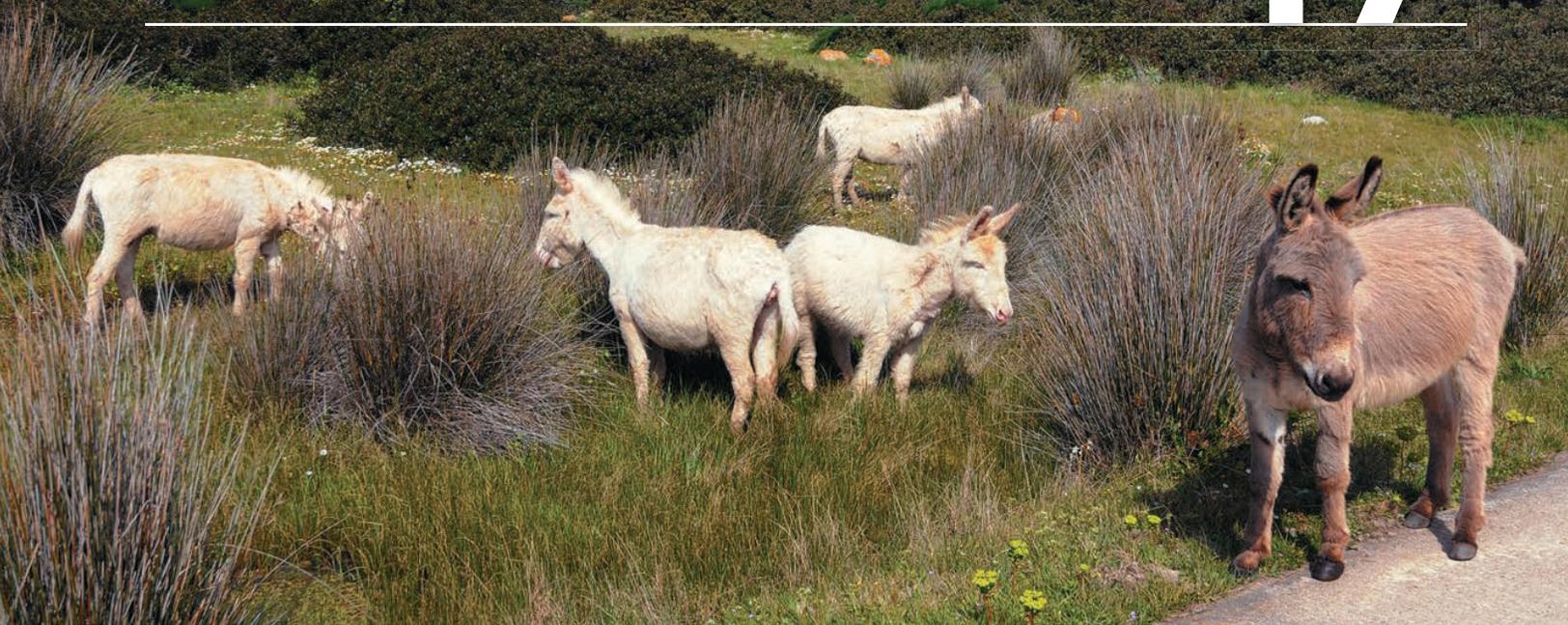


Gene Expression: From Gene to Protein

17



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

KEY CONCEPTS

- 17.1** Genes specify proteins via transcription and translation
- 17.2** Transcription is the DNA-directed synthesis of RNA: *a closer look*
- 17.3** Eukaryotic cells modify RNA after transcription
- 17.4** Translation is the RNA-directed synthesis of a polypeptide: *a closer look*
- 17.5** Mutations of one or a few nucleotides can affect protein structure and function

▼ **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 8.1). 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.

