, non-overlapping, three-letter "words." The three "letters" represent three adjacent nucleotide bases in the mRNA polynucleotide. Each three-letter "word" is called a **codon**, and each codon specifies a particular amino acid. Each codon is complementary to the corresponding triplet of bases in the DNA molecule from which it was transcribed. In short, the genetic code relates codons to their specifically encoded amino acids.

The genetic code is redundant but not ambiguous

Molecular biologists "broke" the genetic code in the early 1960s. The problem they addressed was perplexing: how could more than 20 "code words" be written with an "alphabet" consisting of only four "letters"? In other words, how could four bases (A, U, G, and C) code for 20 different amino acids?

A triplet code, based on three-letter codons, was considered likely. Since there are only four letters (A, G, C, and U), a one-letter code clearly could not unambiguously encode 20 amino acids; it could encode only four of them. A two-letter code could have only $4 \times 4 = 16$ unambiguous codons—still not enough. But a triplet code could have $4 \times 4 \times 4 = 64$ codons, more than enough to encode the 20 amino acids.

Marshall W. Nirenberg and J. Heinrich Matthaei, at the U.S. National Institutes of Health, made the first decoding breakthrough in 1961 when they realized that the code would be easier to break if they were working with a very simple, known mRNA sequence rather than with a complex natural mRNA molecule. They set out to synthesize mRNA molecules consisting of just one type of nucleotide base; poly U mRNA, for example, consisted of just uracil nucleotides. Nirenberg and Matthaei's goal was to then identify, through a translation process conducted in a test tube, the polypeptide that the artificial messenger encoded. Their experiment, which is laid out in **Investigating Life: Deciphering the Genetic Code**, led to the identification of the first codons. Soon after this experiment, the rest of the code was identified. This was a major achievement, linking the information in DNA (the gene) to its expression in a protein (the phenotype). The understanding that the amino acids for each protein are spelled out via codons not only led to our understanding of the fundamentals of genetics and mutation, but to investigations into the genetic underpinnings of disease, such as the development of resistance in MRSA described in the opening of this chapter.

View in Achieve
Animation 14.3
Deciphering the Genetic Code

The complete genetic code is shown in Figure 14.11. Notice that there are many more codons than there are different amino acids in proteins. Proteins are built from just 20 amino acids, but all possible combinations of the four available "letters" (the bases) yield 64 (4^3) different three-letter codons. Why are there more codons than amino acids? One reason is that there is more than one codon for almost all amino acids. For example, leucine is represented by six different codons (see Figure 14.11). Only methionine and tryptophan are represented by just one codon each. Thus we say that the genetic code is redundant (or degenerate). Also, a few select codons serve functions other than coding for amino acids. AUG, for example, not only codes for methionine but is also the **start codon**, the initiation signal for translation. Sixty one codons encode amino acid information and are termed **sense codons**. Three codons (UAA, UAG, UGA) do not encode amino acid information and are termed **nonsense** or **stop codons**, as they act as termination signals for translation. When the translation machinery reaches one of these codons, translation stops and the polypeptide is released from the translation complex.

Don't confuse a *redundant* code with an *ambiguous* code. If the code were ambiguous, a single codon could specify two (or more) different amino acids, and there would be doubt about which amino acid should be incorporated into a growing polypeptide chain. Redundancy in the code simply means there is more than one clear way to say "Put leucine here." The genetic code is not ambiguous: a given amino acid may be encoded by more than one codon, but a codon can code for only one amino acid.

The genetic code is (nearly) universal

One of the most amazing discoveries following the deciphering of the genetic code in bacteria was that the genetic code is essentially identical across all the species on our planet. Thus the code must be an ancient one that has been maintained intact throughout the evolution of living organisms. Exceptions are known: for example, the code for mitochondrial DNA (see Key Concept 12.5) and for chloroplast DNA differs slightly from that used by prokaryotes and for the nuclear DNA of eukaryotic cells; and in one group of protists, UAA and UAG code for glutamine rather than for stop codons. The significance of these differences is not yet clear. What is clear is that the exceptions are exceedingly few, and involve only a few codons when present.

Second letter

First letter	U	C	A	G	Third letter
U	UUU Phenylalanine UUC Phenylalanine UUA Leucine UUG Leucine	UCU Serine UCC Serine UCA Serine UCG Serine	UAU Tyrosine UAC Tyrosine **UAA** Stop codon **UAG** Stop codon	UGU Cysteine UGC Cysteine **UGA** Stop codon UGG Tryptophan	U C A G
C	CUU Leucine CUC Leucine CUA Leucine CUG Leucine	CCU Proline CCC Proline CCA Proline CCG Proline	CAU Histidine CAC Histidine CAA Glutamine CAG Glutamine	CGU Arginine CGC Arginine CGA Arginine CGG Arginine	U C A G
A	AUU Isoleucine AUC Isoleucine AUA Isoleucine **AUG** Methionine; start codon	ACU Threonine ACC Threonine ACA Threonine ACG Threonine	AAU Asparagine AAC Asparagine AAA Lysine AAG Lysine	AGU Serine AGC Serine AGA Arginine AGG Arginine	U C A G
G	GUU Valine GUC Valine GUA Valine GUG Valine	GCU Alanine GCC Alanine GCA Alanine GCG Alanine	GAU Aspartic acid GAC Aspartic acid GAA Glutamic acid GAG Glutamic acid	GGU Glycine GGC Glycine GGA Glycine GGG Glycine	U C A G

Figure 14.11 The Genetic Code Genetic information is encoded in mRNA in three-letter units—codons—made up of nucleoside monophosphates with the bases uracil (U), cytosine (C), adenine (A), and guanine (G) and is read in a 5′-to-3′ direction on mRNA. To decode a codon, find its first letter in the left column, then read across the top to its second letter, then read down the right column to its third letter. The amino acid the codon specifies is given in the corresponding row. For example, AUG codes for methionine, and GUA codes for valine.

View in Achieve
Activity 14.2 The Genetic Code

InvestigatingLIFE Deciphering the Genetic Code

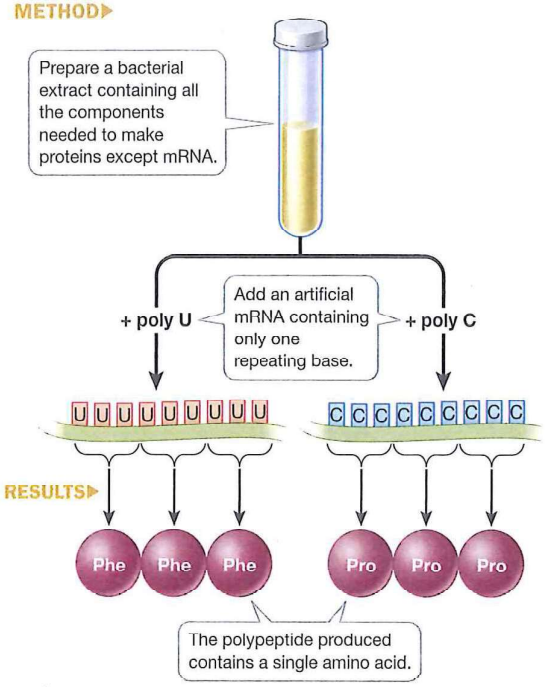
Experiment

Original Paper: M. Nirenberg and J. H. Matthaei. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci USA* 47: 1588–1602.

Nirenberg and Matthaei used a test tube protein synthesis system to determine the amino acids specified by synthetic mRNAs of known compositions.

HYPOTHESIS▶ An artificial mRNA containing only one repeating base will direct the synthesis of a protein containing only one repeating amino acid.

METHOD▶



CONCLUSION▶ Poly U contains codons for phenylalanine. Poly C contains codons for proline.

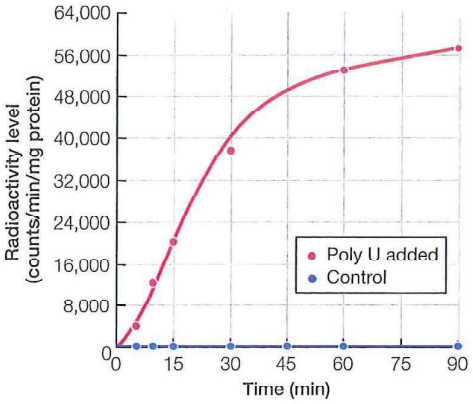
QUESTIONS▶

1. Poly U, an artificial mRNA, was added to a test tube with radiolabeled phenylalanine and all the other components for protein synthesis ("complete system"). Other test tubes differed from the complete system as indicated in the table. Samples were tested for incorporation of radioactivity into the resulting protein, with the results in Table A. Explain the results for each of the conditions.

Table A

Condition	Counts/minute (units of radioactivity)
Complete system	29,500
Minus poly U mRNA	70
Minus ribosomes	52
Minus ATP	83
Plus RNase (hydrolyzes RNA)	120
Plus DNase (hydrolyzes DNA)	27,600
Complete system minus all amino acids except radioactive phenylalanine	31,700

2. Poly U (red dots) was added to the test tubes and then samples were tested for protein synthesis by radioactive amino acid incorporation after various times (and were compared with results of a control run, in which no RNA was added, indicated by blue dots). The results are shown in the graph. What do these data show about the dependence of protein synthesis on added RNA?



Work with the Data

After the relationship between DNA and proteins was established, genetic evidence pointed to triplets of nucleotides on RNA specifying each amino acid. The race was on to identify which triplet coded for which amino acid. Test tube systems were developed in which cell extracts were made and protein synthesis occurred. Marshall Nirenberg, a scientist at the U.S. National Institutes of Health, and J. Heinrich Matthaei, a postdoctoral fellow from Germany, made a synthetic RNA consisting of the base uracil only (called poly U, codon UUU) and tested it in 20 tubes. Each test tube was supplied with all 20 amino acids, but in each one a different amino acid was tagged with a radioactive marker. In each tube the same polypeptide was made: a protein consisting of the amino acid phenylalanine bonded repeatedly to itself. However, only in the test tube with radiolabeled phenylalanine was the resulting protein radioactive.

3. The experiment described in Question 2 was repeated with different amino acids; the results are in Table B. Explain these results in terms of the codon specificity of poly U.

Table B

Radioactive amino acid(s)	Counts/minute/mg protein (radioactivity level)
Phenylalanine	38,300
Glycine, alanine, serine, aspartic acid, glutamic acid	33
Leucine, isoleucine, threonine, methionine, arginine, histidine, valine, lysine, tyrosine, proline, tryptophan	899
Cysteine	113

Go to **Achieve** for a companion **Data in Depth** exercise.

Connect the Concepts As discussed in Key Concept 12.5, some cytoplasmic organelles, notably the mitochondria and chloroplasts, contain small numbers of genes that are remnants of the genomes of the prokaryotes that eventually gave rise to these organelles. The evolutionary process of endosymbiosis that is responsible for the assimilation of these organelles into eukaryotic cells is discussed in Key Concept 26.1.

A common genetic code suggests that it originated early in the evolution of life and is shared by descent from these early species. The common code has profound implications for genetic engineering, as we will see in Chapter 18, since it means that the code for a human gene is the same as that for a bacterial gene. It is therefore impressive, but not surprising, that a human gene can be quite easily expressed in *E. coli* via laboratory manipulations, since these cells speak the same "molecular language."

The codons illustrated in the Nirenberg and Matthaei experiment in Investigating Life: Deciphering the Genetic Code are mRNA codons. The base sequence of the template DNA strand that is transcribed to produce the mRNA is complementary and antiparallel to these codons. Thus, for example, 3′-AAA-5′ in the template DNA strand corresponds to phenylalanine (which is encoded by the mRNA codon 5′-UUU-3′). In contrast, the coding strand of DNA has the same sequence as the mRNA (but with T's instead of U's). By convention, DNA sequences are usually shown beginning with the 5′ end of the coding sequence.

You might be thinking that in a long DNA molecule with many protein-coding regions, one strand is the coding strand for all the genes and the other strand is the template strand. In fact, some genes are transcribed from one strand and some genes are transcribed from the other. There is strand switching along a long DNA molecule in terms of the roles of the two strands. So while it is incorrect to say that one strand is "the" coding strand for all genes in a particular DNA molecule, it is correct to say that for a given gene one strand is coding and the other is template.

KEY CONCEPT
14.5 Recap and Assess

The genetic code relates the information in mRNA (as a linear sequence of codons) to protein (a linear sequence of amino acids).

1. Explain why the genetic code has triplets (e.g., AUA) of nucleotides, rather than singlets (e.g., A) or doublets (e.g., AU).
2. Explain what it means to say the genetic code is redundant.
3. A short mRNA is translated into the following peptide: methionine-tryptophan-tryptophan-glycine-tryptophan. What are the possible sequences of mRNAs (from start to stop) that could encode this peptide?

Now that we understand the rules by which mRNA encodes protein information, let's see how the cell is able to synthesize proteins from mRNA molecules using the genetic code.

KEY CONCEPT
14.6 The Coding Sequence in mRNA Is Translated into Proteins by Ribosomes

Learning Objectives

- **14.6.1** Describe initiation, elongation, and termination of translation.
- **14.6.2** Explain how tRNAs are able to translate mRNA codons into the corresponding amino acids.
- **14.6.3** Predict the consequences of an aminoacyl-tRNA synthetase attaching the wrong amino acid to a tRNA.
- **14.6.4** Describe each step experienced by a tRNA, starting with an uncharged tRNA in the cytosol and ending with the release of the uncharged tRNA from a ribosome.
- **14.6.5** Explain the biological significance of polysomes.

Transcription uses complementary base pairing to synthesize an mRNA molecule from a DNA template strand. The information in both molecules is equivalent—a sequence of nucleotides. In contrast, translation uses mRNA to synthesize a polypeptide; nucleotide information encodes amino acid information, as we showed with the genetic code. How is this translation accomplished? How is the genetic code "read" by the translation machinery? The answer is a special kind of RNA molecule called transfer RNA (tRNA) that can (1) read mRNA codons and (2) bring the corresponding amino acid to the ribosome.

Once the tRNAs "decode" the mRNA and deliver the appropriate amino acids, components of the ribosome catalyze the formation of peptide bonds between amino acids. Let's now look at how the tRNAs read codons and get the appropriate amino acids into the ribosome.

View in Achieve
Animation 14.4 Translation

A transfer RNA carries a specific amino acid and binds to a specific mRNA codon

There is at least one specific tRNA molecule for each of the 20 amino acids. Each tRNA has three functions that are fulfilled by its structure and base sequence (Figure 14.12):

1. *tRNAs bind to particular amino acids.* Each tRNA binds to a specific enzyme that attaches it to only 1 of the 20 amino acids. The covalent attachment to an amino acid takes place at the 3′ end of the tRNA. When it is carrying an amino acid, the tRNA is said to be "charged."
2. *tRNAs bind to mRNA.* At about the midpoint on the tRNA polynucleotide chain there is a triplet of bases called the **anticodon**, which is complementary to the mRNA codon for the particular amino acid that the tRNA carries. For example, the mRNA codon for arginine is 5′-CGG-3′, and the complementary tRNA anticodon is 3′-GCC-5′. Like the

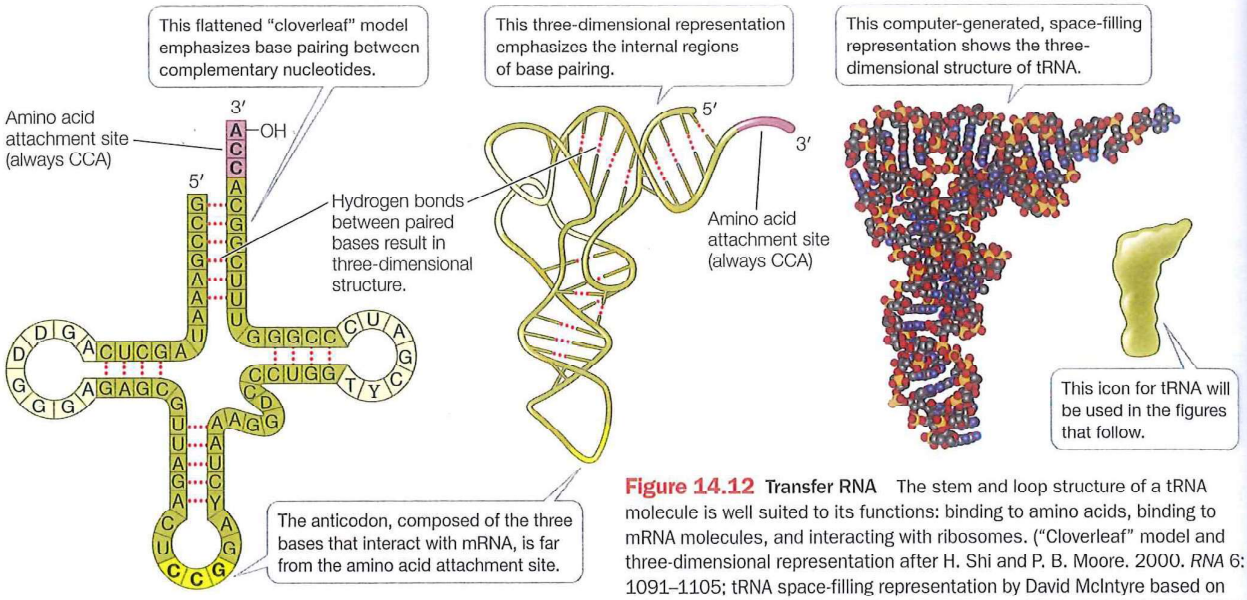


Figure 14.12 Transfer RNA The stem and loop structure of a tRNA molecule is well suited to its functions: binding to amino acids, binding to mRNA molecules, and interacting with ribosomes. ("Cloverleaf" model and three-dimensional representation after H. Shi and P. B. Moore. 2000. *RNA* 6: 1091–1105; tRNA space-filling representation by David McIntyre based on data from PDB ID: 1EHZ. H. Shi and P. B. Moore. 2000. *RNA* 6: 1091–1105.)

two strands of DNA, the codon and anticodon bind together via noncovalent hydrogen bonds.

3. *tRNAs interact with ribosomes.* The ribosome has several sites on its surface that just fit the three-dimensional structure of a tRNA molecule. Interaction between the ribosome and the tRNA is noncovalent.

Recall that 61 different codons encode the 20 amino acids in proteins (see Figure 14.11). Does this mean that the cell must produce 61 different tRNA species, each with a different anticodon? No. The cell gets by with about two-thirds of that number of tRNA species because the interaction between the bases at the third position of the codon–anticodon interaction, which is the 3′ end of the codon and the 5′ end of the anticodon, is not as specific as it is for other base pairs. This phenomenon is called wobble base pairing—base pairing that does not strictly follow the standard rules (i.e., A with U, and G with C). Wobble base pairing occurs only at the third position of the codon–anticodon interaction. Sometimes an unusual purine base, inosine (I), which can pair with A, C, and U, is found at the third position. For example, the presence of inosine in the tRNA with the anticodon 3′-UAI-5′ allows it to recognize and bind to the three isoleucine codons: AUU, AUC, and AUA. Wobble base pairing also occurs for some normal bases in the third position; for example, G in the third position of the anticodon can pair with C or U in the codon. Wobble base pairing explains many of the patterns observed in the genetic code (see Figure 14.11); you may have noticed that all codon pairs that start with the same two bases but end in C or U always code for the same amino acid—such codons can be read by tRNAs with a G in their third position. It is important to recognize that wobble base pairing allows fewer tRNAs to read the code than would be required with strict base pairing, but it does not allow the genetic code to be ambiguous. That is, *each mRNA codon can bind only to tRNAs that carry a specific amino acid.*

Each tRNA is specifically attached to an amino acid

The charging of each tRNA with its correct amino acid is achieved by a family of enzymes known as aminoacyl-tRNA synthetases. Each enzyme is specific for one amino acid and for its corresponding tRNA. The reaction uses ATP, forming a high-energy bond between the amino acid and the tRNAs (Figure 14.13). The energy in this bond is later used in the formation of peptide bonds between amino acids in a growing polypeptide chain.

The specificity between the tRNA and its corresponding amino acid is essential. For example, tRNAcys (with anticodon ACA), which is the tRNA that carries cysteine, always has cysteine added to it by its synthetase to form cys-tRNAcys. A clever experiment by Seymour Benzer and his colleagues at Purdue University demonstrated the importance of this specificity. They took the cys-tRNAcys molecule and chemically modified the cysteine, converting it into alanine (to form ala-tRNAcys). Which component—the amino acid or the tRNA—would be recognized when this hybrid charged tRNA was put into a protein synthesizing system? The answer was the tRNA. Everywhere in the synthesized protein where cysteine was supposed to be, alanine appeared instead. The cysteine-specific tRNA had delivered its cargo (alanine) to every mRNA codon for cysteine. This experiment showed that the protein synthesis machinery recognizes the anticodon of the charged tRNA, not the amino acid attached to it.

The ribosome is the workbench for translation

The ribosome is the molecular workbench where the task of translation is accomplished. Its structure enables it to hold mRNA

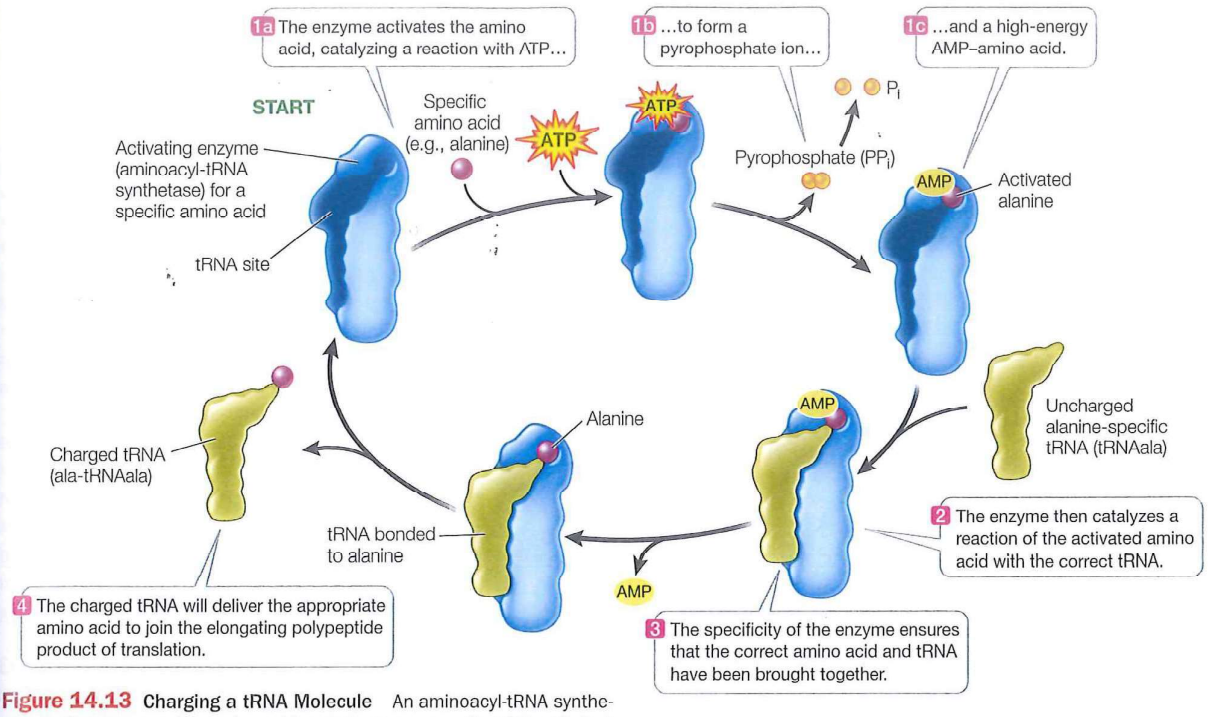


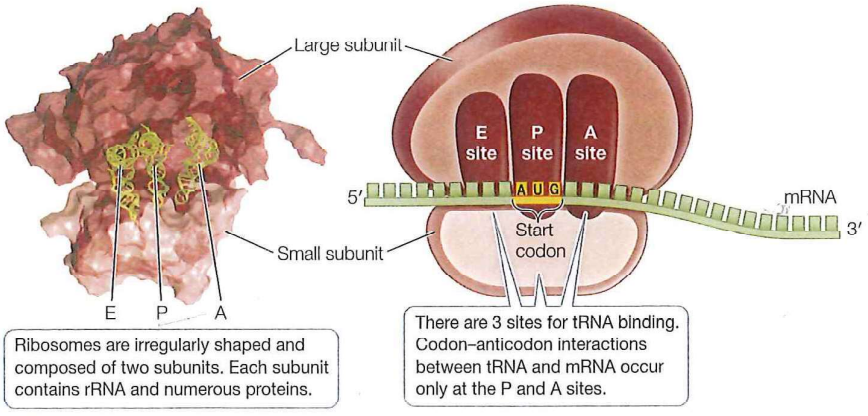
Figure 14.13 Charging a tRNA Molecule An aminoacyl-tRNA synthetase activates a specific amino acid and charges a specific tRNA with that amino acid.

and charged tRNAs in the correct positions, thus allowing a polypeptide chain to be assembled efficiently. A given ribosome does not specifically produce just one kind of protein. A ribosome can use any mRNA and all species of charged tRNAs, and thus can be used to make many different polypeptide products. Ribosomes can be used over and over again, and there are thousands of them in a typical cell.

Although ribosomes are small relative to other cellular structures, their mass of several million daltons makes them large in comparison with charged tRNAs. Each ribosome consists of two subunits, a large one and a small one (Figure 14.14). The two subunits and several dozen other molecules interact noncovalently. In fact, when hydrophobic interactions between the proteins and RNAs are disrupted, the ribosome falls apart. The two subunits separate and all the RNAs and proteins separate from one another. If the disrupting agent is removed, the complex structure self-assembles perfectly! This is amazing: think of throwing the pieces of a jigsaw puzzle up in the air and having them fit together when they land. The ribosome reflects the high specificity of molecular machines in a cell composed of many molecules.

Figure 14.14 Ribosome Structure Each ribosome consists of a large and a small subunit. The subunits remain separate when they are not being used for protein synthesis. (Left by David McIntyre based on data from PDB 1GIX and 1G1Y. M. M. Yusupov et al. 2001. *Science* 292: 883–896.)

Q: The ribosome consists of several dozen proteins and several RNA molecules, held together noncovalently. What are the chemical forces involved? How can these forces be disrupted and the molecules separated from one another?



In eukaryotes, the large subunit consists of three different molecules of ribosomal RNA (rRNA) and about 45 different protein molecules, arranged in a precise configuration. The small subunit consists of 1 rRNA molecule and about 32 different protein molecules. The exact number of proteins present in each subunit varies a little across species. The ribosomes of prokaryotes are somewhat smaller than those of eukaryotes, and their ribosomal proteins and RNAs are different. Mitochondria and chloroplasts also contain ribosomes, which have features similar to those of prokaryotes (see Chapter 5).

There are three sites to which a tRNA can bind on the large subunit of the ribosome, designated A, P, and E (see Figure 14.14). The mRNA and ribosome move in relation to one another, and as they do so, an initially charged tRNA traverses these three sites in order:

1. The *A (aminoacyl-tRNA) site* is where the charged tRNA anticodon binds to the mRNA codon, thus lining up the correct amino acid to be added to the growing polypeptide chain.
2. The *P (peptidyl-tRNA) site* is where the tRNA carrying the growing peptide chain resides.
3. The *E (exit) site* is where the uncharged tRNA, having given up its amino acid, resides before being released from the ribosome into the cytosol to pick up another amino acid and begin the process again.

The ribosome has a fidelity function that ensures that the mRNA–tRNA interactions are accurate; that is, that a charged tRNA with the correct anticodon (e.g., 3′-UAC-5′) forms hydrogen bonds with the appropriate codon in the mRNA (e.g., 5′-AUG-3′). The rRNA of the small ribosomal subunit plays a role in validating the three-base-pair match. If hydrogen bonds have not formed between all three base pairs, the tRNA must be the wrong one for that mRNA codon, and the incorrect tRNA is ejected from the A site of the ribosome.

Translation takes place in three steps

Translation is the process by which the information in mRNA (derived from DNA) is used to specify and link a specific sequence of amino acids, producing a polypeptide. Like transcription, translation occurs in three steps: initiation, elongation, and termination.

INITIATION The translation of mRNA begins with the formation of an **initiation complex**, which consists of a charged tRNA and a small ribosomal subunit, both bound to the mRNA (Figure 14.15).

In prokaryotes, the rRNA of the small ribosomal subunit first binds to a complementary ribosome binding site (AGGAGG; known as the Shine–Dalgarno sequence) on the mRNA. This sequence is less than ten bases upstream of the actual start codon but lines up the start codon so that it is adjacent to the P site of the large subunit:

mRNA 5′.....A G G A G G.....(start codon).....3′

rRNA 3′..... U C C U C C.....(P site).....5′

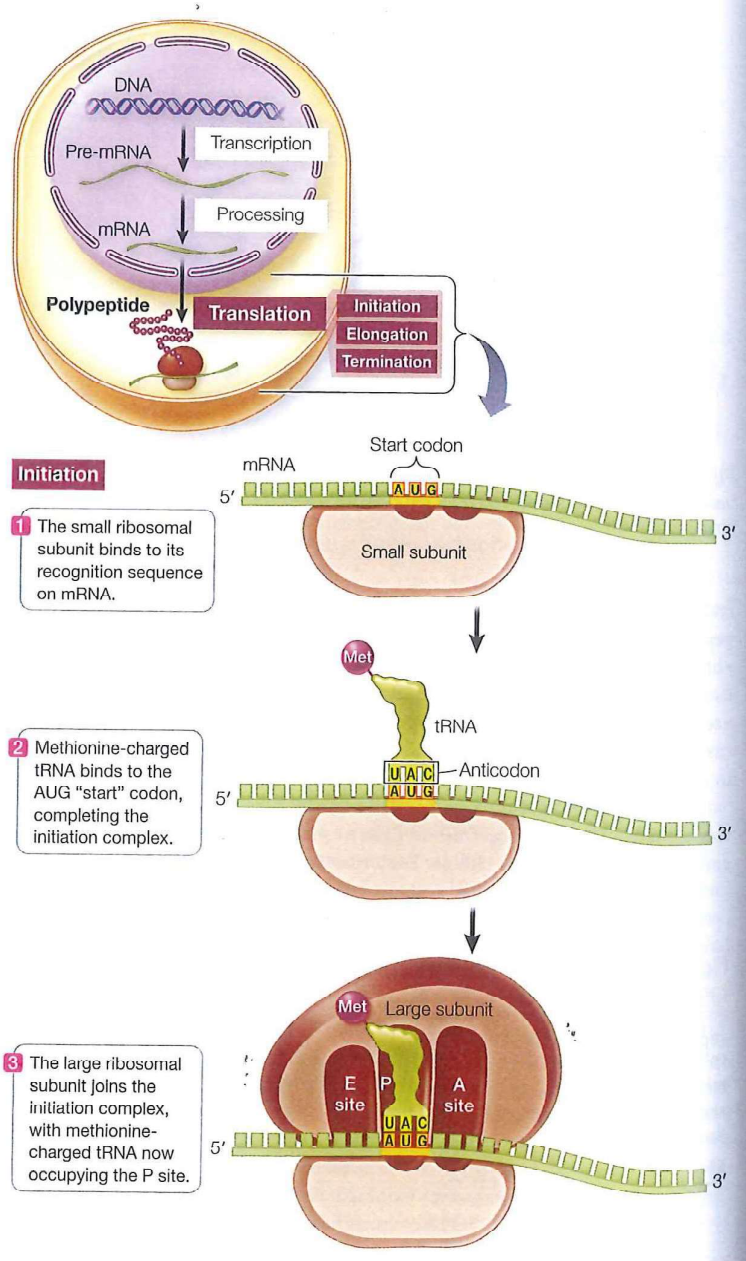


Figure 14.15 The Initiation of Translation Translation begins with the formation of an initiation complex. In prokaryotes, the small ribosomal subunit binds to the Shine–Dalgarno sequence to begin the process, whereas in eukaryotes it binds to the 5′ cap.

Eukaryotes load the mRNA onto the ribosome somewhat differently: the small ribosomal subunit binds to the 5′ cap on the mRNA and then moves along the mRNA until it reaches the start codon.

Recall that the mRNA start codon in the genetic code is AUG (see Figure 14.11). The anticodon (UAC) of a methionine-charged tRNA binds to this start codon by complementary base pairing to complete the initiation complex. Thus the first amino acid in a polypeptide chain is always methionine. However, not all mature proteins have methionine as their N-terminal amino acid. In many cases, the initial methionine is removed by an enzyme after translation.

After the methionine-charged tRNA has bound to the mRNA, the large subunit of the ribosome joins the complex. The methionine-charged tRNA now lies in the P site of the ribosome, and the A site is aligned with the second mRNA codon. These ingredients—mRNA, two ribosomal subunits, and methionine-charged tRNA—are assembled by a group of proteins called initiation factors.

ELONGATION A charged tRNA now enters the open A site of the large ribosomal subunit. If it does not have the correct anticodon, corresponding to the codon in the A site, it exits. At some point, a charged tRNA whose anticodon is complementary to the A site codon of the mRNA enters the A site (Figure 14.16). The large subunit then catalyzes two reactions:

1. It breaks the bond between the tRNA and its amino acid in the P site.
2. It catalyzes the formation of a peptide bond between the amino acid that has just been released from the P site tRNA and the one attached to the tRNA in the A site.

Because the large ribosomal subunit performs these two actions, it is said to have **peptidyl transferase** activity. In this way, methionine (the amino acid in the P site) becomes the N terminus of the new protein. The second amino acid is now bound to methionine but remains attached to its tRNA at the A site.

How does the large ribosomal subunit catalyze peptide bond formation? Harry Noller and his colleagues at the University of California at Santa Cruz did a series of experiments and found that:

- if they removed almost all of the proteins from the large subunit, it still catalyzed peptide bond formation.

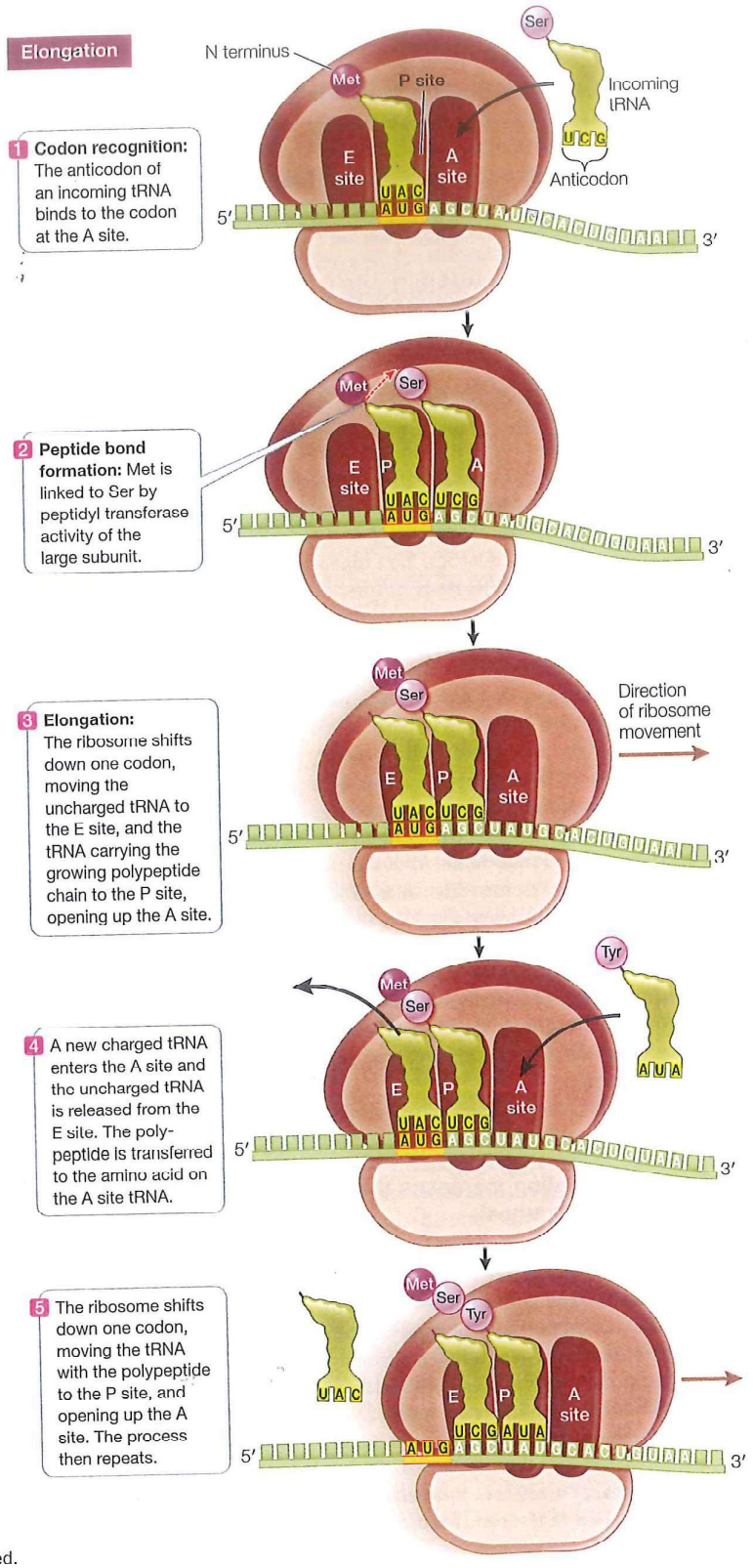


Figure 14.16 The Elongation of Translation
The polypeptide chain elongates as the mRNA is translated.

- if the rRNA was extensively modified, peptidyl transferase activity was destroyed.

The experiment showed that *rRNA is the catalyst*. The purification and crystallization of ribosomes have allowed scientists to examine ribosome structure in detail, and the catalytic role of rRNA in peptidyl transferase activity has been confirmed. These findings support the hypothesis that RNA, and catalytic RNA in particular, evolved before DNA (see Key Concept 4.3).

Connect the Concepts As discussed in Key Concept 4.3, the folded, three-dimensional surface of an RNA molecule can be just as specific as that of a protein, and may thus take on a catalytic function. The "RNA world" hypothesis, which proposed that RNA serves as a catalyst for its own replication, was boosted by the discovery of ribozymes, catalytic RNAs that can speed up biological reactions, including those that involve their own nucleotides.

After the first tRNA releases its methionine, it moves to the E site and is then dissociated from the ribosome, returning to the cytosol to become charged with another methionine. The second tRNA, now bearing a dipeptide (a two-amino acid chain), is shifted to the P site as the ribosome moves one codon along the mRNA in the 5′-to-3′ direction. The elongation process continues, and the polypeptide chain grows, as these steps are repeated. Follow the process in Figure 14.16. All of these steps are assisted by ribosomal proteins called elongation factors.

TERMINATION The elongation cycle ends, and translation is terminated, when a stop codon—UAA, UAG, or UGA—enters the A site (Figure 14.17). Stop codons do not correspond with any amino acids and thus do not bind any tRNAs. Rather, they bind a protein release factor, which allows hydrolysis of the bond between the polypeptide chain and the tRNA in the P site. The newly completed polypeptide thereupon separates from the ribosome. Its C terminus is the last amino acid to join the chain. Its N terminus, at least initially, is methionine, as a consequence of the AUG start codon. In its amino acid sequence, it contains information specifying its conformation, as well as its ultimate cellular destination.

Table 14.3 lists the nucleic acid signals for initiation and termination of transcription and translation.

Polysome formation increases the rate of protein synthesis

Several ribosomes can work simultaneously at translating a single mRNA molecule, producing multiple polypeptides at the same time. As soon as the first ribosome has moved far enough from the 5′ end of the mRNA, a second initiation complex can form, then a third, and so on. An assemblage consisting of a strand of mRNA with its beadlike ribosomes and their growing polypeptide chains is called a **polyribosome**, or **polysome** (Figure 14.18). Cells that are actively synthesizing proteins contain large numbers of polysomes and few free ribosomes or ribosomal subunits.

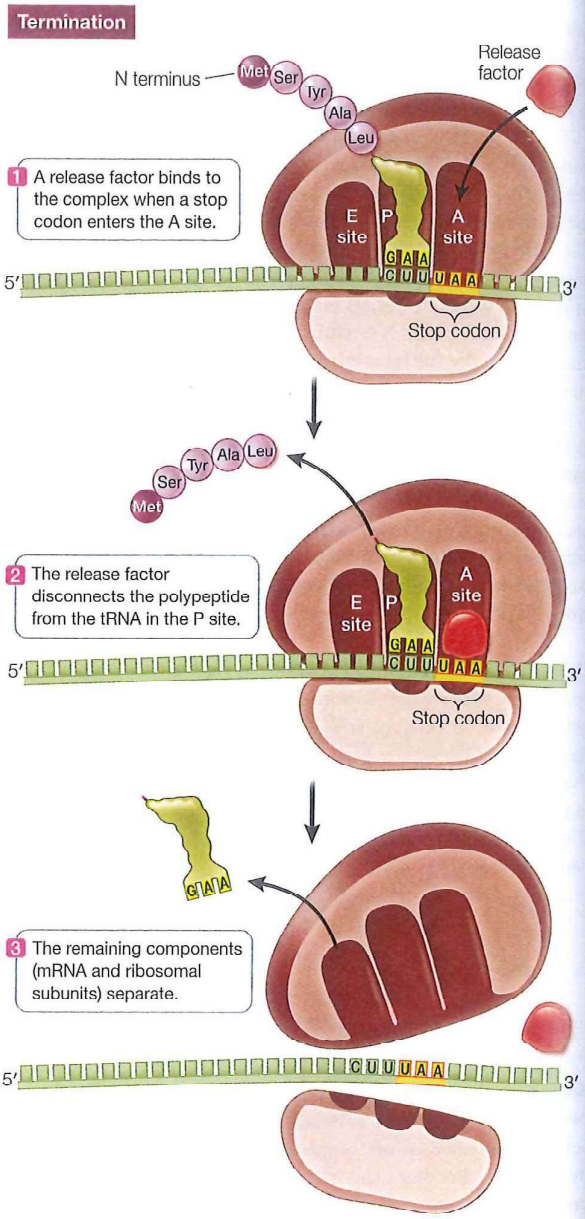


Figure 14.17 The Termination of Translation Translation terminates when the A site of the ribosome encounters a stop codon on the mRNA.

