CHAPTER **SEVEN** 7

From DNA to Protein: How Cells Read the Genome

Once the double-helical structure of DNA (deoxyribonucleic acid) had been determined in the early 1950s, it became clear that the hereditary information in cells is encoded in the linear order—or *sequence*—of the four different nucleotide subunits that make up the DNA. We saw in Chapter 6 how this information can be passed on unchanged from a cell to its descendants through the process of DNA replication. But how does the cell decode and use the information? How do genetic instructions written in an alphabet of just four "letters" direct the formation of a bacterium, a fruit fly, or a human? We still have a lot to learn about how the information stored in an organism's genes produces even the simplest unicellular bacterium, let alone how it directs the development of complex multicellular organisms like ourselves. But the DNA code itself has been deciphered, and we have come a long way in understanding how cells read it.

Even before the DNA code was broken, it was known that the information contained in genes somehow directed the synthesis of proteins. Proteins are the principal constituents of cells and determine not only cell structure but also cell function. In previous chapters, we encountered some of the thousands of different kinds of proteins that cells can make. We saw in Chapter 4 that the properties and function of a protein molecule are determined by the sequence of the 20 different amino acid subunits in its polypeptide chain: each type of protein has its own unique amino acid sequence, which dictates how the chain will fold to form a molecule with a distinctive shape and chemistry. The genetic instructions carried by DNA must therefore specify the amino acid sequences of proteins. We will see in this chapter exactly how this is done.

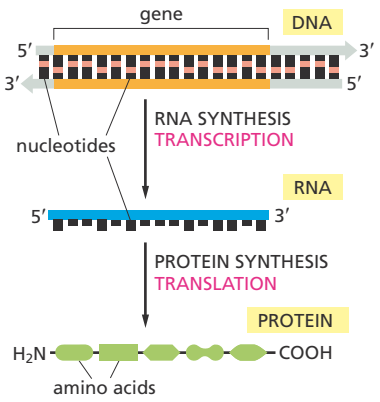


Figure 7–1 Genetic information directs the synthesis of proteins. The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells. It was Francis Crick who dubbed this flow of information "the central dogma." The segments of DNA that are transcribed into RNA are called genes.

DNA does not synthesize proteins itself, but it acts like a manager, delegating the various tasks to a team of workers. When a particular protein is needed by the cell, the nucleotide sequence of the appropriate segment of a DNA molecule is first copied into another type of nucleic acid—*RNA* (*ribonucleic acid*). That segment of DNA is called a **gene**, and the resulting RNA copies are then used to direct the synthesis of the protein. Many thousands of these conversions from DNA to protein occur every second in each cell in our body. The flow of genetic information in cells is therefore from DNA to RNA to protein (**Figure 7–1**). All cells, from bacteria to humans, express their genetic information in this way—a principle so fundamental that it has been termed the *central dogma* of molecular biology.

In this chapter, we explain the mechanisms by which cells copy DNA into RNA (a process called *transcription*) and then use the information in RNA to make protein (a process called *translation*). We also discuss a few of the key variations on this basic scheme. Principal among these is *RNA splicing*, a process in eukaryotic cells in which segments of an *RNA transcript* are removed—and the remaining segments stitched back together—before the RNA is translated into protein. In the final section, we consider how the present scheme of information storage, transcription, and translation might have arisen from much simpler systems in the earliest stages of cell evolution.

FROM DNA TO RNA

Transcription and translation are the means by which cells read out, or *express*, the instructions in their *genes*. Many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules. This successive amplification enables cells to rapidly synthesize large amounts of protein whenever necessary. At the same time, each gene can be transcribed, and its RNA translated, at different rates, providing the cell with a way to make vast quantities of some proteins and tiny quantities of others (**Figure 7–2**). Moreover, as we discuss in Chapter 8, a cell can change (or regulate) the expression of each of its genes according to the needs of the moment. In this section, we discuss the production of RNA, the first step in *gene expression*.

QUESTION 7–1

Consider the expression "central dogma," which refers to the flow of genetic information from DNA to RNA to protein. Is the word "dogma" appropriate in this context?

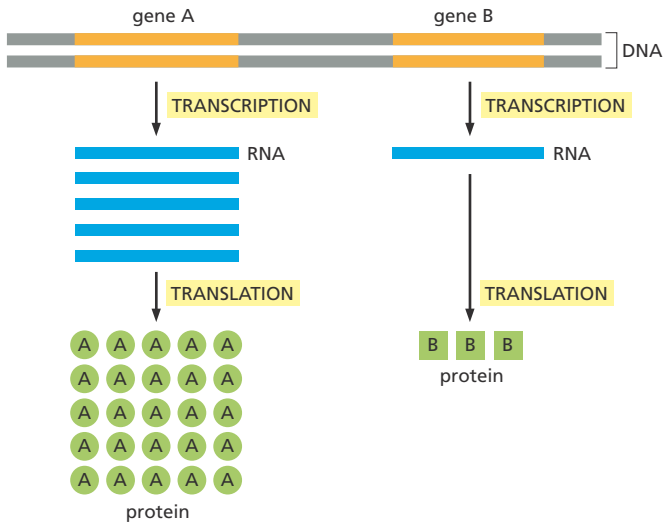


Figure 7–2 A cell can express different genes at different rates. In this and later figures, the untranscribed portions of the DNA are shown in *gray*.

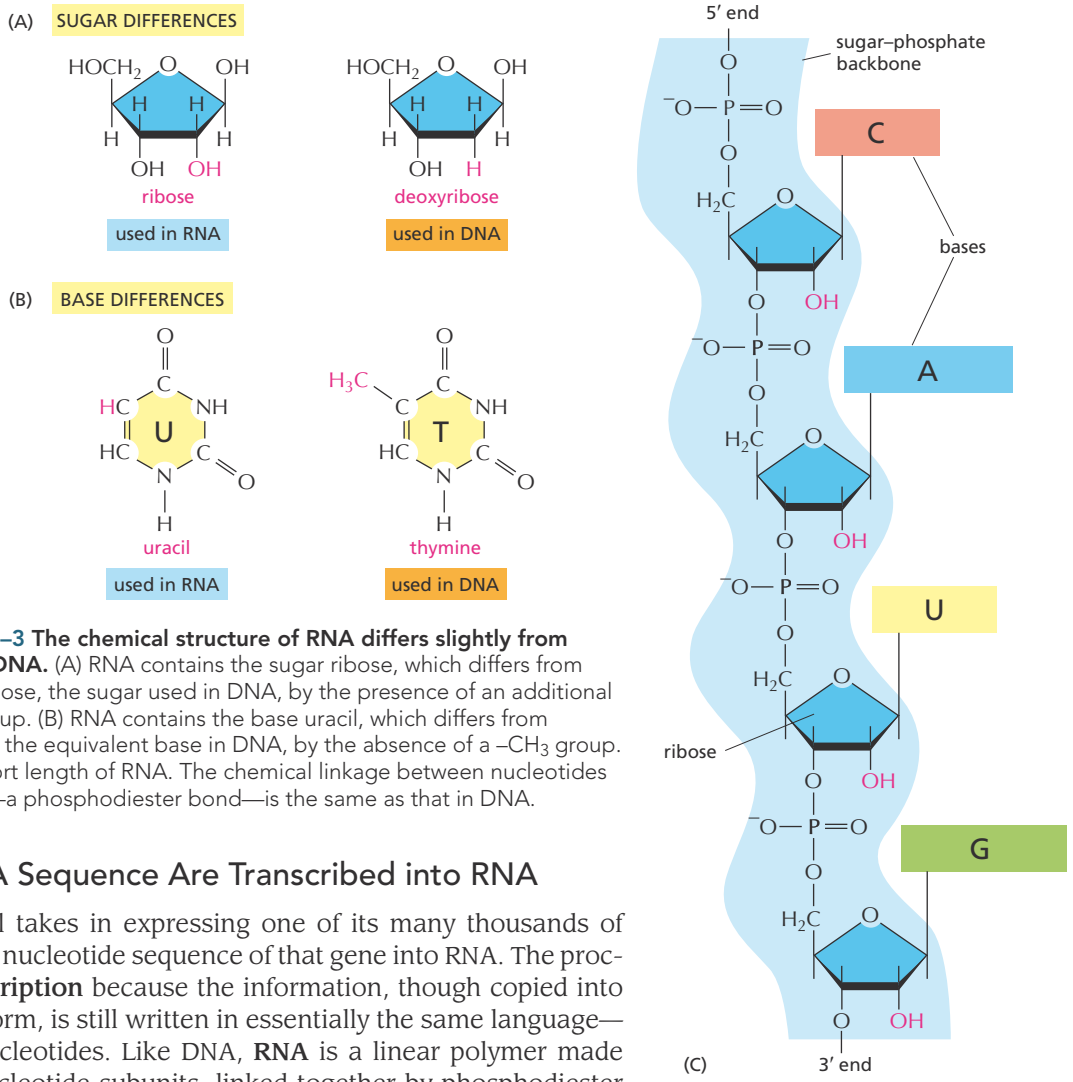


Figure 7–3 The chemical structure of RNA differs slightly from that of DNA. (A) RNA contains the sugar ribose, which differs from deoxyribose, the sugar used in DNA, by the presence of an additional –OH group. (B) RNA contains the base uracil, which differs from thymine, the equivalent base in DNA, by the absence of a $–CH_3$ group. (C) A short length of RNA. The chemical linkage between nucleotides in RNA—a phosphodiester bond—is the same as that in DNA.

Portions of DNA Sequence Are Transcribed into RNA

The first step a cell takes in expressing one of its many thousands of genes is to copy the nucleotide sequence of that gene into RNA. The process is called **transcription** because the information, though copied into another chemical form, is still written in essentially the same language—the language of nucleotides. Like DNA, **RNA** is a linear polymer made of four different nucleotide subunits, linked together by phosphodiester bonds. It differs from DNA chemically in two respects: (1) the nucleotides in RNA are *ribonucleotides*—that is, they contain the sugar ribose (hence the name *ribo*nucleic acid) rather than deoxyribose; (2) although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains uracil (U) instead of the thymine (T) found in DNA (**Figure 7–3**). Because U, like T, can base-pair by hydrogen-bonding with A (**Figure 7–4**), the complementary base-pairing properties described for DNA in Chapter 5 apply also to RNA.

Although their chemical differences are small, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. This difference has important functional consequences. Because an RNA chain is single-stranded, it can fold up into a variety of shapes, just as a polypeptide chain folds up to form the final shape of a protein (**Figure 7–5**); double-stranded DNA cannot fold in this fashion. As we discuss later, the ability to fold into a complex three-dimensional shape allows RNA to carry out various functions in cells, in addition to conveying information between DNA and protein. Whereas DNA functions solely as an information store, some RNAs have structural, regulatory, or catalytic roles.

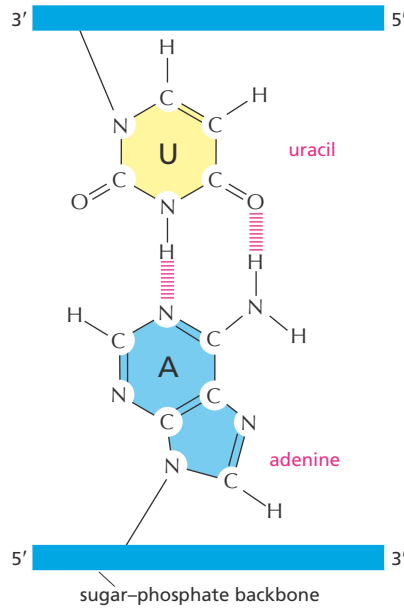


Figure 7–4 Uracil forms a base pair with adenine. The hydrogen bonds that hold the base pair together are shown in *red*. Uracil has the same base-pairing properties as thymine. Thus U-A base pairs in RNA closely resemble T-A base pairs in DNA (see Figure 5–6A).

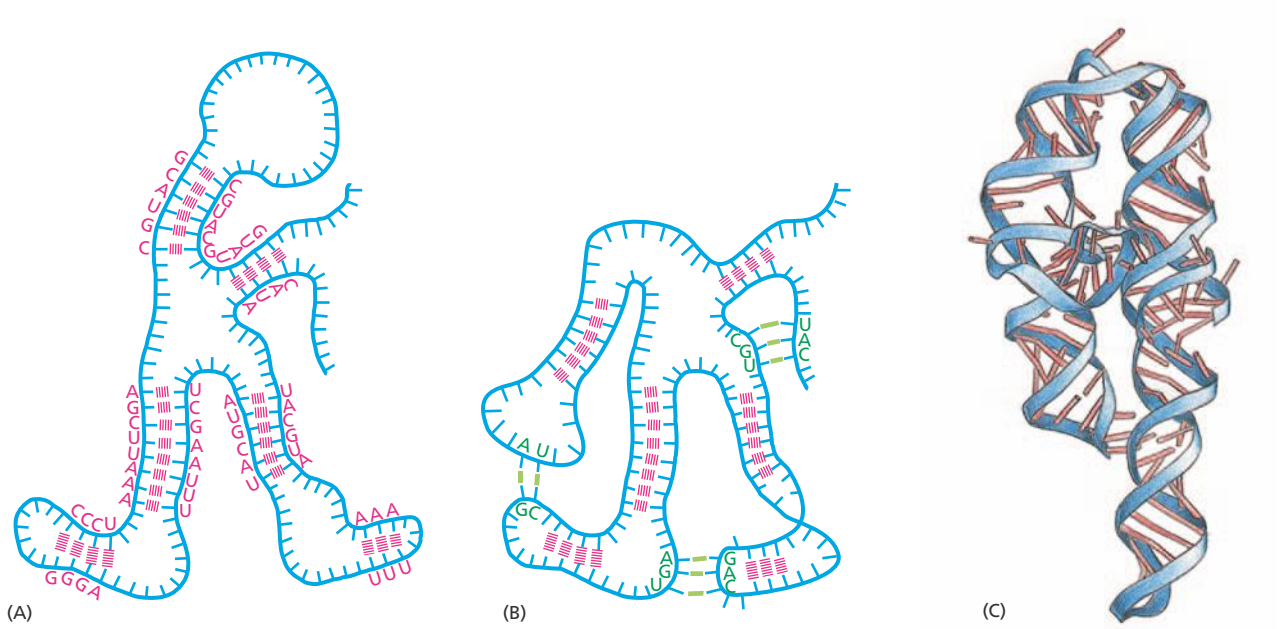


Figure 7–5 RNA molecules can form intramolecular base pairs and fold into specific structures. RNA is single-stranded, but it often contains short stretches of nucleotides that can base-pair with complementary sequences found elsewhere on the same molecule. These interactions—along with some "nonconventional base-pair interactions (e.g., A-G)—allow an RNA molecule to fold into a three-dimensional structure that is determined by its sequence of nucleotides. (A) A diagram of a hypothetical, folded RNA structure showing only conventional (G-C and A-U) base-pair interactions. (B) Incorporating nonconventional base-pair interactions (*green*) changes the structure of the hypothetical RNA shown in (A). (C) Structure of an actual RNA molecule that is involved in RNA splicing. This RNA contains a considerable amount of double-helical structure. The sugar–phosphate backbone is *blue* and the bases are *red*; the conventional base-pair interactions are indicated by red "rungs" that are continuous, and nonconventional base pairs are indicated by broken red rungs. For an additional view of RNA structure, see **Movie 7.1**.

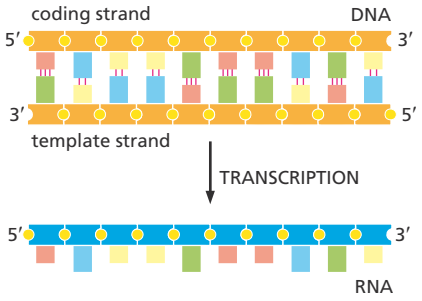


Figure 7–6 Transcription of a gene produces an RNA complementary to one strand of DNA. The transcribed strand of the gene, the *bottom* strand in this example, is called the template strand. The nontemplate strand of the gene (here, shown at the *top*) is sometimes called the *coding strand* because its sequence is equivalent to the RNA product, as shown. Which DNA strand serves as the template varies, depending on the gene, as we discuss later. By convention, an RNA molecule is always depicted with its 5′ end—the first part to be synthesized—to the left.

Transcription Produces RNA That Is Complementary to One Strand of DNA

All the RNA in a cell is made by transcription, a process that has certain similarities to DNA replication (discussed in Chapter 6). Transcription begins with the opening and unwinding of a small portion of the DNA double helix to expose the bases on each DNA strand. One of the two strands of the DNA double helix then acts as a template for the synthesis of RNA. Ribonucleotides are added, one by one, to the growing RNA chain; as in DNA replication, the nucleotide sequence of the RNA chain is determined by complementary base-pairing with the DNA template. When a good match is made, the incoming ribonucleotide is covalently linked to the growing RNA chain by the enzyme *RNA polymerase*. The RNA chain produced by transcription—the **RNA transcript**—is therefore elongated one nucleotide at a time and has a nucleotide sequence exactly complementary to the strand of DNA used as the template (**Figure 7–6**).

Transcription differs from DNA replication in several crucial respects. Unlike a newly formed DNA strand, the RNA strand does not remain hydrogen-bonded to the DNA template strand. Instead, just behind the region where the ribonucleotides are being added, the RNA chain is displaced and the DNA helix re-forms. For this reason—and because only one strand of the DNA molecule is transcribed—RNA molecules are single-stranded. Further, because RNAs are copied from only a limited region of DNA, RNA molecules are much shorter than DNA molecules; DNA molecules in a human chromosome can be up to 250 million nucleotide pairs long, whereas most mature RNAs are no more than a few thousand nucleotides long, and many are much shorter than that.

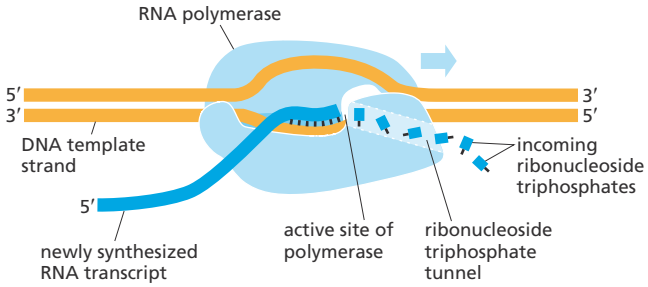


Figure 7–7 DNA is transcribed into RNA by the enzyme RNA polymerase. RNA polymerase (*pale blue*) moves stepwise along the DNA, unwinding the DNA helix in front of it. As it progresses, the polymerase adds ribonucleotides one by one to the RNA chain, using an exposed DNA strand as a template. The resulting RNA transcript is thus single-stranded and complementary to this template strand (see Figure 7–6). As the polymerase moves along the DNA template (in the 3′-to-5′ direction), it displaces the newly formed RNA, allowing the two strands of DNA behind the polymerase to rewind. A short region of hybrid DNA/RNA helix (approximately nine nucleotides in length) therefore forms only transiently, causing a "window" of DNA/RNA helix to move along the DNA with the polymerase (**Movie 7.2**).

Like the DNA polymerase that carries out DNA replication (discussed in Chapter 6), **RNA polymerases** catalyze the formation of the phosphodiester bonds that link the nucleotides together and form the sugar–phosphate backbone of the RNA chain (see Figure 7–3). The RNA polymerase moves stepwise along the DNA, unwinding the DNA helix just ahead to expose a new region of the template strand for complementary base-pairing. In this way, the growing RNA chain is extended by one nucleotide at a time in the 5′-to-3′ direction (**Figure 7–7**). The incoming ribonucleoside triphosphates (ATP, CTP, UTP, and GTP) provide the energy needed to drive the reaction forward (see Figure 6–11).

The almost immediate release of the RNA strand from the DNA as it is synthesized means that many RNA copies can be made from the same gene in a relatively short time; the synthesis of the next RNA is usually started before the first RNA has been completed (**Figure 7–8**). A medium-sized gene—say, 1500 nucleotide pairs—requires approximately 50 seconds for a molecule of RNA polymerase to transcribe it (**Movie 7.2**). At any given time, there could be dozens of polymerases speeding along this single stretch of DNA, hard on one another's heels, allowing more than 1000 transcripts to be synthesized in an hour. For most genes, however, the amount of transcription is much less than this.

Although RNA polymerase catalyzes essentially the same chemical reaction as DNA polymerase, there are some important differences between the two enzymes. First, and most obviously, RNA polymerase uses ribonucleoside for phosphates as substrates, so it catalyzes the linkage of ribonucleotides, not deoxyribonucleotides. Second, unlike the DNA polymerase involved in DNA replication, RNA polymerases can start an RNA chain without a primer. This difference likely evolved because transcription need not be as accurate as DNA replication; unlike DNA, RNA is not used as the permanent storage form of genetic information in cells, so mistakes in RNA transcripts have relatively minor consequences for a cell. RNA polymerases make about one mistake for every 10^4 nucleotides copied into RNA, whereas DNA polymerase makes only one mistake for every 10^7 nucleotides copied.

Cells Produce Various Types of RNA

The vast majority of genes carried in a cell's DNA specify the amino acid sequences of proteins. The RNA molecules encoded by these genes—which

QUESTION 7–2

In the electron micrograph in Figure 7–8, are the RNA polymerase molecules moving from right to left or from left to right? Why are the RNA transcripts so much shorter than the DNA segments (genes) that encode them?

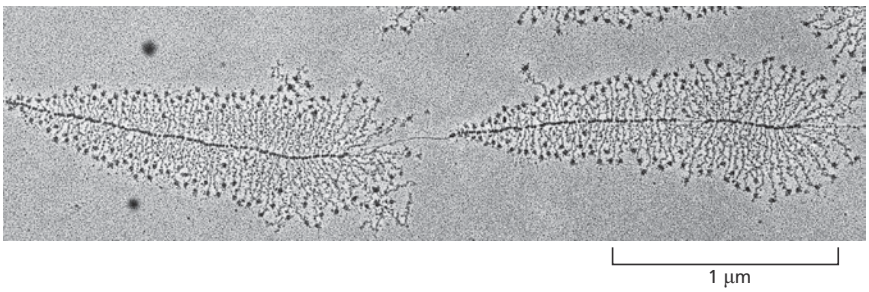


Figure 7–8 Transcription can be visualized in the electron microscope. The micrograph shows many molecules of RNA polymerase simultaneously transcribing two adjacent ribosomal genes on a single DNA molecule. Molecules of RNA polymerase are barely visible as a series of tiny dots along the spine of the DNA molecule; each polymerase has an RNA transcript (a short, fine thread) radiating from it. The RNA molecules being transcribed from the two ribosomal genes—ribosomal RNAs (rRNAs)—are not translated into protein, but are instead used directly as components of ribosomes, macromolecular machines made of RNA and protein. The large particles that can be seen at the free, 5′ end of each rRNA transcript are believed to be ribosomal proteins that have assembled on the ends of the growing transcripts. (Courtesy of Ulrich Scheer.)

ultimately direct the synthesis of proteins—are called **messenger RNAs** (**mRNAs**). In eukaryotes, each mRNA typically carries information transcribed from just one gene, which codes for a single protein; in bacteria, a set of adjacent genes is often transcribed as a single mRNA, which therefore carries the information for several different proteins.

The final product of other genes, however, is the RNA itself. As we see later, these nonmessenger RNAs, like proteins, have various roles, serving as regulatory, structural, and catalytic components of cells. They play key parts, for example, in translating the genetic message into protein: *ribosomal RNAs* (*rRNAs*) form the structural and catalytic core of the ribosomes, which translate mRNAs into protein, and *transfer RNAs* (*tRNAs*) act as adaptors that select specific amino acids and hold them in place on a ribosome for their incorporation into protein. Other small RNAs, called *microRNAs* (*miRNAs*), serve as key regulators of eukaryotic gene expression, as we discuss in Chapter 8. The most common types of RNA are summarized in Table 7–1.

In the broadest sense, the term **gene expression** refers to the process by which the information encoded in a DNA sequence is translated into a product that has some effect on a cell or organism. In cases where the final product of the gene is a protein, gene expression includes both transcription and translation. When an RNA molecule is the gene's final product, however, gene expression does not require translation.

Signals in DNA Tell RNA Polymerase Where to Start and Finish Transcription

The initiation of transcription is an especially critical process because it is the main point at which the cell selects which proteins or RNAs are to be produced. To begin transcription, RNA polymerase must be able to recognize the start of a gene and bind firmly to the DNA at this site. The way in which RNA polymerases recognize the *transcription start site* of a gene differs somewhat between bacteria and eukaryotes. Because the situation in bacteria is simpler, we describe it first.

When an RNA polymerase collides randomly with a DNA molecule, the enzyme sticks weakly to the double helix and then slides rapidly along its length. RNA polymerase latches on tightly only after it has encountered a gene region called a **promoter**, which contains a specific sequence of nucleotides that lies immediately upstream of the starting point for RNA synthesis. Once bound tightly to this sequence, the RNA polymerase opens up the double helix immediately in front of the promoter to expose the nucleotides on each strand of a short stretch of DNA. One of the two exposed DNA strands then acts as a template for complementary base-pairing with incoming ribonucleoside triphosphates, two of which are

TABLE 7–1 TYPES OF RNA PRODUCED IN CELLS

Type of RNA	Function
messenger RNAs (mRNAs)	code for proteins
ribosomal RNAs (rRNAs)	form the core of the ribosome's structure and catalyze protein synthesis
microRNAs (miRNAs)	regulate gene expression
transfer RNAs (tRNAs)	serve as adaptors between mRNA and amino acids during protein synthesis
other noncoding RNAs	used in RNA splicing, gene regulation, telomere maintenance, and many other processes

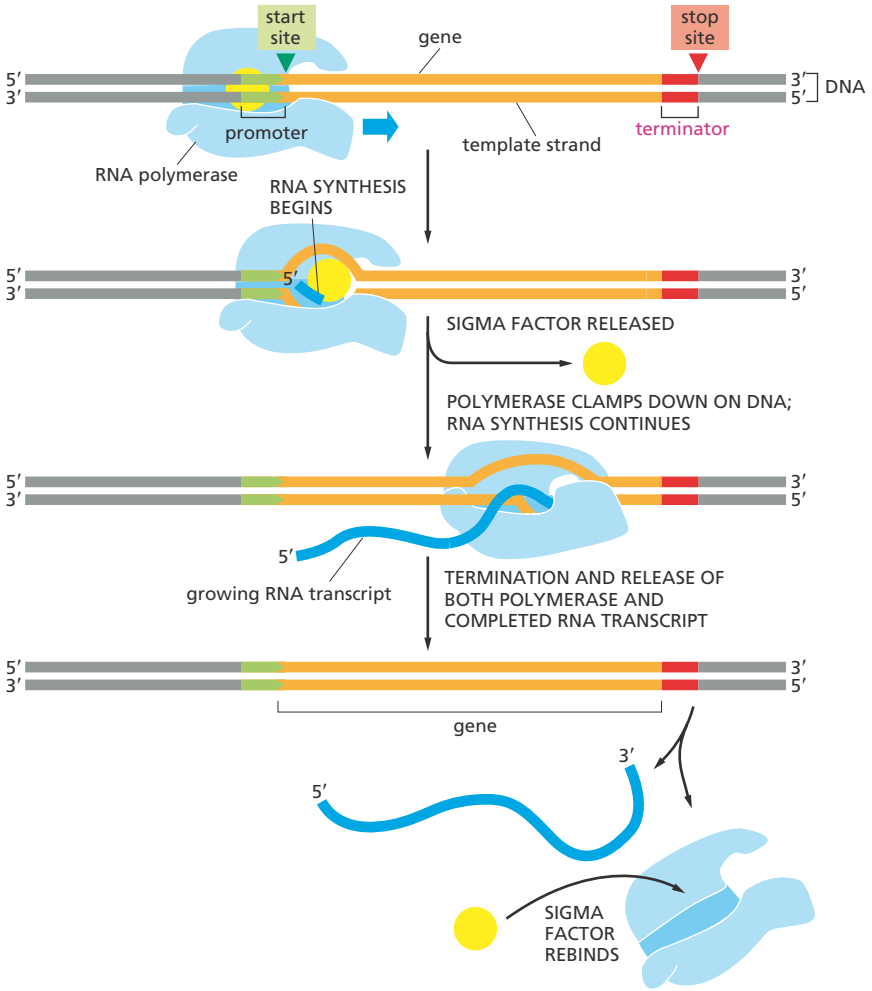


Figure 7–9 Signals in the nucleotide sequence of a gene tell bacterial RNA polymerase where to start and stop transcription. Bacterial RNA polymerase (*light blue*) contains a subunit called sigma factor (*yellow*) that recognizes the promoter of a gene (*green*). Once transcription has begun, sigma factor is released, and the polymerase moves forward and continues synthesizing the RNA. Chain elongation continues until the polymerase encounters a sequence in the gene called the terminator (*red*). There the enzyme halts and releases both the DNA template and the newly made RNA transcript. The polymerase then reassociates with a free sigma factor and searches for another promoter to begin the process again.

joined together by the polymerase to begin synthesis of the RNA chain. Chain elongation then continues until the enzyme encounters a second signal in the DNA, the *terminator* (or stop site), where the polymerase halts and releases both the DNA template and the newly made RNA transcript (**Figure 7–9**). This terminator sequence is contained within the gene and is transcribed into the 3′ end of the newly made RNA.

