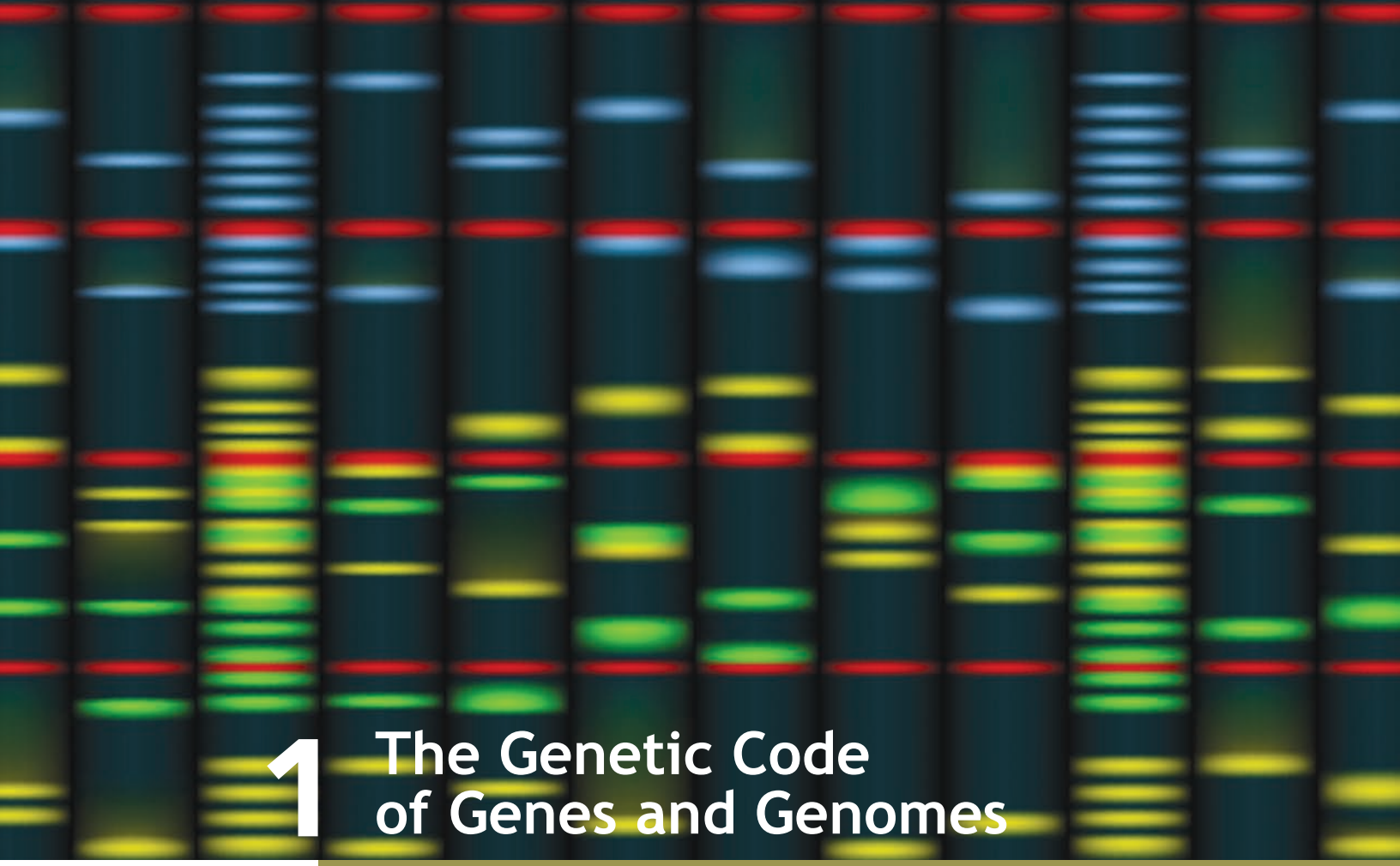


1 The Genetic Code of Genes and Genomes

CHAPTER ORGANIZATION

DNA fingerprints. [© Alila Sao Mai/ ShutterStock, Inc.]

Every science occasionally undergoes a major advance that completely changes perspectives on the field. This happened in physics with the discovery of subatomic particles and in chemistry with the understanding of the nature of the chemical bond. In genetics it has happened most recently with the development of genomics.

Genetics is the study of biologically inherited traits. A related field, **genomics**, is the study of all the genes in an organism to understand their molecular organization, function, interaction, and evolutionary history. Each species of living organism is united by a common set of inherited traits that set it apart from all other species of organisms. For example, a human being habitually stands upright and has long legs, relatively little body hair, a large brain, and a flat face with a prominent nose, jutting chin, distinct lips, and small teeth. Some traits present in human beings are shared with other animals to whom we are more distantly related. In common with other mammals, human beings are warm-blooded, and human mothers feed their young with milk secreted by mammary glands. In common with other vertebrates, human beings have a backbone and a spinal cord. Every normal human being exhibits these biological characteristics, and all of these traits are inherited.

The fundamental concept of genetics and genomics is that

KEY CONCEPT

Inherited traits are determined by the elements of heredity that are transmitted from parents to offspring in reproduction; these elements of heredity are called **genes**.

What makes genomics so important to the study of genetics is that it represents significant advances in discovering genes and analyzing their functions. One noteworthy application of genomics led to the development of methods to determine the complete sequence of the constituents that make up the DNA in an organism. A second noteworthy application provided the means to study the patterns of expression of all of an organism's genes simultaneously, to learn which groups of genes act together to result in normal function or disease.

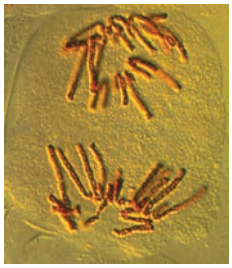
Genomics is merely the latest in a continuing series of major advances in *molecular genetics*, the study of the chemical nature of genes and of how genes function to affect the traits of living organisms. This chapter serves as an introduction to molecular genetics. First we discuss key experiments showing that DNA is the genetic material. Then we provide an overview of how DNA is duplicated in going from one cell generation to the next, how it functions to determine the chemical makeup of enzymes and other proteins in the cell, and how it undergoes mutation to produce defective proteins that are often associated with inherited diseases. Later in this chapter we return to genomics and give an overview of the extent to which living organisms share similar proteins and cellular processes.

The existence of genes and the rules governing their transmission from generation to generation were discovered by Gregor Mendel in the 1860s. His work with garden peas represents the beginning of what would become the science of genetics. Mendel's formulation of inheritance was in terms of the abstract rules by which hereditary elements are transmitted from parents to offspring. The approach to the study of genetics through the analysis of offspring from matings is sometimes referred to as *classical* genetics.

Molecular genetics got its start only three years after Mendel reported his experiments. In 1869 Friedrich Miescher discovered a new type of weak acid, abundant in the nuclei of white blood cells, that turned out to be the chemical substance of which genes are made. Miescher's weak acid is now called **deoxyribonucleic acid** or **DNA**. Nevertheless, even though the two main pieces of the puzzle of heredity—genes and DNA—had been discovered, the pieces were not put together until about the middle of the twentieth century when the chemical identity between genes and DNA was conclusively demonstrated. The next section shows how this connection was made.

1.1 DNA is the molecule of heredity.

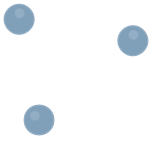
The importance of the cell nucleus in inheritance became clear in the 1870s when the nuclei of the male and female reproductive cells were observed to fuse in the process of fertilization. The next major advance was the discovery of **chromosomes**, thread-like objects inside the nucleus that become visible in the light microscope when stained with certain dyes. Chromosomes exhibit a characteristic "splitting" behavior, in which each daughter cell formed by cell division receives an identical complement of chromosomes. More evidence for the importance of chromosomes was provided by the observation that, whereas the number of chromosomes in each cell differs from one biological species to the next, the number of chromosomes is nearly always constant within the cells of any particular species. These features of chromosomes were well understood by about 1900, and they made it seem likely that chromosomes were the carriers of the genes.



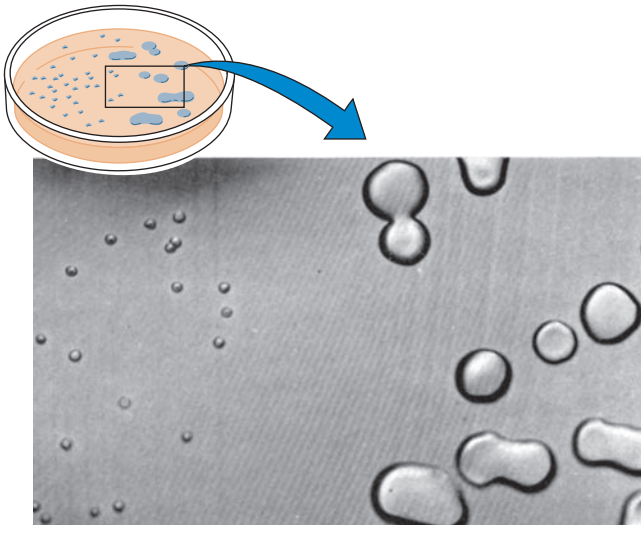
By the 1920s, several lines of indirect evidence suggested a close relationship between chromosomes and DNA. Microscopic studies with special stains showed that DNA is present in chromosomes. Various types of proteins are present in chromosomes too. But whereas most of the DNA in cells of higher organisms is present in chromosomes, and the amount of DNA per cell is constant, the amount and kinds of proteins and other large molecules differ greatly from one type of cell to another. The indirect evidence for DNA as the genetic material was unconvincing, because crude chemical analyses had suggested (erroneously, as it turned out) that DNA lacked the chemical diversity needed in a genetic substance. The favored candidate for the genetic material was protein, because proteins were known to be an exceedingly diverse collection of molecules. Proteins therefore became widely accepted as the genetic material, and DNA was thought to provide only the structural framework of chromosomes. Any researcher who hoped to demonstrate that DNA was the genetic material had a double handicap. Such experiments had to demonstrate not only that DNA *is* the genetic material but also that proteins are *not* the genetic material. Some of the experiments regarded as decisive in implicating DNA are described in this section.

Genetic traits can be altered by treatment with pure DNA.

One type of bacterial pneumonia in mammals is caused by strains of *Streptococcus pneumoniae* able to synthesize a gelatinous capsule composed of polysaccharide (complex carbohydrate). This capsule surrounds the bacterium and protects it from the defense mechanisms of the infected animal; thus it enables the bacterium to cause disease. When a bacterial cell is grown on solid medium, it undergoes repeated cell divisions to form a visible clump of cells called a **colony**. The enveloping capsule makes the size of each colony large and gives it a glistening or smooth (S) appearance (**FIGURE 1.1**). Certain strains of *S. pneumoniae*, however, are unable to synthesize the capsular polysaccharide, and they form small colonies that have a rough (R) surface. The R strains do not cause pneumonia; lacking the capsule, these bacteria are inactivated by the immune system of the host. Both types of bacteria "breed true" in the sense that the progeny formed by cell division have the capsular type of the parent, either S or R.



When mice are injected with living R cells, or with S cells that have been killed with extreme heat, the animals remain healthy. However, in 1928 Frederick Griffith showed that when mice are injected with a *mixture* of living R cells and heat-killed S cells, they often die of pneumonia (**FIGURE 1.2**). Bacteria isolated from blood samples of the dead mice produce S cultures with a capsule typical of the injected S cells, even though the injected S cells had been killed by heat. Evidently, the injected material from the dead S cells includes a substance that can enter living R



R colonies. breed true S colonies. breed true

FIGURE 1.1 Colonies of *Streptococcus pneumoniae*. The small colonies on the left are from a rough (R) strain, and the large colonies on the right are from a smooth (S) strain. The S colonies are larger because of the capsule on the S cells. [© Averry et al., Originally published in ***The Journal of Experimental Medicine***, 79; 137–158. Copyright 1944, Rockefeller University Press.]

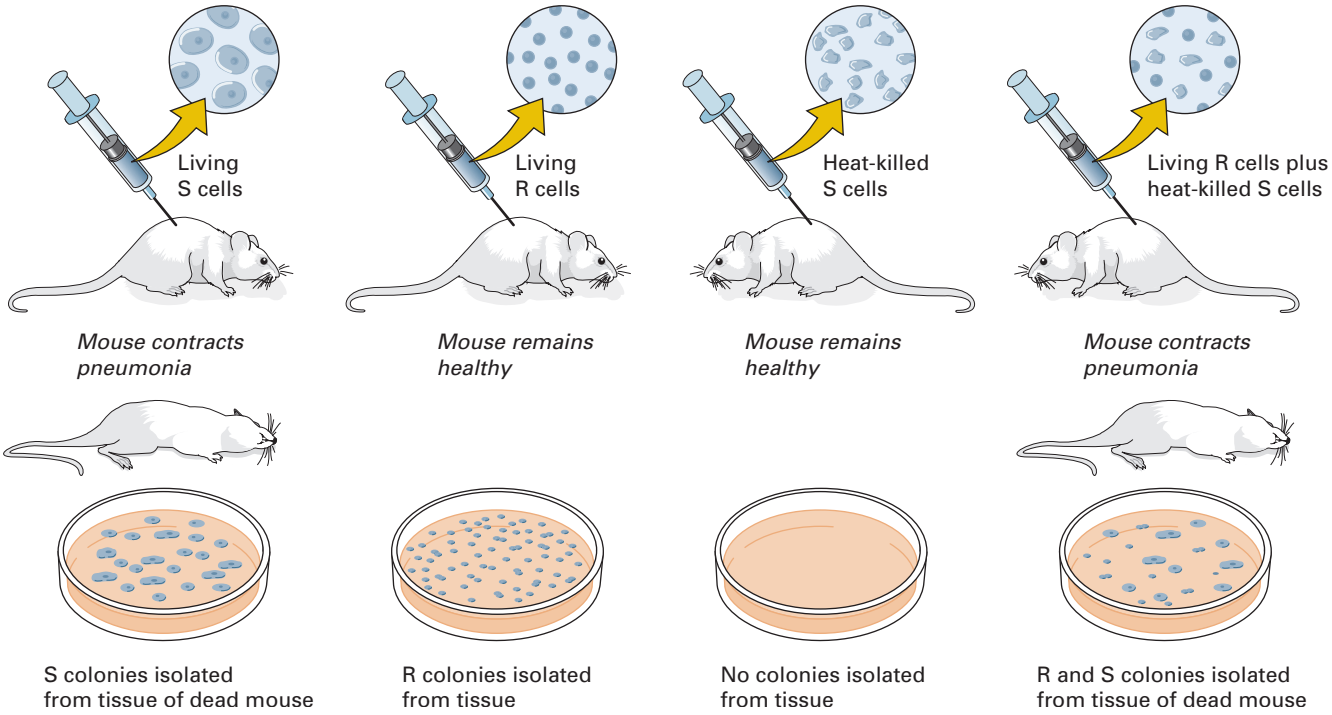


FIGURE 1.2 The Griffith's experiment demonstrating bacterial transformation. A mouse remains healthy if injected with either the nonvirulent R strain of *S. pneumoniae* or heat-killed cell fragments of the usually virulent S strain. R cells in the presence of heat-killed S cells are transformed into the virulent strain, causing pneumonia in the mouse.

bacterial cells and give them the ability to synthesize the S-type capsule. In other words, the R bacteria can be changed—or undergo **transformation**—into S bacteria, and the new characteristics are inherited by descendants of the transformed bacteria.

Griffith's transformation of *Streptococcus* was not in itself definitive, but in 1944 the chemical substance responsible for changing the R cells into S cells was identified as DNA. In a milestone experiment, Oswald Avery, Colin MacLeod, and Maclyn McCarty showed that the substance causing the transformation of R cells into S cells was DNA. In preparation for the experiment, they had to develop chemical procedures for obtaining DNA in almost pure form from bacterial cells, which had not been done before. When they added DNA isolated from S cells to growing cultures of R cells, they observed that a few type-S cells were produced. Although the DNA preparations contained traces of protein and RNA (ribonucleic acid, an abundant cellular macromolecule chemically related to DNA), the transforming activity was not altered by treatments that destroy either protein or RNA. However, treatments that destroy DNA eliminated the transforming activity (**FIGURE 1.3**). These experiments implied that the substance responsible for genetic transformation was the DNA of the cell—and hence that DNA is the genetic material.

Transmission of DNA is the link between generations.

A second pivotal finding was reported by Alfred Hershey and Martha Chase in 1952. They studied cells of the intestinal bacterium *Escherichia coli* after infection by the virus T2. A virus that attacks bacterial cells is called a **bacteriophage**, often shortened to **phage**. (*Bacteriophage* means "bacteria eater.") The T2 particle is exceedingly small, yet it has a complex structure composed of a head containing the phage DNA, a tail, and tail fibers.

The intricate color patterns on butterfly wings demonstrate the intricate complexity that can evolve in developmental processes. [© Ervin Monn/ShutterStock, Inc.]

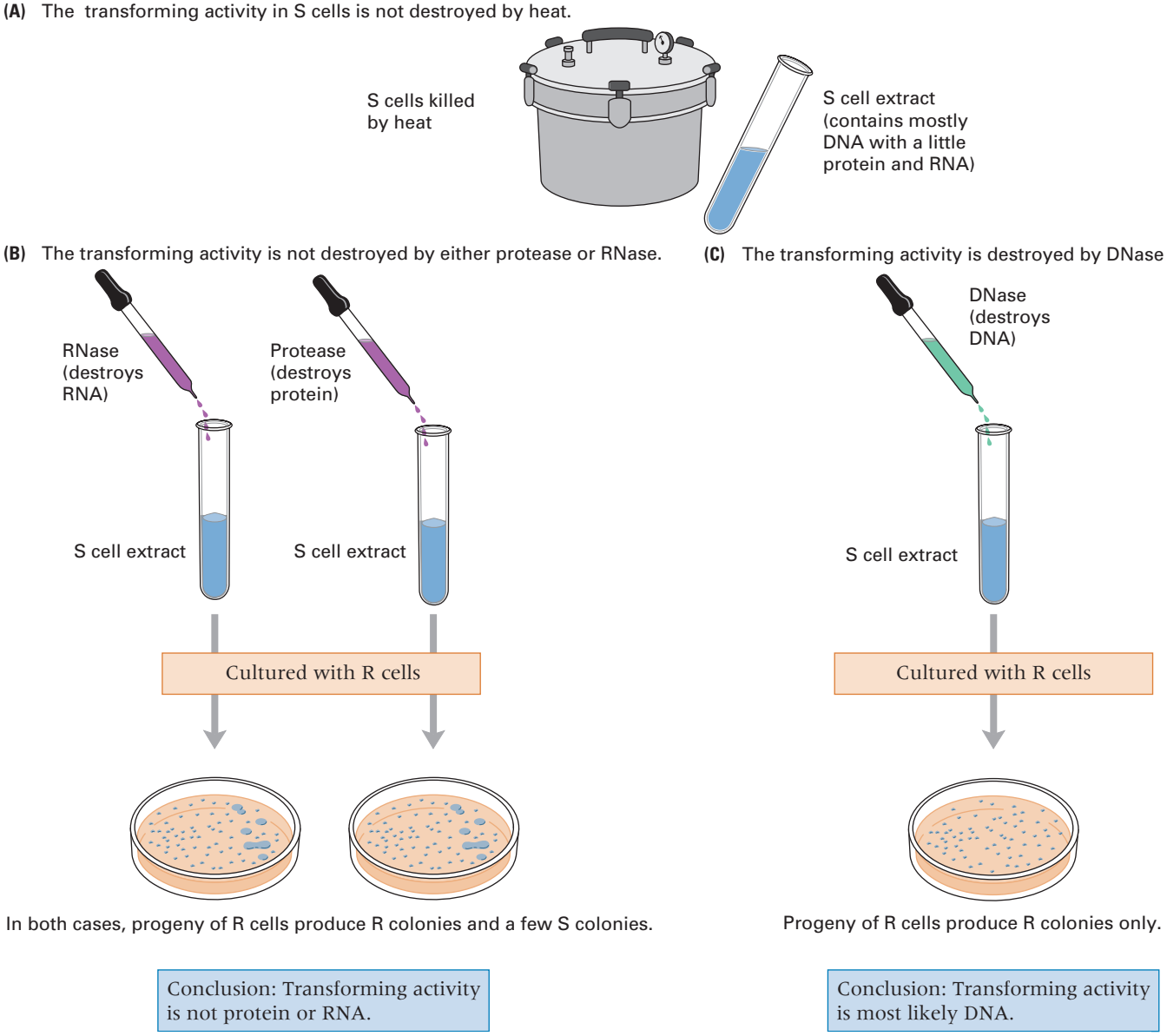


FIGURE 1.3 A diagram of the experiment demonstrating that DNA is the active material in bacterial transformation. (A) Purified DNA extracted from heat-killed S cells can convert some living R cells into S cells, but the extract may still contain undetectable traces of protein and/or RNA. (B) The transforming activity is not destroyed by either protease or RNase. (C) The transforming activity is destroyed by DNase and so probably consists of DNA.