Q: What happens if there is not a stop codon in the mRNA?

TABLE 14.3 | Signals that Start and Stop Transcription and Translation

	Transcription	Translation
Initiation	Promoter sequence in DNA	AUG start codon in the mRNA
Termination	Terminator sequence in DNA	UAA, UAG, or UGA in the mRNA

KEY CONCEPT 16.1 Prokaryotic Gene Expression Is Regulated in Operons

Learning Objectives

- **16.1.1** Describe the conditions that cause high expression of genes of the *lac* and *trp* operons.
- **16.1.2** Describe the roles of structural or regulatory DNA sequences in the *lac* and *trp* operons.
- **16.1.3** Determine how mutations in structural or regulatory genes of the *lac* or *trp* operons would alter the phenotype of a cell.
- **16.1.4** Explain how sigma factors and promoter sequences are involved in regulating classes of genes in prokaryotes.

Prokaryotes conserve energy and resources by making certain proteins only when they are needed. The protein content of a bacterium can change rapidly when conditions warrant. Based on what you learned about gene expression in Chapter 14, you might suggest several ways in which a prokaryotic cell could shut off the supply of an unneeded protein. The cell could:

- decrease the rate of transcription of mRNA for that protein;
- hydrolyze the mRNA after it is made, preventing translation;
- prevent translation of the mRNA at the ribosome;
- hydrolyze the protein after it is made; or
- temporarily inhibit the function of the protein in the cell.

Whichever mechanism is used, it must be both responsive to environmental signals and efficient. The earlier the cell intervenes in the process of protein synthesis, the less energy it wastes making an unneeded protein. Selective blocking of transcription is far more efficient than transcribing the gene, translating the message, and then degrading or inhibiting the protein. While all five mechanisms for regulating protein levels are found in nature, prokaryotes generally use the most efficient one: transcriptional regulation.

Gene expression begins at the promoter (see Key Concept 14.3), where RNA polymerase binds to initiate transcription. Note, however, that in a given cell at a given point in time, not all promoters are active. This observation suggests that gene transcription must be selective. The "decision" regarding which genes to activate involves two types of regulatory proteins that bind to DNA: repressor proteins and activator proteins. In both cases, these proteins bind to DNA to regulate the gene (Figure 16.1):

- In **negative regulation**, binding of a repressor protein prevents transcription.
- In **positive regulation**, an activator protein binds DNA to stimulate transcription.

Connect the Concepts A key DNA sequence involved in regulation is the promoter, where transcription begins and various regulatory proteins bind. Learn details about proteins at the promoter in Key Concept 14.3.

You will see examples of these mechanisms, or combinations of them, as we examine regulation in prokaryotes, eukaryotes, and viruses. We'll focus first on a regulatory system for the use of the sugar lactose.

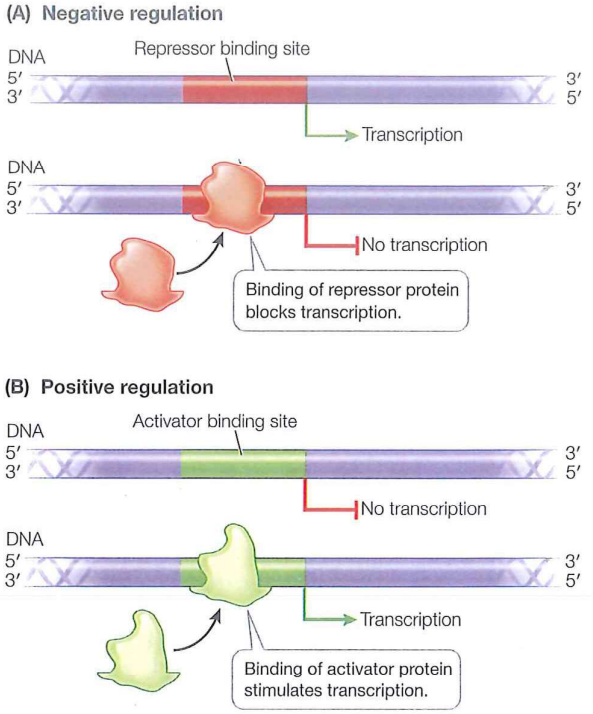


Figure 16.1 Positive and Negative Regulation Proteins regulate gene expression by binding to DNA and preventing or allowing RNA polymerase to bind DNA at the promoter region to control transcription of the gene.

Q: Could a gene be under both positive and negative regulation?

Regulating gene transcription conserves energy

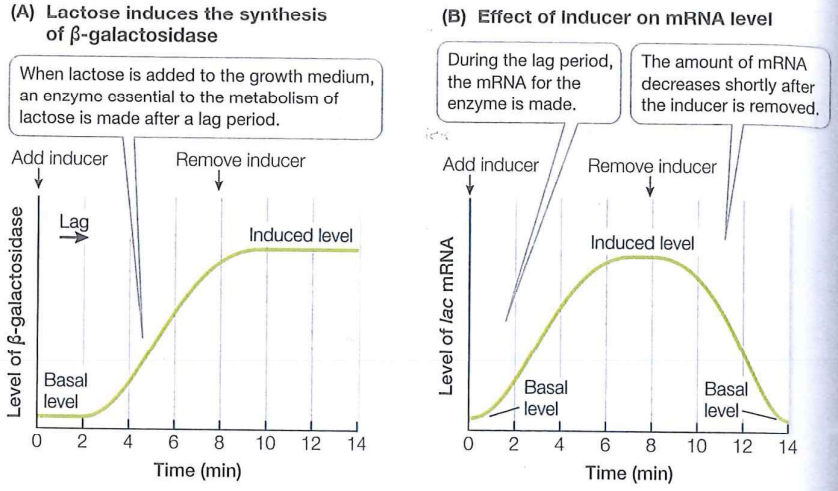
As a normal inhabitant of the human intestine, *E. coli* must be able to adjust to sudden changes in its chemical environment. Its host may present it with one foodstuff one hour (e.g., glucose in fruit) and another the next (e.g., lactose in milk). Such changes in nutrients present the bacterium with a metabolic challenge. Glucose is the easiest sugar for the bacterium to metabolize, and is its preferred energy source. Lactose is a β-galactoside—a disaccharide containing galactose β-linked to glucose (see Key Concept 3.3). Three proteins are upregulated by *E. coli* in response to an increase in lactose:

1. β-Galactoside permease is a carrier protein in the bacterial cell membrane that transports the sugar into the cell.
2. β-Galactosidase is an enzyme that hydrolyses lactose to glucose and galactose.
3. β-Galactoside transacetylase transfers acetyl groups from acetyl CoA to certain β-galactosides. Its role in the metabolism of lactose is not clear.

When *E. coli* grows and reproduces in a lab medium that contains glucose but no lactose or other β-galactosides, the levels of these

Figure 16.2 Lactose Induces the Expression of a Gene for an Enzyme (A) When the inducer lactose is added to the growth medium for the bacterium *E. coli*, the synthesis of β-galactosidase begins only after an initial lag period. (B) There is a lag period because the mRNA for β-galactosidase has to be made before the protein can be made. The amount of mRNA decreases rapidly after the lactose is removed, indicating that transcription is no longer occurring. These changes in mRNA levels indicate that the mechanism of induction by lactose is transcriptional regulation.

Q: Why doesn't the level of β-galactosidase drop following removal of the inducer, like the level of mRNA does?



three proteins are extremely low—the cell does not waste energy and materials making the unneeded enzymes. But if the environment changes such that lactose is the predominant sugar available and very little glucose is present, the bacterium promptly begins making all three enzymes after a short lag period. While few molecules of β-galactosidase (and the other two enzymes) are present in an *E. coli* cell in the presence of glucose, in the absence of glucose the addition of lactose can induce the synthesis of about 1,500 times more molecules of β-galactosidase per cell (Figure 16.2A)!

What's behind this dramatic increase? An important clue comes from measuring the amount of mRNA for β-galactosidase. The mRNA level increases during the lag period after lactose is added to the medium, and this mRNA is translated into protein (Figure 16.2B). Moreover, the high mRNA level depends on the presence of lactose, because if the lactose is removed, the mRNA level goes down. *The response of the bacterial cell to lactose is clearly at the level of transcription.*

Compounds such as lactose that stimulate the synthesis of a protein are called **inducers**. The proteins that are produced are called **inducible proteins**, whereas proteins that are made all the time at a constant rate are called **constitutive proteins**.

We have now seen two basic ways of regulating the rate of a metabolic pathway. In Key Concept 8.5 we described the allosteric regulation of enzyme activity, which allows the rapid fine-tuning of metabolism. Regulation of protein synthesis—that is, regulation of the concentration of enzymes—is slower but results in greater savings of energy and resources. Protein synthesis is a highly endergonic process, since assembling mRNA, charging tRNA, and moving the ribosomes along mRNA all require the hydrolysis of nucleoside triphosphates such as ATP. Figure 16.3 compares these two modes of regulation.

Operons are units of transcriptional regulation in prokaryotes

The genes in *E. coli* that encode the three enzymes for using lactose are structural genes; **structural genes** specify the primary structures (the amino acid sequences) of protein molecules that act as enzymes or cytoskeletal proteins. The three genes are adjacent to one another on the *E. coli* chromosome. This arrangement is no coincidence: the genes share a single promoter, and their DNA is transcribed into a single, continuous molecule of mRNA. Because this particular mRNA governs the synthesis of all three lactose-metabolizing enzymes, either all or none of these enzymes are made, depending on whether their common message—their mRNA—is present in the cell.

A cluster of genes with a single promoter is called an **operon**, and the operon that encodes the three lactose-metabolizing enzymes in *E. coli* is called the *lac* operon. The *lac* operon promoter

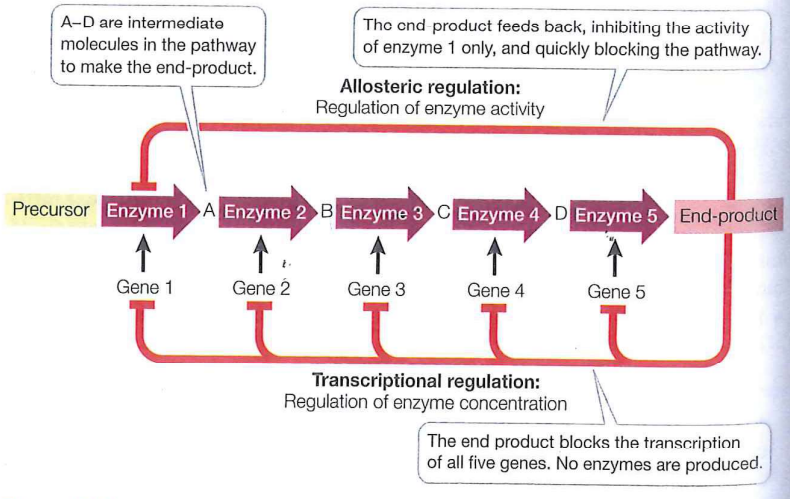


Figure 16.3 Two Ways to Regulate a Metabolic Pathway Feedback from the end-product of a metabolic pathway can block enzyme activity (allosteric regulation), or it can stop the transcription of genes that code for the enzymes in the pathway (transcriptional regulation).

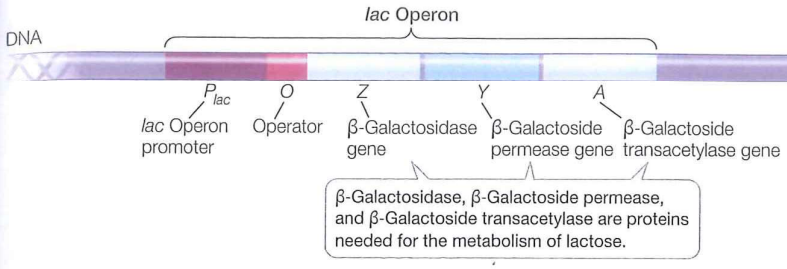


Figure 16.4 The *lac* Operon of E. coli
The *lac* operon of *E. coli* is a segment of DNA that includes a promoter, an operator, and the three structural genes that code for lactose-metabolizing enzymes.

can be very efficient (the maximum rate of mRNA synthesis can be high), but mRNA synthesis is very low when the enzymes are not needed. In addition to having a promoter, an operon has other **regulatory sequences** (DNA sequences to which the protein products of regulatory genes bind) that are not transcribed. A typical operon consists of a promoter, an operator, and two or more structural genes (Figure 16.4). The **operator** is a short stretch of DNA that lies between the promoter and the structural genes. It can bind very tightly with regulatory proteins that either activate or repress transcription.

There are numerous mechanisms to control the transcription of operons; we describe three examples:

1. An inducible operon regulated by a repressor protein
2. A repressible operon regulated by a repressor protein
3. An operon regulated by an activator protein

Operator–repressor interactions control transcription in the *lac* and *trp* operons

The *lac* operon contains a promoter, to which RNA polymerase binds to initiate transcription, and an operator, to which a **repressor** protein can bind. The gene that encodes this repressor is located near the *lac* operon on the *E. coli* chromosome. When the repressor is bound, transcription of the operon is blocked. This example of negative regulation was elegantly worked out by Nobel Prize winners François Jacob and Jacques Monod.

The repressor protein has two binding sites: one that binds to the operator DNA and the other that binds to the carbohydrate inducer. The environmental signal that induces the *lac* operon (for example, in the human digestive tract) is lactose, but the actual inducer is allolactose, a molecule that forms from lactose once it enters the cell. In the absence of the inducer, the repressor protein fits into the major groove of the operator DNA and recognizes and binds to a specific nucleotide base sequence. This prevents the binding of RNA polymerase to the promoter, and the operon is not transcribed (Figure 16.5A).

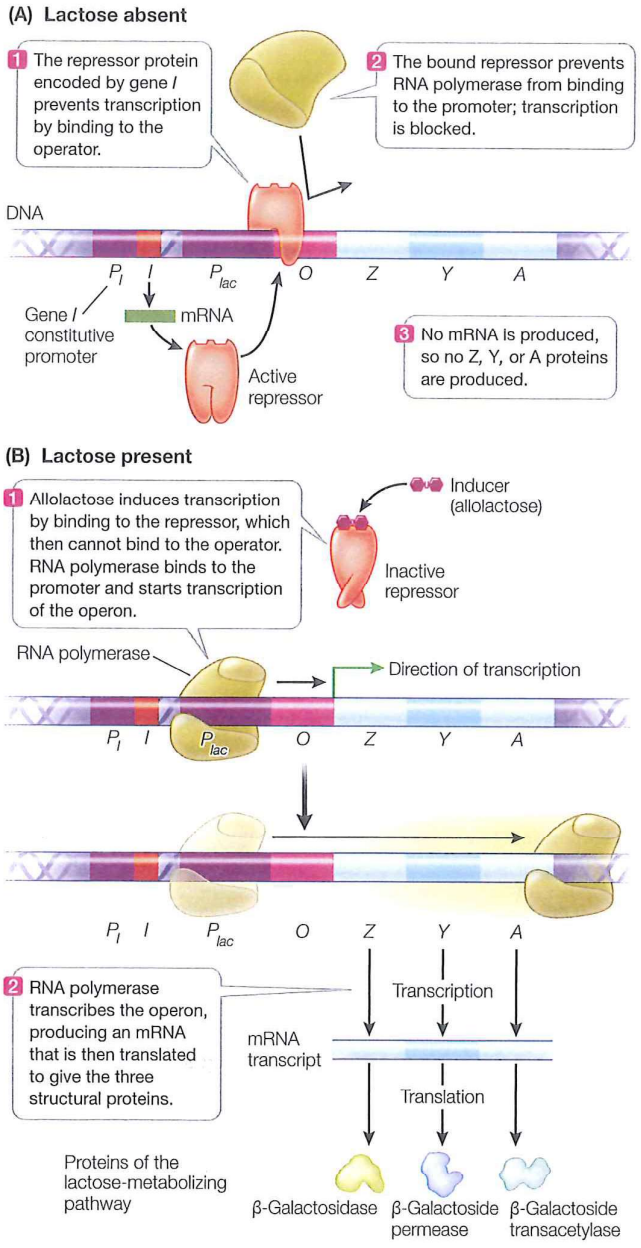


Figure 16.5 The *lac* Operon: An Inducible System (A) When lactose is absent, the synthesis of enzymes for its metabolism is inhibited. (B) Lactose leads to synthesis of the enzymes in the lactose-metabolizing pathway. The actual inducer is allolactose, which binds to the repressor protein and prevents its binding to the operator.

Q: If part B of this figure were not labeled "Lactose present," how could you determine that lactose is present?

View in Achieve
Animation 16.1 The *lac* Operon

When the inducer is present, it binds to the repressor and changes the shape of the repressor. This change in three-dimensional structure (conformation) prevents the repressor from binding to the operator. As a result, RNA polymerase can bind to the promoter and start transcribing the structural genes of the *lac* operon (Figure 16.5B).

You can see from this example that a key to transcriptional control of gene expression is the presence of regulatory sequences that do not code for proteins but are binding sites for regulatory proteins and other proteins involved in transcription.

In contrast to the inducible system of the *lac* operon, other operons in *E. coli* are repressible; that is, they are usually expressed but can be repressed under specific conditions. In such a system, the repressor is not normally bound to the operator. But if another molecule called a **co-repressor** binds to the repressor, the repressor changes shape and binds to the operator, thereby inhibiting transcription. An example is the *trp* operon, whose five structural genes catalyze the synthesis of the amino acid tryptophan. When tryptophan is absent, the repressor does not bind to the operator and the structural genes are transcribed and translated (Figure 16.6A). However, if tryptophan is present in the cell, it is advantageous to stop making the enzymes for tryptophan synthesis. To do this, the repressor binds to tryptophan (the co-repressor) and then binds to the operator in the *trp* operon, shutting off transcription (Figure 16.6B).

Let's pause to summarize the differences between these two types of operons:

- In *inducible* systems, the substrate of a metabolic pathway (the inducer) interacts with a regulatory protein (the repressor), rendering the repressor incapable of binding to the operator and thus allowing transcription.
- In *repressible* systems, the product of a metabolic pathway (the co-repressor) binds to a regulatory protein (the repressor), which is then able to bind to the operator and block transcription.

Usually, inducible systems control catabolic pathways (which are turned on only when the substrate is available), whereas repressible systems control anabolic pathways (which are turned on until the concentration of the product becomes excessive). In both systems, the regulatory protein is a repressor that functions by binding to the operator. Next we consider an example of positive control involving an activator.

Connect the Concepts As discussed in Key Concept 8.1, there are two kinds of metabolism: Catabolic pathways break down complex molecules into simpler ones, releasing energy formerly stored in the chemical bonds. Anabolic pathways link simple molecules to form more complex molecules, a process that requires an input of energy.

View in Achieve
Animation 16.2 The *trp* Operon

Protein synthesis can be controlled by increasing promoter efficiency

In negative control, transcription is *decreased* in the presence of a repressor protein. *E. coli* can also use positive control to *increase* transcription through the presence of an **activator** protein. For an example we return to the *lac* operon, where the relative levels of glucose and lactose determine the amount of transcription. We have seen that in the presence of lactose the *lac* repressor is unable to bind

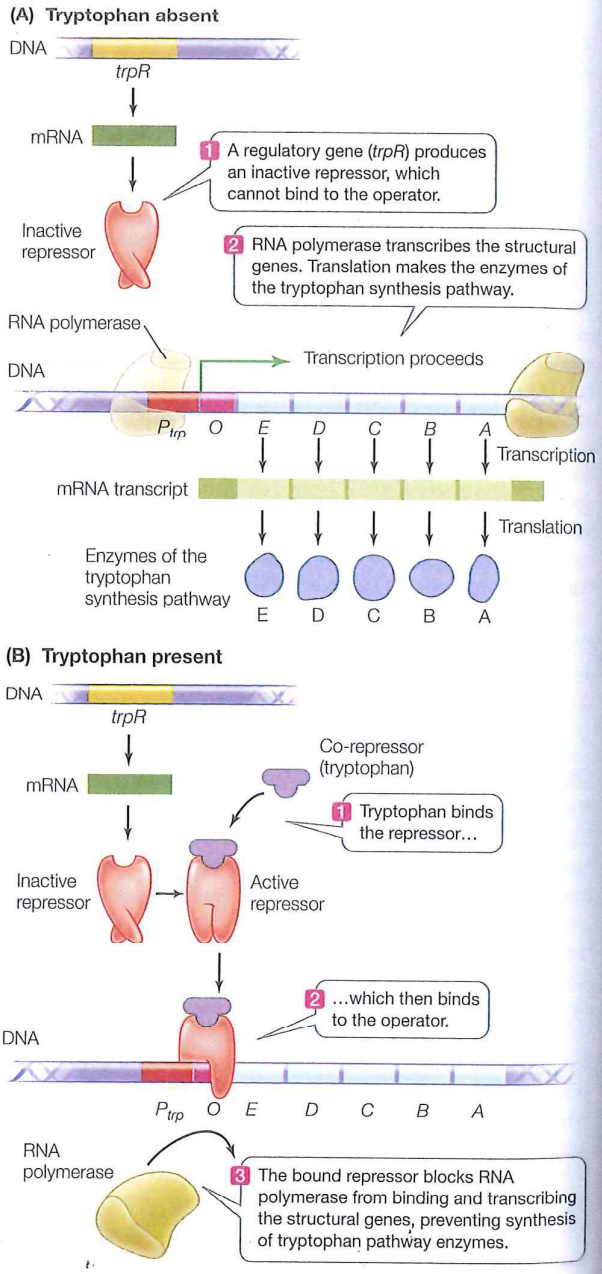


Figure 16.6 The *trp* Operon The tryptophan (*trp*) operon has a promoter, an operator, and five structural genes that encode the enzymes of the tryptophan synthesis pathway. A repressor protein, encoded by the *trpR* gene, produces an inactive repressor that becomes active when it binds to tryptophan and then is able to bind to the operator. The operon is thus transcribed when tryptophan is absent **(A)** but not when tryptophan is present **(B)**.

to the *lac* operator to repress transcription (see Figure 16.5B). But glucose is the preferred source of energy for the cell, so if glucose and lactose levels are both high, the *lac* operon is still not transcribed efficiently. This is because efficient transcription of the *lac* operon requires binding of an activator protein to its promoter.

Low levels of glucose in the cell set off a signaling pathway that leads to increased levels of the second messenger cyclic AMP (cAMP) (see Key Concept 7.3). Cyclic AMP binds to an activator protein called *c*AMP *r*eceptor *p*rotein (CRP), producing a conformational change in CRP that allows it to bind to the *lac* promoter. CRP is an activator of transcription, because its binding results in more efficient binding of RNA polymerase to the promoter, and thus increased transcription of the structural genes (Figure 16.7). In the presence of abundant glucose, cAMP levels are low, CRP does not bind to the promoter, and the efficiency of transcription of the *lac* operon is reduced. This is an example of **catabolite repression**, a system of gene regulation in which the presence of the preferred energy source represses other catabolic pathways. The mechanisms controlling positive and negative regulation of the *lac* operon are summarized in Table 16.1.

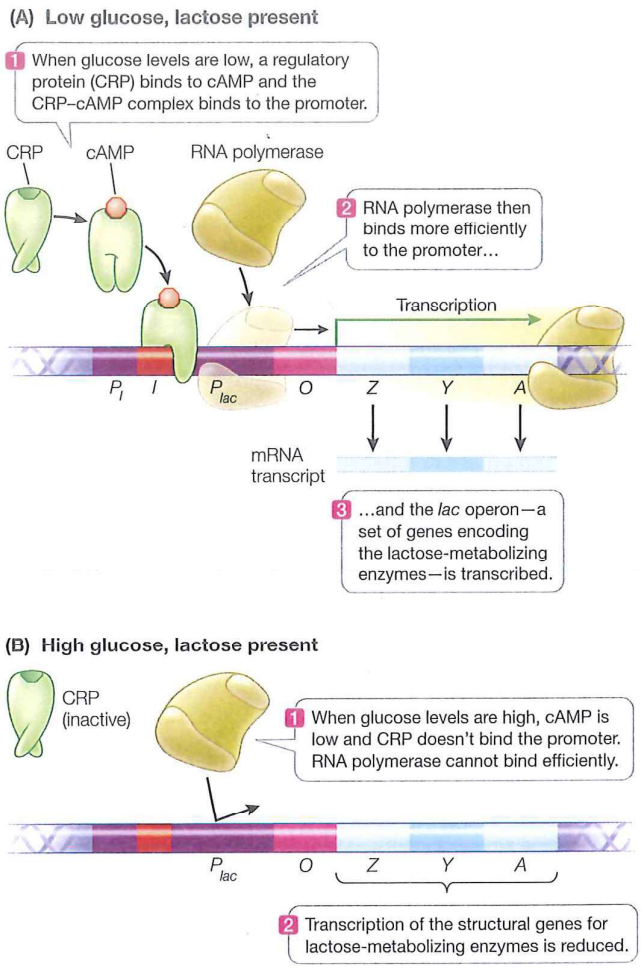


Figure 16.7 Catabolite Repression Regulates the *lac* Operon
(A) The promoter for the *lac* operon binds RNA polymerase more efficiently in the presence of cAMP, as occurs when glucose levels are low. (B) High glucose levels thus lead to low levels of transcription of the *lac* operon.

RNA polymerases can be directed to particular classes of promoters

Thus far we have described a promoter as a specific DNA sequence located upstream of a transcription initiation site. The promoter binds RNA polymerase so that the enzyme can then catalyze the synthesis of RNA from a gene-coding region of DNA. The promoter also orients the polymerase so that it transcribes the correct one of the two DNA strands for that gene. Not all promoters are identical, but they all have similar sequences by which they are recognized by the RNA polymerase and other proteins. Prokaryotic promoters generally have two sites for these recognition sequences, which begin 10 and 35 base pairs upstream of the transcription start site (the –10 element and the –35 element). Different classes of promoters have different recognition sequences at these two sites. The largest class consists of promoters for "housekeeping genes," which are all the genes that are normally expressed in actively growing cells. In these genes, the –10 element is 5′-TATAAT-3′, and the –35 element is 5′-TTGACAT-3′ (N stands for any nucleotide):

5′-NNNTTGACATNNNNNNNNNNNNNNNNNNTATAATNNNN*NNNNNNNNNNNN-3′
3′-NNNAACTGTANNNNNNNNNNNNNNNNNNATATTANNNN*NNNNNNNNNNNN-5′
–35 Element –10 Element Transcription start site

TABLE 16.1 | Positive and Negative Regulation in the *lac* Operon

Glucose	cAMP levels	CRP regulator	Lactose	*lac* Repressor	RNA polymerase binding?	Transcription of *lac* genes?	Lactose used by cells?
Present	Low	Not bound to promoter	Absent	Bound to operator	Absent	No	No
Present	Low	Not bound to promoter	Present	Not bound to operator	Present, not efficient	Low level	No
Absent	High	Bound to promoter	Absent	Bound to operator	Absent	No	No
Absent	High	Bound to promoter	Present	Not bound to operator	Present, very efficient	High level	Yes

Other classes of genes have different recognition sequences at their –10 and –35 elements. Why do DNA recognition sequences differ in different classes of promoters? After all, they are all binding the same protein, RNA polymerase. The answer lies in the fact that these DNA sequences bind not just RNA polymerase but other proteins as well. And it is those other proteins that control how well RNA polymerase binds to certain promoter sequences, thereby giving the system the ability to regulate classes of genes together.

Sigma factors are the proteins in prokaryotic cells that bind to RNA polymerase and direct it to specific classes of promoters. The RNA polymerase must be bound to a sigma factor before it can recognize a promoter and begin transcription. For example, the sigma-70 factor is active most of the time and binds to the recognition sequences of housekeeping genes; other sigma factors are activated only under specific conditions and bind to other promoter sequences. For example, when *E. coli* cells experience conditions such as DNA damage or osmotic stress, the sigma-38 factor is activated, and it directs RNA polymerase to the promoters of various genes that are expressed under stress conditions. *E. coli* has seven sigma factors; this number varies in other prokaryotes.

Regulation of proteins directing RNA polymerase to certain promoters is not unusual. In fact, you'll see in the next key concept that it is also common in eukaryotes.

KEY CONCEPT

16.1 Recap and Assess

Gene expression in prokaryotes is most commonly regulated through control of transcription. An operon consists of two or more adjacent structural genes and a single set of regulatory sequences (promoter and operator) that control their transcription. Operons can be regulated by both negative and positive controls. Sigma factors control the expression of specific classes of prokaryotic genes that share recognition sequences in their promoters.

1. A prokaryotic cell can metabolize sugar "X," using the enzyme "Xase." When there is a low concentration of X in the environment, there is low activity of Xase in the cells; but when the X concentration is high, Xase activity is also high. What five mechanisms could the cell use to reduce the activity of Xase in the absence of X?
2. Assuming that glucose is absent, compare the *lac* operon promoter in the presence versus absence of lactose in terms of what proteins can and cannot bind.
3. How do sigma factors and recognition sequences act to affect the expression of classes of genes?
4. The repressor protein that acts on the *lac* operon of *E. coli* is encoded by a regulatory gene. Suppose that a mutation occurs in the regulatory gene that renders it unable to bind to DNA. What would the effect of such a mutation be on the regulation of the *lac* operon?

Studies of bacteria have provided a basic understanding of mechanisms that regulate gene expression and of the roles of regulatory proteins in both positive and negative regulation. You'll see these same types of mechanisms again as we now turn to the transcriptional control of gene expression in eukaryotes.

KEY CONCEPT

16.2 Eukaryotic Gene Expression Is Regulated by Transcription Factors

Learning Objectives

- **16.2.1** Distinguish between general and specific transcription factors.
- **16.2.2** Describe the function of mediator.
- **16.2.3** State the similarities and differences between enhancers and silencers.
- **16.2.4** Contrast coordinated regulation in prokaryotes and eukaryotes.
- **16.2.5** Give an example of a structural motif that is present in some transcription factors that allows them to recognize and bind specific regions of DNA.

For cell function in single-celled eukaryotes as well as the normal development of a multicellular organism from fertilized egg to adult, certain proteins must be made at just the right times and in just the right cells; these proteins must not be made at other times in other cells. Here are two examples from humans:

1. In human pancreatic exocrine cells, the digestive enzyme procarboxypeptidase A makes up 7.6% of all the protein in the cell; in other cell types it is usually undetectable.
2. In human breast duct cells, alpha-lactalbumin, a protein in breast milk, is made only late in pregnancy and during lactation. Alpha-lactalbumin is not made in any other cell types or at any other times.

Clearly the expression of eukaryotic genes is regulated.

As in prokaryotes, gene expression in eukaryotes can be regulated at several different points in the process of transcribing and translating the gene into a protein (Figure 16.8). In this key concept we describe the mechanisms that result in the selective transcription of specific genes. The mechanisms for regulating gene expression in eukaryotes have similar themes as in prokaryotes. Both types of cells use DNA–protein interactions and negative and positive regulation. However, there are many differences, some of them dictated by the presence in eukaryotes of a nucleus, which physically separates transcription and translation (Table 16.2). In the following discussion we restrict our attention to the initiation of transcription for protein-coding genes, which are transcribed by RNA polymerase II, one of the three RNA polymerases found in eukaryotes.

General transcription factors act at eukaryotic promoters

As in prokaryotes, a promoter in a eukaryotic gene is a sequence of DNA near the 5′ end of the coding region, where RNA polymerase binds and initiates transcription. Although eukaryotic promoters are more diverse than those of prokaryotes, many protein-coding genes contain a nucleotide sequence similar to the –10 element in prokaryotic promoters. This element is usually located close to the transcription start site and is called the **TATA box** because it is rich in A-T base pairs. When present, the TATA box is the site

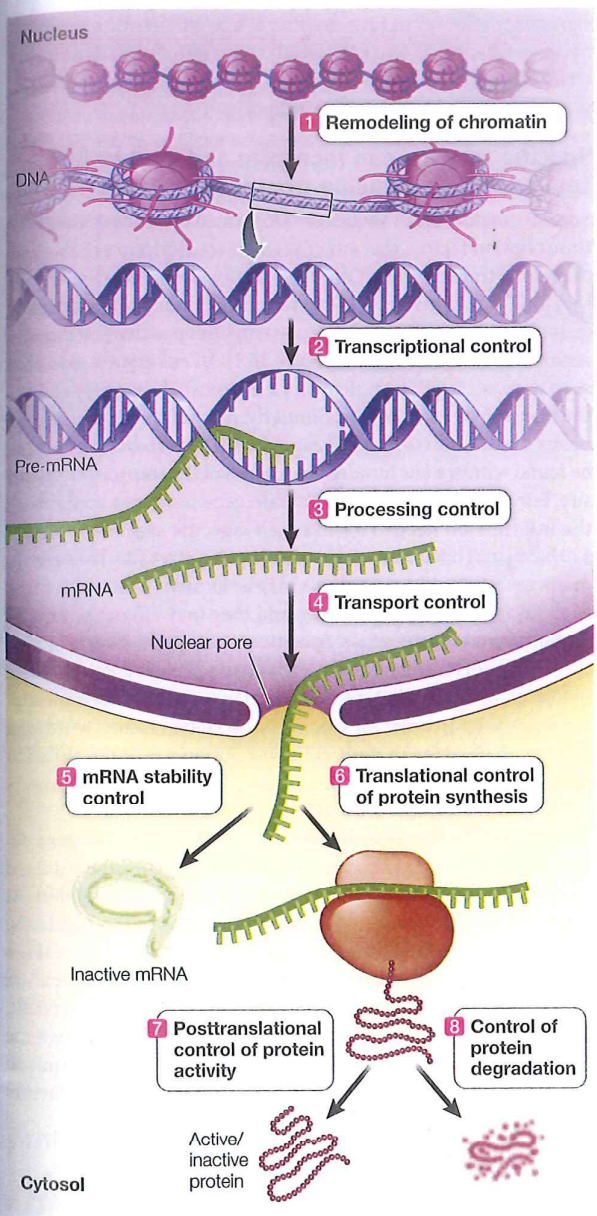


Figure 16.8 Potential Points for the Regulation of Gene Expression in Eukaryotes Gene expression can be regulated before transcription (1), during transcription (2), after transcription but before translation (3, 4, 5), at translation (6), or after translation (7, 8).

Q: In prokaryotes, transcription and translation are often coupled in time and space, but in eukaryotes they are separated. What are the advantages of the nucleus as a compartment?

View in Achieve
Activity 16.1 Eukaryotic Gene Expression Control Points

where DNA begins to denature so that the template strand can be exposed. In addition to having a TATA box, eukaryotic promoters typically include multiple regulatory sequences that are recognized and bound by **transcription factors**: regulatory proteins that help control transcription.

Like the prokaryotic RNA polymerase, eukaryotic RNA polymerase II cannot simply bind to the promoter and initiate transcription. Rather, it does so only after various general transcription factors have assembled on the chromosome to form the basal transcription apparatus (Figure 16.9). **General transcription factors** are proteins that bind to most promoters and are distinct from **specific transcription factors** that act only at certain promoters or classes of promoters. First, the protein complex called TFIID ("TF" stands for transcription factor) binds to the TATA box. Binding of TFIID changes both its own shape and that of the DNA, presenting a new surface that attracts the binding of other general transcription factors to form a basal transcription apparatus, which is the minimal set of proteins needed to initiate transcription. RNA polymerase II binds only after several other proteins have bound to this apparatus.

Each general transcription factor has a role in gene expression:

- TFIIB binds both RNA polymerase II and TFIID and helps identify the transcription initiation site.
- TFIIF prevents nonspecific binding of the basal transcription apparatus to DNA and helps recruit RNA polymerase II; it is similar in function to a bacterial sigma factor.
- TFIIE binds to the promoter and stabilizes the denaturation of the DNA.
- TFIIH opens up the DNA for transcription.

Some regulatory DNA sequences, such as the TATA box, are common to the promoters of many eukaryotic genes and are

TABLE 16.2 | Transcription in Prokaryotes and Eukaryotes

	Prokaryotes	Eukaryotes
Locations of functionally related genes	Often clustered in operons with one promoter	Often distant from one another with separate promoters
RNA polymerases	One	Three (RNA pol I, II, and III)
		I transcribes rRNA
		II transcribes mRNA
		III transcribes tRNA and small RNAs
Regulatory sequences besides promoter	Few	Many
Initiation of transcription	Binding of RNA polymerase to promoter	Binding of general and specific transcription factors, and of RNA polymerase

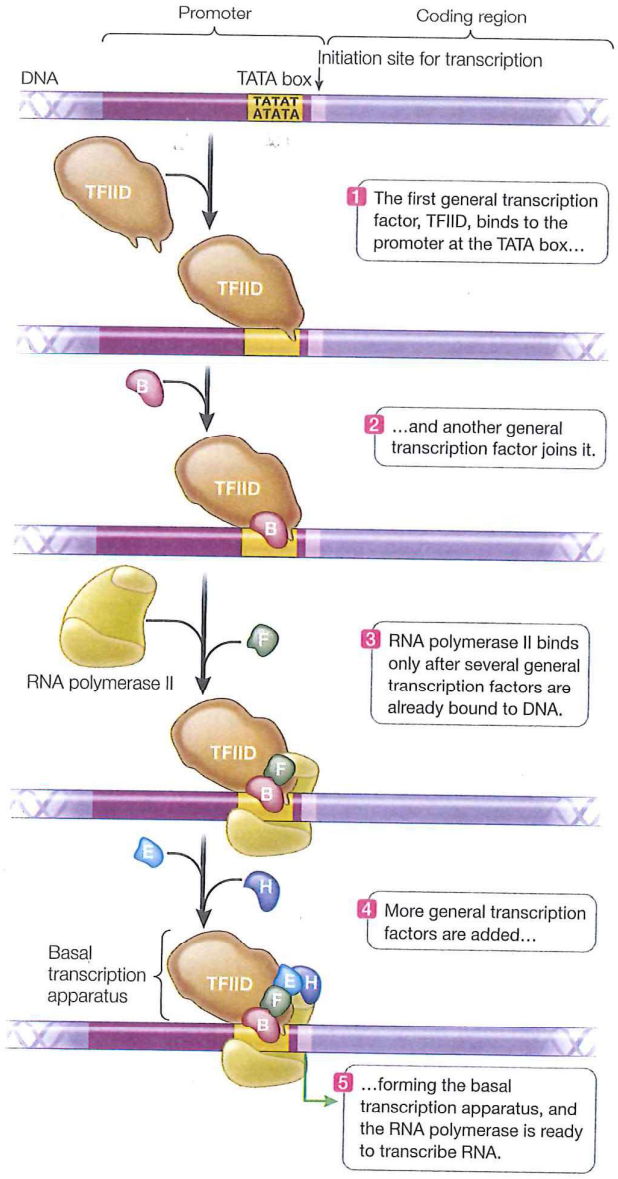


Figure 16.9 The Initiation of Transcription in Eukaryotes Apart from TFIID, which binds to the TATA box, each general transcription factor in this basal transcription apparatus has binding sites only for the other proteins in the apparatus and does not bind directly to DNA. B, E, F, and H are general transcription factors.

View in Achieve
Animation 16.3 Initiation of Transcription

recognized by general transcription factors that are found in all the cells of an organism. Other regulatory sequences are present in only a few genes and are recognized by specific transcription factors. These factors may be found only in certain types of cells or at certain stages of the cell cycle, or they may be activated by signaling pathways in response to cellular or environmental signals (see Chapter 7). Once the basal transcription apparatus forms, transcription can begin. However, other transcription factors also play a role in regulating transcription.