Because the polymerase must bind tightly before transcription can begin, a segment of DNA will be transcribed only if it is preceded by a promoter. This ensures that those portions of a DNA molecule that contain a gene will be transcribed into RNA. The nucleotide sequences of a typical promoter—and a typical terminator—are presented in **Figure 7–10**.

In bacteria, it is a subunit of RNA polymerase, the *sigma* (σ) *factor* (see Figure 7–9), that is primarily responsible for recognizing the promoter sequence on the DNA. But how can this factor "see" the promoter, given that the base-pairs in question are situated in the interior of the DNA double helix? It turns out that each base presents unique features to the outside of the double helix, allowing the sigma factor to find the promoter sequence without having to separate the entwined DNA strands.

The next problem an RNA polymerase faces is determining which of the two DNA strands to use as a template for transcription: each strand has a different nucleotide sequence and would produce a different RNA transcript. The secret lies in the structure of the promoter itself. Every promoter has a certain polarity: it contains two different nucleotide sequences upstream of the transcriptional start site that position the RNA polymerase, ensuring that it binds to the promoter in only one orientation

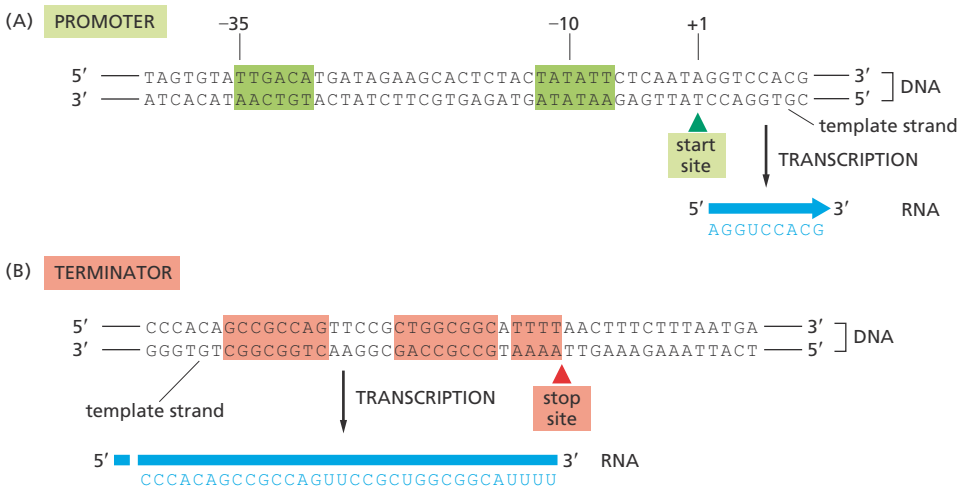


Figure 7–10 Bacterial promoters and terminators have specific nucleotide sequences that are recognized by RNA polymerase. (A) The *green*-shaded regions represent the nucleotide sequences that specify a promoter. The numbers above the DNA indicate the positions of nucleotides counting from the first nucleotide transcribed, which is designated +1. The polarity of the promoter orients the polymerase and determines which DNA strand is transcribed. All bacterial promoters contain DNA sequences at –10 and –35 that closely resemble those shown here. (B) The *red*-shaded regions represent sequences in the gene that signal the RNA polymerase to terminate transcription. Note that the regions transcribed into RNA contain the terminator but not the promoter nucleotide sequences. By convention, the sequence of a gene is that of the non-template strand, as this strand has the same sequence as the transcribed RNA (with T substituting for U).

(see Figure 7–10A). Because the polymerase can only synthesize RNA in the 5′-to-3′ direction once the enzyme is bound it must use the DNA strand oriented in the 3′-to-5′ direction as its template.

This selection of a template strand does not mean that on a given chromosome, transcription always proceeds in the same direction. With respect to the chromosome as a whole, the direction of transcription varies from gene to gene. But because each gene typically has only one promoter, the orientation of its promoter determines in which direction that gene is transcribed and therefore which strand is the template strand (**Figure 7–11**).

Initiation of Eukaryotic Gene Transcription Is a Complex Process

Many of the principles we just outlined for bacterial transcription also apply to eukaryotes. However, transcription initiation in eukaryotes differs in several important ways from that in bacteria:

- The first difference lies in the RNA polymerases themselves. While bacteria contain a single type of RNA polymerase, eukaryotic cells have three—*RNA polymerase I*, *RNA polymerase II*, and *RNA polymerase III*. These polymerases are responsible for transcribing different types of genes. RNA polymerases I and III transcribe the genes encoding transfer RNA, ribosomal RNA, and various other RNAs that play structural and catalytic roles in the cell (**Table 7–2**). RNA polymerase II transcribes the vast majority of eukaryotic genes, including all those that encode proteins and miRNAs (**Movie 7.3**). Our subsequent discussion will therefore focus on RNA polymerase II.
- A second difference is that, whereas the bacterial RNA polymerase (along with its sigma subunit) is able to initiate transcription on its own, eukaryotic RNA polymerases require the assistance of a large set of accessory proteins. Principal among these are the *general transcription factors*, which must assemble at each promoter, along with the polymerase, before the polymerase can begin transcription.

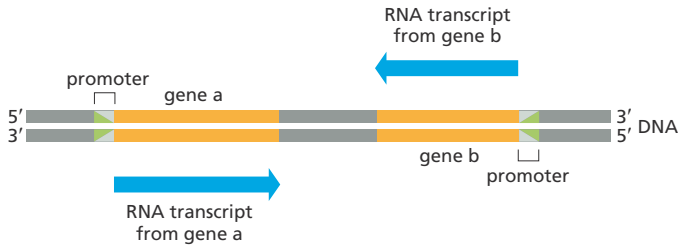


Figure 7–11 On an individual chromosome, some genes are transcribed using one DNA strand as a template, and others are transcribed from the other DNA strand. RNA polymerase always moves in the 3′-to-5′ direction and the selection of the template strand is determined by the orientation of the promoter (*green* arrowheads) at the beginning of each gene. Thus the genes transcribed from left to right use the bottom DNA strand as the template (see Figure 7–10); those transcribed from right to left use the top strand as the template.

TABLE 7–2 THE THREE RNA POLYMERASES IN EUKARYOTIC CELLS

Type of Polymerase	Genes Transcribed
RNA polymerase I	most rRNA genes
RNA polymerase II	all protein-coding genes, miRNA genes, plus genes for other noncoding RNAs (e.g., those in spliceosomes)
RNA polymerase III	tRNA genes 5S rRNA gene genes for many other small RNAs

- A third distinctive feature of transcription in eukaryotes is that the mechanisms that control its initiation are much more elaborate than those in prokaryotes—a point we discuss in detail in Chapter 8. In bacteria, genes tend to lie very close to one another in the DNA, with only very short lengths of nontranscribed DNA between them. But in plants and animals, including humans, individual genes are spread out along the DNA, with stretches of up to 100,000 nucleotide pairs between one gene and the next. This architecture allows a single gene to be controlled by a large variety of *regulatory DNA sequences* scattered along the DNA, and it enables eukaryotes to engage in more complex forms of transcriptional regulation than do bacteria.
- Last but not least, eukaryotic transcription initiation must take into account the packing of DNA into *nucleosomes* and more compact forms of chromatin structure, as we describe in Chapter 8.

We now turn to the general transcription factors and discuss how they help eukaryotic RNA polymerase II initiate transcription.

Eukaryotic RNA Polymerase Requires General Transcription Factors

The initial finding that, unlike bacterial RNA polymerase, purified eukaryotic RNA polymerase II could not initiate transcription on its own in a test tube led to the discovery and purification of the **general transcription factors**. These accessory proteins assemble on the promoter, where they position the RNA polymerase and pull apart the DNA double helix to expose the template strand, allowing the polymerase to begin transcription. Thus the general transcription factors have a similar role in eukaryotic transcription as sigma factor has in bacterial transcription.

Figure 7–12 shows how the general transcription factors assemble at a promoter used by RNA polymerase II. The assembly process typically begins with the binding of the general transcription factor TFIID to a short

QUESTION 7–3

Could the RNA polymerase used for transcription be used as the polymerase that makes the RNA primer required for DNA replication (discussed in Chapter 6)?

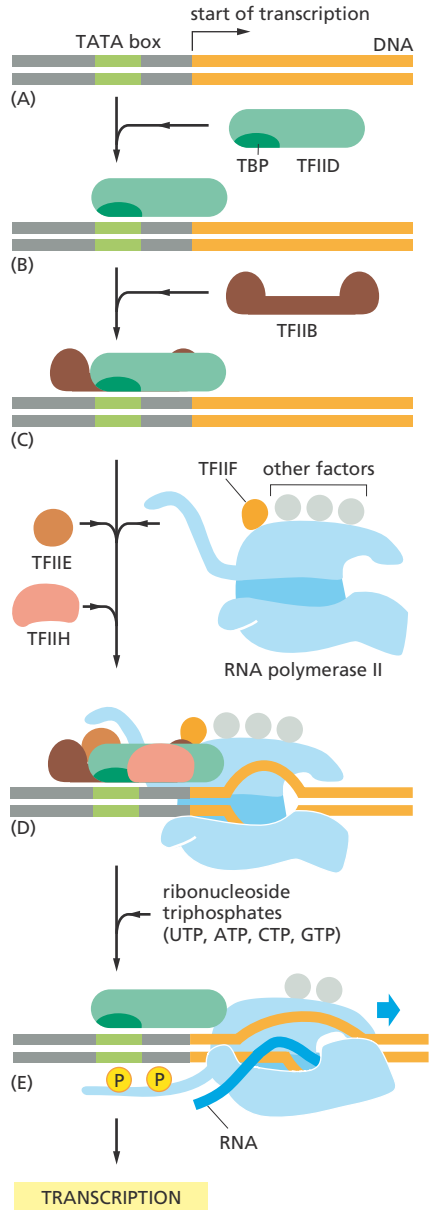


Figure 7–12 To begin transcription, eukaryotic RNA polymerase II requires a set of general transcription factors. These transcription factors are called TFIIB, TFIID, and so on. (A) Many eukaryotic promoters contain a DNA sequence called the TATA box. (B) The TATA box is recognized by a subunit of the general transcription factor TFIID, called the TATA-binding protein (TBP). For simplicity, the DNA distortion produced by the binding of the TBP (see Figure 7–13) is not shown. (C) The binding of TFIID enables the adjacent binding of TFIIB. (D) The rest of the general transcription factors, as well as the RNA polymerase itself, assemble at the promoter. (E) TFIIH then pries apart the double helix at the transcription start point, using the energy of ATP hydrolysis, which exposes the template strand of the gene (not shown). TFIIH also phosphorylates RNA polymerase II, releasing the polymerase from most of the general transcription factors, so it can begin transcription. The site of phosphorylation is a long polypeptide "tail" that extends from the polymerase.

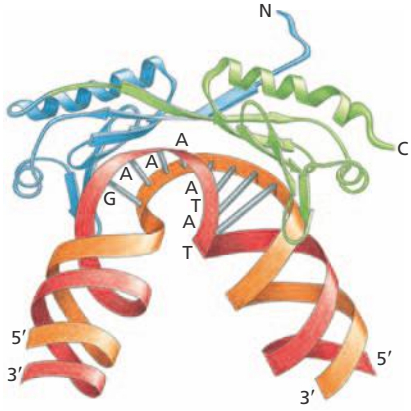


Figure 7–13 TATA-binding protein (TBP) binds to the TATA box (indicated by letters) and bends the DNA double helix. The unique distortion of DNA caused by TBP, which is a subunit of TFIID (see Figure 7–12), helps attract the other general transcription factors. TBP is a single polypeptide chain that is folded into two very similar domains (*blue* and *green*). The protein sits atop the DNA double helix like a saddle on a bucking horse (**Movie 7.4**). (Adapted from J.L. Kim et al., *Nature* 365:520–527, 1993. With permission from Macmillan Publishers Ltd.)

segment of DNA double helix composed primarily of T and A nucleotides; because of its composition, this part of the promoter is known as the *TATA box*. Upon binding to DNA, TFIID causes a dramatic local distortion in the DNA double helix (**Figure 7–13**), which helps to serve as a landmark for the subsequent assembly of other proteins at the promoter. The TATA box is a key component of many promoters used by RNA polymerase II, and it is typically located 25 nucleotides upstream from the transcription start site. Once TFIID has bound to the TATA box, the other factors assemble, along with RNA polymerase II, to form a complete *transcription initiation complex*. Although Figure 7–12 shows the general transcription factors piling onto the promoter in a certain order, the actual order of assembly probably differs from one promoter to the next.

After RNA polymerase II has been positioned on the promoter, it must be released from the complex of general transcription factors to begin its task of making an RNA molecule. A key step in liberating the RNA polymerase is the addition of phosphate groups to its "tail" (see Figure 7–12E). This liberation is initiated by the general transcription factor TFIIH, which contains a protein kinase as one of its subunits. Once transcription has begun, most of the general transcription factors dissociate from the DNA and then are available to initiate another round of transcription with a new RNA polymerase molecule. When RNA polymerase II finishes transcribing a gene, it too is released from the DNA; the phosphates on its tail are stripped off by protein phosphatases, and the polymerase is then ready to find a new promoter. Only the dephosphorylated form of RNA polymerase II can initiate RNA synthesis.

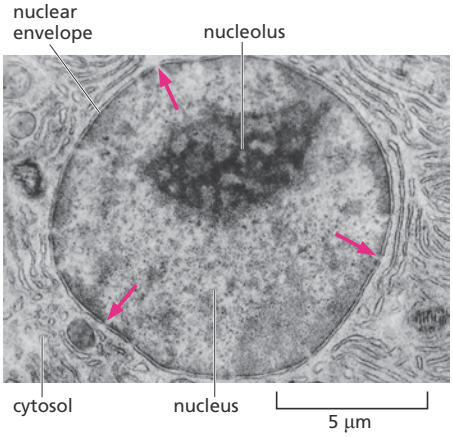


Figure 7–14 Before they can be translated, mRNA molecules made in the nucleus must be exported to the cytosol via pores in the nuclear envelope (*red arrows*). Shown here is a section of a liver cell nucleus. The nucleolus is where ribosomal RNAs are synthesized and combined with proteins to form ribosomes, which are then exported to the cytoplasm. (From D.W. Fawcett, A Textbook of Histology, 11th ed. Philadelphia: Saunders, 1986. With permission from Elsevier.)

Eukaryotic mRNAs Are Processed in the Nucleus

Although the templating principle by which DNA is transcribed into RNA is the same in all organisms, the way in which the RNA transcripts are handled before they can be used by the cell to make protein differs greatly between bacteria and eukaryotes. Bacterial DNA lies directly exposed to the cytoplasm, which contains the *ribosomes* on which protein synthesis takes place. As an mRNA molecule in a bacterium starts to be synthesized, ribosomes immediately attach to the free 5′ end of the RNA transcript and begin translating it into protein.

In eukaryotic cells, by contrast, DNA is enclosed within the *nucleus*. Transcription takes place in the nucleus, but protein synthesis takes place on ribosomes in the cytoplasm. So, before a eukaryotic mRNA can be translated into protein, it must be transported out of the nucleus through small pores in the nuclear envelope (**Figure 7–14**). Before it can be exported to the cytosol, however, a eukaryotic RNA must go through several **RNA processing** steps, which include *capping*, *splicing*, and *polyadenylation*, as we discuss shortly. These steps take place as the RNA is being synthesized. The enzymes responsible for RNA processing ride on the phosphorylated tail of eukaryotic RNA polymerase II as it synthesizes an RNA molecule (see Figure 7–12), and they process the transcript as it emerges from the polymerase (**Figure 7–15**).

Different types of RNA are processed in different ways before leaving the nucleus. Two processing steps, capping and polyadenylation, occur only on RNA transcripts destined to become mRNA molecules (called *precursor mRNAs*, or *pre-mRNAs*).

1. **RNA capping** modifies the 5′ end of the RNA transcript, the end that is synthesized first. The RNA is capped by the addition of an atypical nucleotide—a guanine (G) nucleotide bearing a methyl group, which is attached to the 5′ end of the RNA in an unusual way (**Figure 7–16**). This capping occurs after RNA polymerase II has produced about 25 nucleotides of RNA, long before it has completed transcribing the whole gene.
2. **Polyadenylation** provides a newly transcribed mRNA with a special structure at its 3′ end. In contrast with bacteria, where the 3′ end of an mRNA is simply the end of the chain synthesized by the RNA polymerase, the 3′ end of a forming eukaryotic mRNA is first trimmed by an enzyme that cuts the RNA chain at a particular sequence of nucleotides. The transcript is then finished off by a second enzyme that adds a series of repeated adenine (A) nucleotides to the cut end. This *poly-A tail* is generally a few hundred nucleotides long (see Figure 7–16A).

These two modifications—capping and polyadenylation—increase the stability of a eukaryotic mRNA molecule, facilitate its export from the nucleus to the cytoplasm, and generally mark the RNA molecule as an mRNA. They are also used by the protein-synthesis machinery to make sure that both ends of the mRNA are present and that the message is therefore complete before protein synthesis begins.

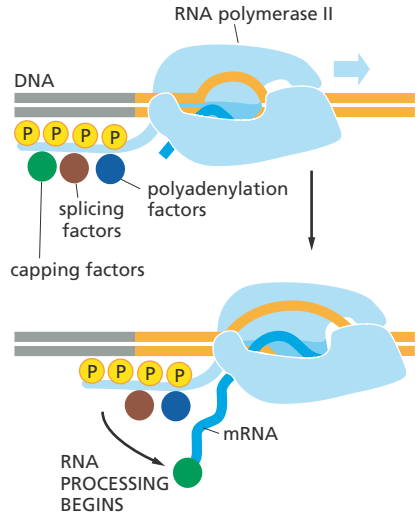
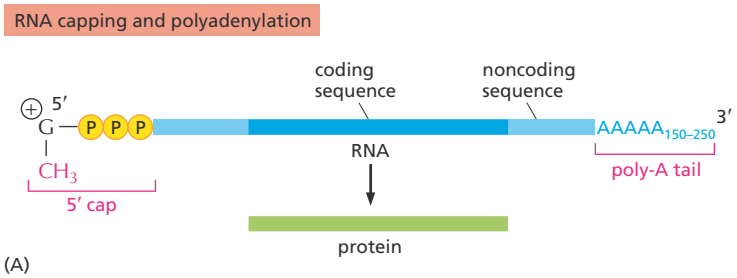


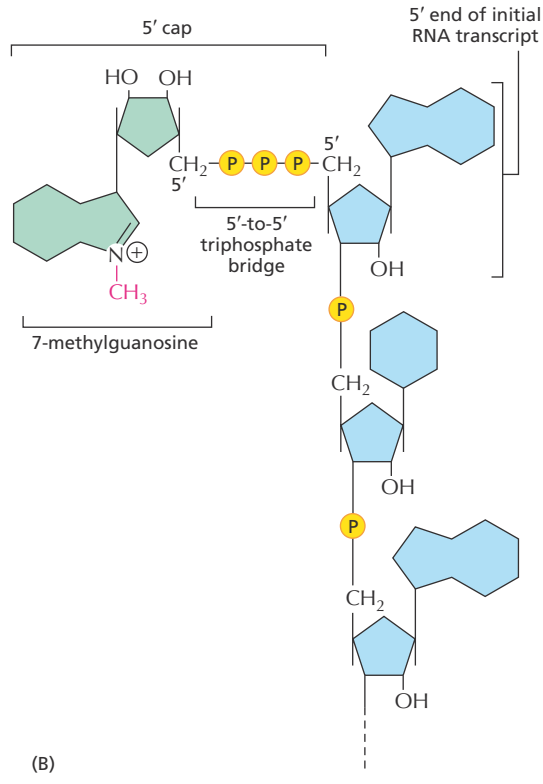
Figure 7–15 Phosphorylation of the tail of RNA polymerase II allows RNA-processing proteins to assemble there. Note that the phosphates shown here are in addition to the ones required for transcription initiation (see Figure 7–12). Capping, polyadenylation, and splicing are all modifications that occur during RNA processing in the nucleus.

In Eukaryotes, Protein-Coding Genes Are Interrupted by Noncoding Sequences Called Introns

Most eukaryotic pre-mRNAs have to undergo an additional processing step before they are functional mRNAs. This step involves a far more radical modification of the pre-mRNA transcript than capping or polyadenylation, and it is the consequence of a surprising feature of most eukaryotic genes. In bacteria, most proteins are encoded by an uninterrupted stretch of DNA sequence that is transcribed into an mRNA that, without any further processing, can be translated into protein. Most protein-coding eukaryotic genes, in contrast, have their coding sequences interrupted by long, noncoding, *intervening sequences* called **introns**. The scattered pieces of coding sequence—called *expressed sequences* or

Figure 7–16 Eukaryotic pre-mRNA molecules are modified by capping and polyadenylation. (A) A eukaryotic mRNA has a cap at the 5′ end and a poly-A tail at the 3′ end. Note that not all of the RNA transcript shown codes for protein. (B) The structure of the 5′ cap. Many eukaryotic mRNA caps carry an additional modification: the 2′-hydroxyl group on the second ribose sugar in the mRNA is methylated (not shown).





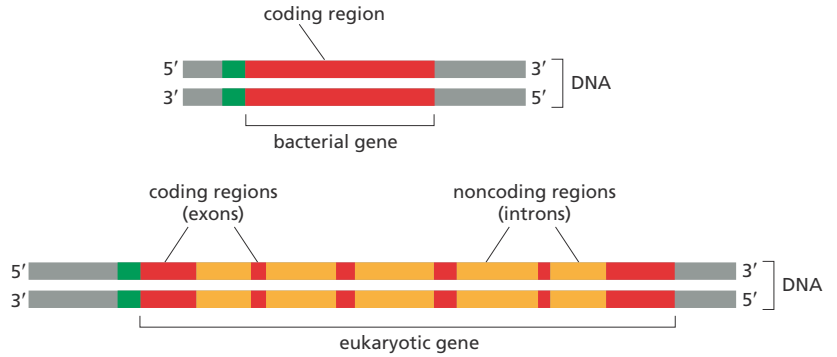


Figure 7–17 Eukaryotic and bacterial genes are organized differently. A bacterial gene consists of a single stretch of uninterrupted nucleotide sequence that encodes the amino acid sequence of a protein (or more than one protein). In contrast, the protein-coding sequences of most eukaryotic genes (*exons*) are interrupted by noncoding sequences (*introns*). Promoters for transcription are indicated in *green*.

exons—are usually shorter than the introns, and they often represent only a small fraction of the total length of the gene (**Figure 7–17**). Introns range in length from a single nucleotide to more than 10,000 nucleotides. Some protein-coding eukaryotic genes lack introns altogether, and some have only a few; but most have many (**Figure 7–18**). Note that the terms "exon" and "intron" apply to both the DNA and the corresponding RNA sequences.

Introns Are Removed From Pre-mRNAs by RNA Splicing

To produce an mRNA in a eukaryotic cell, the entire length of the gene, introns as well as exons, is transcribed into RNA. After capping, and as RNA polymerase II continues to transcribe the gene, the process of **RNA splicing** begins, in which the introns are removed from the newly synthesized RNA and the exons are stitched together. Each transcript ultimately receives a poly-A tail; in some cases, this happens after splicing, whereas in other cases, it occurs before the final splicing reactions have been completed. Once a transcript has been spliced and its 5′ and 3′ ends have been modified, the RNA is now a functional mRNA molecule that can leave the nucleus and be translated into protein.

How does the cell determine which parts of the RNA transcript to remove during splicing? Unlike the coding sequence of an exon, most of the nucleotide sequence of an intron is unimportant. Although there is little overall resemblance between the nucleotide sequences of different introns, each intron contains a few short nucleotide sequences that act as cues for its removal from the pre-mRNA. These special sequences are found at or near each end of the intron and are the same or very similar in all introns (**Figure 7–19**). Guided by these sequences, an elaborate splicing machine cuts out the intron in the form of a "lariat" structure (**Figure 7–20**), formed by the reaction of the "A" nucleotide highlighted in red in Figures 7–19 and 7–20.

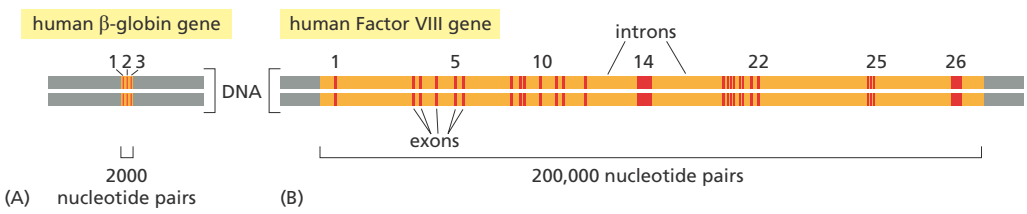


Figure 7–18 Most protein-coding human genes are broken into multiple exons and introns. (A) The β-globin gene, which encodes one of the subunits of the oxygen-carrying protein hemoglobin, contains 3 exons. (B) The Factor VIII gene, which encodes a protein (Factor VIII) that functions in the blood-clotting pathway, contains 26 exons. Mutations in this large gene are responsible for the most prevalent form of the blood disorder hemophilia.

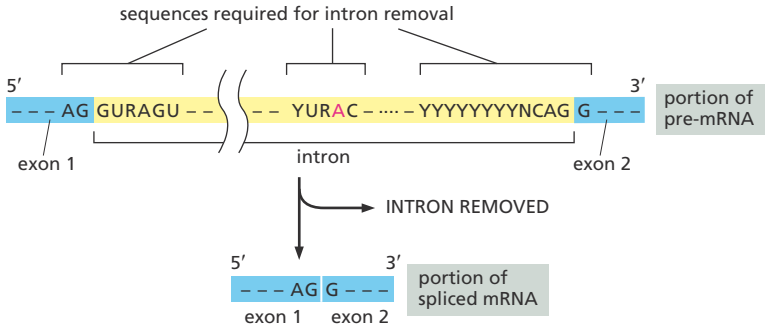


Figure 7–19 Special nucleotide sequences in a pre-mRNA transcript signal the beginning and the end of an intron. Only the nucleotide sequences shown are required to remove an intron; the other positions in an intron can be occupied by any nucleotide. The special sequences are recognized primarily by small nuclear ribonucleoproteins (snRNPs), which direct the cleavage of the RNA at the intron–exon borders and catalyze the covalent linkage of the exon sequences. Here, in addition to the standard symbols for nucleotides (A, C, G, U), R stands for either A or G; Y stands for either C or U; N stands for any nucleotide. The A shown in *red* forms the branch point of the lariat produced in the splicing reaction shown in Figure 7–20. The distances along the RNA between the three splicing sequences are highly variable; however, the distance between the branch point and the 5′ splice junction is typically much longer than that between the 3′ splice junction and the branch point (see Figure 7–20). The splicing sequences shown are from humans; similar sequences direct RNA splicing in other eukaryotes.