(The head of a human sperm is about 30–50 times larger in both length and width than the head of T2.) Hershey and Chase were already aware that T2 infection proceeds via the attachment of a phage particle by the tip of its tail to the bacterial cell wall, entry of phage material into the cell, multiplication of this material to form a hundred or more progeny phage, and release of the progeny phage by bursting (lysis) of the bacterial host cell. They also knew that T2 particles are composed of DNA and protein in approximately equal amounts.

Because DNA contains phosphorus but no sulfur, whereas most proteins contain sulfur but no phosphorus, it is possible to label DNA and proteins differentially by the use of radioactive isotopes of the two elements. Hershey and Chase produced particles containing radioactive DNA by infecting *E. coli* cells that had been grown for several generations in a medium that included ^{32}P (a radioactive isotope of phosphorus) and then collecting the phage progeny. They obtained other particles containing labeled proteins in the same way, using medium that included ^{35}S (a radioactive isotope of sulfur).

In the experiments summarized in **FIGURE 1.4**, nonradioactive *E. coli* cells were infected with phage labeled with *either* ^{32}P (part A) *or* ^{35}S (part B) in order to follow the DNA and proteins separately. Infected cells were separated from unattached phage particles by centrifugation, resuspended in fresh medium, and then swirled violently in a kitchen blender to shear

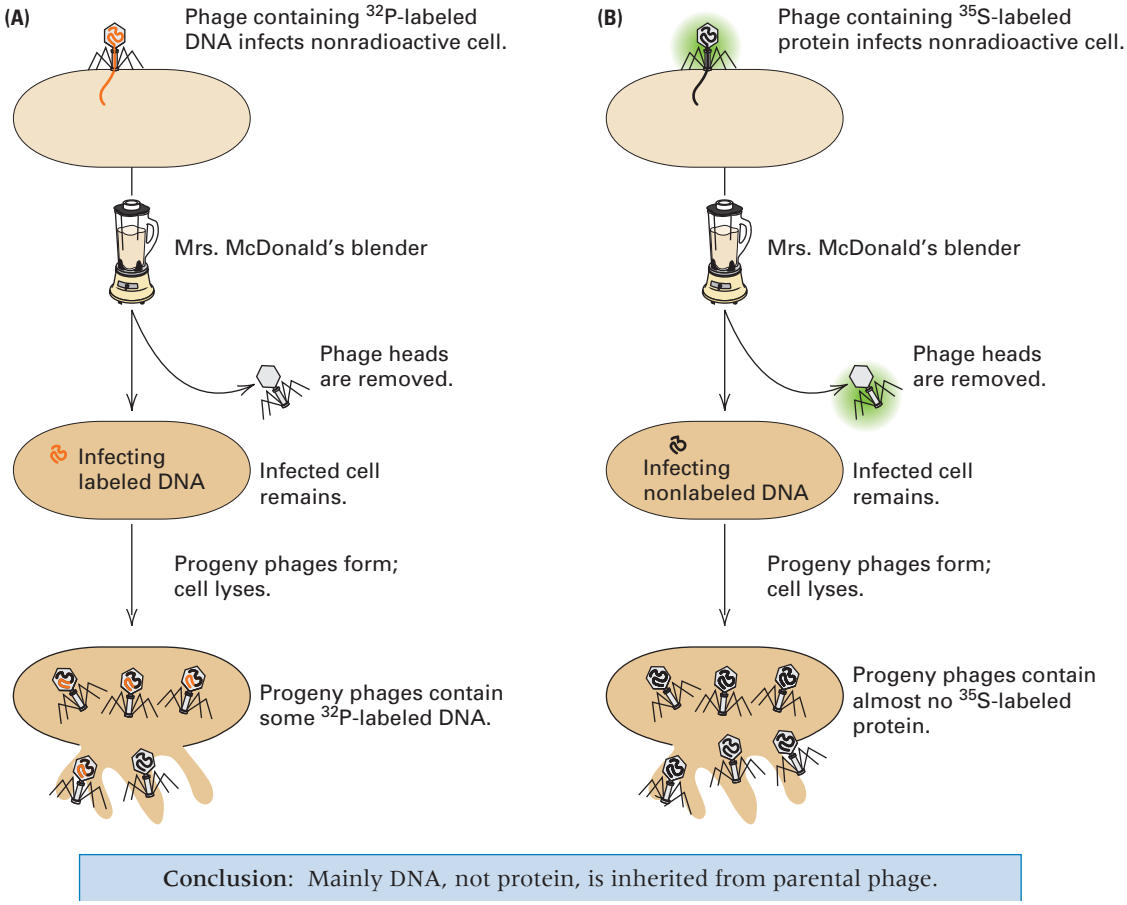


FIGURE 1.4 The Hershey–Chase ("blender") experiment, which demonstrated that DNA, not protein, is responsible for directing the reproduction of phage T2 in infected *E. coli* cells. (A) Radioactive DNA is transmitted to progeny phage in substantial amounts. (B) Radioactive protein is transmitted to progeny phage in negligible amounts.

attached phage material from the cell surfaces. This treatment was found to have no effect on the subsequent course of the infection, which implies that the genetic material must enter the infected cells very soon after phage attachment (**FIGURE 1.5**). The kitchen blender turned out to be the critical piece of equipment. Other methods had been tried to tear the phage heads from the bacterial cell surface, but nothing had worked reliably. Hershey later explained, "We tried various grinding arrangements, with results that weren't very encouraging. When Margaret McDonald loaned us her kitchen blender, the experiment promptly succeeded."

After the phage heads were removed by blending, the infected bacteria were examined. Most of the radioactivity from ^{32}P-labeled phage was found to be associated with the bacteria, whereas only a small fraction of the ^{35}S radioactivity was present in the infected cells. The retention of most of the labeled DNA, contrasted with the loss of most of the labeled protein, implied that a T2 phage transfers most of its DNA, but very little of its protein, to the cell it infects. The critical finding (Figure 1.4) was that about 50 percent of the transferred ^{32}P-labeled DNA, but less than 1 percent

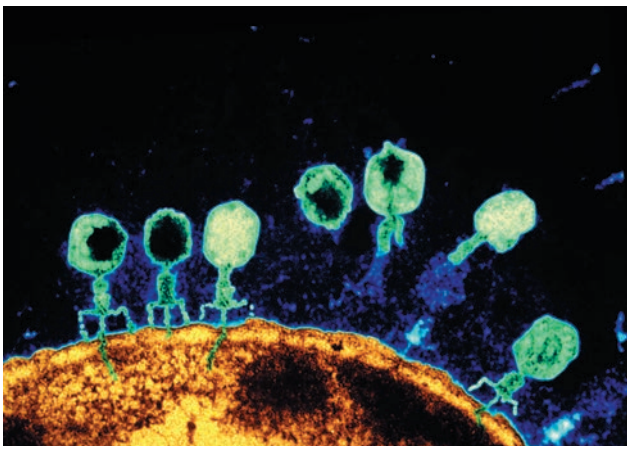


FIGURE 1.5 T2 phages infecting a cell of *E. coli*. Each phage attaches to the bacterial cell wall and injects its DNA into the host. The image has been color-enhanced to show the injected phage DNA in green. [© Oliver Meckes/E.O.S./MPI Tubingen/ Photo Researchers, Inc.]

of the transferred ^{35}S-labeled protein, was inherited by the *progeny* phage particles. Hershey and Chase interpreted this result to mean that the genetic material in T2 phage is DNA.

The transformation experiment and the Hershey–Chase experiment are regarded as classic demonstration that genes consist of DNA. At the present time, the equivalent of the transformation experiment is carried out daily in many research laboratories throughout the world, usually with bacteria, yeast, or animal or plant cells grown in culture. These experiments indicate that DNA is the genetic material in these organisms as well as in phage T2.

KEY CONCEPT

There are no known exceptions to the generalization that DNA is the genetic material in all cellular organisms.

It is worth noting, however, that in a few types of viruses, the genetic material consists of the other type of nucleic acid called RNA.

1.2 The structure of DNA is a double helix composed of two intertwined strands.

Even after it was shown that genes consist of DNA, many questions remained. How is the DNA in a gene duplicated when a cell divides? How does the DNA in a gene control a hereditary trait? What happens to the DNA when a mutation (a change in the DNA) takes place in a gene? Important clues to the answers to these questions emerged from the discovery of the three-dimensional structure of the DNA molecule itself. This structure is discussed next.

A central feature of double-stranded DNA is complementary base pairing.

In the early 1950s, a number of researchers began to try to understand the detailed molecular structure of DNA. The first essentially correct three-dimensional structure of the DNA molecule was proposed in 1953 by James Watson and Francis Crick at Cambridge University. The structure was dazzling in its elegance and revolutionary in suggesting how DNA duplicates itself, controls hereditary traits, and undergoes mutation. Even while their tin sheet and wire model of the DNA molecule was still incomplete, Crick announced in his favorite pub that, "We have discovered the secret of life."

In the Watson–Crick structure, DNA consists of two long chains of subunits twisted around one another to form a double-stranded helix. The double helix is right-handed, which means that as one looks along the barrel, each chain follows a clockwise path as it progresses. You can see the right-handed coiling in part A of **FIGURE 1.6** if you imagine yourself looking up into the structure from the bottom: The smaller spheres outline the "backbone" of each individual strand, and they coil in a clockwise direction. The subunits of each strand are **nucleotides**, each of which contains any one of four chemical constituents called **bases**. The four bases in DNA are

Adenine (A) **Guanine (G)**
Thymine (T) **Cytosine (C)**

The chemical structures of the nucleotides and bases need not concern us at this point. A key point for our present purposes is that the bases in the double helix are paired as shown in Figure 1.6, part B. That is,

KEY CONCEPT

At any position on the paired strands of a DNA molecule, if one strand has an A, then the partner strand has a T; and if one strand has a G, then the partner strand has a C.

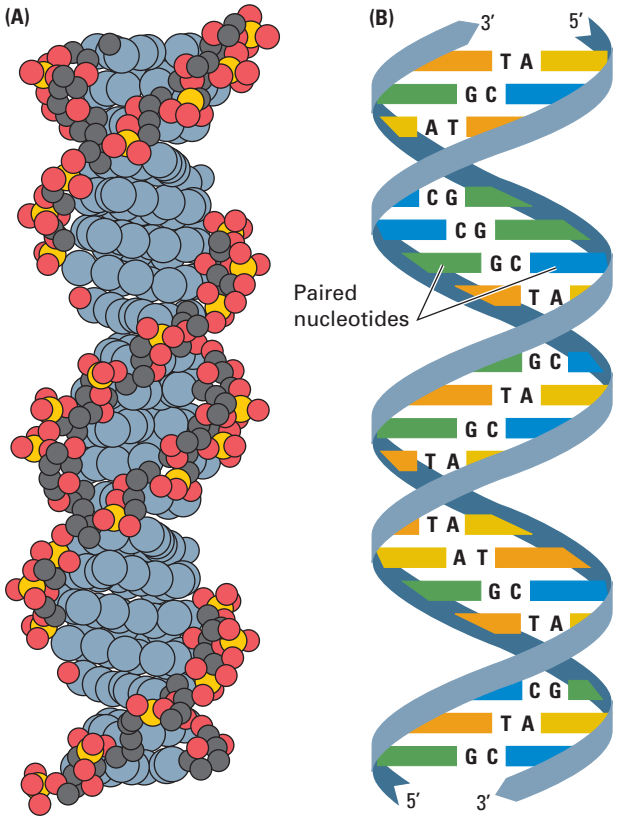


FIGURE 1.6 Molecular structure of a DNA double helix. (A) A "space-filling" model, in which each atom is depicted as a sphere. (B) A diagram highlighting the helical backbones on the outside of the molecule and the stacked A—T and G—C base pairs inside.

The base pairing between A and T and between G and C is said to be **complementary base pairing**; the complement of A is T, and the complement of G is C. The complementary pairing in the duplex molecule means that each base along one strand of the DNA is matched with a base in the opposite position on the other strand. Furthermore,

KEY CONCEPT

Nothing restricts the sequence of bases in a single strand, so any sequence could be present along one strand.

This principle explains how only four bases in DNA can code for the huge amount of information needed to make an organism. It is the linear order or *sequence* of bases along the DNA that encodes the genetic information, and the sequence is completely unrestricted.

The complementary pairing is also called *Watson–Crick base pairing*. In the three-dimensional structure (Figure 1.6, part A), the base pairs are represented by the larger spheres filling the interior of the double helix. The base pairs lie almost flat, stacked on top of one another perpendicular to the long axis of the double helix, like pennies in a roll. When discussing a DNA molecule, biologists frequently refer to the individual strands as **single-stranded DNA** and to the double helix as **double-stranded DNA** or **duplex DNA**.

Each DNA strand has a **polarity**, or directionality, like a chain of circus elephants linked trunk to tail. In this analogy, each elephant corresponds to one nucleotide along the DNA strand. The polarity is determined by the direction in which the nucleotides are pointing. The "trunk" end of the strand is called the 5′ *end* of the strand, and the "tail" end is called the 3′ *end*. In double-stranded DNA, the paired strands are oriented in opposite directions: The 5′ end of one strand is aligned with the 3′ end of the other. The oppositely oriented strands are said to be **antiparallel**. In illustrating DNA molecules, we use an arrow-like ribbon to represent the backbone, and we use tabs jutting off the ribbon to represent the nucleotides. The polarity of a DNA strand is indicated by the direction of the arrow-like ribbon. The tail of the arrow represents the 5′ end of the DNA strand, the head the 3′ end.

Beyond the most optimistic hopes, knowledge of the structure of DNA immediately gave clues to its function:

1. The sequence of bases in DNA could be copied by using each of the separate "partner" strands as a pattern for the creation of a new partner strand with a complementary sequence of bases.
2. The DNA could contain genetic information in coded form in the sequence of bases, analogous to letters printed on a strip of paper.
3. Changes in genetic information (mutations) could result from errors in copying in which the base sequence of the DNA became altered.

In the remainder of this chapter, we discuss some of the implications of these clues.

In replication, each parental DNA strand directs the synthesis of a new partner strand.

"It has not escaped our notice," wrote Watson and Crick, "that the specific base pairing we have postulated immediately suggests a copying mechanism for the genetic material." The copying process in which a single DNA molecule becomes two identical molecules is called **replication**. The replication mechanism that Watson and Crick had in mind is illustrated in **FIGURE 1.7**. The strands of the original (parent) duplex separate, and each individual strand serves as a pattern, or **template**, for the synthesis of a new strand (replica). The replica strands are synthesized by the addition of successive nucleotides in such a way that each base in the replica is complementary (in the Watson–Crick pairing sense) to the base across the way in the template strand. Although the mechanism in Figure 1.7 is simple

Q A MOMENT TO THINK 1

Problem 1: When the DNA sequence of the bacteriophage λ (lambda), which infects *E. coli*, was first determined, geneticists were surprised to find a 12-base single-stranded overhang at the 5′ end of each strand. The structure of the ends is diagrammed below. The dots represent the remaining 48,501 base pairs of phage sequence that are not shown.

Using your knowledge that, in double-stranded DNA, A pairs with T and G pairs with C, and that the paired strands in duplex DNA are antiparallel, identify what is special about these single-stranded ends and suggest a manner in which they might interact. *Hint*: The single-stranded regions are called cohesive ends. (The answer can be found at the end of the chapter.)

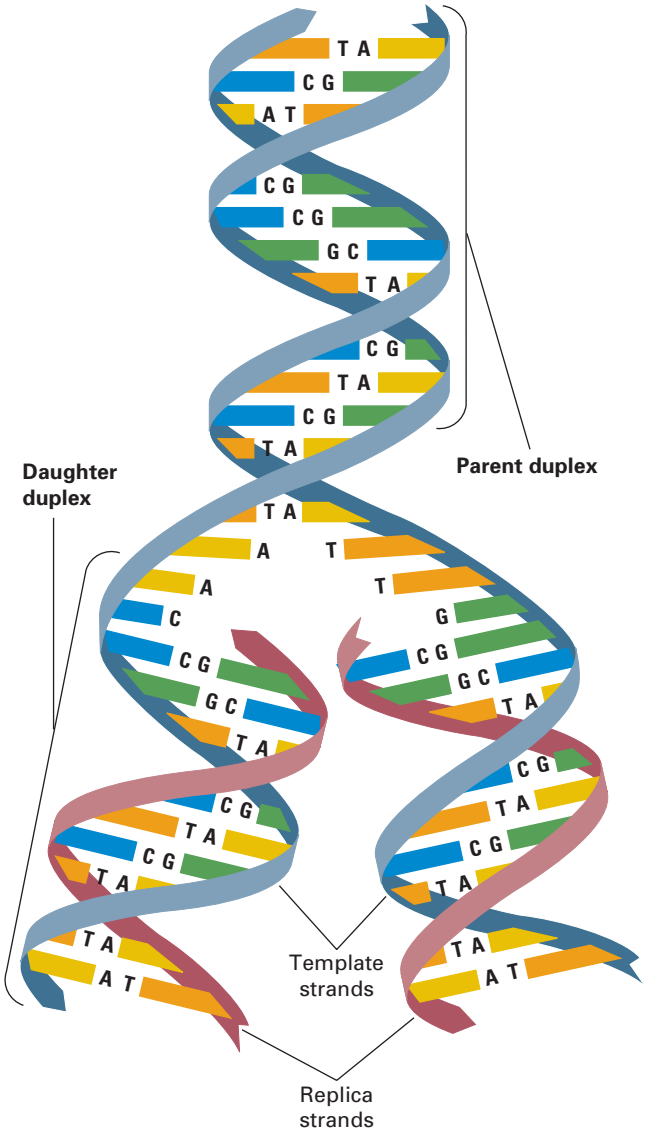


FIGURE 1.7 Replication in a long DNA duplex as originally proposed by Watson and Crick. The parental strands separate, and each parental strand serves as a template for the formation of a new daughter strand by means of A—T and G—C base pairing.

in principle, it is a complex process that is fraught with geometric problems and requires a variety of enzymes and other proteins. The end result of replication is that a single double-stranded molecule becomes replicated into two copies with identical sequences:

```
5′-ATGCCGTA-3′
3′-TACGGCAT-5′

5′-ATGCCGTA-3′        5′-ATGCCGTA-3′
3′-TACGGCAT-5′        3′-TACGGCAT-5′
```

Here the bases in the newly synthesized strands are shown in red. In the duplex on the left, the top strand is the template from the parental molecule and the bottom strand is newly synthesized; in the duplex on the right, the bottom strand is the template from the parental molecule and the top strand is newly synthesized.