Specific proteins can recognize and bind to DNA sequences and regulate transcription

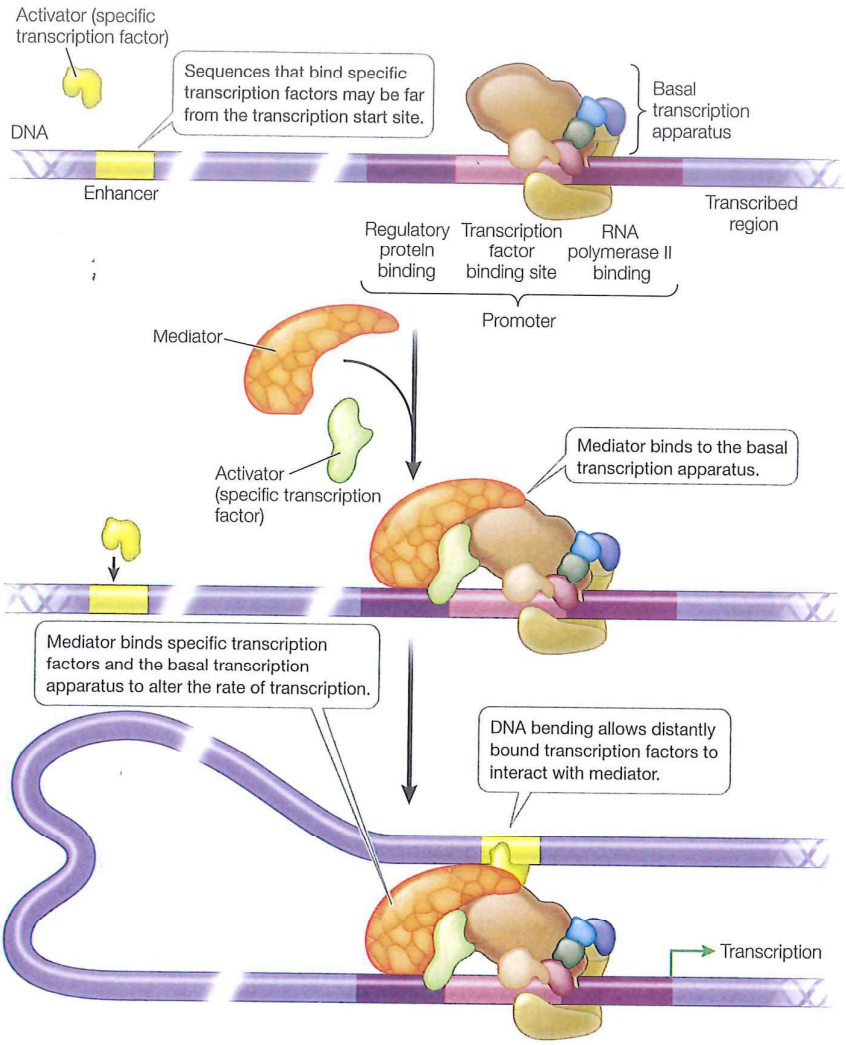
Some regulatory DNA sequences are termed **enhancers**: they bind transcription factors that either activate transcription or increase the rate of transcription. Other regulatory elements are **silencers**: they bind factors that repress transcription. Binding of transcription factors to enhancers and silencers exemplifies positive and negative regulation respectively (see Figure 16.1). In eukaryotic microbes such as yeast, regulatory elements are almost always very close to the transcriptional start site. Similarly, most of the regulatory elements needed for correct expression of many vertebrate genes can be found within a few hundred base pairs of the transcription start site. For example, the mouse albumin gene promoter contains all the information needed for liver cell–specific expression within 170 base pairs upstream of the transcription start site. However, in many genes, regulatory elements may be located thousands or tens of thousands of base pairs away, and they may affect the expression of several nearby genes. Specific transcription factors bind to their regulatory elements and then bind to a protein complex called **mediator**, which facilitates the interaction with the basal transcription apparatus. In order for specific transcription factors to interact through mediator with the basal transcription apparatus, the DNA bends, bringing distant regulatory sequences into close proximity to the promoter (Figure 16.10).

The combination of transcription factors binding to a gene determines the rate of transcription. For example, the immature red blood cells in bone marrow make large amounts of β-globin. At least 13 different transcription factors are involved in regulating transcription of the β-globin gene in these cells. Not all of these factors are present or active in other cells, such as the immature white blood cells produced by the same bone marrow. As a result, the β-globin gene is not transcribed in those cells. So although the same genes are present in all cells, the fate of the cell is determined by which of its genes are expressed. How do transcription factors recognize specific DNA sequences?

Specific protein–DNA interactions underlie binding

As we have seen, transcription factors with specific DNA-binding domains are involved in the activation and inactivation of specific genes. There are several common structural themes in the protein domains that bind to DNA in both eukaryotes and prokaryotes. These themes, or **structural motifs**, consist of different combinations of structural elements (protein conformations) and may include special components such as zinc. One common structural motif is the helix-turn-helix, in which two α helices are connected via a non-helical turn. The interior-facing "recognition" helix interacts with the bases inside the DNA. The exterior-facing helix sits on the sugar–phosphate backbone, ensuring that the interior helix is presented to the bases in the correct configuration (Figure 16.11).

Figure 16.10 Transcription Factors and Transcription Initiation The actions of many proteins determine whether and where RNA polymerase II will transcribe DNA. This example shows binding of two specific transcription factor activators, one at the regulatory promoter and one at an enhancer element.



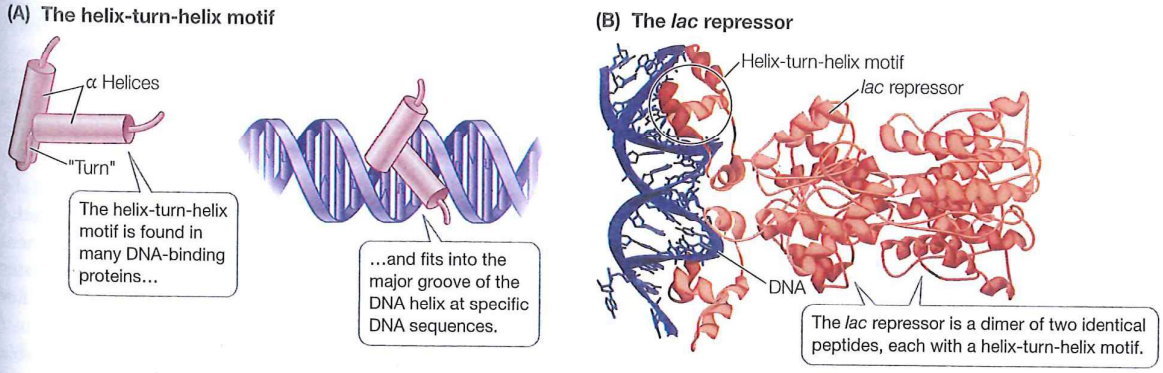


Figure 16.11 The Helix-Turn-Helix DNA-Binding Motif **(A)** The helix-turn-helix is one of several structural motifs seen in protein domains that bind to DNA. Such motifs facilitate recognition of, and binding to, specific DNA sequences. **(B)** The *lac* repressor binds to its operator using helix-turn-helix motifs. Data from PBD 2PE5. R. Daber et al. 2007. *J Mol Biol* 370: 609–619.

Connect the Concepts The structure and chemistry of DNA are key to its recognition by proteins. How the shapes and chemical structures of proteins allow them to bind noncovalently to other molecules is covered in Key Concept 3.2.

How does a protein recognize a sequence in DNA? As you learned in Key Concept 3.2, the complementary bases in DNA not only form hydrogen bonds with each other, but also can form additional hydrogen bonds with proteins, particularly at points exposed in the major and minor grooves. In this way, an intact DNA double helix can be recognized by a protein motif whose structure:

- fits into the major or minor groove;
- has amino acids that can project into the interior of the double helix; and
- has amino acids that can form hydrogen bonds with the interior bases.

Transcription factors underlie cell differentiation

During the development of a complex organism from fertilized egg to adult, cells become more and more differentiated (specialized). Since all differentiated cells contain the entire genome, their specific characteristics must arise from differential gene expression. Changes in gene expression result from the activation (and inactivation) of various transcription factors.

One well-studied example of cell differentiation is the conversion of undifferentiated muscle precursor cells into cells destined to form muscle (Figure 16.12). A key event in the commitment of these cells to become muscle is that they stop dividing. Indeed, during development, *cell division and cell differentiation are often mutually exclusive.* Cell signaling activates the gene for a transcription factor called MyoD (for *myo*blast-*d*etermining gene); this in turn activates the gene for p21, which is an inhibitor of the cyclin-dependent kinases (CDKs) that normally stimulate the cell cycle at G1 phase (see Figure 11.6). Expression of the *p21* gene causes the cell cycle to stop, and other transcription factors then enter the picture so that differentiation can proceed. Interestingly, MyoD is also activated in the stem cells that are present in adult muscle, indicating a role for this transcription factor in the repair of muscle tissue as it gets damaged and worn out.

Genes such as *myoD* that direct the most fundamental decisions in development (often by regulating other genes on other chromosomes) usually encode transcription factors. In some cases, a single transcription factor can cause a cell to differentiate in a certain way. In others, complex interactions between genes and proteins determine a sequence of transcriptional events that leads to differentiation.

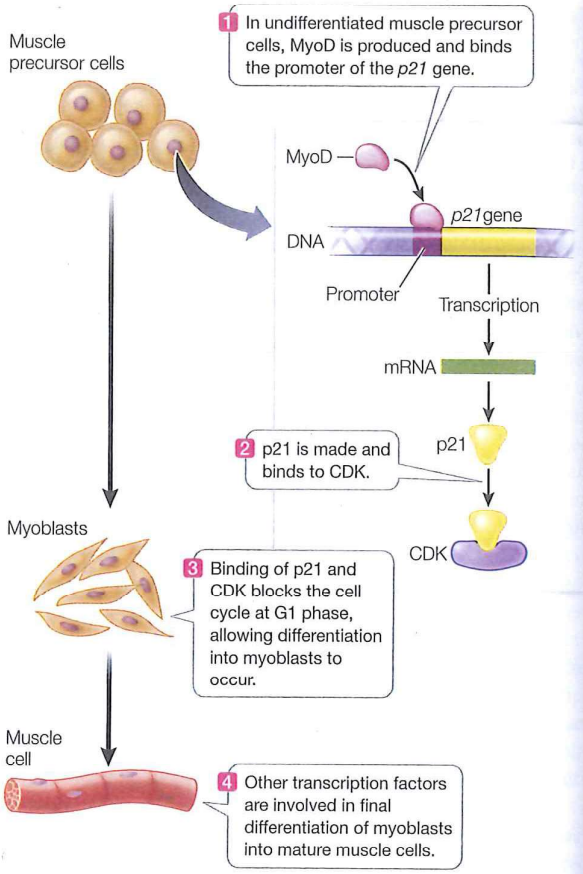


Figure 16.12 Transcription and Differentiation in the Formation of Muscle Cells Production of the transcription factor MyoD is important in muscle cell differentiation.

The expression of sets of genes can be coordinately regulated by transcription factors

How do eukaryotic cells coordinate the regulation of several genes whose transcription must be turned on at the same time? Prokaryotes solve this problem by arranging multiple genes in an operon that is controlled by a single promoter, and by using sigma factors to recognize particular classes of promoters. Most eukaryotic genes have their own separate promoters, and genes that are coordinately regulated may be far apart. In these cases, the expression of genes can be coordinated if they share regulatory sequences that bind the same transcription factors.

Shared regulatory sequences enable organisms to respond to stress—plants, for example, use shared regulatory sequences to respond to drought. Under conditions of drought stress, a plant must simultaneously synthesize several proteins whose genes are scattered throughout the genome. To coordinate expression of the stress response, each of the associated genes has a specific regulatory sequence near its promoter called the stress response element (SRE). A transcription factor binds to this element and stimulates mRNA synthesis (Figure 16.13). The stress response proteins not only help the plant conserve water, but also protect the plant against excess salt in the soil and freezing. This finding has considerable importance for agriculture because crops are often grown under less than optimal conditions or are affected by weather.

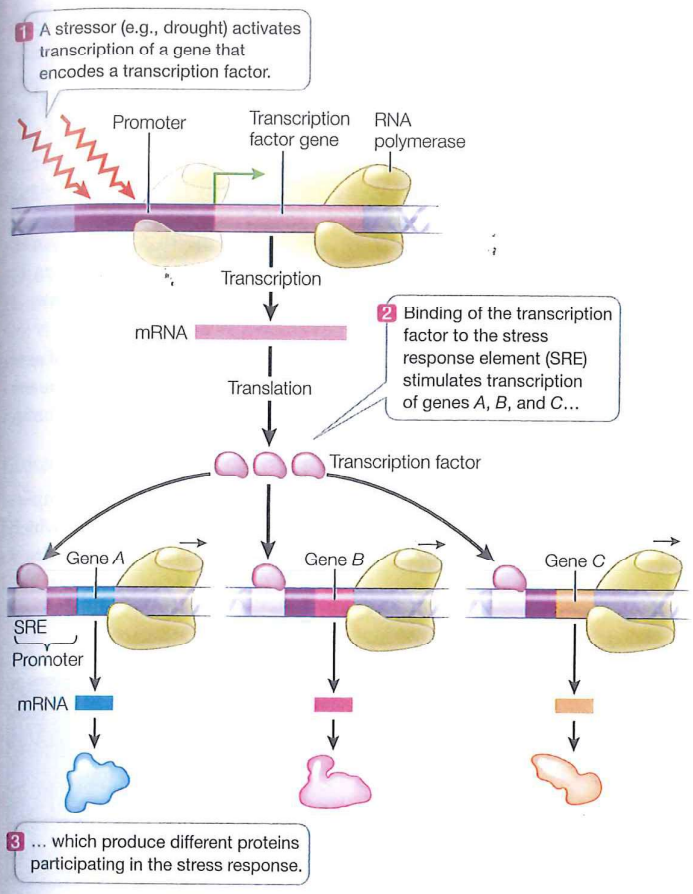


Figure 16.13 Coordinating Gene Expression A single environmental signal, such as drought stress, causes the synthesis of a transcription factor that acts on many genes.

KEY CONCEPT
16.2 Recap and Assess

Eukaryotes can increase or decrease transcription in various ways to help regulate gene expression. Several general transcription factors must bind to a eukaryotic promoter before RNA polymerase II will bind to it. Once the basal transcription apparatus forms, RNA polymerase II can begin transcription. Other, specific transcription factors bind to regulatory DNA sequences and interact with the basal transcription apparatus through mediator to control the level of gene expression. The regulation of gene expression determines cell differentiation, explaining why cells with the same DNA content can have different phenotypes.

1. How do specific and general transcription factors regulate the rate of gene transcription?
2. How do transcription factors recognize specific DNA sequences?
3. How can more than one gene show the same regulation in a eukaryote? How does this differ from coordinated gene regulation in prokaryotes?

We have seen how prokaryotes and eukaryotes regulate the transcription of their genes and operons. Next we will see how viruses can hijack prokaryotic and eukaryotic transcription mechanisms in order to complete their life cycles.

KEY CONCEPT
16.3 Viruses Regulate Their Gene Expression during the Reproductive Cycle

Learning Objectives

- **16.3.1** Distinguish between the lytic and lysogenic cycles of a virus.
- **16.3.2** Explain the control of early versus late gene transcription during viral infection.
- **16.3.3** Give an example of how host transcription is turned off during viral infection.
- **16.3.4** Explain how expression of some HIV genes is regulated during the elongation stage of transcription.
- **16.3.5** Detail the steps at which different anti-HIV drugs work.

"A virus is a piece of bad news wrapped in protein." This quote from immunologist Sir Peter Medawar is certainly true for the cells that viruses infect. As we described in Chapter 13, a bacterial virus (**bacteriophage**) injects its genetic material into a host bacterium and turns that cell into a virus factory (see Figure 13.3). Other viruses enter cells intact and then shed their coats and take over the cell's replication machinery. Viral life cycles can be very efficient. An example is the poliovirus: a single poliovirus infecting a single mammalian cell can produce more than 100,000 new virus particles!

Understanding the reproductive cycle of a virus makes it possible to design therapeutic agents to fight infections by the virus. **Viruses** are small infective agents that make copies of themselves inside cellular organisms and that cannot reproduce outside host cells. Most virus particles, called **virions**, consist of only two or three components: the genetic material made up of DNA or RNA, a protein coat that protects the genetic material, and in some cases, an envelope of lipids that surrounds the protein coat. As we will see in this key concept, viral genomes include sequences that encode regulatory proteins. These proteins hijack the host cells' transcriptional machinery, allowing the viruses to complete their reproductive cycles.

Viruses undertake two kinds of reproductive cycles

After a virus infects a cell, typically it takes over the cell's molecular genetic machinery and begins the production of new virus. But in some cases there is an alternate series of events, in which the viral genome becomes integrated into the host genome.

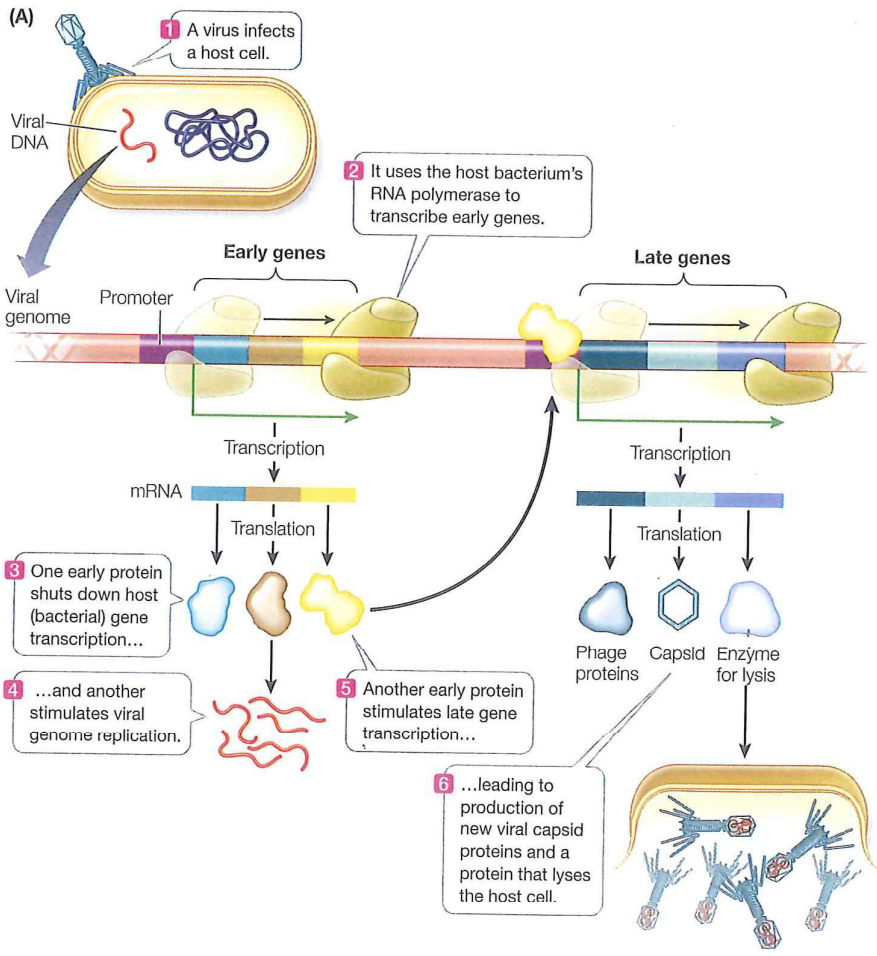
LYTIC CYCLE The Hershey–Chase experiment (see Figure 13.4) involved a typical **lytic** viral reproductive cycle, so named because soon after infection, the host cell bursts (lyses), releasing progeny viruses. In this cycle, the viral genome takes over the host's genetic machinery for its own reproduction immediately after infection. In the case of some bacteriophages, the process is extremely rapid—within 15 minutes, new phage particles appear in the bacterial cell. Ten minutes later, the "game is over," and these particles are released from the lysed cell. What happened?

At the molecular level, the reproductive cycle of a typical lytic virus has two stages: early and late, as illustrated in Figure 16.14. Follow along in the text and Figure 16.14 and you'll see examples of both stimulation and inhibition of host and viral gene expression:

- The viral genome contains a promoter that is recognized by a host RNA polymerase that is bound to the housekeeping sigma factor, sigma-70. As a result, in the early stage (1–2 min after phage DNA entry), viral genes that lie adjacent to this promoter are transcribed.
- These early genes often encode proteins that shut down host transcription. In some bacteriophages, one of these genes encodes a protein that binds to host sigma factors, especially sigma-70, inactivating them. In addition, proteins from the early genes stimulate viral genome replication and transcription of viral late genes. Three minutes after DNA entry, viral nuclease enzymes digest the host's chromosome, providing nucleotides for the synthesis of viral genomes.
- In the late stage, viral late genes are transcribed; they encode the proteins that make up the **capsid** (the outer shell of the virus) and other protein components of the virus and enzymes that lyse the host cell to release the new virions. This begins 9 minutes after DNA entry and 6 minutes before the first new phage particles appear.

The entire process—from binding and infection to release of new phage—takes about 30 minutes. During this period, the sequence of transcriptional events is carefully controlled to produce complete, infective virions.

LYSOGENIC CYCLE Some viruses have evolved a more complex reproductive cycle that includes a process called **lysogeny**, which postpones the lytic cycle. In lysogeny, the viral DNA becomes



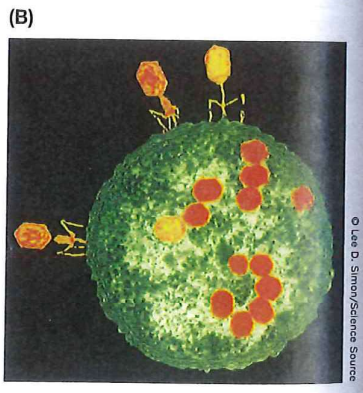


Figure 16.14 The Lytic Cycle: A Strategy for Viral Reproduction
(A) In a host cell infected with a virus, the viral genome uses its early genes to shut down host transcription while it replicates itself. Once the viral genome is replicated, its late genes produce capsid proteins that package the genome and other proteins that lyse the host cell. **(B)** Bacteriophages have attached to this *E. coli* cell, and the reproductive cycle is underway, producing new phage particles. The cell is viewed in transverse section.

Q: How does the virus co-express early genes and then co-express late genes?

integrated into the host DNA and is not expressed. A bacteriophage genome that is inserted into the host genome and is not currently producing progeny phage is termed a **prophage** (Figure 16.15). As the host cell divides, the viral DNA gets replicated as part of the host chromosome. The prophage can remain inactive within the bacterial genome for thousands of generations, producing many copies of the original viral DNA, but no progeny bacteriophage.

However, under certain conditions, such as if the host cell is not growing well, the virus "cuts its losses." It switches to a lytic cycle, in which the prophage excises itself from the host chromosome and reproduces, producing progeny phage that lyse the cell. Lysogeny allows a virus to make copies of its genome without harming its host by inserting its DNA into the host chromosome, where it sits as a silent passenger until conditions are right for lysis.

Eukaryotic viruses can have complex life cycles

Eukaryotes are susceptible to infection by various kinds of viruses (see Key Concept 24.4), which can be broadly categorized based on whether their genome is single- or double-stranded DNA or RNA and how they replicate.

- *DNA viruses.* Some viruses contain double-stranded DNA genomes. Others have single-stranded DNA, and a complementary strand is made after the viral genome enters the host cell. Like some bacteriophages, DNA viruses that infect eukaryotes are capable of undergoing both lytic and lysogenic life cycles. Examples include the herpes viruses and papillomaviruses (which cause warts).
- *RNA viruses.* Most RNA virus genomes are single-stranded, though some are double-stranded. The RNA is translated by the host's machinery to produce viral proteins, some of which are involved in replication of the RNA genome. The influenza virus has an RNA genome.
- *Retroviruses.* As we described in Key Concept 14.2, a **retrovirus** is an RNA virus that carries a gene for **reverse transcriptase**, a protein that synthesizes DNA from an RNA template. The retrovirus uses this protein to make a DNA copy of its genome, which then becomes integrated into the host genome. The integrated DNA acts as a template for both mRNA and new viral genomes. HIV is a retrovirus that infects cells of the immune system and causes acquired immune deficiency syndrome (AIDS; see Key Concept 40.6).

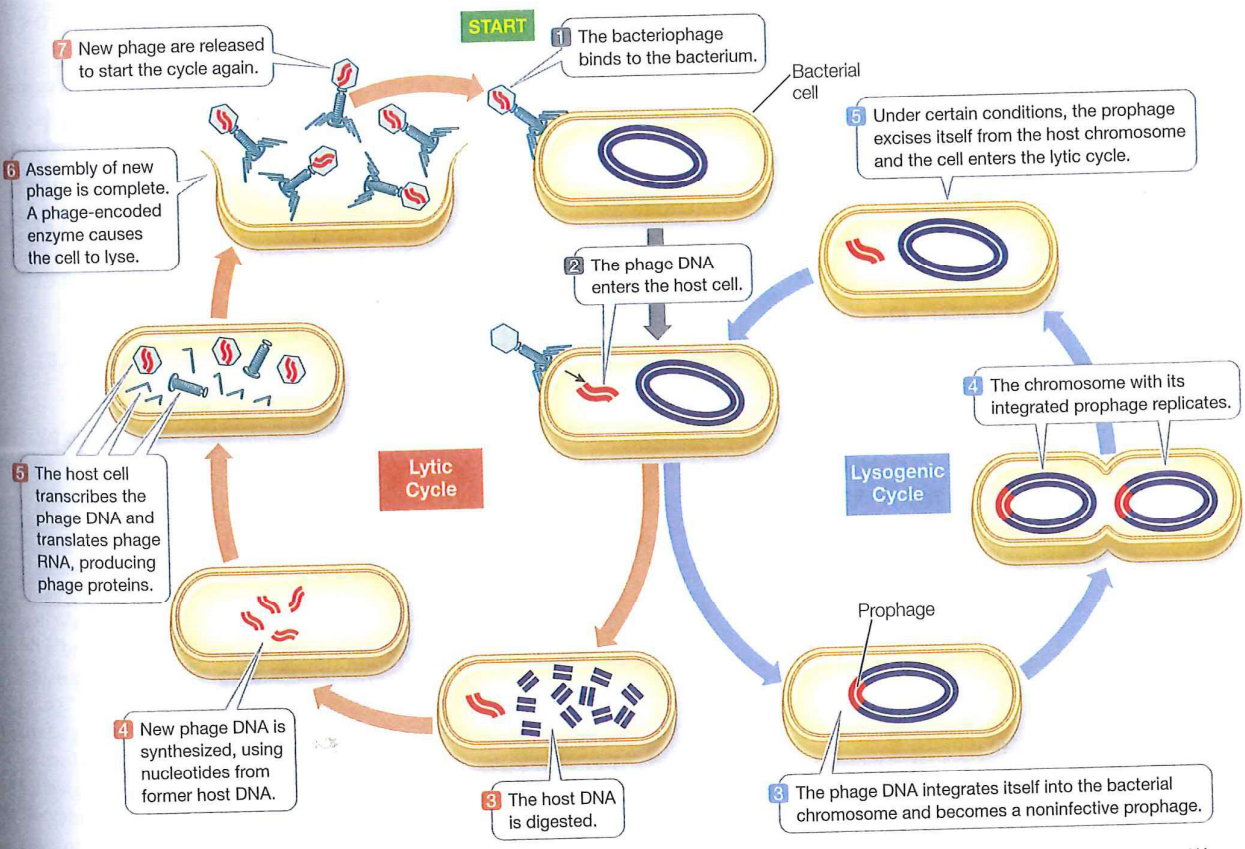


Figure 16.15 Some Bacteriophages Have Both Lytic and Lysogenic Cycles In the lytic cycle, infection of a bacterium by viral DNA leads directly to multiplication of the virus and lysis of the host cell. In the lysogenic cycle, an inactive prophage is integrated into the host DNA, where it is replicated during the bacterial life cycle.

Connect the Concepts Viral diversity is discussed in Key Concept 25.4, which explains why the genomes of some viruses consist of single-stranded RNA and how RNA retroviruses reproduce themselves by reverse transcription.

HIV gene regulation occurs at the level of transcription elongation

As we have discussed so far, many instances of gene regulation occur at the level of transcription *initiation*, involving both activator and repressor proteins that bind to the promoters of genes. However, studies of HIV and other viruses have revealed that transcription can also be controlled at the *elongation* stage.

HIV is an **enveloped virus**; it is enclosed within a phospholipid membrane derived from its host cells (a specific type of immune system cell) (Figure 16.16). During infection, proteins in this membrane interact with proteins on the host cell surface, and the viral envelope fuses with the host cell membrane. After the virus enters the cell, its capsid is broken down. The viral reverse transcriptase then uses the virus's RNA template to produce a complementary DNA (cDNA) strand, while at the same time degrading the viral RNA. The enzyme then makes a complementary copy of the cDNA, and the resulting double-stranded DNA is inserted into the host's chromosome by an enzyme appropriately named integrase. The integrated DNA is referred to as the **provirus**. Both the reverse transcriptase and the integrase are carried inside the HIV virion, along with other proteins needed at the very early stages of infection.

The HIV provirus resides permanently in the host chromosome and can remain in a latent (inactive) state for many years. During this time transcription of the viral DNA is initiated, but host cell proteins prevent the RNA from elongating, and transcription is terminated prematurely. Under some circumstances, such as when the host immune cell is activated, the level of transcription initiation increases, and some viral RNA is made. One of these viral genes encodes a protein called tat (*trans*activator of *t*ranscription), which binds to a stem-and-loop structure at the 5′ end of the viral RNA. As a result of tat binding, the production of full-length viral RNA is dramatically increased, and the rest of the viral reproductive cycle is able to proceed. It was only after the discovery of this mechanism in HIV and similar viruses that researchers found that many eukaryotic genes are regulated at the level of transcription elongation.

Almost every step in the reproductive cycle of HIV is, in principle, a potential target for drugs to treat AIDS. The classes of anti-HIV drugs currently in use include:

- Inhibitors of virus binding to target cells (at step 1 in Figure 16.16)
- Inhibitors of internalization of virus following binding (at step 1)
- Reverse transcriptase inhibitors that block viral DNA synthesis from RNA (at step 2)
- Integrase inhibitors that block the incorporation of viral DNA into the host chromosome (at step 3)
- Protease inhibitors that block the posttranslational processing of viral proteins (at step 5)

Combinations of drugs from these classes have been spectacularly successful in treating HIV infection.

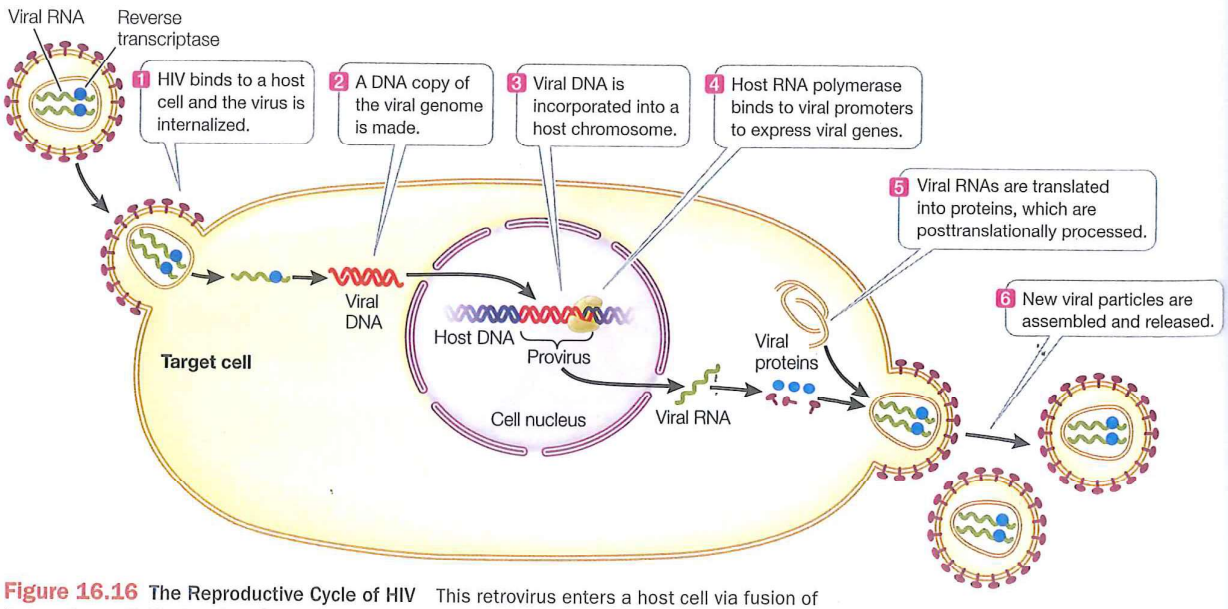


Figure 16.16 The Reproductive Cycle of HIV This retrovirus enters a host cell via fusion of its envelope with the host's cell membrane. Reverse transcription of retroviral RNA then produces a DNA provirus—a molecule of complementary DNA that inserts itself into the host's genome.

Q: Looking at Figures 16.15 and 16.16, would you classify HIV as a lytic or a lysogenic virus?

MOLECULAR BIOLOGY OF THE GENE

SEVENTH EDITION

JAMES D. WATSON
Cold Spring Harbor Laboratory

ALEXANDER GANN
Cold Spring Harbor Laboratory

TANIA A. BAKER
Massachusetts Institute of Technology

MICHAEL LEVINE
University of California, Berkeley

STEPHEN P. BELL
Massachusetts Institute of Technology

RICHARD LOSICK
Harvard University

With
STEPHEN C. HARRISON
Harvard Medical School
(Chapter 6: The Structure of Proteins)

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

CSH PRESS
COLD SPRING HARBOR LABORATORY PRESS
Cold Spring Harbor, New York

PEARSON

Editor-in-Chief: Beth Wilbur
Senior Acquisitions Editor: Josh Frost
Executive Director of Development: Deborah Gale
Assistant Editor: Katherine Harrison-Adcock
Managing Editor: Michael Early
Production Project Manager: Lori Newman
Illustrators: Dragonfly Media Group
Manufacturing Buyer: Michael Penne
Director of Marketing: Christy Lesko
Executive Marketing Manager: Lauren Harp
Executive Media Producer: Laura Tommasi
Editorial Media Producer: Lee Ann Doctor
Supervising Media Project Manager: David Chavez
Director of Content Development, MasteringBiology: Natania Mlawer
Content Specialist, MasteringBiology: J. Zane Barlow, PhD

COLD SPRING HARBOR LABORATORY PRESS

Publisher and Sponsoring Editor: John Inglis
Editorial Director: Alexander Gann
Director of Editorial Development: Jan Argentine
Managing Editor and Developmental Editor: Kaaren Janssen
Project Manager: Inez Sialiano
Production Manager: Denise Weiss
Production Editor: Kathleen Bubbeo
Permissions Coordinator: Carol Brown
Crystal Structure Images: Leemor Joshua-Tor and Stephen C. Harrison
Cover Designer: Mike Albano

Front and Back Cover Images: Far left, drawing by Francis Crick, Wellcome Library, London. Second from left, from Watson J.D. and Crick F.H.C. 1953. *Nature* **171:** 737–738. Second from right, Irving Geis illustration. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission. Far right, structure by Leemor Joshua-Tor (image prepared with PyMOL).

Credits and acknowledgments for materials borrowed from other sources and reproduced, with permission, in this textbook appear on the appropriate page within the text.

Library of Congress Cataloging-in-Publication Data

Watson, James D.
Molecular biology of the gene / James D. Watson, Cold Spring Harbor Laboratory, Tania A. Baker, Massachusetts Institute of Technology, Alexander Gann, Cold Spring Harbor Laboratory, Michael Levine, University of California, Berkeley, Richard Losick, Harvard University.
pages cm
Includes bibliographical references and index.
ISBN-13: 978-0-321-76243-6 (hardcover (student ed))
ISBN-10: 0-321-76243-6 (hardcover (student ed))
ISBN-13: 978-0-321-90537-6 (paper (a la carte))
ISBN-10: 0-321-90537-7 (paper (a la carte))
[etc.]
1. Molecular biology--Textbooks. 2. Molecular genetics--Textbooks. I. Title.
QH506.M6627 2013
572'.33--dc23

2012046093

6 16

PEARSON

www.pearsonhighered.com

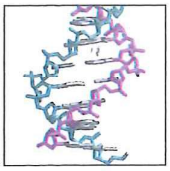
COLD SPRING HARBOR LABORATORY PRESS
www.cshlpress.org

ISBN 10: 0-321-76243-6 (Student Edition)
ISBN 13: 978-0-321-76243-6 (Student Edition)

ISBN 10: 0-321-90264-5 (Instructor's Review Copy)
ISBN 13: 978-0-321-90264-1 (Instructor's Review Copy)

ISBN 10: 0-321-90537-7 (Books à la Carte Edition)
ISBN 13: 978-0-321-90537-6 (Books à la Carte Edition)

CHAPTER 4



The Structure of DNA

THE DISCOVERY THAT DNA IS THE PRIME GENETIC molecule, carrying all of the hereditary information within chromosomes, immediately focused attention on its structure. It was hoped that knowledge of the structure would reveal how DNA carries the genetic messages that are replicated when chromosomes divide to produce two identical copies of themselves. During the late 1940s and early 1950s, several research groups in the United States and in Europe engaged in serious efforts—both cooperative and rival—to understand how the atoms of DNA are linked together by covalent bonds and how the resulting molecules are arranged in three-dimensional space. Not surprisingly, there initially were fears that DNA might have very complicated and perhaps bizarre structures that differed radically from one gene to another. Great relief, if not general elation, was thus expressed when the fundamental DNA structure was found to be the double helix. This told us that all genes have roughly the same three-dimensional form and that the differences between two genes reside in the order and number of their four nucleotide building blocks along the complementary strands.

Now, some 50 years after the discovery of the double helix, this simple description of the genetic material remains true and has not had to be appreciably altered to accommodate new findings. Nevertheless, we have come to realize that the structure of DNA is not quite as uniform as was first thought. For example, the chromosomes of some small viruses have single-stranded, not double-stranded, molecules. Moreover, the precise orientation of the base pairs varies slightly from base pair to base pair in a manner that is influenced by the local DNA sequence. Some DNA sequences even permit the double helix to twist in the left-handed sense, as opposed to the right-handed sense originally formulated for DNA's general structure. And some DNA molecules are linear, whereas others are circular. Still additional complexity comes from the supercoiling (further twisting) of the double helix, often around cores of DNA-binding proteins. Clearly, the structure of DNA is richer and more intricate than was at first appreciated. Indeed, there is no one generic structure for DNA. As we see in this chapter, there are, in fact, variations on common themes of structure that arise from the unique physical, chemical, and topological properties of the polynucleotide chain.

OUTLINE

DNA STRUCTURE

DNA Is Composed of Polynucleotide Chains

WEB STRUCTURAL TUTORIAL

The most important feature of DNA is that it is usually composed of two **polynucleotide chains** twisted around each other in the form of a double helix (Fig. 4-1; see Structural Tutorial 4-1). Figure 4-1a presents the structure of the double helix in a schematic form. Note that if inverted 180° (e.g., by turning this book upside down), the double helix looks superficially the same, because of the complementary nature of the two DNA strands. The space-filling model of the double helix in Figure 4-1b shows the components of the DNA molecule and their relative positions in the helical structure. The backbone of each strand of the helix is composed of alternating sugar and phosphate residues; the bases project inward but are accessible through the major and minor grooves.

Let us begin by considering the nature of the nucleotide, the fundamental building block of DNA. The **nucleotide** consists of a phosphate joined to a sugar, known as **2′-deoxyribose**, to which a base is attached. The phosphate and the sugar have the structures shown in Figure 4-2. The sugar is called 2′-deoxyribose because there is no hydroxyl at position 2′ (just two hydrogens). Note that the positions on the sugar are designated with primes to distinguish them from positions on the bases (see the discussion below).

We can think of how the base is joined to 2′-deoxyribose by imagining the removal of a molecule of water between the hydroxyl on the 1′ carbon of the sugar and the base to form a glycosidic bond (Fig. 4-2). The sugar and base alone are called a **nucleoside.** Likewise, we can imagine linking the phosphate to 2′-deoxyribose by removing a water molecule from between the phosphate and the hydroxyl on the 5′ carbon to make a 5′ phosphomonoester. Adding a phosphate (or more than one phosphate) to a **nucleoside** creates a **nucleotide.** Thus, by making a glycosidic bond between the base

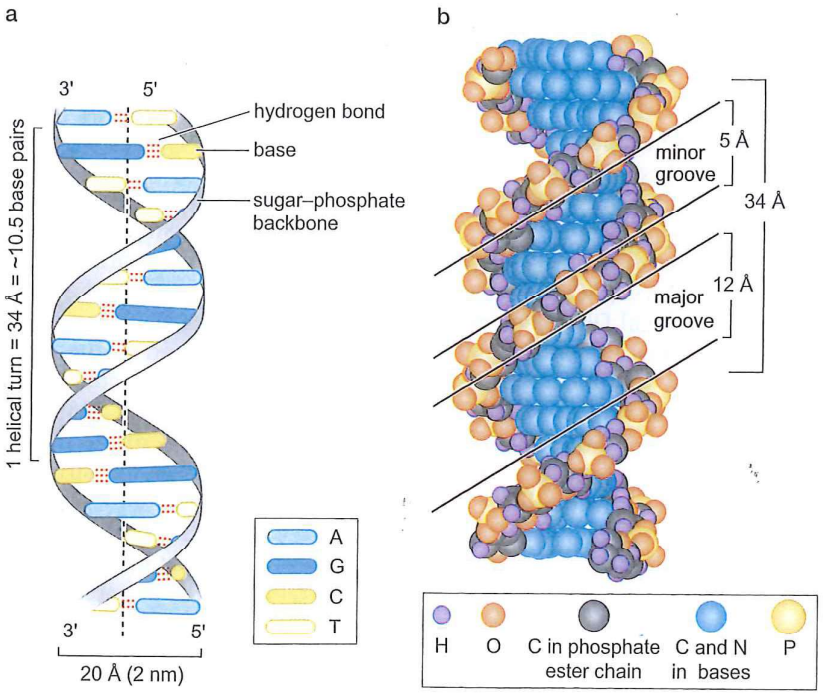


FIGURE 4-1 The helical structure of DNA. (a) Schematic model of the double helix. One turn of the helix (34 Å or 3.4 nm) spans ~10.5 bp. (b) Space-filling model of the double helix. The sugar and phosphate residues in each strand form the backbone, which is traced by the yellow, gray, and red circles, showing the helical twist of the overall molecule. The bases project inward but are accessible through major (12 Å or 1.2 nm) and minor (5 Å or 0.5 nm) grooves. These dimensions are approximate and vary with the sequence of DNA.

