17 Expression of Genes

Figure 17.1 How does a single faulty gene result in the dramatic appearance of these albino donkeys?

KEY CONCEPTS

An albino raccoon.

The Flow of Genetic Information

The island of Asinara lies off the coast of the Italian island of Sardinia. The name Asinara probably originated from the Latin work *sinuaria*, which means "sinus-shaped." A second meaning of Asinara is "donkey-inhabited," which is perhaps even more appropriate because Asinara is home to a wild population of albino donkeys **(Figure 17.1)**. What factors are responsible for the albino phenotype?

Inherited traits are determined by genes, and the trait of albinism is caused by a recessive allele of a pigmentation gene (see Concept 14.4). The information content of genes is in the form of specific sequences of nucleotides along strands of DNA, the genetic material. The albino donkey has a faulty version of a key protein, an enzyme required for pigment synthesis, and this protein is faulty because the gene that codes for it contains incorrect information.

This example illustrates the main point of this chapter: The DNA inherited by an organism leads to specific traits by dictating the synthesis of proteins and of RNA molecules involved in protein synthesis. In other words, proteins are the link between genotype and phenotype. **Gene expression** is the process by which DNA directs the synthesis of proteins (or, in some cases, just RNAs). The expression of genes that code for proteins includes two stages: transcription and translation. This chapter describes the flow of information from gene to protein and explains how genetic mutations affect organisms through their proteins. Understanding the processes of gene expression will allow us to revisit the concept of the gene in more detail at the end of the chapter.

When you see this blue icon, log in to **MasteringBiology** and go to the Study Area for digital resources.

Get Ready for This Chapter

CONCEPT 17.1

Genes specify proteins via transcription and translation

Before going into the details of how genes direct protein synthesis, let's step back and examine how the fundamental relationship between genes and proteins was discovered.

Evidence from Studying Metabolic Defects

In 1902, British physician Archibald Garrod was the first to suggest that genes dictate phenotypes through enzymes, proteins that catalyze specific chemical reactions in the cell. He postulated that the symptoms of an inherited disease reflect an inability to make a particular enzyme. He later referred to such diseases as "inborn errors of metabolism." For example, people with a disease called alkaptonuria have black urine because it contains a chemical called alkapton, which darkens upon exposure to air. Garrod reasoned that most people have an enzyme that breaks down alkapton, whereas people with alkaptonuria have inherited an inability to make that enzyme, so alkapton is expelled in their urine.

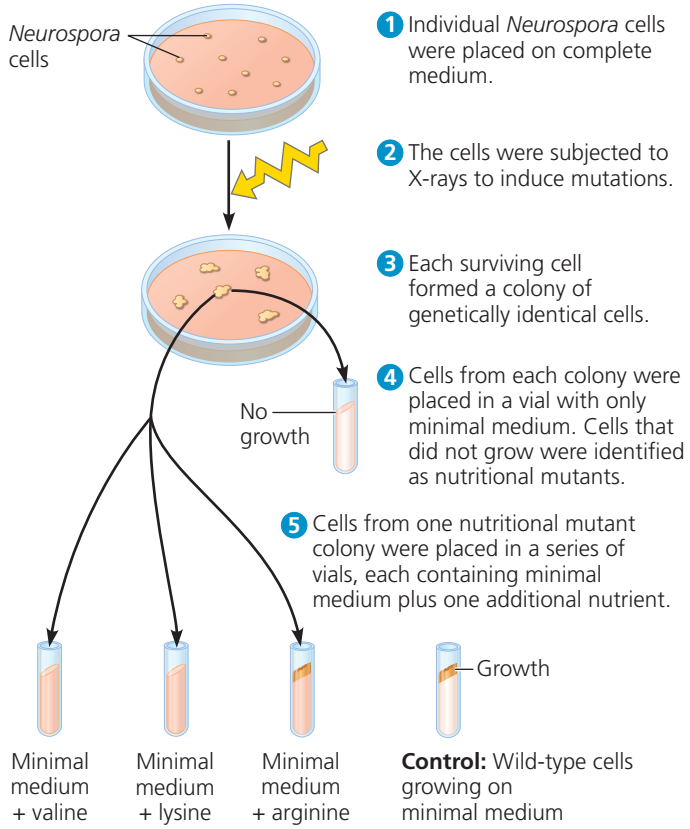
Several decades later, research supported Garrod's hypothesis that a gene dictates the production of a specific enzyme, later named the *one gene–one enzyme hypothesis*. Biochemists learned that cells synthesize and degrade most organic molecules via metabolic pathways, in which each chemical reaction in a sequence is catalyzed by a specific enzyme (see Concept 6.1). Such metabolic pathways lead, for instance, to the synthesis of the pigments that give the brown donkey in Figure 17.1 its fur color or fruit flies (*Drosophila*) their eye color (see Figure 15.3). In the 1930s, the American biochemist and geneticist George Beadle and his French colleague Boris Ephrussi speculated that in *Drosophila*, each mutation affecting eye color blocks pigment synthesis at a specific step by preventing production of the enzyme that catalyzes that step. But neither the chemical reactions nor the enzymes that catalyze them were known at the time.

Nutritional Mutants in Neurospora: Scientific Inquiry

A breakthrough came a few years later at Stanford University, where Beadle and Edward Tatum began working with a bread mold, *Neurospora crassa*, a haploid species. To observe a change in a mutant's phenotype, Beadle and Tatum needed to disable just one allele (rather than two, as in a diploid species) of a protein-coding gene required for a specific metabolic activity. They bombarded *Neurospora* with X-rays, known to cause mutations, and looked among the survivors for mutants that differed in their nutritional needs from the wild-type bread mold.

Wild-type *Neurospora* has modest food requirements. It can grow in the laboratory on a simple solution containing

Figure 17.2 Beadle and Tatum's experimental approach. To obtain nutritional mutants, Beadle and Tatum exposed *Neurospora* cells to X-rays, inducing mutations, then screened mutants that had new nutritional requirements, such as arginine, as shown here.



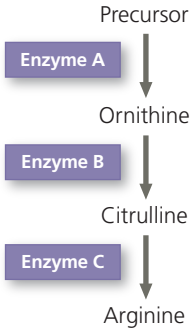
minimal nutrients for growth—inorganic salts, glucose, and the vitamin biotin—incorporated into agar, a support medium. From this so-called *minimal medium*, wild-type mold cells use their metabolic pathways to produce all the other molecules they need for growth, dividing repeatedly and forming visible colonies of genetically identical cells. As shown in **Figure 17.2**, Beadle and Tatum generated different "nutritional mutants" of *Neurospora* cells, each of which was unable to synthesize a particular essential nutrient. Such cells could not grow on minimal medium but could grow on *complete medium*, which contains all nutrients needed for growth. For *Neurospora*, the complete medium consists of the minimal medium supplemented with all 20 amino acids and a few other nutrients. Beadle and Tatum hypothesized that in each nutritional mutant, the gene for the enzyme that synthesizes a particular nutrient had been disabled.

This approach resulted in a valuable collection of mutant strains of *Neurospora*, catalogued by their defect in a particular pathway. Two colleagues of theirs, Adrian Srb and Norman Horowitz, used a collection of arginine-requiring mutants to investigate the biochemical pathway for arginine synthesis in *Neurospora* **(Figure 17.3)**. Srb and Horowitz pinned down each

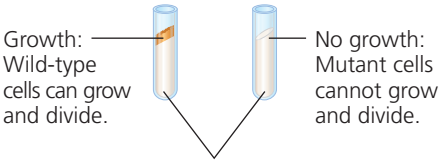
Figure 17.3

Inquiry Do individual genes specify the enzymes that function in a biochemical pathway?

Experiment Working with the mold *Neurospora crassa*, Adrian Srb and Norman Horowitz, then at Stanford University, used Beadle and Tatum's experimental approach (see Figure 17.2) to isolate mutants that required arginine in their growth medium. The researchers showed that these mutants fell into three classes, each defective in a different gene. From studies by others on mammalian liver cells, they suspected that the metabolic pathway of arginine biosynthesis involved a precursor nutrient and the intermediate molecules ornithine and citrulline, as shown in the diagram on the right.



Their most famous experiment, shown here, tested both the *one gene–one enzyme hypothesis* and their postulated arginine-synthesizing pathway. In this experiment, they grew their three classes of mutants under the four different conditions shown in the Results Table below. They included minimal medium (MM) as a control, knowing that wild-type cells could grow on MM but mutant cells could not. (See test tubes below.)



Control: Minimal medium

Results As shown in the table on the right, the wild-type strain was capable of growth under all experimental conditions, requiring only the minimal medium. The three classes of mutants each had a specific set of growth requirements. For example, class II mutants could not grow when ornithine alone was added but could grow when either citrulline or arginine was added.

Results Table — Classes of *Neurospora crassa*

Condition	Wild type	Class I mutants	Class II mutants	Class III mutants
Minimal medium (MM) (control)				
MM + ornithine				
MM + citrulline				
MM + arginine (control)				
Summary of results	Can grow with or without any supplements	Can grow on ornithine, citrulline, or arginine	Can grow only on citrulline or arginine	Require arginine to grow

Conclusion From the growth requirements of the mutants, Srb and Horowitz deduced that each class of mutant was unable to carry out one step in the pathway for synthesizing arginine, presumably because it lacked the necessary enzyme, as shown in the table on the right. Because each of their mutants was mutated in a single gene, they concluded that each mutated gene must normally dictate the production of one enzyme. Their results supported the one gene–one enzyme hypothesis, proposed by Beadle and Tatum, and also confirmed that the arginine pathway described in the mammalian liver also operates in *Neurospora*. (Notice in the Results Table that a mutant can grow only if supplied with a compound made *after* the defective step because this bypasses the defect.)

Gene (codes for enzyme)	**Wild type**	**Class I mutants** (mutation in gene *A*)	**Class II mutants** (mutation in gene *B*)	**Class III mutants** (mutation in gene *C*)
Gene *A* → Enzyme A Gene *B* → Enzyme B Gene *C* → Enzyme C	Precursor → Enzyme A → Ornithine → Enzyme B → Citrulline → Enzyme C → Arginine	Precursor → ~~Enzyme A~~ → Ornithine → Enzyme B → Citrulline → Enzyme C → Arginine	Precursor → Enzyme A → Ornithine → ~~Enzyme B~~ → Citrulline → Enzyme C → Arginine	Precursor → Enzyme A → Ornithine → Enzyme B → Citrulline → ~~Enzyme C~~ → Arginine

WHAT IF? ➤ *Suppose the experiment had shown that class I mutants could grow only in MM supplemented by ornithine or arginine and that class II mutants could grow in MM supplemented by citrulline, ornithine, or arginine. What conclusions would the researchers have drawn from those results regarding the biochemical pathway and the defect in class I and class II mutants?*

Data from A. M. Srb and N. H. Horowitz, The ornithine cycle in *Neurospora* and its genetic control, *Journal of Biological Chemistry* 154:129–139 (1944).

mutant's defect more specifically, using additional tests to distinguish among three classes of arginine-requiring mutants. Mutants in each class required a different set of compounds along the arginine-synthesizing pathway, which has three steps. These results, and those of many similar experiments done by Beadle and Tatum, suggested that each class was blocked at a different step in this pathway because mutants in that class lacked the enzyme that catalyzes the blocked step.

Because Beadle and Tatum set up their experimental conditions so that each mutant was defective in a single gene, the collected results, taken together, provided strong support for a working hypothesis they had proposed earlier. The *one gene–one enzyme hypothesis*, as they dubbed it, states that the function of a gene is to dictate the production of a specific enzyme. Further support for this hypothesis came from experiments that identified the specific enzymes lacking in the mutants. Beadle and Tatum shared a Nobel Prize in 1958 for "their discovery that genes act by regulating definite chemical events" (in the words of the Nobel committee).

Today, we know of countless examples in which a mutation in a gene causes a faulty enzyme that in turn leads to an identifiable condition. The albino donkey in Figure 17.1 lacks a key enzyme called tyrosinase in the metabolic pathway that produces melanin, a dark pigment. The absence of melanin causes white fur and other effects throughout the donkey's body. Its nose, ears, and hooves, as well as its eyes, are pink because no melanin is present to mask the reddish color of the blood vessels that run through those structures.

The Products of Gene Expression: A Developing Story

As researchers learned more about proteins, they made revisions to the one gene–one enzyme hypothesis. First of all, not all proteins are enzymes. Keratin, the structural protein of animal hair, and the hormone insulin are two examples of nonenzyme proteins. Because proteins that are not enzymes are nevertheless gene products, molecular biologists began to think in terms of one gene–one protein. However, many proteins are constructed from two or more different polypeptide chains, and each polypeptide is specified by its own gene. For example, hemoglobin—the oxygen-transporting protein of vertebrate red blood cells—contains two kinds of polypeptides (see Figure 5.18), and thus two genes code for this protein, one for each type of polypeptide. Beadle and Tatum's idea was therefore restated as the *one gene–one polypeptide hypothesis*. Even this description is not entirely accurate, though. First, in many cases, a eukaryotic gene can code for a set of closely related polypeptides via a process called alternative splicing, which you will learn about later in this chapter. Second, quite a few genes code for RNA molecules that have important functions in cells even though they are never translated into protein. For now, we will focus on genes that do code for polypeptides. (Note that it is common to refer to these gene products as proteins—a practice you will encounter in this book—rather than more precisely as polypeptides.)

Basic Principles of Transcription and Translation

Genes provide the instructions for making specific proteins. But a gene does not build a protein directly. The bridge between DNA and protein synthesis is the nucleic acid RNA. RNA is chemically similar to DNA except that it contains ribose instead of deoxyribose as its sugar and has the nitrogenous base uracil rather than thymine (see Figure 5.23). Thus, each nucleotide along a DNA strand has A, G, C, or T as its base, and each nucleotide along an RNA strand has A, G, C, or U as its base. An RNA molecule usually consists of a single strand.

It is customary to describe the flow of information from gene to protein in linguistic terms. Just as specific sequences of letters communicate information in a language such as English, both nucleic acids and proteins are polymers with specific sequences of monomers that convey information. In DNA or RNA, the monomers are the four types of nucleotides, which differ in their nitrogenous bases. Genes are typically hundreds or thousands of nucleotides long, each gene having a specific sequence of nucleotides. Each polypeptide of a protein also has monomers arranged in a particular linear order (the protein's primary structure; see Figure 5.18), but its monomers are amino acids. Thus, nucleic acids and proteins contain information written in two different chemical languages. Getting from DNA to protein requires two major stages: transcription and translation.

Transcription is the synthesis of RNA using information in the DNA. The two nucleic acids are written in different forms of the same language, and the information is simply transcribed, or "rewritten," from DNA to RNA. Just as a DNA strand provides a template for making a new complementary strand during DNA replication (see Concept 16.2), it also can serve as a template for assembling a complementary sequence of RNA nucleotides. For a protein-coding gene, the resulting RNA molecule is a faithful transcript of the gene's protein-building instructions. This type of RNA molecule is called **messenger RNA (mRNA)** because it carries a genetic message from the DNA to the protein-synthesizing machinery of the cell. (Transcription is the general term for the synthesis of *any* kind of RNA on a DNA template. Later, you will learn about some other types of RNA produced by transcription.)

Translation is the synthesis of a polypeptide using the information in the mRNA. During this stage, there is a change in language: The cell must translate the nucleotide sequence of an mRNA molecule into the amino acid sequence of

a polypeptide. The sites of translation are **ribosomes**, molecular complexes that facilitate the orderly linking of amino acids into polypeptide chains.

Transcription and translation occur in all organisms. Because most studies have involved bacteria and eukaryotic cells, they are our main focus in this chapter. Our understanding of transcription and translation in archaea lags behind, but we do know that archaeal cells share some features of gene expression with bacteria and others with eukaryotes.

The basic mechanics of transcription and translation are similar for bacteria and eukaryotes, but there is an important difference in the flow of genetic information within the cells. Bacteria do not have nuclei. Therefore, nuclear membranes do not separate bacterial DNA and mRNA from ribosomes and the other protein-synthesizing equipment **(Figure 17.4a)**. As you'll see later, this lack of compartmentalization allows translation of an mRNA to begin while its transcription is still in progress. By contrast, eukaryotic cells have nuclei. The presence of a nuclear envelope separates transcription from translation in space and time **(Figure 17.4b)**. Transcription occurs in the nucleus, but the mRNA must be transported to the cytoplasm for translation. Before eukaryotic RNA transcripts from protein-coding genes can leave the nucleus, they are modified in various ways to produce the final, functional mRNA. The transcription of a protein-coding eukaryotic gene results in *pre-mRNA*, and further processing yields the finished mRNA. The initial RNA transcript from any gene, including those specifying RNA that is not translated into protein, is more generally called a **primary transcript**.

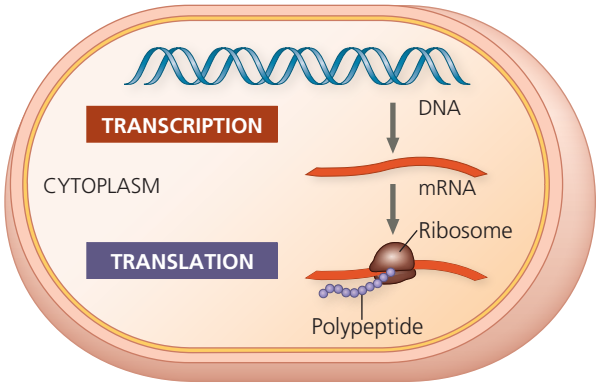
To summarize: Genes program protein synthesis via genetic messages in the form of messenger RNA. Put another way, cells are governed by a molecular chain of command with a directional flow of genetic information:

This concept was dubbed the *central dogma* by Francis Crick in 1956. But in the 1970s, scientists were surprised to discover some enzymes that use RNA molecules as templates for DNA synthesis (which we'll cover in Concept 26.2). However, these exceptions do not invalidate the idea that, in general, genetic information flows from DNA to RNA to protein. In the next section, we discuss how the instructions for assembling amino acids into a specific order are encoded in nucleic acids.