1.3 Genes affect organisms through the action of proteins.

One of the important principles of molecular genetics is that genes exert their effects on organisms indirectly. For most genes, the genetic information contained in the nucleotide sequence specifies a particular type of *protein*. Proteins control the chemical and physical processes of cells known as **metabolism**. Many proteins are **enzymes**, a term introduced in 1878 to refer to the biological catalysts that accelerate biochemical reactions. Enzymes are essential for the breakdown of organic molecules, generating the chemical energy needed for cellular activities; they are also essential for the synthesis of small molecules and for their assembly into larger molecules and complex cellular structures.

Although the fundamental connection between genes and proteins was not widely appreciated until the 1940s, the first evidence for a relationship came much earlier. The pioneering observations were made by Archibald Garrod, a British physician, who studied genetic diseases caused by inherited defects in metabolism. He concluded that an inherited defect in metabolism results from an inherited defect in an enzyme. The key observations on which Garrod based this conclusion are summarized in the following sections.

Enzyme defects result in inborn errors of metabolism.

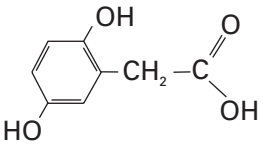
In 1908 Garrod gave a series of lectures in which he proposed this fundamental hypothesis about the relationship between enzymes and disease:

KEY CONCEPT

Any hereditary disease in which cellular metabolism is abnormal results from an inherited defect in an enzyme.

Such diseases became known as **inborn errors of metabolism**, a term still in use today.

Garrod studied a number of inborn errors of metabolism in which the patients excreted abnormal substances in the urine. One of these was **alkaptonuria**. In this case, the abnormal substance excreted is **homogentisic acid**:



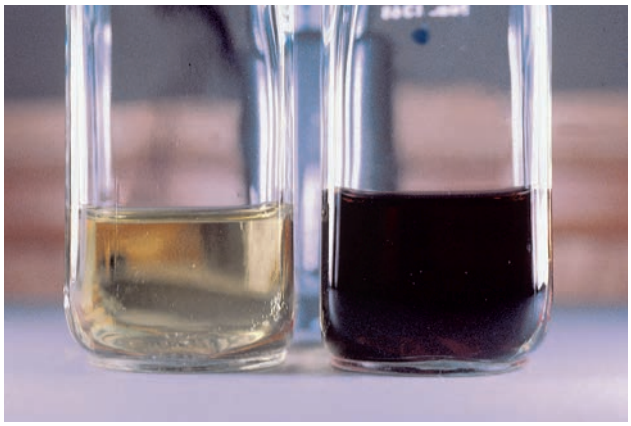


FIGURE 1.8 Urine from a person with alkaptonuria turns black because of the oxidation of the homogentisic acid that it contains. [Courtesy of Daniel De Aguiar.]

This is a conventional chemical representation in which each corner of the hexagon represents a carbon atom, and hydrogen atoms attached to the ring are not shown. The six-carbon ring is called a *phenyl* ring. An early name for homogentisic acid was *alkapton*—hence the name *alkaptonuria*. Even though alkaptonuria is rare, with an incidence of about one in 200,000 people, it was well known even before Garrod studied it. The disease itself is relatively mild, but it has one striking symptom: The urine of the patient turns black because of the oxidation of homogentisic acid (**FIGURE 1.8**). This is why alkaptonuria is also called *black urine disease*. The passing of black urine can hardly escape being noticed. One case was described in the year 1649:

> The patient was a boy who passed black urine and who, at the age of fourteen years, was submitted to a drastic course of treatment that had for its aim the subduing of the fiery heat of his viscera, which was supposed to bring about the condition in question by charring and blackening his bile. Among the measures prescribed were bleedings, purgation, baths, a cold and watery diet, and drugs galore. None of these had any obvious effect, and eventually the patient, who tired of the futile and superfluous therapy, resolved to let things take their natural course. None of the predicted evils ensued. He married, begat a large family, and lived a long and healthy life, always passing urine black as ink. (Quotation from Garrod, 1908.)

Garrod was primarily interested in the biochemistry of alkaptonuria, but he took note of family studies that indicated that the disease was inherited as though it were due to a defect in a single gene. As to the biochemistry, he deduced that the problem in alkaptonuria was the patients' inability to break down the phenyl ring of six carbons that is present in homogentisic acid. Where does this ring come from? Mammals are unable to synthesize it and must obtain it from their diet. Garrod proposed that homogentisic acid originates as a breakdown product of two amino acids, phenylalanine and tyrosine, which also contain a phenyl ring. An **amino acid** is one of the "building blocks" from which proteins are made. Phenylalanine and tyrosine are constituents of normal proteins. The scheme that illustrates the relationship between the molecules is shown in **FIGURE 1.9**. Any such sequence of biochemical reactions is called a **biochemical pathway** or a **metabolic pathway**. Each arrow in the pathway represents a single step depicting the transition from the "input" or **substrate molecule**, shown at the tail of the arrow, to the "output" or **product molecule**, shown at the tip. Biochemical pathways are usually oriented either vertically with the arrows pointing down, as in Figure 1.9, or horizontally, with the arrows pointing from left to right. Garrod did not know all of the details of the pathway in Figure 1.9, but he did understand that the key step in the breakdown of homogentisic acid is the breaking open of the phenyl ring and that the phenyl ring in homogentisic acid comes from dietary phenylalanine and tyrosine.

What allows each step in a biochemical pathway to occur? Garrod's insight was to see that each step requires a specific enzyme to catalyze the reaction and allow the chemical transformation to take place. Persons with an inborn error of metabolism, such as alkaptonuria, have a defect in one step of a metabolic pathway because they lack a functional enzyme for that step. When an enzyme in a pathway is defective, the pathway is said to have a **block** at that step. One frequent result of a blocked pathway is that the substrate of the defective enzyme accumulates. Observing the accumulation of homogentisic acid in patients with alkaptonuria, Garrod proposed that there must be an enzyme whose function is to open the phenyl ring of homogentisic acid and that this enzyme is missing in these patients. Discovery of all the enzymes in the pathway in Figure 1.9 took a long time. The enzyme that opens the phenyl ring of homogentisic acid was not actually isolated until 50 years after Garrod's lectures. In normal people it is found in cells of the liver. Just as Garrod had predicted, the enzyme is defective in patients with alkaptonuria.

The pathway for the breakdown of phenylalanine and tyrosine, as it is understood today, is shown in **FIGURE 1.10**. In this figure the emphasis is on the enzymes rather than on the structures of the **metabolites**, or small molecules, on which the enzymes act. As Garrod would have predicted, each step in the pathway requires the presence of a particular enzyme that catalyzes that step. Although Garrod knew only about alkaptonuria, in which the defective enzyme is homogentisic acid 1,2-dioxygenase, we now know the clinical consequences of defects in the other enzymes. Unlike alkaptonuria, which is a relatively benign inherited disease, the others are very serious. The condition known as **phenylketonuria (PKU)**

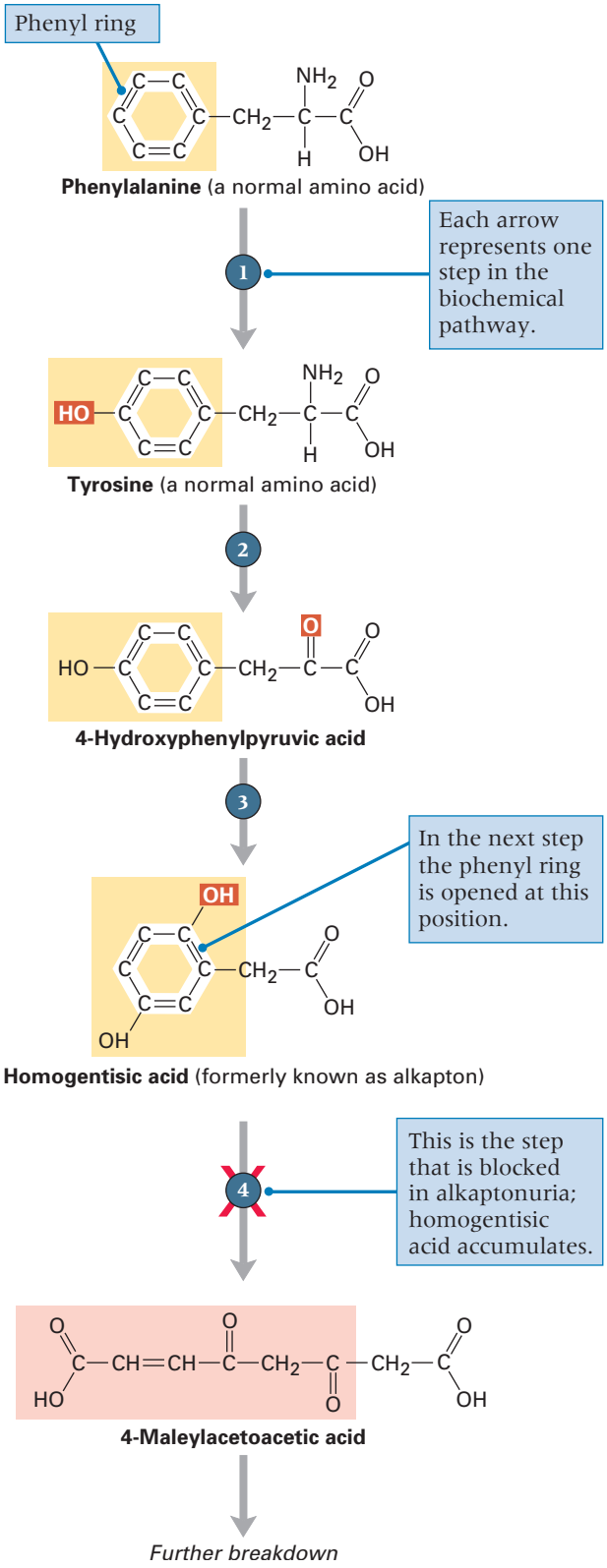


FIGURE 1.9 Metabolic pathway for the breakdown of phenylalanine and tyrosine. Each step in the pathway, represented by an arrow, requires a particular enzyme to catalyze the reaction. The key step in the breakdown of homogentisic acid is the breaking open of the phenyl ring.

results from the absence of (or a defect in) the enzyme **phenylalanine hydroxylase (PAH)**. When this step in the pathway is blocked, phenylalanine accumulates. The excess phenylalanine is broken down into harmful metabolites that cause defects in myelin formation that damage a child's developing nervous system and lead to severe mental retardation.

If PKU is diagnosed in children soon enough after birth, they can be placed on a specially formulated diet low in phenylalanine. The child is allowed only as much phenylalanine as can be used in the synthesis of proteins, so excess phenylalanine does not accumulate. The special diet is very strict. It excludes meat, poultry, fish, eggs, milk and milk products, legumes, nuts, and bakery goods manufactured with regular flour. These foods are replaced by a synthetic formula that is very expensive. With the special diet, however, the detrimental effects of excess phenylalanine on mental development can largely be avoided. In many countries, including the United States, all newborn babies have their blood tested for chemical signs of PKU. Routine screening is cost-effective because PKU is relatively common. In the United States, the incidence is about one in 8000 among Caucasian births. The disease is less common in other ethnic groups.

In the metabolic pathway in Figure 1.10, defects in the breakdown of tyrosine or of 4-hydroxyphenylpyruvic acid lead to types of tyrosinemia. These are also severe diseases. Type II is associated with skin lesions and mental retardation, type III with severe liver dysfunction.

A defective enzyme results from a mutant gene.

It follows from Garrod's work that a defective enzyme results from a mutant gene. How does a mutant gene result in a defective enzyme? Garrod did not speculate. For all he knew, genes *were* enzymes. This would have been a logical hypothesis at the time. We now know that the relationship between genes and enzymes is somewhat indirect. With a few exceptions, each enzyme is *encoded* in a particular sequence of nucleotides present in a region of DNA. The DNA region that codes for the enzyme, as well as adjacent regions that regulate when and in which cells the enzyme is produced, make up the "gene" that encodes the enzyme.

The genes for the enzymes in the biochemical pathway in Figure 1.10 have all been identified and the nucleotide sequence of the DNA determined. In the following list, and throughout this book, we use the typographical convention that the names of *genes* are printed in *italic* type, whereas gene products are printed in regular type. In Figure 1.10 the numbers 1 through 4 correspond to the following genes and enzymes:

1. The gene *PAH* on the long arm of chromosome 12 encodes phenylalanine hydroxylase (PAH).

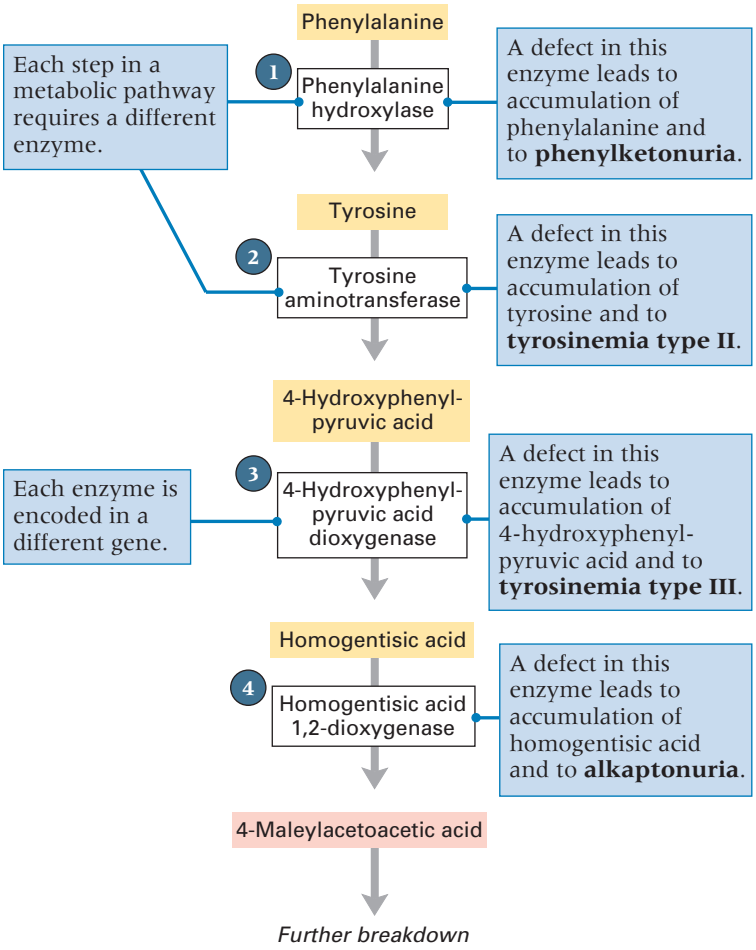


FIGURE 1.10 Inborn errors of metabolism in the breakdown of phenylalanine and tyrosine. A different inherited disease results when each of the enzymes is missing or defective. Alkaptonuria results from a defective homogentisic acid 1,2-dioxygenase, phenylketonuria from a defective phenylalanine hydroxylase.

2. The gene *TAT* on the long arm of chromosome 16 encodes tyrosine aminotransferase (TAT).
3. The gene *HPD* on the long arm of chromosome 12 encodes 4-hydroxyphenylpyruvic acid dioxygenase (HPD).
4. The gene *HGD* on the long arm of chromosome 3 encodes homogentisic acid 1,2-dioxygenase (HGD).

Genetic analysis led to the one gene-one enzyme hypothesis.

Garrod's thinking was far ahead of his time, and his conclusions about inborn errors of metabolism were largely ignored. The influential experiments connecting genes with enzymes were carried out in the 1940s by George W. Beadle and Edward L. Tatum using a filamentous fungus *Neurospora crassa*, commonly called red bread mold, an organism they chose because both genetic and biochemical analysis could be done with ease. In these experiments they identified new mutations that each caused a block in the metabolic pathway for the synthesis of some needed nutrient and showed that each of these blocks corresponded to a defective enzyme needed for one step in the pathway. The experimental approach, now called **genetic analysis**, was important because it solidified the link between genetics and biochemistry. Equally as important, the experimental approach is widely applicable to understanding any complex biological process, ranging from the genetic control of the cell cycle or cancer to that of development or behavior. For this reason the methods of genetic analysis warrant a closer examination.

N. crassa grows in the form of filaments on a great variety of substrates including laboratory medium containing only inorganic salts, a sugar, and one vitamin. Such a medium is known as a **minimal medium** because it contains only the nutrients that are essential for growth of the organism. The filaments consist of a mass of branched threads separated into interconnected, multinucleate compartments allowing free interchange of nuclei and cytoplasm. Each nucleus contains a single set of seven chromosomes. Beadle and Tatum recognized that the ability of *Neurospora* to grow in minimal medium implied that the organism must be able to synthesize all of the other small molecules needed for growth, such as amino acids. If the biosynthetic pathways needed for growth are controlled by genes, then a mutation in a gene responsible for synthesizing an essential nutrient would be expected to render a strain unable to grow unless the strain were provided with the nutrient.

These ideas were tested in the following way. Spores of nonmutant *Neurospora* were irradiated with either x-rays or ultraviolet light to produce mutant strains with various nutritional requirements. The isolation of a set of mutants affecting any biological process, in this case metabolism, is called a **mutant screen**. In the initial step for identifying mutants, summarized in **FIGURE 1.11**, the irradiated spores (purple) were used in crosses with an untreated strain (green). Ascospores produced by the sexual cycle in fruiting bodies were individually germinated in **complete medium**, a complex medium enriched with a variety of amino acids, vitamins, and other substances expected to be essential metabolites whose synthesis could be blocked by a mutation. Even those ascospores containing a

THE HUMAN CONNECTION
One Gene, One Enzyme?

George W. Beadle and Edward L. Tatum 1941
Stanford University, Stanford, California
Genetic Control of Biochemical Reactions in Neurospora

How do genes control metabolic processes? The suggestion that genes control enzymes was made very early in the history of genetics, most notably by the British physician Archibald Garrod in his 1908 book *Inborn Errors of Metabolism*. But the precise relationship between genes and enzymes was still uncertain. Perhaps each enzyme is controlled by more than one gene, or perhaps each gene contributes to the control of several enzymes. The classic experiments of Beadle and Tatum showed that the relationship is usually remarkably simple: One gene codes for one enzyme. The pioneering experiments united genetics and biochemistry, and for the "one gene, one enzyme" concept, Beadle and Tatum were awarded a Nobel Prize in 1958 (Joshua Lederberg shared the prize for his contributions to microbial genetics). Because we now know that some enzymes contain polypeptide chains encoded by two (or occasionally more) different genes, a more accurate statement of the principle is "one gene, one polypeptide." Beadle and Tatum's experiments also demonstrate the importance of choosing the right organism. *Neurospora* had been introduced as a genetic organism only a few years earlier, and Beadle and Tatum realized that they could take advantage of the ability of this organism to grow on a simple medium composed of known substances.