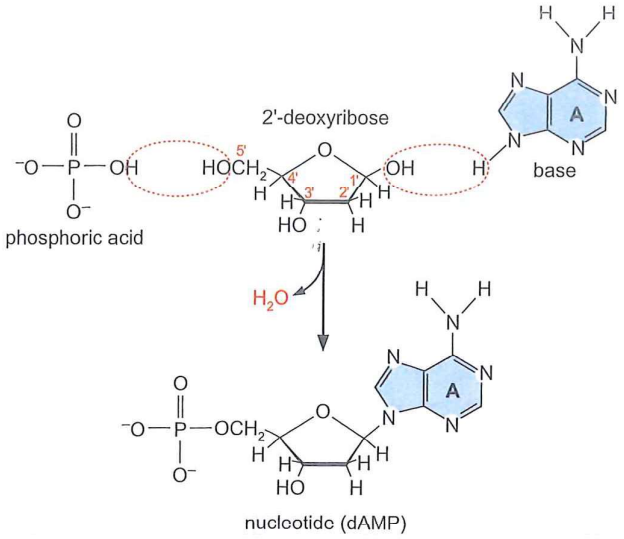


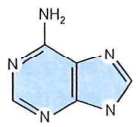
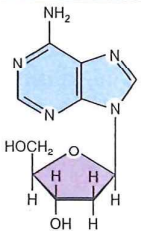
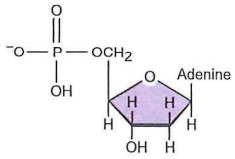
FIGURE 4-2 Formation of nucleotide by removal of water. The numbers of the carbon atoms in 2′-deoxyribose are labeled in red.

and the sugar, and by making a phosphoester bond between the sugar and the phosphoric acid, we have created a nucleotide (Table 4-1).

Nucleotides are, in turn, joined to each other in polynucleotide chains through the 3′-hydroxyl of 2′-deoxyribose of one nucleotide and the phosphate attached to the 5′-hydroxyl of another nucleotide (Fig. 4-3). This is a **phosphodiester linkage** in which the phosphoryl group between the two nucleotides has one sugar esterified to it through a 3′-hydroxyl and a second sugar esterified to it through a 5′-hydroxyl. Phosphodiester linkages create the repeating, sugar–phosphate backbone of the polynucleotide chain, which is a regular feature of DNA. In contrast, the order of the bases along the polynucleotide chain is irregular. This irregularity as well as the long length is, as we shall see, the basis for the enormous information content of DNA.

The phosphodiester linkages impart an inherent polarity to the DNA chain. This polarity is defined by the asymmetry of the nucleotides and the way they are joined. DNA chains have a free 5′-phosphate or 5′-hydroxyl at one end and a free 3′-phosphate or 3′-hydroxyl at the other end. The convention is to write DNA sequences from the 5′ end (on the left) to the 3′ end, generally with a 5′-phosphate and a 3′-hydroxyl.

TABLE 4-1 Adenine and Related Compounds

	Base Adenine	Nucleoside 2′-Deoxyadenosine	Nucleotide 2′-Deoxyadenosine 5′-Phosphate
Structure			
Molecular weight	135.1	251.2	331.2

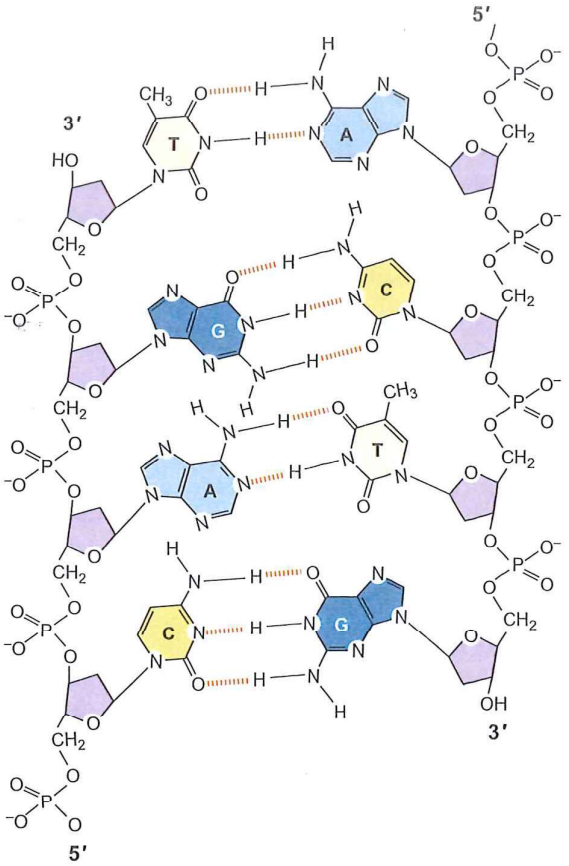


FIGURE 4-3 Detailed structure of polynucleotide polmer. The structure shows base pairing between purines (blue) and pyrimidines (yellow), and the phosphodiester linkages of the backbone. (Adapted from Dickerson R.E. 1983. *Sci. Am.* **249:** 94. Illustration, Irving Geis. Image from Irving Geis Collection/Howard Hughes Medical Institution. Not to be reproduced without permission.)

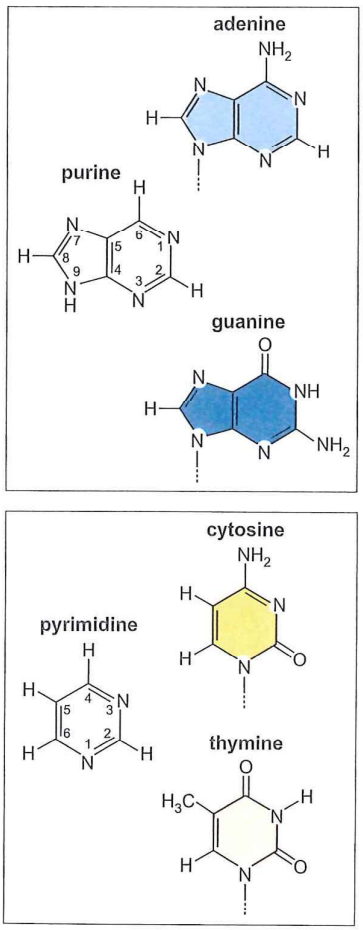


FIGURE 4-4 Purines and pyrimidines. The dotted lines indicate the sites of attachment of the bases to the sugars. For simplicity, hydrogens are omitted from the sugars and bases in subsequent figures, except where pertinent to the illustration.

Each Base Has Its Preferred Tautomeric Form

The bases in DNA are flat, heterocyclic rings, consisting of carbon and nitrogen atoms. The bases fall into two classes, **purines** and **pyrimidines.** The purines are **adenine** and **guanine,** and the pyrimidines are **cytosine** and **thymine.** The purines are derived from the double-ringed structure shown in Figure 4-4. Adenine and guanine share this essential structure but with different groups attached. Likewise, cytosine and thymine are variations on the single-ringed structure shown in Figure 4-4. The figure also shows the numbering of the positions in the purine and pyrimidine rings. The bases are attached to the deoxyribose by glycosidic linkages at N1 of the pyrimidines or at N9 of the purines.

Each of the bases exists in two alternative **tautomeric states**, which are in equilibrium with each other. The equilibrium lies far to the side of the conventional structures shown in Figure 4-4, which are the predominant states and the ones important for base pairing. The nitrogen atoms attached to the purine and pyrimidine rings are in the amino form in the predominant state and only rarely assume the imino configuration. Likewise, the oxygen atoms attached to the guanine and thymine normally have the keto form and only rarely take on the enol configuration. As examples, Figure 4-5 shows tautomerization of cytosine into the imino form (Fig. 4-5a) and guanine into the

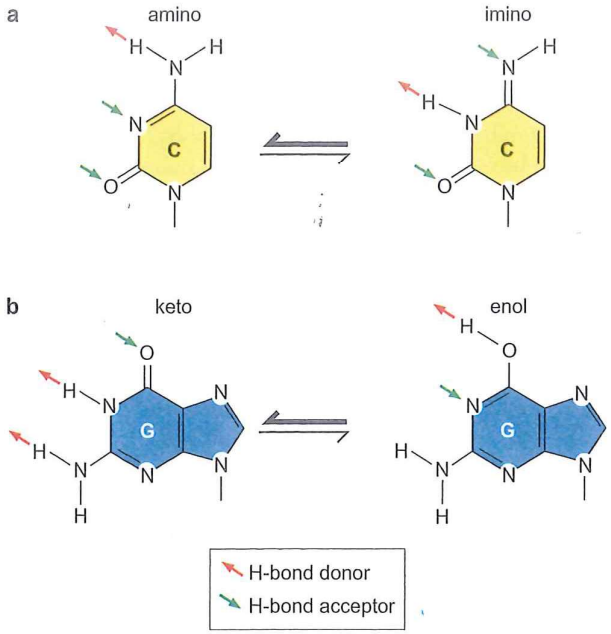


FIGURE 4-5 Base tautomers. Amino ⇌ imino and keto ⇌ enol tautomerism. (a) Cytosine is usually in the amino form but rarely forms the imino configuration. (b) Guanine is usually in the keto form but is rarely found in the enol configuration.

enol form (Fig. 4-5b). As we shall see, the capacity to form an alternative tautomer is a frequent source of errors during DNA synthesis.

The Two Strands of the Double Helix Are Wound around Each Other in an Antiparallel Orientation

The double helix consists of two polynucleotide chains that are aligned in opposite orientation. The two chains have the same helical geometry but have opposite 5′ to 3′ orientations. That is, the 5′ to 3′ orientation of one chain is antiparallel to the 5′ to 3′ orientation of the other strand, as shown in Figures 4-1 and 4-3. The two chains interact with each other by pairing between the bases, with adenine on one chain pairing with thymine on the other chain and, likewise, guanine pairing with cytosine. Thus, the base at the 5′ end of one strand is paired with the base at the 3′ end of the other strand. The antiparallel orientation of the double helix is a stereochemical consequence of the way that adenine and thymine, and guanine and cytosine, pair with each together.

The Two Chains of the Double Helix Have Complementary Sequences

The pairing between adenine and thymine, and between guanine and cytosine, results in a complementary relationship between the sequence of bases on the two intertwined chains and gives DNA its self-encoding character. For example, if we have the sequence 5′-ATGTC-3′ on one chain, the opposite chain must have the complementary sequence 3′-TACAG-5′.

The strictness of the rules for this "Watson–Crick" pairing derives from the complementarity both of shape and of hydrogen-bonding properties between adenine and thymine and between guanine and cytosine (Fig. 4-6). Adenine and thymine match up so that a hydrogen bond can form between the exocyclic amino group at C6 on adenine and the carbonyl at C4 in thymine; and likewise, a hydrogen bond can form between N1 of

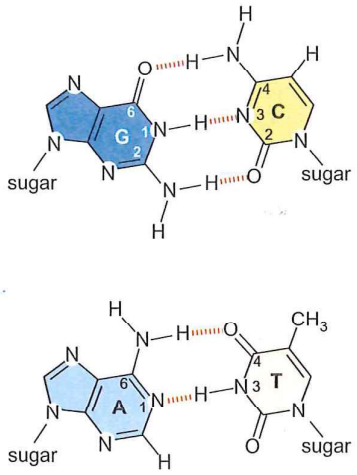


FIGURE 4-6 A:T and G:C base pairs. The figure shows hydrogen bonding between the bases.

adenine and N3 of thymine. A corresponding arrangement can be drawn between a guanine and a cytosine, so that there is both hydrogen bonding and shape complementarity in this base pair as well. A G:C base pair has three hydrogen bonds, because the exocyclic NH_2 at C2 on guanine lies opposite to, and can hydrogen-bond with, a carbonyl at C2 on cytosine. Likewise, a hydrogen bond can form between N1 of guanine and N3 of cytosine and between the carbonyl at C6 of guanine and the exocyclic NH_2 at C4 of cytosine. Watson–Crick base pairing requires that the bases be in their preferred tautomeric states.

An important feature of the double helix is that the two base pairs have exactly the same geometry; having an A:T base pair or a G:C base pair between the two sugars does not perturb the arrangement of the sugars because the distance between the sugar attachment points is the same for both base pairs. Neither does T:A or C:G. In other words, there is an approximately twofold axis of symmetry that relates the two sugars, and all four base pairs can be accommodated within the same arrangement without any distortion of the overall structure of the DNA. In addition, the base pairs can stack neatly on top of each other between the two helical sugar–phosphate backbones. Thus, the irregularity in the order of base pairs in DNA is embedded in an overall architecture that is relatively regular. This is in contrast to proteins (see Chapter 6) in which the irregular order of amino acids results in enormous diversity in protein structures.

The Double Helix Is Stabilized by Base Pairing and Base Stacking

The hydrogen bonds between complementary bases are a fundamental feature of the double helix, contributing to the thermodynamic stability of the helix and the specificity of base pairing. Hydrogen bonding might not, at first glance, appear to contribute importantly to the stability of DNA for the following reason: An organic molecule in aqueous solution has all of its hydrogen-bonding properties satisfied by water molecules that come on and off very rapidly. As a result, for every hydrogen bond that is made when a base pair forms, a hydrogen bond with water is broken that was there before the base pair formed. Thus, the net energetic contribution of hydrogen bonds to the stability of the double helix would appear to be modest. However, when polynucleotide strands are separate, water molecules are lined up on the bases. When strands come together in the double helix, the water molecules are displaced from the bases. This creates disorder and increases entropy, thereby stabilizing the double helix. Hydrogen bonds are not the only force that stabilizes the double helix.

A second important contribution comes from stacking interactions between the bases. The bases are flat, relatively water-insoluble molecules, and they tend to stack above each other roughly perpendicular to the direction of the helical axis. Electron cloud interactions ($\pi-\pi$) between bases in the helical stacks contribute significantly to the stability of the double helix. The stacked bases are attracted to each other by transient, induced dipoles between the electron clouds, a phenomenon known as van der Waals interactions. Base stacking also contributes to the stability of the double helix, a hydrophobic effect. Briefly put, water molecules interact more favorably with each other than with the "greasy" or hydrophobic surfaces of the bases. These hydrophobic surfaces are buried by base stacking in the double helix (as compared with the relative lack of stacking in single-stranded DNA), minimizing the exposure of base surfaces to water molecules and hence lowering the free energy of the double helix.

Hydrogen Bonding Is Important for the Specificity of Base Pairing

As we have seen, hydrogen bonding per se does not contribute importantly to the stability of DNA. It is, however, particularly important for the specificity of base pairing. Suppose we tried to pair an adenine with a cytosine. If so, we would have a hydrogen-bond acceptor (N1 of adenine) lying opposite a hydrogen-bond acceptor (N3 of cytosine) with no room to put a water molecule in between to satisfy the two acceptors (Fig. 4-7). Likewise, two hydrogen-bond donors, the NH_2 groups at C6 of adenine and C4 of cytosine, would lie opposite each other. Thus, an A:C base pair would be unstable because water would have to be stripped off the donor and acceptor groups without restoring the hydrogen bond formed within the base pair.

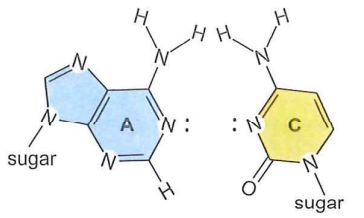


FIGURE 4-7 A:C incompatibility. The structure shows the inability of adenine to form the proper hydrogen bonds with cytosine. The base pair is therefore unstable.

Bases Can Flip Out from the Double Helix

As we have seen, the energetics of the double helix favor the pairing of each base on one polynucleotide strand with the complementary base on the other strand. Sometimes, however, individual bases can protrude from the double helix in a remarkable phenomenon known as **base flipping** (Fig. 4-8). As we shall see in Chapter 10, certain enzymes that methylate bases or remove damaged bases do so with the base in an extrahelical configuration in which it is flipped out from the double helix, enabling the base to sit in the catalytic cavity of the enzyme. Furthermore, enzymes involved in homologous recombination and DNA repair are believed to scan DNA for homology or lesions by flipping out one base after another. This is not energetically expensive because only one base is flipped out at a time. Clearly, DNA is more flexible than might be assumed at first glance.

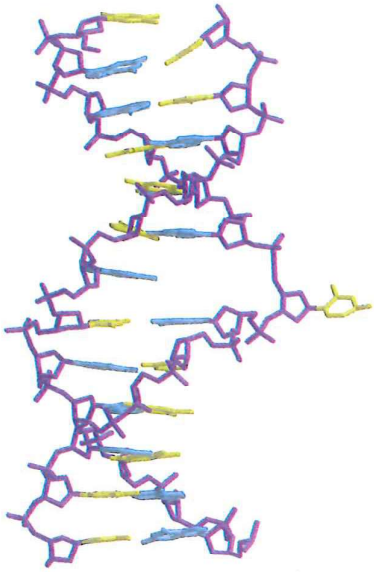


FIGURE 4-8 Base flipping. Structure of isolated DNA from the methylase structure, showing the flipped cytosine residue and the small distortions to the adjacent base pairs. (Klimasauskas S. et al. 1994. *Cell* **76:** 357.) Image prepared with BobScript, MolScript, and Raster3D.

DNA Is Usually a Right-Handed Double Helix

Applying the handedness rule from physics, we can see that each of the polynucleotide chains in the double helix is right-handed. In your mind's eye, hold your right hand up to the DNA molecule in Figure 4-9 with your thumb pointing up and along the long axis of the helix and your fingers following the grooves in the helix. Trace along one strand of the helix in the direction in which your thumb is pointing. Notice that you go around the helix in the same direction as your fingers are pointing. This does not work if you use your left hand. Try it!

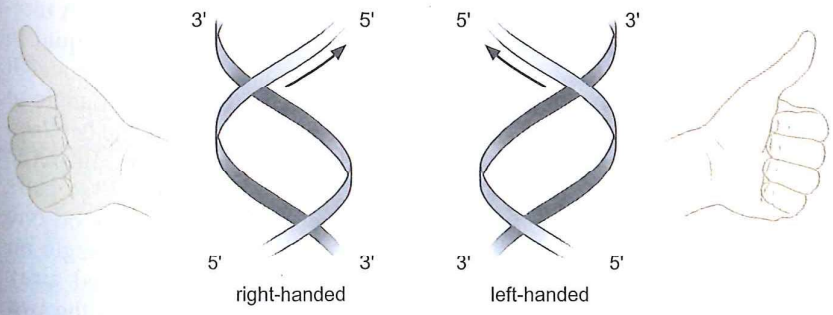
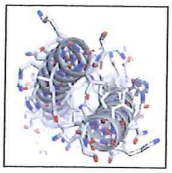


FIGURE 4-9 Left- and right-handed helices. The two polynucleotide chains in the double helix wrap around one another in a right-handed manner.

CHAPTER 6



The Structure of Proteins

PROTEINS ARE POLYMERS. THAT IS, THEY ARE molecules that contain many copies of a smaller building block, covalently linked. The building blocks of proteins are α-amino acids, of which there are 20 that occur regularly in the proteins of living organisms and that are specified by the genetic code. Some of these amino acids can undergo modification when already part of a protein, so the actual variety in proteins isolated from cells or tissues is somewhat greater.

OUTLINE

THE BASICS

Amino Acids

The α-carbon (C_α) of an amino acid has four substituents (Fig. 6-1), distinct from each other except in the case of the simplest amino acid, glycine. An amino group, a carboxyl group, and a proton are three of these substituents on all of the naturally occurring amino acids. The fourth, often symbolized by R and sometimes called the "R group," is the only distinguishing feature. The R-group is also called the "side chain," for reasons that will be clear in the next section. Because its four substituents are distinct (except for glycine), the C_α is a chiral center. Amino acids that occur in ordinary proteins all have the L-configuration at that center; D-amino acids are present in other kinds of molecules (including small protein-like polypeptides in microorganisms).

The properties of its R group determine the specific characteristics of an amino acid. The polarity of the group, which correlates with its solubility in water, is one critical property; size is another. It is useful to cluster the R groups of the 20 genetically encoded amino acids into the following categories: (1) neutral (i.e., uncharged) and nonpolar; (2) neutral and polar; and (3) charged (Fig. 6-2). The size (volume) of the side chain is of particular consequence for nonpolar amino acids because, as we shall see later, these side chains pack into the compact interior of a protein, and therefore the functional roles in proteins of glycine and alanine are quite different from those of phenylalanine and tryptophan. Note also that tryptophan, although largely nonpolar, has a hydrogen-bonding group that gives it a degree of polar character, and that tyrosine, although classified in Figure 6-2 as polar

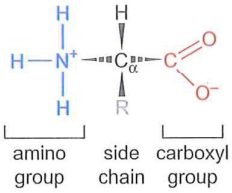


FIGURE 6-1 Structural features of an amino acid.

because of its OH group, is much less so than serine. In short, the boundaries between groups are less sharp than nomenclature might imply. The charged R groups are either negatively charged at neutral pH (aspartic acid and glutamic acid) or positively charged at neutral pH (lysine, arginine, and histidine). The pK_a of histidine is ~6.5, so even at neutral pH, histidine loses most of its charge. This property is particularly important for its role at the catalytic site of many enzymes.

The Peptide Bond

Peptide bonds are the covalent links between amino acids in a protein. A peptide bond forms by a condensation reaction, with elimination of a water molecule (Fig. 6-3a). It is a special case of an amide bond. Each amino acid can form two such bonds, so that successive links of the same kind can create a linear (i.e., unbranched) **polypeptide chain**. Because formation of each peptide bond includes elimination of a water, the components of the chain are known as **amino acid residues**, or sometimes just "residues" when "amino acid" is evident from context. The peptide bond has partial

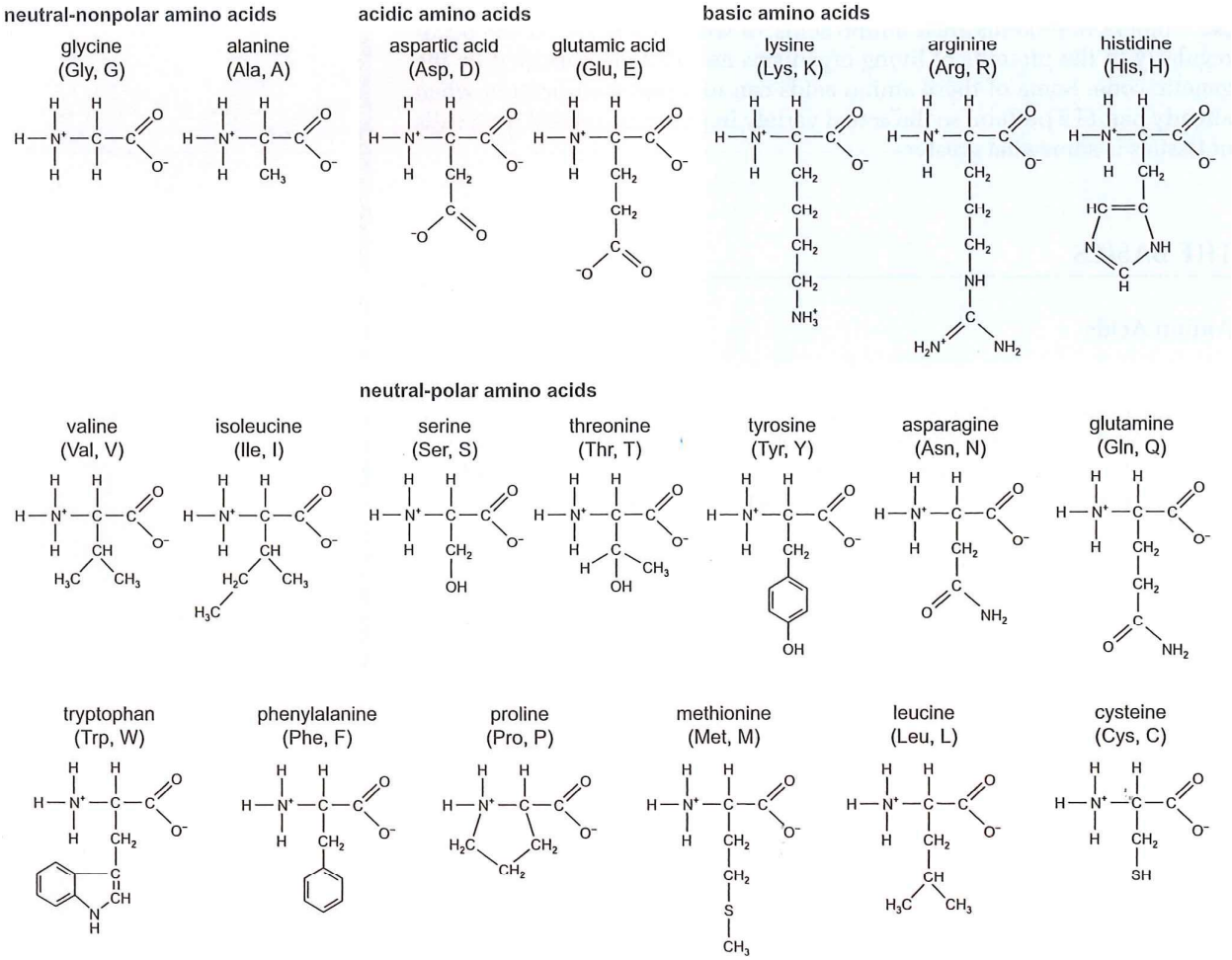


FIGURE 6-2 The 20 naturally occurring amino acids in proteins. Commonly used abbreviations for amino acids, including the single letter code, are shown in parentheses.

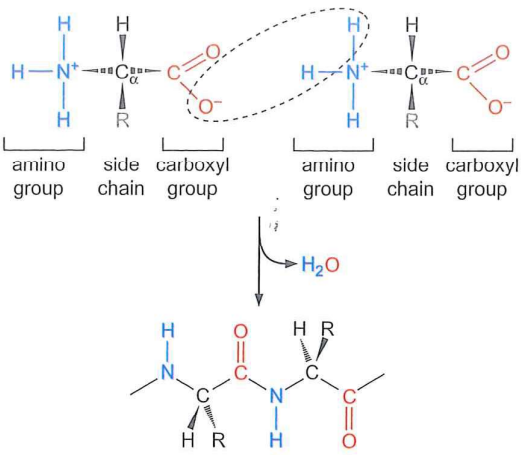


FIGURE 6-3 Peptide-bond formation.

double-bond character; the carbonyl and amide components are nearly coplanar and almost always in a *trans* configuration (Fig. 6-4).

Polypeptide Chains

The word **conformation** describes an arrangement of chemically bonded atoms in three dimensions, and we therefore speak of the conformation of a polypeptide chain, or more simply, of its "folded structure" or fold. If we follow the sequence of covalent linkages along a polypeptide chain,

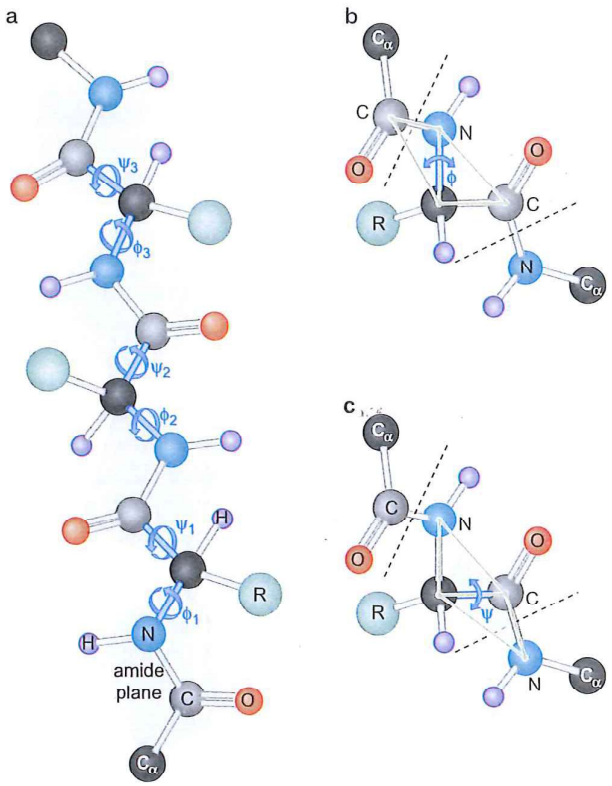


FIGURE 6-4 **The backbone torsion angles ϕ and ψ.** (a) This diagram shows the swivel points of the peptide backbone. (b) The ϕ torsion angle corresponds to the rotation about the N—C_α bond; here the conformation corresponds to a value of ϕ = 180°. (c) The ψ torsion angle corresponds to the rotation about the C_α—C bond; the conformation shown here represents ψ=0°. (Adapted, with permission, from Kuriyan J. et al. 2012. *The molecules of life.* © Garland Science/Taylor & Francis LLC; Branden C. and Tooze J. 1999. *An introduction to protein structure,* p. 8, Fig. 1.6. © Garland Science/Taylor & Francis LLC.)

there are three bonds per amino acid residue—one that joins the NH group to the C_α, another that joins the C_α to the carbonyl, and finally the peptide bond to the next residue in the chain. The first two are single bonds with relatively free torsional rotation about them (Fig. 6-4). But the peptide bond has very little rotational freedom, because of its partial double-bond character. The polypeptide chain conformation is therefore specified by the values of the torsion angles about the first two backbone bonds of each residue, plus the torsion angles for each single bond in each side chain. The two backbone angles are conventionally denoted φ and ψ. Many combinations of these two angles lead to atomic collisions, restricting the conformational freedom of a polypeptide chain to certain ranges of each angle (see Box 6-1, Ramachandran Plot).

Three Amino Acids with Special Conformational Properties

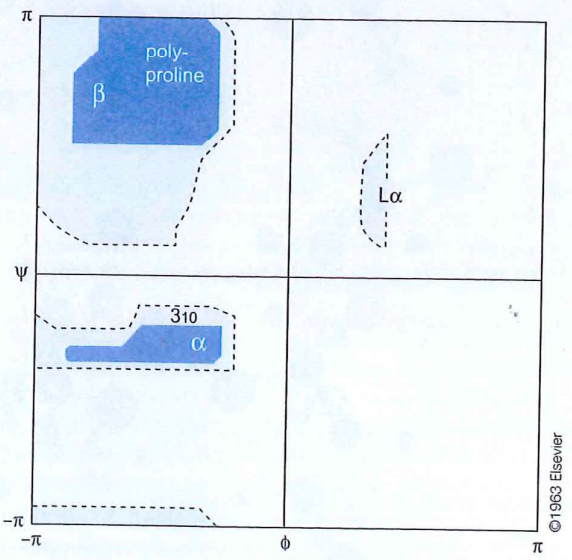
Glycine, proline, and cysteine have special properties. Because its R group is just a proton, glycine is not chiral, and it has much more conformational freedom than any other amino acid. Conversely, proline, in which the side chain has a covalent bond with N as well as C_α (making it, strictly speaking, an *imino* acid), has less conformational freedom than many other amino acids. Moreover, absence of the hydrogen-bonding potential of an NH group restricts its participation in secondary structures (see the section Protein Structure Can Be Described at Four Levels).

Cysteine, with a sulfhydryl group (—SH) on its side chain, is the one amino acid that is sensitive to oxidation–reduction under roughly physiological conditions. Two cysteines, correctly positioned across from each other in a folded protein, can form a **disulfide bond** by oxidation of the two —SH groups to S—S (Fig. 6-5). (The resulting pair of amino acids, linked by the S—S covalent bridge, is sometimes called *cystine*.) Proteins on the cell surface and proteins secreted into the extracellular space are exposed

ADVANCED CONCEPTS

BOX 6-1 Ramachandran Plot: Permitted Combinations of Main-Chain Torsion Angles ϕ and ψ

G.N. Ramachandran and coworkers (1963) studied all possible combinations of the torsion angles ϕ and ψ (shown in Fig. 6-4) and determined which combinations avoided atomic collisions ("allowed") and which combinations led to clashes ("forbidden"). The two-dimensional plot that shows the allowed and forbidden combinations is now known as a "Ramachandran plot" (Box 6-1 Fig. 1). The backbone conformations of regular secondary structures have the ϕ and ψ values indicated: right-handed α helix; β strand; polyproline helix (a threefold screw structure adopted preferentially by continuous stretches of proline); 3_{10} helix (a helix with 3.3 residues per turn, closely related to the α helix, which has 3.6 residues per turn); and left-handed α helix, Lα (permitted for glycine only, because it has no side chain).



BOX 6-1 FIGURE 1 The Ramachandran plot. The "allowed" areas are shown shaded in blue. (Modified, with permission, from Ramachandran G.N., et al. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7: 95–99.)

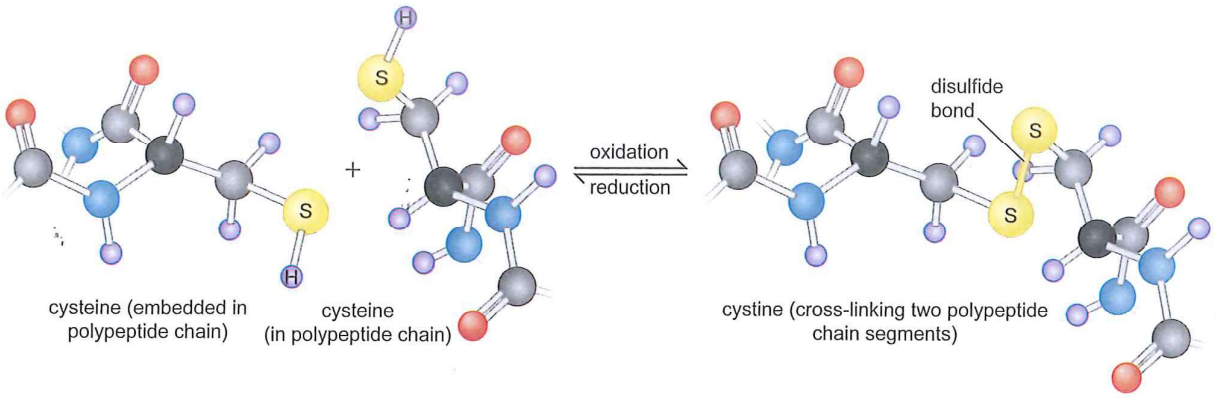


FIGURE 6-5 Formation of the disulfide bond. (Adapted, with permission, from Kuriyan J. et al. 2012. *The molecules of life.* © Garland Science/Taylor & Francis LLC.)

to an environment with a redox potential that favors disulfide formation; most such proteins have disulfide bonds and no unoxidized cysteines. Living cells maintain a more reducing internal environment, and intracellular proteins very rarely have disulfide bonds.

Disulfide bonds enhance the stability of a folded protein by adding covalent cross-links. They are particularly critical for stabilizing small, secreted proteins, such as some hormones, and for reinforcing the extracellular domains of membrane proteins, which face a much less controlled environment than do proteins that remain in the cell interior.

IMPORTANCE OF WATER

All molecular phenomena in living systems depend on their aqueous environment. The importance of the distinction between polar and nonpolar amino acid side chains comes from their properties with respect to water as a solvent. Compare the side chains of aspartic acid and phenylalanine, which resemble acetic acid and toluene, respectively, linked to the peptide main chain. Acetic acid is very soluble in water; toluene is very insoluble. An aspartic acid side chain is therefore called **hydrophilic**, and a phenylalanyl side chain **hydrophobic**. Even hydrophilic side chains can have hydrophobic parts (e.g., the three methylene groups of a lysyl side chain).

Water is an extensively hydrogen-bonded liquid (Fig. 6-6). Each water molecule can donate two hydrogen bonds and accept two hydrogen bonds. The way in which a solute affects the hydrogen bonding of the surrounding water determines its hydrophilic or hydrophobic character. Hydrophobic molecules perturb the network of hydrogen bonds; hydrophilic molecules participate in it. Thus, it is more favorable for hydrophobic molecules to remain adjacent to each other (insolubility) than to disperse into an aqueous medium (solubility). The hydrophobic character of many amino acid side chains makes it favorable for them to cluster away from water, and the hydrophilic character of others allows them to project into water. The sequence of amino acids in a real protein has evolved so that these tendencies cause the polypeptide chain to fold up, sequestering residues of the former kind and exposing residues of the latter. Many of the huge number of possible sequences for an average-sized polypeptide chain cannot fold up in this way—if

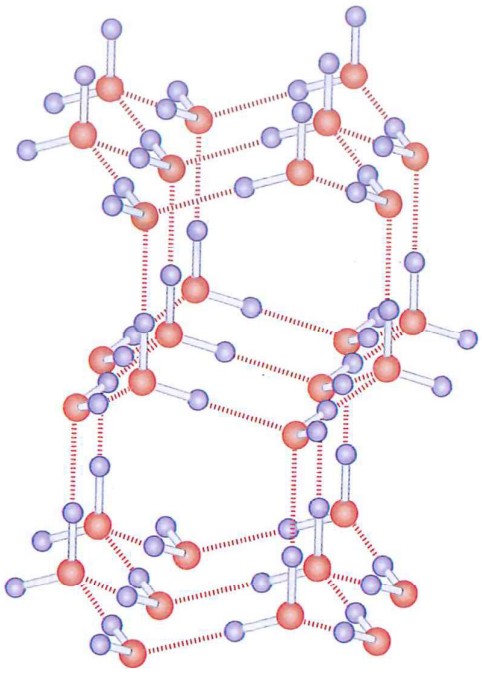


FIGURE 6-6 Water: the hydrogen-bonded structure of ice. In ice, each water molecule donates two hydrogen bonds (from its two protons [lavender]) to lone-pair electrons on an oxygen of its neighbor (red) and accepts two hydrogen bonds at lone pairs of its own oxygen. The hydrogen bonds are shown as dashed red lines. When ice melts, the network of hydrogen bonds fluctuates and breaks apart transiently, but individual water molecules retain (on average) most of the four hydrogen bonds with their neighbors. Thus, the structure of liquid water resembles a fluctuating and distorted version of the ice lattice shown here.

made, they either remain as randomly fluctuating, extended chains in solution (sometimes called **random coils**) or else they aggregate because the hydrophobic groups on one polypeptide chain cluster together with hydrophobic groups on other chains.

PROTEIN STRUCTURE CAN BE DESCRIBED AT FOUR LEVELS

In analyzing and describing the structure of proteins, it is useful to distinguish four levels of organization (Fig. 6-7). The first level, the **primary structure** of a protein, is simply the sequence of amino acid residues in the polypeptide chain. As we have seen, the genetic code specifies the primary structure of a protein directly. The primary structure is thus just a one-dimensional (1D) string, specifying a pattern of chemical bonds; the remaining three levels depend on a protein's three-dimensional (3D) characteristics.

The **secondary structure** of a protein refers to the *local* conformation of its polypeptide chain—the 3D arrangement of a short stretch of amino acid residues. There are two very regular secondary structures found frequently in naturally occurring proteins, because these two local conformations are particularly stable ones for a chain of L-amino acids (Box 6-1). One of these is called the **α helix** (Fig. 6-8a). The polypeptide backbone spirals in a right-handed sense around a helical axis, so that hydrogen bonds form between the main-chain carbonyl group of one residue and the main-chain amide group of a residue four positions further along in the chain. The other regular conformation is called a **β strand** (Fig. 6-8b). It is an extended conformation, in which the side chains project alternately to either side of the backbone, and the amide and carbonyl groups project laterally, also alternating. The backbone is not quite fully stretched, so that the strand has a slightly zigzag or pleated character. In folded proteins, β strands form

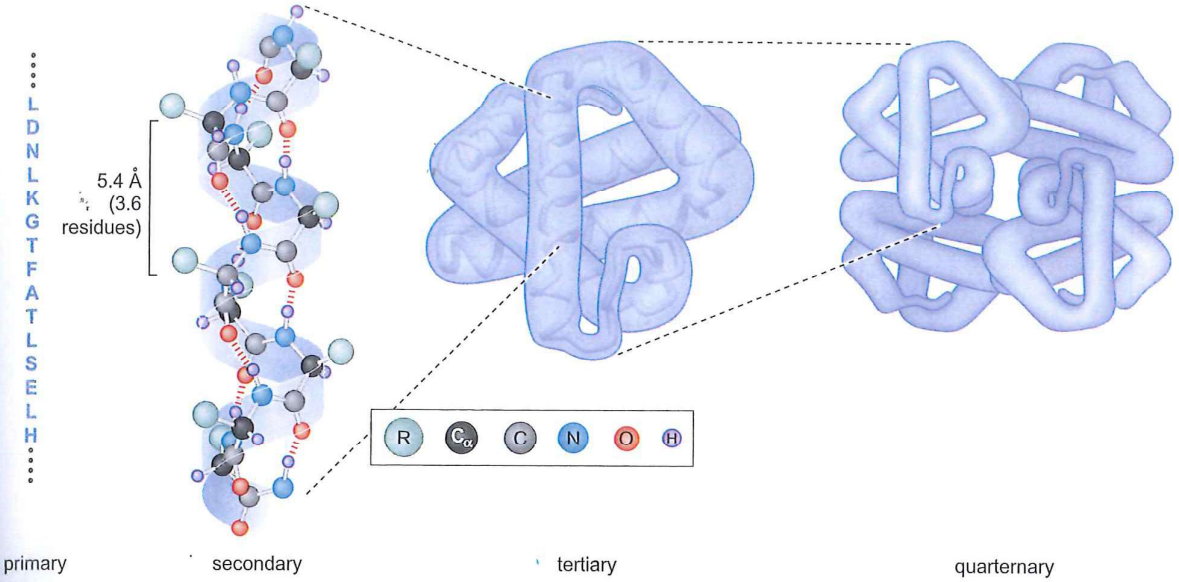


FIGURE 6-7 Levels of protein structure, illustrated by hemoglobin. "Primary structure" refers to the sequence of amino acids in the polypeptide chain. The primary structure of a segment of a hemoglobin subunit is shown in single-letter code. "Secondary structure" refers to regular local structures, with repeated backbone hydrogen bonds. Shown here is a part of one of the long α helices from the hemoglobin subunit. "Tertiary structure" refers to the folded structure of an entire polypeptide chain (or of a single domain of a multidomain protein). The drawing shows one of the four hemoglobin protein subunits. Dashed lines demarcate the segment of α helix corresponding to the primary and secondary structures shown to the left. "Quaternary structure" refers to the arrangement of multiple protein subunits in a larger complex. Hemoglobin is a tetramer of two "α chains" and two "β chains," but the two kinds of chain have very similar tertiary structures, as can be seen in the drawing. (Modified from an illustration by Irving Geis. Rights owned by the Howard Hughes Medical Institute.)

sheets joined by main-chain hydrogen bonds. Either parallel or antiparallel hydrogen-bonding patterns are possible, sometimes called parallel or antiparallel β-pleated sheets, respectively. In real proteins, various mixed sheets are often found—rather than either strictly alternating strand directions or strictly unidirectional ones.