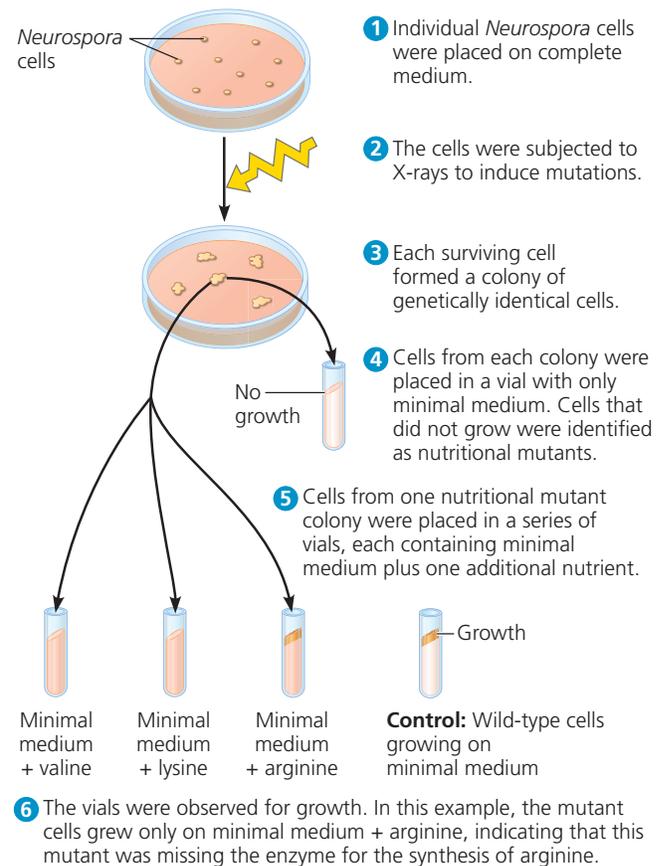
Nutritional Mutants in *Neurospora*: Scientific Inquiry

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.



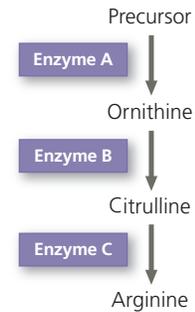
minimal nutrients for growth—inorganic salts, glucose, and the vitamin biotin—incorporated into agar, a support medium. From this so-called *minimal medium*, wild-type mold cells use their metabolic pathways to produce all the other molecules they need for growth, dividing repeatedly and forming visible 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

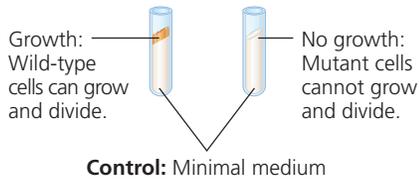
▼ **Figure 17.3**

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.)



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 <i>Neurospora crassa</i>			
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	Precursor → Ornithine	Precursor → Ornithine	Precursor → Ornithine	Precursor → Ornithine
Gene B → Enzyme B	Ornithine → Citrulline	Ornithine → Citrulline	Ornithine → Citrulline	Ornithine → Citrulline
Gene C → Enzyme C	Citrulline → Arginine	Citrulline → Arginine	Citrulline → Arginine	Citrulline → Arginine

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).

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?

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:

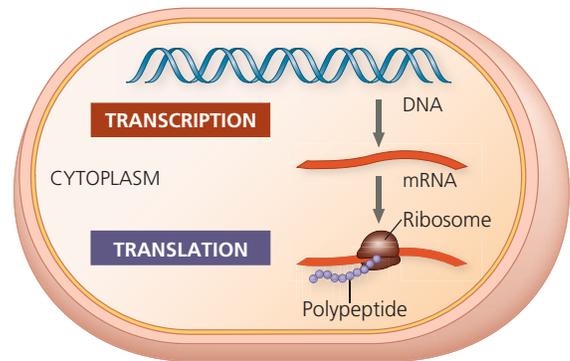


This concept was dubbed the *central dogma* by Francis Crick in 1956. But in the 1970s, scientists were surprised to discover some enzymes that use RNA molecules as templates for DNA synthesis (which we'll cover in Concept 19.2). However, these exceptions do not invalidate the idea that, in general, 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.