From the standpoint of physiological genetics the development and functioning of an organism consist essentially of an integrated system of chemical reactions controlled in some manner by genes. . . . In investigating the roles of genes, the physiological geneticist usually attempts to determine the physiological and biochemical bases of already known hereditary traits. . . . There are, however, a number of limitations inherent in this approach. Perhaps the most serious of these is that the investigator must in general confine himself to the study of non-lethal heritable characters. Such characters are likely to involve more or less non-essential so-called "terminal" reactions. . . .

These preliminary results appear to us to indicate that the approach may offer considerable promise as a method of learning more about how genes regulate development and function.

A second difficulty is that the standard approach to the problem implies the use of characters with visible manifestations. Many such characters involve morphological variations, and these are likely to be based on systems of biochemical reactions so complex as to make analysis exceedingly difficult. . . . Considerations such as those just outlined have led us to investigate the general problem of the genetic control of development and metabolic reactions by reversing the ordinary procedure and, instead of attempting to work out the chemical bases of known genetic characters, to set out to determine if and how genes control known biochemical reactions. The ascomycete *Neurospora* offers many advantages for such an approach and is well suited to genetic studies. Accordingly, our program has been built around this organism. The procedure is based on the assumption that x-ray treatment will induce mutations in genes concerned with the control of known specific chemical reactions. If the organism must be able to carry out a certain chemical reaction to survive on a given medium, a mutant unable to do this will obviously be lethal on this medium. Such a mutant can be maintained and studied, however, if it will grow on a medium to which has been added the essential product of the genetically blocked reaction. . . . Among approximately 2000 strains [derived from single cells after x-ray treatment], three mutants have been found that grow essentially normally on the complete medium and scarcely at all on the minimal medium. One of these strains proved to be unable to synthesize vitamin B_6 (pyridoxine). A second strain turned out to be unable to synthesize vitamin B_1 (thiamine). A third strain has been found to be unable to synthesize para-aminobenzoic acid. . . . These preliminary results appear to us to indicate that the approach may offer considerable promise as a method of learning more about how genes regulate development and function. For example, it should be possible, by finding a number of mutants unable to carry out a particular step in a given synthesis, to determine whether only one gene is ordinarily concerned with the immediate regulation of a given specific chemical reaction.

Source: G. W. Beadle and E. L. Tatum, *Proc. Natl. Acad. Sci. USA* 27 (1941): 499–506.

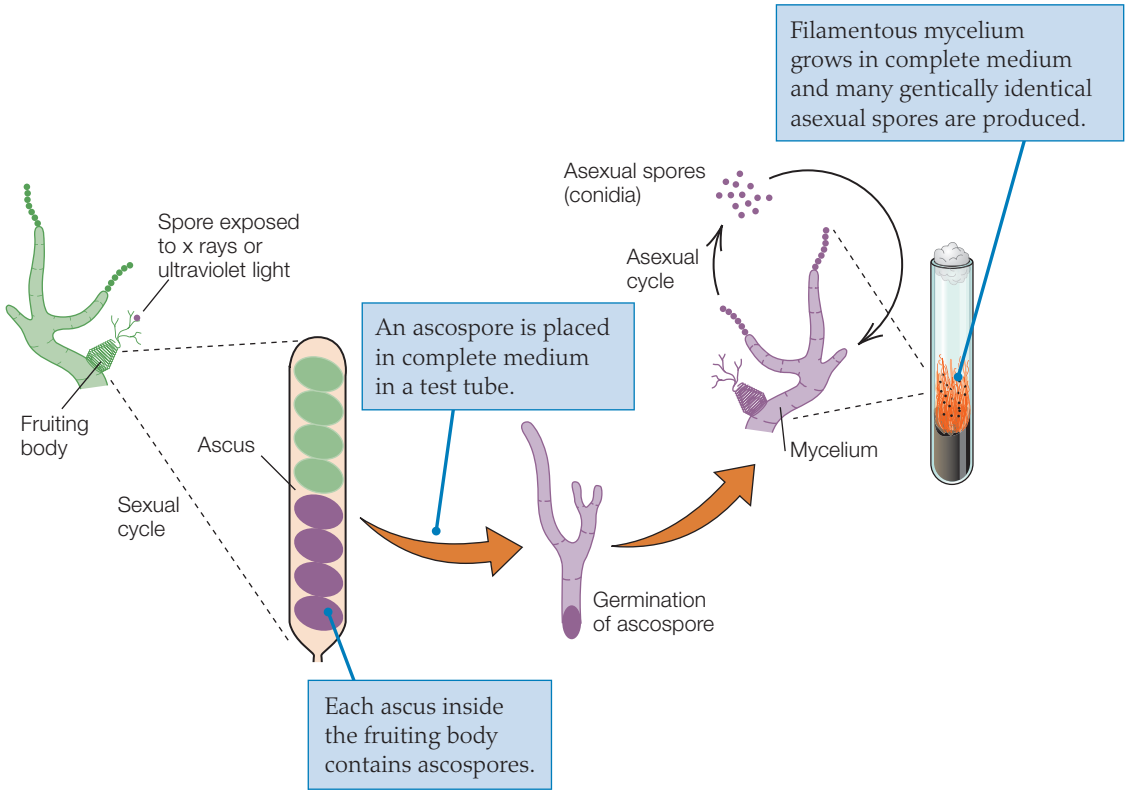


FIGURE 1.11 Beadle and Tatum obtained mutants of the filamentous fungus *Neurospora crassa* by exposing asexual spores to x-rays or ultraviolet light. The treated spores were used to start the sexual cycle in fruiting bodies. After any pair of cells and their nuclei undergo fusion, meiosis takes place almost immediately and results in eight sexual spores (ascospores) included in a single ascus. These are removed individually and cultured in complete medium. Ascospores that carry new nutritional mutants are identified later by their inability to grow in minimal medium.

new mutation affecting synthesis of an essential nutrient would be expected to germinate and grow in complete medium.

To identify which of the irradiated ascospores contained a new mutation affecting the synthesis of an essential nutrient, spores from each culture were transferred to minimal medium (**FIGURE 1.12A**). The vast majority of cultures yielded spores that could grow on minimal medium; these cultures lacked any new mutation of the desired type and were discarded. The cultures that were kept were the small number producing spores unable to grow on minimal medium, because these were mutant cultures that contained a new mutation blocking the synthesis of some essential nutrient.

Spores from each mutant culture were then transferred to a series of media to determine whether the mutation results in a requirement for a vitamin, an amino acid, or some other substance. In the example illustrated in **FIGURE 1.12B**, the mutant strain requires one (or possibly more than one) amino acid, because a mixture of all amino acids added to the minimal medium allows growth. Because the proportion of irradiated cultures with new mutations was very small, only a negligible number of cultures would contain two or more new mutations that had occurred simultaneously.

For nutritional mutants requiring amino acids, further experiments testing each of the amino acid individually usually revealed that only one amino acid was required to be added to minimal medium to support growth. In **FIGURE 1.12C**, the mutant strain requires the amino acid arginine. Even in the 1940s some of the possible intermediates in amino acid biosynthesis had been identified. These were recognized by their chemical resemblance to the amino acid and by being present at low levels in the cells of organisms. In the case of arginine, two candidates were ornithine and citrulline. All mutants requiring arginine were, therefore, tested in medium supplemented with either ornithine alone or citrulline alone (**FIGURE 1.12D**). One class of arginine-requiring mutants, designated Class I, was able to grow in minimal medium supplemented with either ornithine, citrulline, or arginine. Other mutants, designated Class II, were able to grow in minimal medium supplemented with either citrulline or arginine but not ornithine. A third class, Class III, was able to grow only in minimal medium supplemented with arginine.

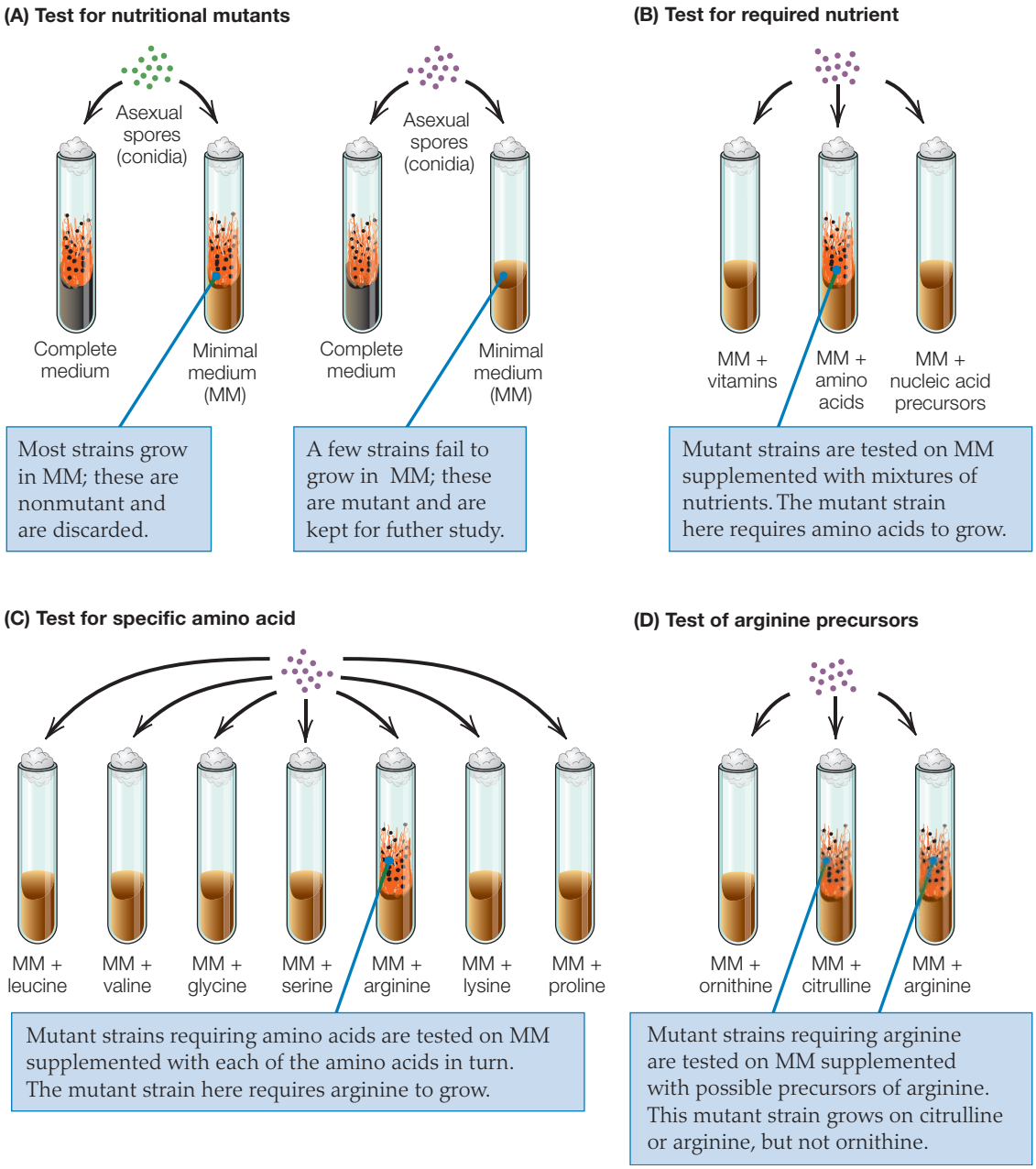


FIGURE 1.12 (A) Mutant spores can grow in complete medium but not in minimal medium. (B) Each new mutant is tested for growth in minimal medium supplemented with a mixture of nutrients. (C) Mutants that can grow on minimal medium supplemented with amino acid are tested with each amino acid individually. (D) Mutants unable to grow in the absence of arginine are tested with likely precursors of arginine.

The types of arginine-requiring mutants illustrate the principle of genetic analysis as applied to metabolic pathways. The basic principle is that

KEY CONCEPT

If a strain with a mutant enzyme that blocks a particular step in a linear metabolic pathway can grow when an intermediate is added to the growth medium, it means that the location of the intermediate in the pathway is *downstream* of the enzymatic step that is blocked.

This principle makes intuitive sense because, if the intermediate were upstream of the metabolic block, then adding the intermediate to the growth medium would not allow growth, because conversion of the intermediate would still be blocked at the point of the mutant enzyme.

Application of the principle to the linear pathway for arginine biosynthesis is shown in **FIGURE 1.13**, where arginine is the end product starting with some precursor metabolite, and ornithine and citrulline are intermediates in the pathway. The

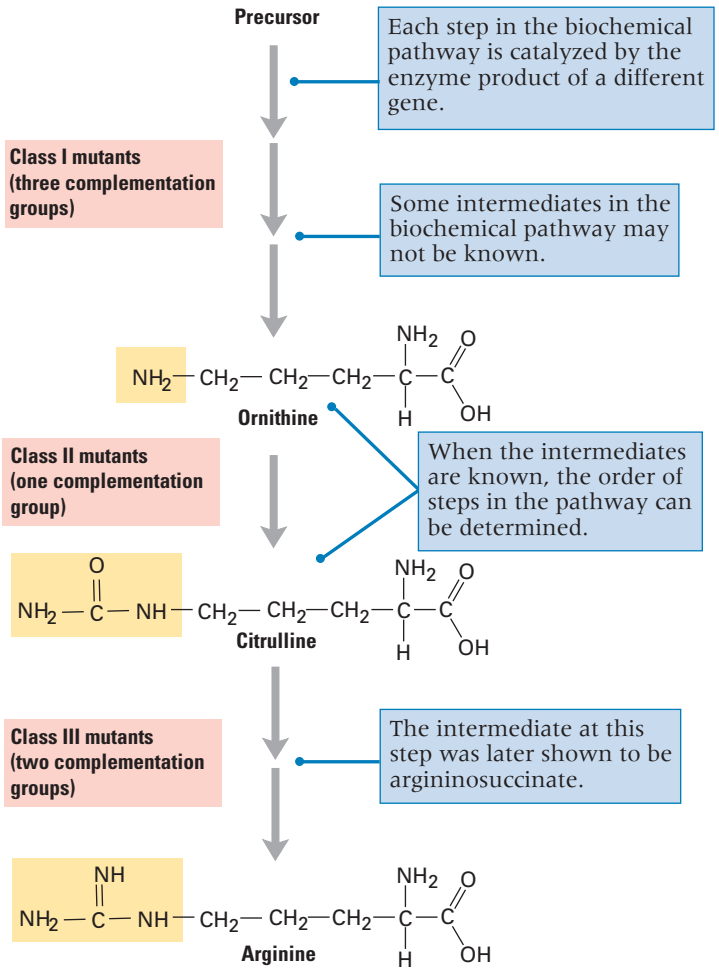


FIGURE 1.13 Metabolic pathway for arginine biosynthesis inferred from genetic analysis of *Neurospora* mutants.

mutants imply the order of the intermediates shown because

- Mutants in Class I are able to grow in the presence of either ornithine or citrulline, which means that both ornithine and citrulline are downstream of any of the enzymes blocked in Class I mutants.
- Mutants in Class II are to grow in the presence of citrulline but not ornithine, which means that citrulline is located downstream of the enzymatic block in Class II mutants and that ornithine is upstream of the metabolic block in Class II mutants.
- Mutants in Class III are unable to grow in the presence of either citrulline or ornithine, which means that these intermediates are upstream of any of the enzymatic steps blocked in Class III mutants.

The structure of the pathway in Figure 1.13 was further confirmed by the observations that Class III mutants accumulate citrulline and Class II mutants accumulate ornithine. Ultimately, direct biochemical experiments demonstrated that the inferred enzymes were actually present in nonmutant strains but defective in mutant strains.

Mutant screens sometimes isolate different mutations in the same gene.

Beadle and Tatum were fortunate to study metabolic pathways in a relatively simple organism in which each gene specifies a single enzyme, a relation often called the **one gene–one enzyme hypothesis**. In such a situation, genetic analysis of the mutants reveals a great deal more about the metabolic pathway than merely the order of the intermediates. By classifying each mutation according to the particular gene it is in and grouping all the mutations in each gene together, each set of mutations and, therefore, each individual gene, correspond to one enzymatic step in the metabolic pathway. In Figure 1.13, for example, Class I includes mutations in any of three different genes, which implies that there are three steps in the pathway between the precursor and ornithine. Similarly, Class III comprises mutations in either of two different genes, which implies that there are two steps in the pathway between the citrulline and arginine. However, Class II consists of mutations in only one gene, which implies only one step in the pathway between ornithine and citrulline.

Mutations that have defects in the same gene are identified by means of a **complementation test**, in which two mutations are brought together into the same cell. In most multicellular organisms (and even some sexual unicellular organisms), the usual way to do this is by means of a mating. When two parents, each carrying one of the two mutations, are crossed, fertilization brings the reproductive cells containing the two mutations together, and through ordinary cell division each cell in the resulting offspring carries one copy of each mutant gene. In *Neurospora* this procedure does not work because nuclear fusion is followed almost immediately by the formation of ascospores, each of which has only one set of chromosomes.

Complementation tests are nevertheless possible in *Neurospora* owing to the multinucleate nature of the filaments. Certain strains, including those studied by Beadle and Tatum, have the property that when the filaments from two mutant organisms come into physical contact, the filaments fuse and the new filament contains multiple nuclei from both of the

participating partners. This sort of hybrid filament is called a **heterokaryon**, and it contains mutant forms of both genes. The word roots of the term *heterokaryon* mean "different nuclei." (A list of the most common word roots used in genetics can be found at the end of the book.)

When a heterokaryon formed from two nutritional mutants is inoculated into minimal medium, it may grow or it may fail to grow. If it grows in minimal medium, the mutant genes are said to undergo **complementation**, and this result indicates that the mutations are in different genes. On the other hand, if the heterokaryon fails to grow in minimal medium, the result indicates **noncomplementation**, and the two mutations are inferred to be in the same gene.

The inferences from complementation or noncomplementation emerge from the logic illustrated in **FIGURE 1.14**. Here the multinucleate filament is shown, and the mutant nuclei are color coded according to which of two different genes (red or purple) is mutant. The red and purple squiggles represent the proteins encoded in the mutant nuclei, and a "sunburst" symbol represents a defect in the protein resulting from a mutation in the corresponding gene.

Part A depicts the situation in which the mutant strains have mutations in different genes. In the heterokaryon, the red nuclei produce mutant forms of the red protein and normal forms of the purple protein, whereas the purple nuclei produce mutant forms of the purple protein and normal forms of the red protein. The result is that the red/purple heterokaryon has normal forms of both the red and purple proteins. It also has mutant forms of both proteins, but these do not matter. What matters is that the normal proteins allow the heterokaryon to grow on minimal medium because all needed nutrients can be synthesized. In other words, the normal purple gene in the red nucleus complements the defective purple gene in the purple nucleus, and the other way around. The logic of complementation is captured in the ancient nursery rhyme "Jack Sprat could eat no fat / His wife could eat no lean / And so between the two of them / They licked the platter clean," because each partner makes up for the defect in the other.