The **tertiary structure** of a protein refers to the usually compact, three-dimensionally folded arrangement that the polypeptide chain adopts under physiological conditions. Segments of the chain may be α helices or β strands; the rest have less regular conformations (e.g., turns or loops between secondary-structure elements that allow these elements to pack tightly against each other). We will outline ways to describe and classify possible tertiary structures in a subsequent section. Usually, the stabilities of the secondary and tertiary structures of a polypeptide chain depend on each other.

Many proteins are composed of more than one polypeptide chain: **quaternary structure** refers to the way individual, folded chains associate with each other. We can distinguish cases in which there are a defined number of copies of a single type of polypeptide chain (generally called a "subunit" in this context, or a "protomer") and cases in which there are defined numbers of each of more than one type of subunit. In simple cases, the way in which the subunits associate does not change how the individual

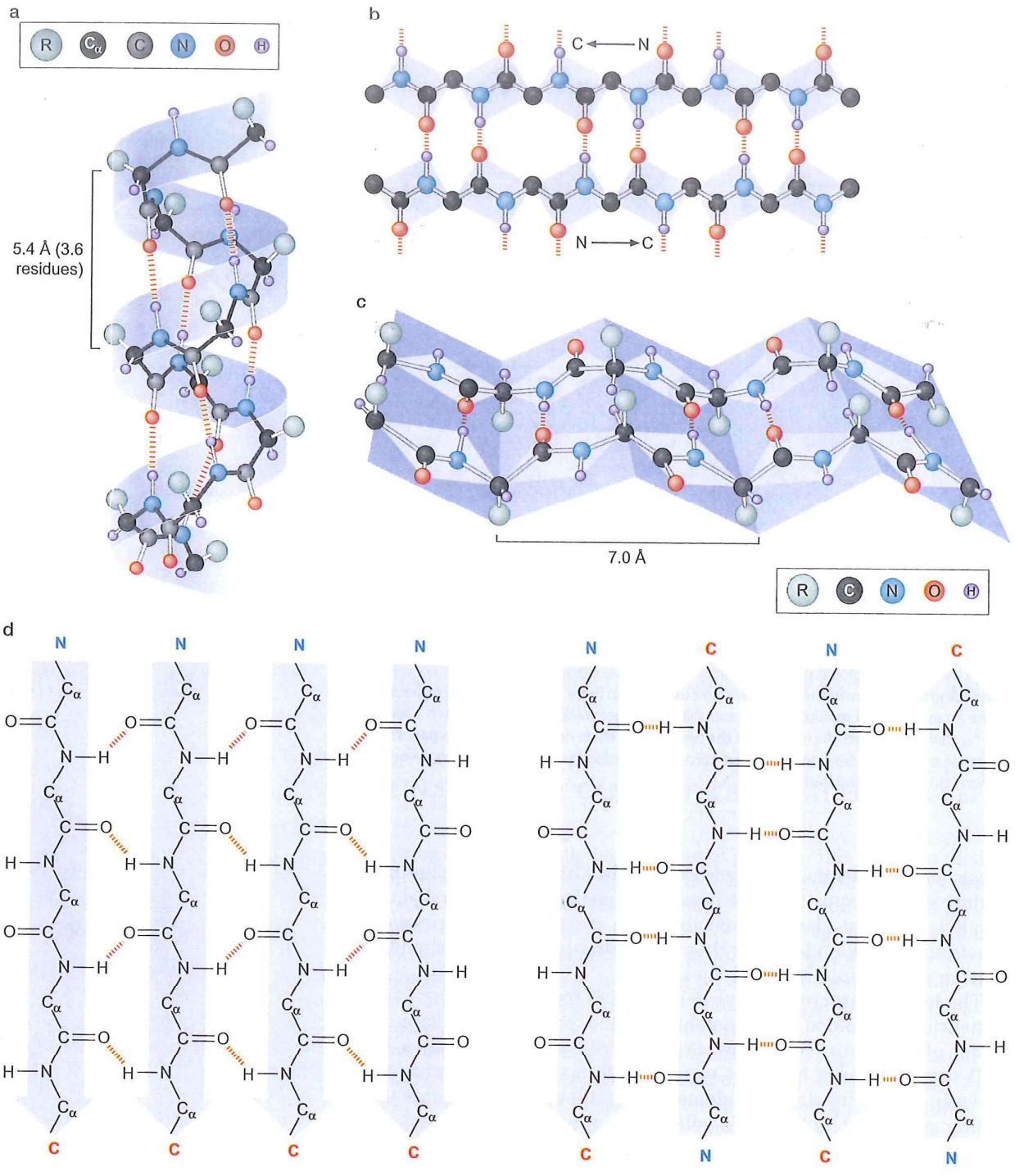


FIGURE 6-8 Protein secondary structures. (a) An α helix. Hydrogen bonds are represented by the series of broken red lines. (b) β sheets. Hydrogen bonds are represented by the series of broken red lines. On the top, a β sheet is shown from above. On the bottom, a β sheet is shown from the side. (c) A parallel β sheet, showing the hydrogen-bonding pattern, in which the chains run in the same amino-to-carboxyl direction. (d) An antiparallel β sheet, showing the hydrogen-bonding pattern, in which the chains run in opposite directions. (a, Modified from illustration by Irving Geis. b,c, Illustrations by Irving Geis. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission. d, Adapted, with permission, from Branden C. and Tooze J. 1999. *Introduction to protein structure*, 2nd ed., p. 19, Fig 2.6a and p. 18, Fig 2.5b. © Garland Science/Taylor & Francis LLC.)

polypeptides fold. Often, however, the tertiary or even secondary structures of the components of a protein **oligomer** (i.e., a protein composed of a small number of subunits) depend on their association with each other. In other words, the individual subunits acquire secondary or tertiary structure only as they also acquire quaternary structure. One common example is the α-helical coiled-coil: two (or sometimes three or even four) polypeptide chains, either identical or different, adopt α-helical conformations and wrap very gently around each other (Fig. 6-9a). The individual chains are not, in general, stable as α helices on their own—if the oligomeric interaction is lost, the separated helices unravel into disordered polypeptide chains.

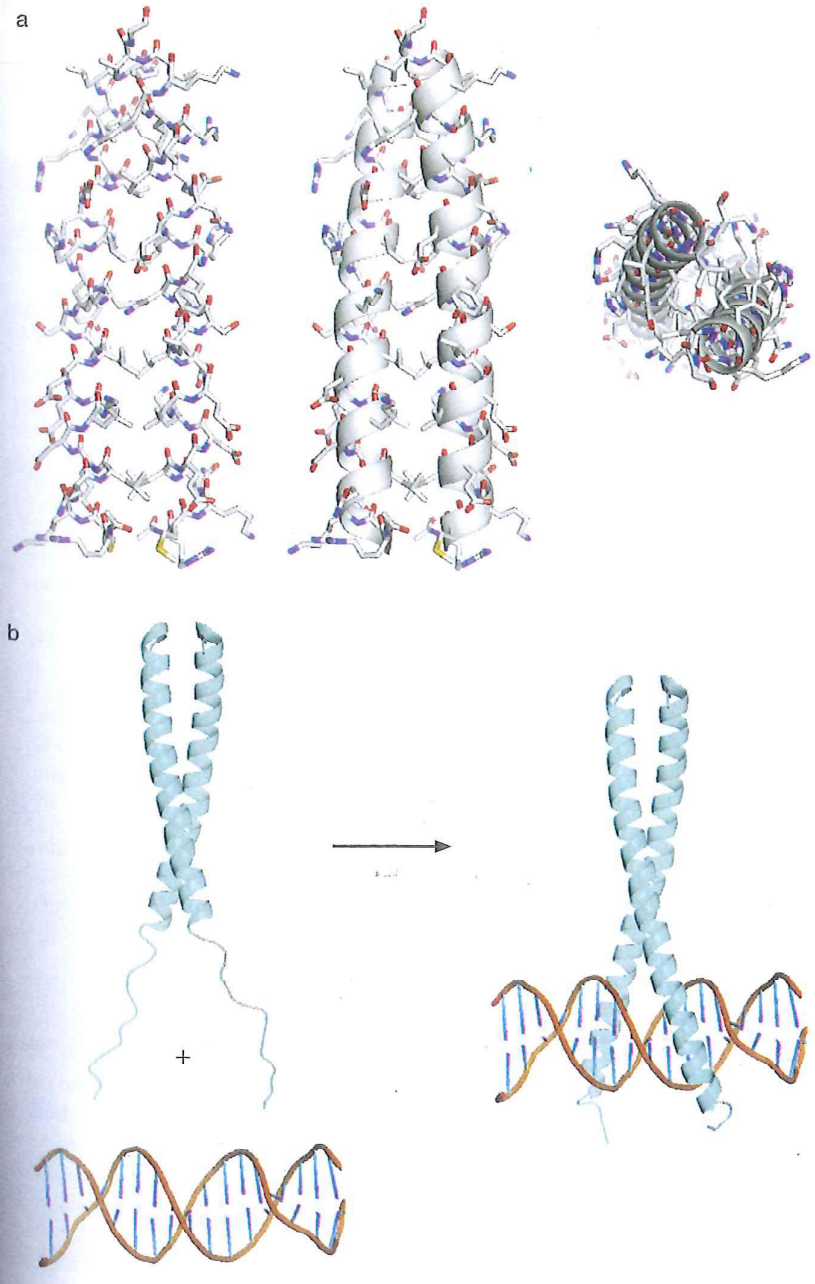


FIGURE 6-9 The yeast transcription factor GCN4. (a) Three views of the structure of the GNC4 coiled-coil. (Left) Representation that shows chemical bonds as sticks and atoms as junction, with carbon in gray, oxygen in red, and nitrogen in blue. The carboxyl termini of the two identical polypeptide chains are at the top. Note the ladder of hydrophobic side chains (mostly gray) at the interface between the two helices. (Center) Representation with polypeptide backbone as an idealized ribbon and side chains as sticks. Note that the two chains coil very gently around each other. (Right) The same representation as in the center, but viewed end-on from the top. (b) Structure of the GCN4 complex with DNA, illustrating the disorder-to-order transition of the so-called "basic region"—the segment amino-terminal to the coiled-coil, which, upon binding, folds into an α helix in the major groove of DNA. Images prepared with PyMOL (Schrödinger, LLC).

PROTEIN DOMAINS

Polypeptide Chains Typically Fold into One or More Domains

Folding of a polypeptide chain creates an "inside" and an "outside" and thus generates **buried** and **exposed** amino acid side chains, respectively. If the polypeptide chain is too short, there are no conformations that bury enough hydrophobic groups to stabilize a folded structure. If the chain is too long, the complexity of the folding process is likely to generate errors. As a result of these restrictions, most stably folded conformations include between about 50 and 300 amino acid residues. Longer polypeptide chains generally fold into discrete modules known as **domains** (see Box 6-2, Glossary of Terms); each domain generally falls within the 50- to 300-residue range just mentioned. The structures of individual domains of such a protein are similar to the structures of smaller, single-domain proteins (Fig. 6-10a).

Each of the two or more domains of a folded polypeptide chain sometimes contains a continuous sequence of amino acid residues. Often,

ADVANCED CONCEPTS

BOX 6-2 Glossary of Terms

Primary structure: Amino acid sequence of a polypeptide chain.

Secondary structure: Elements of regular polypeptide-chain structure with main-chain hydrogen bonds satisfied. The secondary structures that occur frequently in proteins are the α helix and the parallel and antiparallel β sheets.

Tertiary structure: The folded, three-dimensional conformation of a polypeptide chain.

Quaternary structure: Multi-subunit organization of an oligomeric protein or protein assembly.

Domain: A part of a polypeptide chain with a folded structure that does not depend for its stability on any of the remaining parts of the protein.

Motif (sequence): A short amino acid sequence with characteristic properties, often those suitable for association with a specific kind of domain on another protein. (Note that the term "domain" is sometimes incorrectly applied to such sequence motifs.)

Motif (structural): A domain substructure that occurs in many different proteins, often having some characteristic amino acid sequences properties (e.g., the helix-turn-helix motif in many DNA-recognition domains).

Topology (or fold): The structure of most protein domains can be represented schematically by the connectivity in three dimensions of their constituent secondary-structural elements and the packing of those elements against each other. Jane Richardson introduced "ribbon diagrams," such as those in many of the figures in this chapter, as convenient ways to visualize the fold of a domain (see the caption to Fig. 6-10). Not all folds are found in naturally occurring proteins (e.g., knotted folds are not found), and some folds are more common than others.

Homologous domains (or proteins): Domains (or proteins) that derive from a common ancestor. They necessarily have the same fold, and they often (but not always) have recognizably similar amino acid sequences.

Homology modeling: Modeling the structure of a domain based on that of a homologous domain.

Ectodomain: The part of a single-pass membrane protein that lies on the exterior side of the cell membrane.

Glycosylation: Addition of a chain, sometimes branched, of one or more sugars (glycans) to a protein side chain. The glycans can be N-linked (attached to the side-chain amide of asparagine) or O-linked (attached to the side-chain hydroxyl of serine or threonine).

Denaturation: Unfolding a protein or a domain of a protein, either by elevated temperature or by agents such as urea, guanidinium hydrochloride, or strong detergent ("denaturants").

Chaperone: A protein that increases the probability of native folding of another protein, usually by preventing aggregation or by unfolding a misfolded polypeptide chain so that it can "try again" to fold correctly.

Active site (or catalytic site): The site on an enzyme that binds the substrate(s), often in a configuration resembling the transition state of the reaction catalyzed.

Allosteric regulation: Control of affinity or of the rate of an enzymatic reaction by a ligand that binds at a site distinct from that of the substrate(s). The mechanism of allosteric regulation often involves a change in quarternary structure—that is, a reorientation or repositioning of subunits with respect to each other.

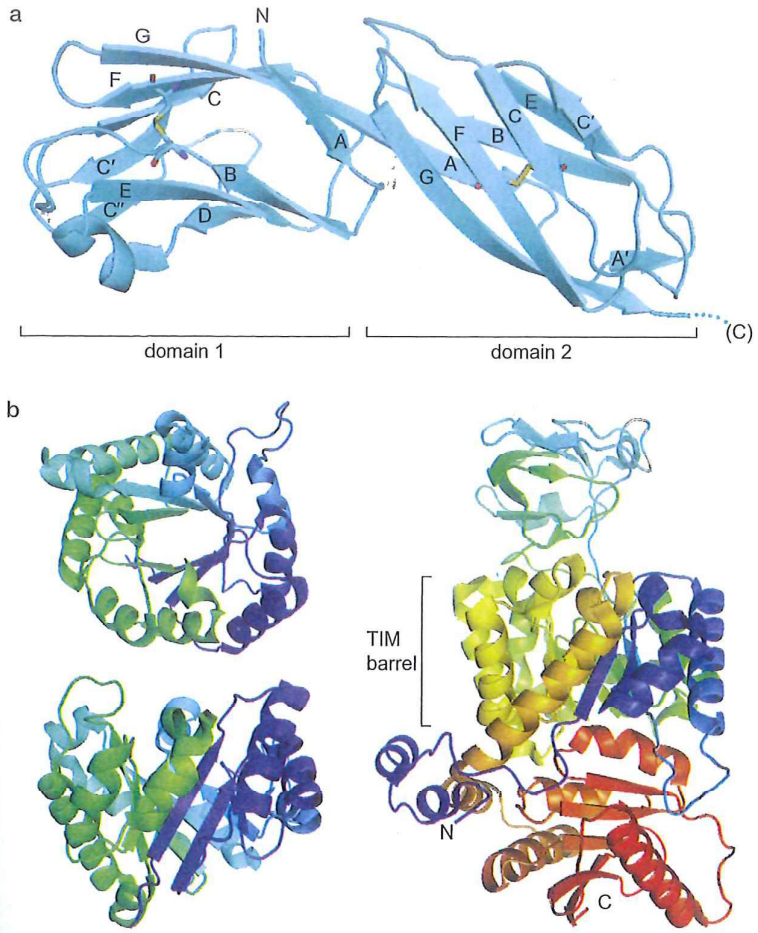


FIGURE 6-10 Protein domains. Polypeptide chains are shown here schematically as "ribbons"—a representation, introduced by Jane Richardson, that emphasizes the role of secondary structural elements in the folded conformation of a domain: α helices are curled ribbons; β strands are gently curved arrows, pointing toward the carboxyl terminus. Intervening loops between secondary structural elements are shown as "worms." (a) Two of the four domains of the protein CD4, which is found on the surface of certain T-cells and macrophages. Each of these domains is a β-sandwich with an immunoglobulin fold (see Box 6-3); the β strands of each domain are designated by letters in the order in which they follow in the polypeptide chain. Each domain has a single disulfide bond, shown in a stick representation with bonds to sulfur atoms in yellow. (b) Two enzymes: triose phosphate isomerase (TIM; left) and pyruvate kinase (PK; right). The figure shows one monomer of the TIM dimer. The TIM subunit is the prototype of a domain called a "TIM barrel"—a short cylinder in which the eight strands that form the inner barrel alternate with helices that cover the periphery. The two views are along the barrel axis (top) and from the side (bottom). The colors run from dark blue at the amino terminus to green at the carboxyl terminus. PK folds into three domains. The central domain is a TIM barrel (compare with the side view of TIM). The "rainbow" colors run from dark blue at the amino terminus to red at the carboxyl terminus. The light blue domain at the top folds from residues that follow strand 3 of the TIM barrel. The orange-red domain at the bottom contains residues carboxy-terminal to the last TIM-barrel helix. The comparison of TIM and PK shows that a domain found as an isolated unit in one protein can join with additional domains in another protein. Moreover, one or more of those additional domains can fold from a polypeptide chain "inserted" between secondary structural elements of the principal domain. Images prepared with PyMOL (Schrödinger, LLC).

however, at least one of the domains folds from two (or more) noncontiguous segment(s), and the intervening part of the chain forms a distinct domain (Fig. 6-10b). The intervening domain then looks like an insertion into the domain that folds from the flanking segments.

Basic Lessons from the Study of Protein Structures

The large number of domain structures that have been determined experimentally allows us to draw the following conclusions. First, most hydrophobic side chains are, indeed, buried, and most polar side chains are exposed. Second, if a functional group that can donate or accept a hydrogen bond is buried, it almost always has a hydrogen-bonding partner. The reason for this property is easy to grasp, when we recall that were the polar group exposed on the domain surface, it would make a similar hydrogen bond with water (which can both donate and accept). If the hydrogen bond were missing in the folded conformation, a favorable energetic contribution would have been lost when water was stripped away from that group as the polypeptide chain folded. Even hydrophobic amino acid residues have two hydrogen-bonding groups, an NH and a CO, in their peptide backbone. These hydrogen bonds are also satisfied in folded structures, in considerable part by formation of secondary structures. Both α helices and β sheets satisfy the main-chain hydrogen bonds of all of the residues within them.

Fulfilling main-chain hydrogen bonding is probably an important reason for the prevalence of regular secondary structures, even within compactly folded protein domains. As a result, it is useful to classify the observed domain structures according to the kinds of secondary structures present within them. We observe that even a relatively short polypeptide chain could, in principle, have an astronomically large number of folded conformations. Only a restricted number of these appear in the large catalog of known 3D protein structures. These not only have a substantial proportion of their amino acid residues in α helices or β sheets (rather than in irregular loops, which would be much less likely to allow main-chain hydrogen bonding), but also have a relatively simple 3D folding pattern. For example, the Ig domains in CD4 (Fig. 6-10a) are composed of two β sheets—a **β sandwich**—with four or five strands in one sheet and four in the other. Although there would be many ways for the polypeptide chain to pass from one of these eight or nine strands to the next, the observed pattern is one in which the chain makes either a sharp turn within one sheet, linking two adjacent strands, or passes across the top or bottom of the domain to the other sheet. One very important property of all known domain structures is that the chain does not form a knot—that is, if you imagined pulling on its ends, the whole thing would open into a straight line.

Classes of Protein Domains

Classifications of protein domains allow simple, summary descriptions. One widely used classification hierarchy, embodied in a database called CATH, starts with separation of proteins into classes according to their principal secondary structures (mostly α helix, mostly β strand, a mixture of the two, and a fourth class for the usually small domains with very little secondary structure). The most important levels in the classification hierarchy are **fold** (also called **topology**) and **homology**. The fold class takes into account not only the secondary structures, but also how the chain passes from one helix or strand to another. The diagrams in Figure 6-11 illustrate this point. A group of homologous proteins are ones with sequence similarities great enough to assume that they have a common evolutionary origin. An unanswered question concerns the likelihood that all domains of a given fold class have a common origin—for very complex domains, a common origin seems intuitively reasonable.

FIGURE 6-11 Examples of the three principal classes of fold. (Left) An all α-helical protein (myoglobin). (Center) A heterodimer of two all β-strand domains (the variable region of an immunoglobulin—see Box 6-3). (Right) A mixed α- and β-domain (the small GTPase, Ras). Colors in each domain run from dark blue at the amino terminus to red at the carboxyl terminus. Images prepared with PyMOL (Schrödinger, LLC).

Linkers and Hinges

The links between two domains of a folded protein can be very short, allowing a tight and rigid interface between them, or quite long, allowing considerable flexibility. Some proteins have extremely long flexible linkers, because their function within a cell requires that the domains at either end interact over long and variable distances. The amino acid sequences of long linkers generally lack large hydrophobic groups, which their extendable, flexible conformation cannot sequester from water, and have other simplified features.

We summarize our discussion of domains and four levels of protein structure with the illustration of an antibody (immunoglobulin) molecule, described in Box 6-3, The Antibody Molecule as an Illustration of Protein Domains.

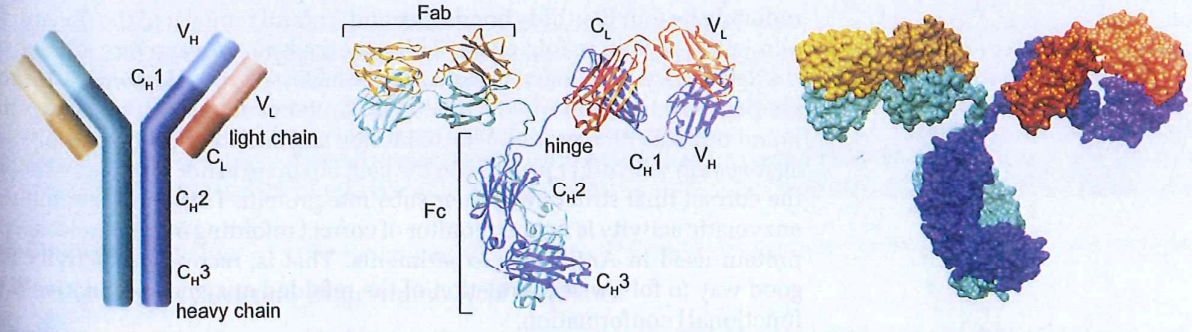
Post-Translational Modifications

Various modifications of amino acid side chains, introduced following emergence of the polypeptide chain from a ribosome, can modulate the

ADVANCED CONCEPTS

Box 6-3 The Antibody Molecule as an Illustration of Protein Domains

Circulating antibodies are immunoglobulin G (IgG) molecules, which contain two identical heavy chains and two identical light chains. The light chains have a variable domain (V_L) and a constant domain (C_L); the heavy chains, a variable domain (V_H) and three constant domains (C_H1, C_H2, and C_H3). Thus, there are a total of 12 independent domains. V_H and V_L are "variable," because there is a large combinatorial library of genes that encode them and because somatic mutations occur in the selected gene during the course of an immune response. The variable domains determine specific affinity for antigen. The C_H1-3 and C_L domains are "constant," because a much smaller number of these domains are linked with one of the many variable domains during maturation of an antibody-producing cell and because they are not prone to somatic mutation. The domains pair in the assembled heterotetramer as shown in Box 6-3 Figure 1: V_H with V_L and C_H1 with C_L, forming an Fab ("antigen-binding") fragment; C_H2 and C_H3 of one heavy chain with C_H2 and C_H3 of the other, respectively, forming an Fc fragment. Controlled proteolytic attack selectively cleaves the hinge, allowing preparation of both the Fab and Fc moieties. Each of the domains has a similar, "Ig-domain" fold, illustrated also in Figure 6-11 as an example of an all-β domain. The short link ("elbow") between variable and constant domains has restricted flexibility. The much longer link (hinge) between C_H2 and C_H3 of each of the heavy chains has much greater flexibility, allowing the antigen binding sites (called "complementarity determining regions") at the tip of each Fab to orient and reorient according to the relative positions of their cognate sites on the antigen.



BOX 6-3 FIGURE 1 **Three different representations of IgG.** The left panel is a schematic diagram of the "Y-like" pattern of association among the four chains of IgG. In the center, a "ribbon" diagram emphasizes the IgG secondary structure. And, in the right panel, a surface rendering shows that side chains of folded proteins pack efficiently to fill the hydrophobic interior of the protein. Images prepared with PyMOL (Schrödinger, LLC) and UCSF Chimera.

structure and function of a protein. One of the most important is glycosylation—addition of one or more sugars ("glycans") to an asparagine side chain or to a serine or threonine side chain. This modification generally takes place in the endoplasmic reticulum of eukaryotic cells, and it is therefore a nearly universal characteristic of the ectodomains of cell-surface proteins and of secreted proteins. Proteins bearing glycans are called glycoproteins. Enzymes that transfer glycans to asparagine side chains recognize a short sequence motif, Asn-X-Ser/Thr, where X can be any amino acid residue.

Phosphorylation of serine, threonine, tyrosine, or histidine side chains is another widespread modification, critical for intracellular regulation. Phosphorylation of the first three residues occurs largely in eukaryotic cells; phosphorylation of the last is more common in prokaryotes.

FROM AMINO-ACID SEQUENCE TO THREE-DIMENSIONAL STRUCTURE

Protein Folding

The amino acid sequence of a domain determines its stable, folded structure. This generalization is an important part of the central dogma of molecular biology, because it means that the nucleotide sequence of a translated gene specifies not only the amino acid sequence of the protein it encodes, but also the 3D structure and function of that protein. A classic experiment concerning refolding of an unfolded protein in the laboratory first established this point (see Box 6-4, Three-Dimensional Structure of a Protein Is Specified by Its Amino Acid Sequence [Anfinsen Experiment]). It also showed that a polypeptide chain can fold correctly without any additional cellular machinery.

The Anfinsen refolding experiment relies on several key points. First, a protein purified from cells or tissues can be unfolded in solution into a random coil. This unfolding is often called **denaturation**, and it is generally accomplished by exposing the protein to high concentrations of certain solutes called **denaturants** (e.g., urea or guanidinium hydrochloride). If the protein is an enzyme, it loses its catalytic activity. If it has a specific binding property (e.g., recognition of a site on DNA), it loses that specificity. That is, almost all of the functional properties of proteins depend on their folded structures. In the case of the protein that Anfinsen and colleagues used in the experiments described in Box 6-4, complete unfolding also required reducing its four disulfide bonds. Second, careful removal of the denaturant allows the protein to fold again. This process is not always very efficient in the laboratory, for many reasons. Cells have enzymes known as **folding chaperones** that can unfold a misfolded protein and allow it to "try again." Some of these chaperones also sequester the unfolded protein to prevent aggregation with other proteins in the cell, but they do not in any way specify the correct final structure of their substrate protein. Third, measurement of enzymatic activity is a good monitor of correct refolding of ribonuclease, the protein used in Anfinsen's experiments. That is, recovery of activity is a good way to follow accumulation of the refolded enzyme in its **native** (i.e., functional) conformation.

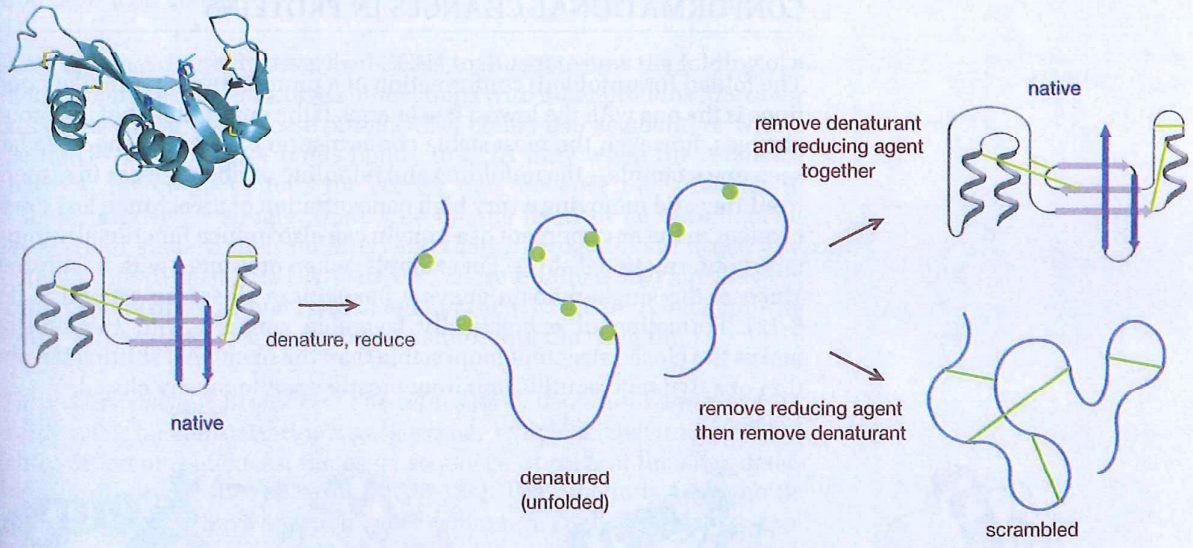
Another conclusion from experiments such as Anfinsen's is that the native structure of a protein is the most stable conformation that its polypeptide chain can adopt, given the particular sequence of amino acids in that chain. In physical chemistry, one would say that the native structure has the lowest free energy of any possible conformation.

KEY EXPERIMENTS

Box 6-4 Three-Dimensional Structure of a Protein Is Specified by Its Amino Acid Sequence (Anfinsen Experiment)

In the early 1960s, Christian Anfinsen and coworkers carried out a classic series of experiments, showing that the amino acid sequence of a protein is sufficient to determine its correctly folded structure and that no external folding "machinery" is necessary. This conclusion is fundamental to our understanding of how the nucleotide sequence of a gene ultimately encodes the information needed to specify protein function.

Ribonuclease A is an enzyme that cleaves the phosphodiester backbone of RNA. The enzyme is active when folded into its native conformation but is inactive when unfolded by a denaturant, such as urea or guanidinium hydrochloride at concentrations of 2–5 M. The 124-residue protein has eight cysteines, which form four disulfide bonds (see Box 6-4 Fig. 1). These disulfides can be reduced to sulfhydryls by adding a reducing agent, such as β-mercaptoethanol. Anfinsen and coworkers found that if they unfolded ribonuclease A in the presence of β-mercaptoethanol and then removed both the denaturant and the reducing agent by dialysis, they could recover a high level of enzymatic activity. Assuming that only a properly folded enzyme can catalyze hydrolysis of phosphodiester bonds, recovery of activity showed that the polypeptide chain contains all the information needed to dictate the folded structure. When Anfinsen et al. *first* removed the β-mercaptoethanol, allowing disulfide bonds to re-form, and *then* dialyzed away the denaturant, they failed to detect activity. Eight cysteines can pair in 105 distinct ways. Formation of disulfide bonds in the presence of denaturant might be expected to allow cysteines to pair randomly, leading primarily to forms with scrambled disulfide bonds rather than to the unique pairing found in the native protein. Thus, oxidation of the unfolded ribonuclease A should yield less than 1% of the activity recovered by oxidizing and refolding at the same time. This expectation agrees with the observations, strengthening the fundamental conclusion that only when the native, noncovalent contacts can form will each cysteine find its proper partner.



BOX 6-4 FIGURE 1 The Anfinsen experiment. Ribonuclease A is represented on the upper left as a ribbon diagram showing the tertiary structure of the enzyme (here the disulfide bonds are shown in yellow). The corresponding schematic below depicts the various secondary structure elements and the locations of the four disulfide bonds. Reducing the disulfides in the presence of a denaturant unfolds the polypeptide chain. Removal of the reducing agent in the presence and in the absence of denaturant leads to two quite different outcomes, as described in the text. In the schematic, the disulfide bonds are represented as green lines and the cysteines as green circles.

Predicting Protein Structure from Amino Acid Sequence

In principle, if amino acid sequence determines the folded structure of a protein, it should be possible to devise a computational method for doing the same thing. But in practice, the computational task is daunting. The following comments illustrate why. First, we might imagine that a computer could calculate the stability (free energy) of every possible conformation of the

polypeptide chain and then pick the one that corresponds to a minimum. It is indeed possible to compute the various forces between atoms in a protein that determine its stability—hydrogen bonds, hydrophobic contacts, and so on. But consider a small protein of 100 amino acid residues and imagine that each residue can have only three configurations (e.g., helix, strand, and other). Then the number of possible conformations is roughly 3^{100} or 10^{47}, an astronomical figure, ruling out this strategy. Second, we might try to simulate the process of protein folding, by some sort of dynamic calculation. Efforts to do so are starting to work, for small proteins and with advanced computational resources; the answers are good approximations for some purposes, but not yet adequate for understanding all aspects of function. Such an approach is not likely in the near future to be a practical way to predict structures of complex proteins just from their amino acid sequences. Third, if we already know the structure of a similar, homologous protein, we might consider starting with it as a first approximation and computationally changing the amino acid residues to match the new protein we wish to understand. Computations of this kind, known as **homology modeling**, have become relatively practical. Their reliability obviously depends on the similarity of the two proteins in question and on the desired accuracy of the prediction.

CONFORMATIONAL CHANGES IN PROTEINS

The folded (or unfolded) conformation of a protein under particular conditions is the one with the lowest free energy. If the environment of the protein changes, however, the most stable conformation can also change. We have seen one example—the unfolding and refolding of ribonuclease in response to adding and removing a very high concentration of urea. Much less drastic changes in the environment of a protein can also induce functionally important, conformational shifts. For example, when presented with its substrate, glucose, the single-domain enzyme hexokinase closes up around it (Fig. 6-12). Formation of energetically favorable contacts with the substrate makes the closed structure more stable than the open one, shifting the position of a dynamic equilibrium from mostly open to mostly closed.

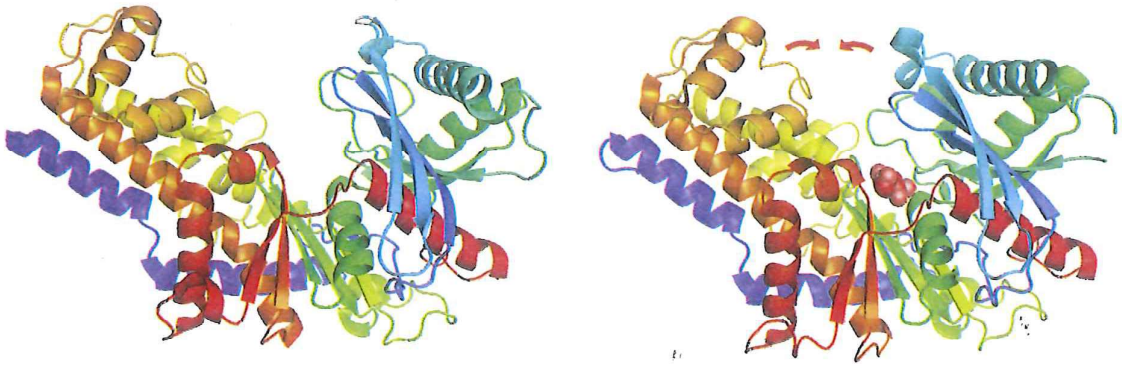


FIGURE 6-12 Domain closure in the enzyme hexokinase. The two lobes of hexokinase, an enzyme that transfers a phosphate to glucose, close up on each other (red arrows) when the substrate (glucose) binds. (Left) Enzyme before binding glucose. (Right) After binding glucose (shown in surface representation, red, in the catalytic cleft of the enzyme). The polypeptide chain is in rainbow colors from blue (amino terminus) to red (carboxyl terminus). Note that the folded chain traverses back and forth twice between the two lobes. Images prepared with PyMOL (Schrödinger, LLC).

Interaction of two proteins with each other can cause one or both partners to undergo a conformational change. Sometimes, the interacting part of one of the partners is unstructured (disordered and flexible) until it associates with the other partner. In other words, the properly folded conformation is stable only in the presence of its target, which can be DNA or RNA as well as another protein. The α helices in the dimeric coiled-coil of the yeast transcription factor GCN4 are stable only when associated with each other. When bound to DNA, an additional segment of the protein forms an α helix in the DNA major groove, but the same part of the protein is unstructured when GCN4 is not associated with its DNA-binding site (Fig. 6-9b).

PROTEINS AS AGENTS OF SPECIFIC MOLECULAR RECOGNITION

Proteins That Recognize DNA Sequence

Regulation of gene expression depends on proteins that bind short DNA segments having a specific nucleotide sequence. We consider here several examples that illustrate some of the principles of protein structure and interaction described above.

i. GCN4 We have already described GCN4 to illustrate how the folding of a protein sometimes depends on its interactions with other proteins (the other chain of the dimer, in the case of the GCN4 coiled-coil segment) or with a target (a DNA site). GCN4 binds tightly to DNA only when the sequence of bases at the contact site is the correct one. Because the α helices in the major groove fit snugly, their side chains need to be complementary—in their shapes, their polarity, and their hydrogen-bond donor and acceptor properties—to the DNA surface. These α helices also have several arginine and lysine residues, which anchor them to the DNA backbone by forming salt bridges with phosphates, reinforcing the snug fit.

ii. The Bacteriophage λ Repressor The repressor of bacteriophage λ has six binding sites on the bacteriophage genome, which all have related but slightly different sequences; the exact sequence of each of the sites determines its affinity for the repressor (Fig. 6-13a). The protein is a symmetric dimer, and the sites have approximately symmetric (**palindromic**) sequences. Each subunit of the protein has two folded domains: an amino-terminal, DNA-binding domain and a carboxy-terminal, dimerization domain.

The DNA-binding domain of λ repressor is a compact bundle of five α helices (Fig. 6-13b). Unlike GCN4, this domain does not undergo any major structural changes when it associates with DNA. Two of its helices (the second and third) form a structural motif, known as a **helix-turn-helix**, seen in many other DNA-binding proteins, especially those from prokaryotes. The way this motif fits against the DNA double helix allows the second of the two helices, sometimes called the **recognition helix**, to fit into the major groove of DNA and to present several of its side chains to the exposed edges of the base pairs (Fig. 6-13c). The major-group edge of each base pair presents a characteristic pattern of hydrogen-bond donor and acceptor groups; the A:T and T:A base pairs also present the hydrophobic surface of a thymine methyl group (Fig. 6-14). The hydrogen-bonding and nonpolar contact properties of side chains on the λ-repressor recognition helix match those of the base sequence recognized. Contacts between the protein and the

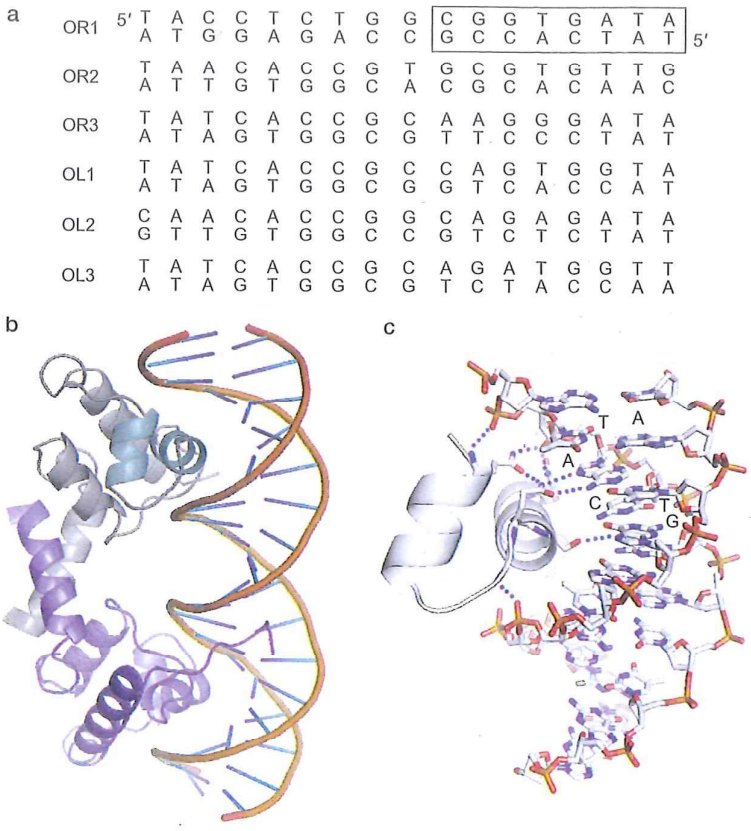


FIGURE 6-13 **DNA recognition by the repressor of bacteriophage λ.** (a) The nucleotide sequences of the six DNA sites ("operators") in the λ genome that bind the λ repressor. Each site is approximately a "palindrome"—the sequence of bases is the same (with some deviations) when read 5′ to 3′ from either end. The right-hand "half site" of the top sequence (OR1) and the left-hand half site of the bottom sequence (OL3) correspond to the best consensus of all the half sites. Because the overall length is an odd number (17 base pairs), the central base pair is necessarily an exception to a perfect palindrome. (b) The DNA-binding (amino-terminal) domain of λ repressor, bound to operator DNA. Each subunit is a cluster of five α helices. Two of these (in light blue on the upper subunit) form a helix-turn-helix motif; the first of the two bridges from one side of the major groove to the other, and the second lies in the groove and nearly parallel to its principal direction. (c) Polar interactions (hydrogen bonds and salt bridges) between residues in the helix-turn-helix motif and DNA (both backbone and bases). The protein fits snugly in the major groove only when the base-pair contacts match the groups on the protein that lie opposite them. Images prepared with PyMOL (Schrödinger, LLC).