We will not describe the splicing machinery in detail, but it is worthwhile to note that, unlike the other steps of mRNA production we have discussed, RNA splicing is carried out largely by RNA molecules rather than proteins. These RNA molecules, called **small nuclear RNAs** (**snRNAs**), are packaged with additional proteins to form *small nuclear ribonucleoproteins* (*snRNPs*, pronounced "snurps"). The snRNPs recognize splice-site sequences through complementary base-pairing between their RNA components and the sequences in the pre-mRNA, and they also participate intimately in the chemistry of splicing (**Figure 7–21**). Together, these snRNPs form the core of the **spliceosome**, the large assembly of RNA and protein molecules that carries out RNA splicing in the nucleus. To watch the spliceosome in action, see **Movie 7.5**.

The intron–exon type of gene arrangement in eukaryotes may, at first, seem wasteful. It does, however, have a number of important benefits. First, the transcripts of many eukaryotic genes can be spliced in different ways, each of which can produce a distinct protein. Such **alternative splicing** thereby allows many different proteins to be produced from the same gene (**Figure 7–22**). About 95% of human genes are thought to undergo alternative splicing. Thus RNA splicing enables eukaryotes to increase the already enormous coding potential of their genomes.

RNA splicing also provides another advantage to eukaryotes, one that is likely to have been profoundly important in the early evolutionary history of genes. As we discuss in detail in Chapter 9, the intron–exon structure of genes is thought to have sped up the emergence of new and useful proteins: novel proteins appear to have arisen by the mixing and matching of different exons of preexisting genes, much like the assembly of a new type of machine from a kit of preexisting functional components. Indeed, many proteins in present-day cells resemble patchworks composed from a common set of protein pieces, called protein *domains* (see Figure 4–51).

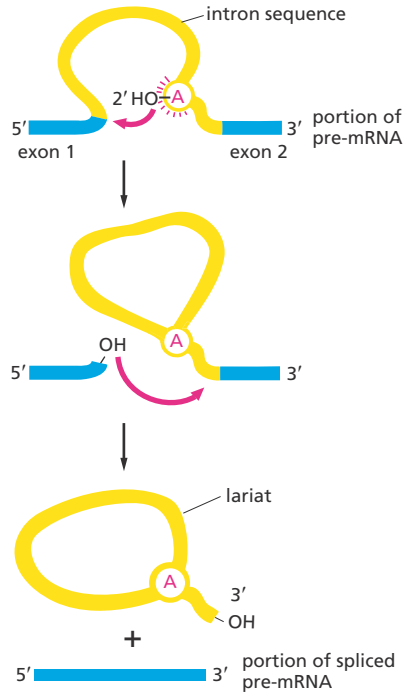


Figure 7–20 An intron in a pre-mRNA molecule forms a branched structure during RNA splicing. In the first step, the branch point adenine (*red* A) in the intron sequence attacks the 5′ splice site and cuts the sugar–phosphate backbone of the RNA at this point (this is the same A highlighted in *red* in Figure 7–19). In this process, the cut 5′ end of the intron becomes covalently linked to the 2′-OH group of the ribose of the A nucleotide to form a branched structure. The free 3′-OH end of the exon sequence then reacts with the start of the next exon sequence, joining the two exons together into a continuous coding sequence and releasing the intron in the form of a lariat structure, which is eventually degraded in the nucleus.

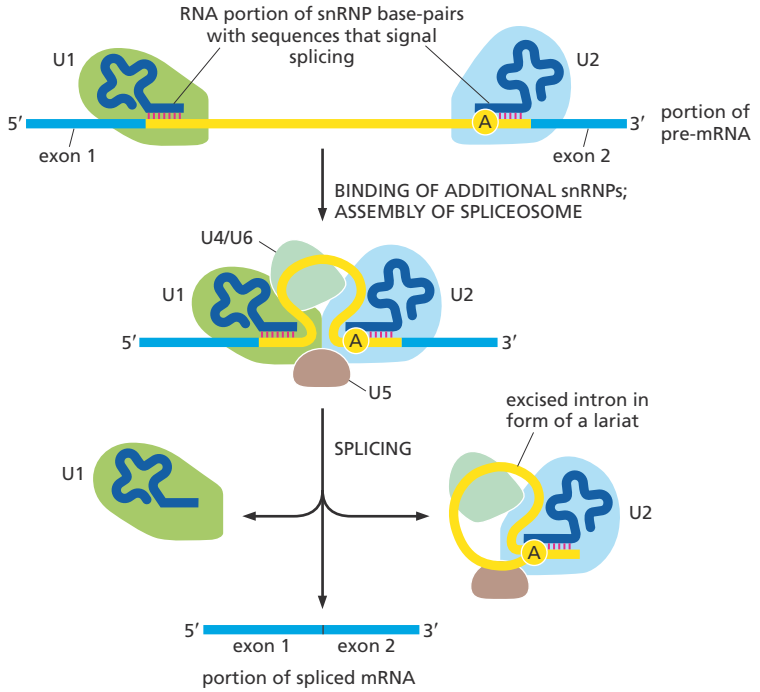


Figure 7–21 Splicing is carried out by a collection of RNA–protein complexes called snRNPs. There are five snRNPs, called U1, U2, U4, U5, and U6. As shown here, U1 and U2 bind to the 5′ splice site (U1) and the lariat branch point (U2) through complementary base-pairing. Additional snRNPs are attracted to the splice site, and interactions between their protein components drive the assembly of the complete spliceosome. Rearrangements in the base pairs that hold together the snRNPs and the RNA transcript then reorganize the spliceosome to form the active site that excises the intron, leaving the spliced mRNA behind (see also Figure 7–20).

Mature Eukaryotic mRNAs Are Exported from the Nucleus

We have seen how eukaryotic pre-mRNA synthesis and processing take place in an orderly fashion within the cell nucleus. However, these events create a special problem for eukaryotic cells: of the total number of pre-mRNA transcripts that are synthesized, only a small fraction—the mature mRNAs—will be useful to the cell. The remaining RNA fragments—excised introns, broken RNAs, and aberrantly spliced transcripts—are not only useless, but they could be dangerous to the cell if allowed to leave the nucleus. How, then, does the cell distinguish between the relatively rare mature mRNA molecules it needs to export to the cytosol and the overwhelming amount of debris generated by RNA processing?

The answer is that the transport of mRNA from the nucleus to the cytosol, where mRNAs are translated into protein, is highly selective: only correctly processed mRNAs are exported. This selective transport is mediated by *nuclear pore complexes*, which connect the nucleoplasm with the cytosol and act as gates that control which macromolecules can enter or leave the nucleus (discussed in Chapter 15). To be "export ready," an mRNA molecule must be bound to an appropriate set of proteins, each of which recognizes different parts of a mature mRNA molecule. These proteins include poly-A–binding proteins, a cap-binding complex, and

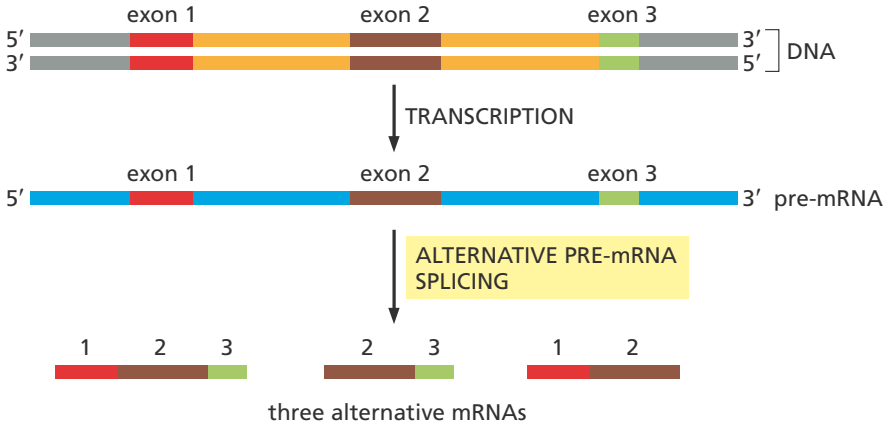


Figure 7–22 Some pre-mRNAs undergo alternative RNA splicing to produce various mRNAs and proteins from the same gene. Whereas all exons are present in a pre-mRNA, some exons can be excluded from the final mRNA molecule. In this example, three of four possible mRNAs are produced. The 5′ caps and poly-A tails on the mRNAs are not shown.

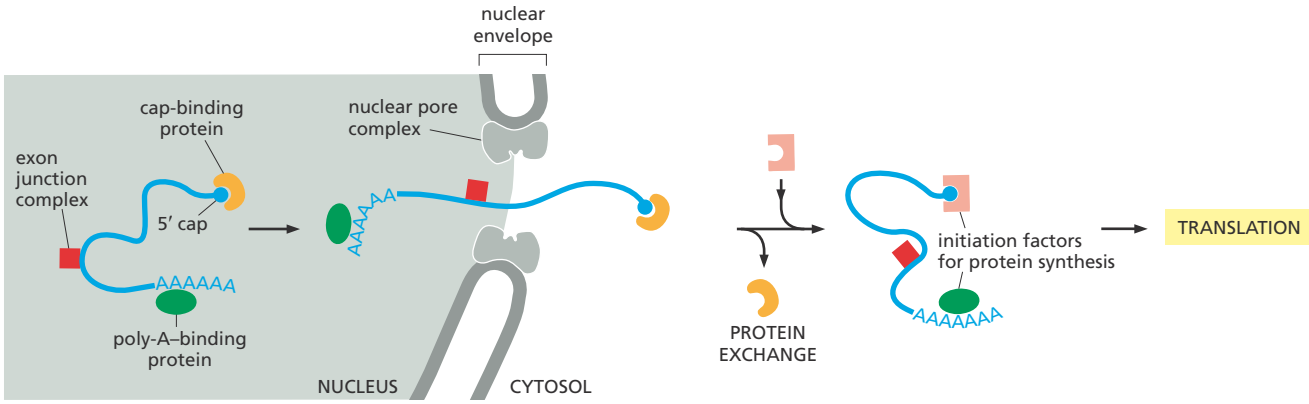


Figure 7–23 A specialized set of RNA-binding proteins signals that a mature mRNA is ready for export to the cytosol. As indicated on the left, the cap and poly-A tail of a mature mRNA molecule are "marked" by proteins that recognize these modifications. In addition, a group of proteins called the *exon junction complex* is deposited on the pre-mRNA after each successful splice has occurred. Once the mRNA is deemed "export ready," a nuclear transport receptor (discussed in Chapter 15) associates with the mRNA and guides it through the nuclear pore. In the cytosol, the mRNA can shed some of these proteins and bind new ones, which, along with poly-A–binding protein, act as initiation factors for protein synthesis, as we discuss later.

proteins that bind to mRNAs that have been appropriately spliced (**Figure 7–23**). The entire set of bound proteins, rather than any single protein, ultimately determines whether an mRNA molecule will leave the nucleus. The "waste RNAs" that remain behind in the nucleus are degraded there, and their nucleotide building blocks are reused for transcription.

mRNA Molecules Are Eventually Degraded in the Cytosol

Because a single mRNA molecule can be translated into protein many times (see Figure 7–2), the length of time that a mature mRNA molecule persists in the cell affects the amount of protein it produces. Each mRNA molecule is eventually degraded into nucleotides by ribonucleases (RNAses) present in the cytosol, but the lifetimes of mRNA molecules differ considerably—depending on the nucleotide sequence of the mRNA and the type of cell. In bacteria, most mRNAs are degraded rapidly, having a typical lifetime of about 3 minutes. The mRNAs in eukaryotic cells usually persist longer: some, such as those encoding β-globin, have lifetimes of more than 10 hours, whereas others have lifetimes of less than 30 minutes.

These different lifetimes are in part controlled by nucleotide sequences in the mRNA itself, most often in the portion of RNA called the *3′ untranslated region*, which lies between the 3′ end of the coding sequence and the poly-A tail. The different lifetimes of mRNAs help the cell control the amount of each protein that it synthesizes. In general, proteins made in large amounts, such as β-globin, are translated from mRNAs that have long lifetimes, whereas proteins made in smaller amounts, or whose levels must change rapidly in response to signals, are typically synthesized from short-lived mRNAs.

The Earliest Cells May Have Had Introns in Their Genes

The process of transcription is universal: all cells use RNA polymerase and complementary base-pairing to synthesize RNA from DNA. Indeed, bacterial and eukaryotic RNA polymerases are almost identical in overall structure and clearly evolved from a shared ancestral polymerase. It may therefore seem puzzling that the resulting RNA transcripts are handled so differently in eukaryotes and in prokaryotes (**Figure 7–24**). In particular, RNA splicing seems to mark a fundamental difference between those two types of cells. But how did this dramatic difference arise?

As we have seen, RNA splicing provides eukaryotes with the ability to produce a variety of proteins from a single gene. It also allows them to evolve new genes by mixing-and-matching exons from preexisting genes, as we discuss in Chapter 9. However, these advantages come with a cost: the cell has to maintain a larger genome and has to discard a

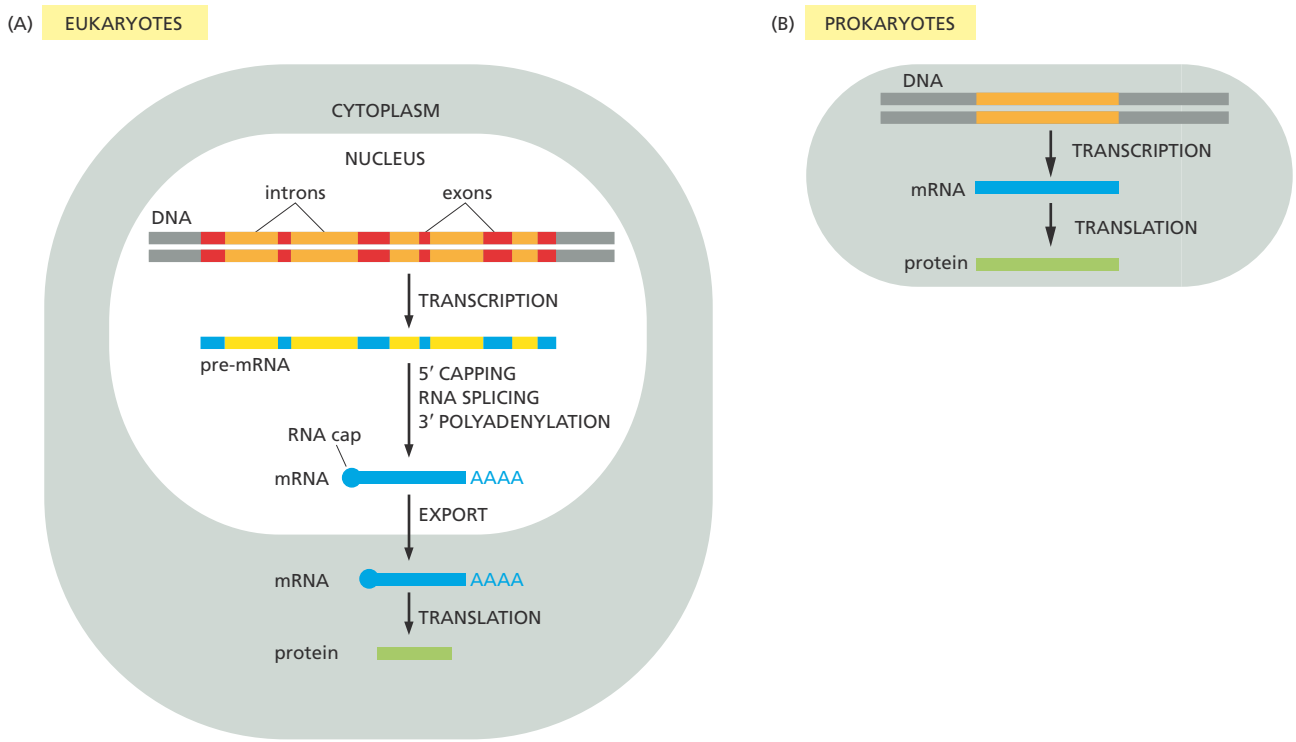


Figure 7–24 Prokaryotes and eukaryotes handle their RNA transcripts differently. (A) In eukaryotic cells, the pre-mRNA molecule produced by transcription contains both intron and exon sequences. Its two ends are modified, and the introns are removed by RNA splicing. The resulting mRNA is then transported from the nucleus to the cytoplasm, where it is translated into protein. Although these steps are depicted as occurring in sequence, one at a time, in reality they occur simultaneously. For example, the RNA cap is usually added and splicing usually begins before transcription has been completed. Because of this overlap, transcripts of the entire gene (including all introns and exons) do not typically exist in the cell. (B) In prokaryotes, the production of mRNA molecules is simpler. The 5′ end of an mRNA molecule is produced by the initiation of transcription by RNA polymerase, and the 3′ end is produced by the termination of transcription. Because prokaryotic cells lack a nucleus, transcription and translation take place in a common compartment. Translation of a bacterial mRNA can therefore begin before its synthesis has been completed. In both eukaryotes and prokaryotes, the amount of a protein in a cell depends on the rates of each of these steps, as well as on the rates of degradation of the mRNA and protein molecules.

large fraction of the RNA it synthesizes without ever using it. According to one school of thought, early cells—the common ancestors of prokaryotes and eukaryotes—contained introns that were lost in prokaryotes during subsequent evolution. By shedding their introns and adopting a smaller, more streamlined genome, prokaryotes would have been able to reproduce more rapidly and efficiently. Consistent with this idea, simple eukaryotes that reproduce rapidly (some yeasts, for example) have relatively few introns, and these introns are usually much shorter than those found in higher eukaryotes.

On the other hand, some argue that introns were originally parasitic mobile genetic elements (discussed in Chapter 9) that happened to invade an early eukaryotic ancestor, colonizing its genome. These host cells then unwittingly replicated the "stowaway" nucleotide sequences along with their own DNA; modern eukaryotes simply never bothered to sweep away the genetic clutter left from that ancient infection. The issue, however, is far from settled; whether introns evolved early—and were lost in prokaryotes—or evolved later in eukaryotes is still a topic of scientific debate, and we return to it in Chapter 9.

FROM RNA TO PROTEIN

By the end of the 1950s, biologists had demonstrated that the information encoded in DNA is copied first into RNA and then into protein. The debate then shifted to the "coding problem": How is the information in a linear sequence of nucleotides in an RNA molecule translated into the linear sequence of a chemically quite different set of subunits—the amino acids in a protein? This fascinating question intrigued scientists at the time. Here was a cryptogram set up by nature that, after more than 3 billion years of evolution, could finally be solved by one of the products of evolution—human beings! Indeed, scientists have not only cracked the code but have revealed, in atomic detail, the precise workings of the machinery by which cells read this code.

codons	GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
amino acids	Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop	
	A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V		

Figure 7–25 The nucleotide sequence of an mRNA is translated into the amino acid sequence of a protein via the genetic code. All the three-nucleotide codons in mRNAs that specify a given amino acid are listed above that amino acid, which is given in both its three-letter and one-letter abbreviations (see Panel 2–5, pp. 74–75, for the full name of each amino acid and its structure). Like RNA molecules, codons are always written with the 5′-terminal nucleotide to the left. Note that most amino acids are represented by more than one codon, and there are some regularities in the set of codons that specify each amino acid. Codons for the same amino acid tend to contain the same nucleotides at the first and second positions and to vary at the third position. There are three codons that do not specify any amino acid but act as termination sites (*stop codons*), signaling the end of the protein-coding sequence in an mRNA. One codon—AUG—acts both as an initiation codon, signaling the start of a protein-coding message, and as the codon that specifies the amino acid methionine.

An mRNA Sequence Is Decoded in Sets of Three Nucleotides

Transcription as a means of information transfer is simple to understand: DNA and RNA are chemically and structurally similar, and DNA can act as a direct template for the synthesis of RNA through complementary base-pairing. As the term transcription signifies, it is as if a message written out by hand were being converted, say, into a typewritten text. The language itself and the form of the message do not change, and the symbols used are closely related.

In contrast, the conversion of the information in RNA into protein represents a **translation** of the information into another language that uses different symbols. Because there are only 4 different nucleotides in mRNA but 20 different types of amino acids in a protein, this translation cannot be accounted for by a direct one-to-one correspondence between a nucleotide in RNA and an amino acid in protein. The rules by which the nucleotide sequence of a gene, through an intermediary mRNA molecule, is translated into the amino acid sequence of a protein are known as the **genetic code**.

In 1961, it was discovered that the sequence of nucleotides in an mRNA molecule is read consecutively in groups of three. And because RNA is made of 4 different nucleotides, there are $4 \times 4 \times 4 = 64$ possible combinations of three nucleotides: AAA, AUA, AUG, and so on. However, only 20 different amino acids are commonly found in proteins. Either some nucleotide triplets are never used, or the code is redundant, with some amino acids being specified by more than one triplet. The second possibility turned out to be correct, as shown by the completely deciphered genetic code shown in **Figure 7–25**. Each group of three consecutive nucleotides in RNA is called a **codon**, and each codon specifies one amino acid. The strategy by which this code was cracked is described in **How We Know**, pp. 240–241.

The same genetic code is used in nearly all present-day organisms. Although a few slight differences have been found, these occur chiefly in the mRNA of mitochondria and of some fungi and protozoa. Mitochondria have their own DNA replication, transcription, and protein-synthesis machinery, which operates independently from the corresponding machinery in the rest of the cell (discussed in Chapter 14), and they have been able to accommodate minor changes to the otherwise universal genetic code. Even in fungi and protozoa, the similarities in the code far outweigh the differences.

In principle, an mRNA sequence can be translated in any one of three different **reading frames**, depending on where the decoding process begins (**Figure 7–26**). However, only one of the three possible reading frames

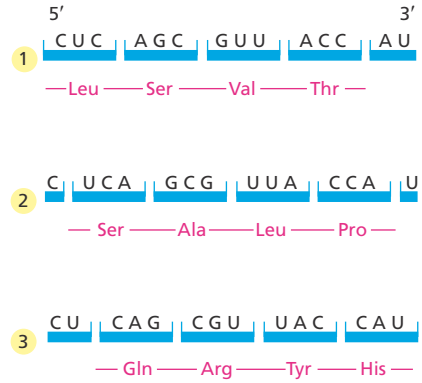


Figure 7–26 In principle, an mRNA molecule can be translated in three possible reading frames. In the process of translating a nucleotide sequence (*blue*) into an amino acid sequence (*red*), the sequence of nucleotides in an mRNA molecule is read from the 5′ to the 3′ end in sequential sets of three nucleotides. In principle, therefore, the same mRNA sequence can specify three completely different amino acid sequences, depending on where translation begins—that is, on the reading frame used. In reality, however, only one of these reading frames encodes the actual message and is therefore used in translation, as we discuss later.

CRACKING THE GENETIC CODE

By the beginning of the 1960s, the *central dogma* had been accepted as the pathway along which information flows from gene to protein. It was clear that genes encode proteins, that genes are made of DNA, and that mRNA serves as an intermediary, carrying the information from DNA to the ribosome, where the RNA is translated into protein.

Even the general format of the genetic code had been worked out: each of the 20 amino acids found in proteins is represented by a triplet codon in an mRNA molecule. But an even greater challenge remained: biologists, chemists, and even physicists set their sights on breaking the genetic code—attempting to figure out which amino acid each of the 64 possible nucleotide triplets designates. The most straightforward path to the solution would have been to compare the sequence of a segment of DNA or of mRNA with its corresponding polypeptide product. Techniques for sequencing nucleic acids, however, would not be devised for another 10 years.

So researchers decided that, to crack the genetic code, they would have to synthesize their own simple RNA molecules. If they could feed these RNA molecules to ribosomes—the machines that make proteins—and then analyze the resulting polypeptide product, they would be on their way to deciphering which triplets encode which amino acids.

Losing the cells

Before researchers could test their synthetic mRNAs, they needed to perfect a cell-free system for protein synthesis. This would allow them to translate their messages into polypeptides in a test tube. (Generally speaking, when working in the laboratory, the simpler the system, the easier it is to interpret the results.) To isolate the molecular machinery they needed for such a cell-free translation system, researchers broke open *E. coli* cells and loaded their contents into a centrifuge tube. Spinning these samples at high speed caused the membranes and other large chunks of cellular debris to be dragged to the bottom of the tube; the lighter cellular components required for protein synthesis—including mRNA, the tRNA adaptors, ribosomes, enzymes, and other small molecules—were left floating in the supernatant. Researchers found that simply adding radioactive amino acids to this cell "soup" would trigger the production of radiolabeled polypeptides. By centrifuging this supernatant again, at a higher speed, the researchers could force the ribosomes, and any newly synthesized peptides attached to them, to the bottom of the tube; the labeled polypeptides could then be detected by measuring the radioactivity in the sediment remaining in the tube after the top layer had been discarded.

The trouble with this particular system was that it produced proteins encoded by the cell's own mRNAs already present in the extract. But researchers wanted to use their own synthetic messages to direct protein synthesis. This problem was solved when Marshall Nirenberg discovered that he could destroy the cells' mRNA in the extract by adding a small amount of ribonuclease—an enzyme that degrades RNA—to the mix. Now all he needed to do was prepare large quantities of synthetic mRNA, add it to the cell-free system, and see what peptides came out.

Faking the message

Producing a synthetic polynucleotide with a defined sequence was not as simple as it sounds. Again, it would be years before chemists and bioengineers developed machines that could synthesize any given string of nucleic acids quickly and cheaply. Nirenberg decided to use polynucleotide phosphorylase, an enzyme that would join ribonucleotides together in the absence of a template. The sequence of the resulting RNA would then depend entirely on which nucleotides were presented to the enzyme. A mixture of nucleotides would be sewn into a random sequence; but a single type of nucleotide would yield a homogeneous polymer containing only that one nucleotide. Thus Nirenberg, working with his collaborator Heinrich Matthaei, first produced synthetic mRNAs made entirely of uracil—poly U.

Together, the researchers fed this poly U to their cell-free translation system. They then added a single type of radioactively labeled amino acid to the mix. After testing each amino acid—one at a time, in 20 different experiments—they determined that poly U directs the synthesis of a polypeptide containing only phenylalanine (**Figure 7–27**). With this electrifying result, the first word in the genetic code had been deciphered (see Figure 7–25).

Nirenberg and Matthaei then repeated the experiment with poly A and poly C and determined that AAA codes for lysine and CCC for proline. The meaning of poly G could not be ascertained by this method because this polynucleotide forms an odd triple-stranded helix that did not serve as a template in the cell-free system.

Feeding ribosomes with synthetic RNA seemed a fruitful technique. But with the single-nucleotide possibilities exhausted, researchers had nailed down only three codons; they had 61 still to go. The other codons, however, were harder to decipher, and a new synthetic approach was needed. In the 1950s, the organic chemist Gobind Khorana had been developing methods for preparing mixed polynucleotides of defined sequence—but his techniques worked only for DNA. When he

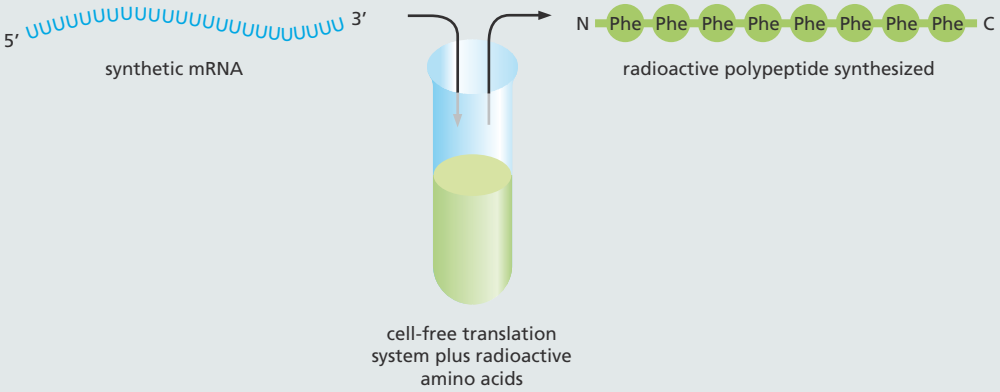


Figure 7–27 UUU codes for phenylalanine. Synthetic mRNAs are fed into a cell-free translation system containing bacterial ribosomes, tRNAs, enzymes, and other small molecules. Radioactive amino acids are added to this mix and the resulting polypeptides analyzed. In this case, poly U is shown to encode a polypeptide containing only phenylalanine.

learned of Nirenberg's work with synthetic RNAs, Khorana directed his energies and skills to producing polyribonucleotides. He found that if he started out by making DNAs of a defined sequence, he could then use RNA polymerase to produce RNAs from those. In this way, Khorana prepared a collection of different RNAs of defined repeating sequence: he generated sequences of repeating dinucleotides (such as poly UC), trinucleotides (such as poly UUC), or tetranucleotides (such as poly UAUC).