Part B in Figure 1.14 shows a heterokaryon formed between mutants with defects in the same gene, in this case purple. Both of the purple nuclei encode a normal form of the red protein, but each purple nucleus encodes a defective purple protein. When the nuclei are together, two different mutant forms of the purple protein are produced, and so the biosynthetic pathway that requires the purple protein is still blocked, and the heterokaryon is unable to grow in minimal medium. In other words, the mutants 2 and 3 in Figure 1.14 fail to complement, and so they are judged to have mutations in the same gene.

The following principle underlies the complementation test.

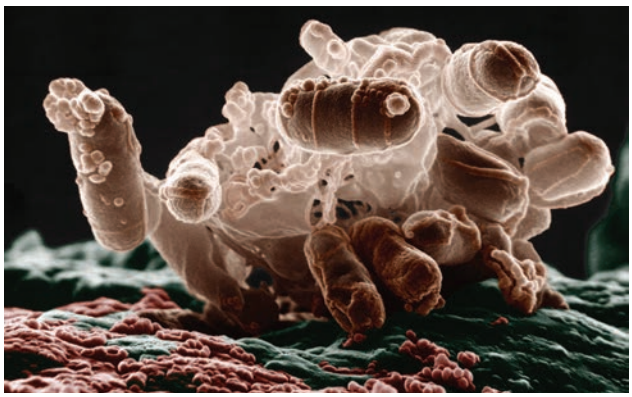
KEY CONCEPT

The Principle of Complementation: A complementation test brings two mutant genes together in the same cell or organism. If this cell or organism is nonmutant, the mutations are said to *complement* one another and it means that the parental strains have mutations in *different* genes. If the cell or organism is mutant, the mutations fail to complement one another, and it means that the parental mutations are in the *same* gene.

A complementation test identifies mutations in the same gene.

In the mutant screen for *Neurospora* mutants requiring arginine, Beadle and Tatum found that mutants in different classes (Class I, Class II, and Class III in Figure 1.13) always complemented one another. This result makes sense, because the genes in each class encode enzymes that act at different levels between the known intermediates. However, some of the mutants in Class I failed to complement others in Class I, and some in Class III failed to complement others in Class III. These results allow the number of genes in each class to be identified.

To illustrate this aspect of genetic analysis, we consider six mutant strains in Class III. These strains were taken in pairs to form heterokaryons and their growth on minimal medium assessed. The data are shown in **FIGURE 1.15A**. The mutant genes in the six strains are denoted *x*1, *x*2, and so forth, and the data are presented in the form of a matrix in which + indicates growth in minimal medium (complementation) and − indicates



Scanning electron micrograph of cells of the bacterium *Escherichia coli*. Some of the foundation studies in modern molecular genetics were carried out with this organism. [Courtesy of Eric Erbe, colorization by Christopher Pooley/USDA ARS.]

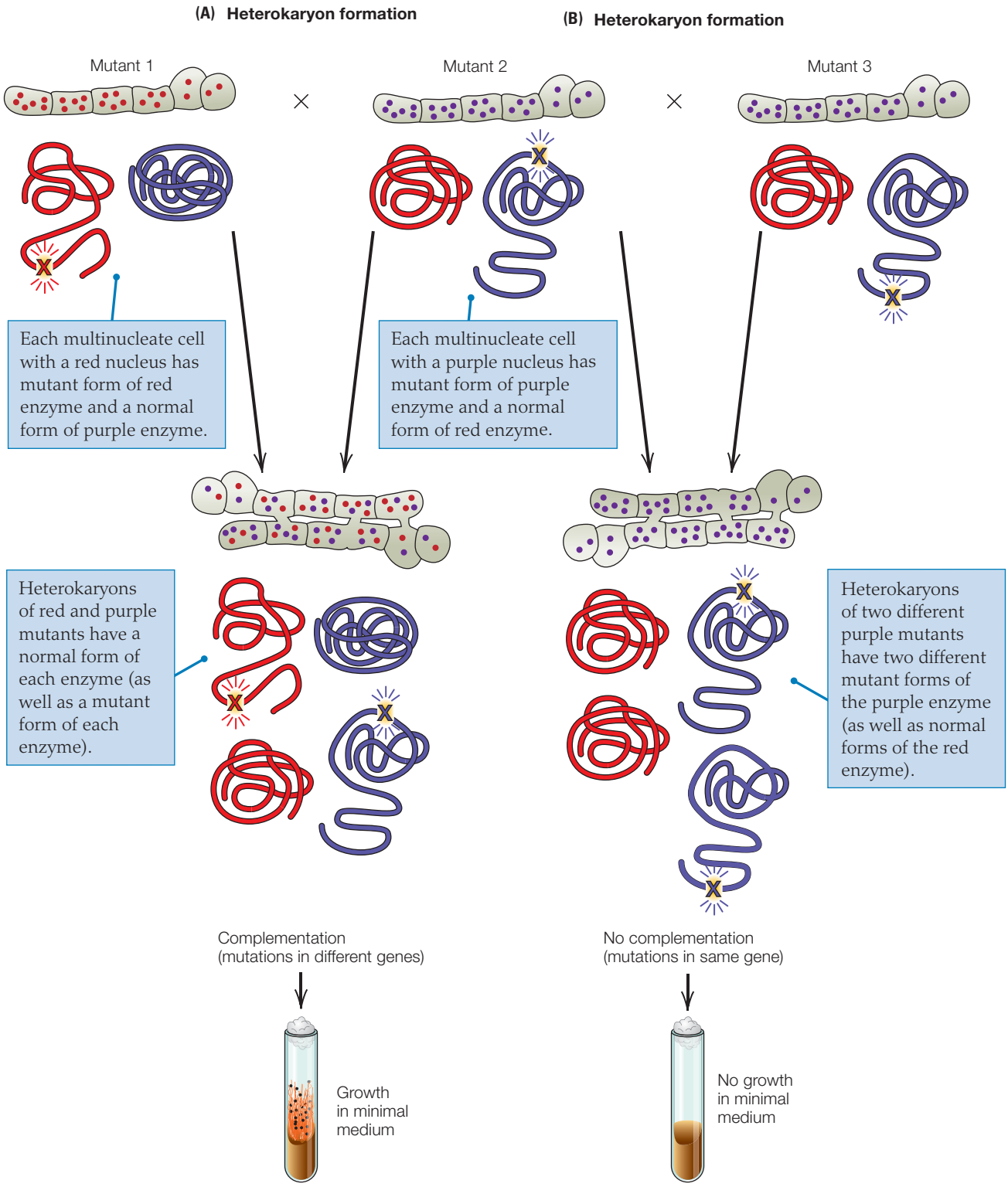


FIGURE 1.14 Molecular interpretation of a complementation test using heterokaryons to determine whether two mutant strains have mutations in different genes (A) or mutations in the same gene (B). In (A) each nucleus contributes a nonmutant form of one or the other polypeptide chain, and so the heterokaryon is able to grow in minimal medium. In (B) both nuclei contribute a mutant form of the same polypeptide chain; hence, no nonmutant form of that polypeptide can be synthesized and the heterokaryon is unable to grow in minimal medium.

lack of growth in minimal medium (lack of complementation). The diagonal entries are all −, which reflects the fact that two copies of the identical mutation cannot show complementation. The pattern of + and − signs in the matrix indicate that mutations *x*1 and *x*5 fail to complement one another; hence, *x*1 and *x*5 are mutations in the same gene. Likewise, mutations *x*2, *x*3, *x*4, and *x*6 fail to complement one another in all possible pairs; hence, *x*2, *x*3, *x*4, and *x*6 are all mutations in the same gene (but a different gene from that represented by *x*1 and *x*5).

Data in a complementation matrix can conveniently be analyzed by arranging the mutant genes in the form of a circle as shown in **FIGURE 1.15B**. Then, for each possible pair of mutations, connect the pair by a straight line if the mutations *fail* to complement (− signs in part A). According to the principle of complementation, these lines connect mutations that are in the same gene. Each of the groups of noncomplementing mutations is called a **complementation group**. As we have seen, each complementation group defines a gene, so the complementation test actually provides the geneticist's operational definition:

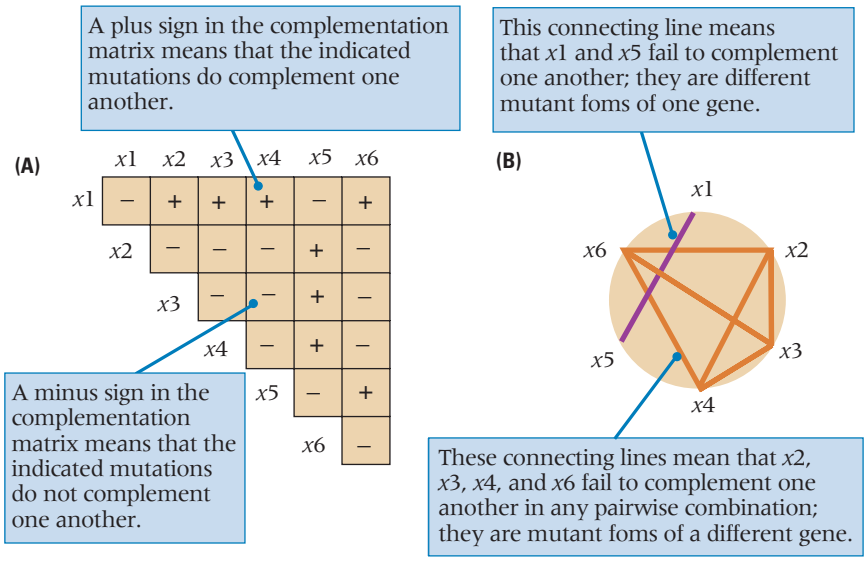


FIGURE 1.15 (A) Results of complementation tests. (B) To interpret the results, arrange the mutations in a circle. Connect by a straight line any pair of mutations that fail to complement—that is, that yield a mutant heterokaryon. Any pair of mutations connected by a straight line are mutations in the same gene, and are more than likely mutations at different nucleotide sites in the gene. This example shows two complementation groups, each of which represents a single gene needed for arginine biosynthesis.

KEY CONCEPT

A *gene* is defined experimentally as a set of mutations that make up a single complementation group. Any pair of mutations within a complementation group fail to complement one another.

The mutations in Figure 1.15, therefore, represent two genes, mutation of any one of which results in the inability of the strain to convert citrulline to arginine. On the basis of the one gene–one enzyme hypothesis, which is largely true for metabolic enzymes in *Neurospora*, the pathway from citrulline to arginine in Figure 1.13 must comprise two steps with an unknown intermediate in between. This intermediate was later found to be argininosuccinate. Likewise, Class I mutants defined three complementation groups; hence, there are three enzymatic steps from the precursor to ornithine. These intermediates were also soon identified. Finally, Class II mutations all failed to complement one another, and the finding of only one complementation group means that there is but a single enzymatic step that converts ornithine to citrulline.

Genetic analysis can be applied to the study of any complex biological process.

The type of genetic analysis pioneered by Beadle and Tatum is immensely powerful for identifying the genetic control of complex biological processes. Their approach lays out a systematic path—a sort of recipe—for gene discovery. First, decide what process you want to study. Next figure out what characteristics mutant organisms with a disruption in that process would display. Then do a mutant screen for mutants showing these characteristics. Carry out complementation tests to find out how many different genes that you have identified. And finally, find out what the products of those genes are, what they do, how they interact with each other, and in what order they function.

Beadle and Tatum themselves analyzed many metabolic pathways for a wide variety of essential nutrients, but their experiments were especially important in deciphering pathways of amino acid biosynthesis. Their findings over just a few years are said to have "contributed more knowledge of amino acid biosynthetic pathways than had been accumulated during decades of traditional study." They were awarded the 1958 Nobel Prize in Physiology or Medicine for their research, and in the intervening years at least five more Nobel Prizes in Physiology or Medicine were awarded in which genetic analysis carried out along the lines of Beadle and Tatum played a significant role. Here is a list,

with quotations from the official citations of the Nobel Foundation.

- 1958—George Beadle and Edward Tatum "for their discovery that genes act by regulating definite chemical events." (It was in doing literature research for his Nobel Prize Lecture that Beadle discovered Garrod's earlier work and brought it to the world's attention.)
- 1965—François Jacob, André Lwoff, and Jacques Monod "for their discoveries concerning genetic control of enzyme and virus synthesis"
- 1995—Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus "for their discoveries concerning the genetic control of early embryonic development"
- 2000—Leland H. Hartwell, Tim Hunt, and Sir Paul Nurse "for their discoveries of key regulators of the cell cycle"
- 2002—Sydney Brenner, H. Robert Horvitz, and John E. Sulston "for their discoveries concerning genetic regulation of organ development and programmed cell death"
- 2007—Mario R. Capecchi, Martin J. Evans, and Oliver Smithies "for their discoveries of principles for introducing specific gene modifications in mice by the use of embryonic stem cells"

1.4 Genes specify proteins by means of a genetic code.

The Beadle and Tatum experiments established that a gene specifies the structure of an enzyme but left open the issue of how this happens. We now know that the relationship between genes and proteins is indirect. The genetic information that specifies a protein is actually contained in the sequence of bases in DNA in a manner analogous to letters printed on a strip of paper. In a region of DNA that directs the synthesis of a protein, the genetic code for the protein is contained in only one strand, and it is decoded in a linear order. The result of protein synthesis is a polypeptide chain, which consists of a linear sequence of amino acids connected end to end. Each polypeptide chain folds into a characteristic three-dimensional configuration that is determined by its particular sequence of amino acids. A typical protein is made up of one or more polypeptide chains. Many proteins function as enzymes that participate in metabolic processes such as amino acid biosynthesis.

One of the DNA strands directs the synthesis of a molecule of RNA.

The details of how genes code for proteins were not understood until the 1960s, and an outline of the process is shown in **FIGURE 1.16**. The decoding of the genetic information takes place in two distinct steps known as *transcription* and *translation*. The indirect route of information transfer

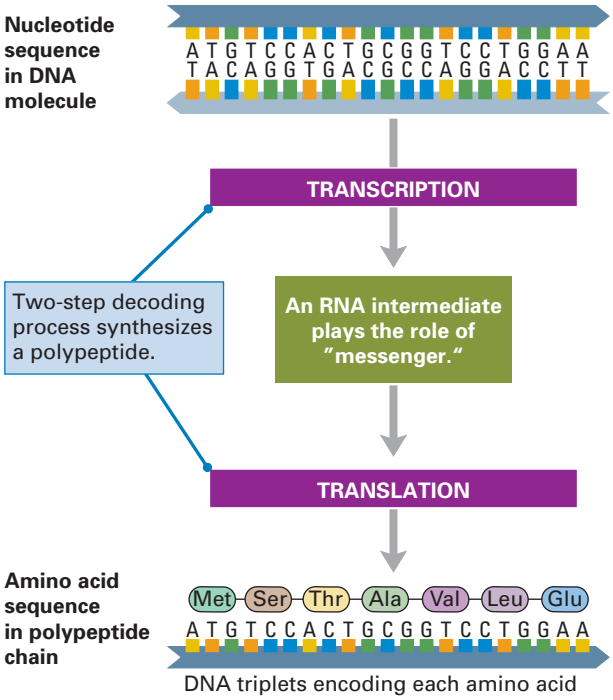


FIGURE 1.16 DNA sequence coding for the first seven amino acids in a polypeptide chain. The DNA sequence specifies the amino acid sequence through a molecule of RNA that serves as an intermediary "messenger." Although the decoding process is indirect, the net result is that each amino acid in the polypeptide chain is specified by a group of three adjacent bases in the DNA. In this example, the polypeptide chain is that of phenylalanine hydroxylase (PAH).

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

is known as the **central dogma** of molecular genetics. The term *dogma* means "set of beliefs"; it dates from the time the idea was first put forward as a theory. Since then the "dogma" has been confirmed experimentally, but the term persists. The main concept in the central dogma is that DNA does not code for protein directly but rather acts through an intermediary molecule of **ribonucleic acid (RNA)**. The structure of RNA is similar to, but not identical with, that of DNA. There is a difference in the sugar (RNA contains the sugar **ribose** instead of deoxyribose), RNA is usually single-stranded (not a duplex), and RNA contains the base **uracil (U)** instead of thymine (T), which is present in DNA. Three types of RNA take part in the synthesis of proteins:

- A molecule of **messenger RNA (mRNA)**, which carries the genetic information from DNA and is used as a template for polypeptide synthesis. In most mRNA molecules, a relatively high

proportion of the nucleotides actually code for amino acids. For example, the mRNA for phenylalanine hydroxylase is 2400 nucleotides in length and codes for a polypeptide of 452 amino acids; in this case, more than 50 percent of the length of the mRNA codes for amino acids.

- Four types of **ribosomal RNA (rRNA)**, which are major constituents of the cellular particles called **ribosomes** on which polypeptide synthesis takes place.
- A set of about 45 **transfer RNA (tRNA)** molecules, each of which carries a particular amino acid as well as a three-base recognition region that base-pairs with a group of three adjacent bases in the mRNA. As each tRNA participates in translation, its amino acid becomes the terminal subunit of the growing polypeptide chain. A tRNA that carries methionine is denoted $tRNA^{Met}$, one that carries serine is denoted $tRNA^{Ser}$, and so forth. (Because there are more than 20 different tRNAs, but only 20 amino acids, some amino acids can be attached to any of several tRNAs.)

The central dogma **FIGURE 1.17** is the fundamental principle of molecular genetics because it summarizes how the genetic information in DNA becomes expressed in the amino acid sequence in a polypeptide chain.

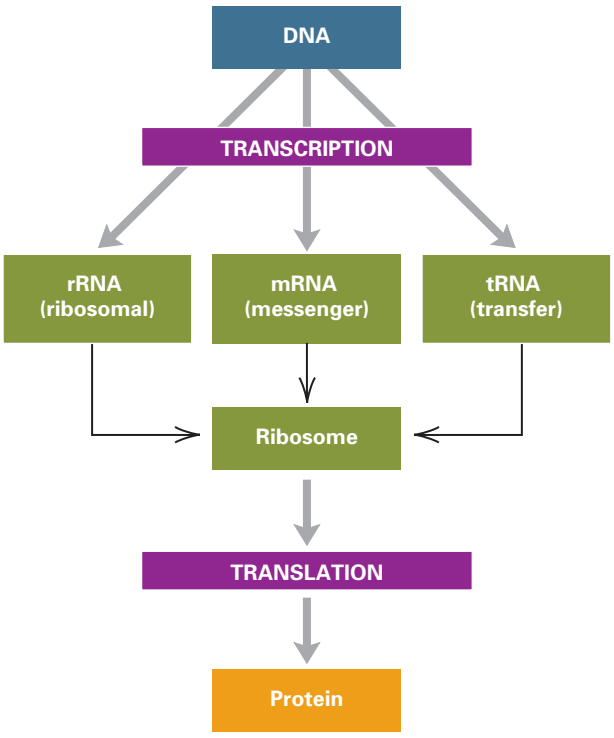


FIGURE 1.17 The "central dogma" of molecular genetics: DNA codes for RNA, and RNA codes for proteins. The DNA → RNA step is transcription, and the RNA → protein step is translation.