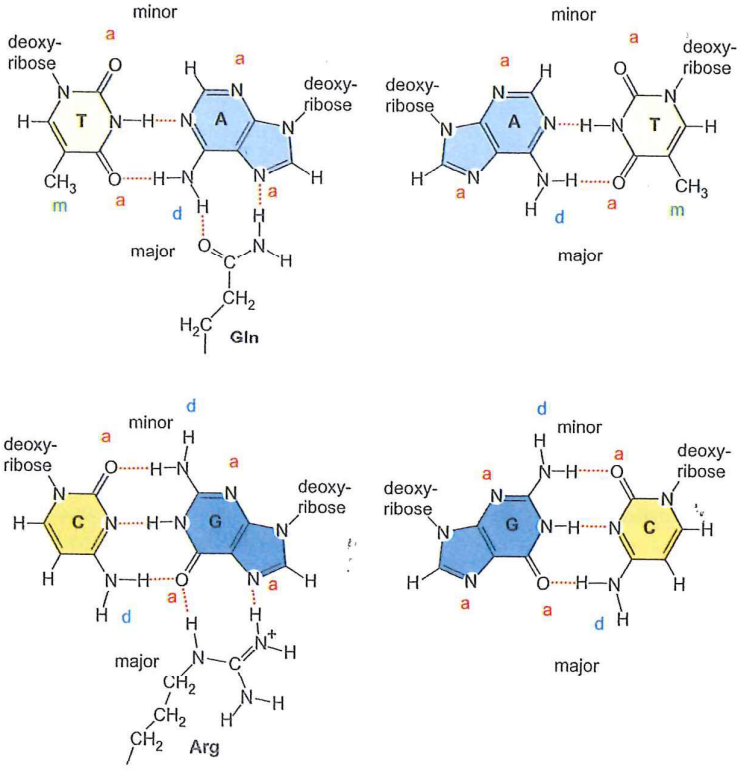


FIGURE 6-14 **Properties of DNA base pairs in the major and minor grooves.** The four DNA base pairs, with labels on groups in the major and minor groove that can determine specific contacts: a, hydrogen-bond acceptor; d, hydrogen-bond donor; m, methyl group (van der Waals contact). Hydrogen bonds are shown as dotted lines. In the major groove, each of the four base pairs presents a distinct pattern: T:A, m-a-d-a; A:T, a-d-a-m; C:G, d-a-a; G:C, a-a-d. Two particular examples of amino acid side-chain complementarity are shown with the T:A and C:G pairs. These two modes of base-pair recognition (pointed out in 1976 by Seeman, Rosenberg, and Rich) do occur with some frequency, but most cases of specific DNA recognition involve a more complex set of contacts. In the minor groove, T:A and A:T present the same pattern of potential contact (a-a); likewise, C:G and G:C (a-d-a). Thus, sequence-specific DNA recognition usually involves major-groove contacts.

sugar–phosphate backbone of DNA position and orient the recognition-helix side chains to ensure this complementarity.

The complementarity of protein side chains and DNA bases differs in an important way from the complementarity of the two bases in a DNA base pair. Each DNA base has a unique complementary base, such that their hydrogen bonding is consistent with the geometry of an undistorted double helix. In contrast, there are several ways in which proteins recognize a particular base or even a particular sequence of bases. Moreover, as illustrated by the different sequences of the repressor-binding sites, the same protein structure can adjust slightly to create complementarity with a slightly altered base sequence (at some cost in affinity). Thus, there is no "code" for DNA recognition by proteins—just a set of recurring themes, such as the presentation of protein side chains by an α helix inserted into the DNA major groove.

The λ repressor illustrates a general feature of proteins that recognize specific DNA sequences: they have relatively small DNA-binding domains, usually linked to one or more additional domains with distinct functions, such as oligomerization or interaction with other proteins.

iii. Zinc-Finger Proteins The most abundant DNA-recognition domain in many eukaryotes is a small module known as a **zinc finger** (Fig. 6-15a). These domains generally occur in tandem, with short linker segments between them. The linkers are flexible; when the proteins bind DNA, they become ordered. The approximately 30 residues of each zinc finger are barely enough to create a hydrophobic core, and the zinc ion in the center is necessary to hold together the folded domain. Two cysteines and two

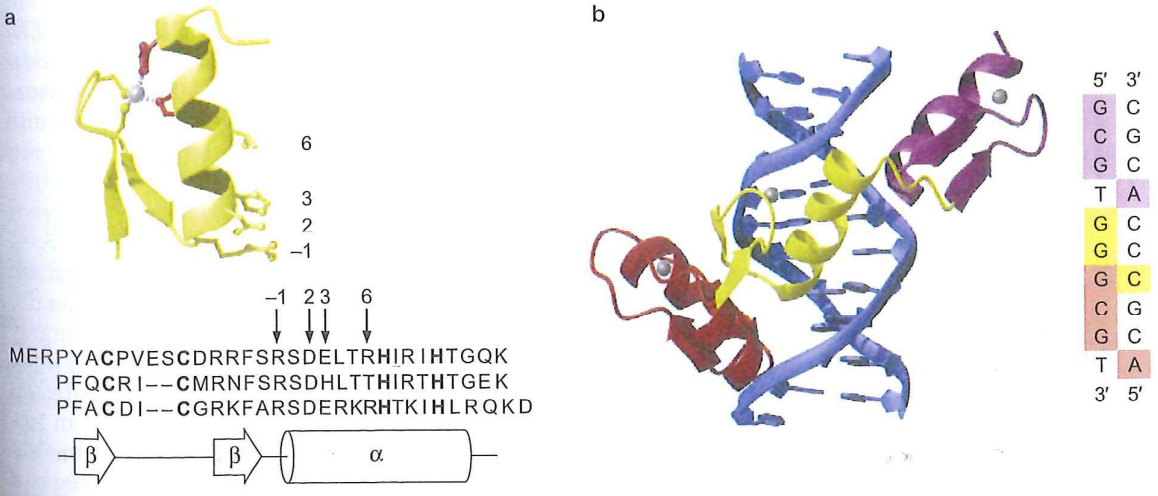


FIGURE 6-15 Zinc-finger motifs. (a) The Cys2His2 zinc finger motif and the Zif268 finger sequences. Shown at the top is a ribbon diagram of finger 2, including the two cysteine side chains (yellow) and two histidine side chains (red) that coordinate the zinc ion (silver sphere). The side chains of key residues make base contacts in the major groove of the DNA (numbers identify their position relative to the start of the recognition helix). Shown below is the amino acid sequence alignment of the three fingers from Zif268 with the conserved cysteines and histidines in boldface. Secondary structure elements are indicated at the bottom of the diagram. (b) To the left is the Zif268–DNA complex, showing the three zinc fingers of Zif268 bound in the major groove of the DNA. Fingers are spaced at 3-bp intervals; DNA (blue); fingers 1 (red), 2 (yellow), and 3 (purple); the coordinated zinc ions (silver spheres). The DNA sequence of the Zif268 binding site on the right is color-coded to indicate base contacts for each finger. (Reproduced, with permission, from Pabo C.O. et al. 2001. *Annu. Rev. Biochem.* **70:** 313–340, Fig. 6-15a is Fig. 1 on p. 315; Fig. 6-15b is Fig. 2 on p. 316. © Annual Reviews.)

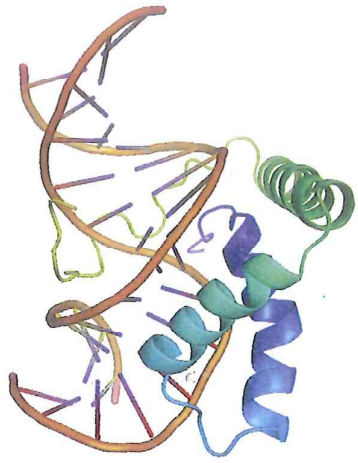


FIGURE 6-16 The LEF-1 protein bound to DNA. Image prepared with PyMOL (Schrödinger, LLC).

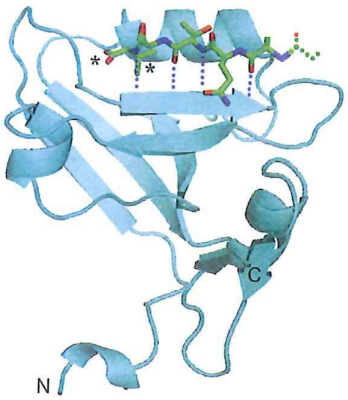


FIGURE 6-17 Peptide recognition. Specific recognition of the carboxy-terminal segment of a protein by a PDZ domain—a repeating module that associates with the carboxy-terminal, cytoplasmic "tails" of membrane proteins. Principal contacts are in pockets (asterisks for the carboxyl group and the nonpolar side chain of the carboxy-terminal valine) and through addition to the antiparallel β sheet in the domain (foreground) by several residues of the ligand that precede the valine (dotted black lines represent β-sheet hydrogen bonds). Image prepared with PyMOL (Schrödinger, LLC).

histidines coordinate the Zn^{2+}. Because intracellular proteins do not have disulfide bonds, Zn^{2+} coordination often serves the same stabilizing purpose for very small domains. When zinc fingers bind DNA, the short α helix lies in the major groove, and successive zinc fingers in a tandem array contact successive sets of base pairs—roughly 3 bp per zinc finger, with some overlap (Fig. 6-15b). There is considerable regularity in the pattern of base-pair contacts: residues −1, 2, 3, and 6 of the helix are the most likely to contact one or more base pairs (Fig. 6-15a). Because of this regularity—and the way in which tandem zinc fingers wind into the DNA major groove—proteins can be designed to recognize relatively long sequences of base pairs. Moreover, libraries of individual modules are now available to make designed proteins specific for DNA sequences 12–18 bp in length.

iv. Lymphocyte Enhancer Factor-1 (LEF-1) Contacts with base pairs in the major groove of DNA are not the only way to create base sequence specificity. The base sequence does not uniquely specify the pattern of hydrogen-bond contacts in the minor groove, because A:T and T:A look the same in this respect, as do G:C and C:G (Fig. 6-14), but base sequence also influences the propensity for the DNA double helix to bend or twist—that is, to adopt conformations that deviate from an ideal Watson–Crick double helix. This sensitivity to the influence of base sequence on the propensity of DNA to bend and twist is sometimes called "indirect readout," to distinguish it from the sequence specificity provided by direct polar and nonpolar contacts with base pairs. Lymphocyte Enhancer Factor-1 (LEF-1), which regulates T-cell gene expression in concert with several other factors, is a three-helix bundle that fits into the substantially widened minor groove of bent DNA (Fig. 6-16). Most of the amino acid side chains that face into the minor groove are nonpolar, and one of them inserts part way between two adjacent base pairs, helping to stabilize the nearly 90° bend in the DNA axis. The bend brings proteins bound upstream and downstream of LEF-1 closer together: It has been called an "architectural protein" for this reason, because part of its role is to enhance contacts between other DNA-bound transcription factors.

Protein–Protein Interfaces

Protein–protein interfaces tend to be even more exquisitely complementary than protein–DNA interfaces. The reason is that the former generally involve considerable hydrophobic surface, whereas the latter are largely polar. Water, which is both a donor and acceptor, can bridge gaps between hydrogen-bonding groups at a DNA–protein interface, but a gap between nonpolar surfaces at a protein interface would leave either a hole or an isolated water—both very unfavorable. As we have seen, a transcription factor such as λ repressor can bind DNA targets with a modest range of sequences, each deviating slightly from a consensus. The same is not true for most protein interfaces. For some transcription factors, alternative pairing of structurally homologous subunits *does* occur, to increase combinatorial diversity. The relevant complementary surfaces are conserved in such cases, which probably arise from gene duplication at some point in the evolutionary history of the protein.

Specific protein recognition can depend on association of prefolded, matching surfaces of two subunits, such as occurs in formation of a hemoglobin tetramer (Fig. 6-7), or on cofolding of two polypeptide chains, as in GCN4 dimerization (Fig. 6-9a), or on docking of an unstructured segment onto the recognition surface of a partner protein (Fig. 6-17). In this last sort of interaction, the segment in question adopts a defined structure in

the complex—that is, its correctly folded conformation is stable only in the presence of the target surface. Binding sometimes depends on a post-translational modification such as phosphorylation or acetylation, so that the interaction can be switched on or off by signals from other cellular processes. The docked segment of polypeptide chain often has a recognizable amino acid sequence motif. Association of this kind is particularly common in the assembly of protein complexes that regulate transcription, probably because it allows considerable variability in longer-range organization. Either the unstructured segment or the domain that binds it, or both, may be embedded in a larger unstructured region with a relatively polar, "low-complexity" amino acid composition (i.e., having many repeated instances of the same, polar residue). These low-complexity regions impart long-range flexibility, so the spacing between the specific interactions can vary, and the same assembly can adapt to different circumstances (e.g., to different arrays of sites on DNA).

Proteins That Recognize RNA

Unlike DNA, RNA can have a great variety of local structures, and tertiary interactions stabilize well-defined 3D conformations, as in tRNA. Protein–RNA interactions are therefore in some respects rather like protein–protein interactions. The shape of the RNA and the way interacting groups (e.g., phosphates or 2′-hydroxyl groups or bases) distribute on its surface are critical determinants of specificity. Two prefolded structures can associate, as in binding of a tRNA to the enzyme that transfers an amino acid to its 5′ end, or one or both partners can have little or no fixed structure until the complex forms.

The RNA-recognition motif (RRM; also known as the ribonuclear protein [RNP] motif) is a sequence that characterizes a domain involved in specific RNA recognition. The RRM sequence of 80–90 amino acids folds into a four-stranded antiparallel β sheet and two α helices that pack against it. This arrangement gives the domain a characteristic split αβ topology. An example of this common domain is found in the U1A protein that interacts with the U1 small nuclear RNA (snRNA), both components of the machinery that splices RNA transcripts (Chapter 14). The structure of the U1A:U1snRNA complex, shown in Figure 6-18, shows that the shape of the RNA-binding surface of U1A is specific for this particular RNA.

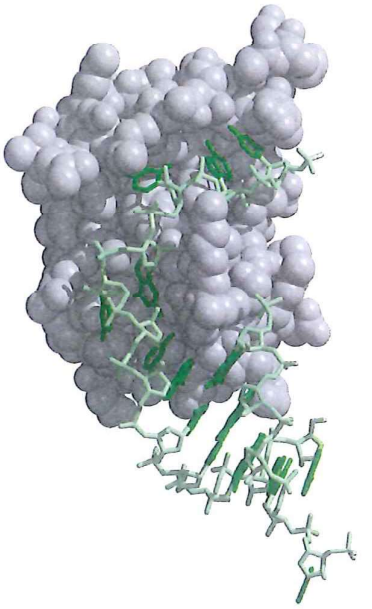


FIGURE 6-18 Structure of spliceosomal protein:RNA complex: U1A binds hairpin II of U1 snRNA. The protein is shown in gray; the U1 snRNA is shown in green. (Oubridge C. et al. 1994. *Nature* **372:** 432.) Image prepared with MolScript, BobScript, and Raster 3D.

ENZYMES: PROTEINS AS CATALYSTS

One of the most important roles for proteins in cells is to catalyze biochemical reactions. Almost all processes that go on in a cell—from transformation of nutrients for generating energy to polymerization of nucleotides for synthesis of DNA and RNA—require catalysis (i.e., enhancement of their rates), because the spontaneous reaction rates are far too slow to support normal cellular activity and survival. Most catalysts in living systems are proteins (enzymes); RNA is a catalyst for certain very ancient reactions (ribozymes).

The barrier to a chemical reaction is formation of a high-energy arrangement of the reactants, known as the **transition state**. Because the transition state has a structure intermediate between those of the reactants and the products, some distortion of the reactants is necessary to reach it. A reaction can be accelerated—often very dramatically—by reducing the energy needed to distort the reactants into their transition-state configurations. Most enzymes do so by having an **active site**—usually a pocket or groove—that

is complementary in shape and interaction properties (e.g., hydrogen bonds and nonpolar contacts) to the transition state of the reaction. The favorable contacts that form when the reactants associate with the active site compensate to some extent for the distortion they undergo to do so. The precision with which evolution of an enzyme structure molds its active site imparts great specificity to this process. For example, enzymes that catalyze polymerization of deoxyribonucleotides into DNA cannot, in general, catalyze polymerization of ribonucleotides into RNA, because the 2′-hydroxyl of the ribose would collide with atoms in the active site of the polymerase.

REGULATION OF PROTEIN ACTIVITY

We have seen that interaction with other molecules—both small molecules such as the substrates of an enzyme or macromolecules such as proteins and nucleic acids—can induce proteins to undergo conformational changes. Molecules that bind a protein (or any other target) in a defined way are known as **ligands**. Ligands can regulate the activity of a protein (e.g., an enzyme) by stabilizing a particular state. For example, if binding of a ligand to an enzyme stabilizes a conformation in which the active site is blocked, the ligand will have turned off the activity of that enzyme. The binding site for the inhibitory ligand need not overlap the active site—it need only

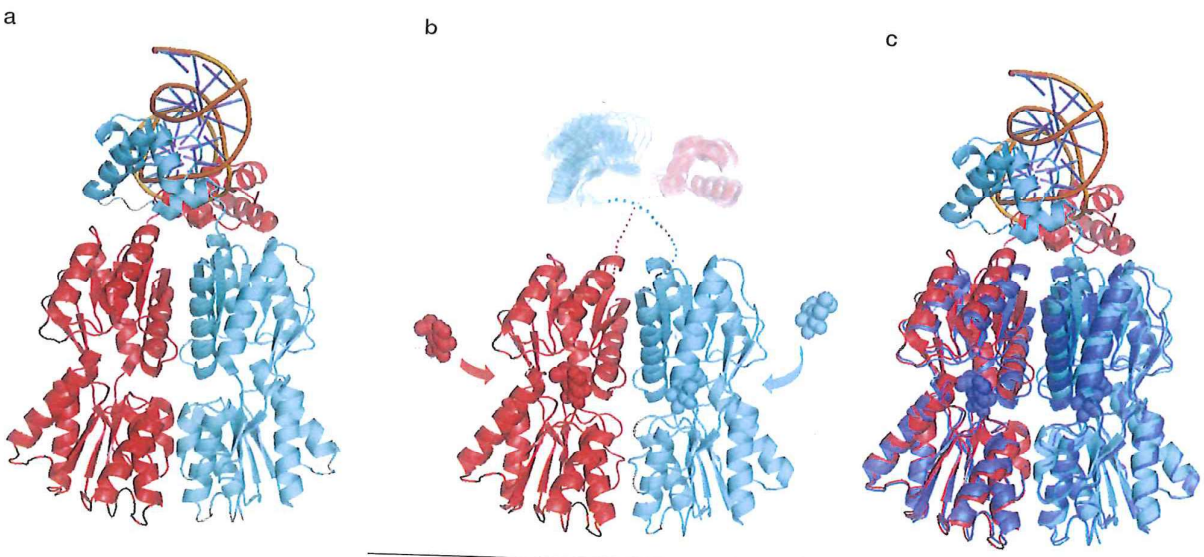


FIGURE 6-19 Allosteric regulation of Lac repressor DNA binding. (a) DNA-bound conformation of dimeric Lac repressor. A short DNA segment, representing the specific binding site ("operator"), is at the top of the figure. The amino-terminal, DNA-binding domain, with a helix-turn-helix recognition motif, interacts with base pairs in the major groove. The body of the protein has a site, located between its two domains, that accommodates molecules related to lactose; the site is empty in the DNA-bound conformation shown here. The two identical repressor subunits are in red and cyan, respectively. (b) Binding of an inducer molecule (any of a variety of galactosides, illustrated in surface rendering, both outside the repressor, as if about to bind, and also at the specific binding site within each repressor subunit) causes the two domains in the body of the repressor to change orientation with respect to each other. As a result, the hinge segments between the DNA-binding domains and the body of the protein become disordered, with the domains themselves now loosely tethered and unable to bind tightly to operator sites. (c) Superposition of the DNA-bound and induced conformations, to show how one of the domains of the repressor shifts with respect to the other. DNA-bound subunits are colored as in panel a; the induced repressor dimer is in dark blue. Images prepared with PyMOL (Schrödinger, LLC).

be such that ligand binding lowers the energy of a conformation in which the reactants cannot reach the active site or in which the active site no longer has the right configuration. Conversely, ligand binding at a remote site might favor a conformation in which the active site is available to substrate and complementary to the transition state of the reaction; the ligand would then be an activator. This kind of regulation is known as **allosteric regulation** or allostery, because the structure of the ligand (its "steric" character) is different from (Greek *allo-*) the structure of any of the reactants.

The Lac repressor (which inhibits expression of the bacterial gene encoding β-galactosidase, an enzyme that hydrolyzes β-galactosides such as lactose) is a good example of allosteric regulation in control of transcription (Fig. 6-19). Lac repressor is a dimer. The dimer has two distinct conformations—one when bound to a specific DNA site (known as its **operator**) and another when bound to an inhibitory metabolite (known as its **inducer**). Because the operator-bound repressor blocks RNA polymerase from synthesizing β-galactosidase mRNA and because a high concentration of the inducer favors a conformation that does not bind well to DNA, the inducer can change DNA affinity and hence influence gene regulation, even though its binding site is at some distance from the DNA-contacting surface of the repressor. Even more complicated allosteric switches are possible, with multiple ligands and multiple binding sites. Allosteric regulation often involves quaternary-structure changes, as in the transition between the two dimer conformations of Lac repressor.

SUMMARY

Proteins are linear chains of amino acids, joined by peptide bonds ("polypeptide chains"). The 20 L-amino acids specified by the genetic code include nine with nonpolar (hydrophobic) side chains, six with polar side chains that do not bear a charge at neutral pH, two with acidic side chains (negatively charged at neutral pH), and three with basic side chains (positively charged at neutral pH, or partially so in the case of histidine). Peptide bonds have partial double-bond character; torsion angles for the N-Cα and Cα-(C=O) bonds specify the three-dimensional conformation of a polypeptide-chain backbone. Three amino acids have special conformational properties: glycine is nonchiral, with greater conformational freedom than the others; proline (technically, an imino acid) has a covalent bond between side chain and amide, restraining its conformational freedom; and cysteine, with a sulfhydryl group on its side chain, can undergo oxidation to form a disulfide bond with a second cysteine, cross-linking a folded polypeptide chain or two neighboring polypeptide chains. The reducing environment of a cell interior restricts disulfide-bond formation to oxidizing organelles and the extracellular milieu.

Protein structure is traditionally described at four levels: primary (the sequence of amino acids in the polypeptide chain—the one level determined directly by the genetic code), secondary (local, repeated backbone conformations, stabilized by main-chain hydrogen bonds—principally α helices and β strands), tertiary (the folded, three-dimensional conformation of a polypeptide chain), and quaternary (association of folded polypeptide chains in a multisubunit assembly). At the tertiary level, polypeptide chains fold into one or more independent domains, which would fold similarly even if excised from the rest of the protein. The structure of a domain can usefully be specified by the way in which its component secondary-structure elements (helices and strands) pack together in three dimensions. Linkers between domains of a multidomain polypeptide chain can be long and flexible or short and stiff. The aqueous environment and the diverse set of naturally occurring amino acids are together critical for the conformational stability of folded domains and of the interfaces between them that create quaternary structure. Nonpolar side chains cluster away from water into the closely packed, hydrophobic core of a folded domain, and any sequestered hydrogen-bonding groups, which lose a hydrogen bond with water, must have a protein-derived partner. Secondary-structure elements satisfy the latter requirement for the main-chain amide and carbonyl groups, thus accounting for their importance in describing and classifying domain structures.

The sequence of amino acids in a polypeptide chain specifies whether and how it will fold. This property allows the genetic code to determine not merely primary structure, but other levels as well, and hence to dictate protein function. The various noncovalent interactions within a correctly folded domain (and in extracellular domains, the covalent disulfide bonds) create a global free-energy minimum (conformation of greatest stability), so that the chain can reach its native conformation spontaneously. Changes in the environment of a protein, including post-translational modifications of one or more of its side chains or binding of ligands, may alter the position of this free-energy minimum and

induce a conformational change. The array of amino acid side chains on the surface of a folded protein, and sometimes even in a segment of unfolded polypeptide chain, can also specify how it recognizes a protein or nucleic-acid partner or a small-molecule ligand. Proteins are thus the key agents of specific molecular recognition, both within a cell and between cells, as well as the specific catalysts of chemical reactions (enzymes).

BIBLIOGRAPHY

Books

Branden C. and Tooze J. 1999. *Introduction to protein structure*. 2nd ed. Garland Publishing, New York.

Kuriyan J., Konforti B., and Wemmer D. 2012. *The molecules of life*. Garland Publishing, New York.

Pauling L. 1960. *The nature of the chemical bond*, 3rd ed. Cornell University Press, Ithaca, New York.

Petsko G.A. and Ringe D. 2003. *Protein structure and function (primers in biology)*. New Science Press, Waltham, Massachusetts.

Williamson M. 2012. *How proteins work*. Garland Publishing, New York.

Protein Structure Can Be Described at Four Levels

Richardson J.S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34:** 167–339.

From Amino Acid Sequence to Three-Dimensional Structure

Anfinsen C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223–230.

QUESTIONS

MasteringBiology® *For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.

Question 1. What is the bond that can form between two cysteines in secreted proteins? Why does this bond not ordinarily form in intracellular proteins? How does this interaction differ from the interactions that can occur between other amino acid side chains?

Question 2. Give an example of two amino acid side chains that can interact with each other through an ionic bond at neutral pH. See Chapter 3 for a review of ionic bonds.

Question 3. A mutation that occurs in DNA can cause an amino acid substitution in the encoded protein. Amino acid substitutions are described as conservative when the amino acid in the mutated protein has chemical properties similar to those of the amino acid it has replaced. Referring to Figure 6-2, identify four different examples of pairs of amino acids that could be involved in conservative substitutions.

Question 4. Peptide bond formation is an example of a condensation reaction. Explain what this statement means and why peptide bond formation is also referred to as a dehydration reaction.

Question 5. Describe how a β strand differs from a β sandwich.

Question 6. The oxygen-binding proteins hemoglobin and myoglobin differ in that hemoglobin functions as a tetramer in red blood cells, whereas myoglobin functions as a monomer in muscle cells. The globular structure of myoglobin and each hemoglobin monomer involves eight α-helical segments. Is it the primary, secondary, tertiary, or quaternary structure that differs most between these two proteins? Explain.

Question 7. For the following amino acids, suggest whether they are more likely to be found buried or exposed in a stably folded protein domain: phenylalanine, arginine, glutamine, methionine. Explain your answers.

Question 8. You treat a protein with the denaturant urea. For each interaction or bond below, state if the interaction or bond is disrupted by the urea treatment.

A. Ionic bonds.

B. Hydrogen bonds.

C. Disulfide bonds.

D. Peptide bonds.

E. van der Waals interactions.

Question 9. From what you learned about the structure of DNA in Chapter 4, explain why Gcn4 interacts with DNA in the major groove rather than in the minor groove. Describe the importance of arginines and lysines in the interaction between Gcn4 and DNA.

Question 10. Predict the effect of substituting one or more of the conserved cysteines or histidines in a Cys2His2 zinc finger with alanine. Explain your answer.

Question 11. Describe the unusual features of the interaction of LEF-1 with DNA.

Question 12. How do enzymes enhance the rate of a reaction?

Question 13. Consider a ligand that is structurally similar to the substrate of an enzyme and that binds tightly in the active site, excluding the normal substrate. What is the difference between such a "competitive inhibitor" and an allosteric inhibitor?

Question 14. A translation initiation factor, called Tif3 or eIRF4B, in yeast cells has the following sequence of elements in its polypeptide chain: an amino-terminal domain containing

KEY CONCEPT
24.1 Bacteria and Archaea Represent the First Split in the Tree of Life

Learning Objectives

24.1.1 Compare and contrast the features of bacteria, archaea, and eukaryotes.

24.1.2 Describe how archaea and bacteria each contributed to the evolution of eukaryotes.

24.1.3 Draw a phylogenetic tree demonstrating the effects of lateral gene transfer.

You may think that you have little in common with a bacterium. But all multicellular eukaryotes, including you, share many attributes with the **prokaryotes**—organisms that lack a nucleus (see Key Concept 5.1). For example, all organisms, whether eukaryotes or prokaryotes,

- have cell membranes and ribosomes (see Chapter 5 and 6).
- have a common set of metabolic pathways, such as glycolysis (see Chapter 9).
- replicate DNA semiconservatively (see Chapter 13).
- use DNA as the genetic material to encode proteins, and use similar genetic codes to produce those proteins by transcription and translation (see Chapter 14).

These shared features support the hypothesis that all living organisms share a common ancestor. If life had multiple origins, there would be little reason to expect all organisms to use overwhelmingly similar genetic codes or to share structures as distinctive as ribosomes. Furthermore, similarities in the DNA sequences of genes that are shared by all organisms confirm the monophyly of life.

The earliest split in the tree of life gave rise to Bacteria and Archaea

Although all living things share many features, major differences have evolved across the diversity of life. Biologists have now sequenced the genomes of many living organisms, and these genomes allow us to reconstruct the details of evolutionary history. These studies clarify that the earliest split in the tree of life resulted in two major groups: Bacteria and Archaea (Figure 24.1).

Although the earliest division of life gave rise to two major lineages, biologists usually refer to *three* major **domains** of life: Bacteria, Archaea, and Eukarya (which includes all plants, animals, and fungi). But if you examine Figure 24.1, you will see that the third domain, Eukarya, arose through phylogenetic contributions from both bacteria and archaea. (Note that we use lowercase when referring to members of these domains and initial capitals when referring to the domains themselves.) Most genes of eukaryotes are more closely related to those of archaea, but several endosymbiotic events involving bacteria also contributed to making eukaryotes a distinctive new lineage. So biologists are beginning to view eukaryotes as a specialized group of archaea that developed some important new features (including a cell nucleus and mitochondria, at least partly through contributions from bacteria). We call all the organisms that lack these specializations prokaryotes (Greek, "before the kernel," or before the evolution of a cellular nucleus).

In Chapter 20 we discussed the concept of monophyly (a group that includes an ancestor and all of its descendants) and the importance of monophyly in biological taxonomy. However, the reticulations of bacterial and archaeal lineages near the base of the tree of life make the recognition of strictly monophyletic groups of life's primary domains impractical. Prokaryotes are clearly not a monophyletic group (because the group excludes some of their descendants—the eukaryotes), but the name "prokaryotes" is nonetheless a useful way of talking about all organisms that are not eukaryotes. In a similar manner, when biologists talk about "archaea" or "bacteria," they are typically excluding their eukaryotic descendants, so neither of those groups are strictly monophyletic either. These complications in our taxonomic terminology are unavoidable, given the reticulations that occur among the major lineages near the base of the tree of life. We will use "bacteria" and "archaea" to refer to the two major groups of prokaryotes, and "eukaryotes" to refer to the lineage that emerged from combinations of some specific bacteria and archaea.

All prokaryotic organisms are unicellular, although they may form large, coordinated colonies or communities consisting of many individuals. Eukaryotes, by contrast, include both unicellular as well as many multicellular life forms. As we saw in Chapter 5, prokaryotic cells differ from eukaryotic cells in some important ways:

- *Prokaryotic cells do not divide by mitosis.* Instead, after replicating their DNA, prokaryotic cells divide by their own method, binary fission (see Key Concept 11.1).
- *The organization of the genetic material differs.* The DNA of the prokaryotic cell is not organized within a membrane-enclosed nucleus. DNA molecules in prokaryotes are often circular. Many (but not all) prokaryotes have only one main chromosome and are effectively haploid, although many have additional smaller DNA molecules, called plasmids (see Key Concept 12.6).
- *Prokaryotes have none of the membrane-enclosed cytoplasmic organelles that are found in most eukaryotes.* However, the cytoplasm of a prokaryotic cell may contain a variety of infoldings of the cell membrane and photosynthetic membrane systems that are not found in eukaryotes.

Although the study and classification of eukaryotic organisms go back centuries, much of our knowledge of the evolutionary relationships of prokaryotes is extremely recent. Not until the final quarter of the twentieth century did advances in molecular genetics and biochemistry enable the research that revealed the deep-seated distinctions between archaea and bacteria.

The two prokaryotic groups differ in significant ways

A glance at Table 24.1 will show you that there are major differences (most of which cannot be seen even under an electron microscope) between archaea and bacteria. Archaea share several features with eukaryotes, but they also retain some ancestral similarities with bacteria. The basic unit of an archaeon (the term for a single archaeal organism) or bacterium (a single bacterial

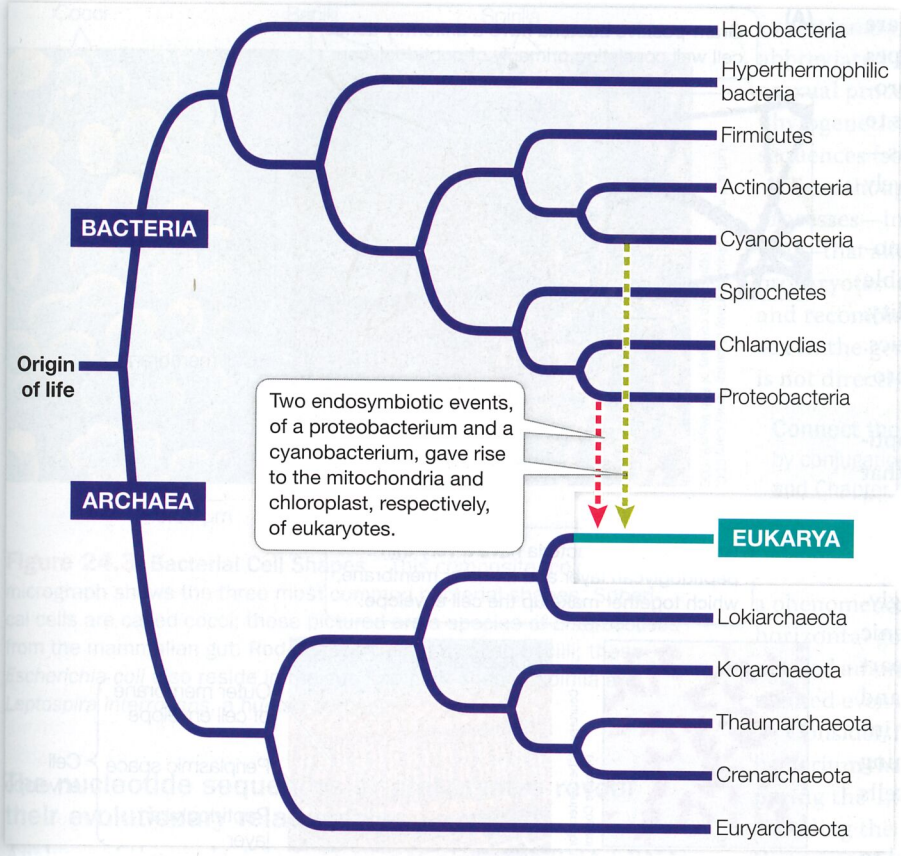


Figure 24.1 The Major Groups of the Living World This phylogenetic tree of Bacteria and Archaea shows their relationships to each other and to their descendants, Eukarya. The relationships among the many clades of bacteria, most of which are not shown here, are incompletely resolved at this time.

View in Achieve
Animation 24.1 The Primary Divisions of Life

TABLE 24.1 | Characteristics of Bacteria, Archaea, and Eukarya

Characteristic	Bacteria	Archaea	Eukarya
Membrane-enclosed nucleus	Absent	Absent	Present
Membrane-enclosed organelles	Few	Absent	Many
Peptidoglycan in cell wall	Present	Absent	Absent
Membrane lipids	Ester-linked Unbranched	Ether-linked Branched	Ester-linked Unbranched
Ribosomes[a]	70S	70S	80S
Initiator tRNA	Formylmethionine	Methionine	Methionine
Operons	Yes	Yes	Rare
Plasmids	Yes	Yes	Rare
Number of RNA polymerases[b]	One	One	Three
Ribosomes sensitive to chloramphenicol and streptomycin	Yes	No	No
Ribosomes sensitive to diphtheria toxin	No	Yes	Yes

[a]70S ribosomes are smaller than 80S ribosomes.

[b]The structure of archaeal RNA polymerase is similar to that of eukaryotic polymerases.

organism) is the prokaryotic cell. Each single-celled prokaryote contains a full complement of genetic and protein-synthesizing systems, including DNA, RNA, and all the enzymes needed to transcribe and translate genetic information into proteins. The prokaryotic cell also contains at least one system for generating the ATP it needs.

Genetic studies clearly indicate that all forms of life share a single common ancestor. As we noted earlier, most of the eukaryotic genome shares a more recent common ancestor with certain groups of archaea than with bacteria (see Figure 24.1). However, the mitochondria of eukaryotes and the chloroplasts of photosynthetic eukaryotes (such as plants) originated through endosymbioses with bacteria. Some biologists prefer to view the origin of eukaryotes as a fusion of two equal partners (one ancestor that was related to modern archaea and another that was more closely related to modern bacteria). Others view the divergence of the early eukaryotes from specific groups of archaea as an event separate from the later endosymbioses. In either case, most eukaryotic genes are more closely related to those of specific groups of archaea, whereas other genes (especially genes related to mitochondria and chloroplasts) are most closely related to those of bacteria.

Connect the Concepts The origin of mitochondria and chloroplasts by endosymbiosis is described in Key Concepts 5.5 and 26.1.

Biologists estimate that the last common ancestor of all living organisms lived about 3 billion years ago. We can deduce that it had DNA as its genetic material, and that its machinery for transcription and translation produced RNAs and proteins, respectively. This ancestor likely had a circular chromosome. All living organisms are the products of billions of years of mutation, natural selection, and genetic drift, and they are all well adapted to present-day environments. The earliest prokaryotic fossils, which date back at least 3.5 billion years, indicate that there was considerable diversity among the prokaryotes even during those earliest days of life.

The small size of prokaryotes has hindered our study of their evolutionary relationships

Until about 300 years ago, nobody had even *seen* an individual prokaryote. Most prokaryotes remained invisible to humans until the

invention of the first simple microscope. Prokaryotes are so small, however, that even the best light microscopes don't reveal much about them. It took advanced microscopic equipment and modern molecular techniques to open up the microbial world. (Microscopic organisms—both prokaryotes and eukaryotes—are often collectively referred to as "microbes.")

Before DNA sequencing became practical, taxonomists based prokaryote classification on observable phenotypic characters such as shape, color, motility, nutritional requirements, and sensitivity to antibiotics. One of the characters most widely used to classify prokaryotes is the structure of their cell walls.

The cell walls of almost all bacteria contain **peptidoglycan**, a cross-linked polymer of amino sugars that produces a firm, protective, meshlike structure around the cell. Peptidoglycan is a substance unique to bacteria; its absence from the cell walls of archaea is a key difference between the two prokaryotic domains. Peptidoglycan is also an excellent target for combating pathogenic (disease-causing) bacteria because it has no counterpart in eukaryotic cells. Antibiotics such as penicillin and ampicillin, as well as other agents that specifically interfere with the synthesis of peptidoglycan-containing cell walls, tend to have little, if any, effect on the cells of humans and other eukaryotes.

The **Gram stain** is a technique that can be used to separate most types of bacteria into two distinct groups. A smear of bacterial cells on a microscope slide is soaked in a violet dye and treated with iodine; it is then washed with alcohol and counterstained with a red dye called safranin. **Gram-positive bacteria** retain the violet dye and appear blue to purple (Figure 24.2A). The alcohol washes the violet stain out of **Gram-negative bacteria**, which then pick up the safranin counterstain and appear pink to red (Figure 24.2B). For most bacteria, the effect of the Gram stain is determined by the chemical structure of the cell wall:

- A *Gram-negative cell wall* usually has a thin peptidoglycan layer, which is surrounded by a second, outer membrane quite distinct in chemical makeup from the cell membrane (see Figure 24.2B). Together the cell wall and the outer membrane are called the cell envelope. The space between the cell membrane and the outer membrane (known as the periplasmic space) contains proteins that are important in digesting some materials, transporting others, and detecting chemical gradients in the environment.
- A *Gram-positive cell wall* usually has about five times as much peptidoglycan as a Gram-negative cell wall. Its thick peptidoglycan layer is a meshwork that may serve some of the same purposes as the periplasmic space of the Gram-negative cell envelope.

Shape is another phenotypic character that is useful for the basic identification of bacteria. The three most common shapes are spheres, rods, and spiral forms (Figure 24.3). Many bacterial names are based on these shapes. A spherical bacterium is called a **coccus** (plural *cocci*). Cocci may live singly or may associate in two- or three-dimensional arrays such as chains, plates, blocks, or clusters of cells. A rod-shaped bacterium is called a **bacillus** (plural *bacilli*). A spiral bacterium (shaped like a corkscrew) is called a **spirillum** (plural *spirilla*). Bacilli and spirilla may be single, form chains, or gather in regular clusters. Among the other bacterial shapes are long filaments and branched filaments.