These mixed polynucleotides, however, yielded results that were much more difficult to decode than the mononucleotide messages that Nirenberg had used. Take poly UG, for example. When this repeating dinucleotide is added to the translation system, researchers discovered that it codes for a polypeptide of alternating cysteines and valines. This RNA, of course, contains two different alternating codons: UGU and GUG. So researchers could say that UGU and GUG code for cysteine and valine, although they could not tell which went with which. Thus these mixed messages provided useful information, but they did not definitively reveal which codons specified which amino acids (**Figure 7–28**).

Trapping the triplets

These final ambiguities in the code were resolved when Nirenberg and a young medical graduate named Phil Leder discovered that RNA fragments that were only three nucleotides in length—the size of a single codon—could bind to a ribosome and attract the appropriate amino-acid-containing tRNA molecule to the protein-making machinery. These complexes—containing one ribosome, one mRNA codon, and one radiolabeled aminoacyl-tRNA—could then be captured on a piece of filter paper and the attached amino acid identified.

Their trial run with UUU—the first word—worked splendidly. Leder and Nirenberg primed the usual cell-free translation system with snippets of UUU. These trinucleotides bound to the ribosomes, and Phe-tRNAs bound to the UUU. The new system was up and running, and the researchers had confirmed that UUU codes for phenylalanine.

All that remained was for researchers to produce all 64 possible codons—a task that was quickly accomplished in both Nirenberg's and Khorana's laboratories. Because these small trinucleotides were much simpler to synthesize chemically, and the triplet-trapping tests were easier to perform and analyze than the previous decoding experiments, the researchers were able to work out the complete genetic code within the next year.

MESSAGE	PEPTIDES PRODUCED	CODON ASSIGNMENTS
poly UG	...Cys–Val–Cys–Val...	UGU, GUG — Cys, Val*
poly AG	...Arg–Glu–Arg–Glu...	AGA, GAG — Arg, Glu
poly UUC	...Phe–Phe–Phe... + ...Ser–Ser–Ser... + ...Leu–Leu–Leu...	UUC, UCU, CUU — Phe, Ser, Leu
poly UAUC	...Tyr–Leu–Ser–Ile...	UAU, CUA, UCU, AUC — Tyr, Leu, Ser, Ile

* One codon specifies Cys, the other Val, but which is which? The same ambiguity exists for the other codon assignments shown here.

Figure 7–28 Using synthetic RNAs of mixed, repeating ribonucleotide sequences, scientists further narrowed the coding possibilities. Although these mixed messages produced mixed polypeptides, they did not permit the unambiguous assignment of a single codon to a specific amino acid. For example, the results of the poly-UG experiment cannot distinguish whether UGU or GUG encodes cysteine. As indicated, the same type of ambiguity confounded the interpretation of all the experiments using di-, tri-, and tetranucleotides.

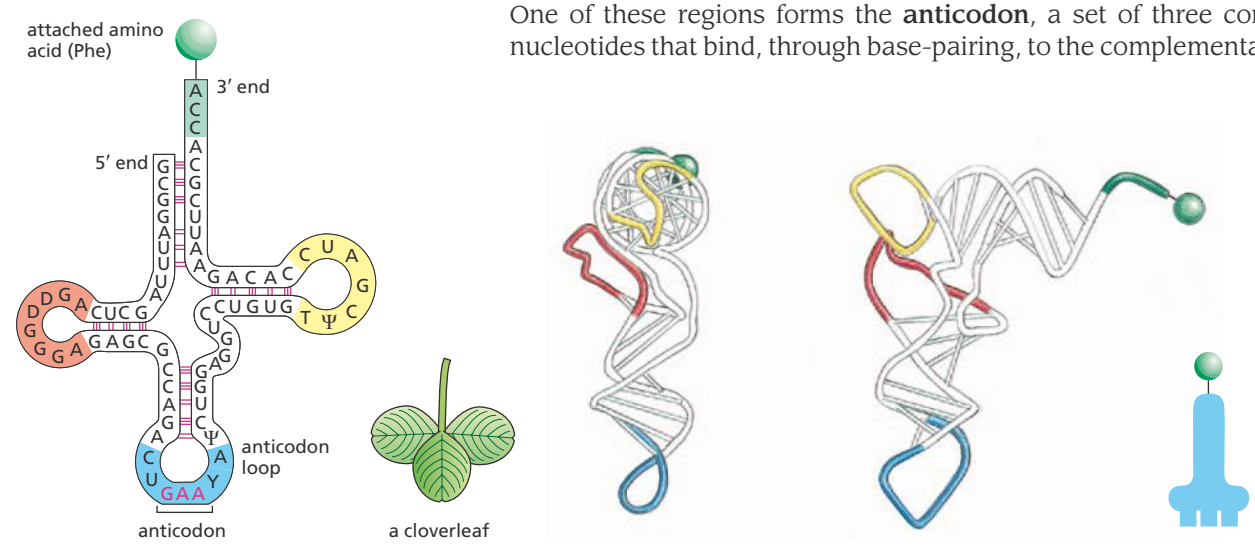
in an mRNA specifies the correct protein. We discuss later how a special punctuation signal at the beginning of each mRNA molecule sets the correct reading frame.

tRNA Molecules Match Amino Acids to Codons in mRNA

The codons in an mRNA molecule do not directly recognize the amino acids they specify: the group of three nucleotides does not, for example, bind directly to the amino acid. Rather, the translation of mRNA into protein depends on adaptor molecules that can recognize and bind to a codon at one site on their surface and to an amino acid at another site. These adaptors consist of a set of small RNA molecules known as **transfer RNAs** (**tRNAs**), each about 80 nucleotides in length.

We saw earlier that an RNA molecule generally folds into a three-dimensional structure by forming base pairs between different regions of the molecule. If the base-paired regions are sufficiently extensive, they will fold back on themselves to form a double-helical structure, like that of double-stranded DNA. The tRNA molecule provides a striking example of this. Four short segments of the folded tRNA are double-helical, producing a molecule that looks like a cloverleaf when drawn schematically (**Figure 7–29A**). For example, a 5′-GCUC-3′ sequence in one part of a polynucleotide chain can base-pair with a 5′-GAGC-3′ sequence in another region of the same molecule. The cloverleaf undergoes further folding to form a compact, L-shaped structure that is held together by additional hydrogen bonds between different regions of the molecule (**Figure 7–29B and C**).

Two regions of unpaired nucleotides situated at either end of the L-shaped tRNA molecule are crucial to the function of tRNAs in protein synthesis. One of these regions forms the **anticodon**, a set of three consecutive nucleotides that bind, through base-pairing, to the complementary codon



5′ GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3′

(E) anticodon

Figure 7–29 tRNA molecules are molecular adaptors, linking amino acids to codons. In this series of diagrams, the same tRNA molecule—in this case, a tRNA specific for the amino acid phenylalanine (Phe)—is depicted in various ways. (A) The conventional "cloverleaf" structure shows the complementary base-pairing (*red lines*) that creates the double-helical regions of the molecule. The anticodon loop (*blue*) contains the sequence of three nucleotides (*red letters*) that base-pairs with a codon in mRNA. The amino acid matching the codon–anticodon pair is attached at the 3′ end of the tRNA. tRNAs contain some unusual bases, which are produced by chemical modification after the tRNA has been synthesized. The bases denoted Ψ (for pseudouridine) and D (for dihydrouridine) are derived from uracil. (B and C) Views of the actual L-shaped molecule, based on X-ray diffraction analysis. These two images are rotated 90° with respect to each other. (D) Schematic representation of tRNA, emphasizing the anticodon, that will be used in subsequent figures. (E) The linear nucleotide sequence of the tRNA molecule, color-coded to match A, B, and C.

in an mRNA molecule. The other is a short single-stranded region at the 3′ end of the molecule; this is the site where the amino acid that matches the codon is covalently attached to the tRNA.

We saw in the previous section that the genetic code is redundant; that is, several different codons can specify a single amino acid (see Figure 7–25). This redundancy implies either that there is more than one tRNA for many of the amino acids or that some tRNA molecules can base-pair with more than one codon. In fact, both situations occur. Some amino acids have more than one tRNA, and some tRNAs are constructed so that they require accurate base-pairing only at the first two positions of the codon and can tolerate a mismatch (or *wobble*) at the third position. This wobble base-pairing explains why so many of the alternative codons for an amino acid differ only in their third nucleotide (see Figure 7–25). Wobble base-pairings make it possible to fit the 20 amino acids to their 61 codons with as few as 31 kinds of tRNA molecules. The exact number of different kinds of tRNAs, however, differs from one species to the next. For example, humans have nearly 500 different tRNA genes, but only 48 anticodons are represented among them.

Specific Enzymes Couple tRNAs to the Correct Amino Acid

For a tRNA molecule to carry out its role as an adaptor, it must be linked—or charged—with the correct amino acid. How does each tRNA molecule recognize the one amino acid in 20 that is its right partner? Recognition and attachment of the correct amino acid depend on enzymes called **aminoacyl-tRNA synthetases**, which covalently couple each amino acid to its appropriate set of tRNA molecules. In most organisms, there is a different synthetase enzyme for each amino acid. That means that there are 20 synthetases in all: one attaches glycine to all tRNAs that recognize codons for glycine, another attaches phenylalanine to all tRNAs that recognize codons for phenylalanine, and so on. Each synthetase enzyme recognizes specific nucleotides in both the anticodon and the amino-acid-accepting arm of the correct tRNA (**Movie 7.6**). The synthetases are thus equal in importance to the tRNAs in the decoding process, because it is the combined action of the synthetases and tRNAs that allows each codon in the mRNA molecule to specify its proper amino acid (**Figure 7–30**).

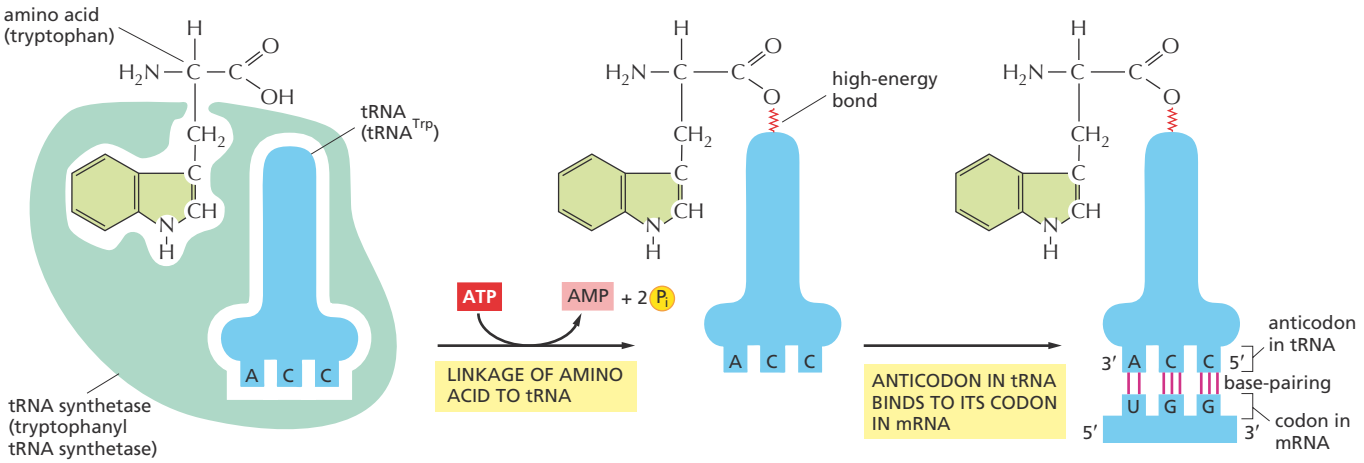


Figure 7–30 The genetic code is translated by the cooperation of two adaptors: aminoacyl-tRNA synthetases and tRNAs. Each synthetase couples a particular amino acid to its corresponding tRNAs, a process called charging. The anticodon on the charged tRNA molecule then forms base pairs with the appropriate codon on the mRNA. An error in either the charging step or the binding of the charged tRNA to its codon will cause the wrong amino acid to be incorporated into a protein chain. In the sequence of events shown, the amino acid tryptophan (Trp) is selected by the codon UGG on the mRNA.

The synthetase-catalyzed reaction that attaches the amino acid to the 3′ end of the tRNA is one of many reactions in cells coupled to the energy-releasing hydrolysis of ATP (see Figure 3–33). The reaction produces a high-energy bond between the charged tRNA and the amino acid. The energy of this bond is later used to link the amino acid covalently to the growing polypeptide chain.

The mRNA Message Is Decoded by Ribosomes

The recognition of a codon by the anticodon on a tRNA molecule depends on the same type of complementary base-pairing used in DNA replication and transcription. However, accurate and rapid translation of mRNA into protein requires a molecular machine that can move along the mRNA, capture complementary tRNA molecules, hold the tRNAs in position, and then covalently link the amino acids that they carry to form a polypeptide chain. In both prokaryotes and eukaryotes, the machine that gets the job done is the **ribosome**—a large complex made from dozens of small proteins (the *ribosomal proteins*) and several crucial RNA molecules called **ribosomal RNAs** (**rRNAs**). A typical eukaryotic cell contains millions of ribosomes in its cytoplasm (**Figure 7–31**).

Eukaryotic and prokaryotic ribosomes are very similar in structure and function. Both are composed of one large subunit and one small subunit, which fit together to form a complete ribosome with a mass of several million daltons (**Figure 7–32**); for comparison, an average-sized protein has a mass of 30,000 daltons. The small ribosomal subunit matches the tRNAs to the codons of the mRNA, while the large subunit catalyzes the formation of the peptide bonds that covalently link the amino acids together into a polypeptide chain. These two subunits come together on an mRNA molecule near its 5′ end to start the synthesis of a protein. The mRNA is then pulled through the ribosome like a long piece of tape. As the mRNA inches forward in a 5′-to-3′ direction, the ribosome translates its nucleotide sequence into an amino acid sequence, one codon at a time, using the tRNAs as adaptors. Each amino acid is thereby added in the correct sequence to the end of the growing polypeptide chain (**Movie 7.7**). When synthesis of the protein is finished, the two subunits of the ribosome separate. Ribosomes operate with remarkable efficiency: a eukaryotic ribosome adds about 2 amino acids to a polypeptide chain each second; a bacterial ribosome operates even faster, adding about 20 amino acids per second.

QUESTION 7–4

In a clever experiment performed in 1962, a cysteine already attached to its tRNA was chemically converted to an alanine. These "hybrid" tRNA molecules were then added to a cell-free translation system from which the normal cysteine-tRNAs had been removed. When the resulting protein was analyzed, it was found that alanine had been inserted at every point in the polypeptide chain where cysteine was supposed to be. Discuss what this experiment tells you about the role of aminoacyl-tRNA synthetases during the normal translation of the genetic code.

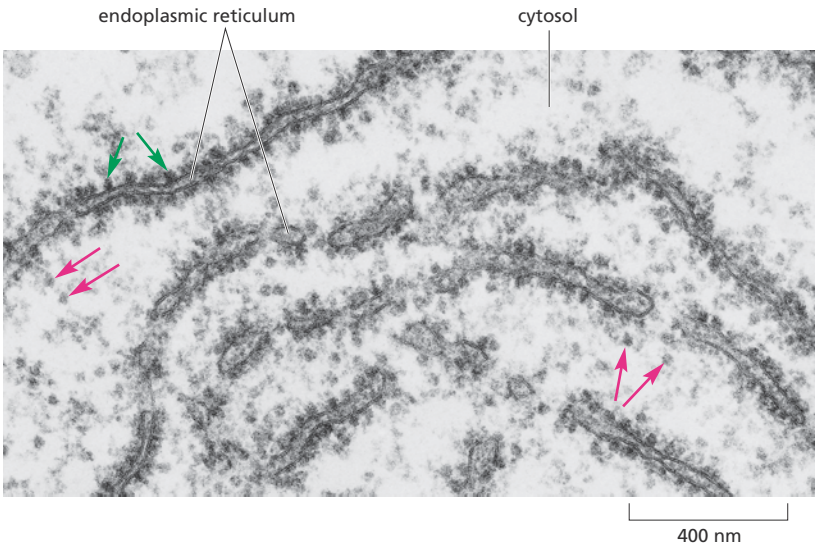


Figure 7–31 Ribosomes are located in the cytoplasm of eukaryotic cells. This electron micrograph shows a thin section of a small region of cytoplasm. The ribosomes appear as small gray blobs. Some are free in the cytosol (*red arrows*); others are attached to membranes of the endoplasmic reticulum (*green arrows*). (Courtesy of George Palade.)

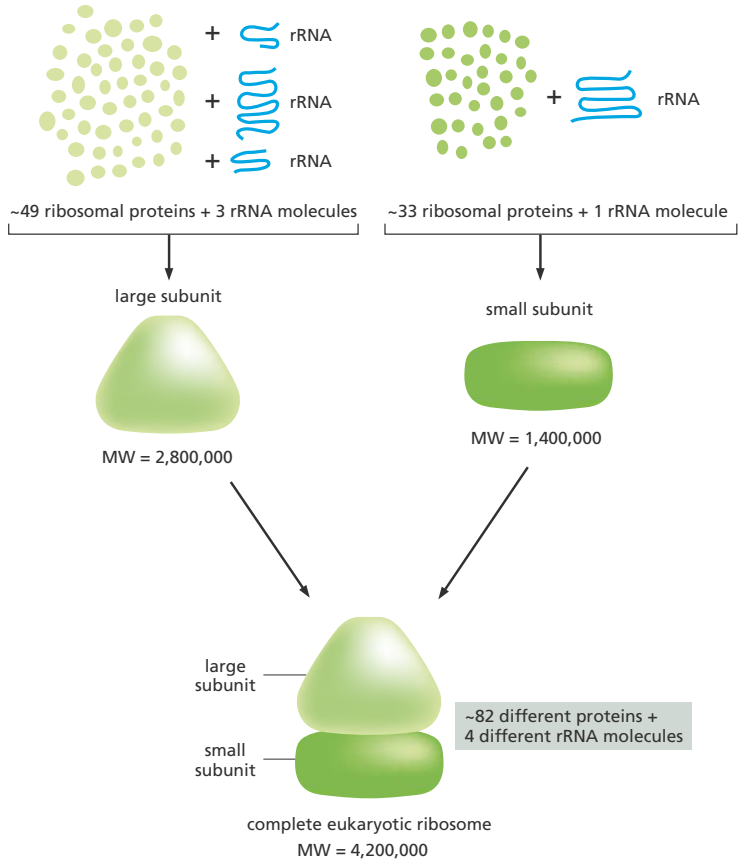


Figure 7–32 The eukaryotic ribosome is a large complex of four rRNAs and more than 80 small proteins. Prokaryotic ribosomes are very similar: both are formed from a large and small subunit, which only come together after the small subunit has bound an mRNA. Although ribosomal proteins greatly outnumber rRNAs, the RNAs account for most of the mass of the ribosome and give it its overall shape and structure.

How does the ribosome choreograph all the movements required for translation? In addition to a binding site for an mRNA molecule, each ribosome contains three binding sites for tRNA molecules, called the A site, the P site, and the E site (**Figure 7–33**). To add an amino acid to a growing peptide chain, the appropriate charged tRNA enters the A site by base-pairing with the complementary codon on the mRNA molecule. Its amino acid is then linked to the peptide chain held by the tRNA in the neighboring P site. Next, the large ribosomal subunit shifts forward, moving the spent tRNA to the E site before ejecting it (**Figure 7–34**). This cycle of reactions is repeated each time an amino acid is added to the polypeptide chain, with the new protein growing from its amino to its carboxyl end until a stop codon in the mRNA is encountered.

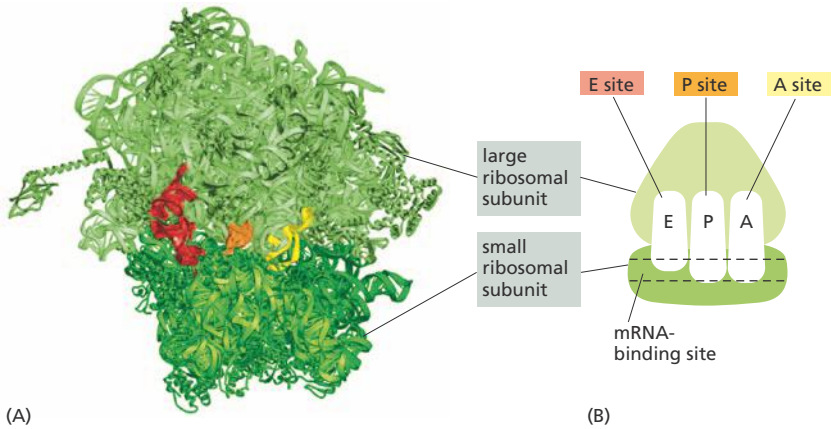


Figure 7–33 Each ribosome has a binding site for mRNA and three binding sites for tRNA. The tRNA sites are designated the A, P, and E sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively). (A) Three-dimensional structure of a bacterial ribosome, as determined by X-ray crystallography, with the small subunit in *dark green* and the large subunit in *light green*. Both the rRNAs and the ribosomal proteins are shown in *green*. tRNAs are shown bound in the E site (*red*), the P site (*orange*), and the A site (*yellow*). Although all three tRNA sites are shown occupied here, during the process of protein synthesis only two of these sites are occupied at any one time (see Figure 7–34). (B) Highly schematized representation of a ribosome (in the same orientation as A), which will be used in subsequent figures. Note that both the large and small subunits are involved in forming the A, P, and E sites, while only the small subunit forms the binding site for an mRNA. (B, adapted from M.M. Yusupov et al., *Science* 292:883–896, 2001, with permission from AAAS. Courtesy of Albion Baucom and Harry Noller.)

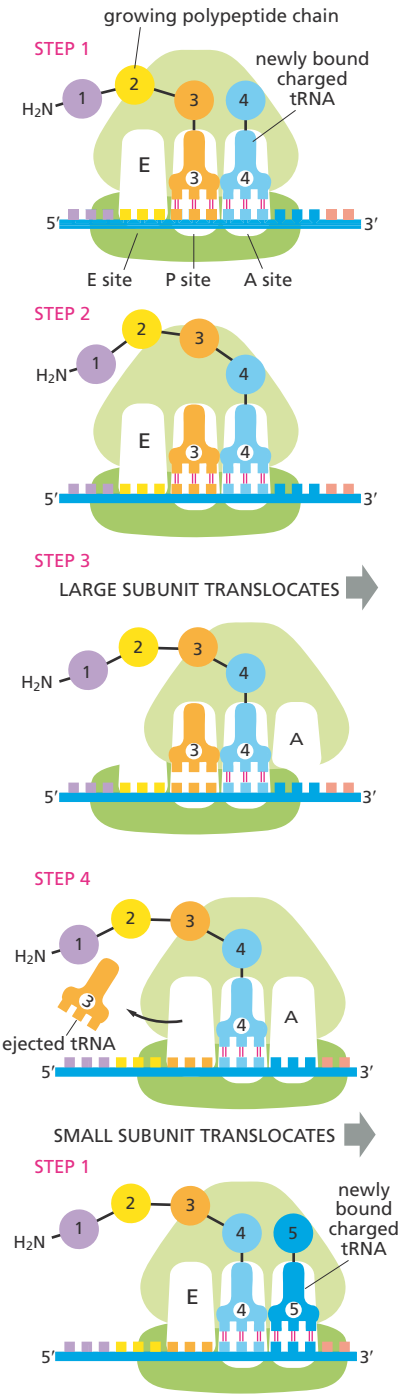


Figure 7–34 Translation takes place in a four-step cycle. This cycle is repeated over and over during the synthesis of a protein. In *step 1*, a charged tRNA carrying the next amino acid to be added to the polypeptide chain binds to the vacant A site on the ribosome by forming base pairs with the mRNA codon that is exposed there. Because only the appropriate tRNA molecules can base-pair with each codon, this codon determines the specific amino acid added. The A and P sites are sufficiently close together that their two tRNA molecules are forced to form base pairs with codons that are contiguous, with no stray bases in between. This positioning of the tRNAs ensures that the correct reading frame will be preserved throughout the synthesis of the protein. In *step 2*, the carboxyl end of the polypeptide chain (amino acid 3 in step 1) is uncoupled from the tRNA at the P site and joined by a peptide bond to the free amino group of the amino acid linked to the tRNA at the A site. This reaction is catalyzed by an enzymatic site in the large subunit. In *step 3*, a shift of the large subunit relative to the small subunit moves the two tRNAs into the E and P sites of the large subunit. In *step 4*, the small subunit moves exactly three nucleotides along the mRNA molecule, bringing it back to its original position relative to the large subunit. This movement ejects the spent tRNA and resets the ribosome with an empty A site so that the next charged tRNA molecule can bind (**Movie 7.8**).

As indicated, the mRNA is translated in the 5′-to-3′ direction, and the N-terminal end of a protein is made first, with each cycle adding one amino acid to the C-terminus of the polypeptide chain. To watch the translation cycle in atomic detail, see **Movie 7.9**.

The Ribosome Is a Ribozyme

The ribosome is one of the largest and most complex structures in the cell, composed of two-thirds RNA and one-third protein by weight. The determination of the entire three-dimensional structure of its large and small subunits in 2000 was a major triumph of modern biology. The structure confirmed earlier evidence that the rRNAs—not the proteins—are responsible for the ribosome's overall structure and its ability to choreograph and catalyze protein synthesis.

The rRNAs are folded into highly compact, precise three-dimensional structures that form the core of the ribosome (**Figure 7–35**). In marked contrast to the central positioning of the rRNAs, the ribosomal proteins are generally located on the surface, where they fill the gaps and crevices of the folded RNA. The main role of the ribosomal proteins seems to be

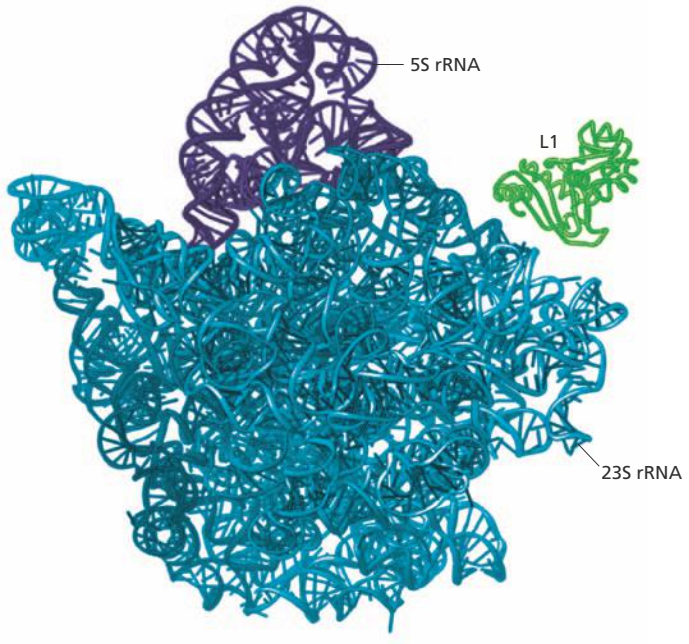


Figure 7–35 Ribosomal RNAs give the ribosome its overall shape. Shown here are the detailed structures of the two rRNAs that form the core of the large subunit of a bacterial ribosome—the 23S rRNA (*blue*) and the 5S rRNA (*purple*). One of the protein subunits of the ribosome (L1) is included as a reference point, as this protein forms a characteristic protrusion on the ribosome surface. Ribosomal components are commonly designated by their "S values," which refer to their rate of sedimentation in an ultracentrifuge. (Adapted from N. Ban et al., *Science* 289:905–920, 2000. With permission from AAAS.)

to help fold and stabilize the RNA core, while permitting the changes in rRNA conformation that are necessary for this RNA to catalyze efficient protein synthesis.