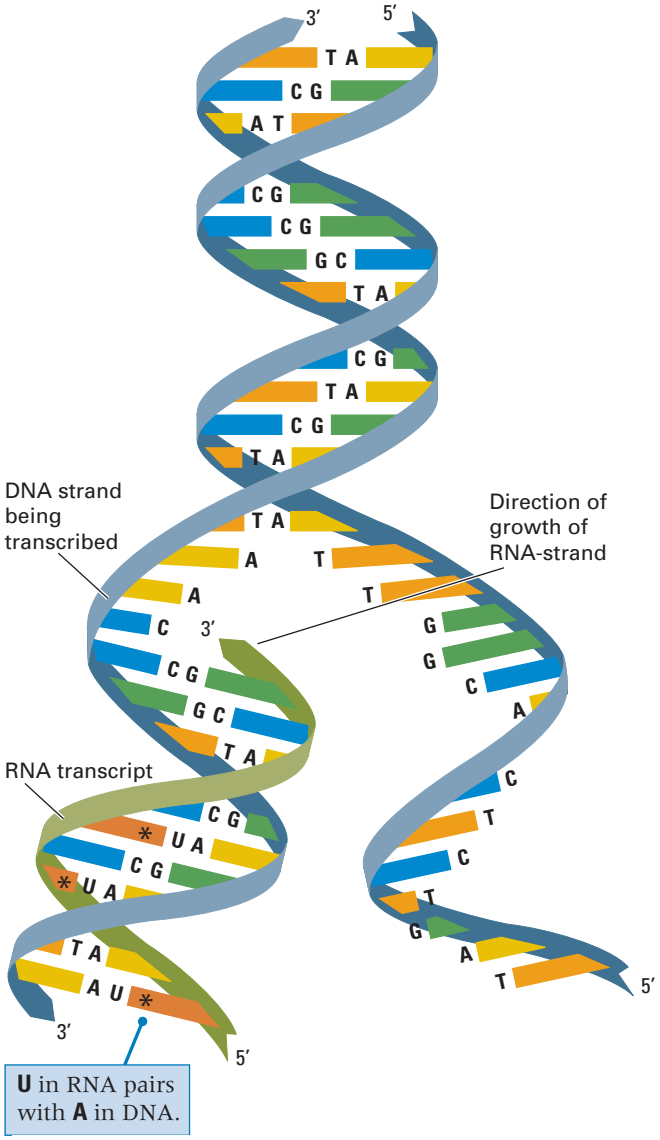


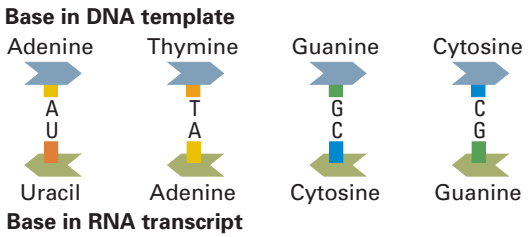
FIGURE 1.18 Transcription is the production of an RNA strand that is complementary in base sequence to a DNA strand. In this example, a DNA strand is being transcribed into an RNA strand at the bottom left. Note that in an RNA molecule, the base U (uracil) plays the role of T (thymine) in that it pairs with A (adenine). Each A–U pair is marked with an asterisk.

KEY CONCEPT

The sequence of nucleotides in a gene specifies the sequence of nucleotides in a molecule of messenger RNA; in turn, the sequence of nucleotides in the messenger RNA specifies the sequence of amino acids in the polypeptide chain.

The manner in which genetic information is transferred from DNA to RNA is shown in **FIGURE 1.18**. The DNA opens up, and one of the strands is used as a template for the synthesis of a complementary strand of RNA.

The process of making an RNA strand from a DNA template is **transcription**, and the RNA molecule that is made is the **transcript**. The base sequence in the RNA is complementary (in the Watson–Crick pairing sense) to that in the DNA template, except that U (which pairs with A) is present in the RNA in place of T. The rules of base pairing between DNA and RNA are summarized below.



Like DNA, an RNA strand also exhibits polarity, its 5′ and 3′ ends determined by the orientation of the nucleotides. The 5′ end of the RNA transcript is synthesized first, and in the RNA–DNA duplex formed in transcription, the polarity of the RNA strand is opposite to that of the DNA strand. Each gene includes particular nucleotide sequences that initiate and terminate transcription. The RNA transcript made from any gene begins at an initiation site in the template strand, which is located "upstream" from the amino-acid coding region, and ends at a termination site, which is located "downstream" from the amino-acid coding region. For any gene, the length of the RNA transcript is very much smaller than the length of the DNA in the entire chromosome. For example, the transcript of the *PAH* gene for phenylalanine hydroxylase is 90,000 nucleotides in length, but the DNA in all of chromosome 12 is about 130,000,000 nucleotide pairs. In this case, the length of the *PAH* transcript is less than 0.1 percent of the length of the DNA in the chromosome. A different gene in chromosome 12 would be transcribed from a different region of the DNA molecule in chromosome 12, but the transcribed region would again be small in comparison with the total length of the DNA in the chromosome.

A molecule of RNA directs the synthesis of a polypeptide chain.

The synthesis of a polypeptide under the direction of an mRNA molecule is known as **translation**. Although the sequence of bases in the mRNA codes for the sequence of amino acids in a polypeptide, the molecules that actually do the "translating" are the tRNA molecules. The mRNA molecule is translated in nonoverlapping groups of three bases called **codons**. For each codon in the mRNA that specifies an amino acid, there is one tRNA molecule containing a complementary group of three adjacent bases that can pair with the bases in the codon. The correct amino acid is attached to the other end of the tRNA, and when this tRNA comes into line, the amino acid attached to it becomes the new terminal end of the growing polypeptide chain.

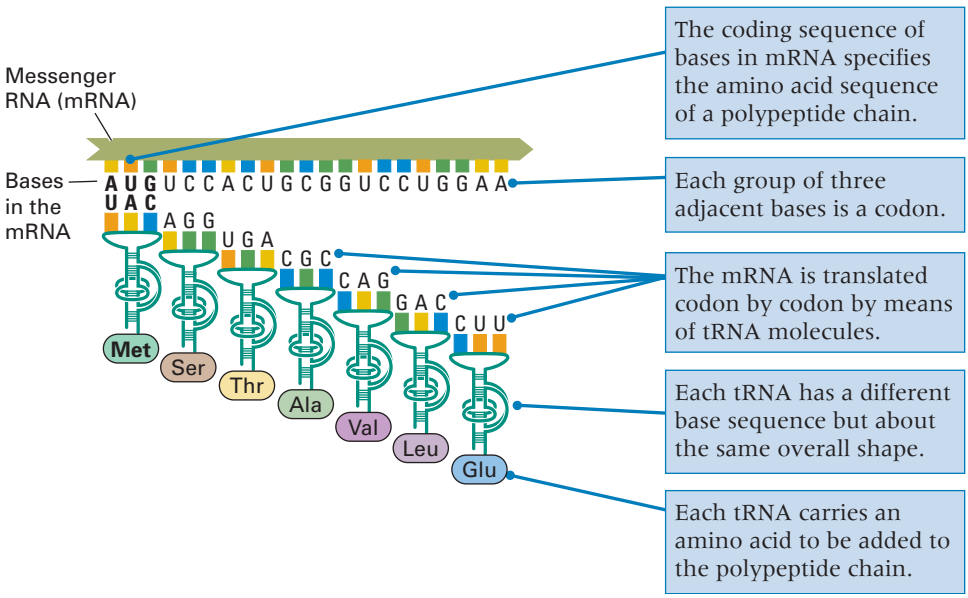


FIGURE 1.19 The role of messenger RNA in translation is to carry the information contained in a sequence of DNA bases to a ribosome, where it is translated into a polypeptide chain. Translation is mediated by transfer RNA (tRNA) molecules, each of which can base-pair with a group of three adjacent bases in the mRNA. Each tRNA also carries an amino acid, and when it is brought to the ribosome by base pairing, its amino acid becomes the growing end of the polypeptide chain.

The role of tRNA in translation is illustrated in **FIGURE 1.19** and can be described as follows:

KEY CONCEPT

The mRNA is read codon by codon. Each codon that specifies an amino acid matches with a complementary group of three adjacent bases in a single tRNA molecule. One end of the tRNA is attached to the correct amino acid, so the correct amino acid is brought into line.

The tRNA molecules used in translation do not line up along the mRNA simultaneously as shown in Figure 1.19. The process of translation takes place on a ribosome, which combines with a single mRNA and moves along it in steps, three nucleotides at a time (codon by codon). As each new codon comes into place, the next tRNA binds with the ribosome, and the growing end of the polypeptide chain becomes attached to the amino acid on the tRNA. In this way, each tRNA in turn serves temporarily to hold the polypeptide chain as it is being synthesized. As the polypeptide chain is transferred from each tRNA to the next in line, the tRNA that previously held the polypeptide is released from the ribosome. The polypeptide chain elongates one amino acid at a step until any one of three particular codons specifying "stop" is encountered. At this point, synthesis of the chain of amino acids is finished, and the polypeptide chain is released from the ribosome.

The genetic code is a triplet code.

Figure 1.19 indicates that the mRNA codon AUG specifies methionine (Met) in the polypeptide chain, UCC specifies Ser (serine), ACU specifies Thr (threonine), and so on. The complete decoding table is called the **genetic code**, and it is shown in **TABLE 1.1**. For any codon, the column on the left corresponds to the first nucleotide in the codon (reading from the 5′ end), the row across the top corresponds to the second nucleotide, and the column on the right corresponds to the third nucleotide. The complete codon is given in the body of the table, along with the amino acid (or "stop") that the codon specifies. Each amino acid is designated by its full name as well as by a three-letter abbreviation and a single-letter abbreviation. Both types of abbreviations are used in molecular genetics. The code in Table 1.1 is the "standard" genetic code used in translation in the cells of nearly all organisms.

In addition to the 61 codons that code only for amino acids, there are 4 codons that have specialized functions:

- The codon AUG, which specifies Met (methionine), is also the "start" codon for polypeptide synthesis. The positioning of a $tRNA^{Met}$ bound to AUG is one of the first steps in the initiation of polypeptide

TABLE 1.1

The Standard Genetic Code

Second Nucleotide in Codon

First nucleotide in codon (5′ end)	U				C				A				G				Third nucleotide in codon (3′ end)
U	UUU	Phe	F	*Phenylalanine*	UCU	Ser	S	*Serine*	UAU	Tyr	Y	*Tyrosine*	UGU	Cys	C	*Cysteine*	U
	UUC	Phe	F	*Phenylalanine*	UCC	Ser	S	*Serine*	UAC	Tyr	Y	*Tyrosine*	UGC	Cys	C	*Cysteine*	C
	UUA	Leu	L	*Leucine*	UCA	Ser	S	*Serine*	UAA		Termination		UGA		Termination		A
	UUG	Leu	L	*Leucine*	UCG	Ser	S	*Serine*	UAG		Termination		UGG	Trp	W	*Tryptophan*	G
C	CUU	Leu	L	*Leucine*	CCU	Pro	P	*Proline*	CAU	His	H	*Histidine*	CGU	Arg	R	*Arginine*	U
	CUC	Leu	L	*Leucine*	CCC	Pro	P	*Proline*	CAC	His	H	*Histidine*	CGC	Arg	R	*Arginine*	C
	CUA	Leu	L	*Leucine*	CCA	Pro	P	*Proline*	CAA	Gln	Q	*Glutamine*	CGA	Arg	R	*Arginine*	A
	CUG	Leu	L	*Leucine*	CCG	Pro	P	*Proline*	CAG	Gln	Q	*Glutamine*	CGG	Arg	R	*Arginine*	G
A	AUU	Ile	I	*Isoleucine*	ACU	Thr	T	*Threonine*	AAU	Asn	N	*Asparagine*	AGU	Ser	S	*Serine*	U
	AUC	Ile	I	*Isoleucine*	ACC	Thr	T	*Threonine*	AAC	Asn	N	*Asparagine*	AGC	Ser	S	*Serine*	C
	AUA	Ile	I	*Isoleucine*	ACA	Thr	T	*Threonine*	AAA	Lys	K	*Lysine*	AGA	Arg	R	*Arginine*	A
	AUG	Met	M	*Methionine*	ACG	Thr	T	*Threonine*	AAG	Lys	K	*Lysine*	AGG	Arg	R	*Arginine*	G
G	GUU	Val	V	*Valine*	GCU	Ala	A	*Alanine*	GAU	Asp	D	*Aspartic acid*	GGU	Gly	G	*Glycine*	U
	GUC	Val	V	*Valine*	GCC	Ala	A	*Alanine*	GAC	Asp	D	*Aspartic acid*	GGC	Gly	G	*Glycine*	C
	GUA	Val	V	*Valine*	GCA	Ala	A	*Alanine*	GAA	Glu	E	*Glutamic acid*	GGA	Gly	G	*Glycine*	A
	GUG	Val	V	*Valine*	GCG	Ala	A	*Alanine*	GAG	Glu	E	*Glutamic acid*	GGG	Gly	G	*Glycine*	G

Codon

Three-letter and single-letter abbreviations

A MOMENT TO THINK 2

Problem 2: A 1975 Nobel Prize was awarded to David Baltimore and Howard Temin for their discovery of an enzyme called *reverse transcriptase*, which can produce a complementary DNA strand from a template of RNA. Immediately recognized as one of the most important enzymes ever discovered, reverse transcriptase could account for the ability of certain animal viruses whose genetic material is RNA to create a complementary DNA molecule that can become inserted into the genetic material of an infected cell. Reverse transcription is similar to ordinary transcription in that the RNA template and DNA transcript are antiparallel and that the DNA transcript grows by the addition of successive nucleotides to the 3′ end. Shown here is part of the RNA sequence of the virus HIV-1 (human immunodeficiency virus-1), the causative agent of AIDS (acquired immune deficiency syndrome), isolated from an infected child in Italy. The sequence encodes part of the HIV-1 reverse transcriptase and is itself reverse-transcribed.

5′-UCCUAUUGAAACUGUACCAGUAAAAUU-3′

What single-stranded DNA sequence would be reverse-transcribed from this stretch of RNA? Would the DNA transcript strand grow from left to right or from right to left? (The answer can be found at the end of the chapter.)

synthesis, so all polypeptide chains begin with Met. In most organisms, the $tRNA^{Met}$ used for initiation of translation is the same $tRNA^{Met}$ used to specify methionine at internal positions in a polypeptide chain.

- The codons UAA, UAG, and UGA, each of which is a "stop," specify the termination of translation and result in release of the completed polypeptide chain from the ribosome. These codons do not have tRNA molecules that recognize them but are instead recognized by protein factors that terminate translation.

How the genetic code table is used to infer the amino acid sequence of a polypeptide chain may be illustrated using phenylalanine hydroxylase again, in particular the DNA sequence coding for amino acid numbers 1 through 7. The DNA sequence is

5′-ATGTCCACTGCGGTCCTGGAA-3′
3′-TACAGGTGACGCCAGGACCTT-5′

This region is transcribed into RNA in a left-to-right direction, and because RNA grows by the addition of successive nucleotides to the 3′ end (Figure 1.18), it is the bottom strand that is transcribed. The nucleotide sequence of the RNA is that of the top strand of the DNA, except that U replaces T, so the mRNA for amino acids 1 through 7 is

5′-AUGUCCACUGCGGUCCUGGAA-3′

The codons are read from left to right according to the genetic code shown in Table 1.1. Codon AUG codes for Met (methionine), UCC codes for Ser (serine), and so on. Altogether, the amino acid sequence of this region of the polypeptide is

5′-AUGUCCACUGCGGUCCUGGAA-3′
MetSerThrAlaValLeuGlu

or, in terms of the single-letter abbreviations,

5′-AUGUCCACUGCGGUCCUGGAA-3′
M S T A V L E

The full decoding operation for this region of the *PAH* gene is shown in **FIGURE 1.20**. In this figure, the initiation codon AUG is highlighted because some patients with PKU have a mutation in this particular codon. As might be expected from the fact that AUG is the initiation codon for polypeptide synthesis, cells in patients with this particular mutation fail to produce any of the PAH polypeptide. Mutation and its consequences are considered next.

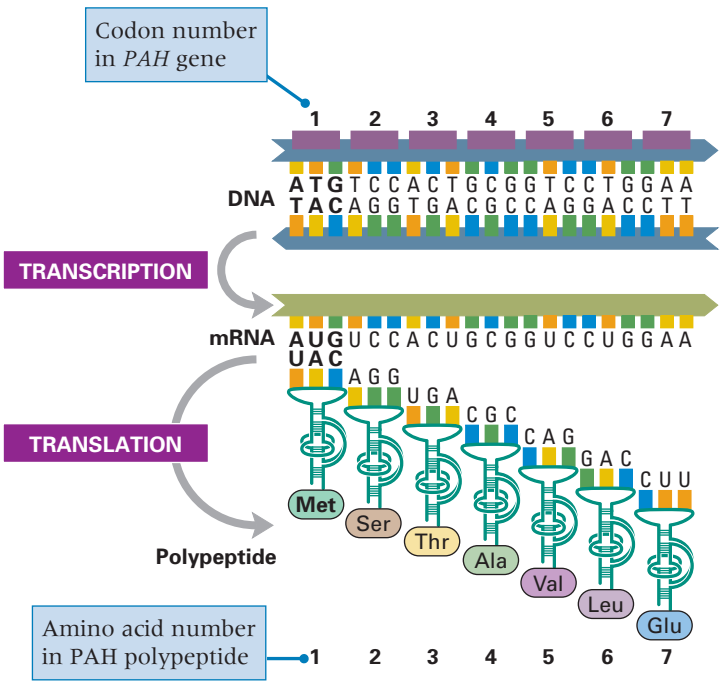


FIGURE 1.20 The central dogma in action. The DNA that encodes PAH serves as a template for the production of a messenger RNA, and the mRNA, in turn, serves to specify the sequence of amino acids in the PAH polypeptide chain through interactions with the tRNA molecules.

1.5 Genes change by mutation.

The term **mutation** refers to any heritable change in a gene (or, more generally, in the genetic material); the term also refers to the process by which such a change takes place. One type of mutation results in a change in the sequence of bases in DNA. The change may be simple, such as the substitution of one pair of bases in a duplex molecule for a different pair of bases. For example, a C—G pair in a duplex molecule may mutate to T—A, A—T, or G—C. The change in base sequence may also be more complex, such as the deletion or addition of base pairs. Geneticists also use the term **mutant**, which refers to the result of a mutation. A mutation yields a mutant gene, which in turn produces a mutant mRNA, a mutant protein, and finally a mutant organism that exhibits the effects of the mutation—for example, an inborn error of metabolism.

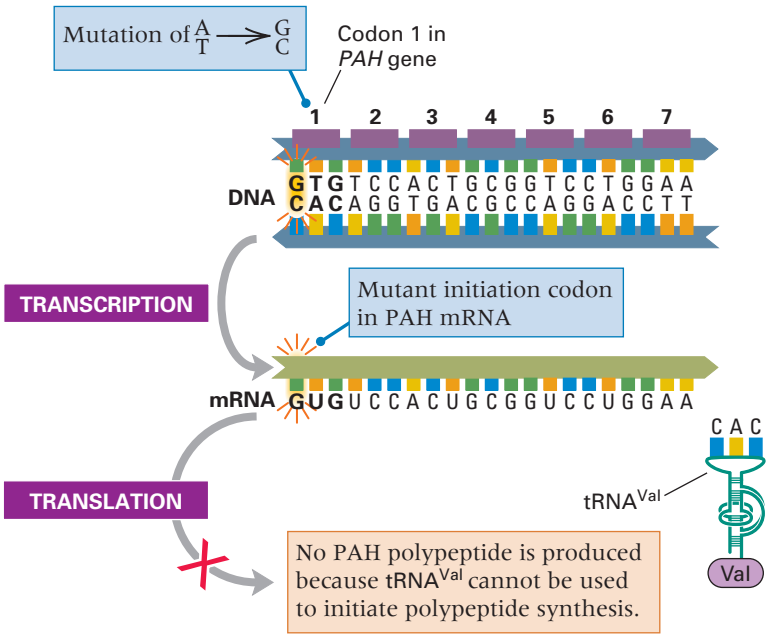


FIGURE 1.21 The M1V mutant in the *PAH* gene. The methionine codon needed for initiation mutates to a codon for valine. Translation cannot be initiated, and no PAH polypeptide is produced.