Less is known about the shapes of archaea because many of these organisms have never been seen. Many archaea are known only from samples of DNA from the environment. However, the species whose morphologies are known include cocci, bacilli, and even triangular and square species. Some flattened species grow on surfaces, arranged like sheets of postage stamps.

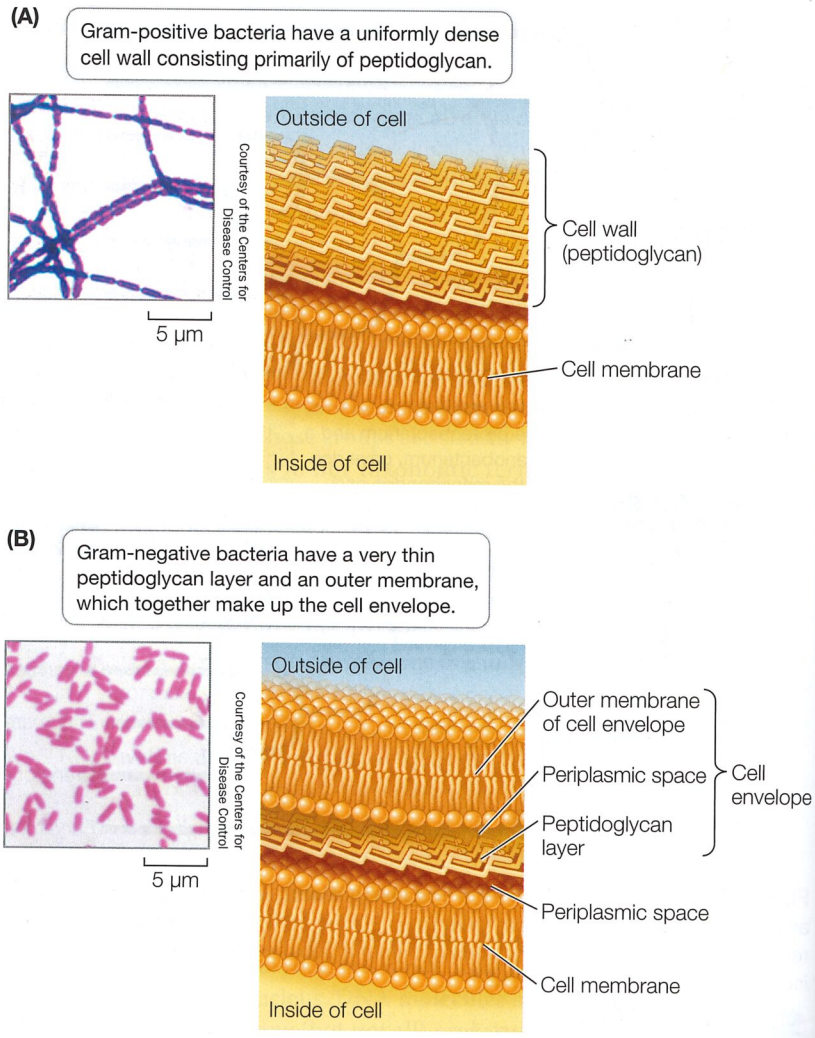
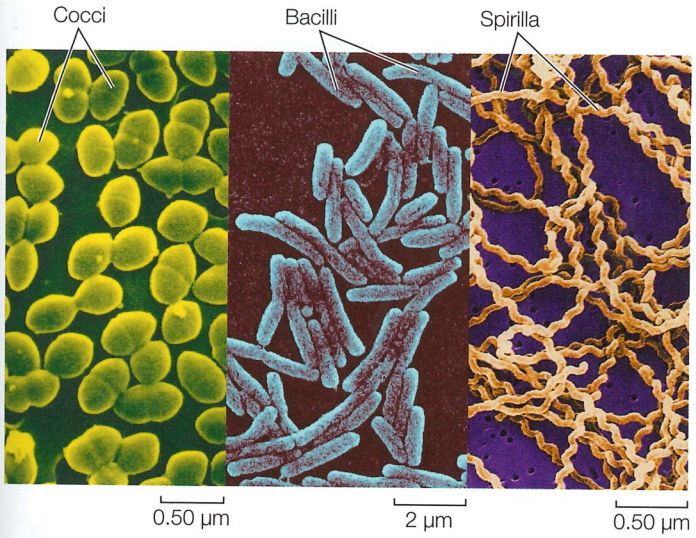


Figure 24.2 The Gram Stain and the Bacterial Cell Wall When treated with Gram-staining reagents, the cell walls of bacteria react in one of two ways. **(A)** Gram-positive bacteria have a thick peptidoglycan cell wall that retains the violet dye and appears deep blue or purple. **(B)** Gram-negative bacteria have a thin peptidoglycan layer that does not retain the violet dye, but picks up the counterstain and appears pink to red.

View in Achieve
Activity 24.1 Gram Stain and Bacteria



Left and middle photo courtesy of Janice Haney Carr/CDC; right photo courtesy of Rob Weyant/CDC

Figure 24.3 Bacterial Cell Shapes This composite, colorized micrograph shows the three most common bacterial shapes. Spherical cells are called cocci; those pictured are a species of *Enterococcus* from the mammalian gut. Rod-shaped cells are called bacilli; these *Escherichia coli* also reside in the gut. The helix-shaped spirilla are *Leptospira interrogans*, a human pathogen.

The nucleotide sequences of prokaryotes reveal their evolutionary relationships

Analyses of the nucleotide sequences of ribosomal RNA (rRNA) genes provided the first comprehensive evidence of evolutionary relationships among prokaryotes. Comparisons of rRNA genes are often used to identify microbes. For several reasons, rRNA is particularly useful for phylogenetic studies and identification purposes:

- rRNA was present in the common ancestor of all life and is therefore evolutionarily ancient.
- No free-living organism lacks rRNA, so rRNA genes can be compared across the tree of life.
- rRNA plays a critical role in translation in all organisms, so lateral transfer of rRNA genes among distantly related species is unlikely.
- rRNA has evolved slowly enough that gene sequences from even distantly related species can be aligned and analyzed.

Although studies of rRNA genes reveal much about the evolutionary relationships of prokaryotes, they don't always reveal the entire evolutionary history of these organisms. In some groups of prokaryotes, analyses of multiple gene sequences have suggested several different phylogenetic patterns. How could such differences among different gene sequences arise? Studies of whole prokaryotic genomes have revealed that even distantly related prokaryotes sometimes exchange genetic material.

Lateral gene transfer can lead to discordant gene trees

As noted earlier, prokaryotes reproduce by binary fission. If we could follow these divisions back through evolutionary time, we would be tracing the complete tree of life. At a much broader scale, these divisions of organisms lead to splits among the major evolutionary lineages, or species of life (represented in highly abbreviated form in Appendix A). Because binary fission is an asexual process that replicates whole genomes, we would expect phylogenetic trees of prokaryotes constructed from most gene sequences (see Chapter 21) to reflect these same relationships.

Even though binary fission is an asexual process, there are other processes—including transformation, conjugation, and transduction—that allow the transfer of genetic information between some prokaryotes without reproduction. Thus prokaryotes can transfer and recombine their DNA with that of other individuals (this is sex in the genetic sense of the word), but this genetic exchange is not directly linked to reproduction, as it is in most eukaryotes.

Connect the Concepts Prokaryotic exchange of genetic material by conjugation and transformation is described in Key Concept 12.6 and Chapter 18, respectively.

From early in evolution to the present day, some genes have been moving "sideways" from one prokaryotic species to another, a phenomenon known as **lateral gene transfer** (sometimes called horizontal gene transfer). Lateral gene transfers are well documented among closely related species, and some have been documented even across the domains of life.

Consider, for example, the genome of *Thermotoga maritima*, a bacterium that can survive extremely high temperatures. By comparing the 1,869 gene sequences of *T. maritima* with sequences encoding the same proteins in other species, investigators found that some of this bacterium's genes have their closest relationships not with the genes of other bacterial species, but with the genes of archaea that live in similar extreme environments.

When genes involved in lateral transfer events are sequenced and analyzed, the resulting **gene trees** will not match the organismal tree in every respect (**Figure 24.4**). The individual gene trees will vary because the history of lateral transfer events is different for each gene. Biologists can reconstruct the underlying organismal phylogeny by comparing multiple genes (to produce a consensus tree) or by concentrating on genes that are unlikely to be involved in lateral gene transfer events. For example, genes that are involved in fundamental cellular processes (such as the rRNA genes discussed in the previous section) are unlikely to be replaced by the same genes from other species because functional, locally adapted copies of these genes are already present.

What kinds of genes are most likely to be involved in lateral gene transfer? Genes that result in a new adaptation that confers higher fitness on a recipient species are most likely to be transferred repeatedly among species. For example, genes that produce antibiotic resistance are often transferred among bacterial species on plasmids, especially under the strong selection pressure such as that imposed by modern antibiotic medications. Improper or overly frequent use of antibiotics can select for resistant strains of bacteria that are much harder to treat. This selection for antibiotic resistance explains why informed physicians have become more careful in prescribing antibiotics.

It is debatable whether lateral gene transfer has seriously complicated our attempts to resolve the tree of prokaryotic life. Recent work suggests that it has not. Lateral gene transfer rarely creates problems at higher taxonomic levels, even though it may

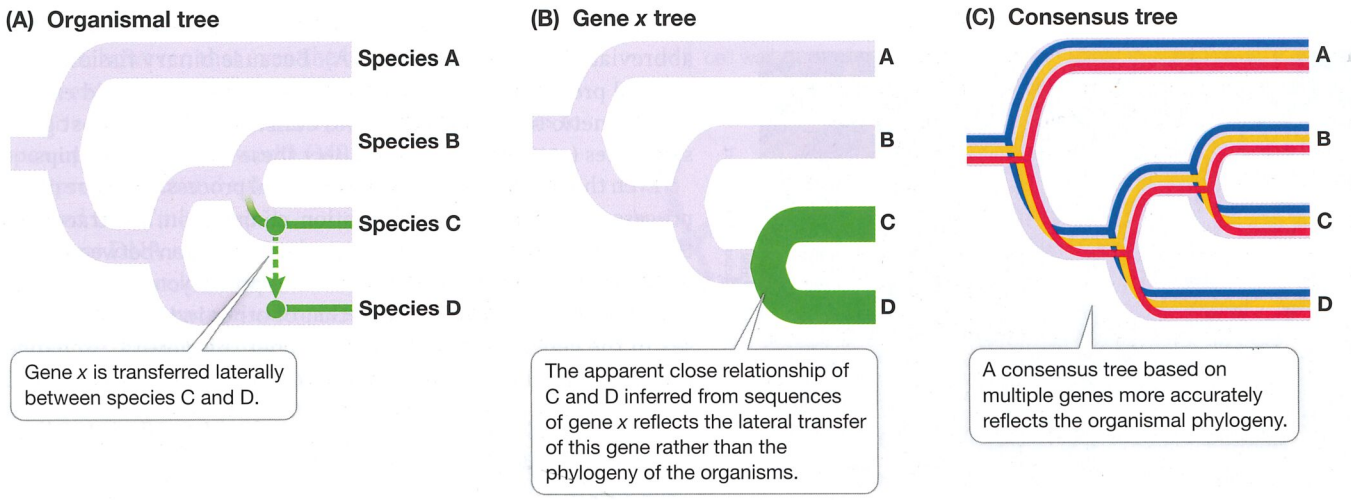


Figure 24.4 Lateral Gene Transfer Complicates Phylogenetic Relationships **(A)** The phylogeny of four hypothetical prokaryotic species, two of which have been involved in a lateral transfer of gene *x*. **(B)** A tree based only on gene *x* shows the phylogeny of the laterally transferred gene, rather than the organismal phylogeny. **(C)** A consensus tree based on multiple genes is more likely to reflect the true organismal phylogeny, especially if those genes come from a stable core of genes involved in fundamental processes.

Q: Why are multiple lateral gene transfers between the same two branches on a phylogeny expected to be rare, at least compared with similarities inherited through the stable core?

complicate our understanding of the relationships among individual species. Some species clearly obtain some of their genes from otherwise distantly related species, so evolutionary histories of individual genes may differ within a single organism. But it is now possible to make nucleotide sequence comparisons involving entire genomes, and these studies are revealing a stable core of crucial genes that are uncomplicated by lateral gene transfer. Gene trees based on this stable core more accurately reveal the organismal phylogeny (see Figure 24.4). The problem remains, however, that only a very small proportion of the prokaryotic world has been described and studied.

The great majority of prokaryotic species have never been studied

Most prokaryotes have defied all attempts to grow them in pure culture, causing biologists to wonder how many species, and possibly even major clades, we might be missing. A window onto this problem was opened with the introduction of a new way of examining nucleic acid sequences. When biologists are unable to work with the whole genome of a single prokaryotic species, they can instead examine genomes collected from an environmental sample (such as a scoop of sediment from the seafloor). This technique is known as **environmental genomics**.

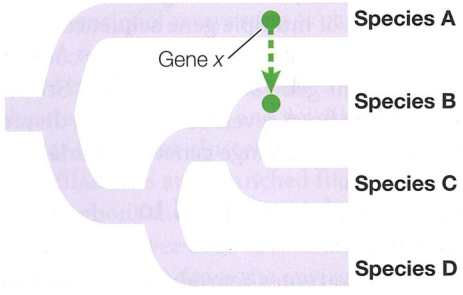
Biologists now routinely isolate gene sequences, or even whole genomes, from environmental samples such as soil and seawater. Comparing such sequences with previously known ones has revealed that an extraordinary number of the sequences represent new, previously unrecognized species. Biologists have described only about 10,000 species of bacteria and only a few hundred species of archaea (see Figure 1.12). The results of some environmental genomic studies suggest that there may be millions—perhaps hundreds of millions—of prokaryotic species. Other biologists put the estimate much lower, arguing that the high dispersal ability of many bacterial species greatly reduces endemism (i.e., the number of species restricted to a particular geographic area). Only the magnitude of these estimates differs, however; all sides agree that we have just begun to uncover Earth's prokaryotic diversity.

KEY CONCEPT

24.1 Recap and Assess

Bacteria and archaea are the two major lineages that arose from the first split in the tree of life. Eukaryotes evolved from a group of archaea that captured and incorporated bacteria through endosymbiosis. Mitochondria and chloroplasts are among the cellular structures of eukaryotes that are derived from endosymbiosis of bacteria. Although only about 10,000 species of bacteria and a few hundred species of archaea have been formally described, genetic evidence suggest there are many millions of species.

1. What findings support the separation of Bacteria and Archaea into separate domains?
2. The figure below shows an organismal tree in which gene *x* has undergone a lateral transfer event. Draw the phylogenetic tree you would expect based on gene *x*, as well as the phylogenetic tree you would expect based on a consensus of non-transferred genes.



3. How did eukaryotes arise through evolutionary contributions from both archaea and bacteria?

Despite the challenges of reconstructing prokaryotic phylogeny, taxonomists are starting to establish evolutionary classification systems for these organisms. With a full understanding that new information requires periodic revisions in these classifications, we next apply a current system of classification to organize our survey of prokaryotic diversity.

KEY CONCEPT 24.2 Prokaryotic Diversity Reflects the Ancient Origins of Life

Learning Objectives

- **24.2.1** Interpret a phylogenetic tree of major lineages of bacteria and archaea and use it to explain the origins of eukaryotes.
- **24.2.2** Discuss two lines of evidence that support the origin of life in a high-temperature environment.
- **24.2.3** Explain the role of bacteria in changing the oxygen content of Earth's atmosphere.

The prokaryotes were alone on Earth for a very long time, adapting to new environments and to changes in existing environments. They have survived to this day, in massive numbers and incredible diversity, and they are found nearly everywhere. In numbers of individuals, prokaryotes are far more abundant than eukaryotes. Individual prokaryotes in the oceans number more than 3×10^{28}—more than the number of stars in the universe. Closer to home, the individual bacteria living in your intestinal tract outnumber all the humans who have ever lived.

Given our still-fragmentary knowledge of prokaryotic diversity, it is not surprising that there are many different hypotheses about the relationships of the major groups of prokaryotes. In this book we use a classification system that is supported by nucleotide sequence data. We discuss a few of the major bacterial groups that have the broadest phylogenetic support and have received the most study, including hadobacteria, hyperthermophilic bacteria, firmicutes, actinobacteria, cyanobacteria, spirochetes, chlamydias, and proteobacteria (see Figure 24.1). Many other major groups of bacteria are known but are less thoroughly studied. We then describe the archaea, whose great diversity is just beginning to be fully understood and appreciated.

Two early-branching lineages of bacteria live at very high temperatures

Several lineages of bacteria and archaea are **extremophiles**: they thrive under extreme conditions that would kill most other organisms. The **hadobacteria**, for example, are thermophiles (Greek, "heat lovers"). The group's name is derived from Hades, the ancient Greek name for the underworld. Hadobacteria of the genus *Deinococcus* are resistant to radiation and can degrade nuclear waste and other toxic materials. They can also survive extremes of cold as well as hot temperatures. Another hadobacterium, *Thermus aquaticus*, was the source of the thermally stable DNA polymerase that was critical for the development of the polymerase chain reaction. *Thermus aquaticus* was originally isolated from a hot spring, but it can be found wherever hot water occurs (including in many residential hot-water heaters).

The **hyperthermophilic bacteria** are another major group of extremophiles. Genera such as *Aquifex* live near volcanic vents and in hot springs, sometimes at temperatures near the boiling point of water. Some species of *Aquifex* need only hydrogen, oxygen, carbon dioxide, and mineral salts to live and grow. Species of the genus *Thermotoga* live deep underground in oil reservoirs as well as in other high-temperature environments.

Biologists have hypothesized that high temperatures characterized the ancestral conditions for life, given that most environments on early Earth were much hotter than those of today. Reconstructions of ancestral bacterial genes have supported this hypothesis by showing that the ancestral sequences functioned best at elevated temperatures. The presence of multiple lineages of extremophiles at the base of the bacterial tree (see Figure 24.1) also provides support for the origin of life in a high-temperature environment.

Firmicutes include some of the smallest cellular organisms

The **firmicutes** are also known as low-GC Gram-positive bacteria. The first part of this description (low-GC) derives from the relatively low ratio of G-C to A-T nucleotide base pairs in their DNA. The second part (Gram-positive) is less accurate: some of the firmicutes are in fact Gram-negative, and some have no cell wall at all. Despite these differences, phylogenetic analyses of DNA sequences support the monophyly of this bacterial group.

One group of firmicutes can produce resting structures called **endospores** (Figure 24.5). When a key nutrient such as nitrogen or carbon becomes scarce, the bacterium replicates its DNA and encapsulates one copy, along with some of its cytoplasm, in a

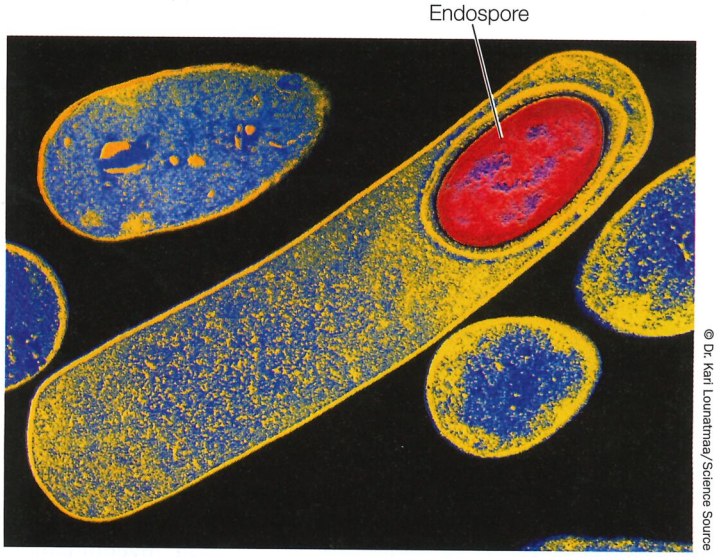
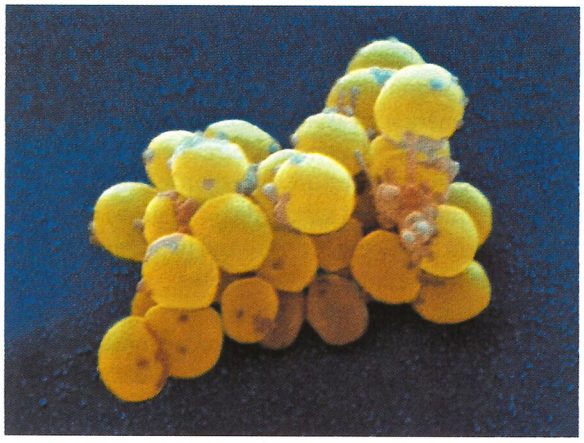


Figure 24.5 A Structure for Waiting Out Bad Times Under harsh conditions, some firmicutes can replicate their DNA and encase it in an endospore. The parent cell then breaks down, and the endospore survives in a dormant state until conditions improve.



Staphylococcus aureus 1 μm

Figure 24.6 Staphylococci "Grape clusters" are the usual arrangement of these firmicutes, which are often the cause of skin or wound infections.

Mycoplasma sp. 0.7 μm

Figure 24.7 Tiny Cells With about one-fifth as much DNA as *E. coli* has, mycoplasmas are among the smallest known bacteria.

tough endospore wall heavily thickened with peptidoglycan and surrounded by a spore coat. The parent cell then breaks down, releasing the endospore. Endospore production is not a reproductive process, as the endospore merely replaces the parent cell. The endospore, however, can survive harsh environmental conditions that would kill the parent cell, such as high or low temperatures or drought, because it is dormant—its normal metabolic activity is suspended. Later, if it encounters favorable conditions, the endospore becomes metabolically active and divides, forming new cells that are like the parent cell. Members of this endospore-forming group include the many species of *Clostridium* and *Bacillus*. Some of their endospores can be reactivated after more than 1,000 years of dormancy. There are even credible claims of reactivation of *Bacillus* endospores that are millions of years old.

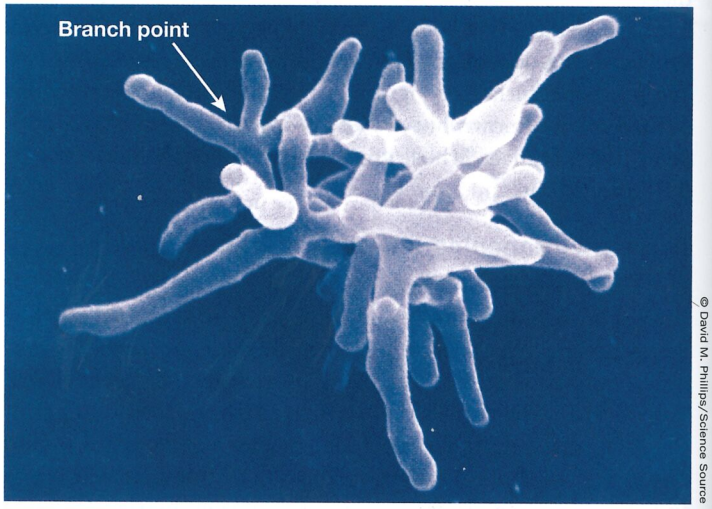
Endospores of *Bacillus anthracis* are the cause of anthrax. Anthrax is primarily a disease of cattle and sheep, but it can be fatal in humans. When the endospores sense macrophage (a type of white blood cell that digests cellular debris and foreign substances in mammalian blood), they reactivate and release toxins into the bloodstream. *Bacillus anthracis* has been used as a bioterrorism agent because it is relatively easy to transport large quantities of its endospores and release them among human populations, where they may be inhaled or ingested.

Members of the genus *Staphylococcus*—the **staphylococci** (Figure 24.6)—are abundant on the human body surface; they are responsible for boils and many other skin problems. *Staphylococcus aureus* is the best-known human pathogen in this genus; it is present in 20% to 40% of normal adults (and in 50%–70% of hospitalized adults). In addition to causing skin diseases, *S. aureus* can cause respiratory, intestinal, and wound infections.

Another interesting group of firmicutes, the **mycoplasmas**, lack cell walls, although some have a stiffening material outside the cell membrane. The mycoplasmas are among the smallest cellular organisms known (Figure 24.7). The smallest mycoplasmas have a diameter of about 0.2 μm. They are small in another crucial sense as well: they have less than half as much DNA as most other prokaryotes. It has been speculated that the DNA in a mycoplasma, which codes for fewer than 500 proteins, may be close to the minimum amount required to encode the essential properties of a living cell.

Actinobacteria include major pathogens as well as valuable sources of antibiotics

Actinobacteria, also known as high-GC Gram-positive bacteria, have a higher ratio of G-C to A-T nucleotide base pairs than do the firmicutes (low-GC Gram-positives). These bacteria develop an elaborately branched system of filaments (Figure 24.8) that resembles the filamentous growth habit of fungi, albeit at a smaller scale. Some actinobacteria reproduce by forming chains of spores at the tips of the filaments. In species that do not form spores, the



Actinomyces sp. 2 μm

Figure 24.8 Actinobacteria Often Produce Branching Filaments The tangled, branching filaments seen in this scanning electron micrograph are typical of this medically important bacterial group.

branched, filamentous growth ceases and the structure breaks up into typical cocci or bacilli, which then reproduce by binary fission.

The actinobacteria include several medically important bacteria. *Mycobacterium tuberculosis* causes tuberculosis, which kills 2 million people each year. Genetic data suggest that this bacterium may have been infecting our ancestors for almost 3 million years, making it the oldest known human bacterial pathogen. The genus *Streptomyces* produces streptomycin as well as hundreds of other antibiotics. We derive most of our antibiotics from actinobacteria.

Cyanobacteria were the first photosynthesizers

Cyanobacteria, sometimes called blue-green bacteria because of their pigmentation, are photosynthetic. They use chlorophyll *a* for photosynthesis and release oxygen gas (O_2); many species also fix nitrogen (discussed in Key Concept 24.3). The production of oxygen by these bacteria transformed the atmosphere of early Earth, eventually leading to the oxygen-rich atmosphere we know today.

Cyanobacteria carry out the same type of photosynthesis that is characteristic of eukaryotic photosynthesizers. They contain elaborate and highly organized internal membrane systems called **photosynthetic lamellae**. As mentioned in Key Concept 24.1, the chloroplasts of photosynthetic eukaryotes are derived from an endosymbiotic cyanobacterium. So all photosynthesis is either directly from cyanobacteria, or from eukaryotes with chloroplasts that are derived from cyanobacteria.

Cyanobacteria may live free as single cells or associate in multicellular colonies. Depending on the species and on growth conditions, these colonies may range from flat sheets one cell thick to filaments to spherical balls of cells. Some filamentous colonies of cyanobacteria differentiate into three specialized cell types: vegetative cells, spores, and heterocysts (Figure 24.9). **Vegetative cells** photosynthesize, **spores** are resting stages that can survive harsh environmental conditions and eventually develop into new filaments, and **heterocysts** are cells specialized for nitrogen fixation. All of the known cyanobacteria with heterocysts fix nitrogen. Heterocysts also have a role in reproduction: when filaments break apart to reproduce, the heterocyst may serve as a breaking point.

Spirochetes move by means of axial filaments

Spirochetes are Gram-negative, motile bacteria characterized by a unique structure known as an axial filament, which is composed of multiple internal flagella running through the periplasmic space (Figure 24.10A). The cell body is a long cylinder coiled into a helix. The internal flagella begin at either end of the cell and overlap in the middle. Motor proteins connect the axial filament to the cell wall, enabling the corkscrew-like movement of the bacterium. Many spirochetes are parasites of humans; a few are pathogens, including those that cause syphilis (Figure 24.10B) and Lyme disease. Others live free in mud or water.

Chlamydias are extremely small parasites

Chlamydias are among the smaller bacteria (0.2–1.5 μm in diameter). They are obligate parasites—that is, they can live only as parasites in the cells of other organisms. It was once believed that their obligate parasitism resulted from an inability to produce ATP—that chlamydias were "energy parasites." However, genome sequencing indicates that chlamydias have the genetic capacity to produce at least some ATP. They can augment this capacity by using an enzyme called a translocase, which allows them to take up ATP from the cytoplasm of their host in exchange for ADP from their own cells.

These tiny, Gram-negative cocci are unique among prokaryotes because of a complex life cycle that involves two different

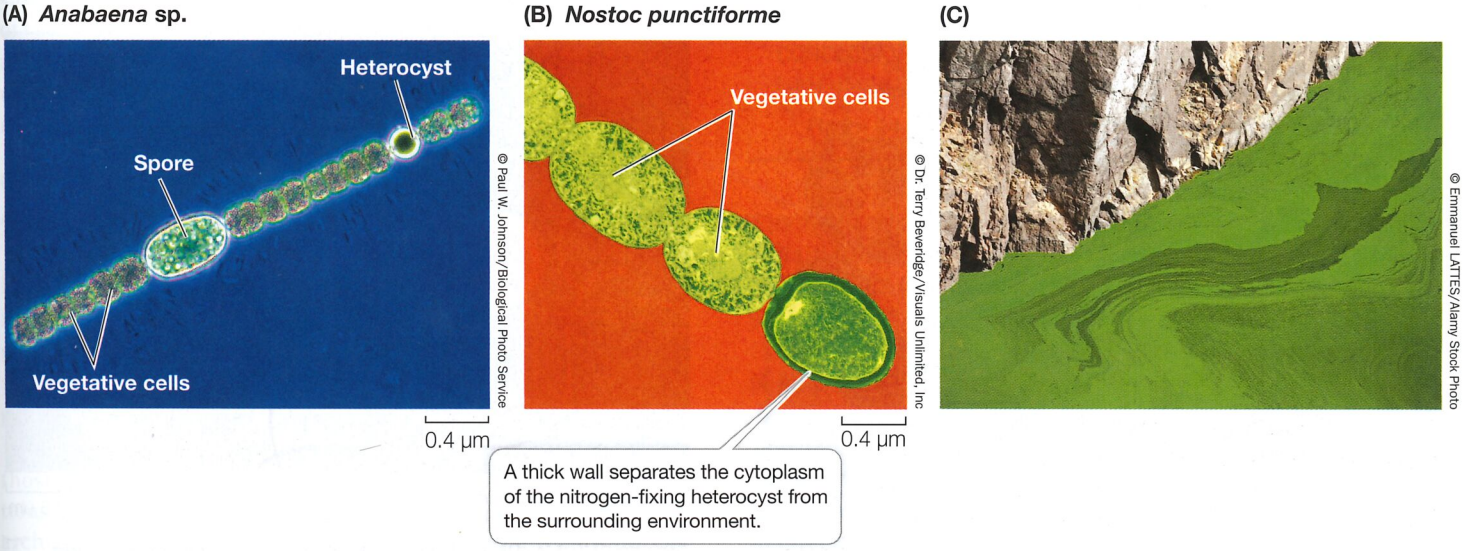


Figure 24.9 Cyanobacteria **(A)** Some cyanobacteria form filamentous colonies containing three cell types. **(B)** Heterocysts are specialized for nitrogen fixation and may serve as a breaking point when filaments reproduce. **(C)** This pond in Canada has experienced eutrophication: phosphorus and other nutrients generated by human activity have accumulated, feeding an immense green mat (commonly referred to as "pond scum") that is made up of several species of free-living cyanobacteria.

Media Clip 24.1 Cyanobacteria
Life12e.com/mc24.1

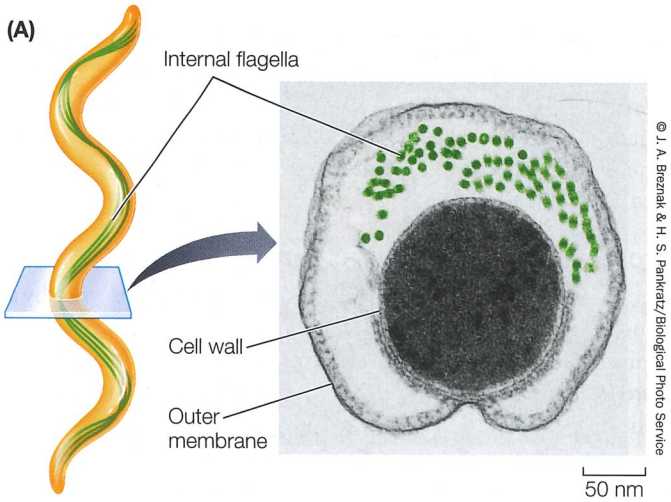


Figure 24.10 Spirochetes Get Their Shape from Axial Filaments **(A)** A spirochete from the gut of a termite, seen in cross section, shows the internal flagella that compose the axial filament, which these helical prokaryotes use to produce a corkscrew-like movement. **(B)** This spirochete species causes syphilis in humans.

forms of cells, elementary bodies and reticulate bodies (Figure 24.11). Various strains of chlamydias cause eye infections (especially trachoma), sexually transmitted diseases, and some forms of pneumonia in humans.

The proteobacteria are a large and diverse group

By far the largest bacterial group, in terms of numbers of described species, is the **proteobacteria**. The proteobacteria include many species of Gram-negative photoautotrophs (see Key Concept 24.3) that use light-driven reactions to metabolize sulfur, as well as dramatically diverse bacteria that bear no phenotypic resemblance to the photoautotrophic species. Genetic and morphological evidence indicates that the mitochondria of eukaryotes were derived from a proteobacterium by endosymbiosis.

Among the proteobacteria are some nitrogen-fixing genera, such as *Rhizobium*, and other bacteria that contribute to the global nitrogen and sulfur cycles. *Escherichia coli*, one of the most studied organisms on Earth, is a proteobacterium. So, too, are many of the most famous human pathogens, such as *Yersinia pestis* (which causes bubonic plague), *Vibrio cholerae* (cholera), and *Salmonella typhimurium* (gastrointestinal disease) (Figure 24.12).

The bioluminescent *Vibrio* shown at the opening of this chapter are also members of this group. There are many potential applications of the genes that encode bioluminescent proteins in

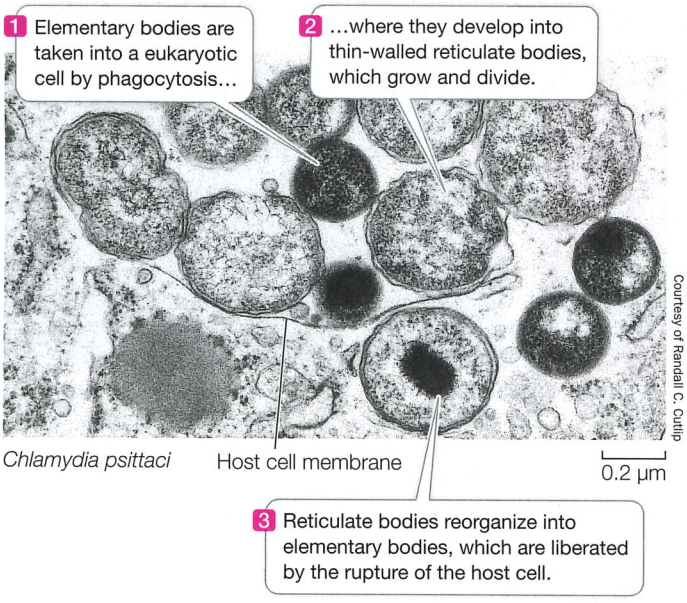


Figure 24.11 Chlamydias Change Form Elementary bodies and reticulate bodies are the two cell forms of the chlamydia life cycle.

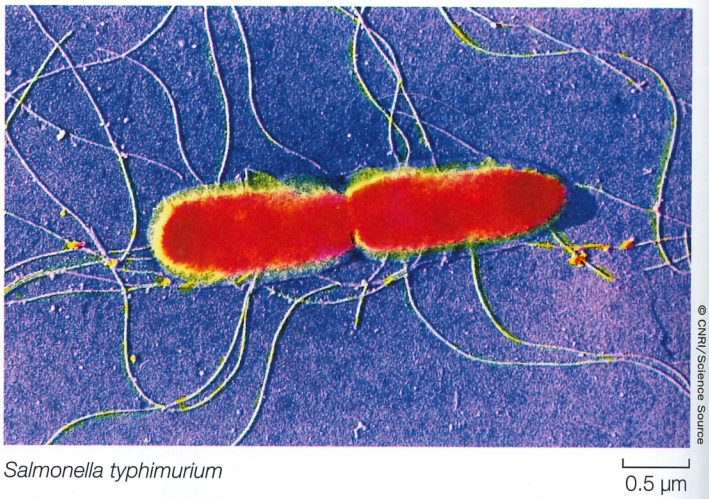
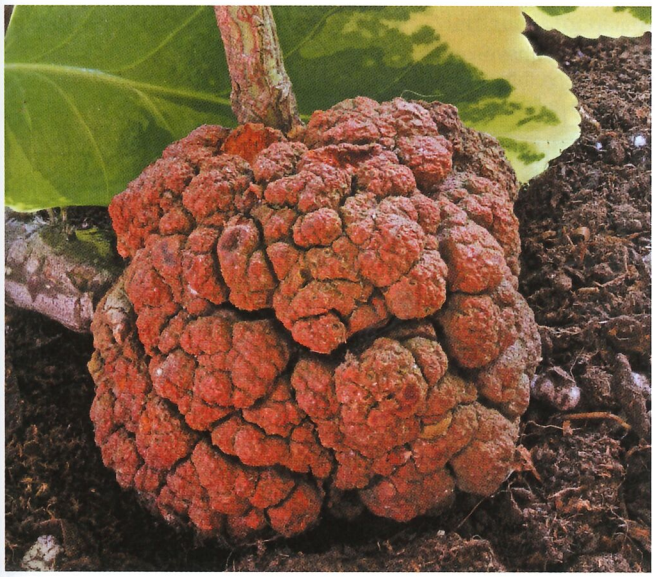


Figure 24.12 Proteobacteria Include Many Familiar Bacteria These conjugating (see Key Concept 12.6) cells of *Salmonella typhimurium* are exchanging genetic material. This pathogen causes a wide range of gastrointestinal illnesses in humans.

Media Clip 24.2 A Swarm of *Salmonella*
Life12e.com/mc24.2



David McIntyre

Euonymus sp.

Figure 24.13 Crown Gall Crown gall, a type of tumor shown here growing on the stem of a *Euonymus* shrub, is caused by the proteobacterium *Agrobacterium tumefaciens*.

bacteria. Already, these genes are being inserted into the genomes of other species in which the resulting bioluminescence is used as a marker of gene expression. Futuristic proposals for making use of bioluminescence in bioengineered organisms include crop plants that glow when they become water-stressed and need to be irrigated, and glowing trees that could light highways at night in place of electric lights.

Although fungi cause most plant diseases, and viruses cause others, about 200 known plant diseases are of bacterial origin. Crown gall, with its characteristic tumors (**Figure 24.13**), is one of the most striking. The causal agent of crown gall is *Agrobacterium tumefaciens*, a proteobacterium that harbors a plasmid used in recombinant DNA studies as a vehicle for inserting genes into new plant hosts.

Gene sequencing enabled biologists to differentiate Archaea from Bacteria

The original identification of Archaea as a group distinct from Bacteria was based on phylogenetic relationships determined from rRNA gene sequences. This separation was supported when biologists sequenced the first complete archaeal genome, which consisted of 1,738 genes—more than half of which were unlike any genes ever found in Bacteria.

Archaea are known for living in extreme habitats such as those with high salinity (salt content), low oxygen concentrations, high temperatures, or high or low pH (**Figure 24.14**). Many archaea are not extremophiles, however—they are common in soil, in many aquatic environments, and in the guts of animals, for example.

Recent studies are revealing many new lineages of archaea, a few of which are shown in Figure 24.1: **Euryarchaeota**, **Crenarchaeota**, **Thaumarchaeota**, **Korarchaeota**, and **Lokiarchaeota**.

Experiment

Figure 24.14A What Is the Highest Temperature Compatible with Life?

Original Paper: K. Kashefi and D. R. Lovley. 2003. Extending the upper temperature limit for life. *Science* 301: 934.

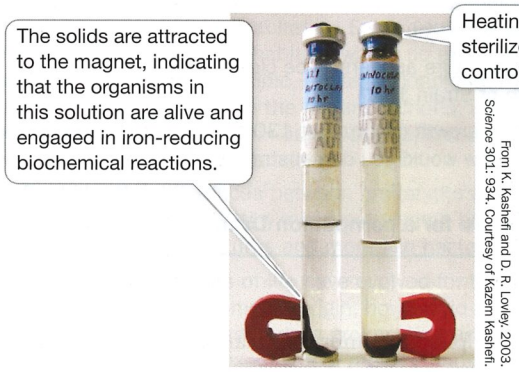
Can any organism thrive at temperatures above 120°C? This is the temperature used for sterilization, known to destroy all previously described organisms. Kazem Kashefi and Derek Lovley isolated an unidentified prokaryote from water samples taken near a hydrothermal vent and found it survived and even multiplied at 121°C. The organism was dubbed "Strain 121," and its gene sequencing results indicate that it is an archaeal species.

HYPOTHESIS▶ Some prokaryotes can survive at temperatures above 120°C.

METHOD▶

1. Seal samples of unidentified, iron-reducing, thermal vent prokaryotes in tubes with a medium containing Fe^{3+} as an electron acceptor. Control tubes contain Fe^{3+} but no organisms.
2. Hold both tubes in a sterilizer at 121°C for 10 hours. If the iron-reducing organisms are metabolically active, they will reduce the Fe^{3+} to Fe^{2+} (as magnetite, which can be detected with a magnet).

RESULTS▶



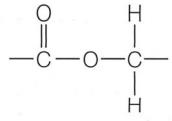
From K. Kashefi and D. R. Lovley. 2003. *Science* 301: 934. Courtesy of Kazem Kashefi.