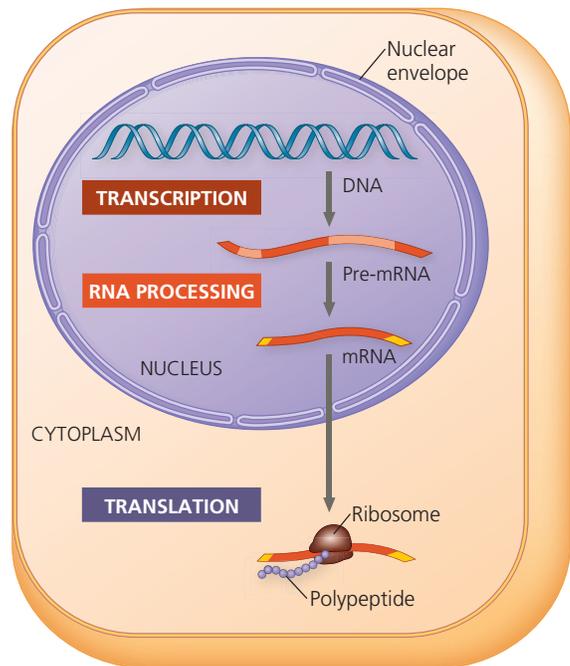
The Genetic Code

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.

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

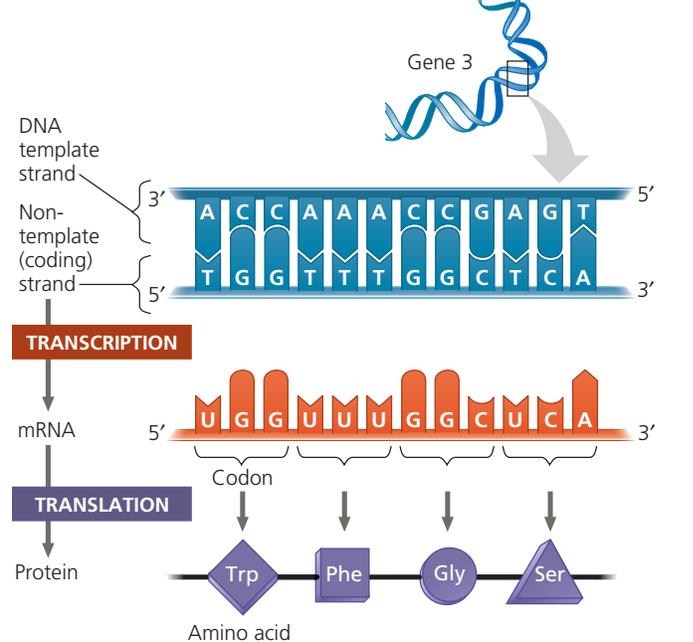
If each kind of nucleotide base were translated into an amino acid, only four amino acids could be specified, one per nucleotide base. Would a language of two-letter code words suffice? The two-nucleotide sequence AG, for example, could specify one amino acid, and GT could specify another. Since there 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

▼ **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

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.

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

Y 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.

		Second mRNA base					
		U	C	A	G		
U	UUU	Phe (F)	UCU UCC UCA UCG Ser (S)	UAU	Tyr (Y)	UGU	Cys (C)
	UUC	Leu (L)		UAA	Stop	UGA	Stop
	UUA		UAG	Stop	UGG	Trp (W)	
	UUG						
C	CUU	Leu (L)	CCU	CAU	His (H)	CGU	Arg (R)
	CUC		CCC	Pro (P)	CAC	Gln (Q)	CGC
	CUA	CCA	CAG				
	CUG	CCG					
A	AUU	Ile (I)	ACU	AAU	Asn (N)	AGU	Ser (S)
	AUC		ACC	Thr (T)	AAC	Lys (K)	AGA
	AUA	ACA	AAG				
	AUG	ACG	Met (M) or start				
G	GUU	Val (V)	GCU	GAU	Asp (D)	GGU	Gly (G)
	GUC		GCC	Ala (A)	GAC	Glu (E)	GGA
	GUA	GCA	GAG				
	GUG	GCG					

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.