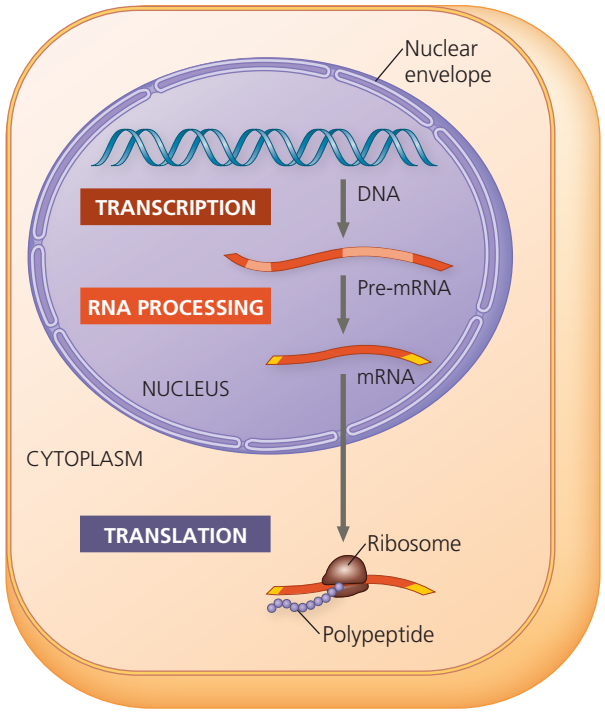
The Genetic Code

When biologists began to suspect that the instructions for protein synthesis were encoded in DNA, they recognized a problem: There are only four nucleotide bases to specify

Figure 17.4 Overview: the roles of transcription and translation in the flow of genetic information. In a cell, inherited information flows from DNA to RNA to protein. The two main stages of information flow are transcription and translation. A miniature version of part **(a)** or **(b)** accompanies several figures later in the chapter as an orientation diagram to help you see where a particular figure fits into the overall scheme of gene expression.



(a) Bacterial cell. In a bacterial cell, which lacks a nucleus, mRNA produced by transcription is immediately translated without additional processing.



(b) Eukaryotic cell. The nucleus provides a separate compartment for transcription. The original RNA transcript, called pre-mRNA, is processed in various ways before leaving the nucleus as mRNA.

MB **Animation: Overview of Protein Synthesis in Bacteria**
Animation: Overview of Protein Synthesis in Eukaryotes

20 amino acids. Thus, the genetic code cannot be a language like Chinese, where each written symbol corresponds to a word. How many nucleotides, then, would turn out to correspond to an amino acid?

Codons: Triplets of Nucleotides

If each kind of nucleotide base were translated into an amino acid, only four amino acids could be specified, one per nucleotide base. Would a language of two-letter code words suffice? The two-nucleotide sequence AG, for example, could specify one amino acid, and GT could specify another. Since there are four possible nucleotide bases in each position, this would give us 16 (that is, 4×4, or 4^2) possible arrangements—still not enough to code for all 20 amino acids.

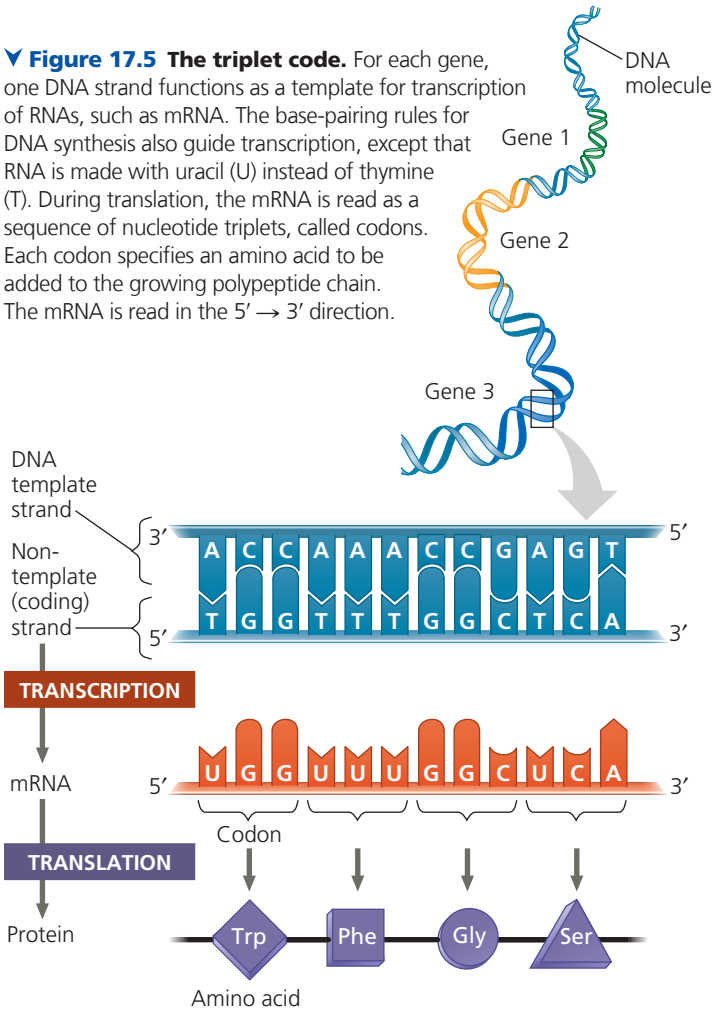
Triplets of nucleotide bases are the smallest units of uniform length that can code for all the amino acids. If each arrangement of three consecutive nucleotide bases specifies an amino acid, there can be 64 (that is, 4^3) possible code words—more than enough to specify all the amino acids. Experiments have verified that the flow of information from gene to protein is based on a **triplet code**: The genetic instructions for a polypeptide chain are written in the DNA as a series of nonoverlapping, three-nucleotide words. The series of words in a gene is transcribed into a complementary series of nonoverlapping, three-nucleotide words in mRNA, which is then translated into a chain of amino acids **(Figure 17.5)**.

During transcription, the gene determines the sequence of nucleotide bases along the length of the RNA molecule that is being synthesized. For each gene, only one of the two DNA strands is transcribed. This strand is called the **template strand** because it provides the pattern, or template, for the sequence of nucleotides in an RNA transcript. For any given gene, the same strand is used as the template every time the gene is transcribed. However, farther along on the same chromosomal DNA molecule, the opposite strand may function as the template for a different gene. The strand that is used as the template is determined by the orientation of the enzyme that transcribes genes, which in turn depends on the particular DNA sequences associated with that gene.

An mRNA molecule is complementary rather than identical to its DNA template because RNA nucleotides are assembled on the template according to base-pairing rules (see Figure 17.5). The pairs are similar to those that form during DNA replication, except that U (the RNA substitute for T) pairs with A and the mRNA nucleotides contain ribose instead of deoxyribose. Like a new strand of DNA, the RNA molecule is synthesized in an antiparallel direction to the template strand of DNA. (To review what is meant by "antiparallel" and the 5′ and 3′ ends of a nucleic acid chain, see Figure 16.7.) In the example in Figure 17.5, the nucleotide triplet ACC along the DNA template strand (written as 3′-ACC-5′) provides a template for 5′-UGG-3′ in the mRNA molecule. The mRNA nucleotide triplets are called **codons**, and they are customarily written in the 5′ → 3′ direction. In our example, UGG is the codon for the amino acid tryptophan (abbreviated Trp, or W). The term *codon* is also used for the DNA nucleotide triplets along the *nontemplate* strand. These codons are complementary to the template strand and thus identical in sequence to the mRNA, except that they have a T wherever there is a U in the mRNA. For this reason, the nontemplate DNA strand is often called the **coding strand**; by convention, the sequence of the coding strand is used when a gene's sequence is reported.

During translation, the sequence of codons along an mRNA molecule is decoded, or translated, into a sequence of amino acids making up a polypeptide chain. The codons are read by the translation machinery in the 5′ → 3′ direction along the mRNA. Each codon specifies which one of the 20 amino acids will be incorporated at the corresponding position along a polypeptide. Because codons are nucleotide triplets, the number of nucleotides making up a genetic message must be three times the number of amino acids in the protein product. For example, it takes 300 nucleotides along an mRNA strand to code for the amino acids in a polypeptide that is 100 amino acids long.

Figure 17.5 The triplet code. For each gene, one DNA strand functions as a template for transcription of RNAs, such as mRNA. The base-pairing rules for DNA synthesis also guide transcription, except that RNA is made with uracil (U) instead of thymine (T). During translation, the mRNA is read as a sequence of nucleotide triplets, called codons. Each codon specifies an amino acid to be added to the growing polypeptide chain. The mRNA is read in the 5′ → 3′ direction.



VISUAL SKILLS ➤ *By convention, the nontemplate strand, also called the coding strand, is used to represent a DNA sequence. Write the sequence of the mRNA strand and the nontemplate strand—in both cases reading from 5′ to 3′—and compare them. Why do you think this convention was adopted? (Hint: Why is this called the coding strand?)*

MP3 Tutor: DNA to RNA to Protein

Cracking the Code

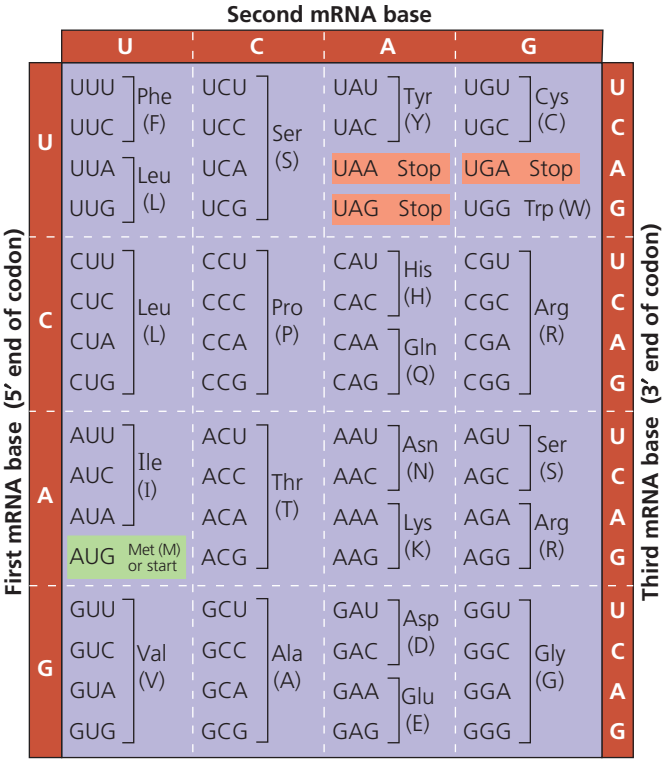
Molecular biologists cracked the genetic code of life in the early 1960s when a series of elegant experiments disclosed the amino acid translations of each of the RNA codons. The first codon was deciphered in 1961 by Marshall Nirenberg, of the National Institutes of Health, along with his colleagues. Nirenberg synthesized an artificial mRNA by linking together many identical RNA nucleotides containing uracil as their base. No matter where the genetic message started or stopped, it could contain only one codon (UUU) over and over. Nirenberg added this "poly-U" polynucleotide to a test-tube mixture containing amino acids, ribosomes, and the other components required for protein synthesis. His artificial system translated the poly-U mRNA into a polypeptide containing many units of the amino acid phenylalanine (Phe, or F), strung together as a long polyphenylalanine chain. Thus, Nirenberg determined that the mRNA codon UUU specifies the amino acid phenylalanine. Soon, the amino acids specified by the codons AAA, GGG, and CCC were also identified.

Although more elaborate techniques were required to decode mixed triplets such as AUA and CGA, all 64 codons were deciphered by the mid-1960s. As **Figure 17.6** shows, 61 of the 64 triplets code for amino acids. The three codons that do not designate amino acids are "stop" signals, or termination codons, marking the end of translation. Notice that the codon AUG has a dual function: It codes for the amino acid methionine (Met, or M) and also functions as a "start" signal, or initiation codon. Genetic messages usually begin with the mRNA codon AUG, which signals the protein-synthesizing machinery to begin translating the mRNA at that location. (Because AUG also stands for methionine, polypeptide chains begin with methionine when they are synthesized. However, an enzyme may subsequently remove this starter amino acid from the chain.)

Notice in Figure 17.6 that there is redundancy in the genetic code, but no ambiguity. For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them ever specifies any other amino acid (no ambiguity). The redundancy in the code is not altogether random. In many cases, codons that are synonyms for a particular amino acid differ only in the third nucleotide base of the triplet. We will consider the significance of this redundancy later in the chapter.

Our ability to extract the intended message from a written language depends on reading the symbols in the correct groupings—that is, in the correct **reading frame**. Consider this statement: "The red dog ate the bug." Group the letters incorrectly by starting at the wrong point, and the result will probably be gibberish: for example, "her edd oga tet heb ug." The reading frame is also important in the molecular language of cells. The short stretch of polypeptide shown in Figure 17.5, for instance, will be made

Figure 17.6 The codon table for mRNA. The three nucleotide bases of an mRNA codon are designated here as the first, second, and third bases, reading in the 5′ → 3′ direction along the mRNA. The codon AUG not only stands for the amino acid methionine (Met, or M) but also functions as a "start" signal for ribosomes to begin translating the mRNA at that point. Three of the 64 codons function as "stop" signals, marking where ribosomes end translation. Both one- and three-letter codes are shown for the amino acids; see Figure 5.14 for their full names.



VISUAL SKILLS ➤ *A segment in the middle of an mRNA has the sequence 5′-AGAGAACCGCGA-3′. Using the codon table, translate this sequence, assuming the first three nucleotides are a codon.*

Animation: Translation: The Genetic Code

correctly only if the mRNA nucleotides are read from left to right (5′ → 3′) in the groups of three shown in the figure: UGG UUU GGC UCA. Although a genetic message is written with no spaces between the codons, the cell's protein-synthesizing machinery reads the message as a series of nonoverlapping three-letter words. The message is *not* read as a series of overlapping words—UGGUUU, and so on—which would convey a very different message.

Evolution of the Genetic Code

EVOLUTION The genetic code is nearly universal, shared by organisms from the simplest bacteria to the most complex plants and animals. The mRNA codon CCG, for instance, is translated as the amino acid proline in all organisms whose genetic code has been examined. In laboratory experiments, genes can be transcribed and translated after being transplanted from one species to another, sometimes with quite

Figure 17.7 Evidence for evolution: expression of genes from different species. Because diverse forms of life share a common genetic code due to their shared ancestry, one species can be programmed to produce proteins characteristic of a second species by introducing DNA from the second species into the first.

(a) Tobacco plant expressing a firefly gene. The yellow glow is produced by a chemical reaction catalyzed by the protein product of the firefly gene.

(b) Pig expressing a jellyfish gene. Researchers injected a jellyfish gene for a fluorescent protein into fertilized pig eggs. One developed into this fluorescent pig.

MB **Video: GFP Transgenic Mice**

striking results, as shown in **Figure 17.7**. Bacteria can be programmed by the insertion of human genes to synthesize certain human proteins for medical use, such as insulin. Such applications have produced many exciting developments in the area of biotechnology (see Concept 19.4).

Despite a small number of exceptions, the evolutionary significance of the code's near universality is clear. A language shared by all living things must have been operating very early in the history of life—early enough to be present in the common ancestor of all present-day organisms. A shared genetic vocabulary is a reminder of the kinship of all life.

CONCEPT CHECK 17.1

1. **MAKE CONNECTIONS ➤** In a research article about alkaptonuria published in 1902, Garrod suggested that humans inherit two "characters" (alleles) for a particular enzyme and that both parents must contribute a faulty version for the offspring to have alkaptonuria. Today, would this disorder be called dominant or recessive? (See Concept 14.4.)
2. Since AUG is the start codon and codes for methionine, do all proteins have methionine as the first amino acid? Explain.
3. **DRAW IT ➤** T The template strand of a gene contains the sequence 3′-TTCAGTCGT-5′. Imagine that the nontemplate sequence was transcribed instead of the template sequence. Draw the mRNA sequence and translate it using Figure 17.6. (Be sure to pay attention to the 5′ and 3′ ends.) Predict how well the protein synthesized from the nontemplate strand would function, if at all.

For suggested answers, see Appendix A.

CONCEPT 17.2

Transcription is the DNA-directed synthesis of RNA: *a closer look*

Now that we have considered the linguistic logic and evolutionary significance of the genetic code, we are ready to reexamine transcription, the first stage of gene expression, in greater detail.

Molecular Components of Transcription

Messenger RNA, the carrier of information from DNA to the cell's protein-synthesizing machinery, is transcribed from the template strand of a gene. An enzyme called an **RNA polymerase** pries the two strands of DNA apart and joins together RNA nucleotides complementary to the DNA template strand, thus elongating the RNA polynucleotide **(Figure 17.8)**. Like the DNA polymerases that function in DNA replication, RNA polymerases can assemble a polynucleotide only in its 5′ → 3′ direction, adding onto its 3′ end. Unlike DNA polymerases, however, RNA polymerases are able to start a chain from scratch; they don't need to add the first nucleotide onto a pre-existing primer.

Specific sequences of nucleotides along the DNA mark where transcription of a gene begins and ends. The DNA sequence where RNA polymerase attaches and initiates transcription is known as the **promoter**; in bacteria, the sequence that signals the end of transcription is called the **terminator**. (The termination mechanism is different in eukaryotes; we'll describe it later.) Molecular biologists refer to the direction of transcription as "downstream" and the other direction as "upstream." These terms are also used to describe the positions of nucleotide sequences within the DNA or RNA. Thus, the promoter sequence in DNA is said to be upstream from the terminator. The stretch of DNA downstream from the promoter that is transcribed into an RNA molecule is called a **transcription unit**.

Bacteria have a single type of RNA polymerase that synthesizes not only mRNA but also other types of RNA that function in protein synthesis, such as ribosomal RNA. In contrast, eukaryotes have at least three types of RNA polymerase in their nuclei; the one used for pre-mRNA synthesis is called RNA polymerase II. The other RNA polymerases transcribe RNA molecules that are not translated into protein. In the discussion that follows, we start with the features of mRNA synthesis common to both bacteria and eukaryotes and then describe some key differences.