Not only are the three tRNA-binding sites (the A, P, and E sites) on the ribosome formed primarily by the rRNAs, but the catalytic site for peptide bond formation is formed by the 23S rRNA of the large subunit; the nearest ribosomal protein is located too far away to make contact with the incoming charged tRNA or with the growing polypeptide chain. The catalytic site in this rRNA—a peptidyl transferase—is similar in many respects to that found in some protein enzymes: it is a highly structured pocket that precisely orients the two reactants—the elongating polypeptide and the charged tRNA—thereby greatly increasing the probability of a productive reaction.

RNA molecules that possess catalytic activity are called **ribozymes**. Later, in the final section of this chapter, we will consider other ribozymes and discuss what the existence of RNA-based catalysis might mean for the early evolution of life on Earth. Here we need only note that there is good reason to suspect that RNA rather than protein molecules served as the first catalysts for living cells. If so, the ribosome, with its catalytic RNA core, could be viewed as a relic of an earlier time in life's history, when cells were run almost entirely by ribozymes.

Specific Codons in mRNA Signal the Ribosome Where to Start and to Stop Protein Synthesis

In the test tube, ribosomes can be forced to translate any RNA molecule (see How We Know, pp. 240–241). In a cell, however, a specific start signal is required to initiate translation. The site at which protein synthesis begins on an mRNA is crucial, because it sets the reading frame for the whole length of the message. An error of one nucleotide either way at this stage will cause every subsequent codon in the mRNA to be misread, resulting in a nonfunctional protein with a garbled sequence of amino acids (see Figure 7–26). And the rate of initiation determines the rate at which the protein is synthesized from the mRNA.

The translation of an mRNA begins with the codon AUG, and a special charged tRNA is required to initiate translation. This **initiator tRNA** always carries the amino acid methionine (or a modified form of methionine, formyl-methionine, in bacteria). Thus newly made proteins all have methionine as the first amino acid at their N-terminal end, the end of a protein that is synthesized first. This methionine is usually removed later by a specific protease.

In eukaryotes, an initiator tRNA, charged with methionine, is first loaded into the P site of the small ribosomal subunit, along with additional proteins called **translation initiation factors** (**Figure 7–36**). The initiator tRNA is distinct from the tRNA that normally carries methionine. Of all the tRNAs in the cell, only a charged initiator tRNA molecule is capable of binding tightly to the P site in the absence of the large ribosomal subunit. Next, the small ribosomal subunit loaded with the initiator tRNA binds to

QUESTION 7–5

A sequence of nucleotides in a DNA strand—5′-TTAACGGCTTTTTTC-3′—was used as a template to synthesize an mRNA that was then translated into protein. Predict the C-terminal amino acid and the N-terminal amino acid of the resulting polypeptide. Assume that the mRNA is translated without the need for a start codon.

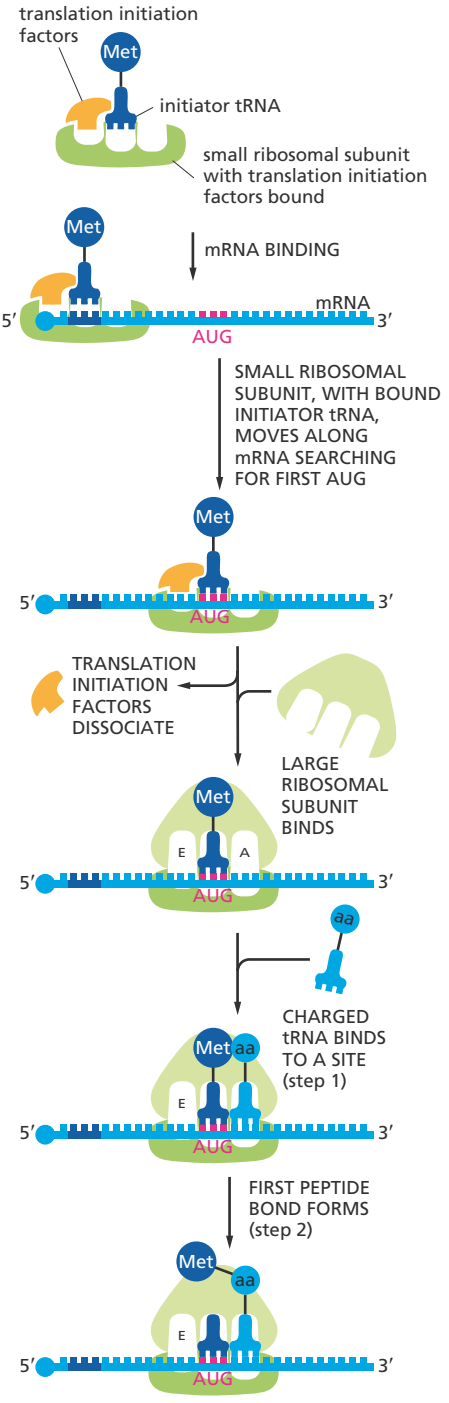


Figure 7–36 Initiation of protein synthesis in eukaryotes requires translation initiation factors and a special initiator tRNA. Although not shown here, efficient translation initiation also requires additional proteins that are bound at the 5′ cap and poly-A tail of the mRNA (see Figure 7–23). In this way, the translation apparatus can ascertain that both ends of the mRNA are intact before initiating translation. Following initiation, the protein is elongated by the reactions outlined in Figure 7–34.

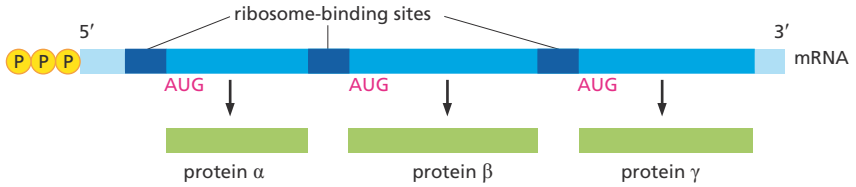
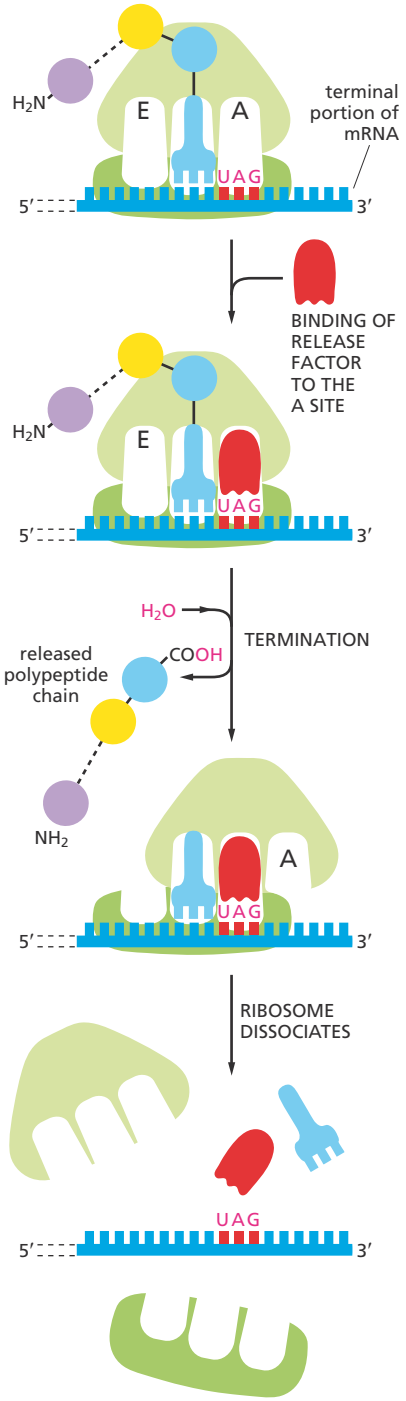


Figure 7–37 A single prokaryotic mRNA molecule can encode several different proteins. In prokaryotes, genes directing the different steps in a process are often organized into clusters (operons) that are transcribed together into a single mRNA. A prokaryotic mRNA does not have the same sort of 5′ cap as a eukaryotic mRNA, but instead has a triphosphate at its 5′ end. Prokaryotic ribosomes initiate translation at ribosome-binding sites (*dark blue*), which can be located in the interior of an mRNA molecule. This feature enables prokaryotes to synthesize different proteins from a single mRNA molecule, with each protein made by a different ribosome.



the 5′ end of an mRNA molecule, which is marked by the 5′ cap that is present on all eukaryotic mRNAs (see Figure 7–16). The small ribosomal subunit then moves forward (5′ to 3′) along the mRNA searching for the first AUG. When this AUG is encountered and recognized by the initiator tRNA, several initiation factors dissociate from the small ribosomal subunit to make way for the large ribosomal subunit to bind and complete ribosomal assembly. Because the initiator tRNA is bound to the P site, protein synthesis is ready to begin with the addition of the next charged tRNA to the A site (see Figure 7–34).

The mechanism for selecting a start codon is different in bacteria. Bacterial mRNAs have no 5′ caps to tell the ribosome where to begin searching for the start of translation. Instead, they contain specific ribosome-binding sequences, up to six nucleotides long, that are located a few nucleotides upstream of the AUGs at which translation is to begin. Unlike a eukaryotic ribosome, a prokaryotic ribosome can readily bind directly to a start codon that lies in the interior of an mRNA, as long as a ribosome-binding site precedes it by several nucleotides. Such ribosome-binding sequences are necessary in bacteria, as prokaryotic mRNAs are often *polycistronic*—that is, they encode several different proteins, each of which is translated from the same mRNA molecule (**Figure 7–37**). In contrast, a eukaryotic mRNA usually carries the information for a single protein.

The end of translation in both prokaryotes and eukaryotes is signaled by the presence of one of several codons, called *stop codons*, in the mRNA (see Figure 7–25). The stop codons—UAA, UAG, and UGA—are not recognized by a tRNA and do not specify an amino acid, but instead signal to the ribosome to stop translation. Proteins known as *release factors* bind to any stop codon that reaches the A site on the ribosome; this binding alters the activity of the peptidyl transferase in the ribosome, causing it to catalyze the addition of a water molecule instead of an amino acid to the peptidyl-tRNA (**Figure 7–38**). This reaction frees the carboxyl end of the polypeptide chain from its attachment to a tRNA molecule; because this is the only attachment that holds the growing polypeptide to the ribosome, the completed protein chain is immediately released. At this point, the ribosome also releases the mRNA and dissociates into its two separate subunits, which can then assemble on another mRNA molecule to begin a new round of protein synthesis.

Figure 7–38 Translation halts at a stop codon. In the final phase of protein synthesis, the binding of release factor to an A site bearing a stop codon terminates translation of an mRNA molecule. The completed polypeptide is released, and the ribosome dissociates into its two separate subunits. Note that only the 3′ end of the mRNA molecule is shown here.

We saw in Chapter 4 that many proteins can fold into their three-dimensional shape spontaneously, and some do so as they are spun out of the ribosome. Most proteins, however, require *chaperone proteins* to help them fold correctly in the cell. Chaperones can "steer" proteins along productive folding pathways and prevent them from aggregating inside the cell (see Figures 4–9 and 4–10). Newly synthesized proteins are typically met by their chaperones as they emerge from the ribosome.

Proteins Are Made on Polyribosomes

The synthesis of most protein molecules takes between 20 seconds and several minutes. But even during this short period, multiple ribosomes usually bind to each mRNA molecule being translated. If the mRNA is being translated efficiently, a new ribosome hops onto the 5′ end of the mRNA molecule almost as soon as the preceding ribosome has translated enough of the nucleotide sequence to move out of the way. The mRNA molecules being translated are therefore usually found in the form of *polyribosomes*, also known as *polysomes*. These large cytoplasmic assemblies are made up of many ribosomes spaced as close as 80 nucleotides apart along a single mRNA molecule (**Figure 7–39**). With multiple ribosomes working simultaneously on a single mRNA, many more protein molecules can be made in a given time than would be possible if each polypeptide had to be completed before the next could be started.

Polysomes operate in both bacteria and eukaryotes, but bacteria can speed up the rate of protein synthesis even further. Because bacterial mRNA does not need to be processed and is also physically accessible to ribosomes while it is being made, ribosomes will typically attach to the free end of a bacterial mRNA molecule and start translating it even before the transcription of that RNA is complete; these ribosomes follow closely behind the RNA polymerase as it moves along DNA.

Inhibitors of Prokaryotic Protein Synthesis Are Used as Antibiotics

The ability to translate mRNAs accurately into proteins is a fundamental feature of all life on Earth. Although the ribosome and other molecules that carry out this complex task are very similar among organisms, we

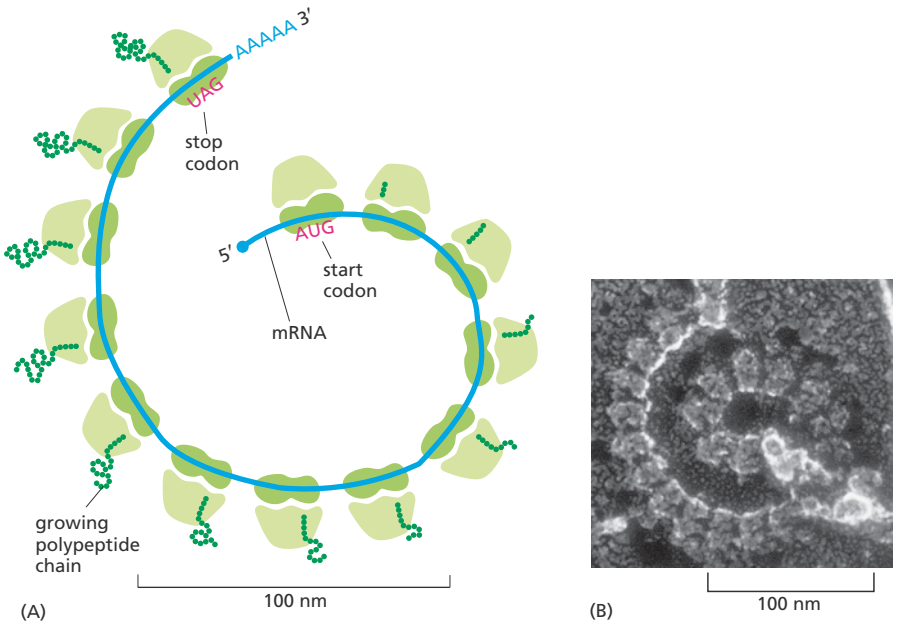


Figure 7–39 Proteins are synthesized on polyribosomes. (A) Schematic drawing showing how a series of ribosomes can simultaneously translate the same mRNA molecule (**Movie 7.10**). (B) Electron micrograph of a polyribosome in the cytosol of a eukaryotic cell. (B, courtesy of John Heuser.)

TABLE 7–3 ANTIBIOTICS THAT INHIBIT BACTERIAL PROTEIN OR RNA SYNTHESIS

Antibiotic	Specific Effect
Tetracycline	blocks binding of aminoacyl-tRNA to A site of ribosome (step 1 in Figure 7–34)
Streptomycin	prevents the transition from initiation complex to chain elongation (see Figure 7–36); also causes miscoding
Chloramphenicol	blocks the peptidyl transferase reaction on ribosomes (step 2 in Figure 7–34)
Cycloheximide	blocks the translocation reaction on ribosomes (step 3 in Figure 7–34)
Rifamycin	blocks initiation of transcription by binding to RNA polymerase

have seen that there are some subtle differences in the way that bacteria and eukaryotes synthesize RNA and proteins. Through a quirk of evolution, these differences form the basis of one of the most important advances in modern medicine.

Many of our most effective antibiotics are compounds that act by inhibiting bacterial, but not eukaryotic, RNA and protein synthesis. Some of these drugs exploit the small structural and functional differences between bacterial and eukaryotic ribosomes, so that they interfere preferentially with bacterial protein synthesis. These compounds can thus be taken in doses high enough to kill bacteria without being toxic to humans. Because different antibiotics bind to different regions of the bacterial ribosome, these drugs often inhibit different steps in protein synthesis. A few of the antibiotics that inhibit bacterial RNA and protein synthesis are listed in **Table 7–3**.

Many common antibiotics were first isolated from fungi. Fungi and bacteria often occupy the same ecological niches; to gain a competitive edge, fungi have evolved, over time, potent toxins that kill bacteria but are harmless to themselves. Because fungi and humans are both eukaryotes, and are thus more closely related to each other than either is to bacteria (see Figure 1–28), we have been able to borrow these weapons to combat our own bacterial foes.

Controlled Protein Breakdown Helps Regulate the Amount of Each Protein in a Cell

After a protein is released from the ribosome, a cell can control its activity and longevity in various ways. The number of copies of a protein in a cell depends, like the human population, not only on how quickly new individuals are made but also on how long they survive. So controlling the breakdown of proteins into their constituent amino acids helps cells regulate the amount of each particular protein. Proteins vary enormously in their life-span. Structural proteins that become part of a relatively stable tissue such as bone or muscle may last for months or even years, whereas other proteins, such as metabolic enzymes and those that regulate cell growth and division (discussed in Chapter 18), last only for days, hours, or even seconds. How does the cell control these lifetimes?

Cells possess specialized pathways that enzymatically break proteins down into their constituent amino acids (a process termed *proteolysis*). The enzymes that degrade proteins, first to short peptides and finally to individual amino acids, are known collectively as **proteases**. Proteases

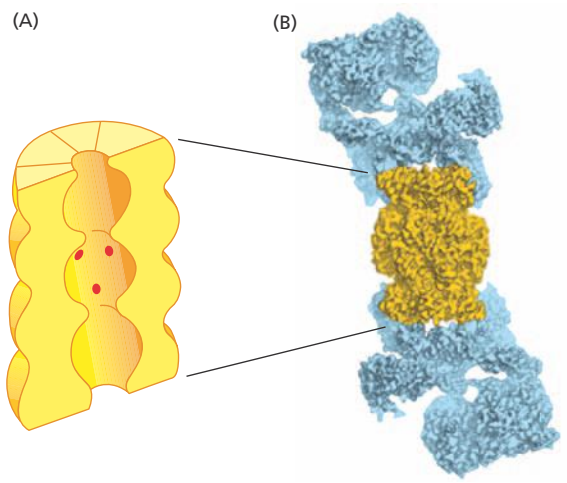


Figure 7–40 A proteasome degrades short-lived and misfolded proteins. The structures shown were determined by X-ray crystallography. (A) A cut-away view of the central cylinder of the proteasome, with the active sites of the proteases indicated by *red dots*. (B) The structure of the entire proteasome, in which access to the central cylinder (*yellow*) is regulated by a stopper (*blue*) at each end. (B, adapted from P.C.A da Fonseca et al., *Mol. Cell* 46:54–66, 2012.)

act by cutting (hydrolyzing) the peptide bonds between amino acids (see Panel 2–5, pp. 74–75). One function of proteolytic pathways is to rapidly degrade those proteins whose lifetimes must be kept short. Another is to recognize and remove proteins that are damaged or misfolded. Eliminating improperly folded proteins is critical for an organism, as misfolded proteins tend to aggregate, and protein aggregates can damage cells and even trigger cell death. Eventually, all proteins—even long-lived ones—accumulate damage and are degraded by proteolysis.

In eukaryotic cells, proteins are broken down by large protein machines called **proteasomes**, present in both the cytosol and the nucleus. A proteasome contains a central cylinder formed from proteases whose active sites face into an inner chamber. Each end of the cylinder is stoppered by a large protein complex formed from at least 10 types of protein subunits (**Figure 7–40**). These protein stoppers bind the proteins destined for degradation and then—using ATP hydrolysis to fuel this activity—unfold the doomed proteins and thread them into the inner chamber of the cylinder. Once the proteins are inside, proteases chop them into short peptides, which are then jettisoned from either end of the proteasome. Housing proteases inside these molecular destruction chambers makes sense, as it prevents the enzymes from running rampant in the cell.

How do proteasomes select which proteins in the cell should be degraded? In eukaryotes, proteasomes act primarily on proteins that have been marked for destruction by the covalent attachment of a small protein called *ubiquitin*. Specialized enzymes tag selected proteins with a short chain of ubiquitin molecules; these ubiquitylated proteins are then recognized, unfolded, and fed into proteasomes by proteins in the stopper (**Figure 7–41**).

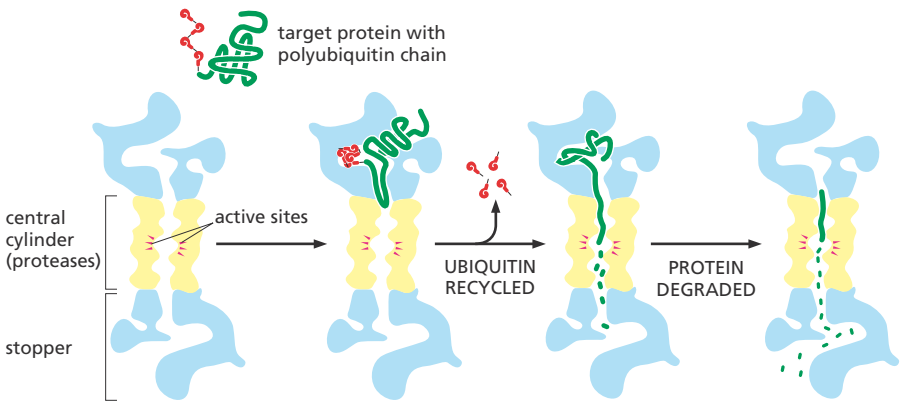


Figure 7–41 Proteins marked by a polyubiquitin chain are degraded by the proteasome. Proteins in the stopper of a proteasome (*blue*) recognize target proteins marked by a specific type of polyubiquitin chain. The stopper then unfolds the target protein and threads it into the proteasome's central cylinder (*yellow*), which is lined with proteases that chop the protein to pieces.

Proteins that are meant to be short-lived often contain a short amino acid sequence that identifies the protein as one to be ubiquitylated and degraded in proteasomes. Damaged or misfolded proteins, as well as proteins containing oxidized or otherwise abnormal amino acids, are also recognized and degraded by this ubiquitin-dependent proteolytic system. The enzymes that add a polyubiquitin chain to such proteins recognize signals that become exposed on these proteins as a result of the misfolding or chemical damage—for example, amino acid sequences or conformational motifs that remain buried and inaccessible in the normal "healthy" protein.

There Are Many Steps Between DNA and Protein

We have seen that many types of chemical reactions are required to produce a protein from the information contained in a gene. The final concentration of a protein in a cell therefore depends on the rate at which each of the many steps is carried out (**Figure 7–42**). In addition, many proteins—once they leave the ribosome—require further attention before they are useful to the cell. Examples of such *post-translational modifications* include covalent modification (such as phosphorylation), the binding of small-molecule cofactors, or association with other protein subunits, which are often needed for a newly synthesized protein to become fully functional (**Figure 7–43**).

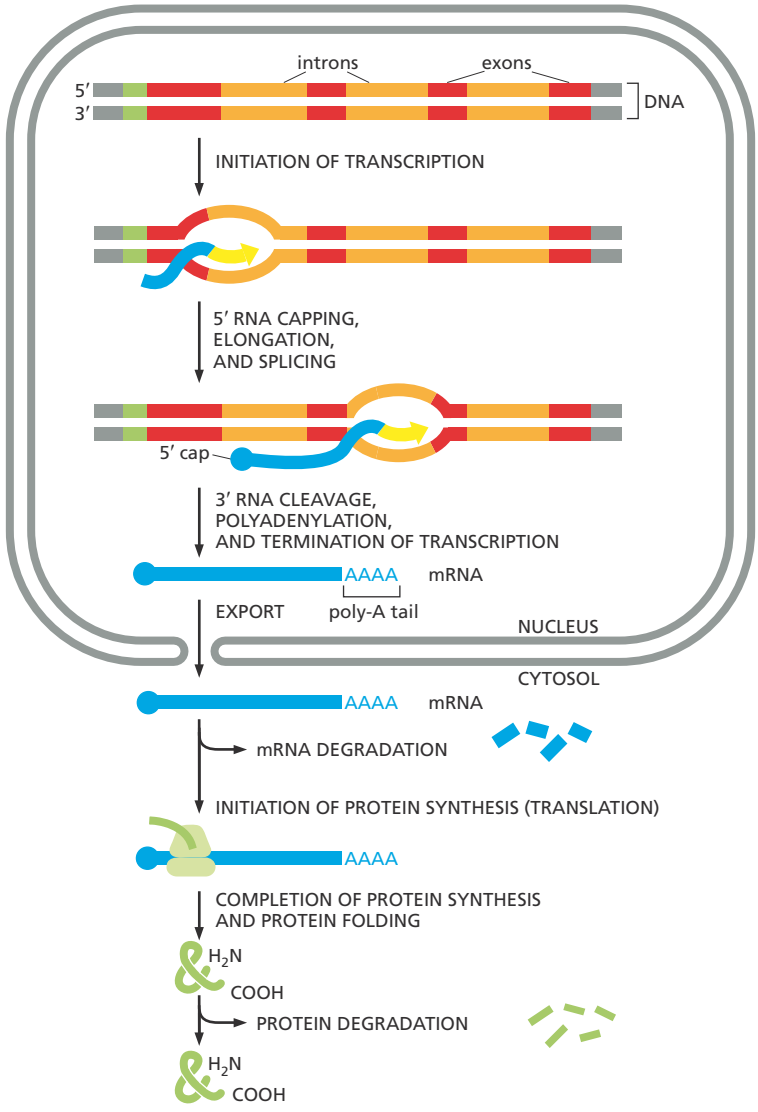


Figure 7–42 Protein production in a eukaryotic cell requires many steps. The final concentration of each protein depends on the rate of each step depicted. Even after an mRNA and its corresponding protein have been produced, their concentrations can be regulated by degradation. Although not shown here, the activity of the protein can also be regulated by other post-translational modifications or the binding of small molecules (see Figure 7–43).

We will see in the next chapter that cells have the ability to change the concentrations of most of their proteins according to their needs. In principle, all of the steps in Figure 7–42 can be regulated by the cell—and many of them, in fact, are. However, as we will see in the next chapter, the initiation of transcription is the most common point for a cell to regulate the expression of its genes.

Transcription and translation are universal processes that lie at the heart of life. However, when scientists came to consider how the flow of information from DNA to protein might have originated, they came to some unexpected conclusions.

RNA AND THE ORIGINS OF LIFE

The central dogma—that DNA makes RNA that makes protein—presented evolutionary biologists with a knotty puzzle: if nucleic acids are required to direct the synthesis of proteins, and proteins are required to synthesize nucleic acids, how could this system of interdependent components have arisen? One view is that an **RNA world** existed on Earth before cells containing DNA and proteins appeared. According to this hypothesis, RNA—which today serves largely as an intermediate between genes and proteins—both stored genetic information and catalyzed chemical reactions in primitive cells. Only later in evolutionary time did DNA take over as the genetic material and proteins become the major catalysts and structural components of cells (**Figure 7–44**). If this idea is correct, then the transition out of the RNA world was never completed; as we have seen, RNA still catalyzes several fundamental reactions in modern cells. These RNA catalysts—or ribozymes—including those that operate in the ribosome and in the RNA-splicing machinery, can thus be viewed as molecular fossils of an earlier world.