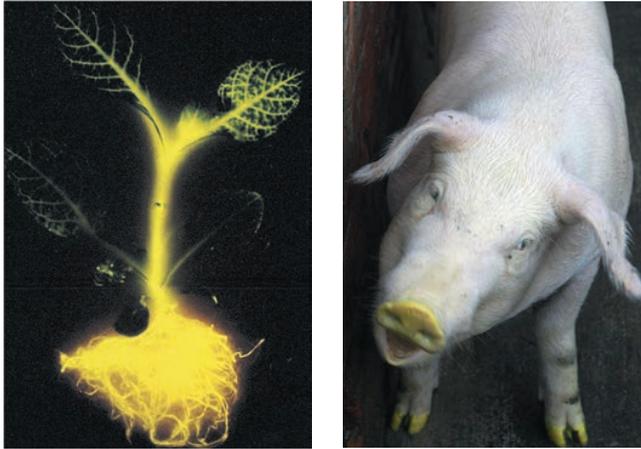
Animation: Translation: The Genetic Code

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.

 **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 20.4).

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.** What polypeptide product would you expect from a poly-G mRNA that is 30 nucleotides long?
- 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.

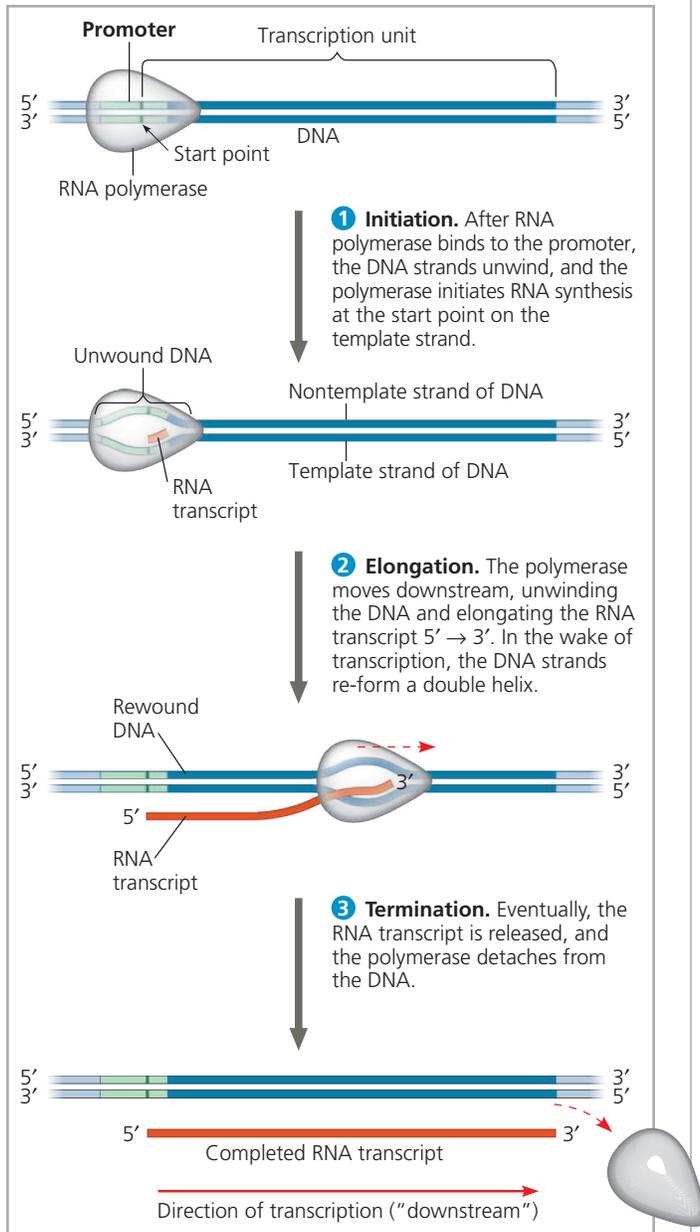
Synthesis of an RNA Transcript

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

▼ 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.