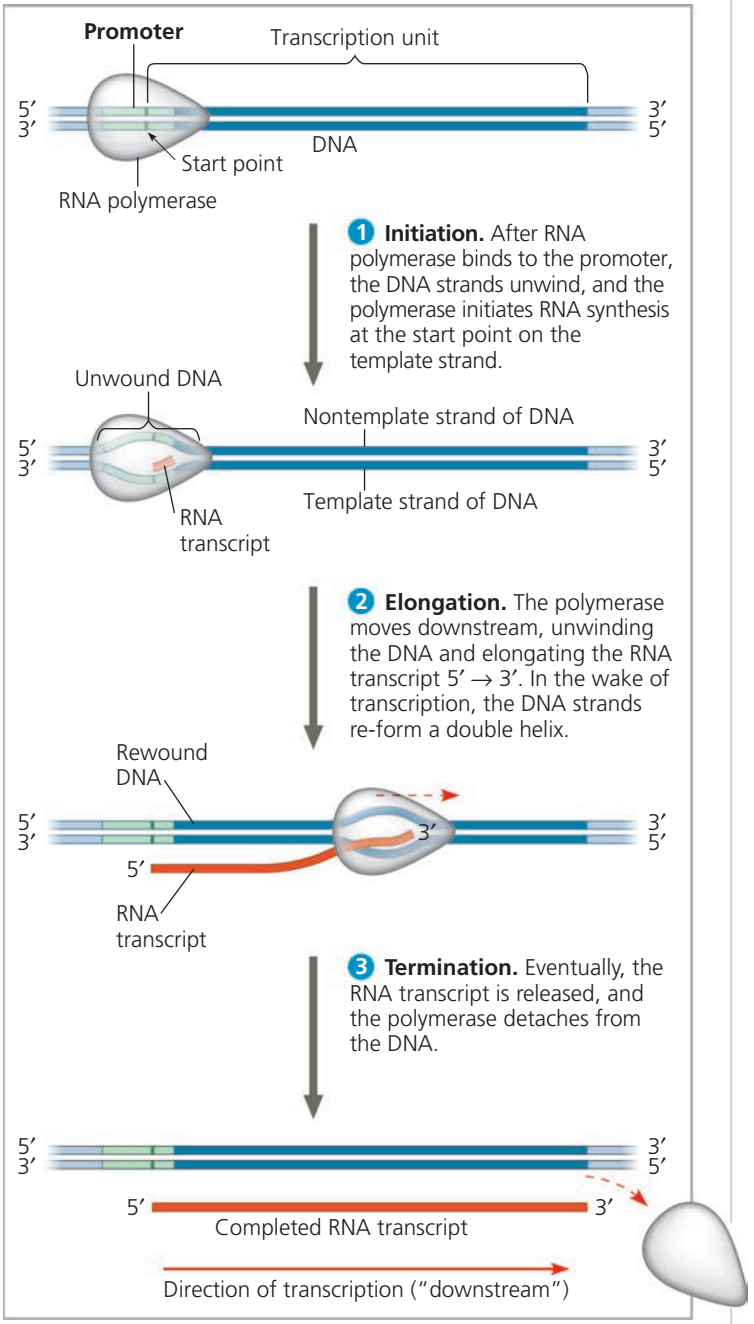
Synthesis of an RNA Transcript

The three stages of transcription, as shown in Figure 17.8 and described next, are initiation, elongation, and termination of the RNA chain. Study Figure 17.8 to familiarize yourself with the stages and the terms used to describe them.

RNA Polymerase Binding and Initiation of Transcription

The promoter of a gene includes within it the transcription **start point**—the nucleotide where RNA polymerase actually begins synthesis of the mRNA—and typically extends several dozen or so nucleotide pairs upstream from the start point **(Figure 17.9)**. Based on interactions with proteins that will be covered shortly, RNA polymerase binds in a precise location and orientation on the promoter. This in turn determines where transcription starts and which of the two strands of the DNA helix is used as the template.

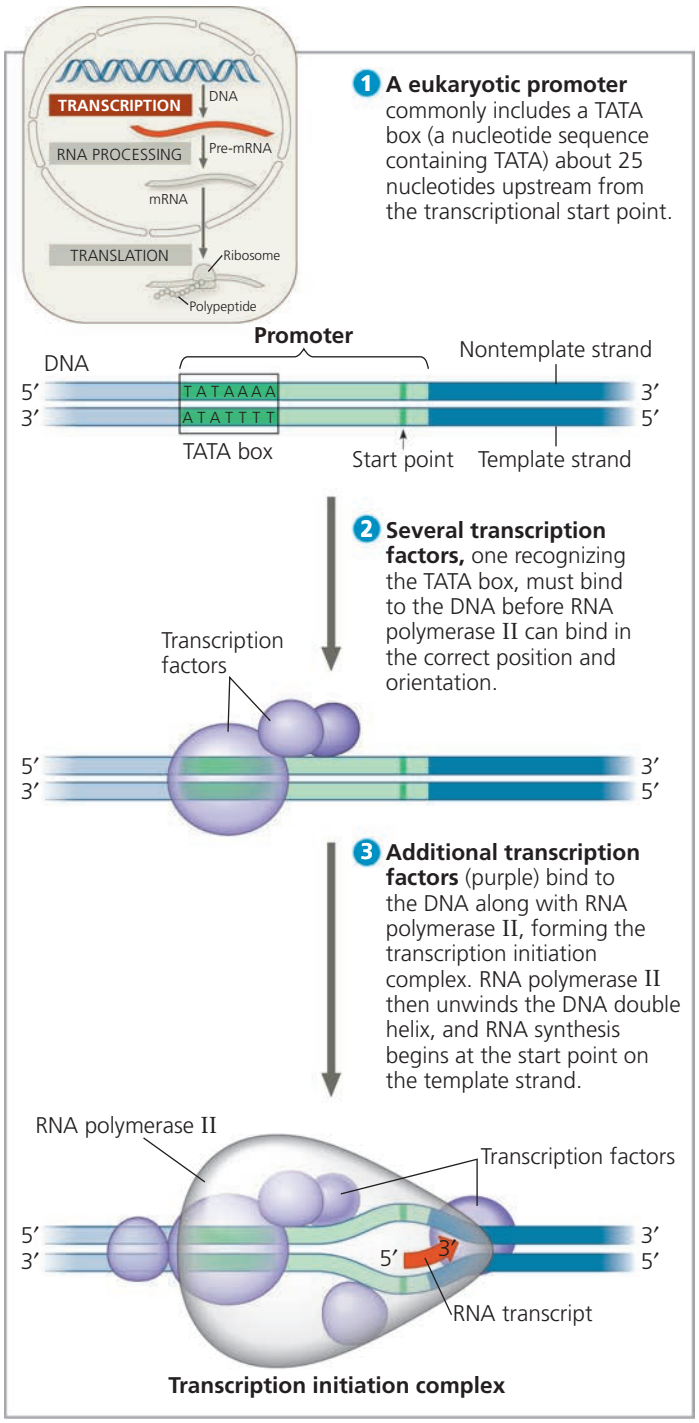
Figure 17.8 The stages of transcription: initiation, elongation, and termination. This general depiction of transcription applies to both bacteria and eukaryotes, but the details of termination differ, as described in the text. Also, in a bacterium, the RNA transcript is immediately usable as mRNA; in a eukaryote, the RNA transcript must first undergo processing.



MAKE CONNECTIONS ➤ *Compare the use of a template strand during transcription and replication. See Figure 16.17.*

Animation: Overview of Transcription
Animation: Overview of Transcription in Bacteria

Figure 17.9 The initiation of transcription at a eukaryotic promoter. In eukaryotic cells, proteins called transcription factors mediate the initiation of transcription by RNA polymerase II.



? *Explain how the interaction of RNA polymerase with the promoter would differ if the figure showed transcription initiation for bacteria.*

Certain sections of a promoter are especially important for binding RNA polymerase in a way that ensures that transcription will begin at the right place. In bacteria, part of the RNA polymerase itself specifically recognizes and binds to the promoter. In eukaryotes, a collection of proteins called **transcription factors** mediate the binding of RNA polymerase and the initiation of transcription. Only after transcription factors are attached to the promoter does RNA polymerase II bind to it. The whole complex of transcription factors and RNA polymerase II bound to the promoter is called a **transcription initiation complex**. Figure 17.9 shows the role of transcription factors and a crucial promoter DNA sequence called the **TATA box** in forming the initiation complex at a eukaryotic promoter.

The interaction between eukaryotic RNA polymerase II and transcription factors is an example of the importance of protein-protein interactions in controlling eukaryotic transcription. Once the appropriate transcription factors are firmly attached to the promoter DNA and the polymerase is bound to them in the correct orientation on the DNA, the enzyme unwinds the two DNA strands and begins transcribing the template strand at the start point.

Elongation of the RNA Strand

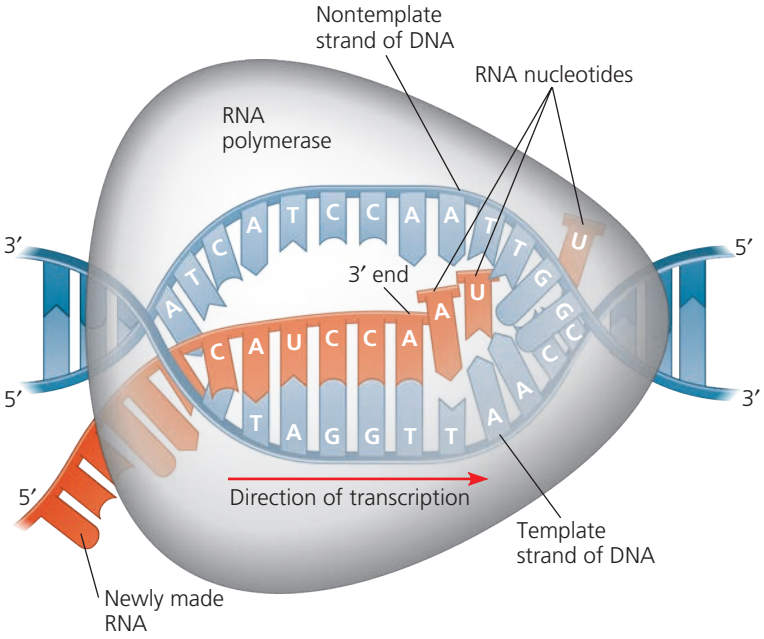
As RNA polymerase moves along the DNA, it untwists the double helix, exposing about 10–20 DNA nucleotides at a time for pairing with RNA nucleotides **(Figure 17.10)**. The enzyme adds nucleotides to the 3′ end of the growing RNA molecule as it continues along the double helix. In the wake of this advancing wave of RNA synthesis, the new RNA molecule peels away from its DNA template, and the DNA double helix re-forms. Transcription progresses at a rate of about 40 nucleotides per second in eukaryotes.

A single gene can be transcribed simultaneously by several molecules of RNA polymerase following each other like trucks in a convoy. A growing strand of RNA trails off from each polymerase, with the length of each new strand reflecting how far along the template the enzyme has traveled from the start point (see the mRNA molecules in Figure 17.23). The congregation of many polymerase molecules simultaneously transcribing a single gene increases the amount of mRNA transcribed from it, which helps the cell make the encoded protein in large amounts.

Termination of Transcription

Bacteria and eukaryotes differ in the way they terminate transcription. In bacteria, transcription proceeds through a terminator sequence in the DNA. The transcribed terminator (an RNA sequence) functions as the termination signal, causing the polymerase to detach from the DNA and release the transcript, which requires no further modification

Figure 17.10 Transcription elongation. RNA polymerase moves along the DNA template strand, joining complementary RNA nucleotides to the 3′ end of the growing RNA transcript. Behind the polymerase, the new RNA peels away from the template strand, which re-forms a double helix with the nontemplate strand.



BioFlix® Animation: Transcription
Animation: Elongation of the RNA Strand

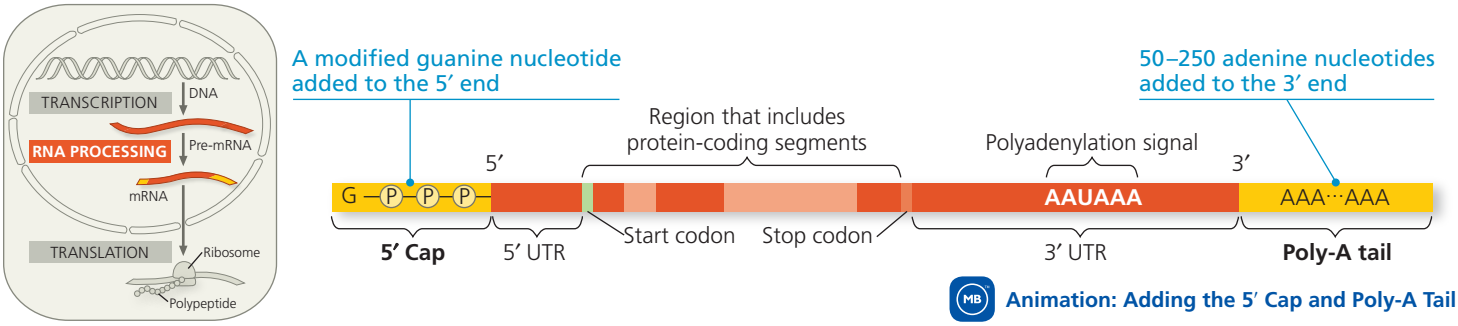
before translation. In eukaryotes, RNA polymerase II transcribes a sequence on the DNA called the polyadenylation signal sequence, which specifies a polyadenylation signal (AAUAAA) in the pre-mRNA. This is called a "signal" because once this stretch of six RNA nucleotides appears, it is immediately bound by certain proteins in the nucleus. Then, at a point about 10–35 nucleotides downstream from the AAUAAA, these proteins cut the RNA transcript free from the polymerase, releasing the pre-mRNA. The pre-mRNA then undergoes processing, the topic of the next section. Although that cleavage marks the end of the mRNA, the RNA polymerase II continues to transcribe. Enzymes begin to degrade the RNA starting at its newly exposed 5′ end. The polymerase continues transcribing, pursued by the enzymes, until they catch up to the polymerase and it dissociates from the DNA.

CONCEPT CHECK 17.2

1. What is a promoter? Is it located at the upstream or downstream end of a transcription unit?
2. What enables RNA polymerase to start transcribing a gene at the right place on the DNA in a bacterial cell? In a eukaryotic cell?
3. **WHAT IF?** ➤ Suppose X-rays caused a sequence change in the TATA box of a particular gene's promoter. How would that affect transcription of the gene? (See Figure 17.9.)

For suggested answers, see Appendix A.

Figure 17.11 RNA processing: addition of the 5′ cap and poly-A tail. Enzymes modify the two ends of a eukaryotic pre-mRNA molecule. The modified ends may promote the export of mRNA from the nucleus, and they help protect the mRNA from degradation. When the mRNA reaches the cytoplasm, the modified ends, in conjunction with certain cytoplasmic proteins, facilitate ribosome attachment. The 5′ cap and poly-A tail are not translated into protein, nor are the regions called the 5′ untranslated region (5′ UTR) and 3′ untranslated region (3′ UTR). The pink segments are introns, which will be described shortly (see Figure 17.12).



Animation: Adding the 5′ Cap and Poly-A Tail

CONCEPT 17.3

Eukaryotic cells modify RNA after transcription

Enzymes in the eukaryotic nucleus modify pre-mRNA in specific ways before the genetic message is dispatched to the cytoplasm. During this **RNA processing**, both ends of the primary transcript are altered. Also, in most cases, certain interior sections of the RNA molecule are cut out and the remaining parts spliced together. These modifications produce an mRNA molecule ready for translation.

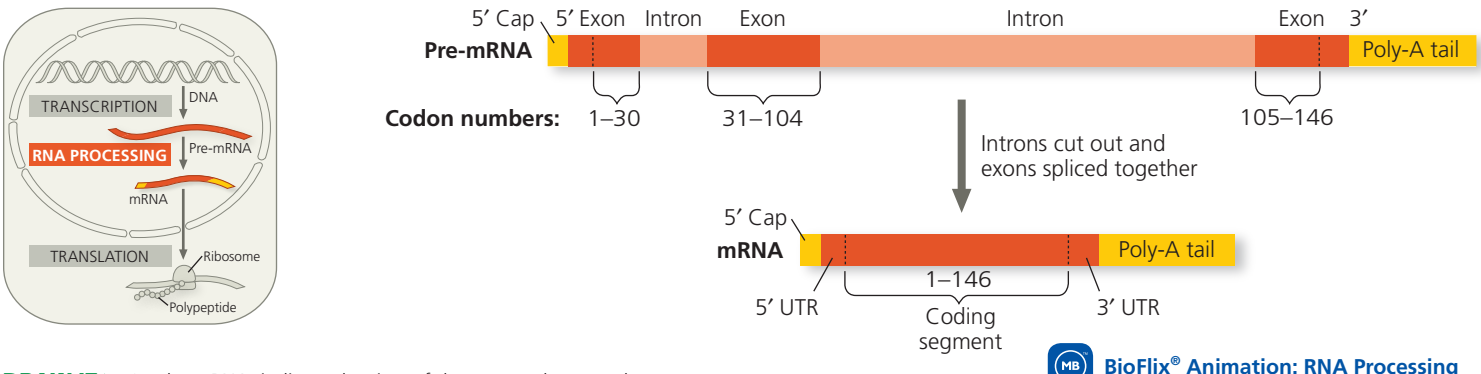
Alteration of mRNA Ends

Each end of a pre-mRNA molecule is modified in a particular way **(Figure 17.11)**. The 5′ end, which is synthesized first, receives a **5′ cap**, a modified form of a guanine (G) nucleotide added onto the 5′ end after transcription of the first 20–40 nucleotides. The 3′ end of the pre-mRNA molecule is also modified before the mRNA exits the nucleus. Recall that the pre-mRNA is cut and released soon after the polyadenylation signal, AAUAAA, is transcribed. At the 3′ end, an enzyme then adds 50–250 more adenine (A) nucleotides, forming a **poly-A tail**. The 5′ cap and poly-A tail share several important functions. First, they seem to facilitate the export of the mature mRNA from the nucleus. Second, they help protect the mRNA from degradation by hydrolytic enzymes. And third, they help ribosomes attach to the 5′ end of the mRNA once the mRNA reaches the cytoplasm. Figure 17.11 also shows the untranslated regions (UTRs) at the 5′ and 3′ ends of the mRNA (referred to as the 5′ UTR and 3′ UTR). The UTRs are parts of the mRNA that will not be translated into protein, but they have other functions, such as ribosome binding.