Life Requires Autocatalysis

The origin of life requires molecules that possess, if only to a small extent, one crucial property: the ability to catalyze reactions that lead—directly or indirectly—to the production of more molecules like themselves. Catalysts with this self-producing property, once they had arisen by chance, would divert raw materials from the production of other substances to make more of themselves. In this way, one can envisage the gradual development of an increasingly complex chemical system of organic monomers and polymers that function together to generate more molecules of the same types, fueled by a supply of simple raw materials in the primitive environment on Earth. Such an *autocatalytic* system would have many of the properties we think of as characteristic of living matter: the system would contain a far-from-random selection of interacting molecules; it would tend to reproduce itself; it would compete with other systems dependent on the same raw materials; and, if deprived of its raw materials or maintained at a temperature that upset the balance of reaction rates, it would decay toward chemical equilibrium and "die."

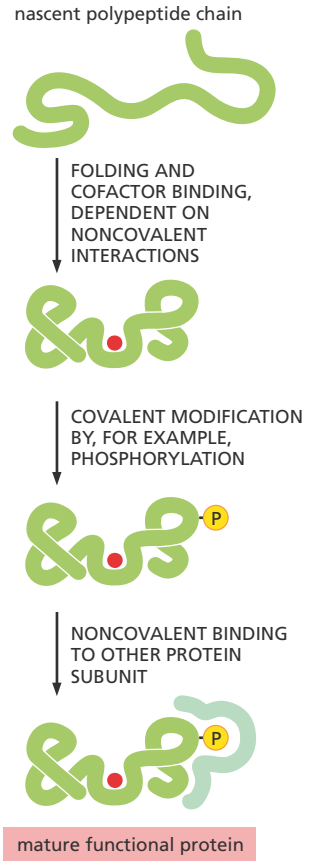


Figure 7–43 Many proteins require various modifications to become fully functional. To be useful to the cell, a completed polypeptide must fold correctly into its three-dimensional conformation and then bind any required cofactors (*red*) and protein partners—all via noncovalent bonding. Many proteins also require one or more covalent modifications to become active—or to be recruited to specific membranes or organelles (not shown). Although phosphorylation and glycosylation are the most common, more than 100 types of covalent modifications of proteins are known.

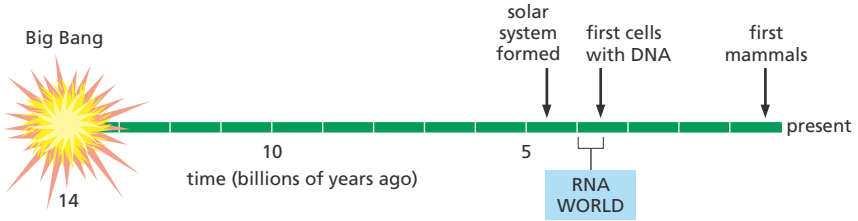


Figure 7–44 An RNA world may have existed before modern cells with DNA and proteins evolved.

But what molecules could have had such autocatalytic properties? In present-day living cells, the most versatile catalysts are proteins, which are able to adopt diverse three-dimensional forms that bristle with chemically reactive sites on their surface. However, there is no known way in which a protein can reproduce itself directly. RNA molecules, by contrast, could—at least, in principle—catalyze their own synthesis.

RNA Can Both Store Information and Catalyze Chemical Reactions

We have seen that complementary base-pairing enables one nucleic acid to act as a template for the formation of another. Thus a single strand of RNA or DNA can specify the sequence of a complementary polynucleotide, which, in turn, can specify the sequence of the original molecule, allowing the original nucleic acid to be replicated (**Figure 7–45**). Such complementary templating mechanisms lie at the heart of both DNA replication and transcription in modern-day cells.

But the efficient synthesis of polynucleotides by such complementary templating mechanisms also requires catalysts to promote the polymerization reaction: without catalysts, polymer formation is slow, error-prone, and inefficient. Today, nucleotide polymerization is catalyzed by protein enzymes—such as DNA and RNA polymerases. But how could this reaction be catalyzed before proteins with the appropriate catalytic ability existed? The beginnings of an answer were obtained in 1982, when it was discovered that RNA molecules themselves can act as catalysts. The unique potential of RNA molecules to act both as information carriers and as catalysts is thought to have enabled them to have a central role in the origin of life.

In present-day cells, RNA is synthesized as a single-stranded molecule, and we have seen that complementary base-pairing can occur between nucleotides in the same chain. This base-pairing, along with nonconventional hydrogen bonds, can cause each RNA molecule to fold up in a unique way that is determined by its nucleotide sequence (see Figure 7–5). Such associations produce complex three-dimensional shapes.

As we discuss in Chapter 4, protein enzymes are able to catalyze biochemical reactions because they have surfaces with unique contours and chemical properties. In the same way, RNA molecules, with their unique folded shapes, can serve as catalysts (**Figure 7–46**). RNAs do not have the same structural and functional diversity as do protein enzymes; they are, after all, built from only four different subunits. Nonetheless, ribozymes can catalyze many types of chemical reactions. Most of the ribozymes that have been studied were constructed in the laboratory and selected for their catalytic activity in a test tube (**Table 7–4**), as relatively few catalytic RNAs exist in present-day cells. But the processes in which catalytic RNAs still seem to have major roles include some of the most

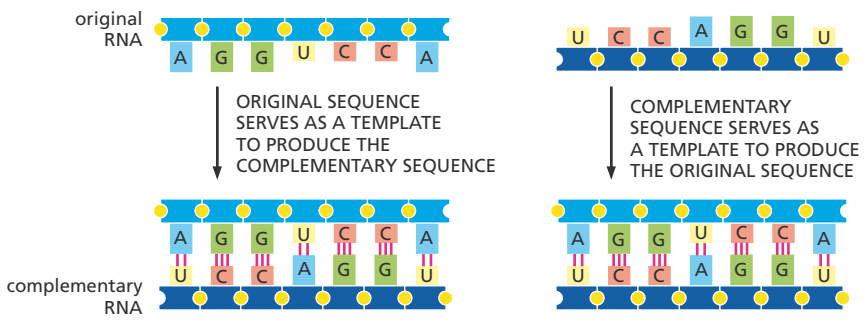


Figure 7–45 An RNA molecule can in principle guide the formation of an exact copy of itself. In the first step, the original RNA molecule acts as a template to form an RNA molecule of complementary sequence. In the second step, this complementary RNA molecule itself acts as a template to form an RNA molecule of the original sequence. Since each template molecule can produce many copies of the complementary strand, these reactions can result in the amplification of the original sequence.

TABLE 7–4 BIOCHEMICAL REACTIONS THAT CAN BE CATALYZED BY RIBOZYMES

Activity	Ribozymes
Peptide bond formation in protein synthesis	ribosomal RNA
DNA ligation	*in vitro* selected RNA
RNA splicing	self-splicing RNAs, small nuclear RNAs
RNA polymerization	*in vitro* selected RNA
RNA phosphorylation	*in vitro* selected RNA
RNA aminoacylation	*in vitro* selected RNA
RNA alkylation	*in vitro* selected RNA
C–C bond rotation (isomerization)	*in vitro* selected RNA

fundamental steps in the expression of genetic information—especially those steps where RNA molecules themselves are spliced or translated into protein.

RNA, therefore, has all the properties required of a molecule that could catalyze its own synthesis (**Figure 7–47**). Although self-replicating systems of RNA molecules have not been found in nature, scientists appear to be well on the way to constructing them in the laboratory. Although this demonstration would not prove that self-replicating RNA molecules were essential to the origin of life on Earth, it would establish that such a scenario is possible.

RNA Is Thought to Predate DNA in Evolution

The first cells on Earth would presumably have been much less complex and less efficient in reproducing themselves than even the simplest present-day cells. They would have consisted of little more than a simple membrane enclosing a set of self-replicating molecules and a few other components required to provide the materials and energy for this autocatalytic replication. If the evolutionary role for RNA proposed above is correct, these earliest cells would also have differed fundamentally from the cells we know today in having their hereditary information stored in RNA rather than DNA.

Evidence that RNA arose before DNA in evolution can be found in the chemical differences between them. Ribose (see Figure 7–3A), like

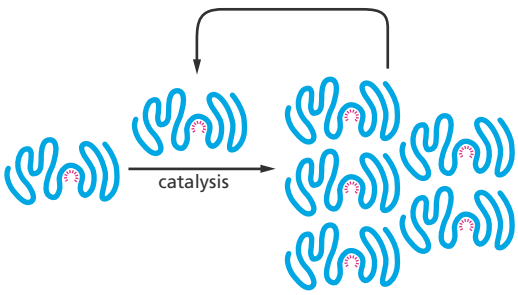


Figure 7–47 Could an RNA molecule catalyze its own synthesis? This hypothetical process would require that the RNA catalyze both steps shown in Figure 7–45. The *red* rays represent the active site of this ribozyme.

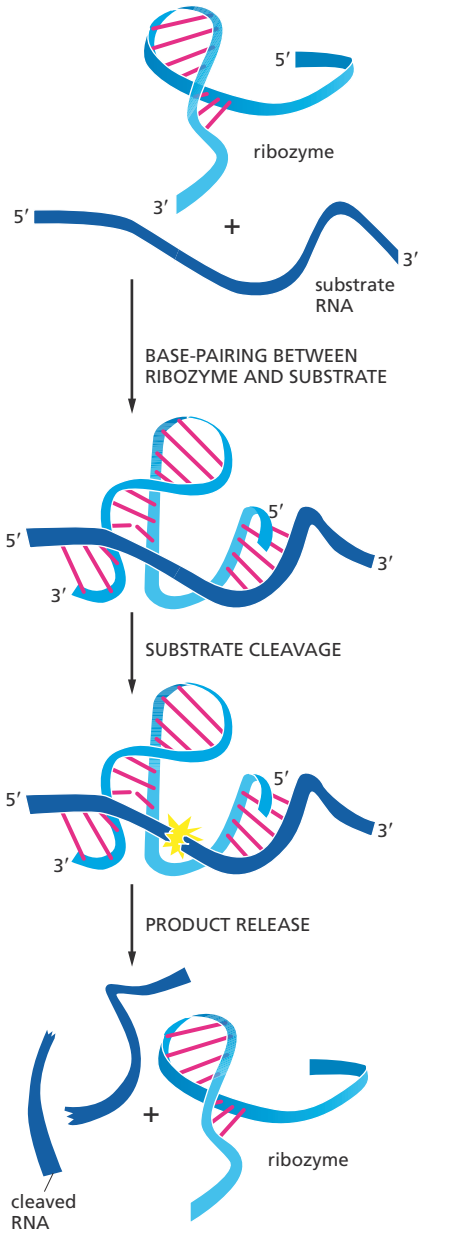


Figure 7–46 A ribozyme is an RNA molecule that possesses catalytic activity. The RNA molecule shown catalyzes the cleavage of a second RNA at a specific site. Similar ribozymes are found embedded in large RNA genomes—called viroids—that infect plants, where the cleavage reaction is one step in the replication of the viroid. (Adapted from T.R. Cech and O.C. Uhlenbeck, *Nature* 372:39–40, 1994. With permission from Macmillan Publishers Ltd.)

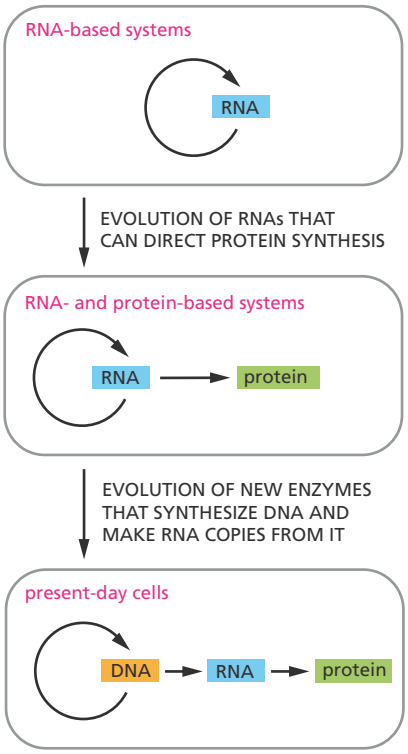


Figure 7–48 RNA may have preceded DNA and proteins in evolution. According to this hypothesis, RNA molecules provided genetic, structural, and catalytic functions in the earliest cells. DNA is now the repository of genetic information, and proteins carry out almost all catalysis in cells. RNA now functions mainly as a go-between in protein synthesis, while remaining a catalyst for a few crucial reactions (including protein synthesis).

QUESTION 7–6

Discuss the following: "During the evolution of life on Earth, RNA lost its glorious position as the first self-replicating catalyst. Its role now is as a mere messenger in the information flow from DNA to protein."

glucose and other simple carbohydrates, is readily formed from formaldehyde (HCHO), which is one of the principal products of experiments simulating conditions on the primitive Earth. The sugar deoxyribose is harder to make, and in present-day cells it is produced from ribose in a reaction catalyzed by a protein enzyme, suggesting that ribose predates deoxyribose in cells. Presumably, DNA appeared on the scene after RNA, and then proved more suited than RNA as a permanent repository of genetic information. In particular, the deoxyribose in its sugar–phosphate backbone makes chains of DNA chemically much more stable than chains of RNA, so that greater lengths of DNA can be maintained without breakage.

The other differences between RNA and DNA—the double-helical structure of DNA and the use of thymine rather than uracil—further enhance DNA stability by making the molecule easier to repair. We saw in Chapter 6 that a damaged nucleotide on one strand of the double helix can be repaired by using the other strand as a template. Furthermore, deamination, one of the most common unwanted chemical changes occurring in polynucleotides, is easier to detect and repair in DNA than in RNA (see Figure 6–23). This is because the product of the deamination of cytosine is, by chance, uracil, which already exists in RNA, so that such damage would be impossible for repair enzymes to detect in an RNA molecule. However, in DNA, which has thymine rather than uracil, any uracil produced by the accidental deamination of cytosine is easily detected and repaired.

Taken together, the evidence we have discussed supports the idea that RNA—with its ability to provide genetic, structural, and catalytic functions—preceded DNA in evolution. As cells more closely resembling present-day cells appeared, it is believed that many of the functions originally performed by RNA were taken over by DNA and proteins: DNA took over the primary genetic function, and proteins became the major catalysts, while RNA remained primarily as the intermediary connecting the two (**Figure 7–48**). With the advent of DNA, cells were able to become more complex, for they could then carry and transmit more genetic information than could be stably maintained by RNA alone. Because of the greater chemical complexity of proteins and the variety of chemical reactions they can catalyze, the shift (albeit incomplete) from RNA to proteins also provided a much richer source of structural components and enzymes. This enabled cells to evolve the great diversity of structure and function that we see in life today.

ESSENTIAL CONCEPTS

- **The flow of genetic information in all living cells is DNA → RNA → protein. The conversion of the genetic instructions in DNA into RNAs and proteins is termed gene expression.**
- **To express the genetic information carried in DNA, the nucleotide sequence of a gene is first transcribed into RNA. Transcription is catalyzed by the enzyme RNA polymerase, which uses nucleotide sequences in the DNA molecule to determine which strand to use as a template, and where to start and stop transcribing.**
- **RNA differs in several respects from DNA. It contains the sugar ribose instead of deoxyribose and the base uracil (U) instead of thymine (T). RNAs in cells are synthesized as single-stranded molecules, which often fold up into complex three-dimensional shapes.**
- **Cells make several functional types of RNAs, including messenger RNAs (mRNAs), which carry the instructions for making proteins; ribosomal RNAs (rRNAs), which are the crucial components of**

ribosomes; and transfer RNAs (tRNAs), which act as adaptor molecules in protein synthesis.

- To begin transcription, RNA polymerase binds to specific DNA sites called promoters that lie immediately upstream of genes. To initiate transcription, eukaryotic RNA polymerases require the assembly of a complex of general transcription factors at the promoter, whereas bacterial RNA polymerase requires only an additional subunit, called sigma factor.
- Most protein-coding genes in eukaryotic cells are composed of a number of coding regions, called exons, interspersed with larger noncoding regions, called introns. When a eukaryotic gene is transcribed from DNA into RNA, both the exons and introns are copied.
- Introns are removed from the RNA transcripts in the nucleus by RNA splicing, a reaction catalyzed by small ribonucleoprotein complexes known as snRNPs. Splicing removes the introns from the RNA and joins together the exons—often in a variety of combinations, allowing multiple proteins to be produced from the same gene.
- Eukaryotic pre-mRNAs go through several additional RNA processing steps before they leave the nucleus as mRNAs, including 5′ RNA capping and 3′ polyadenylation. These reactions, along with splicing, take place as the pre-mRNA is being transcribed.
- Translation of the nucleotide sequence of an mRNA into a protein takes place in the cytoplasm on large ribonucleoprotein assemblies called ribosomes. As the mRNA moves through the ribosome, its message is translated into protein.
- The nucleotide sequence in mRNA is read in sets of three nucleotides called codons; each codon corresponds to one amino acid.
- The correspondence between amino acids and codons is specified by the genetic code. The possible combinations of the 4 different nucleotides in RNA give 64 different codons in the genetic code. Most amino acids are specified by more than one codon.
- tRNAs act as adaptor molecules in protein synthesis. Enzymes called aminoacyl-tRNA synthetases covalently link amino acids to their appropriate tRNAs. Each tRNA contains a sequence of three nucleotides, the anticodon, which recognizes a codon in an mRNA through complementary base-pairing.
- Protein synthesis begins when a ribosome assembles at an initiation codon (AUG) in an mRNA molecule, a process that depends on proteins called translation initiation factors. The completed protein chain is released from the ribosome when a stop codon (UAA, UAG, or UGA) in the mRNA is reached.
- The stepwise linking of amino acids into a polypeptide chain is catalyzed by an rRNA molecule in the large ribosomal subunit, which thus acts as a ribozyme.
- The concentration of a protein in a cell depends on the rate at which the mRNA and protein are synthesized and degraded. Protein degradation in the cytosol and nucleus occurs inside large protein complexes called proteasomes.
- From our knowledge of present-day organisms and the molecules they contain, it seems likely that life on Earth began with the evolution of RNA molecules that could catalyze their own replication.
- It has been proposed that RNA served as both the genome and the catalysts in the first cells, before DNA replaced RNA as a more stable molecule for storing genetic information, and proteins replaced RNAs as the major catalytic and structural components. RNA catalysts in modern cells are thought to provide a glimpse into an ancient, RNA-based world.

KEY TERMS

alternative splicing
aminoacyl-tRNA synthetase
anticodon
codon
exon
gene
gene expression
general transcription factors
genetic code
initiator tRNA
intron
messenger RNA (mRNA)
polyadenylation
promoter
protease
proteasome
reading frame
ribosomal RNA (rRNA)
ribosome
ribozyme
RNA
RNA capping
RNA polymerase
RNA processing
RNA splicing
RNA transcript
RNA world
small nuclear RNA (snRNA)
spliceosome
transcription
transfer RNA (tRNA)
translation
translation initiation factor

QUESTIONS

QUESTION 7–7

Which of the following statements are correct? Explain your answers.

A. An individual ribosome can make only one type of protein.

B. All mRNAs fold into particular three-dimensional structures that are required for their translation.

C. The large and small subunits of an individual ribosome always stay together and never exchange partners.

D. Ribosomes are cytoplasmic organelles that are encapsulated by a single membrane.

E. Because the two strands of DNA are complementary, the mRNA of a given gene can be synthesized using either strand as a template.

F. An mRNA may contain the sequence **ATTGACCCCGGTCAA**.

G. The amount of a protein present in a cell depends on its rate of synthesis, its catalytic activity, and its rate of degradation.

QUESTION 7–8

The Lacheinmal protein is a hypothetical protein that causes people to smile more often. It is inactive in many chronically unhappy people. The mRNA isolated from a number of different unhappy individuals in the same family was found to lack an internal stretch of 173 nucleotides that is present in the Lacheinmal mRNA isolated from happy members of the same family. The DNA sequences of the *Lacheinmal* genes from the happy and unhappy family members were determined and compared. They differed by a single nucleotide substitution, which lay in an intron. What can you say about the molecular basis of unhappiness in this family?

(Hints: [1] Can you hypothesize a molecular mechanism by which a single nucleotide substitution in a gene could cause the observed deletion in the mRNA? Note that the deletion is *internal* to the mRNA. [2] Assuming the 173-base-pair deletion removes coding sequences from the Lacheinmal mRNA, how would the Lacheinmal protein differ between the happy and unhappy people?)

QUESTION 7–9

Use the genetic code shown in Figure 7–25 to identify which of the following nucleotide sequences would code for the polypeptide sequence arginine-glycine-aspartate:

1. 5′-AGA-GGA-GAU-3′
2. 5′-ACA-CCC-ACU-3′
3. 5′-GGG-AAA-UUU-3′
4. 5′-CGG-GGU-GAC-3′

QUESTION 7–10

"The bonds that form between the anticodon of a tRNA molecule and the three nucleotides of a codon in mRNA are _____." Complete this sentence with each of the following options and explain why each of the resulting statements is correct or incorrect.

A. Covalent bonds formed by GTP hydrolysis

B. Hydrogen bonds that form when the tRNA is at the A site

C. Broken by the translocation of the ribosome along the mRNA

QUESTION 7–11

List the ordinary, dictionary definitions of the terms *replication*, *transcription*, and *translation*. By their side, list the special meaning each term has when applied to the living cell.

QUESTION 7–12

In an alien world, the genetic code is written in pairs of nucleotides. How many amino acids could such a code specify? In a different world, a triplet code is used, but the sequence of nucleotides is not important; it only matters which nucleotides are present. How many amino acids could this code specify? Would you expect to encounter any problems translating these codes?

QUESTION 7–13

One remarkable feature of the genetic code is that amino acids with similar chemical properties often have similar codons. Thus codons with U or C as the second nucleotide tend to specify hydrophobic amino acids. Can you suggest a possible explanation for this phenomenon in terms of the early evolution of the protein-synthesis machinery?

QUESTION 7–14

A mutation in DNA generates a UGA stop codon in the middle of the mRNA coding for a particular protein. A second mutation in the cell's DNA leads to a single nucleotide change in a tRNA that allows the correct translation of the protein; that is, the second mutation "suppresses" the defect caused by the first. The altered tRNA translates the UGA as tryptophan. What nucleotide change has probably occurred in the mutant tRNA molecule? What consequences would the presence of such a mutant tRNA have for the translation of the normal genes in this cell?

QUESTION 7–15

The charging of a tRNA with an amino acid can be represented by the following equation:

amino acid + tRNA + ATP $\longrightarrow$ aminoacyl-tRNA + AMP + PP_i

where PP_i is pyrophosphate (see Figure 3–40). In the aminoacyl-tRNA, the amino acid and tRNA are linked with a high-energy covalent bond; a large portion of the energy derived from the hydrolysis of ATP is thus stored in this bond and is available to drive peptide bond formation at the later stages of protein synthesis. The free-energy change of the charging reaction shown in the equation is close to zero and therefore would not be expected to favor attachment of the amino acid to tRNA. Can you suggest a further step that could drive the reaction to completion?

QUESTION 7–16

A. The average molecular weight of a protein in the cell is about 30,000 daltons. A few proteins, however, are much larger. The largest known polypeptide chain made by any cell is a protein called titin (made by mammalian muscle cells), and it has a molecular weight of 3,000,000 daltons. Estimate how long it will take a muscle cell to translate an mRNA coding for titin (assume the average molecular weight of an amino acid to be 120, and a translation rate of two amino acids per second for eukaryotic cells).

B. Protein synthesis is very accurate: for every 10,000 amino acids joined together, only one mistake is made. What is the fraction of average-sized protein molecules and of titin molecules that are synthesized without any errors? (Hint: the probability P of obtaining an error-free protein is given by $P = (1 - E)^n$, where E is the error frequency and n the number of amino acids.)

C. The molecular weight of all eukaryotic ribosomal proteins combined is about 2.5×10^6 daltons. Would it be advantageous to synthesize them as a single protein?

D. Transcription occurs at a rate of about 30 nucleotides per second. Is it possible to calculate the time required to synthesize a titin mRNA from the information given here?

QUESTION 7–17

Which of the following types of mutations would be predicted to harm an organism? Explain your answers.

A. Insertion of a single nucleotide near the end of the coding sequence.

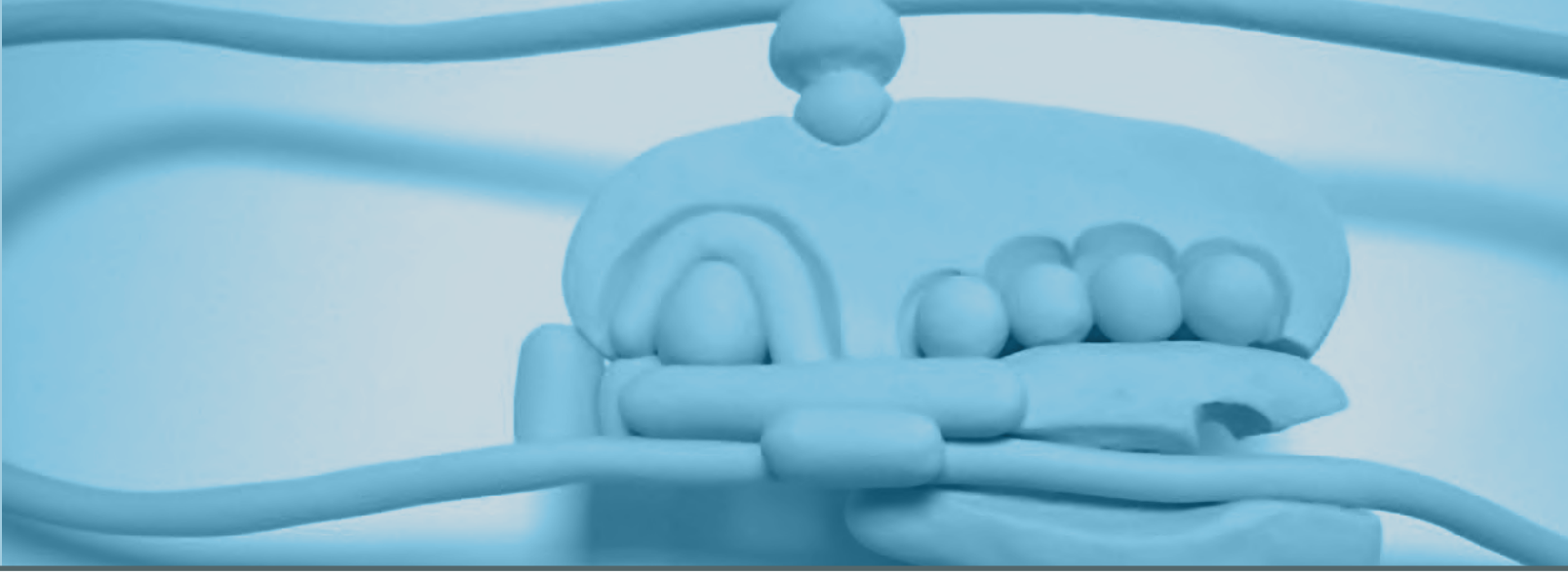
B. Removal of a single nucleotide near the beginning of the coding sequence.

C. Deletion of three consecutive nucleotides in the middle of the coding sequence.

D. Deletion of four consecutive nucleotides in the middle of the coding sequence.

E. Substitution of one nucleotide for another in the middle of the coding sequence.

Page left intentionally blank



CHAPTER **EIGHT** 8

Control of Gene Expression

An organism's DNA encodes all of the RNA and protein molecules that are needed to make its cells. Yet a complete description of the DNA sequence of an organism—be it the few million nucleotides of a bacterium or the few billion nucleotides in each human cell—does not enable us to reconstruct that organism any more than a list of all the English words in a dictionary enables us to reconstruct a play by Shakespeare. We need to know how the elements in the DNA sequence or the words on a list work together to make the masterpiece.