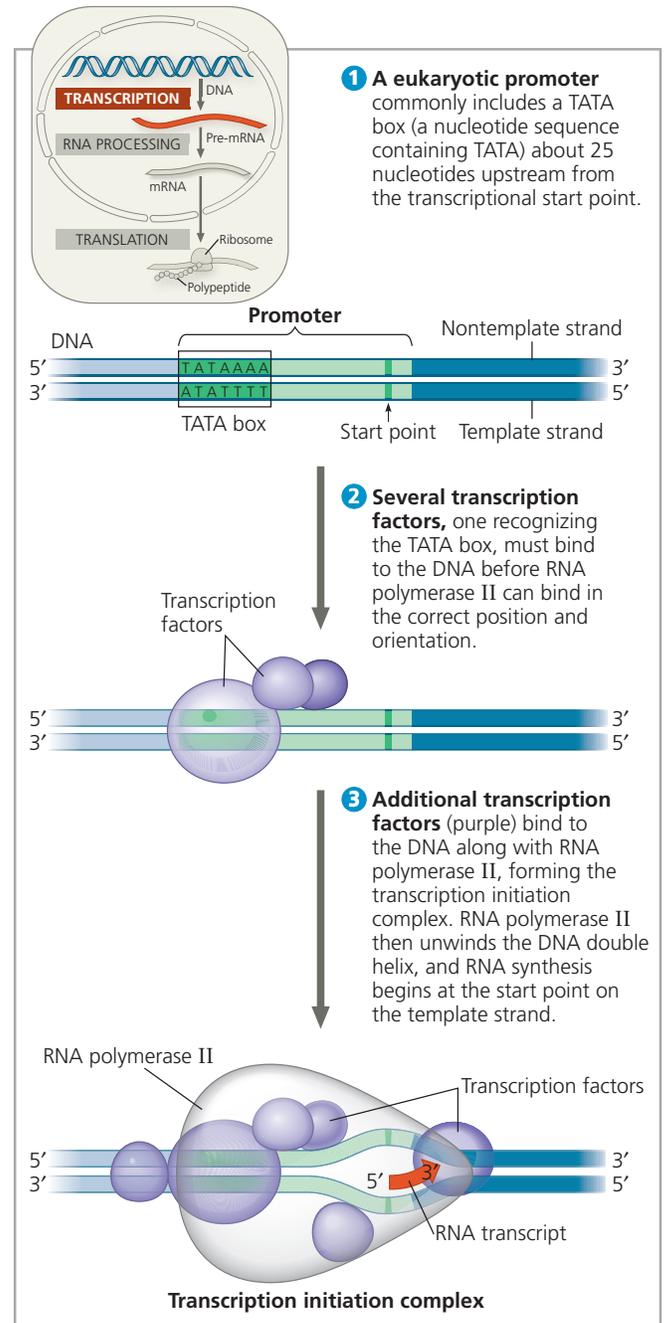


MAKE CONNECTIONS ▶ Compare the use of a template strand during transcription and replication. See Figure 16.17.

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

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 17.9 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