Split Genes and RNA Splicing

A remarkable stage of RNA processing in the eukaryotic nucleus is **RNA splicing (Figure 17.12)**, where large portions of the RNA molecules are removed and the remaining portions

Figure 17.12 RNA processing: RNA splicing. The RNA molecule shown here codes for β-globin, one of the polypeptides of hemoglobin. The numbers under the RNA refer to codons; β-globin is 146 amino acids long. The β-globin gene and its pre-mRNA transcript have three exons, corresponding to sequences that will leave the nucleus as mRNA. (The 5′ UTR and 3′ UTR are parts of exons because they are included in the mRNA; however, they do not code for protein.) During RNA processing, the introns are cut out and the exons spliced together. In many genes, the introns are much longer than the exons.



DRAW IT ➤ *On the mRNA, indicate the sites of the start and stop codons.*

BioFlix® Animation: RNA Processing

are reconnected. This cut-and-paste job is similar to editing a movie. The average length of a transcription unit along a human DNA molecule is about 27,000 nucleotide pairs, so the primary RNA transcript is also that long. However, the average-sized protein of 400 amino acids requires only 1,200 nucleotides in RNA to code for it. (Remember, each amino acid is encoded by a *triplet* of nucleotides.) This is because most eukaryotic genes and their RNA transcripts have long noncoding stretches of nucleotides, regions that are not translated. Even more surprising is that most of these noncoding sequences are interspersed between coding segments of the gene and thus between coding segments of the pre-mRNA. In other words, the sequence of DNA nucleotides that codes for a eukaryotic polypeptide is usually not continuous; it is split into segments. The noncoding segments of nucleic acid that lie between coding regions are called *int*ervening sequences, or **introns**. The other regions are called **exons**, because they are eventually *ex*pressed, usually by being translated into amino acid sequences. (Exceptions include the UTRs of the exons at the ends of the RNA, which make up part of the mRNA but are not translated into protein. Because of these exceptions, you may prefer to think of exons as sequences of RNA that *ex*it the nucleus.) The terms *intron* and *exon* are used for both RNA sequences and the DNA sequences that specify them.

In making a primary transcript from a gene, RNA polymerase II transcribes both introns and exons from the DNA, but the mRNA molecule that enters the cytoplasm is an abridged version. In RNA splicing, the introns are cut out from the molecule and the exons joined together, forming an mRNA molecule with a continuous coding sequence.

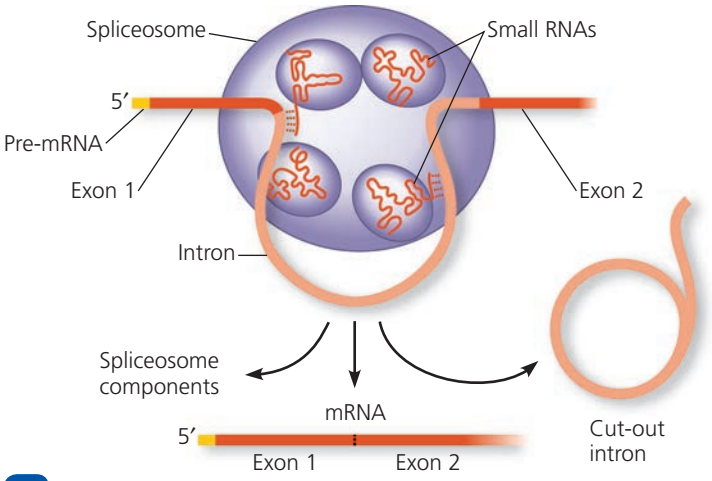
How is pre-mRNA splicing carried out? The removal of introns is accomplished by a large complex made of proteins and small RNAs called a **spliceosome**. This complex binds to several short nucleotide sequences along an intron, including key sequences at each end **(Figure 17.13)**. The intron is then released (and rapidly degraded), and the spliceosome joins together the two exons that flanked the intron. It turns out that the small RNAs in the spliceosome not only participate in spliceosome assembly and splice site recognition, but also catalyze the splicing reaction.

MB **Interview with Joan Steitz: Studying RNA, her "favorite molecule"**

Ribozymes

The idea of a catalytic role for the RNAs in the spliceosome arose from the discovery of **ribozymes**, RNA molecules that function as enzymes. In some organisms, RNA splicing can occur without proteins or even additional RNA molecules: The intron RNA functions as a ribozyme and catalyzes its own excision! For example, in the ciliate protist *Tetrahymena*, self-splicing occurs in the production of ribosomal RNA (rRNA), a component of the organism's

Figure 17.13 A spliceosome splicing a pre-mRNA. The diagram shows a portion of a pre-mRNA transcript, with an intron (pink) flanked by two exons (red). Small RNAs within the spliceosome base-pair with nucleotides at specific sites along the intron. Next, small spliceosome RNAs catalyze cutting of the pre-mRNA and the splicing together of the exons, releasing the intron for rapid degradation.



MB **Animation: A Spliceosome**

ribosomes. The pre-rRNA actually removes its own introns. The discovery of ribozymes rendered obsolete the idea that all biological catalysts are proteins.

Three properties of RNA enable some RNA molecules to function as enzymes. First, because RNA is single-stranded, a region of an RNA molecule may base-pair, in an antiparallel arrangement, with a complementary region elsewhere in the same molecule; this gives the molecule a particular three-dimensional structure. A specific structure is essential to the catalytic function of ribozymes, just as it is for enzymatic proteins. Second, like certain amino acids in an enzymatic protein, some of the bases in RNA contain functional groups that can participate in catalysis. Third, the ability of RNA to hydrogen-bond with other nucleic acid molecules (either RNA or DNA) adds specificity to its catalytic activity. For example, complementary base pairing between the RNA of the spliceosome and the RNA of a primary RNA transcript precisely locates the region where the ribozyme catalyzes splicing. Later in this chapter, you will see how these properties of RNA also allow it to perform important noncatalytic roles in the cell, such as recognition of the three-nucleotide codons on mRNA.

The Functional and Evolutionary Importance of Introns

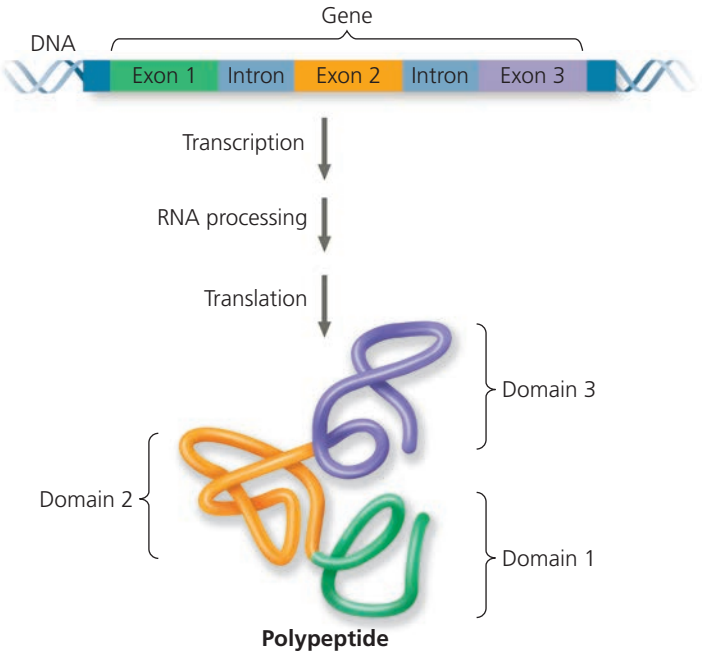
EVOLUTION Whether or not RNA splicing and the presence of introns have provided selective advantages during evolutionary history is a matter of some debate. In any case, it is informative to consider their possible adaptive benefits. Specific functions have not been identified for most introns, but at least some contain sequences that regulate gene expression, and many affect gene products.

One important consequence of the presence of introns in genes is that a single gene can encode more than one kind of polypeptide. Many genes are known to give rise to two or more different polypeptides, depending on which segments are treated as exons during RNA processing; this is called **alternative RNA splicing** (see Figure 18.13). Results from the Human Genome Project (discussed in Concept 20.1) suggest that alternative RNA splicing is one reason humans can get along with about the same number of genes as a nematode (roundworm). Because of alternative splicing, the number of different protein products an organism produces can be much greater than its number of genes.

Proteins often have a modular architecture consisting of discrete structural and functional regions called **domains**. One domain of an enzyme, for example, might include the active site, while another might allow the enzyme to bind to a cellular membrane. In quite a few cases, different exons code for the different domains of a protein **(Figure 17.14)**.

The presence of introns in a gene may facilitate the evolution of new and potentially beneficial proteins as a result of a process known as *exon shuffling* (see Figure 20.16). Introns increase the probability of crossing over between the exons of alleles of a gene—simply by providing more terrain for crossovers without interrupting coding sequences. This might result in new combinations of exons and proteins with altered structure and function. We can also imagine the occasional mixing and matching of exons between completely different (nonallelic) genes. Exon shuffling of either sort could lead to new proteins with novel combinations of functions. While most of the shuffling would result in nonbeneficial changes, occasionally a beneficial variant might arise.

Figure 17.14 Correspondence between exons and protein domains.



CONCEPT CHECK 17.3

1. There are about 20,000 human protein-coding genes. How can human cells make 75,000–100,000 different proteins?
2. How is RNA splicing similar to how you would watch a recorded television show? What would introns be?
3. **WHAT IF?** ➤ What would be the effect of treating cells with an agent that removed the cap from mRNAs?

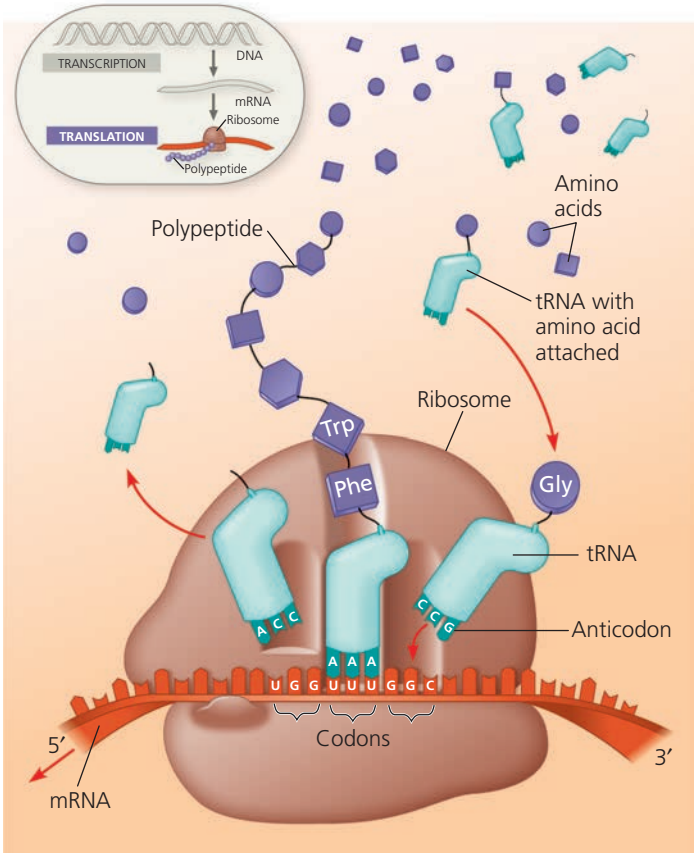
For suggested answers, see Appendix A.

CONCEPT 17.4

Translation is the RNA-directed synthesis of a polypeptide: *a closer look*

We will now examine how genetic information flows from mRNA to protein—the process of translation **(Figure 17.15)**. We'll focus on the basic steps of translation that occur in both bacteria and eukaryotes, while pointing out key differences.

Figure 17.15 Translation: the basic concept. As a molecule of mRNA is moved through a ribosome, codons are translated into amino acids, one by one. The translators, or interpreters, are tRNA molecules, each type with a specific anticodon at one end and a corresponding amino acid at the other end. A tRNA adds its amino acid cargo to a growing polypeptide chain when the anticodon hydrogen-bonds to the complementary codon on the mRNA.



BioFlix® Animation: Translation
Animation: Overview of Translation

Molecular Components of Translation

In the process of translation, a cell "reads" a genetic message and builds a polypeptide accordingly. The message is a series of codons along an mRNA molecule, and the translator is called a **transfer RNA (tRNA)**. The function of a tRNA is to transfer an amino acid from the cytoplasmic pool of amino acids to a growing polypeptide in a ribosome. A cell keeps its cytoplasm stocked with all 20 amino acids, either by synthesizing them from other compounds or by taking them up from the surrounding solution. The ribosome, a structure made of proteins and RNAs, adds each amino acid brought to it by a tRNA to the growing end of a polypeptide chain (see Figure 17.15).

Translation is simple in principle but complex in its biochemistry and mechanics, especially in the eukaryotic cell. In dissecting translation, we'll focus on the slightly less complicated version of the process that occurs in bacteria. We'll first look at the major players in this process.

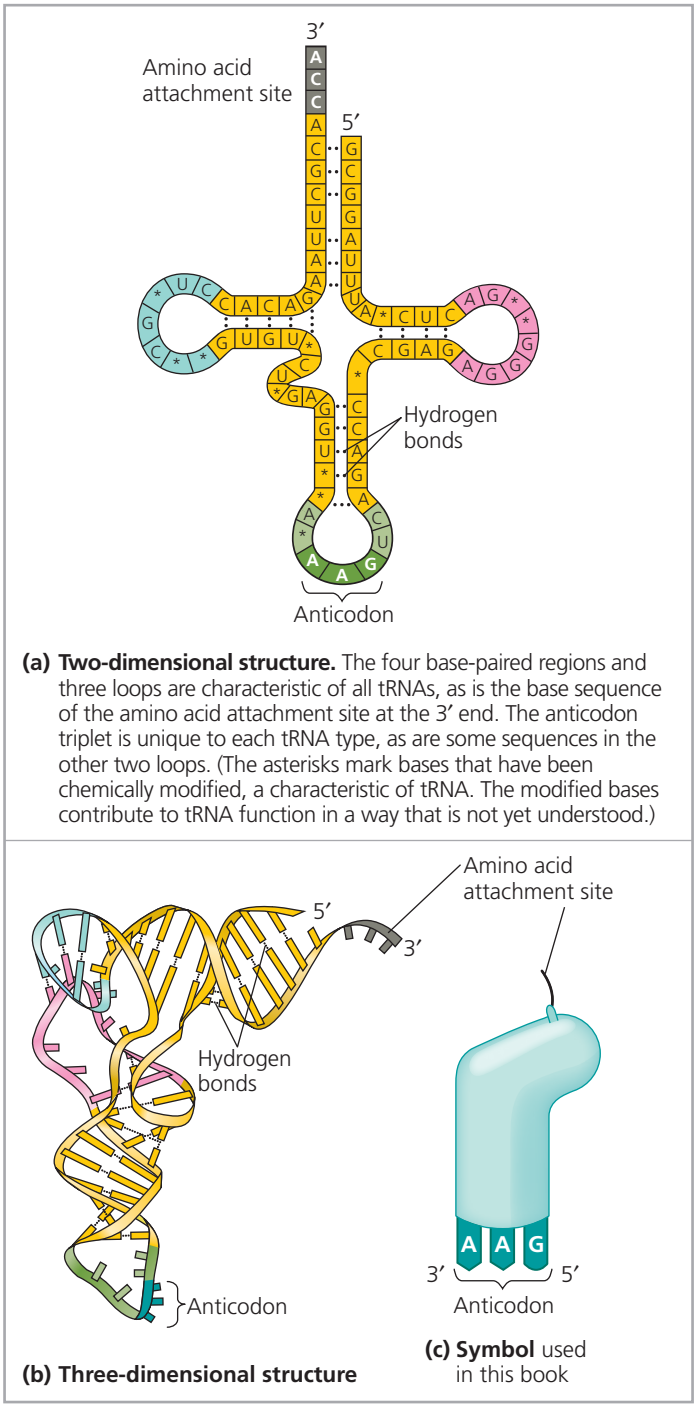
The Structure and Function of Transfer RNA

The key to translating a genetic message into a specific amino acid sequence is the fact that each tRNA molecule enables translation of a given mRNA codon into a certain amino acid. This is possible because a tRNA bears a specific amino acid at one end of its three-dimensional structure, while at the other end is a nucleotide triplet that can base-pair with the complementary codon on mRNA.

A tRNA molecule consists of a single RNA strand that is only about 80 nucleotides long (compared to hundreds of nucleotides for most mRNA molecules). Because of the presence of complementary stretches of nucleotide bases that can hydrogen-bond to each other, this single strand can fold back on itself and form a molecule with a three-dimensional structure. Flattened into one plane to clarify this base pairing, a tRNA molecule looks like a cloverleaf **(Figure 17.16a)**. The tRNA actually twists and folds into a compact three-dimensional structure that is roughly L-shaped **(Figure 17.16b)**, with the 5′ and 3′ ends of the linear tRNA both located near one end of the structure. The protruding 3′ end acts as the attachment site for an amino acid. The loop extending from the other end of the L includes the **anticodon**, the particular nucleotide triplet that base-pairs to a specific mRNA codon. Thus, the structure of a tRNA molecule fits its function.

Anticodons are conventionally written 3′ → 5′ to align properly with codons written 5′ → 3′ (see Figure 17.15). (For base pairing, RNA strands must be antiparallel, like DNA.) As an example of how tRNAs work, consider the mRNA codon 5′-GGC-3′, which is translated as the amino acid glycine. The tRNA that base-pairs with this codon by hydrogen bonding has 3′-CCG-5′ as its anticodon and carries glycine at its other end (see the incoming tRNA approaching the ribosome in Figure 17.15). As an mRNA molecule is moved through a ribosome, glycine will be added to the polypeptide chain whenever the codon 5′-GGC-3′ is presented for translation. Codon by codon, the genetic message is translated as tRNAs

Figure 17.16 The structure of transfer RNA (tRNA).



(a) Two-dimensional structure. The four base-paired regions and three loops are characteristic of all tRNAs, as is the base sequence of the amino acid attachment site at the 3′ end. The anticodon triplet is unique to each tRNA type, as are some sequences in the other two loops. (The asterisks mark bases that have been chemically modified, a characteristic of tRNA. The modified bases contribute to tRNA function in a way that is not yet understood.)