DNA from patients from all over the world who have phenylketonuria has been studied to determine what types of mutations are responsible for the inborn error. There are a large variety of mutant types. More than 400 different mutations have been described. In some cases part of the gene is missing, so the genetic information to make a complete PAH enzyme is absent. In other cases the genetic defect is more subtle, but the result is still either the failure to produce a PAH protein or the production of a PAH protein that is inactive. In the mutation shown in **FIGURE 1.21**, substitution of a G—C base pair for the normal A—T base pair at the very first position in the coding sequence changes the normal codon AUG (Met) used for the initiation of translation into the codon GUG, which normally specifies valine (Val) and cannot be used as a "start" codon. The result is that translation of the PAH mRNA cannot occur, so no PAH polypeptide is made. This mutant is designated M1V because the codon for M (methionine) at amino acid position 1 in the PAH polypeptide has been changed to a codon for V (valine). Although the M1V mutant is quite rare worldwide, it is common in some localities, such as in Québec Province in Canada.

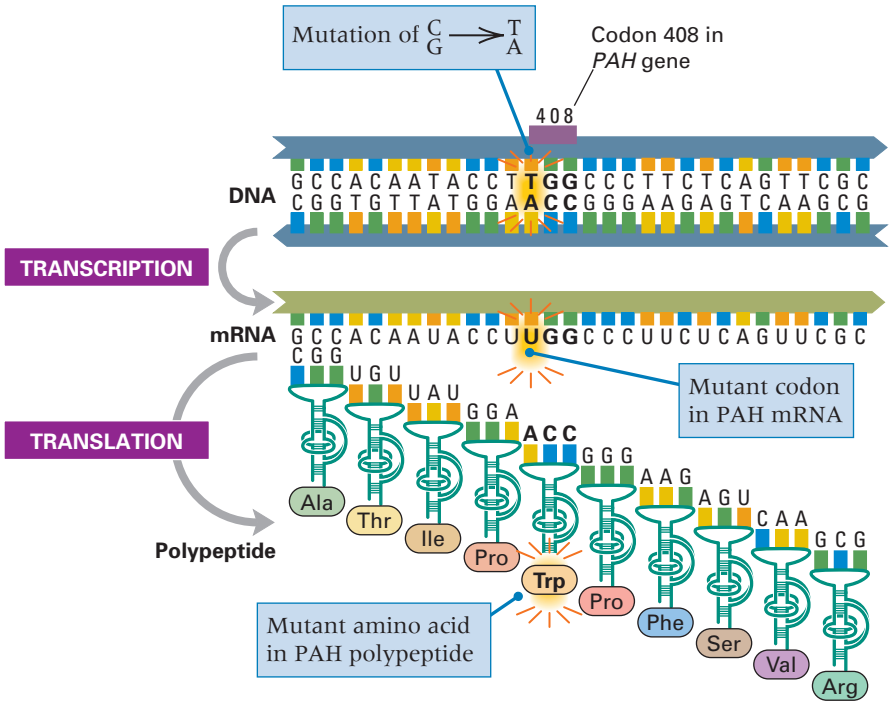


FIGURE 1.22 The R408W mutant in the *PAH* gene. Codon 408 for arginine (R) is mutated into a codon for tryptophan (W). The result is that position 408 in the mutant PAH polypeptide is occupied by tryptophan rather than by arginine. The mutant protein has no PAH enzyme activity.

One PAH mutant that is quite common is designated R408W, which means that codon 408 in the PAH polypeptide chain has been changed from one coding for arginine (R) to one coding for tryptophan (W). This mutant is one of the four most common in cases of PKU among European Caucasians. The molecular basis of the mutation is shown in **FIGURE 1.22**. In this case, the first base pair in codon 408 is changed

from a C—G base pair into a T—A base pair. The result is that the PAH mRNA has a mutant codon at position 408; specifically, it has UGG instead of CGG. Translation does occur in this mutant because everything else about the mRNA is normal, but the result is that the mutant PAH carries a tryptophan (Trp) instead of an arginine (Arg) at position 408 in the polypeptide chain. The consequence of the seemingly minor change of one amino acid is very drastic, because the mutant PAH has no enzyme activity and so is unable to catalyze its metabolic reaction. In other words, the mutant PAH protein is complete but inactive. With PAH, as with other proteins, some amino acid replacements result in a polypeptide chain that is unable to fold properly. The incorrectly folded polypeptides are digested by proteases in the cell, which recycles the amino acids for use in the synthesis of other proteins.

1.6 Traits are affected by environment as well as by genes.

Inborn errors of metabolism illustrate the general principle that genes code for proteins and that mutant genes code for mutant proteins. In cases such as PKU, mutant proteins cause such a drastic change in metabolism that a severe genetic defect results. But biology is not necessarily destiny. Organisms are also affected by the environment. PKU serves as an example of this principle, because patients who adhere to a diet restricted in the amount of phenylalanine develop mental capacities within the normal range. What is true in this example is true in general. Most traits are determined by the interaction of genes and environment.

It is also true that most traits are affected by multiple genes. No one knows how many genes are involved in the development and maturation of the brain and nervous system, but the number must be in the thousands. This number is in addition to the genes that are required in all cells to carry out metabolism and other basic life functions. It is easy to lose sight of the multiplicity of genes when considering extreme examples, such as PKU, in which a single mutation can have such a drastic effect on mental development. The situation is the same as that with any complex machine. An airplane can function only if thousands of parts are working together in harmony, but it takes only one defective part, if it affects a vital system, to bring it down. Likewise, the development and functioning of every trait require a large number of genes working in harmony, but in some cases a single mutant gene can have catastrophic consequences.

In other words, the relationship between a gene and a trait is not necessarily simple. Three principles govern the relationships between genes and traits:

1. One gene can affect more than one trait through secondary or indirect effects. The various, sometimes seemingly unrelated effects of a mutant gene are called **pleiotropic effects**, and the phenomenon itself is known as **pleiotropy**. For example, among cats with white fur and blue eyes, about 40 percent are born deaf (**FIGURE 1.23**). This form of deafness can be regarded as a pleiotropic effect of white fur and blue eyes.
2. Any trait can be affected by more than one gene. We discussed this principle earlier in connection with the large number of genes that are required for the normal development and functioning of the brain and nervous system. Multiple genes affect even simpler traits such as hair color and eye color.
3. Many traits are affected by environmental factors as well as by genes. Consider again the low-phenylalanine diet. Children with PKU are not "doomed" to severe mental deficiency. Their capabilities can be brought into the normal range by dietary treatment. PKU serves as an example of what motivates geneticists to search to discover the molecular basis of inherited disease. The hope is that knowing the metabolic basis of the disease will eventually make it possible to develop methods for clinical intervention through diet, medication, or other treatments that will reduce the severity of the disease.

FIGURE 1.23 Among cats with white fur and blue eyes, about 40 percent are born deaf. The reason is that pigment cells derived from the neural crest migrate to various tissues including hair follicles, the eyes, and the middle ear, where their function is essential for hearing. Defective pigment cells resulting in white fur and blue eyes can, therefore, lead to deafness, which may be regarded as a pleiotropic effect of white fur and blue eyes. (© Medioimages/Alamy Images)

1.7 Evolution means continuity of life with change.

The pathway for the breakdown and excretion of phenylalanine is by no means unique to human beings. One of the remarkable generalizations to have emerged

from molecular genetics is that organisms that are very distinct—for example, plants and animals—share many features in their genetics and biochemistry. These similarities indicate a fundamental "unity of life":

KEY CONCEPT

All creatures on Earth share many features of the genetic apparatus, including genetic information encoded in the sequence of bases in DNA, transcription into RNA, and translation into protein on ribosomes with the use of transfer RNAs. All creatures also share certain characteristics in their biochemistry, including many enzymes and other proteins that are similar in amino acid sequence, three-dimensional structure, and function.

Groups of related organisms descend from a common ancestor.

Organisms share a common set of similar genes and proteins because they evolved by descent from a common ancestor. The process of **evolution** takes place when a population of organisms gradually changes in genetic composition through time. Evolutionary changes in genes and proteins result in differences in metabolism, development, and behavior among organisms, which allows them to become progressively better adapted to their environments. From an evolutionary perspective, the unity of fundamental molecular processes in organisms alive today reflects inheritance from a distant common ancestor in which the molecular mechanisms were already in place.

Not only the unity of life but also many other features of living organisms become comprehensible from an evolutionary perspective. For example, the interposition of an RNA intermediate in the basic flow of genetic information from DNA to RNA to protein makes sense if the earliest forms of life used RNA for both genetic information and enzyme catalysis. The importance of the evolutionary perspective in understanding aspects of biology that seem pointless or needlessly complex is summed up in a famous aphorism of the evolutionary biologist Theodosius Dobzhansky: "Nothing in biology makes sense except in the light of evolution."

Biologists distinguish three major kingdoms of organisms:

1. ***Bacteria*** This group includes most bacteria and cyanobacteria (formerly called blue-green algae). Cells of these organisms lack a membrane-bounded nucleus and mitochondria, are surrounded by a cell wall, and divide by binary fission.
2. ***Archaea*** This group was initially discovered among microorganisms that produce methane gas or that live in extreme environments, such as hot springs or pools with high salt concentrations. They are widely distributed in more normal environments as well. Superficially resembling bacteria, the cells of archaeans show important differences in the manner in which their membrane lipids are chemically linked. The machinery for DNA replication and transcription in archaeans resembles that of eukaryans, whereas their metabolism strongly resembles that of bacteria. DNA sequence analysis indicates that about half of the genes found in the kingdom Archaea are unique to this group.
3. ***Eukarya*** This group includes all organisms whose cells contain an elaborate network of internal membranes, a membrane-bounded nucleus, and mitochondria. Their DNA is present in the form of linear molecules organized into true chromosomes, and cell division takes place by means of mitosis. The eukaryotes include plants and animals as well as fungi and many single-celled organisms, such as amoebae and ciliated protozoa.

The members of the groups Bacteria and Archaea are often grouped together into a larger assemblage called **prokaryotes**, which literally means "before [the evolution of] the nucleus." This terminology is convenient for designating prokaryotes as a group in contrast with **eukaryotes**, which literally means "good [well-formed] nucleus."

The earliest forms of life may have used RNA for both information storage and enzyme catalysis.

The core cellular processes of bacteria and archaeans are quite similar. Both use double-stranded DNA as their genetic material, transcribe DNA into messenger RNA, and use the messenger RNA to specify the amino acid sequence in proteins by means of the same genetic code. Likewise, most of their metabolic pathways for the synthesis or breakdown of small molecules are similar. These resemblances imply that bacteria and archaeans most likely shared a common ancestor. Because one of the key differences between the groups is in the chemistry of their membrane lipids, it seems likely that bacteria and archaeans had a common ancestor prior to the time that life became cellularized by the synthesis of a surrounding membrane.

The common ancestor of bacteria and archaea (and eukaryotes as well) has a name. It is called **LUCA,** which stands for *last universal common ancestor*. Where did LUCA come from? Some scientists have speculated that it might have come from another planet by way of contaminated space debris such as an asteroid. But if LUCA lacked a membrane, then it must have been a fragile creature; therefore, it probably was home grown from natural chemical and physical processes.

About 9 billion years after the universe came into being, planet Earth was formed through the aggregation

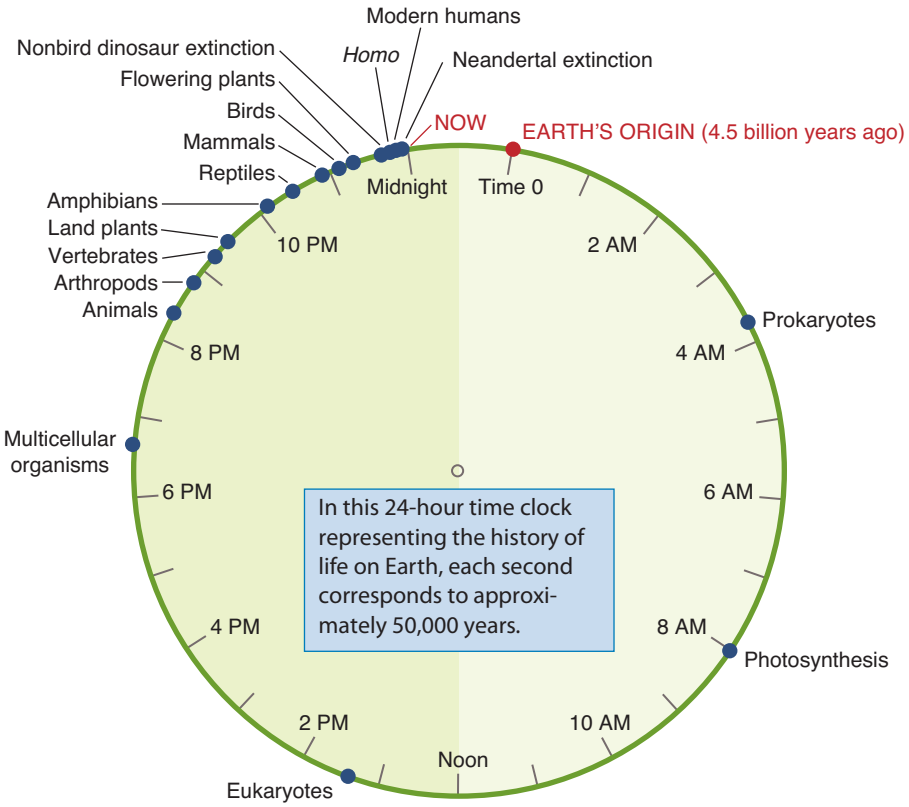


FIGURE 1.24 Timing of major events in life history calibrated to a 24-hour clock beginning with the formation of Earth 4.5 billion years ago.

of small particles of ice and dust that, as it grew larger, was able to attract larger bodies through gravitational attraction. This was about 4.5 billion years ago, and in **FIGURE 1.24** the origin of Earth is shown as the 0 point of a 24-hour clock ending now. Each second of this virtual clock corresponds to about 50 thousand years of real time. The newborn planet is thought to have cooled rather quickly and acquired an atmosphere of gases such as hydrogen, methane, carbon dioxide, ammonia, and nitrogen escaping from volcanoes, with liquid water added by collisions with icy asteroids. Note the almost complete absence of free oxygen, which came much later with the evolution of photosynthesis. The absence of free oxygen means that the atmosphere was reducing, and the ready availability of electrons was conducive to forming chemical bonds. In the early 1950s, Stanley Miller and Harold Urey prepared an apparatus that simulated lighting (electrical sparks) acting on an atmosphere of hydrogen, methane, and ammonia in the presence of liquid water and demonstrated the creation of many small organic molecules including the amino acids alanine, glycine, valine, serine, phenylalanine, aspartic acid, and glutamic acid. These results supported earlier speculation that life originally arose in a sort of rich broth of accumulated, spontaneously synthesized, small organic molecules. The central role of RNA in extant life has given rise to the view that the very earliest forms of life depended on RNA for both information storage and enzyme catalysis, but that later in evolution, as LUCA gradually came into being, the informational role of RNA was taken over by DNA and most of the catalytic role by proteins.

A popular, current hypothesis is that life on Earth may have originated in microscopic porous chambers within giant carbonate spires reaching as high as 10-story buildings formed at alkaline hydrothermal vents near mid-ocean ridges where the sea floor spreads and sea water swoops down and accumulates minerals before heating up and rising again (**FIGURE 1.25**). These tiny carbonate chambers are coated with a thin layer of metal sulfide that in the reducing environment can serve as catalyst to promote chemical reactions among the chemicals in the water moving through. The chambers act to concentrate the reaction products and to partially isolate one chamber from the next. Membranes were not needed as separators until after all of the basic cellular processes had evolved, but their evolution allowed the earliest cells to become free living.

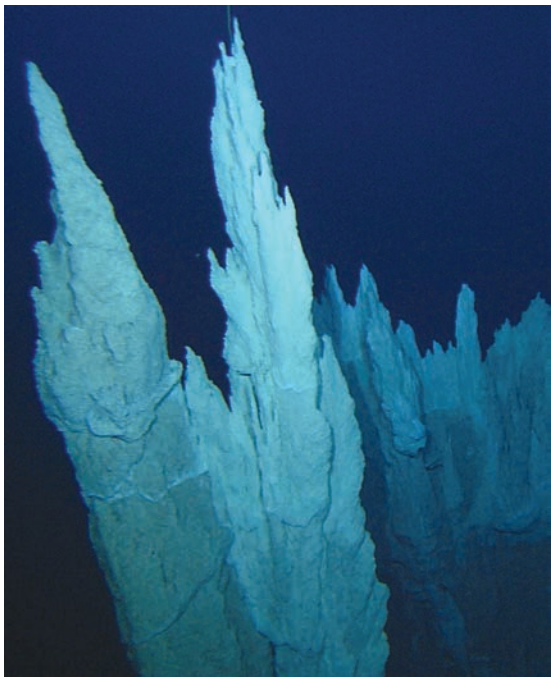


FIGURE 1.25 The core chemical processes of life may have originated among millions of tiny chambers in giant carbonate spires like this, which are formed from hydrothermal vents near mid-ocean ridges.

However and wherever life originated, biologists know something about its chronology. The first evidence of prokaryotes is in the form of fossilized biofilms dating back 3.8 billion years. In the 24-hour virtual clock in Figure 1.24, the time is about 3:45 AM. The earliest fossils of photosynthetic bacteria (cyanobacteria) date to 3 billion year ago, which in our clock is 8:00 AM. At around this time the atmosphere began to change as photosynthetic organisms took up carbon dioxide and water and released free oxygen. Eukaryotes do not appear in the fossil record until about 2 billion years ago, about 1:20 PM on our clock. Eukaryotes appear to have originated as a fusion between an archaean cell and a bacterial cell. The archaean genome eventually evolved into the nucleus, and the bacterial genome into the mitochondria, although most of the bacterial genes were eventually transferred to the nucleus. Evidence for multicellular organisms dates to 1 billion years (6:40 PM on our clock), and photosynthetic plants eventually arose by incorporating a cyanobacterium that became the chloroplast. And so life goes on. The hour before midnight in the virtual clock includes the rise of birds (from dinosaur ancestors), the proliferation of flowering plants, the extinction of the dinosaurs, the debut of the genus Homo that includes our species about 2.5 million years ago (48 seconds before midnight on the clock), anatomically modern humans (3.8 seconds before midnight), and the extinction of our nearest relative the Neandertals (0.5 seconds before midnight). (Neandertals are not completely extinct, as some of the genes in humans alive today derive from interbreeding between our ancestors and the Neandertals.) What about the 5 thousand years of recorded human history? In our clock it is a mere snap of the fingers: one-tenth of a second!