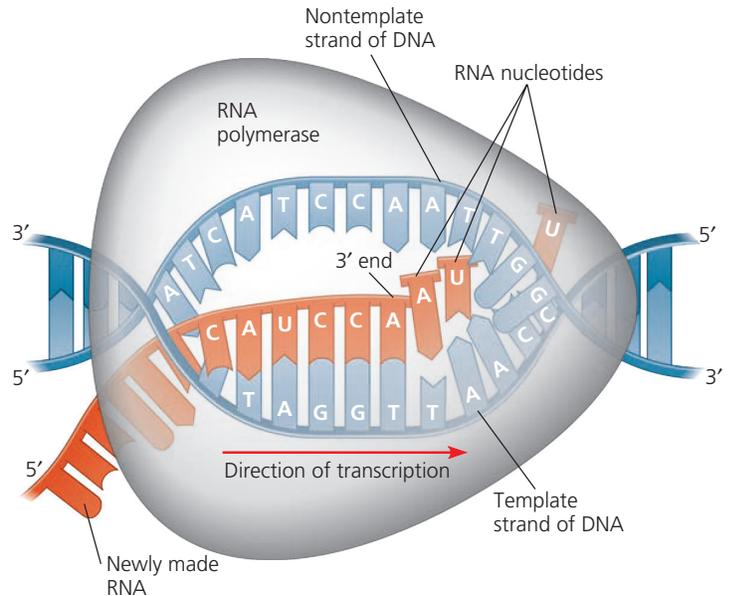
As RNA polymerase moves along the DNA, it untwists the double helix, exposing about 10–20 DNA nucleotides at a time for pairing with RNA nucleotides (**Figure 17.10**). The enzyme adds nucleotides to the 3' end of the growing RNA molecule as it continues along the double helix. In the wake of this advancing 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

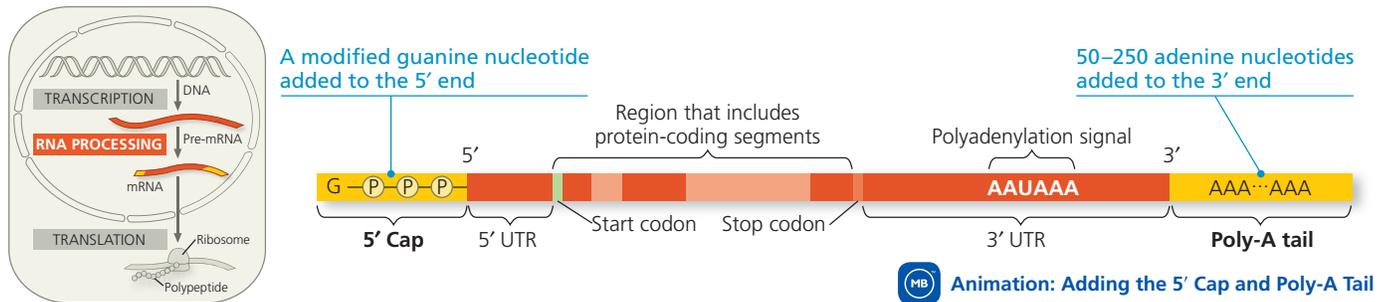
1. What is a promoter? Is it located at the upstream or downstream end of a transcription unit?
2. What enables RNA polymerase to start transcribing a gene at the right place on the DNA in a bacterial cell? In a eukaryotic cell?
3. **WHAT IF? >** Suppose X-rays caused a sequence change in the TATA box 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.

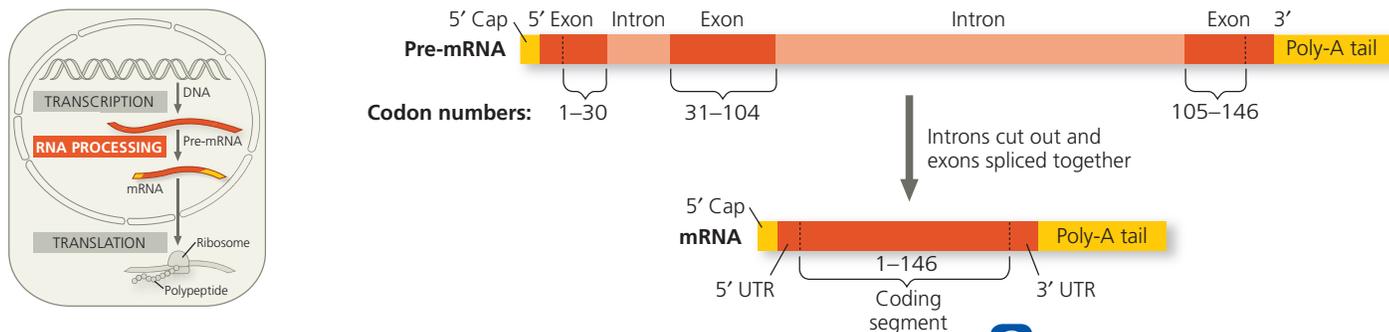
Split Genes and RNA Splicing

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 *intervening sequences*, or **introns**. The other regions are called **exons**, because they are eventually expressed, 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 exit 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.