For cells, the question involves *gene expression*. Even the simplest single-celled bacterium can use its genes selectively—for example, switching genes on and off to make the enzymes needed to digest whatever food sources are available. In multicellular plants and animals, however, gene expression is under much more elaborate control. Over the course of embryonic development, a fertilized egg cell gives rise to many cell types that differ dramatically in both structure and function. The differences between an information-processing nerve cell and an infection-fighting white blood cell, for example, are so extreme that it is difficult to imagine that the two cells contain the same DNA (**Figure 8–1**). For this reason, and because cells in an adult organism rarely lose their distinctive characteristics, biologists originally suspected that certain genes might be selectively lost when a cell becomes specialized. We now know, however, that nearly all the cells of a multicellular organism contain the same genome. Cell *differentiation* is instead achieved by changes in gene expression.

In mammals, hundreds of different cell types carry out a range of specialized functions that depend upon genes that are switched on in that

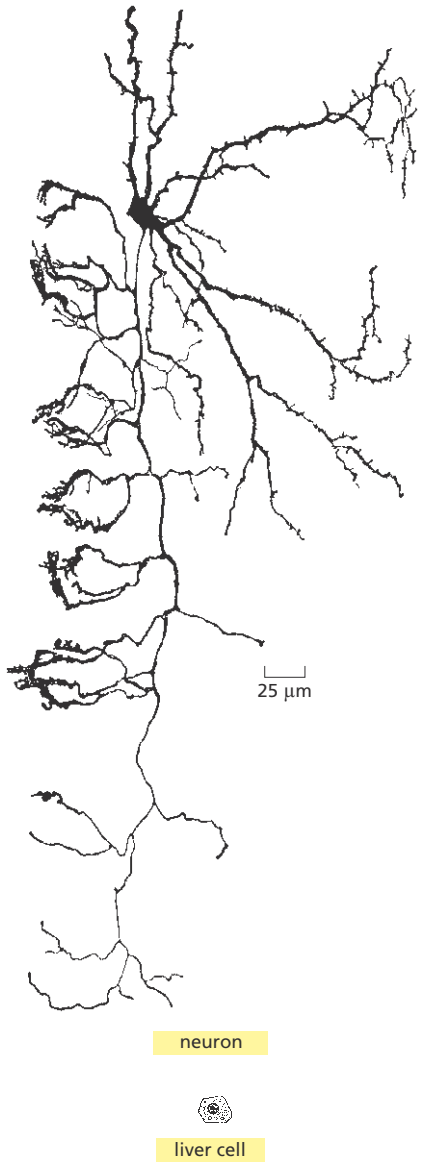


Figure 8–1 A neuron and a liver cell share the same genome. The long branches of this neuron from the retina enable it to receive electrical signals from many other neurons and carry them to many neighboring neurons. The liver cell, which is drawn to the same scale, is involved in many metabolic processes, including digestion and the detoxification of alcohol and other drugs. Both of these mammalian cells contain the same genome, but they express many different RNAs and proteins. (Neuron adapted from S. Ramón y Cajal, Histologie du Système Nerveux de l'Homme et de Vertébrés, 1909–1911. Paris: Maloine; reprinted, Madrid: C.S.I.C., 1972.)

cell type but not in most others: for example, the β cells of the pancreas make the protein hormone insulin, while the α cells of the pancreas make the hormone glucagon; the B lymphocytes of the immune system make antibodies, while developing red blood cells make the oxygen-transport protein hemoglobin. The differences between a neuron, a white blood cell, a pancreatic β cell, and a red blood cell depend upon the precise control of gene expression. A typical differentiated cell expresses only about half the genes in its total repertoire.

In this chapter, we discuss the main ways in which gene expression is regulated, with a focus on those genes that encode proteins as their final product. Although some of these control mechanisms apply to both eukaryotes and prokaryotes, eukaryotic cells—with their more complex chromosomal structure—have some ways of controlling gene expression that are not available to bacteria.

AN OVERVIEW OF GENE EXPRESSION

Gene expression is a complex process by which cells selectively direct the synthesis of the many thousands of proteins and RNAs encoded in their genome. But how do cells coordinate and control such an intricate process—and how does an individual cell specify which of its genes to express? This decision is an especially important problem for animals because, as they develop, their cells become highly specialized, ultimately producing an array of muscle, nerve, and blood cells, along with the hundreds of other cell types seen in the adult. Such cell **differentiation** arises because cells make and accumulate different sets of RNA and protein molecules: that is, they express different genes.

The Different Cell Types of a Multicellular Organism Contain the Same DNA

The evidence that cells have the ability to change which genes they express without altering the nucleotide sequence of their DNA comes from experiments in which the genome from a differentiated cell is made to direct the development of a complete organism. If the chromosomes of the differentiated cell were altered irreversibly during development, they would not be able to accomplish this feat.

Consider, for example, an experiment in which the nucleus is taken from a skin cell in an adult frog and injected into a frog egg from which the nucleus has been removed. In at least some cases, that doctored egg will develop into a normal tadpole (**Figure 8–2**). Thus, the transplanted skin-cell nucleus cannot have lost any critical DNA sequences. Nuclear transplantation experiments carried out with differentiated cells taken from adult mammals—including sheep, cows, pigs, goats, and mice—have shown similar results. And in plants, individual cells removed from a carrot, for example, can regenerate an entire adult carrot plant. These experiments all show that the DNA in specialized cell types of multicellular organisms still contains the entire set of instructions needed to form

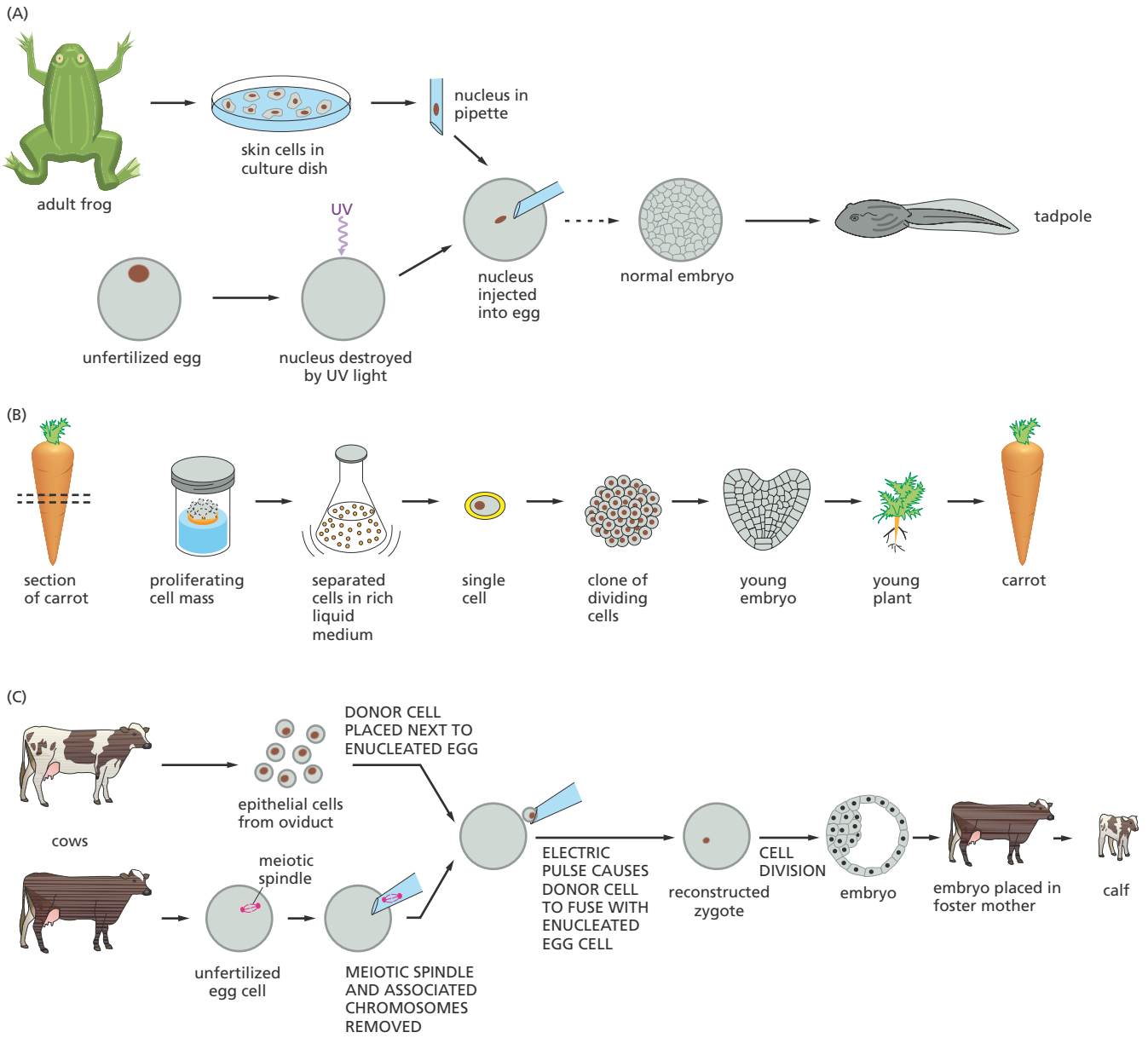


Figure 8–2 Differentiated cells contain all the genetic instructions necessary to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an egg whose nucleus has been destroyed can give rise to an entire tadpole. The broken arrow indicates that to give the transplanted genome time to adjust to an embryonic environment, a further transfer step is required in which one of the nuclei is taken from the early embryo that begins to develop and is put back into a second enucleated egg. (B) In many types of plants, differentiated cells retain the ability to "de-differentiate," so that a single cell can proliferate to form a clone of progeny cells that later give rise to an entire plant. (C) A nucleus removed from a differentiated cell from an adult cow can be introduced into an enucleated egg from a different cow to give rise to a calf. Different calves produced from the same differentiated cell donor are all clones of the donor and are therefore genetically identical. (A, modified from J.B. Gurdon, *Sci. Am.* 219:24–35, 1968, with permission from the Estate of Bunji Tagawa.)

a whole organism. The various cell types of an organism therefore differ not because they contain different genes, but because they express them differently.

Different Cell Types Produce Different Sets of Proteins

The extent of the differences in gene expression between different cell types may be roughly gauged by comparing the protein composition of cells in liver, heart, brain, and so on. In the past, such analysis was performed by two-dimensional gel electrophoresis (see Panel 4–5, p. 167). Nowadays, the total protein content of a cell can be rapidly analyzed by

a method called mass spectrometry (see Figure 4–49). This technique is much more sensitive than electrophoresis and it enables the detection of even proteins that are produced in minor quantities.

Both techniques reveal that many proteins are common to all the cells of a multicellular organism. These *housekeeping* proteins include, for example, the structural proteins of chromosomes, RNA polymerases, DNA repair enzymes, ribosomal proteins, enzymes involved in glycolysis and other basic metabolic processes, and many of the proteins that form the cytoskeleton. In addition, each different cell type also produces specialized proteins that are responsible for the cell's distinctive properties. In mammals, for example, hemoglobin is made almost exclusively in developing red blood cells.

Gene expression can also be studied by cataloging a cell's RNAs, including the mRNAs that encode protein. The most comprehensive methods for such analyses involve determining the nucleotide sequence of every RNA molecule made by the cell, an approach that can also reveal their relative abundance. Estimates of the number of different mRNA sequences in human cells suggest that, at any one time, a typical differentiated human cell expresses perhaps 5000–15,000 protein-coding genes from a total of about 21,000. It is the expression of a different collection of genes in each cell type that causes the large variations seen in the size, shape, behavior, and function of differentiated cells.

A Cell Can Change the Expression of Its Genes in Response to External Signals

The specialized cells in a multicellular organism are capable of altering their patterns of gene expression in response to extracellular cues. For example, if a liver cell is exposed to the steroid hormone cortisol, the production of several proteins is dramatically increased. Released by the adrenal gland during periods of starvation, intense exercise, or prolonged stress, cortisol signals liver cells to boost the production of glucose from amino acids and other small molecules. The set of proteins whose production is induced by cortisol includes enzymes such as tyrosine aminotransferase, which helps convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins returns to its resting level.

Other cell types respond to cortisol differently. In fat cells, for example, the production of tyrosine aminotransferase is reduced, while some other cell types do not respond to cortisol at all. The fact that different cell types often respond in different ways to the same extracellular signal contributes to the specialization that gives each cell type its distinctive character.

Gene Expression Can Be Regulated at Various Steps from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? As we saw in the last chapter, there are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it contains by (1) controlling when and how often a given gene is transcribed, (2) controlling how an RNA transcript is spliced or otherwise processed, (3) selecting which mRNAs are exported from the nucleus to the cytosol, (4) regulating how quickly certain mRNA molecules are degraded, (5) selecting which mRNAs are translated into protein by ribosomes, or

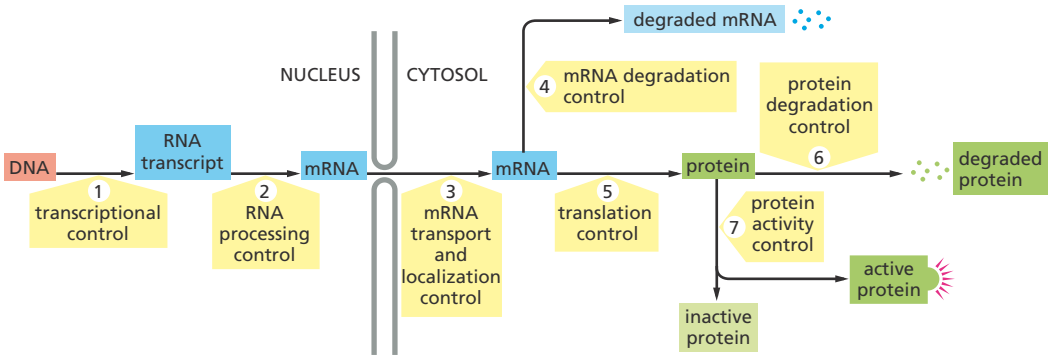


Figure 8–3 Gene expression in eukaryotic cells can be controlled at various steps. Examples of regulation at each of these steps are known, although for most genes the main site of control is step 1—transcription of a DNA sequence into RNA.

(6) regulating how rapidly specific proteins are destroyed after they have been made; in addition, the activity of individual proteins can be further regulated in a variety of ways. These steps are illustrated in **Figure 8–3**.

Gene expression can be regulated at each of these steps. For most genes, however, the control of transcription (step number 1 in Figure 8–3) is paramount. This makes sense because only transcriptional control can ensure that no unnecessary intermediates are synthesized. So it is the regulation of transcription—and the DNA and protein components that determine which genes a cell transcribes into RNA—that we address first.

HOW TRANSCRIPTIONAL SWITCHES WORK

Until 50 years ago, the idea that genes could be switched on and off was revolutionary. This concept was a major advance, and it came originally from studies of how *E. coli* bacteria adapt to changes in the composition of their growth medium. Many of the same principles apply to eukaryotic cells. However, the enormous complexity of gene regulation in higher organisms, combined with the packaging of their DNA into chromatin, creates special challenges and some novel opportunities for control—as we will see. We begin with a discussion of the *transcription regulators,* proteins that bind to DNA and control gene transcription.

Transcription Regulators Bind to Regulatory DNA Sequences

Control of transcription is usually exerted at the step at which the process is initiated. In Chapter 7, we saw that the **promoter** region of a gene binds the enzyme *RNA polymerase* and correctly orients the enzyme to begin its task of making an RNA copy of the gene. The promoters of both bacterial and eukaryotic genes include a *transcription initiation site*, where RNA synthesis begins, plus a sequence of approximately 50 nucleotide pairs that extends upstream from the initiation site (if one likens the direction of transcription to the flow of a river). This upstream region contains sites that are required for the RNA polymerase to recognize the *promoter*, although they do not bind to RNA polymerase directly. Instead, these sequences contain recognition sites for proteins that associate with the active polymerase—sigma factor in bacteria (see Figure 7–9) or the general transcription factors in eukaryotes (see Figure 7–12).

In addition to the promoter, nearly all genes, whether bacterial or eukaryotic, have **regulatory DNA sequences** that are used to switch the gene on or off. Some regulatory DNA sequences are as short as 10 nucleotide pairs and act as simple switches that respond to a single signal; such simple regulatory switches predominate in bacteria. Other regulatory DNA sequences, especially those in eukaryotes, are very long (sometimes spanning more than 10,000 nucleotide pairs) and act as molecular

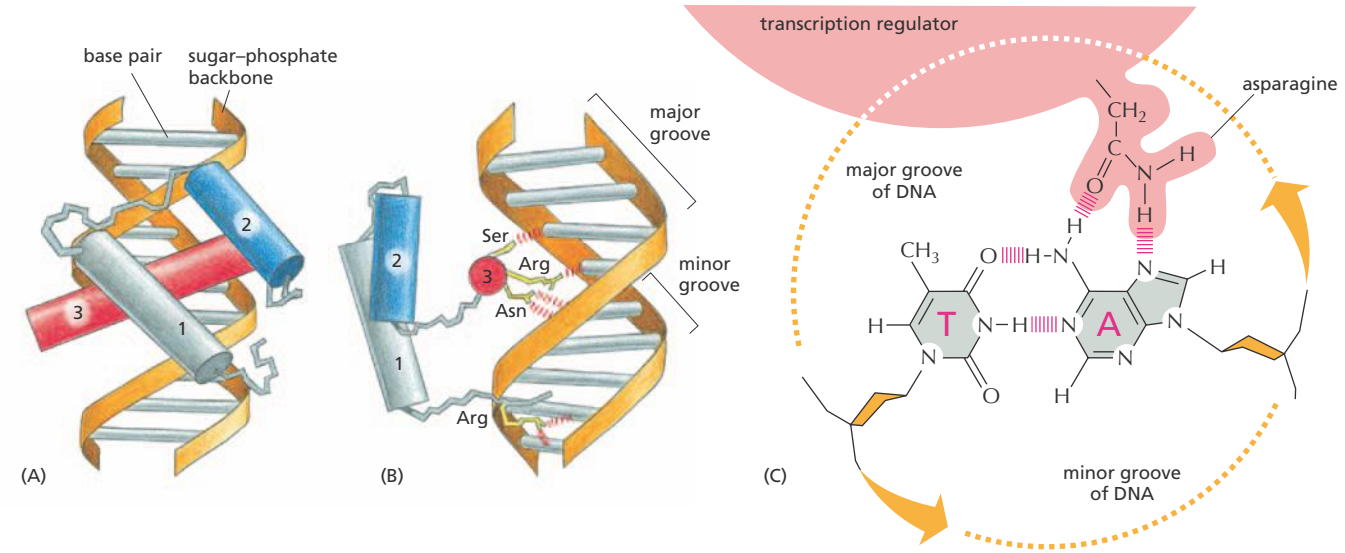


Figure 8–4 A transcription regulator interacts with the major groove of a DNA double helix. (A) This regulator recognizes DNA via three α helices, shown as numbered cylinders, which allow the protein to fit into the major groove and form tight associations with the base pairs in a short stretch of DNA. This particular structural motif, called a *homeodomain*, is found in many eukaryotic DNA-binding proteins (Movie 8.1). (B) Most of the contacts with the DNA bases are made by helix 3 (*red*), which is shown here end-on. The protein interacts with the edges of the nucleotides without disrupting the hydrogen bonds that hold the base pairs together. (C) An asparagine residue from helix 3 forms two hydrogen bonds with the adenine in an A-T base pair. The view is end-on looking down the DNA double helix, and the protein contacts the base pair from the major groove side. For simplicity, only one amino acid–base contact is shown; in reality, transcription regulators form hydrogen bonds (as shown here), ionic bonds, and hydrophobic interactions with individual bases in the major groove. Typically, the protein–DNA interface would consist of 10–20 such contacts, each involving a different amino acid and each contributing to the overall strength of the protein–DNA interaction.

microprocessors, integrating information from a variety of signals into a command that dictates how often transcription of the gene is initiated.

Regulatory DNA sequences do not work by themselves. To have any effect, these sequences must be recognized by proteins called **transcription regulators**. It is the binding of a transcription regulator to a regulatory DNA sequence that acts as the switch to control transcription. The simplest bacterium produces several hundred different transcription regulators, each of which recognizes a different DNA sequence and thereby regulates a distinct set of genes. Humans make many more—several thousand—indicating the importance and complexity of this form of gene regulation in the development and function of a complex organism.

Proteins that recognize a specific nucleotide sequence do so because the surface of the protein fits tightly against the surface features of the DNA double helix in that region. Because these surface features will vary depending on the nucleotide sequence, different DNA-binding proteins will recognize different nucleotide sequences. In most cases, the protein inserts into the major groove of the DNA helix and makes a series of intimate molecular contacts with the nucleotide pairs within the groove (**Figure 8–4**). Although each individual contact is weak, the 10 to 20 contacts that are typically formed at the protein–DNA interface combine to ensure that the interaction is both highly specific and very strong; indeed, protein–DNA interactions are among the tightest and most specific molecular interactions known in biology.

Many transcription regulators bind to the DNA helix as dimers (**Figure 8–5**). Such dimerization roughly doubles the area of contact with the DNA, thereby greatly increasing the strength and specificity of the protein–DNA interaction.

Transcriptional Switches Allow Cells to Respond to Changes in Their Environment

The simplest and best understood examples of gene regulation occur in bacteria and in the viruses that infect them. The genome of the bacterium *E. coli* consists of a single circular DNA molecule of about 4.6×10^6 nucleotide pairs. This DNA encodes approximately 4300 proteins, although only a fraction of these are made at any one time. Bacteria regulate the expression of many of their genes according to the food sources that are available in the environment. For example, in *E. coli*, five genes code for enzymes that manufacture the amino acid tryptophan. These genes are arranged in a cluster on the chromosome and are transcribed from a single promoter as one long mRNA molecule; such coordinately transcribed clusters are called *operons* (**Figure 8–6**). Although operons are common in bacteria, they are rare in eukaryotes, where genes are transcribed and regulated individually (see Figure 7–2).

When tryptophan concentrations are low, the operon is transcribed; the resulting mRNA is translated to produce a full set of biosynthetic enzymes, which work in tandem to synthesize tryptophan. When tryptophan is abundant, however—for example, when the bacterium is in the gut of a mammal that has just eaten a protein-rich meal—the amino acid is imported into the cell and shuts down production of the enzymes, which are no longer needed.

We now understand in considerable detail how this repression of the tryptophan operon comes about. Within the operon's promoter is a short DNA sequence, called the operator (see Figure 8–6), that is recognized by a transcription regulator. When this regulator binds to the *operator*, it blocks access of RNA polymerase to the promoter, preventing transcription of the operon and production of the tryptophan-producing enzymes. The transcription regulator is known as the *tryptophan repressor*, and it is controlled in an ingenious way: the repressor can bind to DNA only if it has also bound several molecules of tryptophan (**Figure 8–7**).

The tryptophan repressor is an allosteric protein (see Figure 4–41): the binding of tryptophan causes a subtle change in its three-dimensional structure so that the protein can bind to the operator sequence. When the concentration of free tryptophan in the bacterium drops, the repressor no longer binds to DNA, and the tryptophan operon is transcribed. The repressor is thus a simple device that switches production of a set of biosynthetic enzymes on and off according to the availability of the end product of the pathway that the enzymes catalyze.

The tryptophan repressor protein itself is always present in the cell. The gene that encodes it is continuously transcribed at a low level, so that a small amount of the repressor protein is always being made. Thus the bacterium can respond very rapidly to a rise in tryptophan concentration.

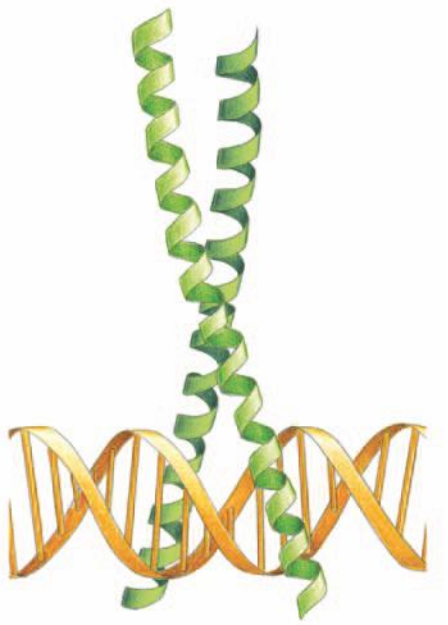


Figure 8–5 Many transcription regulators bind to DNA as dimers. This transcription regulator contains a *leucine zipper* motif, which is formed by two α helices, each contributed by a different protein subunit. Leucine zipper proteins thus bind to DNA as dimers, gripping the double helix like a clothespin on a clothesline (Movie 8.2).

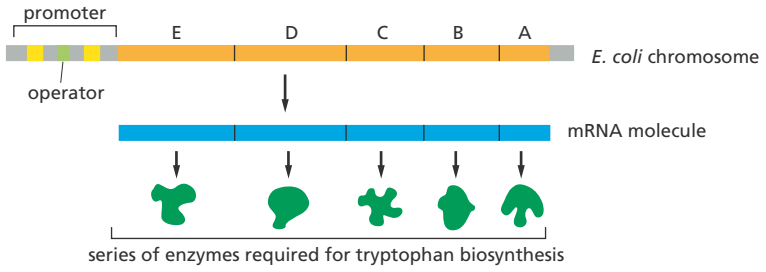


Figure 8–6 A cluster of bacterial genes can be transcribed from a single promoter. Each of these five genes encodes a different enzyme; all of the enzymes are needed to synthesize the amino acid tryptophan. The genes are transcribed as a single mRNA molecule, a feature that allows their expression to be coordinated. Clusters of genes transcribed as a single mRNA molecule are common in bacteria. Each of these clusters is called an *operon* because its expression is controlled by a regulatory DNA sequence called the *operator* (*green*), situated within the promoter. The *yellow* blocks in the promoter represent DNA sequences that bind RNA polymerase.

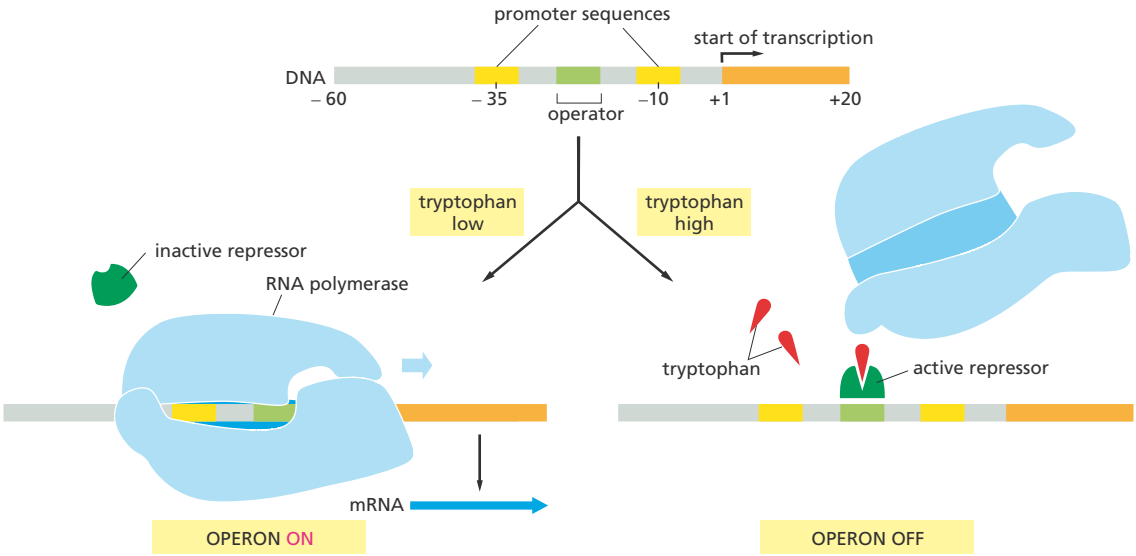


Figure 8–7 Genes can be switched off by repressor proteins. If the concentration of tryptophan inside a bacterium is low (*left*), RNA polymerase (*blue*) binds to the promoter and transcribes the five genes of the tryptophan operon. However, if the concentration of tryptophan is high (*right*), the repressor protein (*dark green*) becomes active and binds to the operator (*light green*), where it blocks the binding of RNA polymerase to the promoter. Whenever the concentration of intracellular tryptophan drops, the repressor falls off the DNA, allowing the polymerase to again transcribe the operon. The promoter contains two key blocks of DNA sequence information, the –35 and –10 regions, highlighted in *yellow*, which are recognized by RNA polymerase (see Figure 7–10). The complete operon is shown in Figure 8–6.