(b) Three-dimensional structure

(c) Symbol used in this book

VISUAL SKILLS ➤ *Look at the tRNA shown in this figure. Based on its anticodon, identify the codon it would bind to, as well as the amino acid that it would carry.*

HHMI Video: RNA Folding

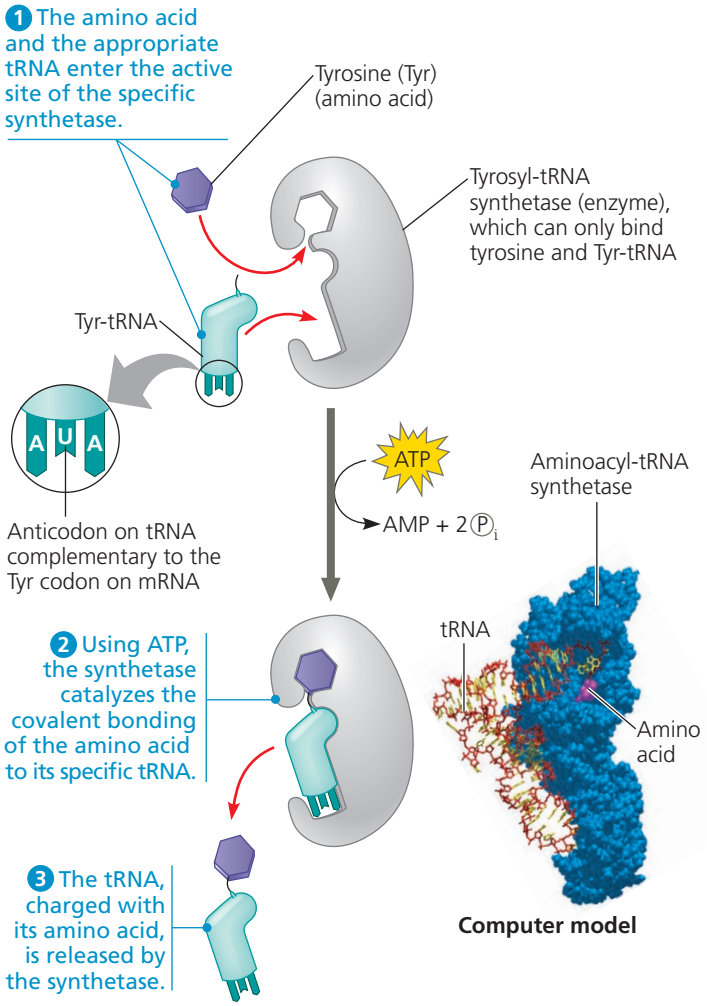
position each amino acid in the order prescribed, and the ribosome adds that amino acid onto the growing polypeptide chain. The tRNA molecule is a translator in the sense that, in the context of the ribosome, it can read a nucleic acid word (the mRNA codon) and interpret it as a protein word (the amino acid).

Like mRNA and other types of cellular RNA, transfer RNA molecules are transcribed from DNA templates. In a eukaryotic cell, tRNA, like mRNA, is made in the nucleus and then travels to the cytoplasm, where it will participate in the process of translation. In both bacterial and eukaryotic cells, each tRNA molecule is used repeatedly, picking up its designated amino acid in the cytosol, depositing this cargo onto a polypeptide chain at the ribosome, and then leaving the ribosome, ready to pick up another of the same amino acid.

The accurate translation of a genetic message requires two instances of molecular recognition. First, a tRNA that binds to an mRNA codon specifying a particular amino acid must carry that amino acid, and no other, to the ribosome. The correct matching up of tRNA and amino acid is carried out by a family of related enzymes that are aptly named **aminoacyl-tRNA synthetases** **(Figure 17.17)**. The active site of each type of aminoacyl-tRNA synthetase fits only a specific combination of amino acid and tRNA. There are 20 different synthetases, one for each amino acid. A synthetase joins a given amino acid to an appropriate tRNA; one synthetase is able to bind to all the different tRNAs for its particular amino acid. The synthetase catalyzes the covalent attachment of the amino acid to its tRNA in a process driven by the hydrolysis of ATP. The resulting aminoacyl tRNA, also called a charged tRNA, is released from the enzyme and is then available to deliver its amino acid to a growing polypeptide chain on a ribosome.

The second instance of molecular recognition is the pairing of the tRNA anticodon with the appropriate mRNA codon. If one tRNA variety existed for each mRNA codon specifying an amino acid, there would be 61 tRNAs (see Figure 17.6). In bacteria, however, there are only about 45 tRNAs, signifying that some tRNAs must be able to bind to more than one codon. Such versatility is possible because the rules for base pairing between the third nucleotide base of a codon and the corresponding base of a tRNA anticodon are relaxed compared to those at other codon positions. For example, the nucleotide base U at the 5′ end of a tRNA anticodon can pair with either A or G in the third position (at the 3′ end) of an mRNA codon. The flexible base pairing at this codon position is called **wobble**. Wobble explains why the synonymous codons for a given amino acid most often differ in their third nucleotide base. Accordingly, a tRNA with the anticodon 3′-UCU-5′ can base-pair with either the mRNA codon 5′-AGA-3′ or 5′-AGG-3′, both of which code for arginine (see Figure 17.6).

Figure 17.17 Aminoacyl-tRNA synthetases provide specificity in joining amino acids to their tRNAs. Linkage of a tRNA to its amino acid is an endergonic process that occurs at the expense of ATP, which loses two phosphate groups, becoming AMP.



MB **Animation: Transfer RNA**

The Structure and Function of Ribosomes

Ribosomes facilitate the specific coupling of tRNA anticodons with mRNA codons during protein synthesis. A ribosome consists of a large subunit and a small subunit, each made up of proteins and one or more **ribosomal RNAs (rRNAs)**. In eukaryotes, the subunits are made in the nucleolus. Ribosomal RNA genes are transcribed, and the RNA is processed and assembled with proteins imported from the cytoplasm. Completed ribosomal subunits are then exported via nuclear pores to the cytoplasm. In both bacteria and eukaryotes, a large and a small subunit join to form a functional ribosome only when attached to an mRNA molecule. About one-third of the mass of a ribosome is made up of proteins; the rest consists of three rRNA molecules (in bacteria) or four (in eukaryotes). Because most cells contain thousands of ribosomes, rRNA is the most abundant type of cellular RNA.

Although the ribosomes of bacteria and eukaryotes are very similar in structure and function, eukaryotic ribosomes are slightly larger, as well as differing somewhat from bacterial ribosomes in their molecular composition. The differences are medically significant. Certain antibiotic drugs can inactivate bacterial ribosomes without affecting eukaryotic ribosomes. These drugs, including tetracycline and streptomycin, are used to combat bacterial infections.

The structure of a ribosome reflects its function of bringing mRNA together with tRNAs carrying amino acids. In addition to a binding site for mRNA, each ribosome has three binding sites for tRNA **(Figure 17.18)**. The **P site** (peptidyl-tRNA binding site) holds the tRNA carrying the growing polypeptide chain, while the **A site** (aminoacyl-tRNA binding site) holds the tRNA carrying the next amino acid to be added to the chain. Discharged tRNAs leave the ribosome from the **E site** (exit site). The ribosome holds the tRNA and mRNA in close proximity and positions the new amino acid so that it can be added to the carboxyl end of the growing polypeptide. It then catalyzes the formation of the peptide bond. As the polypeptide becomes longer, it passes through an *exit tunnel* in the ribosome's large subunit. When the polypeptide is complete, it is released through the exit tunnel.

The widely accepted model is that rRNAs, rather than ribosomal proteins, are primarily responsible for both the structure and the function of the ribosome. The proteins, which are largely on the exterior, support the shape changes of the rRNA molecules as they carry out catalysis during translation. Ribosomal RNA is the main constituent of the A and P sites and of the interface between the two subunits; it also acts as the catalyst of peptide bond formation. Thus, a ribosome could actually be considered one colossal ribozyme!

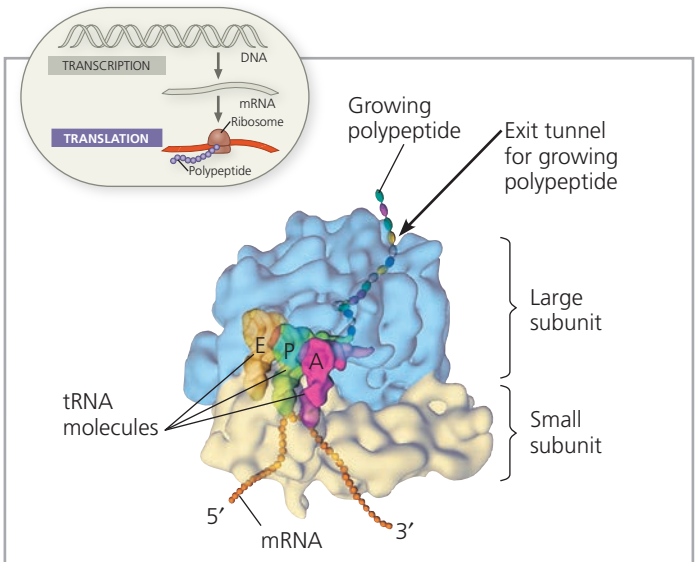
Building a Polypeptide

We can divide translation, the synthesis of a polypeptide, into three stages: initiation, elongation, and termination. All three require protein "factors" that aid in the translation process. Some steps of initiation and elongation also require energy, provided by the hydrolysis of guanosine triphosphate (GTP).

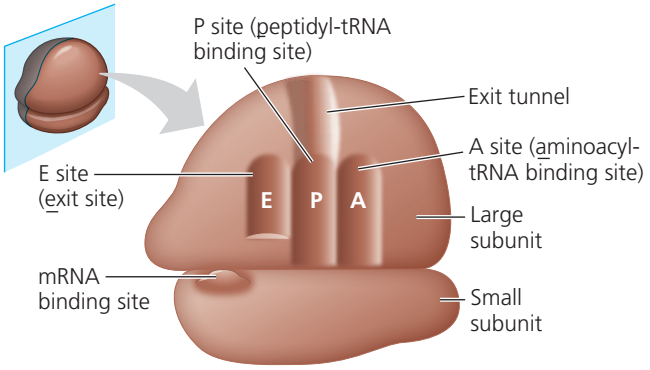
Ribosome Association and Initiation of Translation

In either bacteria or eukaryotes, the start codon (AUG) signals the start of translation; this is important because it establishes the codon reading frame for the mRNA. In the first step of translation, a small ribosomal subunit binds to both the mRNA and a specific initiator tRNA, which carries the amino acid methionine. In bacteria, the small subunit can bind the two in either order; it binds the mRNA at a specific RNA sequence, just upstream of the AUG start codon. In the **Scientific Skills Exercise**, you can work with DNA sequences encoding the ribosomal binding sites on the mRNAs of a group of *Eschericia coli*

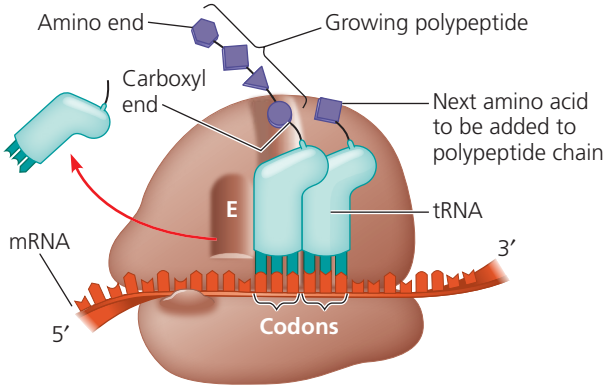
Figure 17.18 The anatomy of a functioning ribosome.



(a) Computer model of functioning ribosome. This is a model of a bacterial ribosome, showing its overall shape. The eukaryotic ribosome is roughly similar. A ribosomal subunit is a complex of ribosomal RNA molecules and proteins.



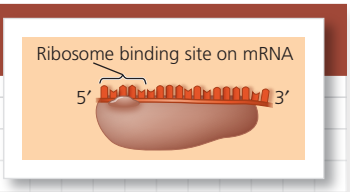
(b) Schematic model showing binding sites. A ribosome has an mRNA binding site and three tRNA binding sites, known as the A, P, and E sites. This schematic ribosome will appear in later diagrams.



(c) Schematic model with mRNA and tRNA. A tRNA fits into a binding site when its anticodon base-pairs with an mRNA codon. The P site holds the tRNA attached to the growing polypeptide. The A site holds the tRNA carrying the next amino acid to be added to the polypeptide chain. Discharged tRNAs leave from the E site. The polypeptide grows at its carboxyl end.

Animation: Ribosomes

SCIENTIFIC SKILLS EXERCISE



Interpreting a Sequence Logo

How Can a Sequence Logo Be Used to Identify Ribosome Binding Sites on Bacterial mRNAs? When initiating translation, ribosomes bind to an mRNA at a ribosome binding site upstream of the AUG start codon. Because mRNAs from different genes all bind to a ribosome, the genes encoding these mRNAs are likely to have a similar base sequence where the ribosomes bind. Therefore, candidate ribosome binding sites on mRNA can be identified by comparing DNA sequences (and thus the mRNA sequences) of multiple genes in a species, searching the region upstream of the start codon for shared (conserved) stretches of bases. In this exercise, you will analyze DNA sequences from multiple such genes, represented by a visual graphic called a sequence logo.

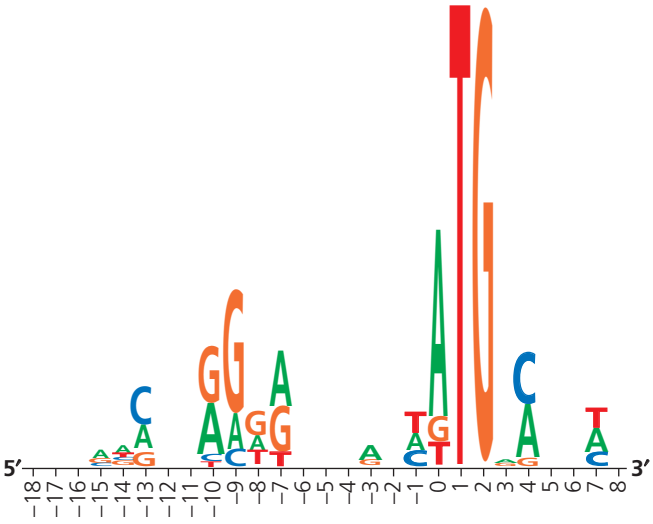
How the Experiment Was Done The DNA sequences of 149 genes from the *E. coli* genome were aligned using computer software. The aim was to identify similar base sequences—at the appropriate location in each gene—as potential ribosome binding sites. Rather than presenting the data as a series of 149 sequences aligned in a column (a sequence alignment), the researchers used a sequence logo.

Data from the Experiment To show how sequence logos are made, the potential ribosome binding regions from 10 *E. coli* genes are shown below in a sequence alignment, followed by the sequence logo derived from the aligned sequences. Note that the DNA shown is the nontemplate (coding) strand, which is how DNA sequences are typically presented.

	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8
thrA	G	G	T	A	A	C	G	A	G	G	T	A	A	C	A	A	C	C	A	T	G	C	G	A	G	T	G
lacA	C	A	T	A	A	C	G	G	A	G	T	G	A	T	C	G	C	A	T	T	G	A	A	C	A	T	G
lacY	C	G	C	G	T	A	A	G	G	A	A	A	T	C	C	A	T	T	A	T	G	T	A	C	T	A	T
lacZ	T	T	C	A	C	A	C	A	G	G	A	A	A	C	A	G	C	T	A	T	G	A	C	C	A	T	G
lacI	C	A	A	T	T	C	A	G	G	G	T	G	G	T	G	A	A	T	G	T	G	A	A	A	C	C	A
recA	G	G	C	A	T	G	A	C	A	G	G	A	G	T	A	A	A	A	A	T	G	G	C	T	A	T	C
galR	A	C	C	C	A	C	T	A	A	G	G	T	A	T	T	T	T	C	A	T	G	G	C	G	A	C	C
metJ	A	A	G	A	G	G	A	T	T	A	A	G	T	A	T	C	T	C	A	T	G	G	C	T	G	A	A
lexA	A	T	A	C	A	C	C	C	A	G	G	G	G	C	G	G	A	A	T	G	A	A	A	A	G	C	G
trpR	T	A	A	C	A	A	T	G	G	C	G	A	C	A	T	A	T	T	A	T	G	G	C	C	C	A	A

5′ 3′

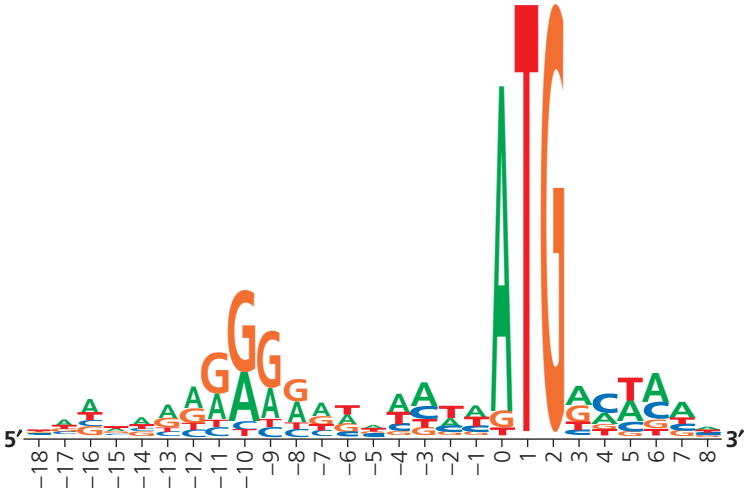
Sequence alignment



Sequence logo

INTERPRET THE DATA

1. In the sequence logo (bottom, left), the horizontal axis shows the primary sequence of the DNA by nucleotide position. Letters for each base are stacked on top of each other according to their relative frequency at that position among the aligned sequences, with the most common base as the largest letter at the top of the stack. The height of each letter represents the relative frequency of that base *at that position*. (a) In the sequence alignment, count the number of each base at position −9 and order them from most to least frequent. Compare this to the size and placement of each base at −9 in the logo. (b) Do the same for positions 0 and 1.
2. The height of a stack of letters in a logo indicates the predictive power of that stack (determined statistically). If the stack is tall, we can be more confident in predicting what base will be in that position if a new sequence is added to the logo. For example, at position 2 in the sequence alignment, all 10 sequences have a G; the probability of finding a G there in a new sequence is very high, as is the stack in the sequence logo. For short stacks, the bases all have about the same frequency, so it's hard to predict a base at those positions. (a) Looking at the sequence logo, which two positions have the most predictable bases? What bases do you predict would be at those positions in a newly sequenced gene? (b) Which 12 positions have the least predictable bases? How do you know? How does this reflect the relative frequencies of the bases shown at these positions in the sequence alignment? Use the two leftmost positions of the 12 as examples in your answer.
3. In the actual experiment, the researchers used 149 sequences to build their sequence logo, which is shown below. There is a stack at each position, even if short, because the sequence logo includes more data. (a) Which three positions in this sequence logo have the most predictable bases? Name the most frequent base at each. (b) Which four positions have the least predictable bases? How can you tell?