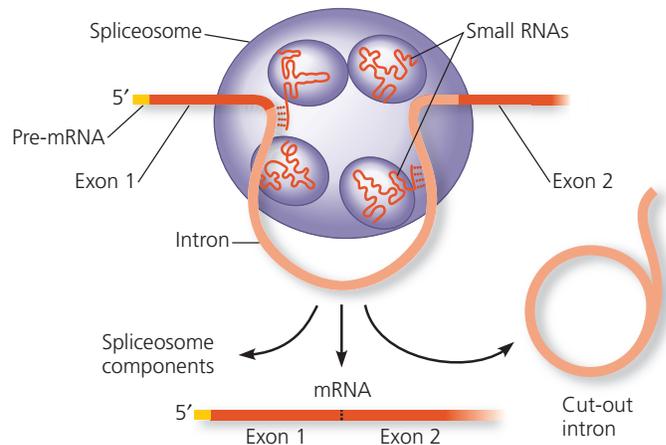
How is pre-mRNA splicing carried out? The removal of introns is accomplished by a large complex made of proteins and small RNAs called a **spliceosome**. This complex binds to several short nucleotide sequences along an intron, including key sequences at each end (**Figure 17.13**). The intron is then 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.

 **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.



 **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

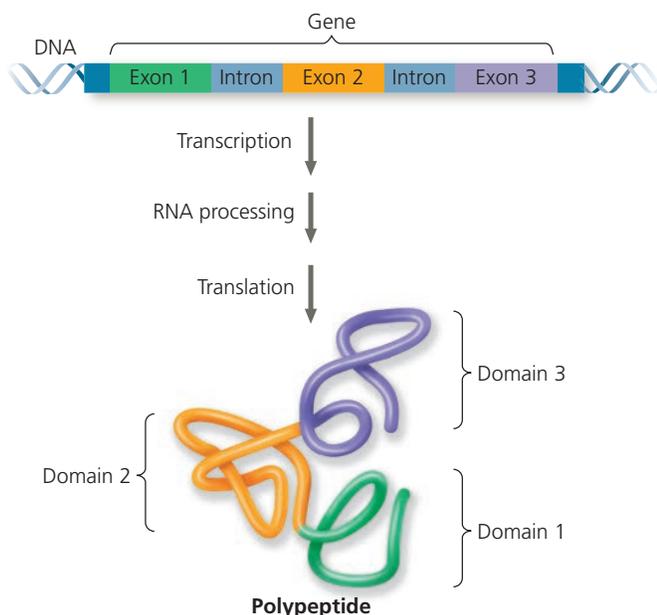
EVOLUTION Whether or not RNA splicing and the presence of introns have provided selective advantages during evolutionary history is a matter of some debate. In any case, it is informative to consider their possible adaptive benefits. Specific functions have not been identified for most introns, but at 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 21.1) 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 21.16). Introns increase the probability of crossing over between the exons of alleles of a gene—simply by providing more terrain for cross-overs 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?