CONCLUSION▶ Archaea of Strain 121 can survive at temperatures above the previously defined sterilization limit.

Figure 24.14B Work with the Data follows on next page.

The Lokiarchaeota were the first discovered members of the **Asgard archaea**—a group of particular interest because recent studies of their genomes show that they include the closest known prokaryotic relatives of eukaryotes.

Two characteristics shared by all archaea are the absence of peptidoglycan in their cell walls and the presence of lipids of distinctive composition in their cell membranes (see Table 24.1). The unusual lipids in the membranes of archaea are found in no bacteria or eukaryotes. Most lipids in bacterial and eukaryotic membranes contain unbranched long-chain fatty acids connected to glycerol molecules by **ester linkages**:



Work with the Data

FIGURE 24.14B What Is the Highest Temperature Compatible with Life?

Original Paper: K. Kashefi and D. R. Lovley. 2003. Extending the upper temperature limit for life. *Science* 301: 934.

After Kashefi and Lovley isolated Strain 121, they examined its growth at various temperatures. The table shows generation time (time between cell divisions) at nine temperatures.

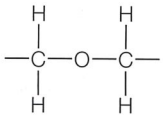
Temperature (°C)	Generation time (hr)
85	10
90	4
95	3
100	2.5
105	2
110	4
115	6
120	20
130	No growth, but cells not killed

QUESTIONS▶

1. Make a graph from these data showing generation time as a function of temperature.
2. Which temperature appears to be closest to the optimum for the growth of Strain 121?
3. Note that no growth occurred at 130°C, but that the cells were not killed. How would you demonstrate that these cells were still alive?

Go to Achieve for a companion Data in Depth exercise.

In contrast, some lipids in membranes of archaea contain long-chain hydrocarbons connected to glycerol molecules by **ether linkages**:



In addition, the long hydrocarbon chains in the lipids of archaea are branched. One class of archaeal lipids contains glycerol at *both* ends of the hydrocarbons (Figure 24.15). These lipids form a lipid monolayer structure that is unique to archaea. They still fit into a biological membrane because they are twice as long as the typical lipids in the bilayers of other membranes. Lipid monolayers and bilayers are both found among the archaea.

Archaea live in extremely diverse environments

Most known crenarchaeotes are either thermophilic, acidophilic (acid loving), or both. Members of the genus *Sulfolobus* live in hot sulfur springs at temperatures of 70°C to 75°C. They become metabolically inactive at 55°C (131°F). Hot sulfur springs are also extremely acidic. *Sulfolobus* grows best in the range from pH 2 to pH 3, but some members of this genus readily tolerate pH values as low as 0.9. Most acidophilic thermophiles maintain an internal

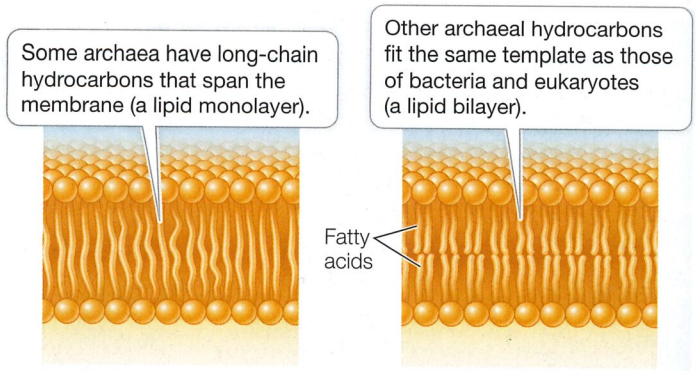


Figure 24.15 Membrane Architecture in Archaea The long-chain hydrocarbons of many archaeal lipids have glycerol molecules at both ends, so that the membranes they form consist of a lipid monolayer. In contrast, the membranes of other archaea, bacteria, and eukaryotes consist of a lipid bilayer.

pH of 5.5 to 7 (close to neutral) in spite of their acidic environment. These and other crenarchaeotes thrive where very few other organisms can even survive (Figure 24.16).

Some species of euryarchaeotes are **methanogens**: they produce methane (CH_4) by reducing carbon dioxide as the key step in their energy metabolism. All of the methanogens are obligate anaerobes (see Key Concept 24.3). Comparison of their rRNA gene sequences has revealed a close evolutionary relationship among these methanogenic species, which were previously assigned to several different groups of bacteria.

Methanogenic euryarchaeotes release approximately 2 billion tons of methane gas into Earth's atmosphere each year, accounting for 80% to 90% of the methane that enters the atmosphere, including that produced in many animal digestive systems (including our own). Approximately one-third of this methane comes from methanogens living in the guts of ruminants such as cattle, sheep, and deer, and another large fraction comes from methanogens living

Figure 24.16 Some Crenarchaeotes Like It Hot Thermophilic crenarchaeotes can thrive in the intense heat of volcanic hot sulfur springs such as these in Yellowstone National Park.

in the guts of termites and cockroaches. Methane is increasing in Earth's atmosphere by about 1% per year and contributes to the greenhouse effect. Part of that increase is due to increases in cattle and rice production and the methanogens associated with both.

Another group of euryarchaeotes, the **extreme halophiles** (salt lovers), live exclusively in very salty environments. Because they contain pink carotenoid pigments, these archaea are sometimes easy to see (Figure 24.17). Extreme halophiles grow in the Dead Sea and in brines of all types. The reddish pink spots that can occur on pickled fishes are colonies of halophilic archaea. Few other organisms can live in the saltiest homes that the extreme halophiles occupy—most would "dry" to death, losing too much water to the hypertonic environment. Extreme halophiles have been found in lakes with pH values as high as 11.5. These are the most alkaline environments inhabited by living organisms, and almost as alkaline as household ammonia.

Some of the extreme halophiles have a unique system for trapping light energy and using it to form ATP—without using any form of chlorophyll—when oxygen is in short supply. They use the pigment retinal (also found in the vertebrate eye) combined with a protein to form a light-absorbing molecule called bacteriorhodopsin.

Another member of Euryarchaeota, *Thermoplasma*, has no cell wall. It is thermophilic and acidophilic, its metabolism is aerobic, and it lives in coal deposits. Its genome of 1,100,000 base pairs is among the smallest (along with that of the mycoplasmas) found in any free-living organism, although some parasitic organisms have even smaller genomes.

Many of the known archaea are crenarchaeotes or euryarchaeotes, but studies of extreme environments have identified several lineages that are not closely related to either of these major groups. For example, the korarchaeotes are known only from DNA isolated directly from hot environments. The thaumarchaeotes were originally found in hot environments as well but have since been found to be common in marine surface waters, where they oxidize ammonia and appear to play an important role in the nitrogen cycle.

The lokiarchaeotes were discovered in 2015 by sequencing environmental samples from near a hydrothermal vent called Loki's Castle deep in the Arctic Ocean. An organism's genome was detected and sequenced from this deep-sea sample. Sequence analysis revealed that the organism was a distinct lineage of archaea, which was named *Lokiarchaeum*. Of particular interest was the finding that the *Lokiarchaeum* genome contains a large number of genes with cell membrane–related functions—genes and functions that had previously been known to occur only in eukaryotes.

After the discovery of *Lokiarchaeum*, biologists began to look for related archaea in other aquatic environments (both marine and freshwater) and discovered many new lineages of archaea with affinities to eukaryotes. Together, these close relatives of eukaryotes are known as the Asgard archaea (named after Asgard, one of the nine worlds of Norse mythology). The genomes of the Asgard archaea include many additional genes that were formerly thought to be restricted to eukaryotes, including genes that function in eukaryotes in controlling cell shape and cytoskeleton formation. These findings suggest that some of the properties that have long been associated exclusively with eukaryotes probably first arose within the Asgard archaea.

© iStock.com/Nancy Nehring

Figure 24.17 Extreme Halophiles Highly saline environments such as these commercial seawater evaporating ponds in San Francisco Bay are home to extreme halophiles. The archaea are easily visible here because of the rich red coloration of their carotenoid pigments.

KEY CONCEPT

24.2 Recap and Assess

Bacteria and archaea are highly diverse groups that survive in almost every imaginable habitat on Earth. Many prokaryotes can survive and even thrive in habitats where no eukaryotes can live, including extremely hot, acidic, or saline conditions. Eukaryotes are most closely related to the Asgard archaea, although endosymbioses of bacteria within eukaryotic cells contributed to the evolution of eukaryotic organelles.

1. Consider the differences between prokaryotes and eukaryotes shown in Table 24.1. Why might eukaryotes be more like archaea in some features and more like bacteria in others?
2. Given that all species of life have evolved for the same amount of time since their common origin, how would you respond to someone who characterizes prokaryotes as "primitive"? Include at least two examples of major groups of prokaryotes to support your answer.
3. How have bacteria changed Earth's atmosphere over the past 3 billion years?
4. What two lines of evidence can you use to support the idea that eukaryotes evolved from a lineage of the Asgard archaea?

Prokaryotes are found almost everywhere on Earth and live in a wide variety of ecosystems. In the next key concept we examine the contributions of prokaryotes to the functioning of those ecosystems.

KEY CONCEPT

24.3 Ecological Communities Depend on Prokaryotes

Learning Objectives

24.3.1 Describe what biofilms are, how they are formed, and why they present problems for humans.

24.3.2 Explain how bacteria communicate through quorum sensing.

24.3.3 Explain the critical role of prokaryotes in the nitrogen cycle.

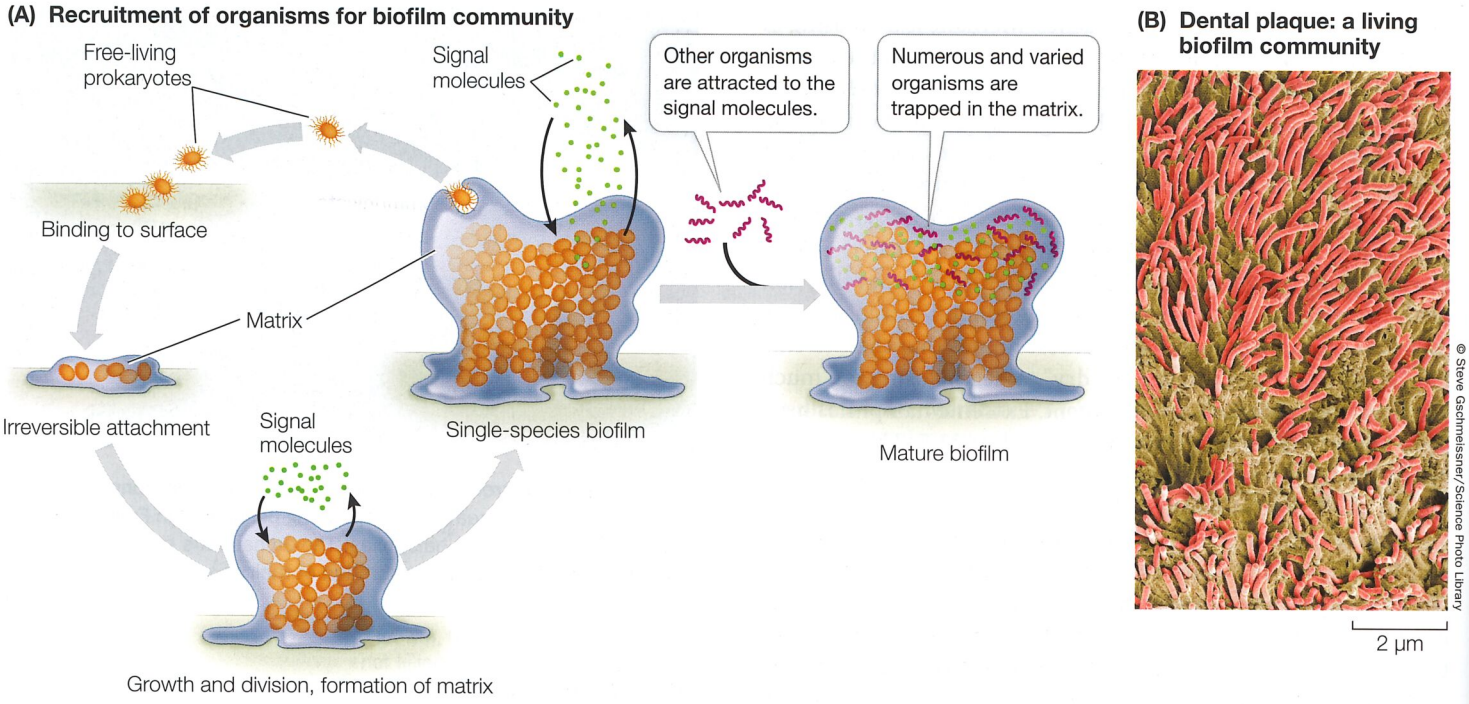


Figure 24.18 Forming a Biofilm **(A)** Free-living prokaryotes readily attach themselves to surfaces and form films that are stabilized and protected by a surrounding matrix. Once the population is large enough, the developing biofilm can send out chemical signals that attract other microorganisms. **(B)** Scanning electron micrography reveals a biofilm of dental plaque. The bacteria (red) are embedded in a matrix consisting of proteins from both bacterial secretions and saliva. (A after D. Monroe. 2007. *PLOS Biol* 5: e307.)

Prokaryotic cells do not usually live in isolation. Rather, they live in communities of many different species, often including microscopic eukaryotes. Whereas some microbial communities are harmful to humans, others provide important services. They help us digest our food, break down municipal waste, and recycle organic matter and chemical elements in the environment.

Many prokaryotes form complex communities

Some microbial communities form layers in sediments, and others form clumps a meter or more in diameter. Many microbial communities tend to form dense **biofilms**. Upon contacting a solid surface, the cells bind to that surface and secrete a sticky, gel-like polysaccharide matrix that traps other cells (Figure 24.18). Once a biofilm forms, the cells become more difficult to kill.

Biofilms are found in many places, and in some of those places they cause problems for humans. The material on our teeth that we call dental plaque is a biofilm. Pathogenic bacteria are difficult for the immune system—and modern medicine—to combat once they form a biofilm, which may be impermeable to antibiotics. Biofilms may form on just about any available surface, including contact lenses and artificial joint replacements. They foul metal pipes and cause corrosion, a major problem in steam-driven electricity generation plants. Fossil stromatolites—large, rocky structures made up of alternating layers of fossilized biofilm and calcium carbonate—are among the oldest remnants of life on Earth (see Figure 23.9A).

Some biologists are studying the chemical signals that prokaryotes use to communicate with one another and that trigger density-linked activities such as biofilm formation. One example of this type of communication—called **quorum sensing**—is found in the bioluminescent *Vibrio* that are shown in the opening photo of this chapter. How does quorum sensing work? As demonstrated in **Investigating Life: How Do Bacteria Communicate with One Another?**, individual *Vibrio* bacteria can excrete a signal that is detected by other individuals, and this signal then functions to turn on the genes that produce luciferase—an enzyme that produces bioluminescence when it is active.

Microbiomes are critical to the health of many eukaryotes

Although only a few bacterial species are pathogens, popular notions of bacteria as "germs" and fear of the consequences of infection cause many people to assume that most bacteria are harmful. Increasingly, however, biologists are discovering that the health of humans (as well as that of most other eukaryotes) depends in large part on the health of our **microbiomes**: the communities of bacteria and archaea that live in and on our bodies. Other communities of microbes live in close association with other multicellular organisms.

Every surface of your body is covered with diverse communities of bacteria (Figure 24.19). A recent study identified more than 1,000 species of bacteria that live on human skin. Inside your body, your digestive system teems with bacteria. When these communities are disrupted, they must be restored before the body can function normally.

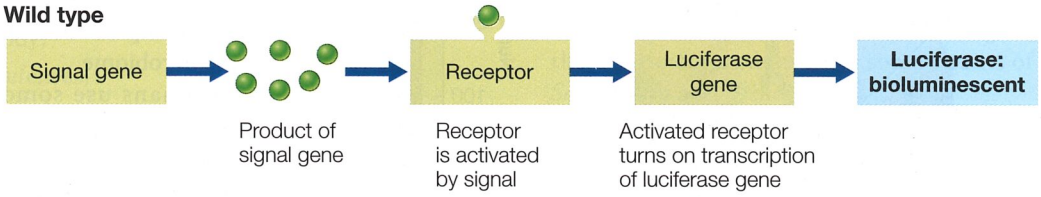
Biologists are discovering that many complex health problems are linked to the disruption of our microbiomes. These diverse

InvestigatingLIFE How Do Bacteria Communicate with One Another?

Experiment

Original Paper: M. B. Miller and B. L. Bassler. 2001. Quorum sensing in bacteria. *Annu Rev Microbiol* 55: 165–199.

Bonnie Bassler and her colleagues at Princeton University investigated how *Vibrio fischeri* bacteria communicate with one another. These bacteria produce bioluminescence when they are present in sufficiently high densities. In a normal *V. fischeri* bacterium, the following pathway produces bioluminescence when a bacterial colony becomes dense enough to produce sufficient signal:

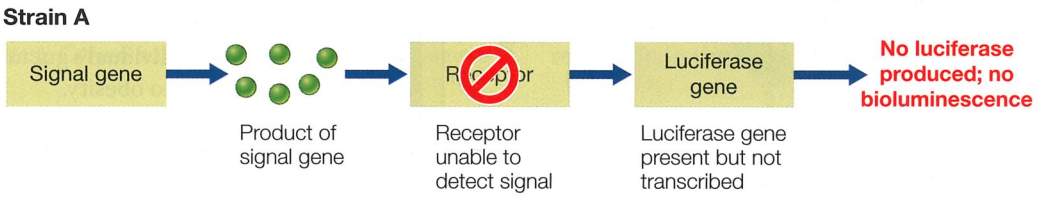


The fact that the bacteria emit light only when they are present in high densities suggests that the signal is used to communicate among nearby bacteria, alerting one another to their presence. But how can we tell that the signal produced by one bacterium is being received by another?

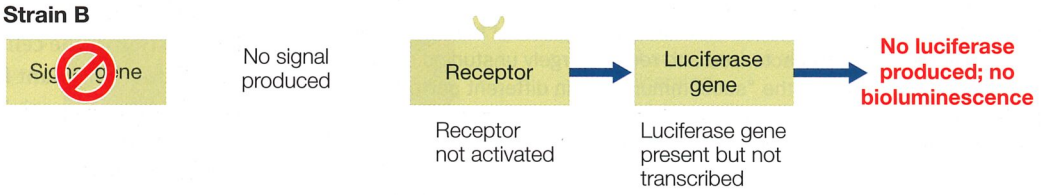
HYPOTHESIS▶ *V. fischeri* can transmit the signal molecule from one individual to another, thus communicating about the presence of other nearby bacteria.

METHOD▶

1. Select two mutant strains of *V. fischeri* incapable of emitting light on their own. In strain A, a mutation for the signal receptor renders the bacteria unable to detect the signal molecule, so they do not bioluminesce:

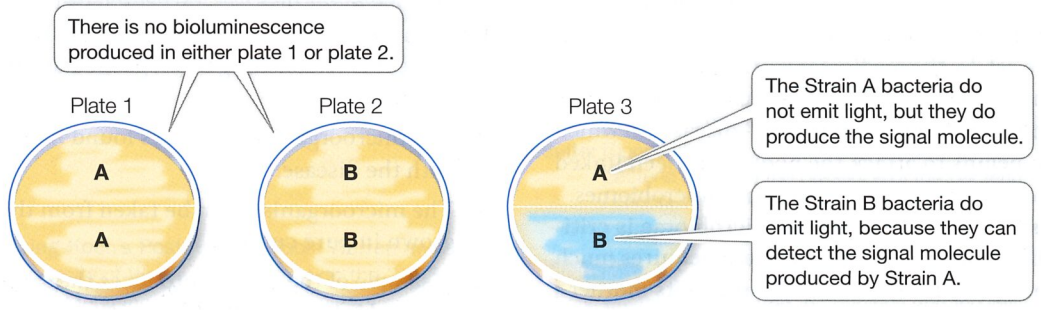


2. Strain B bacteria have a mutated signal gene, so they do not produce a signal molecule, although the receptor and the luciferase gene are normal:



3. Incubate samples of *V. fischeri* on agar petri plates, as follows:
   - Plate 1: Strain A only
   - Plate 2: Strain B only
   - Plate 3: Half of plate with strain A, half with strain B

RESULTS▶



CONCLUSION▶ *Vibrio* bacteria can detect the signal molecule produced by other nearby bacteria *and* use this information to detect the presence of other bacteria.

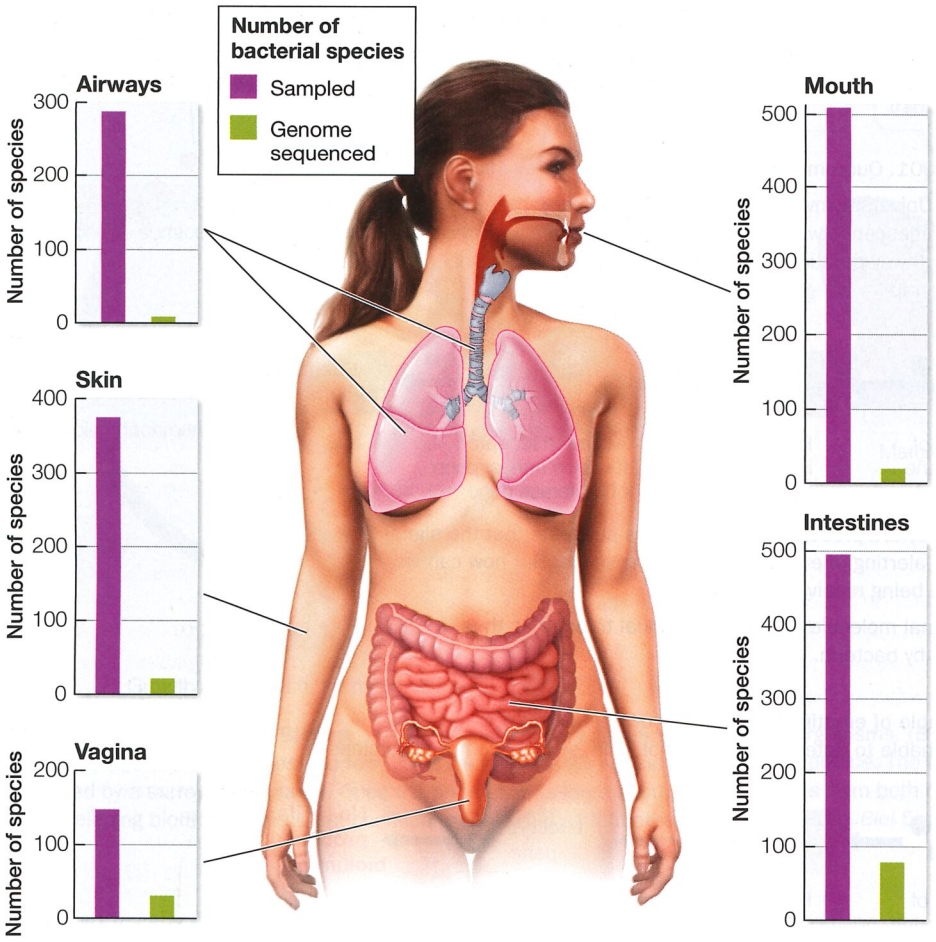


Figure 24.19 The Body's Microbiome Is Critical to the Maintenance of Health Surveys of the human microbiome have shown that this community includes thousands of diverse bacterial species that are adapted to grow in or on various parts of the body. Although we now know that the composition of this microbiome is closely associated with many aspects of human health, most of the component species are poorly characterized and remain largely unstudied by biologists. What has become clear is that, although the "subcommunities" in different parts of the body share similarities, each is a site-specific assemblage of many distinctive species. (After Y. K. Lee and S. K. Mazmanian. 2010. *Science* 330: 1768–1773.)

microbial communities affect the expression of our genes and play a critical role in the development and maintenance of a healthy immune system. When our microbiomes contain an appropriate community of beneficial species, our bodies function normally. But these communities are strongly affected by our life experiences, by the food we eat, by the medicines we take, and by our exposure to various environmental toxins. The recent rapid increase in the rate of autoimmune diseases in humans—diseases in which the immune system begins to attack the body—has been linked to the changing diversity and composition of our microbiomes.

The early acquisition of an appropriate microbiome is critical for lifelong health. Normally, a human infant acquires much of its microbiome at birth, from the microbiome in its mother's vagina. Other components of the microbiome are also acquired from the mother, especially through breast feeding. Recent studies have shown that babies born by cesarean section, as well as babies that are bottle-fed on artificial milk formula, typically acquire microbes from a wider variety of sources. Many of the bacteria acquired in this way are not well suited for human health. Biologists have discovered that the incidence of many autoimmune diseases is much higher in people who were born by cesarean section and in those who were fed on formula as infants, compared with individuals who were born vaginally and breast-fed as infants. The difference appears to be related to the composition of the individual's original microbiome.

Humans use some of the metabolic products—especially vitamins B_{12} and K—produced by the microbiome living in the large intestine. Communities of bacteria line our intestines with a dense biofilm that is in intimate contact with the mucosal lining of the gut. This biofilm facilitates nutrient transfer from the intestine into the body, functioning like a specialized "tissue" that is essential to our health. This biofilm has a complex ecology that scientists have just begun to explore in detail—including the possibility that the species composition of an individual's gut microbiome may contribute to obesity.

Animals harbor a variety of microbes in their digestive tracts, many of which play important roles in digestion. Cattle depend on prokaryotes to break down plant material. Like most animals, cattle cannot produce cellulase, the enzyme needed to start the digestion of the cellulose that makes up the bulk of their plant food. However, bacteria living in a special section of the gut, called the rumen, produce enough cellulase to process the daily diet for the cattle.

A small minority of bacteria are pathogens

The late nineteenth century was a productive era in the history of medicine—a time when bacteriologists, chemists, and physicians proved that many diseases are caused by microbial agents. During this time, the German physician Robert Koch laid down a set of four rules for establishing that a particular microorganism causes a particular disease:

1. The microorganism is always found in individuals with the disease.
2. The microorganism can be taken from the host and grown in pure culture.
3. A sample of the culture produces the same disease when injected into a new, healthy host.
4. The newly infected host yields a new, pure culture of microorganisms identical to those obtained in the second step.

These rules, called **Koch's postulates**, were important tools in a time when it was not widely understood that microorganisms cause disease. Although modern medical science has more powerful diagnostic tools, Koch's postulates remain useful. For example, physicians were taken aback in the 1980s when stomach ulcers—long accepted and treated as the result of excess stomach acid—were shown by Koch's postulates to be caused by the bacterium *Helicobacter pylori* (Figure 24.20).

For an organism to be a successful pathogen, it must:

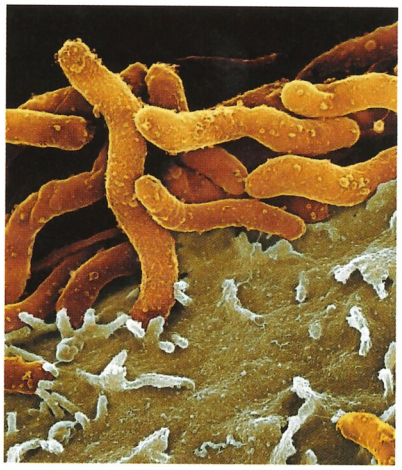
- arrive at the body surface of a potential host;
- enter the host's body;
- evade the host's defenses;
- reproduce inside the host; and
- infect a new host.

Failure to complete any of these steps ends the disease cycle of a pathogenic organism. Yet in spite of the many defenses available to potential hosts, some bacteria are very successful pathogens. Pathogenic bacteria are often surprisingly difficult to combat, even with today's arsenal of antibiotics. One source of this difficulty is their ability to form biofilms.

For the host, the consequences of a bacterial infection depend on several factors. One is the invasiveness of the pathogen: its ability to multiply in the host's body. Another is its toxigenicity: its ability to produce toxins (chemical substances that are harmful to the host's tissues). *Corynebacterium diphtheriae*, the agent that causes diphtheria, has low invasiveness and multiplies only in the throat, but its toxigenicity is so great that the entire body is affected. In contrast, *Bacillus anthracis*, which causes anthrax, has low toxigenicity but is so invasive that the entire bloodstream ultimately teems with the bacteria.

There are two general types of bacterial toxins: exotoxins and endotoxins. **Endotoxins** are released when certain bacteria grow or lyse (burst). Endotoxins are lipopolysaccharides (complexes consisting of a polysaccharide and a lipid component) that form part of the outer bacterial membrane. Endotoxins are rarely fatal to the host; they normally cause fever, vomiting, and diarrhea. Among the endotoxin producers are some strains of the proteobacteria *Salmonella* and *Escherichia*.

Exotoxins are soluble proteins released by living, multiplying bacteria. They are highly toxic—often fatal—to the host. Human diseases induced by bacterial exotoxins include tetanus (*Clostridium tetani*), cholera (*Vibrio cholerae*), and bubonic plague (*Yersinia pestis*). Anthrax is caused by three exotoxins produced by *Bacillus anthracis*. Botulism is caused by exotoxins produced by *Clostridium botulinum*; these exotoxins are among the most poisonous ever discovered. The lethal dose for humans of one exotoxin of *C. botulinum* is about one-millionth of a gram. Nonetheless, much smaller doses of this exotoxin, marketed under various trade names (e.g., Botox®), are used to treat muscle spasms and for cosmetic purposes (temporary wrinkle reduction in the skin).



Helicobacter pylori

Marshall and Warren set out to satisfy Koch's postulates:

Test 1

The microorganism must be present in every case of the disease.

Results: Biopsies from the stomachs of many patients revealed that the bacterium was always present if the stomach was inflamed or ulcerated.

Test 2

The microorganism must be cultured from a sick host.

Results: The bacterium was isolated from biopsy material and eventually grown in culture media in the laboratory.

Test 3

The isolated and cultured bacteria must be able to induce the disease.

Results: Marshall was examined and found to be free of bacteria and inflammation in his stomach. After drinking a pure culture of the bacterium, he developed stomach inflammation (gastritis).

Test 4

The bacteria must be recoverable from newly infected individuals.

Results: Biopsy of Marshall's stomach 2 weeks after he ingested the bacteria revealed the presence of the bacterium, now christened *Helicobacter pylori*, in the inflamed tissue.

Conclusion

Antibiotic treatment eliminated the bacteria and the inflammation in Marshall's stomach. The experiment was repeated on healthy volunteers, and many patients with gastric ulcers were cured with antibiotics. Thus Marshall and Warren demonstrated that the stomach inflammation leading to ulcers is caused by *H. pylori* infections in the stomach.

Figure 24.20 Satisfying Koch's Postulates Robin Warren and Barry Marshall of the University of Western Australia won the 2005 Nobel Prize in Physiology or Medicine for showing that ulcers are caused not by the action of stomach acid but by infection with the bacterium *Helicobacter pylori*. (After B. J. Marshall and J. R. Warren. 1983. *Lancet* 321: 1273–1275; B. J. Marshall et al. 1985. *Med J Aust* 142: 436–439.)

Prokaryotes have amazingly diverse metabolic pathways

Archaea and bacteria outdo the eukaryotes in terms of metabolic diversity. Although they are much more diverse in size and shape, eukaryotes draw on fewer metabolic mechanisms for their energy needs. In fact, much of the eukaryotes' energy metabolism is carried out in organelles—mitochondria and chloroplasts—that are endosymbiotic descendants of bacteria. The long evolutionary history of prokaryotes, during which they have had time to explore a wide variety of habitats, has led to the extraordinary diversity of their metabolic "lifestyles"—their use or nonuse of oxygen, their energy sources, their sources of carbon atoms, and the materials they release as waste products.

ANAEROBIC VERSUS AEROBIC METABOLISM Some prokaryotes can live only by anaerobic metabolism because oxygen is poisonous to them. These oxygen-sensitive organisms are called **obligate anaerobes**. By definition, an anaerobe does not use oxygen as an electron acceptor for its respiration. Other prokaryotes, called **facultative anaerobes**, can alternate between an anaerobic mode of metabolism (such as fermentation) and an aerobic mode (such as cellular respiration) as conditions dictate. **Aerotolerant anaerobes** cannot conduct cellular respiration, but they are not damaged by oxygen when it is present. At the other extreme from the obligate anaerobes, some prokaryotes are **obligate aerobes**, unable to survive for extended periods in the *absence* of oxygen. Obligate aerobes require oxygen for cellular respiration.

NUTRITIONAL CATEGORIES All living organisms face the same nutritional challenges: they must synthesize energy-rich compounds such as ATP to power their life-sustaining metabolic reactions, and they must obtain carbon atoms to build their own organic molecules. Biologists recognize four broad nutritional categories of organisms: photoautotrophs, photoheterotrophs, chemoautotrophs, and chemoheterotrophs. Prokaryotes are represented in all four groups (Table 24.2).

Photoautotrophs perform photosynthesis. They use light as their energy source and carbon dioxide (CO_2) as their carbon source. The cyanobacteria, like green plants and other photosynthetic eukaryotes, use chlorophyll *a* as their key photosynthetic pigment and produce oxygen gas (O_2) as a by-product of noncyclic electron transport.

There are other photoautotrophs among the bacteria, but these organisms use bacteriochlorophyll as their key photosynthetic pigment, and they do not produce O_2. Instead, some of these photosynthesizers produce particles of pure sulfur, because hydrogen sulfide (H_2S), rather than H_2O, is their electron donor for photophosphorylation. Many proteobacteria fit into this category. Bacteriochlorophyll molecules absorb light of longer wavelengths than the chlorophyll molecules used by other photosynthesizing organisms. As a result, bacteria using this pigment can grow in water under fairly dense layers of algae, using light of wavelengths that are not absorbed by the algae (Figure 24.21).

Photoheterotrophs use light as their energy source but must obtain their carbon atoms from organic compounds made by other organisms. Their "food" consists of organic compounds such as carbohydrates, fatty acids, and alcohols. For example, compounds released from plant roots (as in rice paddies) or from decomposing photosynthetic bacteria in hot springs are taken up by photoheterotrophs and metabolized to form building blocks for other compounds. Sunlight provides the ATP necessary for metabolism through photophosphorylation.

Chemoautotrophs obtain their energy by oxidizing inorganic substances, and they use some of that energy to fix carbon (convert inorganic carbon into organic carbon-based molecules). Some chemoautotrophs use reactions identical to those of the typical photosynthetic cycle, but others use alternative pathways for carbon fixation. Some bacteria oxidize ammonia or nitrite ions to form nitrate ions. Others oxidize hydrogen gas, hydrogen sulfide, sulfur, and other materials. Many archaea are chemoautotrophs.

Finally, **chemoheterotrophs** obtain both energy and carbon atoms from one or more complex organic compounds that have been synthesized by other organisms. Most known bacteria and archaea are chemoheterotrophs—as are all animals and fungi and many protists.

Although most chemoheterotrophs rely on the breakdown of organic compounds for energy, some chemoheterotrophic prokaryotes obtain their energy by breaking down inorganic substances. Organisms that obtain energy from oxidizing inorganic substances (both chemoautotrophs as well as some chemoheterotrophs) are also known as lithotrophs (Greek, "rock consumers").

TABLE **24.2** | **How Organisms Obtain Their Energy and Carbon**

Nutritional category	Energy source	Carbon source
Photoautotrophs (some bacteria, some eukaryotes)	Light	Carbon dioxide
Photoheterotrophs (some bacteria)	Light	Organic compounds
Chemoautotrophs (some bacteria, many archaea)	Inorganic substances	Carbon dioxide
Chemoheterotrophs (some bacteria, some archaea, most eukaryotes)	Usually organic compounds; sometimes inorganic substances	Organic compounds

Prokaryotes play important roles in element cycling

The metabolic diversity of the prokaryotes makes them key players in the cycles that keep elements moving through ecosystems. Many prokaryotes are decomposers: organisms that metabolize organic compounds in dead organic material and return the products to the environment as inorganic substances. Prokaryotes, along with fungi, return tremendous quantities of carbon to the atmosphere as carbon dioxide, thus carrying out a key step in the carbon cycle.

The key metabolic reactions of many prokaryotes involve nitrogen or sulfur. For example, some bacteria carry out respiratory electron transport without using oxygen as an electron acceptor. These organisms use oxidized inorganic ions such as nitrate, nitrite, or sulfate as electron acceptors.

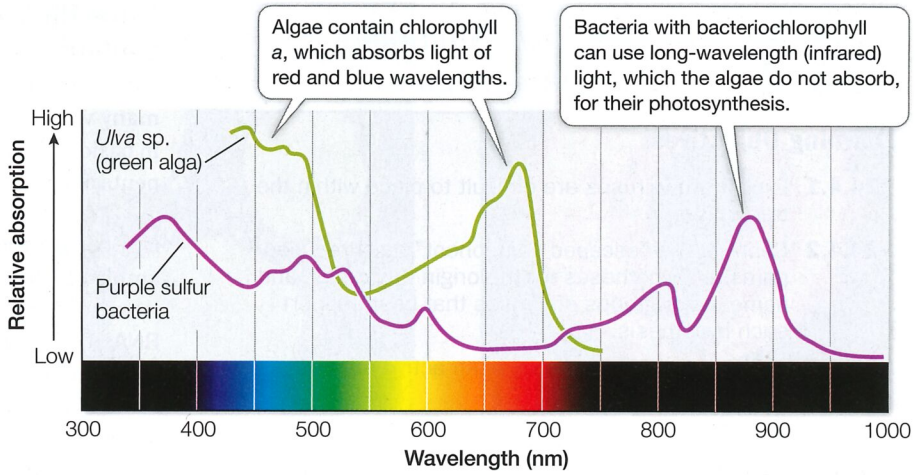


Figure 24.21 Bacteriochlorophyll Absorbs Long-Wavelength Light The green alga *Ulva* contains chlorophyll, which absorbs no light of wavelengths longer than 750 nanometers (nm). Purple sulfur bacteria, which contain bacteriochlorophyll, can conduct photosynthesis using longer infrared wavelengths. As a result, these bacteria can grow under layers of algae. (After F. Haxo and L. R. Blinks. 1950. *J Gen Physiol* 33: 389–422; S. Mehrabi et al. 2001. *Biomol Eng* 18: 49–56.)

Examples include the **denitrifiers**, which release nitrogen to the atmosphere as nitrogen gas (N_2). These normally aerobic bacteria, mostly species of the genera *Bacillus* and *Pseudomonas*, use nitrate (NO_3^-) as an electron acceptor in place of oxygen if they are kept under anaerobic conditions:

$$2\,NO_3^- + 10\,e^- + 12\,H^+ \rightarrow N_2 + 6\,H_2O$$

Denitrifiers play a key role in the cycling of nitrogen through ecosystems. Without denitrifiers, which convert nitrate ions back into nitrogen gas, all forms of nitrogen would leach from the soil and end up in lakes and oceans, making life on land much more difficult.

Nitrogen fixers convert atmospheric nitrogen gas into a chemical form (ammonia) that is usable by the nitrogen fixers themselves as well as by other organisms:

$$N_2 + 6\,H \rightarrow 2\,NH_3$$

All organisms require nitrogen in order to build proteins, nucleic acids, and other important compounds. Nitrogen fixation is thus vital to life as we know it. This all-important biochemical process is carried out by a wide variety of archaea and bacteria (including cyanobacteria) but by no eukaryotes, so we depend on these prokaryotes for our very existence.

Connect the Concepts For descriptions of the role of nitrogen in plant nutrition and in the global nitrogen cycle, see Key Concepts 34.3 and 56.4.

Ammonia is oxidized to nitrate in soil and in seawater by chemoautotrophic bacteria called **nitrifiers**. Bacteria of two genera, *Nitrosomonas* and *Nitrosococcus*, convert ammonia (NH_3) to nitrite ions (NO_2^-), and *Nitrobacter* oxidizes nitrite to nitrate (NO_3^-), the form of nitrogen most easily used by many plants. What do the nitrifiers get out of these reactions? Their metabolism is powered by the energy released by the oxidation of ammonia or nitrite. For example, by passing the electrons from nitrite through an electron transport system, *Nitrobacter* can make ATP and, using some of this ATP, can also make NADH. With this ATP and NADH, the bacterium can convert CO_2 and H_2O into glucose.

We have already seen the importance of the cyanobacteria in the cycling of oxygen: in ancient times, the oxygen generated by their photosynthesis converted Earth's atmosphere from an anaerobic to an aerobic environment (see Key Concept 23.2). Other prokaryotes—both bacteria and archaea—contribute to the cycling of sulfur. Deep-sea hydrothermal vent ecosystems depend on chemoautotrophic prokaryotes that are incorporated into large communities of crabs, mollusks, and giant worms, all living at a depth of 2,500 meters—below any hint of sunlight. These bacteria obtain energy by oxidizing hydrogen sulfide and other substances released in the near-boiling water flowing from volcanic vents in the ocean floor.

KEY CONCEPT

24.3 Recap and Assess

Many prokaryotes are beneficial and even necessary to other forms of life. Most animals, including humans, depend on a complex community of prokaryotes—a microbiome—to maintain health, especially of the immune and digestive systems. Prokaryotes play critical roles in cycling many elements through Earth's ecosystems, including carbon, nitrogen, and oxygen. Some bacteria are pathogenic—the direct causes of diseases. Finding cures for pathogenic diseases entails understanding how pathogens enter and reproduce in the body.

1. How do biofilms form, and why are they of special interest to researchers?
2. Why would elimination of all bacteria from a human gut be problematic from a health standpoint?
3. Why is nitrogen metabolism in prokaryotes vital to other organisms?

Before moving on to discuss the diversity of eukaryotic life, it is appropriate to consider another category of life that includes some pathogens: the viruses. Although they are not cellular, viruses are numerically among the most abundant forms of life on Earth. Their effects on other organisms are enormous. Where did viruses come from, and how do they fit into the tree of life? Biologists are still working to answer these questions.