4. A consensus sequence identifies the base occurring most often at each position in the set of sequences. (a) Write out the consensus sequence of this (the nontemplate) strand. In any position where the base can't be determined, put a dash. (b) Which provides more information—the consensus sequence or the sequence logo? What is lost in the less informative method?
5. (a) Based on the logo, what five adjacent base positions in the 5′ UTR region are most likely to be involved in ribosome binding? Explain. (b) What is represented by the bases in positions 0–2?

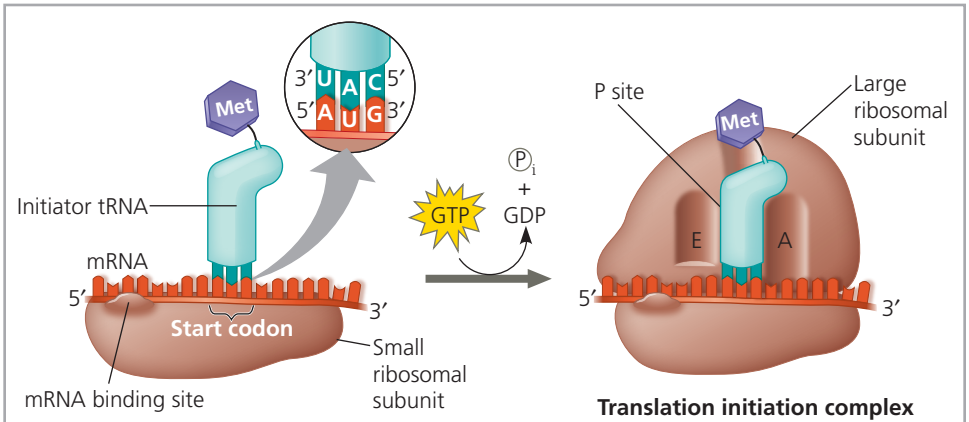
Instructors: A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Further Reading T. D. Schneider and R. M. Stephens, Sequence logos: A new way to display consensus sequences, *Nucleic Acids Research* 18:6097–6100 (1990).

genes. In eukaryotes, the small subunit, with the initiator tRNA already bound, binds to the 5′ cap of the mRNA and then moves, or *scans*, downstream along the mRNA until it reaches the start codon; the initiator tRNA then hydrogen-bonds to the AUG start codon.

Thus, the first components to associate with each other during the initiation stage of translation are mRNA, a tRNA bearing the first amino acid of the polypeptide, and the small ribosomal subunit **(Figure 17.19)**. This is followed by the attachment of a large ribosomal subunit, completing the *translation initiation complex*. Proteins called *initiation factors* are required to bring all these components together. The cell also expends energy obtained by hydrolysis of a GTP molecule to form the initiation complex. At the completion of the initiation process, the initiator tRNA sits in the P site of the ribosome, and the vacant A site is ready for the next aminoacyl tRNA. Note that a polypeptide is always synthesized in one direction, from the initial methionine at the amino end, also called the N-terminus, toward the final amino acid at the carboxyl end, also called the C-terminus (see Figure 5.15).

Figure 17.19 The initiation of translation. Animation: Initiation of Translation



1 A small ribosomal subunit binds to a molecule of mRNA. In a bacterial cell, the mRNA binding site on this subunit recognizes a specific nucleotide sequence on the mRNA just upstream of the start codon. An initiator tRNA, with the anticodon UAC, base-pairs with the start codon, AUG. This tRNA carries the amino acid methionine (Met).

2 The arrival of a large ribosomal subunit completes the initiation complex. Proteins called initiation factors (not shown) are required to bring all the translation components together. Hydrolysis of GTP provides the energy for the assembly. The initiator tRNA is in the P site; the A site is available to the tRNA bearing the next amino acid.

Elongation of the Polypeptide Chain

In the elongation stage of translation, amino acids are added one by one to the previous amino acid at the C-terminus of the growing chain. Each addition involves several proteins called *elongation factors* and occurs in a three-step cycle described in **Figure 17.20**. Energy expenditure occurs in the first and third steps. Codon recognition requires hydrolysis of one molecule of GTP, which increases the accuracy and efficiency of this step. One more GTP is hydrolyzed to provide energy for the translocation step.

The mRNA is moved through the ribosome in one direction only, 5′ end first; this is equivalent to the ribosome moving 5′ → 3′ on the mRNA. The main point is that the ribosome and the mRNA move relative to each other, unidirectionally, codon by codon. The elongation cycle takes less than a tenth of a second in bacteria and is repeated as each amino acid is added until the polypeptide is complete. The empty tRNAs that are released from the E site return to the cytoplasm, where they will be reloaded with the appropriate amino acid (see Figure 17.17).

Termination of Translation

The final stage of translation is termination **(Figure 17.21)**. Elongation continues until a stop codon in the mRNA reaches the A site. The nucleotide base triplets UAG, UAA, and UGA (all written 5′ → 3′) do not code for amino acids but instead act as signals to stop translation. A *release factor*, a protein shaped like an aminoacyl tRNA, binds directly to the stop codon in the A site. The release factor causes the addition of a water molecule instead of an amino acid to the polypeptide chain. (Water molecules are abundant in the cytosol.) This reaction breaks (hydrolyzes) the bond between the completed polypeptide and the tRNA in the P site, releasing the polypeptide through the exit tunnel of the ribosome's large subunit. The remainder of the translation assembly then comes apart in a multistep process, aided by other protein factors. Breakdown of the translation assembly requires the hydrolysis of two more GTP molecules.

Completing and Targeting the Functional Protein

The process of translation is often not sufficient to make a functional protein. In this section, you will learn about modifications that polypeptide chains undergo after the translation process as well as some of the mechanisms used to target completed proteins to specific sites in the cell.

Protein Folding and Post-Translational Modifications

During its synthesis, a polypeptide chain begins to coil and fold spontaneously as a consequence of its amino acid sequence (primary structure), forming a protein with a specific shape: a three-dimensional molecule with secondary and tertiary structure (see Figure 5.18). Thus, a gene determines primary structure, which in turn determines shape.

Figure 17.20 The elongation cycle of translation. The hydrolysis of GTP plays an important role in the elongation process; elongation factors are not shown.

Animation: Elongation Cycle of Translation

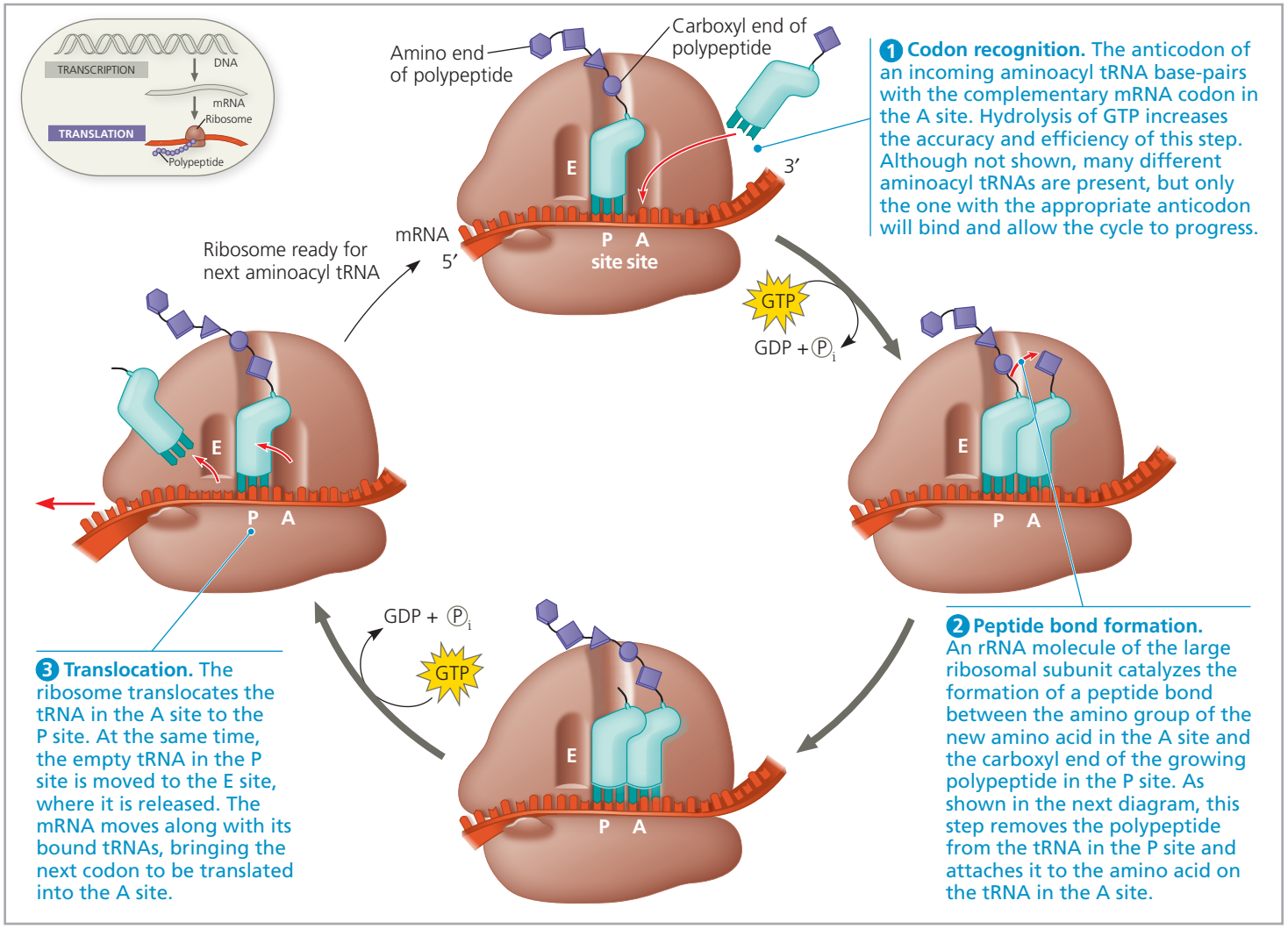
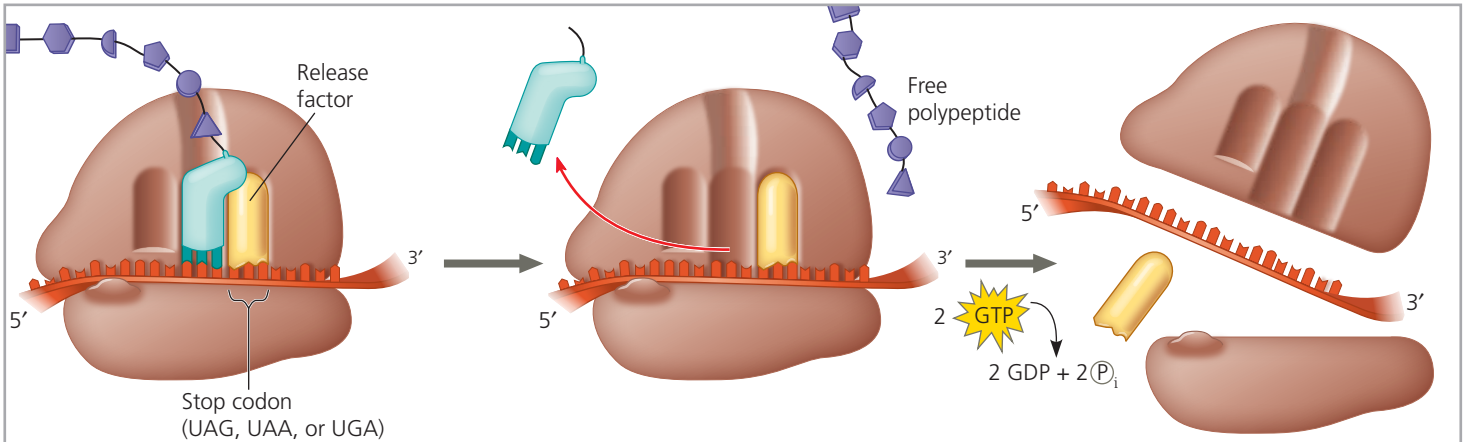


Figure 17.21 The termination of translation. Like elongation, termination requires GTP hydrolysis as well as additional protein factors, which are not shown here.

Animation: Termination of Translation



Additional steps—*post-translational modifications*—may be required before the protein can begin doing its particular job in the cell. Certain amino acids may be chemically modified by the attachment of sugars, lipids, phosphate groups, or other additions. Enzymes may remove one or more amino acids from the leading (amino) end of the polypeptide chain. In some cases, a polypeptide chain may be enzymatically cleaved into two or more pieces. In other cases, two or more polypeptides that are synthesized separately may come together, if the protein has quaternary structure; an example is hemoglobin (see Figure 5.18).

BioFlix® Animation: Protein Processing

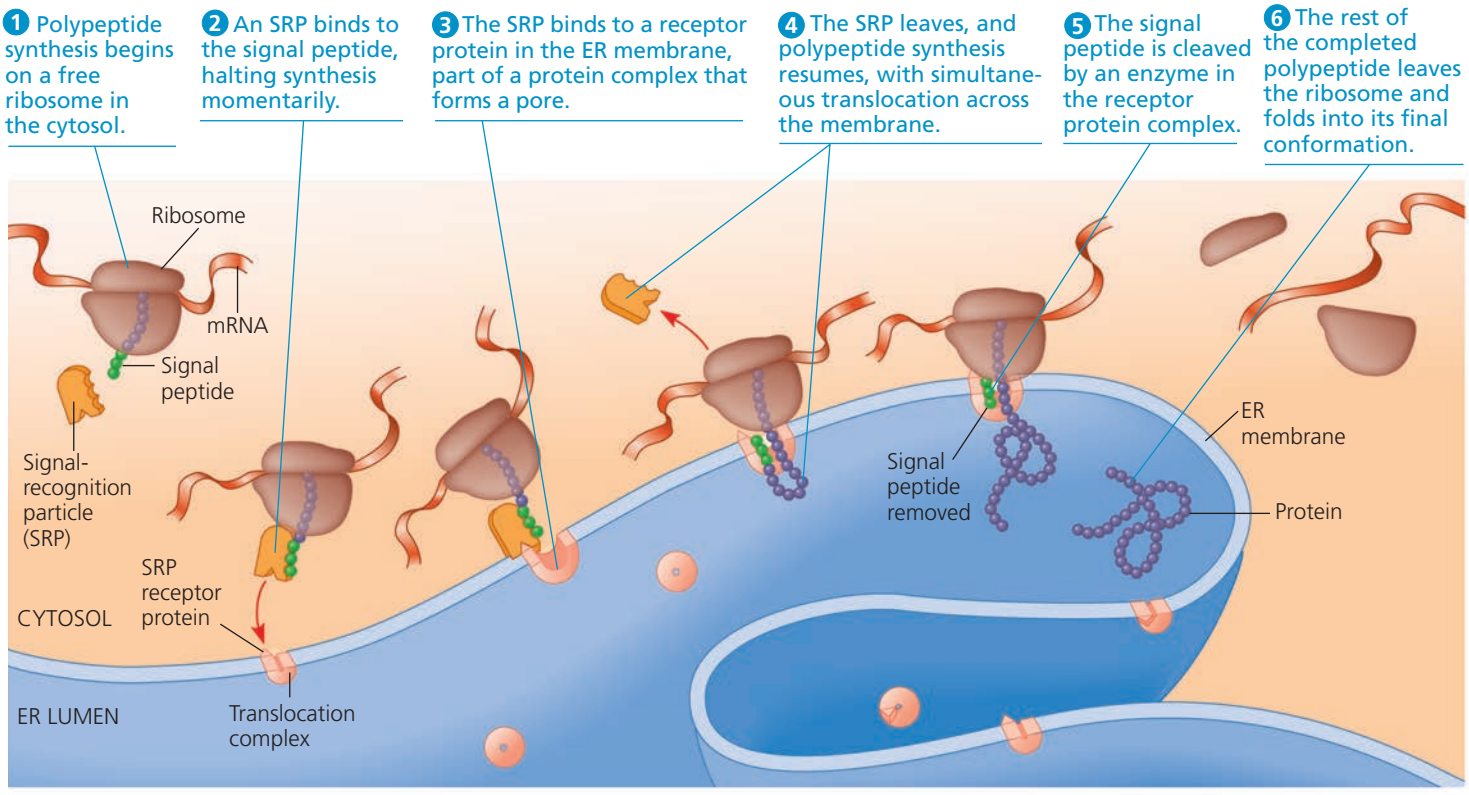
Targeting Polypeptides to Specific Locations

In electron micrographs of eukaryotic cells active in protein synthesis, two populations of ribosomes are evident: free and bound (see Figure 7.10). Free ribosomes are suspended in the cytosol and mostly synthesize proteins that stay in the cytosol and function there. In contrast, bound ribosomes are attached to the cytosolic side of the endoplasmic reticulum (ER) or to the nuclear envelope. Bound ribosomes make proteins of the endomembrane system (see Figure 7.15) as well as proteins secreted from the cell, such as insulin. It is important to note that the ribosomes themselves are identical and can alternate between being free ribosomes one time they are used and being bound ribosomes the next.