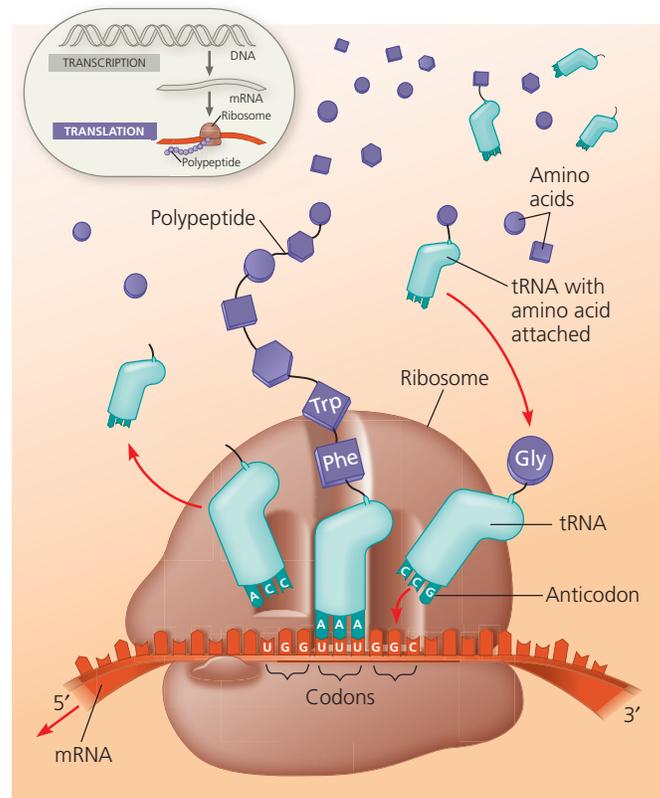
For suggested answers, see Appendix A.

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.



MB 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.

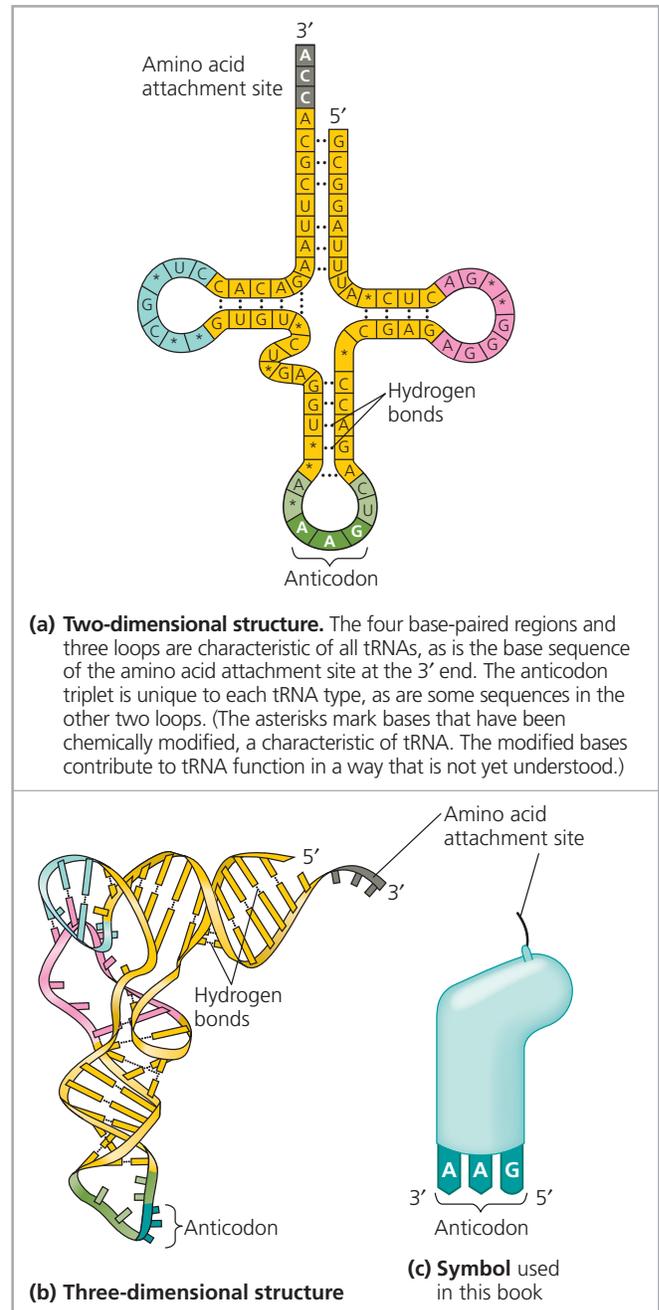
The Structure and Function of Transfer RNA

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 17.16 The structure of transfer RNA (tRNA).



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** 

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