The molecular unity of life is seen in comparisons of genomes and proteomes.

The totality of DNA in a cell, nucleus, or organelle is called its **genome**. When used with reference to a species of organism, for example, in phrases such as "the human genome," the term *genome* is defined as the DNA present in a normal reproductive cell.

Modern methods for sequencing DNA are so rapid and efficient that the complete DNA sequence is known for hundreds of different species of organisms. These include the genomes of multiple representatives of all of the groups of organisms in Figure 1.24, including Neandertals (sequenced from DNA extracted from fossil bones). **TABLE 1.2** shows a small sample of sequenced genomes. Genome size is given in megabases (Mb), or millions of base pairs. The genome of the bacterium *Hemophilus influenzae*, like that of most bacteria, is very compact in that most of it codes for proteins. A high density of genes, relative to the amount of DNA, is also found in the budding yeast, the nematode worm, the fruit fly, and the diminutive flowering plant *Arabidopsis thaliana*. The human genome, by contrast, contains large amounts of noncoding DNA. Comparison with the nematode is illuminating. Whereas the human genome is about 30 times larger than that of the worm, the number of genes is not much different. This discrepancy reflects the fact that only about 1.5 percent of the human genome sequence codes for protein. (About 27 percent of the human genome is present in genes, but much of the DNA sequence present in genes is not protein-coding.)

The complete set of proteins encoded in the genome is known as the **proteome**. In less complex genomes, such as the bacteria and yeast, the number of proteins in an organism's proteome is approximately the same as the number of genes. However, some genes encode two or more proteins through a process called *alternative splicing* in which segments of the original RNA transcript are joined together in a variety of combinations to produce different messenger RNAs. Alternative splicing is especially prevalent in the human genome. At least one-third of human genes, and possibly as many as two-thirds, undergo alternative splicing, and among the genes that undergo alternative splicing,

TABLE 1.2

Comparisons of Genomes and Proteomes

Organism	Genome size, Mb[a] (approximate)	Number of genes (approximate)	Number of distinct proteins in proteome[b] (approximate)	Shared protein families
Hemophilus influenzae (causes bacterial meningitis)	1.9	1700	1400	
Saccharomyces cerevisiae (budding yeast)	13	6000	4400	3000
Caenorhabditis elegans (soil nematode)	100	20,000	9500	5000
Drosophila melanogaster (fruit fly)	120[c]	16,000	8000	7000[d]
Mus musculus (laboratory mouse)	2500	25,000	10,000	9900[f]
Homo sapiens (human being)	2900[e]	25,000	10,000	

[a] Millions of base pairs.
[b] Excludes "families" of proteins with similar sequences (and hence related functions).
[c] Excludes 60 Mb of specialized DNA ("heterochromatin") that has a very low content of genes.
[d] Based on similarity with sequences in messenger RNA (mRNA).
[e] For convenience, this estimate is rounded to 3000 Mb elsewhere in this book.
[f] Based on the observation that only about 1% of mouse genes lack a similar gene in the human genome, and *vice versa*.

the number of different messenger RNAs per gene ranges from 2 to 7. Hence, with its seemingly limited repertoire of 25,000 genes, the human genome can create approximately 60,000–90,000 different mRNAs. The widespread use of alternative splicing to multiply the coding capacity of genes is one source of human genetic complexity.

The molecular unity of life can be seen in the similarity of proteins in the proteome among diverse types of organisms. Such comparisons are shown in the right-hand column of Table 1.2. In this tabulation, each family of proteins that are similar in amino acid sequence is counted only once, in order to estimate the number of proteins in the proteome that are "distinct" in the sense that their sequences are dissimilar. In yeast, worms, and flies, the number of distinct proteins is approximately 4400, 9500, and 8000, respectively. The brackets in Table 1.2 indicate the number of distinct proteins that share sequence similarity between species. From these comparisons, it appears that most multicellular animals share 5,000–10,000 proteins that are similar in sequence and function. Approximately 3000 of these are shared with eukaryotes as distantly related as yeast, and approximately 1000 with prokaryotes as distantly related as bacteria. What these comparisons among proteomes imply is that biological systems are based on protein components numbering in the thousands. This is a challenging level of complexity to understand, but the challenge is much less intimidating than it appeared to be at an earlier time when human cells were thought to produce as many as a million different proteins.

CHAPTER SUMMARY

- Inherited traits are affected by genes.
- Genes are composed of the chemical deoxyribonucleic acid (DNA).
- DNA replicates to form (usually identical) copies of itself.
- DNA contains a code specifying what types of enzymes and other proteins are made in cells.
- DNA occasionally mutates, and the mutant forms specify altered proteins.
- A mutant enzyme is an "inborn error of metabolism" that blocks one step in a biochemical pathway for the metabolism of small molecules.
- Genetic analysis of mutants of the fungus *Neurospora* unable to synthesize an essential nutrient led to the one gene–one enzyme hypothesis.
- Different mutations in the same gene can be identified by means of a complementation test, in which the mutants are brought together in the same cell or organism. Mutations in the same gene fail to complement one another, whereas mutations in different genes show complementation.
- Traits are affected by environment as well as by genes.
- Organisms change genetically through generations in the process of biological evolution.
- Because of their common descent, organisms share many features of their genetics and biochemistry.

LEARNING OUTCOMES

- Given the base sequence in a transcribed strand of protein-coding DNA, specify the sequence of bases in the corresponding region of messenger RNA and the corresponding sequence of amino acids in the protein. For a mutation in which a specified base is replaced with another, deduce the resulting mRNA and protein sequence.
- Given a linear metabolic pathway for an essential nutrient, deduce which intermediates will restore the ability to grow mutant strains that are defective for any of the enzymes in the pathway. Conversely, using data that specify which intermediates in a linear pathway restore the ability of mutants to grow, infer the order of the enzymes and intermediates in the pathway.
- Given data on the complementation or lack of complementation among all pairs of a set of mutations affecting a biological process, sort the mutations into complementation groups, each corresponding to a different gene.

ISSUES AND IDEAS

- What special feature of the structure of DNA allows each strand to be replicated without regard to the other?
- What does it mean to say that a strand of DNA specifies the structure of a molecule of RNA?
- What types of RNA participate in protein synthesis, and what is the role of each type of RNA?
- What is meant by the phrase *the genetic code*, and how is the genetic code relevant to the translation of a polypeptide chain from a molecule of messenger RNA?
- What is meant by the term *genetic analysis*, and how is genetic analysis exemplified by the work of Beadle and Tatum using *Neurospora*?
- What is a complementation test, and what is it used for in genetic analysis?
- In what way does phenylketonuria demonstrate the importance of the environment even for traits that are "determined" by genes?

SOLUTIONS: STEP BY STEP

Problem 1

In the human gene for the beta chain of hemoglobin, the oxygen-carrying protein in the red blood cells, the first 30 nucleotides in the protein-coding region are as shown here.

3′-TACCACGTGGACTGAGGACTCCTCTTCAGA-5′

(a) What is the sequence of the partner strand?
(b) If the DNA duplex of this gene were transcribed from left to right, what is the base sequence of the RNA across this part of the coding region?
(c) What is the sequence of amino acids in this part of the beta-globin polypeptide chain?
(d) In the mutation responsible for sickle-cell anemia, the red T indicated is replaced with an A. The mutant is present at relatively high frequency in some human populations because carriers of the gene are more resistant to falciparum malaria than are noncarriers. What is the amino acid replacement associated with this mutation?

Solution. (a) The partner strand is deduced from the rule that A pairs with T and G pairs with C; however, keep in mind that the paired DNA strands have opposite polarity (that is,

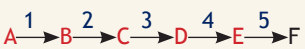
their 5′-to-3′ orientations are reversed). (b) The RNA strand is synthesized in the 5′-to-3′ direction, which means that the template DNA strand is transcribed in the 3′-to-5′ direction, which happens to be the same left-to-right orientation of the strand shown above. The base sequence is deduced from the usual base-pairing rules, except that A in DNA pairs with U in RNA. (c) The polypeptide chain is translated in successive groups of three nucleotides (each group constituting a codon), starting at the 5′ end of the coding sequence in the RNA and moving in the 5′-to-3′ direction. The amino acid corresponding to each codon can be found in the genetic code table. (d) The change from T to A in the transcribed strand alters a GAG codon into a GUG codon in the RNA transcript, resulting in the replacement of the normal glutamic acid (GAG) with valine (V). The nonmutant duplex, the RNA transcript, and the amino acid sequence are as shown below. The amino acid that is replaced in the sickle-cell mutant is indicated in red.

```
3′-TACCACGTGGACTGAGGACTCCTCTTCAGA-5′
5′-ATGGTGCACCTGACTCCTGAGGAGAAGTCT-3′

5′-AUGGUGCACCUGACUCCUGAGGAGAAGUCU-3′
   MetValHisLeuThrProGluGluLysSer
```

Problem 2

The accompanying diagram shows a linear bioynthetic pathway for an essential nutrient designated F in an organism, such as *Neurospora*, able to grow in a minimal medium. Each red letter indicates one intermediate in the pathway, and each blue number indicates a mutant that blocks one step in the pathway.



Make a table in which the columns correspond to the intermediates, arranged in alphabetical order, and the rows correspond to the mutants, arranged in numerical order. In the body of the table, insert a plus sign if the mutant will grow on minimal medium supplemented with the nutrient and a minus sign if the mutant will not grow under these conditions. Assume that all intermediates can be transported into the cell from the growth medium.

Solution. This is a classic type of genetic analysis pioneered by Beadle and Tatum. The principle is that a mutant will grow on any intermediate whose position in the pathway is *downstream* of the metabolic block. Hence, mutant 1 will grow on any intermediate except A, mutant 2 will grow on any intermediate except A or B, and so forth. The complete matrix is as shown. It looks exceptionally simple because both the rows (mutants) and columns (intermediates) are arranged in the same order as their constituents appear in the pathway. Normally this will not be the case.

	A	B	C	D	E	F
1	–	+	+	+	+	+
2	–	–	+	+	+	+
3	–	–	–	+	+	+
4	–	–	–	–	+	+
5	–	–	–	–	–	+

Problem 3

A complementation test is used to sort a set of mutants into groups, each group corresponding to a subset of the mutants that have defects in the same gene. Shown here are the genes (1–5) from the previous problem and 10 mutants (*a*–*j*) grouped according to the gene they affect.

1	2	3	4	5
a	*e*, *g*	*c*, *b*, *h*, *j*	*f*	*d*, *i*

Gene 1 is represented by mutant *a* only, gene 2 by mutants *e* and *g*, and so forth.

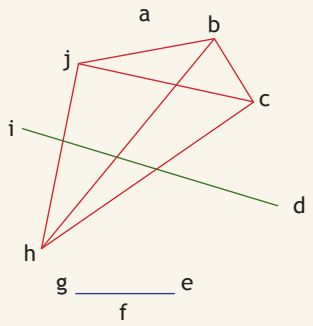
(a) Prepare a square complementation matrix of data, with the rows and columns representing the mutants in alphabetical order. Each entry in the matrix should be a plus sign if the row mutant and the column mutant do show complementation (that is, if they are mutants of different genes) or a minus sign if they do not show complementation (that is, if they are mutants of the same gene).
(b) What is special about the principal diagonal of the matrix? (The principal diagonal is the diagonal that runs from upper left to lower right.) What does this result mean biologically?
(c) What is special about the triangular parts of the matrix above and below the diagonal? What does this result mean biologically?
(d) Prepare a circular diagram of the mutants as discussed in the text, showing which of the mutants form complementation groups.

Solution. (a) The complementation matrix is as shown here. (b) The principal diagonal consists exclusively of minus signs; biologically, this means that a mutant cannot undergo complementation with itself, because two copies of the identical mutation must be in the same gene. (c) The upper and lower triangular matrices are symmetrical, mirror images of one another; biologically, this means that the parent of origin of the mutant makes no difference to whether the mutants undergo complementation. Because of the symmetry of the data matrix, complementation data are often presented only in the form of the upper diagonal. (d) The circular type of the complementation test is also shown. It indicates that the complementation groups are {*a*}, {*b*, *c*, *h*, *j*}, {*d*, *i*}, {*e*, *g*}, and {*f*}. The complementation groups are not informative about where the product of each gene acts in the pathway; this information must come from the type of analysis illustrated in the previous problem.

(a–c)

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
a	–	+	+	+	+	+	+	+	+	+
b	+	–	–	+	+	+	+	–	+	–
c	+	–	–	+	+	+	+	–	+	–
d	+	+	+	–	+	+	+	+	–	+
e	+	+	+	+	–	+	–	+	+	+
f	+	+	+	+	+	–	+	+	+	+
g	+	+	+	+	–	+	–	+	+	+
h	+	–	–	+	+	+	+	–	+	–
i	+	+	+	–	+	+	+	+	–	+
j	+	–	–	+	+	+	+	–	+	–

(d)



CONCEPTS IN ACTION: PROBLEMS FOR SOLUTION

1.1 Prior to the Avery, MacLeod, and McCarty experiment, what features of cells and chromosomes were already known that could have been interpreted as evidence that DNA is an important constituent of the genetic material?

1.2 In the early years of the twentieth century, why did most biologists and biochemists believe that proteins were probably the genetic material?

1.3 From their examination of the structure of DNA, what were Watson and Crick able to infer about the probable mechanisms of DNA replication, coding capability, and mutation?

1.4 What are three principal structural differences between RNA and DNA?

1.5 A region along an RNA transcript contains no U. What base will be missing in the corresponding region of the template strand of DNA?

1.6 When the base composition of a DNA sample from the bacterium *Salinicoccus roseus* was determined, 23.6 percent of the bases were found to be guanine. The DNA of this organism is known to be double-stranded. What is the percentage of adenine in its DNA?

1.7 DNA extracted from a certain virus has the following base composition: 15 percent adenine, 25 percent thymine, 20 percent guanine, and 40 percent cytosine. How would you interpret this result in terms of the structure of the viral DNA?

1.8 A duplex DNA molecule contains 532 occurrences of the dinucleotide 5′-GT-3′ in one or the other of the paired strands. What other dinucleotide is also present exactly 532 times?

1.9 A repeating polymer with the sequence

5′-GAUGAUGAUGAU . . .-3′

was found to produce only two types of polypeptides in a translation system that uses cellular components but not living cells (called an *in vitro* translation system). One polypeptide consisted of repeating Asp and the other of repeating Met. How can you explain this result?

1.10 If one strand of a DNA duplex has the sequence 5′-GTCAT-3′, what is the sequence of the complementary strand. (Write the answer with the 5′ end at the left.)

1.11 Consider a region along one strand of a double-stranded DNA molecule consists of tandem repeats of the trinucleotide 5′-CTA-3′, so that the sequence in this strand is 5′-CTAC-TACTACTA . . .-3′. What is the sequence in the other strand? (Write the answer with the 5′ end at the left.)

1.12 Part of the protein-coding region in a gene has the base sequence 3′-ACAGCATAAACGTTC-5′. What is the sequence of the partner DNA strand?

1.13 If the DNA sequence in Problem 1.12 is the template strand that is transcribed in the synthesis of messenger RNA, would it be transcribed from left to right or from right to left? What base sequence would this region of the RNA contain?

1.14 What amino acid sequence would be synthesized from the messenger RNA region in Problem 1.12?

1.15 If a mutation occurs in the DNA sequence in Problem 1.12 in which the red C is replaced with T, what amino acid sequence would result?

1.16 A polymer is made that has a random sequence consisting of 25 percent U's and 75 percent C's. Among the amino acids in the polypeptide chains resulting from in vitro translation, what is the expected frequency of Pro? Of Phe?

1.17 With *in vitro* translation of an RNA into a polypeptide chain, the translation can begin anywhere along the RNA molecule. A synthetic RNA molecule has the sequence

5′-CGCUUACCACAUGUCGCGAAC-3′

How many reading frames are possible if this molecule is translated *in vitro*? How many reading frames are possible if this molecule is translated *in vivo*, in which translation starts with the codon AUG?

1.18 The coding sequence in the messenger RNA for amino acids 1 through 10 of human phenylalanine hydroxylase is

5′-AUGUCCACUGCGGUCCUGGAAAACCCAGGC-3′

(a) What are the first ten amino acids?
(b) What sequence would result from a mutant RNA in which the red A was changed to G?
(c) What sequence would result from a mutant RNA in which the red C was changed to G?

(d) What sequence would result from a mutant RNA in which the red U was changed to C?
(e) What sequence would result from a mutant RNA in which the red G was changed to U?

1.19 How is it possible for a gene with a mutation in the coding region to encode a polypeptide with the same amino acid sequence as the nonmutant gene?

1.20 Shown here is part of a metabolic pathway in a bacterium in which a substrate metabolite (small molecule) X is converted into a final product metabolite W through a sequence of three steps catalyzed by the enzymes A, B, and C. Each of the enzymes is the product of a different gene.

$$X \xrightarrow{A} Y \xrightarrow{B} Z \xrightarrow{C} W$$

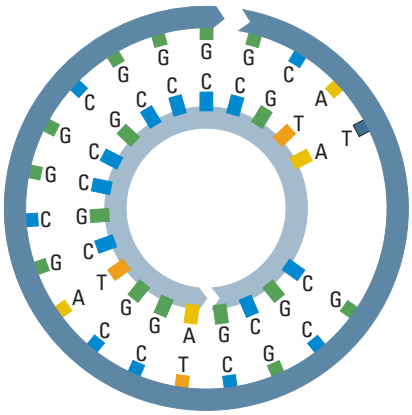
Which metabolites would be expected to be missing, and which present in excess, in cells that are mutant for:

(a) Enzyme A?
(b) Enzyme B?
(c) Enzyme C?

1.21 A mutant is isolated with a defect in one of the enzymes in the metabolic pathway in Problem 20, but it is not known which step (A, B, or C) is blocked. The final product W of the pathway is essential for growth. When mutant cells are placed in cultures lacking W, they cannot grow; but when W is added to the medium, they can grow. Experiments are carried out to determine whether any of the intermediates can substitute for W in supporting growth. The mutant cells are found to grow in the presence of Z but not in the presence of X or Y. Deduce from these data what step in the pathway is blocked in the mutant.

A MOMENT TO THINK 1

Answer to Problem 1: The single-stranded regions are complementary in sequence and polarity, so each end can loop around and form base pairs with the other end. The result is that the phage DNA can form a circle, as illustrated in the diagram. (The term cohesive ends comes from the fact that the ends can stick together, or cohere.)



A MOMENT TO THINK 2

Answer to Problem 2: The RNA and the reverse-transcribed DNA match as follows:

```
5′-UCCUAUUGAAACUGUACCAGUAAAAUU-3′
3′-AGGATAACTTTGACATGGTCATTTTAA-5′
```

Because the DNA strand grows by the addition of nucleotides to the 3′ end, this strand would grow from right to left. Reverse transcriptase also uses the first DNA strand as a template to produce a second DNA strand; the result is the production of a double-stranded DNA molecule from a single strand of RNA.)

GENETICS on the web

GeNETics on the Web will introduce you to some of the most important sites for finding genetics information on the Internet. To explore these sites, visit the Jones & Bartlett companion site to accompany *Essential Genetics: A Genomic Perspective, Sixth Edition*, at http://biology.jbpub.com/Hartl/EssentialGenetics.

There you will find a chapter-by-chapter list of highlighted keywords. When you select one of the keywords, you will be linked to a Web site containing information related to that keyword.