What determines whether a ribosome is free in the cytosol or bound to rough ER? Polypeptide synthesis always begins in the cytosol as a free ribosome starts to translate an mRNA molecule. There, the process continues to completion—*unless* the growing polypeptide itself cues the ribosome to attach to the ER. The polypeptides of proteins destined for the endomembrane system or for secretion are marked by a **signal peptide**, which targets the protein to the ER **(Figure 17.22)**. The signal peptide, a sequence of about 20 amino acids at or near the leading end (N-terminus) of the polypeptide, is recognized as it emerges from the ribosome by a protein-RNA complex called a **signal-recognition particle (SRP)**. This particle escorts the ribosome to a receptor protein built into the ER membrane. The receptor is part of a multiprotein translocation complex. Polypeptide synthesis continues there, and the growing polypeptide snakes across the membrane into the ER lumen via a protein pore. The rest of the completed polypeptide, if it is to be secreted from the cell, is released into solution within the ER lumen. Alternatively, if the polypeptide is to be a membrane protein, it remains partially embedded in the ER membrane. In either case, it travels in a transport vesicle to its destination (see, for example, Figure 8.9).

Other kinds of signal peptides are used to target polypeptides to mitochondria, chloroplasts, the interior of the nucleus, and other organelles that are not part of the endomembrane system. The critical difference in these cases is that translation is completed in the cytosol before the polypeptide is imported

Figure 17.22 The signal mechanism for targeting proteins to the ER.



MAKE CONNECTIONS ➤ *If this protein were destined for secretion, what would happen to it after its synthesis was completed? See Figure 8.9.*

Video: Cotranslational Translocation

into the organelle. Translocation mechanisms also vary, but in all cases studied to date, the "postal zip codes" that address proteins for secretion or to cellular locations are signal peptides of some sort. Bacteria also employ signal peptides to target proteins to the plasma membrane for secretion.

Making Multiple Polypeptides in Bacteria and Eukaryotes

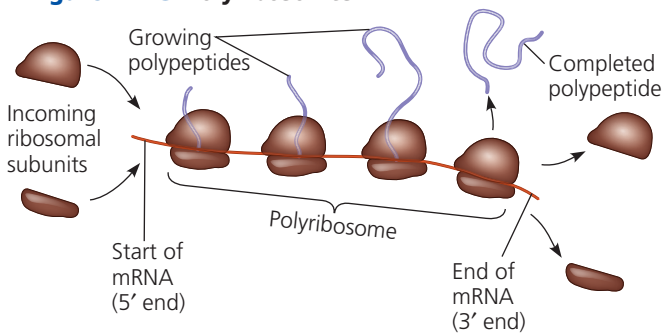
In previous sections, you learned how a single polypeptide is synthesized using the information encoded in an mRNA molecule. When a polypeptide is required in a cell, though, the need is for many copies, not just one.

A single ribosome can make an average-sized polypeptide in less than a minute. In both bacteria and eukaryotes, however, multiple ribosomes translate an mRNA at the same time **(Figure 17.23)**; that is, a single mRNA is used to make many copies of a polypeptide simultaneously. Once a ribosome is far enough past the start codon, a second ribosome can attach to the mRNA, eventually resulting in a number of ribosomes trailing along the mRNA. Such strings of ribosomes, called **polyribosomes** (or **polysomes**), can be seen with an electron microscope; they can be either free or bound. They enable a cell to rapidly make many copies of a polypeptide.

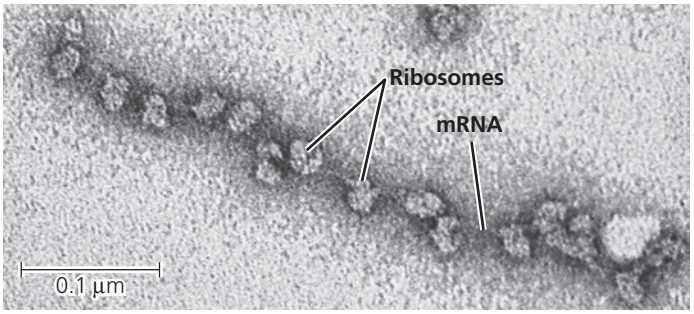
Another way both bacteria and eukaryotes augment the number of copies of a polypeptide is by transcribing multiple mRNAs from the same gene. However, the coordination of the two processes—transcription and translation—differs in the two groups. The most important differences between bacteria and eukaryotes arise from the bacterial cell's lack of compartmental organization. Like a one-room workshop, a bacterial cell ensures a streamlined operation by coupling the two processes. In the absence of a nucleus, it can simultaneously transcribe and translate the same gene **(Figure 17.24)**, and the newly made protein can quickly diffuse to its site of function.

In contrast, the eukaryotic cell's nuclear envelope segregates transcription from translation and provides a compartment for extensive RNA processing. This processing stage includes additional steps, discussed earlier, the regulation of which can help coordinate the eukaryotic cell's elaborate activities. **Figure 17.25** summarizes the path from gene to polypeptide in a eukaryotic cell.

Figure 17.23 Polyribosomes.

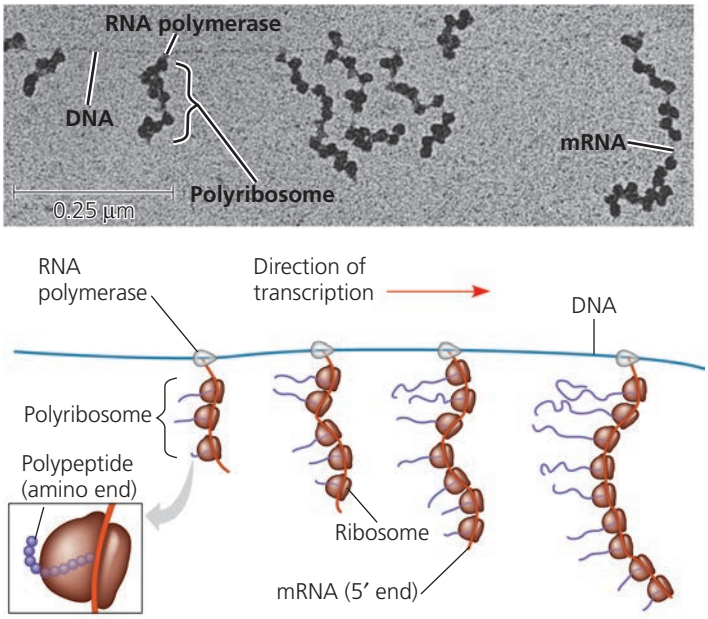


(a) An mRNA molecule is generally translated simultaneously by several ribosomes in clusters called polyribosomes.



(b) This micrograph shows a large polyribosome in a bacterial cell. Growing polypeptides are not visible here (TEM).

Figure 17.24 Coupled transcription and translation in bacteria. In bacterial cells, the translation of mRNA can begin as soon as the leading (5′) end of the mRNA molecule peels away from the DNA template. The micrograph (TEM) shows a strand of *E. coli* DNA being transcribed by RNA polymerase molecules. Attached to each RNA polymerase molecule is a growing strand of mRNA, which is already being translated by ribosomes. The newly synthesized polypeptides are not visible in the micrograph but are shown in the diagram.



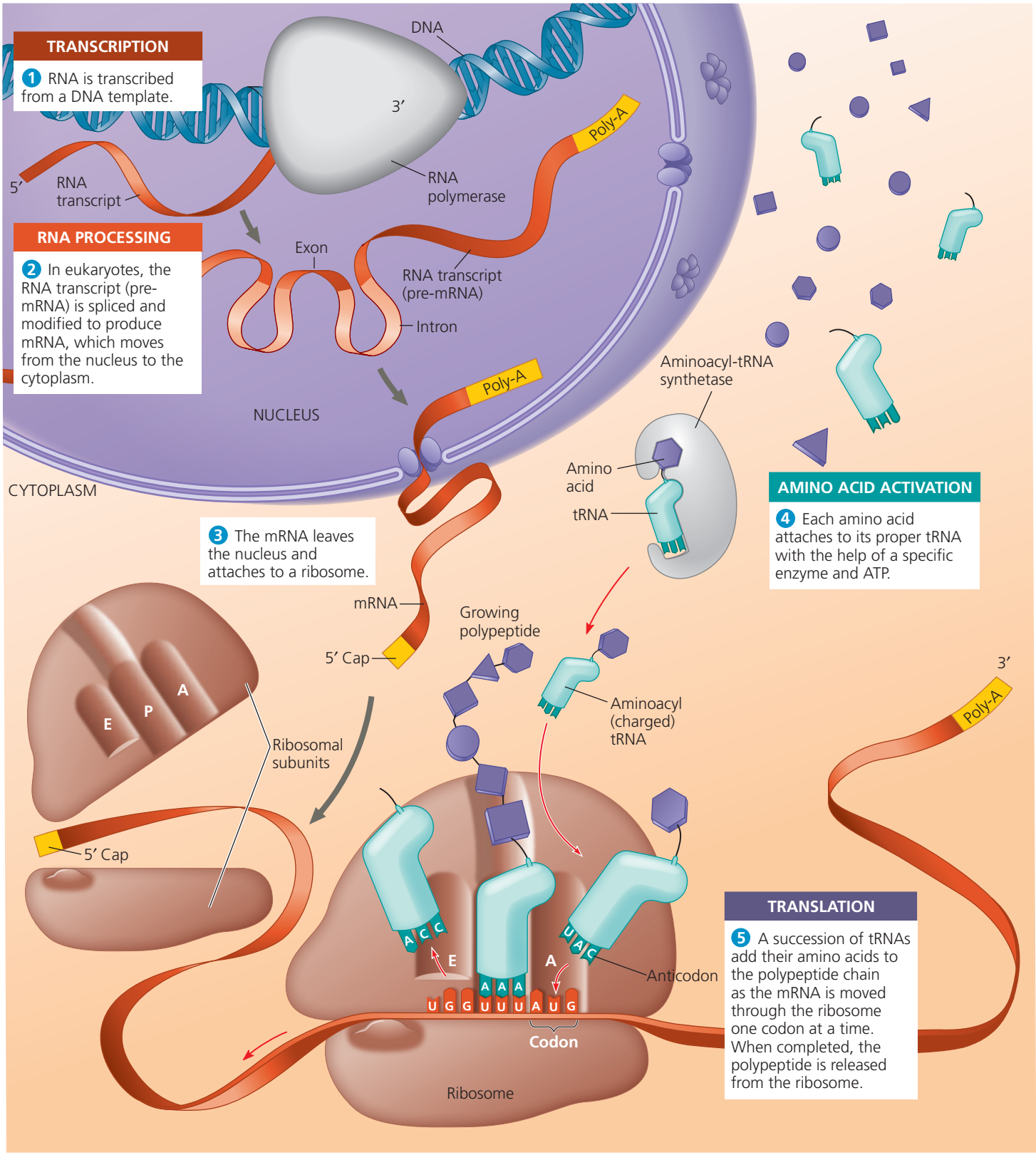
VISUAL SKILLS ➤ *Which one of the mRNA molecules started being transcribed first? On that mRNA, which ribosome started translating the mRNA first?*

CONCEPT CHECK 17.4

1. What two processes ensure that the correct amino acid is added to a growing polypeptide chain?
2. Discuss the different post-translational changes that may be needed to make a functional protein.
3. **WHAT IF? DRAW IT** ➤ Draw a tRNA with the anticodon 3′-CGU-5′. What two different codons could it bind to? Draw each codon on an mRNA, labeling all 5′ and 3′ ends, the tRNA, and the amino acid it carries.
4. **WHAT IF?** ➤ In eukaryotic cells, mRNAs have been found to have a circular arrangement in which proteins hold the poly-A tail near the 5′ cap. How might this increase translation efficiency?

For suggested answers, see Appendix A.

Figure 17.25 A summary of transcription and translation in a eukaryotic cell. This diagram shows the path from one gene to one polypeptide. Each gene in the DNA can be transcribed repeatedly into many identical RNA molecules and each mRNA can be translated repeatedly to yield many identical polypeptide molecules. (Also, remember that the final products of some genes are not polypeptides but RNA molecules that don't get translated, including tRNA and rRNA.) In general, the steps of transcription and translation are similar in bacterial, archaeal, and eukaryotic cells. The major difference is the occurrence of RNA processing in the eukaryotic nucleus. Other significant differences are found in the initiation stages of both transcription and translation and in the termination of transcription. To visualize these processes in their cellular context, see Figure 7.32.



BioFlix® Animation: Protein Synthesis

CONCEPT 17.5

Mutations of one or a few nucleotides can affect protein structure and function

Now that you have explored the process of gene expression, you are ready to understand the effects of changes to the genetic information of a cell. These changes, called **mutations**, are responsible for the huge diversity of genes found among organisms because mutations are the ultimate source of new genes. Earlier, we considered chromosomal rearrangements that affect long segments of DNA (see Figure 15.14); these are considered large-scale mutations. Here we examine small-scale mutations of one or a few nucleotide pairs, including **point mutations**, changes in a single nucleotide pair of a gene.

If a point mutation occurs in a gamete or in a cell that gives rise to gametes, it may be transmitted to offspring and to future generations. If the mutation has an adverse effect on the phenotype of a person, the mutant condition is referred to as a genetic disorder or hereditary disease. For example, we can trace the genetic basis of sickle-cell disease to the mutation of a single nucleotide pair in the gene that encodes the β-globin polypeptide of hemoglobin. The change of a single nucleotide in the DNA's template strand leads to an altered mRNA and the production of an abnormal protein **(Figure 17.26**; also see Figure 5.19). In individuals who are homozygous for the mutant allele, the sickling of red blood cells caused by the altered hemoglobin produces the multiple symptoms associated with sickle-cell disease (see Concept 14.4 and Figure 23.18). Another disorder caused by a point mutation is a heart condition called familial cardiomyopathy, which is responsible for some of the tragic incidents of sudden death in young athletes. Point mutations in several genes encoding muscle proteins have been identified, any of which can lead to this disorder.

Types of Small-Scale Mutations

Let's now consider how small-scale mutations affect proteins. We should first note that many mutations occur in regions outside of protein-coding genes, and any potential effect they have on the phenotype of the organism may be subtle and hard to detect. For this reason, here we'll concentrate on mutations within protein-coding genes. Small-scale mutations within a gene can be divided into two general categories: (1) single nucleotide-pair substitutions and (2) nucleotide-pair insertions or deletions. Insertions and deletions can involve one or more nucleotide pairs.

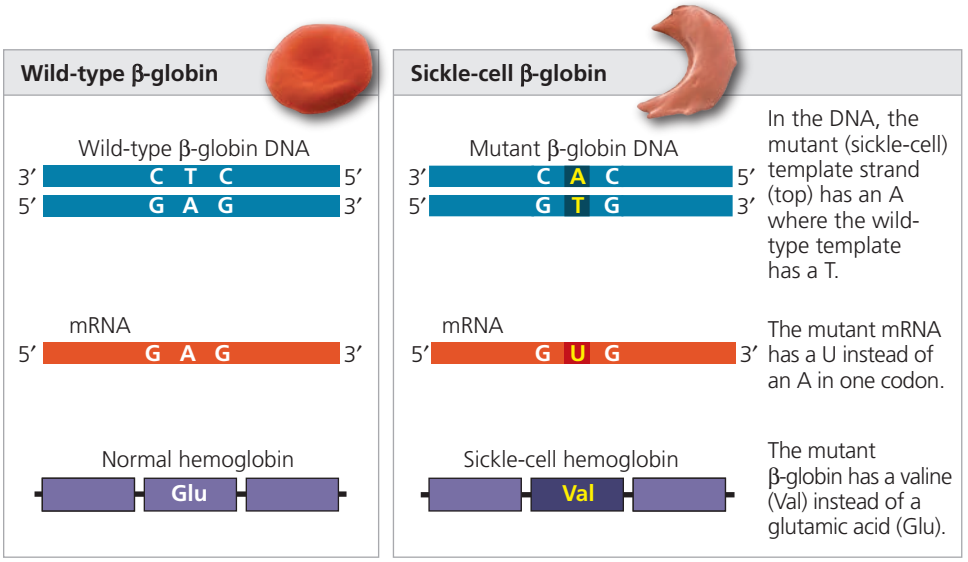
Substitutions

A **nucleotide-pair substitution** is the replacement of one nucleotide and its partner with another pair of nucleotides **(Figure 17.27a)**. Some substitutions have no effect on the encoded protein, owing to the redundancy of the genetic code. For example, if 3′-CCG-5′ on the template strand mutated to 3′-CCA-5′, the mRNA codon that used to be GGC would become GGU, but a glycine would still be inserted at the proper location in the protein (see Figure 17.6). In other words, a change in a nucleotide pair may transform one codon into another that is translated into the same amino acid. Such a change is an example of a **silent mutation**, which has no observable effect on the phenotype. (Silent mutations can occur outside genes as well.) Interestingly, there is evidence that some silent mutations may indirectly affect where or at what level the gene gets expressed, even though the actual protein is the same.

Substitutions that change one amino acid to another one are called **missense mutations**. Such a mutation may have little effect on the protein: The new amino acid may have properties similar to those of the amino acid it replaces, or it may be in a region of the protein where the exact sequence of amino acids is not essential to the protein's function.

However, the nucleotide-pair substitutions of greatest interest are those that cause a major change in a protein. The alteration of a single amino acid in a crucial area of a protein—such as in the part

Figure 17.26 The molecular basis of sickle-cell disease: a point mutation. The allele that causes sickle-cell disease differs from the wild-type (normal) allele by a single DNA nucleotide pair. The micrographs are SEMs of a normal red blood cell (on the left) and a sickled red blood cell (right) from individuals homozygous for wild-type and mutant alleles, respectively.



HHMI Animation: Sickle-Cell Disease
Animation: Mutation Types

Figure 17.27 Types of small-scale mutations that affect mRNA sequence. All but one of the types shown here also affect the amino acid sequence of the encoded polypeptide.

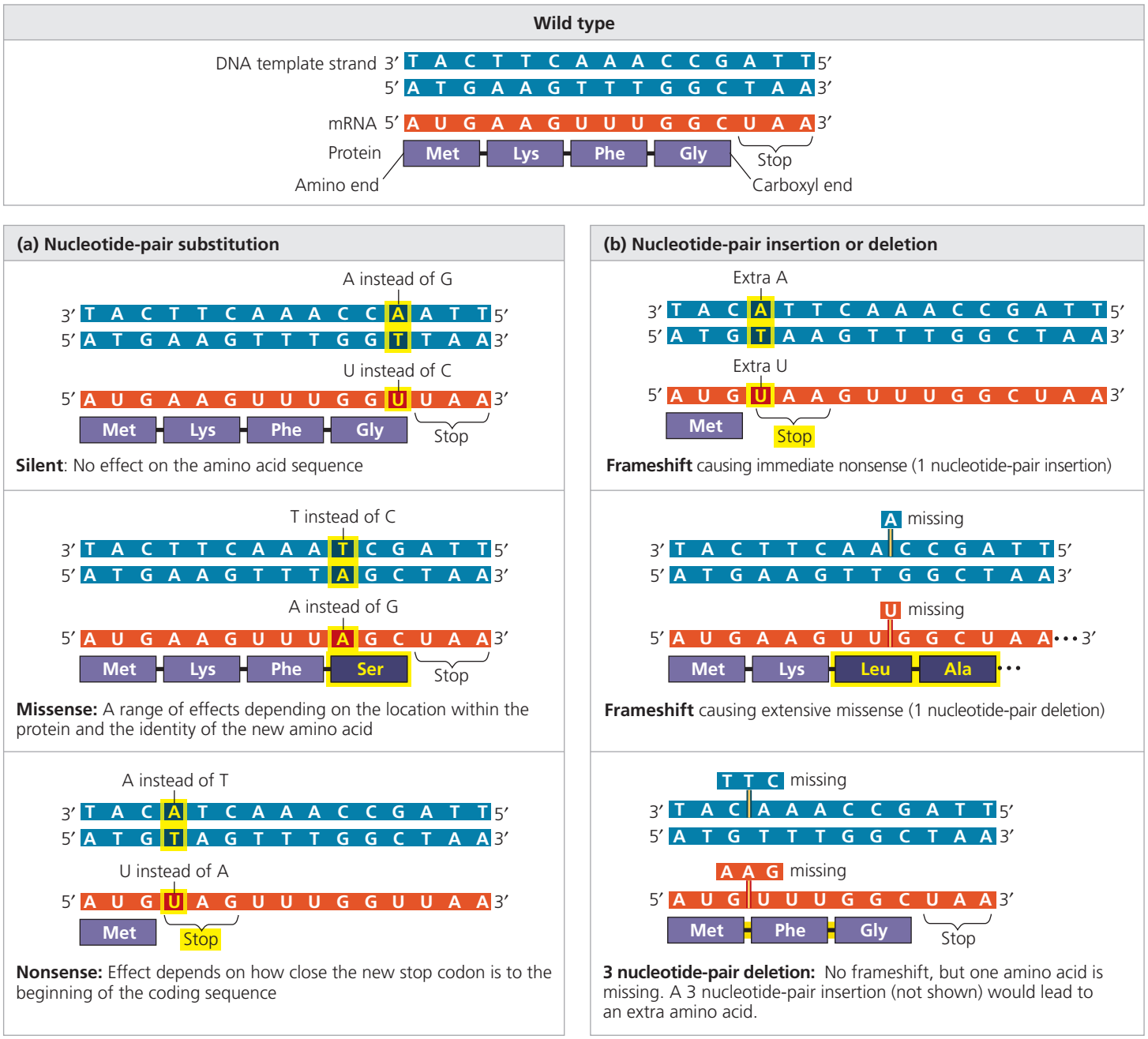


Figure Walkthrough

of the β-globin subunit of hemoglobin shown in Figure 17.26 or in the active site of an enzyme as shown in Figure 6.19—can significantly alter protein activity. Occasionally, such a mutation leads to an improved protein or one with novel capabilities, but much more often such mutations are neutral or detrimental, leading to a useless or less active protein that impairs cellular function.

Substitution mutations are usually missense mutations; that is, the altered codon still codes for an amino acid and thus makes sense, although not necessarily the *right* sense. But a point mutation can also change a codon for an amino acid into a stop codon. This is called a **nonsense mutation**, and it causes translation to be terminated prematurely; the resulting polypeptide will be shorter than the polypeptide encoded by the normal gene. Most nonsense mutations lead to nonfunctional proteins.

In the **Problem-Solving Exercise**, you'll work with a few common single nucleotide-pair substitution mutations in the gene encoding insulin, some or all of which may lead to diabetes. You will classify these mutations into one of

PROBLEM-SOLVING EXERCISE

Are insulin mutations the cause of three infants' neonatal diabetes?

Insulin is a hormone that acts as a key regulator of blood glucose level. In some cases of neonatal diabetes, the gene coding for the insulin protein has a nucleotide-pair substitution mutation that alters the protein structure enough to cause it to malfunction. How can you identify a nucleotide-pair substitution and determine its effect on the amino acid sequence?

Now that it's possible to sequence an individual's whole genome, doctors can use that DNA sequence information to diagnose diseases and identify new treatments. For example, the insulin gene sequence of a patient with neonatal diabetes can be analyzed to determine if it has a mutation and, if so, its effect.

Watch the video in the MasteringBiology Study Area to see how genome sequencing is changing medicine.

MB **ABC News Video: Using Genome Sequencing to Diagnose Gene-Based Diseases**

In this exercise, you will determine the effect of mutations present in a portion of diabetes patients' insulin gene sequences.

Your Approach Suppose you are a medical geneticist presented with three infant patients, all of whom have a nucleotide-pair substitution in their insulin gene. It is your job to analyze each mutation to figure out the effect of the mutation on the amino acid sequence of the insulin protein. To identify the mutation in each patient, you will compare his or her individual insulin complementary DNA (cDNA) sequence to that of the wild-type cDNA. (cDNA is a double-stranded DNA molecule that is based on the mRNA sequence and thus contains only the portion of a gene that is translated—introns are not included. cDNA sequences are commonly used to compare the coding regions of genes.) Identifying the codons that have been changed will tell you which, if any, amino acids are altered in the patient's insulin protein.

Your Data You will analyze the cDNA codons for amino acids 35–54 (of the 110 amino acids) of each patient's insulin protein, so the start codon (AUG) is not present. The sequences of the wild-type cDNA and the patients' cDNA are shown below, arranged in codons.

Wild-type cDNA	5′–CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TTC TTC TAC ACA CCC AAG ACC–3′
Patient 1 cDNA	5′–CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TGC TTC TAC ACA CCC AAG ACC–3′
Patient 2 cDNA	5′–CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TCC TTC TAC ACA CCC AAG ACC–3′
Patient 3 cDNA	5′–CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TTC TTG TAC ACA CCC AAG ACC–3′

Data from N. Nishi and K. Nanjo, Insulin gene mutations and diabetes, *Journal of Diabetes Investigation* 2:92–100 (2011).

Your Analysis

1. Comparing each patient's cDNA sequence to the wild-type cDNA sequence, circle the codons where a nucleotide-pair substitution mutation has occurred.
2. Use a codon table (see Figure 17.6) to identify the amino acid that will be made by the codon with the mutation in each patient's insulin sequence, and compare it to the amino acid made by the codon in the corresponding wild-type sequence. As is standard practice with DNA sequences, the cDNA *coding* (nontemplate) strand has been provided, so to convert it to mRNA for use with the codon table, you just need to change T to U. Classify each patient's nucleotide-pair substitution mutation: Is it a silent, missense, or nonsense mutation? Explain, for each answer.
3. Compare the structure of the amino acid you identified in each patient's insulin sequence to that of the corresponding amino acid in the wild-type insulin sequence (see Figure 5.14). Given that each patient has neonatal diabetes, discuss how the change of amino acid in each might have affected the insulin protein and thus resulted in the disease.

MB **Instructors**: A version of this Problem-Solving Exercise can be assigned in MasteringBiology. Or a more extensive investigation called "Solve It: Which Insulin Mutations May Result in Disease?" can be assigned.

the types we just described and characterize the change in amino acid sequence.

Insertions and Deletions

Insertions and **deletions** are additions or losses of nucleotide pairs in a gene **(Figure 17.27b)**. These mutations have a disastrous effect on the resulting protein more often than substitutions do. Insertion or deletion of nucleotides may alter the reading frame of the genetic message, the triplet grouping of nucleotides on the mRNA that is read during translation. Such a mutation, called a **frameshift mutation**, occurs whenever the number of nucleotides inserted or deleted is not a multiple of three. All nucleotides downstream of the deletion or insertion will be improperly grouped into codons; the result will be

extensive missense mutations, usually ending sooner or later in a nonsense mutation that leads to premature termination. Unless the frameshift is very near the end of the gene, the protein is almost certain to be nonfunctional. Insertions and deletions also occur outside of coding regions; these are not called frameshift mutations, but can have effects on the phenotype—for instance, they can affect how a gene is expressed.

New Mutations and Mutagens

Mutations can arise in a number of ways. Errors during DNA replication or recombination can lead to nucleotide-pair substitutions, insertions, or deletions, as well as to mutations affecting longer stretches of DNA. If an incorrect nucleotide is added to a growing chain during replication, for example, the base on that nucleotide will then be mismatched with the nucleotide base on the other strand. In many cases, the error will be corrected by DNA proofreading and repair systems (see Concept 16.2). Otherwise, the incorrect base will be used as a template in the next round of replication, resulting in a mutation. Such mutations are called *spontaneous mutations*. It is difficult to calculate the rate at which such mutations occur. Rough estimates have been made of the rate of mutation during DNA replication for both *E. coli* and eukaryotes, and the numbers are similar: About one nucleotide in every 10^{10} is altered, and the change is passed on to the next generation of cells.

A number of physical and chemical agents, called **mutagens**, interact with DNA in ways that cause mutations. In the 1920s, Hermann Muller discovered that X-rays caused genetic changes in fruit flies, and he used X-rays to make *Drosophila* mutants for his genetic studies. But he also recognized an alarming implication of his discovery: X-rays and other forms of high-energy radiation pose hazards to the genetic material of people as well as laboratory organisms. Mutagenic radiation, a physical mutagen, includes ultraviolet (UV) light, which can cause disruptive thymine dimers in DNA (see Figure 16.19).

Chemical mutagens fall into several categories. Nucleotide analogs are chemicals similar to normal DNA nucleotides but that pair incorrectly during DNA replication. Other chemical mutagens interfere with correct DNA replication by inserting themselves into the DNA and distorting the double helix. Still other mutagens cause chemical changes in bases that change their pairing properties.

Researchers have developed a variety of methods to test the mutagenic activity of chemicals. A major application of these tests is the preliminary screening of chemicals to identify those that may cause cancer. This approach makes sense because most carcinogens (cancer-causing chemicals) are mutagenic, and conversely, most mutagens are carcinogenic.

What Is a Gene? *Revisiting the Question*

Our definition of a gene has evolved over the past few chapters, as it has through the history of genetics. We began with the Mendelian concept of a gene as a discrete unit of inheritance that affects a phenotypic character (Chapter 14). We saw that Morgan and his colleagues assigned such genes to specific loci on chromosomes (Chapter 15). We went on to view a gene as a region of specific nucleotide sequence along the length of the DNA molecule of a chromosome (Chapter 16). Finally, in this chapter, we have considered a functional definition of a gene as a DNA sequence that codes for a specific polypeptide chain or a functional RNA molecule, such as a tRNA. All these definitions are useful, depending on the context in which genes are being studied.

We have noted that merely saying a gene codes for a polypeptide is an oversimplification. Most eukaryotic genes contain noncoding segments (such as introns), so large portions of these genes have no corresponding segments in polypeptides. Molecular biologists also often include promoters and certain other regulatory regions of DNA within the boundaries of a gene. These DNA sequences are not transcribed, but they can be considered part of the functional gene because they must be present for transcription to occur. Our definition of a gene must also be broad enough to include the DNA that is transcribed into rRNA, tRNA, and other RNAs that are not translated. These genes have no polypeptide products but play crucial roles in the cell. Thus, we arrive at the following definition: *A gene is a region of DNA that can be expressed to produce a final functional product that is either a polypeptide or an RNA molecule*.

When considering phenotypes, however, it is often useful to start by focusing on genes that code for polypeptides. In this chapter, you have learned in molecular terms how a typical gene is expressed—by transcription into RNA and then translation into a polypeptide that forms a protein of specific structure and function. Proteins, in turn, bring about an organism's observable phenotype.

A given type of cell expresses only a subset of its genes. This is an essential feature in multicellular organisms: You'd be in trouble if the lens cells in your eyes started expressing the genes for hair proteins, which are normally expressed only in hair follicle cells! Gene expression is precisely regulated, which we'll explore in the next chapter, beginning with the simpler case of bacteria and continuing with eukaryotes.

CONCEPT CHECK 17.5

1. What happens when one nucleotide pair is lost from the middle of the coding sequence of a gene?
2. **MAKE CONNECTIONS ➤** Individuals heterozygous for the sickle-cell allele are generally healthy but show phenotypic effects of the allele under some circumstances (see Figure 14.17). Explain in terms of gene expression.
3. **WHAT IF? DRAW IT ➤** The template strand of a gene includes this sequence:
   3′-TACTTGTCCGATATC-5′. It is mutated to
   3′-TACTTGTCCAATATC-5′. For both wild-type and mutant sequences, draw the double-stranded DNA, the resulting mRNA, and the amino acid sequence each encodes. What is the effect of the mutation on the amino acid sequence?

For suggested answers, see Appendix A.

17 Chapter Review

Go to **MasteringBiology**™ for Videos, Animations, Vocab Self-Quiz, Practice Tests, and more in the Study Area.

SUMMARY OF KEY CONCEPTS

CONCEPT 17.1

VOCAB SELF-QUIZ goo.gl/Rn5Uax

Genes specify proteins via transcription and translation *(pp. 386–392)*

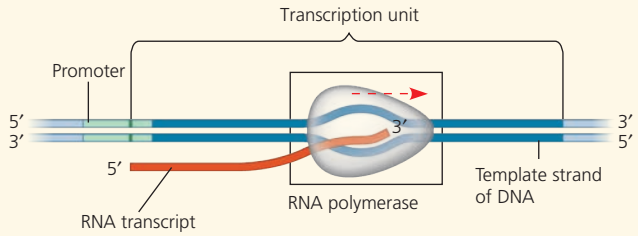
- Beadle and Tatum's studies of mutant strains of *Neurospora* led to the one gene–one polypeptide hypothesis. During **gene expression**, the information encoded in genes is used to make specific polypeptide chains (enzymes and other proteins) or RNA molecules.
- **Transcription** is the synthesis of RNA complementary to a **template strand** of DNA. **Translation** is the synthesis of a polypeptide whose amino acid sequence is specified by the nucleotide sequence in **messenger RNA (mRNA)**.
- Genetic information is encoded as a sequence of nonoverlapping nucleotide triplets, or **codons**. A codon in mRNA either is translated into an amino acid (61 of the 64 codons) or serves as a stop signal (3 codons). Codons must be read in the correct **reading frame**.

? *Describe the process of gene expression, by which a gene affects the phenotype of an organism.*

CONCEPT 17.2

Transcription is the DNA-directed synthesis of RNA: *a closer look* *(pp. 392–394)*

- RNA synthesis is catalyzed by **RNA polymerase**, which links together RNA nucleotides complementary to a DNA template strand. This process follows the same base-pairing rules as DNA replication, except that in RNA, uracil substitutes for thymine.



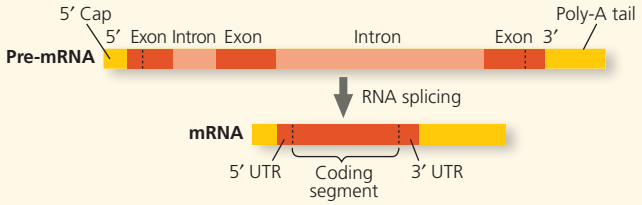
- The three stages of transcription are initiation, elongation, and termination. A **promoter**, often including a **TATA box** in eukaryotes, establishes where RNA synthesis is initiated. **Transcription factors** help eukaryotic RNA polymerase recognize promoter sequences, forming a **transcription initiation complex**. Termination differs in bacteria and eukaryotes.

? *What are the similarities and differences in the initiation of gene transcription in bacteria and eukaryotes?*

CONCEPT 17.3

Eukaryotic cells modify RNA after transcription *(pp. 395–397)*

- Eukaryotic mRNAs undergo **RNA processing**, which includes RNA splicing, the addition of a modified nucleotide **5′ cap** to the 5′ end, and the addition of a **poly-A tail** to the 3′ end. The processed mRNA includes an untranslated region (5′ UTR or 3′ UTR) at each end of the coding segment.
- Most eukaryotic genes are split into segments: They have **introns** interspersed among the **exons** (the regions included in the mRNA). In **RNA splicing**, introns are removed and exons joined. RNA splicing is typically carried out by **spliceosomes**, but in some cases, RNA alone catalyzes its own splicing. The catalytic ability of some RNA molecules, called **ribozymes**, derives from the inherent properties of RNA. The presence of introns allows for **alternative RNA splicing**.

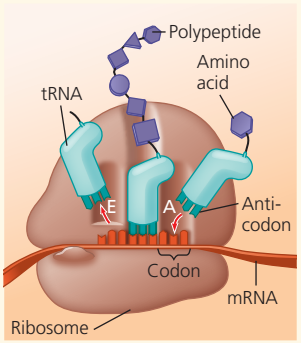


? *What function do the 5′ cap and the poly-A tail serve on a eukaryotic mRNA?*

CONCEPT 17.4

Translation is the RNA-directed synthesis of a polypeptide: *a closer look* *(pp. 397–406)*

- A cell translates an mRNA message into protein using **transfer RNAs (tRNAs)**. After being bound to a specific amino acid by an **aminoacyl-tRNA synthetase**, a tRNA lines up via its **anticodon** at the complementary codon on mRNA. A **ribosome**, made up of **ribosomal RNAs (rRNAs)** and proteins, facilitates this coupling with binding sites for mRNA and tRNA.
- Ribosomes coordinate the three stages of translation: initiation, elongation, and termination. The formation of peptide bonds between amino acids is catalyzed by rRNAs as tRNAs move through the **A** and **P sites** and exit through the **E site**.



- After translation, during protein processing, proteins may be modified by cleavage or by attachment of sugars, lipids, phosphates, or other chemical groups.
- Free ribosomes in the cytosol initiate synthesis of all proteins, but proteins with a **signal peptide** are synthesized on the ER.
- A gene can be transcribed by multiple RNA polymerases simultaneously. Also, a single mRNA molecule can be translated simultaneously by a number of ribosomes, forming a **polyribosome**. In bacteria, these processes are coupled, but in eukaryotes they are separated in space and time by the nuclear membrane.

? *Describe how tRNAs function in the context of the ribosome in building a polypeptide.*