Repressors Turn Genes Off and Activators Turn Them On

The tryptophan repressor, as its name suggests, is a **transcriptional repressor** protein: in its active form, it switches genes off, or *represses* them. Some bacterial transcription regulators do the opposite: they switch genes on, or *activate* them. These **transcriptional activator** proteins work on promoters that—in contrast to the promoter for the tryptophan operon—are only marginally able to bind and position RNA polymerase on their own. However, these poorly functioning promoters can be made fully functional by activator proteins that bind nearby and contact the RNA polymerase to help it initiate transcription (**Figure 8–8**).

Like the tryptophan repressor, activator proteins often have to interact with a second molecule to be able to bind DNA. For example, the bacterial activator protein *CAP* has to bind cyclic AMP (cAMP) before it can bind to DNA (see Figure 4–19). Genes activated by CAP are switched on in response to an increase in intracellular cAMP concentration, which rises when glucose, the bacterium's preferred carbon source, is no longer available; as a result, CAP drives the production of enzymes that allow the bacterium to digest other sugars.

An Activator and a Repressor Control the *Lac* Operon

In many instances, the activity of a single promoter is controlled by two different transcription regulators. The *Lac operon* in *E. coli*, for example,

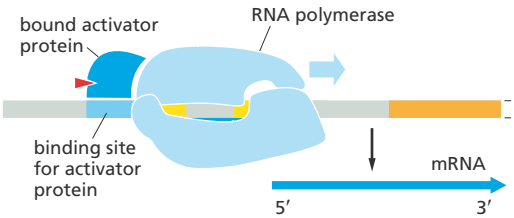


Figure 8–8 Genes can be switched on by activator proteins. An activator protein binds to a regulatory sequence on the DNA and then interacts with the RNA polymerase to help it initiate transcription. Without the activator, the promoter fails to initiate transcription efficiently. In bacteria, the binding of the activator to DNA is often controlled by the interaction of a metabolite or other small molecule (*red* triangle) with the activator protein. The *Lac* operon works in this manner, as we discuss shortly.

is controlled by both the *Lac repressor* and the CAP activator that we just discussed. The *Lac* operon encodes proteins required to import and digest the disaccharide lactose. In the absence of glucose, the bacterium makes cAMP, which activates CAP to switch on genes that allow the cell to utilize alternative sources of carbon—including lactose. It would be wasteful, however, for CAP to induce expression of the *Lac* operon if lactose itself were not present. Thus the Lac repressor shuts off the operon in the absence of lactose. This arrangement enables the control region of the *Lac* operon to integrate two different signals, so that the operon is highly expressed only when two conditions are met: glucose must be absent and lactose must be present (**Figure 8–9**). This genetic circuit thus behaves much like a switch that carries out a logic operation in a computer. When lactose is present AND glucose is absent, the cell executes the appropriate program—in this case, transcription of the genes that permit the uptake and utilization of lactose.

The elegant logic of the *Lac* operon first attracted the attention of biologists more than 50 years ago. The molecular basis of the switch in *E. coli* was uncovered by a combination of genetics and biochemistry, providing the first insight into how transcription is controlled. In a eukaryotic cell, similar transcription regulatory devices are combined to generate increasingly complex circuits, including those that enable a fertilized egg to form the tissues and organs of a multicellular organism.

QUESTION 8–1

Bacterial cells can take up the amino acid tryptophan (Trp) from their surroundings, or if there is an insufficient external supply they can synthesize tryptophan from other small molecules. The Trp repressor is a transcription regulator that shuts off the transcription of genes that code for the enzymes required for the synthesis of tryptophan (see Figure 8–7).

A. What would happen to the regulation of the tryptophan operon in cells that express a mutant form of the tryptophan repressor that (1) cannot bind to DNA, (2) cannot bind tryptophan, or (3) binds to DNA even in the absence of tryptophan?

B. What would happen in scenarios (1), (2), and (3) if the cells, in addition, produced normal tryptophan repressor protein from a second, normal gene?

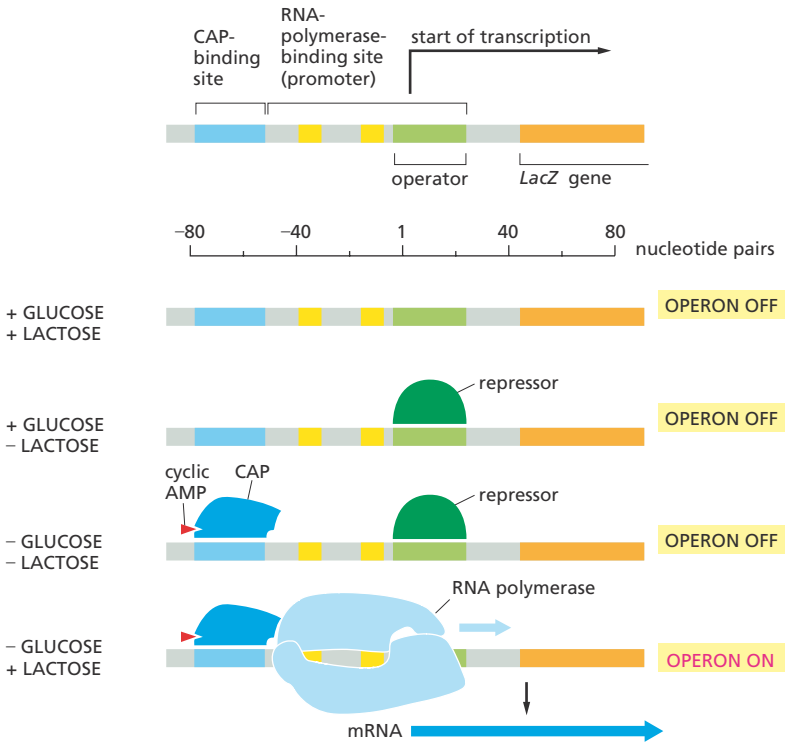


Figure 8–9 The *Lac* operon is controlled by two transcription regulators, the Lac repressor and CAP. When lactose is absent, the Lac repressor binds to the *Lac* operator and shuts off expression of the operon. Addition of lactose increases the intracellular concentration of a related compound, allolactose; allolactose binds to the Lac repressor, causing it to undergo a conformational change that releases its grip on the operator DNA (not shown). When glucose is absent, cyclic AMP (*red* triangle) is produced by the cell, and CAP binds to DNA. *LacZ*, the first gene of the operon, encodes the enzyme β-galactosidase, which breaks down lactose to galactose and glucose.

QUESTION 8–2

Explain how DNA-binding proteins can make sequence-specific contacts to a double-stranded DNA molecule without breaking the hydrogen bonds that hold the bases together. Indicate how, through such contacts, a protein can distinguish a T-A from a C-G pair. Indicate the parts of the nucleotide base pairs that could form noncovalent interactions—hydrogen bonds, electrostatic attractions, or hydrophobic interactions (see Panel 2–7, pp. 78–79)—with a DNA-binding protein. The structures of all the base pairs in DNA are given in Figure 5–6.

Eukaryotic Transcription Regulators Control Gene Expression from a Distance

Eukaryotes, too, use transcription regulators—both activators and repressors—to regulate the expression of their genes. The DNA sites to which eukaryotic gene activators bind are termed *enhancers*, because their presence dramatically enhances the rate of transcription. It was surprising to biologists when, in 1979, it was discovered that these activator proteins could enhance transcription even when they are bound thousands of nucleotide pairs away from a gene's promoter. They also work when bound either upstream or downstream from the gene. These observations raised several questions. How do enhancer sequences and the proteins bound to them function over such long distances? How do they communicate with the promoter?

Many models for this "action at a distance" have been proposed, but the simplest of these seems to apply in most cases. The DNA between the enhancer and the promoter loops out to allow eukaryotic activator proteins to influence directly events that take place at the promoter (**Figure 8–10**). The DNA thus acts as a tether, allowing a protein that is bound to an enhancer—even one that is thousands of nucleotide pairs away—to interact with the proteins in the vicinity of the promoter—including RNA polymerase and the general transcription factors (see Figure 7–12). Often, additional proteins serve to link the distantly bound transcription regulators to these proteins at the promoter; the most important of these regulators is a large complex of proteins known as *Mediator* (see Figure 8–10). One of the ways in which these proteins function is by aiding the assembly of the general transcription factors and RNA polymerase to form a large *transcription complex* at the promoter. Eukaryotic repressor proteins do the opposite: they decrease transcription by preventing the assembly of the same protein complex.

In addition to promoting—or repressing—the assembly of a transcription initiation complex directly, eukaryotic transcription regulators have an additional mechanism of action: they attract proteins that modify chromatin structure and thereby affect the accessibility of the promoter to the general transcription factors and RNA polymerase, as we discuss next.

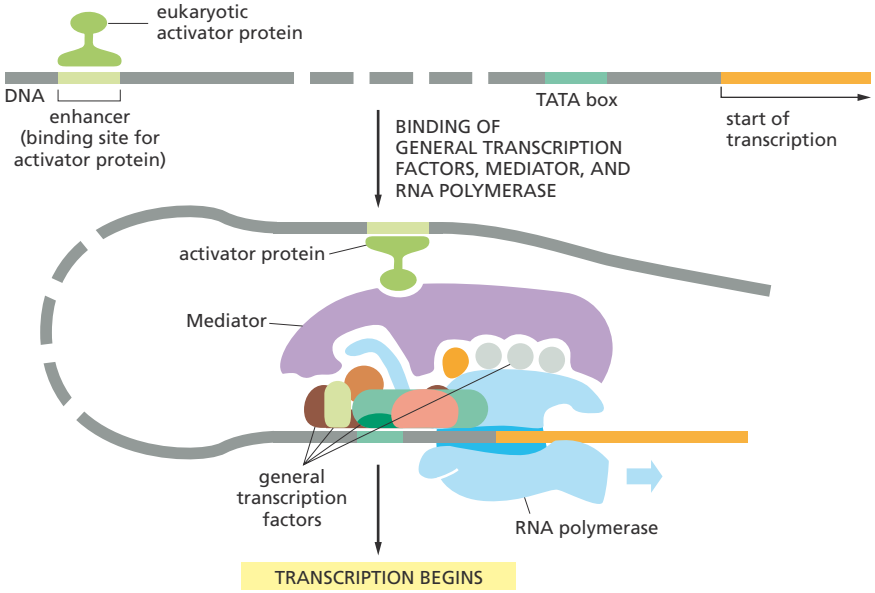


Figure 8–10 In eukaryotes, gene activation can occur at a distance. An activator protein bound to a distant enhancer attracts RNA polymerase and general transcription factors to the promoter. Looping of the intervening DNA permits contact between the activator and the transcription initiation complex bound to the promoter. In the case shown here, a large protein complex called Mediator serves as a go-between. The broken stretch of DNA signifies that the length of DNA between the enhancer and the start of transcription varies, sometimes reaching tens of thousands of nucleotide pairs in length. The TATA box is a DNA recognition sequence for the first general transcription factor that binds to the promoter (see Figure 7–12).

Eukaryotic Transcription Regulators Help Initiate Transcription by Recruiting Chromatin-Modifying Proteins

Initiation of transcription in eukaryotic cells must also take into account the packaging of DNA into chromosomes. As discussed in Chapter 5, eukaryotic DNA is packed into nucleosomes, which, in turn, are folded into higher-order structures. How do transcription regulators, general transcription factors, and RNA polymerase gain access to such DNA? Nucleosomes can inhibit the initiation of transcription if they are positioned over a promoter, because they physically block the assembly of the general transcription factors or RNA polymerase on the promoter. Such chromatin packaging may have evolved in part to prevent leaky gene expression by blocking the initiation of transcription in the absence of the proper activator proteins.

In eukaryotic cells, activator and repressor proteins exploit chromatin structure to help turn genes on and off. As we saw in Chapter 5, chromatin structure can be altered by *chromatin-remodeling complexes* and by enzymes that covalently modify the histone proteins that form the core of the nucleosome (see Figures 5–26 and 5–27). Many gene activators take advantage of these mechanisms by recruiting such chromatin-modifying proteins to promoters. For example, the recruitment of *histone acetyltransferases* promotes the attachment of acetyl groups to selected lysines in the tail of histone proteins. This modification alters chromatin structure, allowing greater accessibility to the underlying DNA; moreover, the acetyl groups themselves attract proteins that promote transcription, including some of the general transcription factors (**Figure 8–11**).

Likewise, gene repressor proteins can modify chromatin in ways that reduce the efficiency of transcription initiation. For example, many repressors attract *histone deacetylases*—enzymes that remove the acetyl groups from histone tails, thereby reversing the positive effects that acetylation has on transcription initiation. Although some eukaryotic repressor proteins work on a gene-by-gene basis, others can orchestrate the formation of large swathes of transcriptionally inactive chromatin containing many

QUESTION 8–3

Some transcription regulators bind to DNA and cause the double helix to bend at a sharp angle. Such "bending proteins" can stimulate the initiation of transcription without contacting either the RNA polymerase, any of the general transcription factors, or any other transcription regulators. Can you devise a plausible explanation for how these proteins might work to modulate transcription? Draw a diagram that illustrates your explanation.

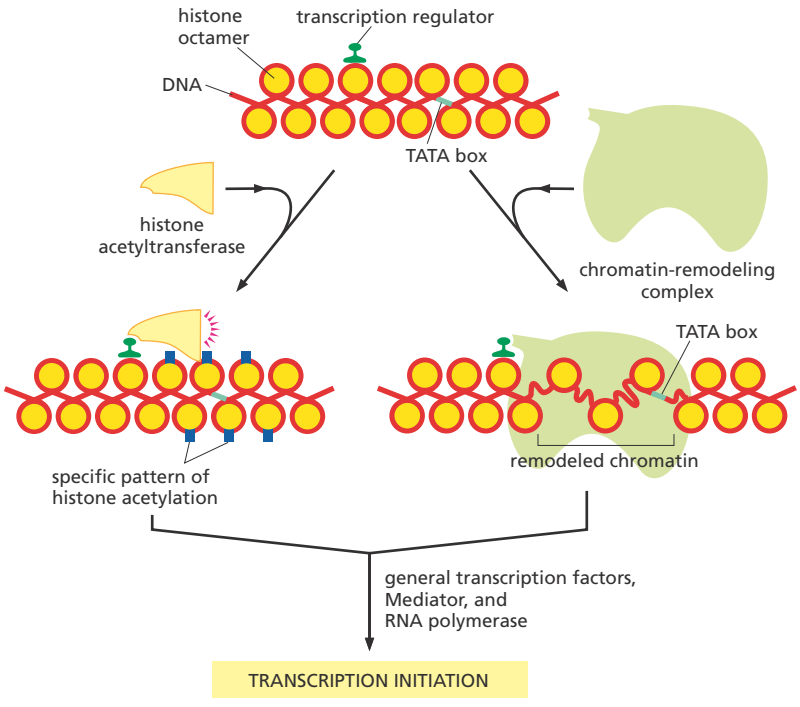


Figure 8–11 Eukaryotic transcriptional activators can recruit chromatin-modifying proteins to help initiate gene transcription. On the right, chromatin-remodeling complexes render the DNA packaged in chromatin more accessible to other proteins in the cell, including those required for transcription initiation; notice, for example, the increased exposure of the TATA box. On the left, the recruitment of histone-modifying enzymes such as histone acetyltransferases adds acetyl groups to specific histones, which can then serve as binding sites for proteins that stimulate transcription initiation (not shown).

genes. As discussed in Chapter 5, these transcription-resistant regions of DNA include the heterochromatin found in interphase chromosomes and the inactive X chromosome in the cells of female mammals.

THE MOLECULAR MECHANISMS THAT CREATE SPECIALIZED CELL TYPES

All cells must be able to switch genes on and off in response to signals in their environment. But the cells of multicellular organisms have evolved this capacity to an extreme degree and in highly specialized ways to form organized arrays of differentiated cell types. In particular, once a cell in a multicellular organism becomes committed to differentiate into a specific cell type, the choice of fate is generally maintained through subsequent cell divisions. This means that the changes in gene expression, which are often triggered by a transient signal, must be remembered by the cell. This phenomenon of *cell memory* is a prerequisite for the creation of organized tissues and for the maintenance of stably differentiated cell types. In contrast, the simplest changes in gene expression in both eukaryotes and bacteria are often only transient; the tryptophan repressor, for example, switches off the tryptophan operon in bacteria only in the presence of tryptophan; as soon as the amino acid is removed from the medium, the genes switch back on, and the descendants of the cell will have no memory that their ancestors had been exposed to tryptophan.

In this section, we discuss some of the special features of transcriptional regulation that are found in multicellular organisms. Our focus will be on how these mechanisms create and maintain the specialized cell types that give a worm, a fly, or a human its distinctive characteristics.

Eukaryotic Genes Are Controlled by Combinations of Transcription Regulators

Because eukaryotic transcription regulators can control transcription initiation when bound to DNA many base pairs away from the promoter, the nucleotide sequences that control the expression of a gene can be spread over long stretches of DNA. In animals and plants, it is not unusual to find the regulatory DNA sequences of a gene dotted over tens of thousands of nucleotide pairs, although much of the intervening DNA serves as "spacer" sequence and is not directly recognized by the transcription regulators.

So far in this chapter, we have treated transcription regulators as though each functions individually to turn a gene on or off. While this idea holds true for many simple bacterial activators and repressors, most eukaryotic transcription regulators work as part of a "committee" of regulatory proteins, all of which are necessary to express the gene in the right place, in the right cell type, in response to the right conditions, at the right time, and in the required amount.

The term **combinatorial control** refers to the way that groups of transcription regulators work together to determine the expression of a single gene. We saw a simple example of such regulation by multiple regulators when we discussed the bacterial *Lac* operon (see Figure 8–9). In eukaryotes, the regulatory inputs have been amplified, and a typical gene is controlled by dozens of transcription regulators. These help assemble chromatin-remodeling complexes, histone-modifying enzymes, RNA polymerase, and general transcription factors via the multiprotein Mediator complex (**Figure 8–12**). In many cases, both repressors and activators will be present in the same complex; how the cell integrates the effects of all of these proteins to determine the final level of gene

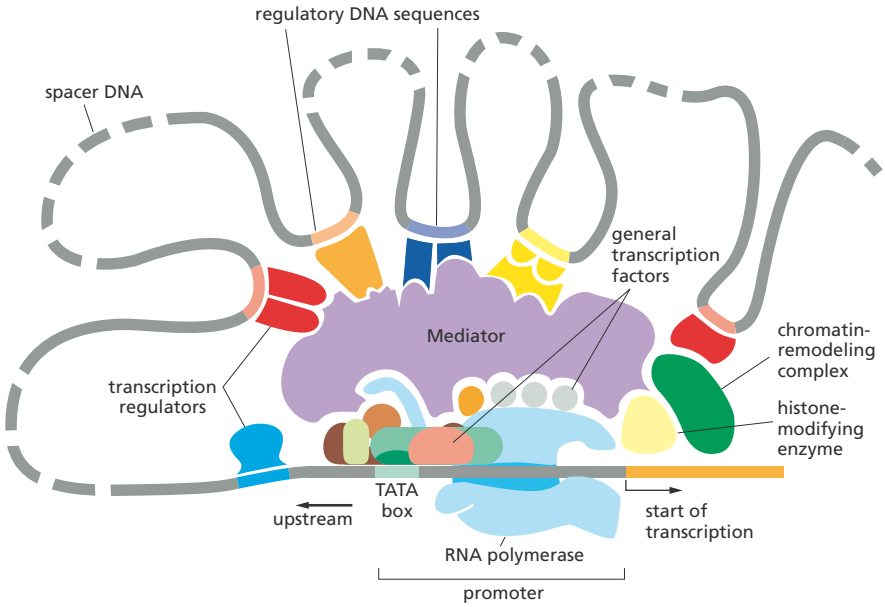


Figure 8–12 Transcription regulators work together as a "committee" to control the expression of a eukaryotic gene. Whereas the general transcription factors that assemble at the promoter are the same for all genes transcribed by RNA polymerase (see Figure 7–12), the transcription regulators and the locations of their DNA binding sites relative to the promoters are different for different genes. These regulators, along with chromatin-modifying proteins, are assembled at the promoter by the Mediator. The effects of multiple transcription regulators combine to determine the final rate of transcription initiation.

expression is only now beginning to be understood. An example of such a complex regulatory system—one that participates in the development of a fruit fly from a fertilized egg—is described in **How We Know**, pp. 274–275.

The Expression of Different Genes Can Be Coordinated by a Single Protein

In addition to being able to switch individual genes on and off, all cells—whether prokaryote or eukaryote—need to coordinate the expression of different genes. When a eukaryotic cell receives a signal to divide, for example, a number of hitherto unexpressed genes are turned on together to set in motion the events that lead eventually to cell division (discussed in Chapter 18). As discussed earlier, one way in which bacteria coordinate the expression of a set of genes is by having them clustered together in an operon under the control of a single promoter (see Figure 8–6). Such clustering is not seen in eukaryotic cells, where each gene is transcribed and regulated individually. So how do these cells coordinate gene expression? In particular, given that a eukaryotic cell uses a committee of transcription regulators to control each of its genes, how can it rapidly and decisively switch whole groups of genes on or off?

The answer is that even though control of gene expression is combinatorial, the effect of a single transcription regulator can still be decisive in switching any particular gene on or off, simply by completing the combination needed to activate or repress that gene. This is like dialing in the final number of a combination lock: the lock will spring open if the other numbers have been previously entered. Just as the same number can complete the combination for different locks, the same protein can complete the combination for several different genes. As long as different genes contain regulatory DNA sequences that are recognized by the same transcription regulator, they can be switched on or off together, as a coordinated unit.

An example of such coordinated regulation in humans is seen with the *cortisol receptor protein*. In order to bind to regulatory sites in DNA, this

GENE REGULATION—THE STORY OF *EVE*

The ability to regulate gene expression is crucial to the proper development of a multicellular organism from a fertilized egg to a fertile adult. Beginning at the earliest moments in development, a succession of transcriptional programs guides the differential expression of genes that allows an animal to form a proper body plan—helping to distinguish its back from its belly, and its head from its tail. These programs ultimately direct the correct placement of a wing or a leg, a mouth or an anus, a neuron or a sex cell.

A central challenge in development, then, is to understand how an organism generates these patterns of gene expression, which are laid down within hours of fertilization. Among the most important genes involved in these early stages of development are those that encode transcription regulators. By interacting with different regulatory DNA sequences, these proteins instruct every cell in the embryo to switch on the genes that are appropriate for that cell at each time point during development. How can a protein binding to a piece of DNA help direct the development of a complex multicellular organism? To see how we can address that large question, we review the story of *Eve*.

Seeing *Eve*

Even-skipped—Eve, for short—is a gene whose expression plays an important part in the development of the *Drosophila* embryo. If this gene is inactivated by mutation, many parts of the embryo fail to form and the fly larva dies early in development. But *Eve* is not expressed uniformly throughout the embryo. Instead, the Eve protein is produced in a striking series of seven neat stripes, each of which occupies a very precise position along the length of the embryo. These seven stripes correspond to seven of the fourteen segments that define the body plan of the fly—three for the head, three for the thorax, and eight for the abdomen.

This pattern never varies: *Eve* can be found in the very same places in every *Drosophila* embryo (see Figure 8–13B). How can the expression of a gene be regulated with such spatial precision—such that one cell will produce a protein while a neighboring cell does not? To find out, researchers took a trip upstream.

Dissecting the DNA

As we have seen in this chapter, regulatory DNA sequences control which cells in an organism will express a particular gene, and at what point during development that gene will be turned on. In eukaryotes, these regulatory sequences are frequently located upstream of the gene itself. One way to locate a regulatory DNA sequence—and study how it operates—is to remove a piece of DNA from the region upstream of a gene of interest and insert that DNA upstream of a **reporter gene**—one that encodes a protein with an activity that is easy to monitor experimentally. If the piece of DNA contains a regulatory sequence, it will drive the expression of the reporter gene. When this patchwork piece of DNA is subsequently introduced into a cell or organism, the reporter gene will be expressed in the same cells and tissues that normally express the gene from which the regulatory sequence was derived (see Figure 10–31).

By excising various segments of the DNA sequences upstream of *Eve*, and coupling them to a reporter gene, researchers found that the expression of the gene is controlled by a series of seven regulatory modules—each of which specifies a single stripe of *Eve* expression. In this way, researchers identified, for example, a single segment of regulatory DNA that specifies stripe 2. They could excise this regulatory segment, link it to a reporter gene, and introduce the resulting DNA segment into the fly. When they examined embryos that carried this engineered DNA, they found that the reporter gene is expressed in the precise position of stripe 2 (**Figure 8–13**). Similar experiments revealed the existence of six other regulatory modules, one for each of the other Eve stripes.

The next question is: How does each of these seven regulatory segments direct the formation of a single stripe in a specific position? The answer, researchers found, is that each segment contains a unique combination of regulatory sequences that bind different combinations of transcription regulators. These regulators, like Eve itself, are distributed in unique patterns within the embryo—some toward the head, some toward the rear, some in the middle.

The regulatory segment that defines stripe 2, for example, contains regulatory DNA sequences for four transcription regulators: two that activate *Eve* transcription and two that repress it (**Figure 8–14**). In the narrow band of tissue that constitutes stripe 2, it just so happens the repressor proteins are not present—so the *Eve* gene is expressed; in the bands of tissue on either side of the stripe, the repressors keep *Eve* quiet. And so a stripe is formed.

The regulatory segments controlling the other stripes are thought to function along similar lines; each regulatory segment reads "positional information" provided

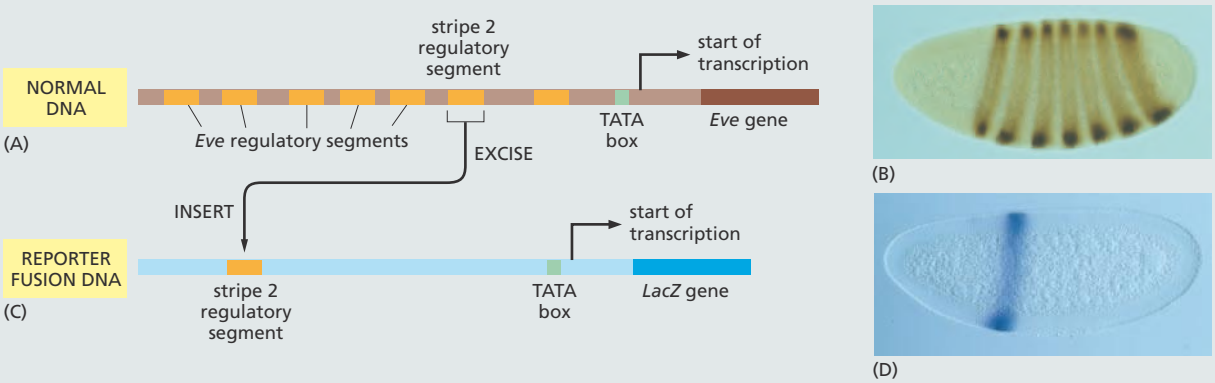


Figure 8–13 An experimental approach that involves the use of a reporter gene reveals the modular construction of the *Eve* gene regulatory region. (A) Expression of the *Eve* gene is controlled by a series of regulatory segments (*orange*) that direct the production of Eve protein in stripes along the embryo. (B) Embryos stained with antibodies to the Eve protein show the seven characteristic stripes of Eve expression. (C) In the laboratory, the regulatory segment that directs the formation of stripe 2 can be excised from the DNA shown in part A and inserted upstream of the *E. coli LacZ* gene, which encodes the enzyme β-galactosidase (see Figure 8–9). (D) When the engineered DNA containing the stripe 2 regulatory segment is introduced into the genome of a fly, the resulting embryo expresses β-galactosidase precisely in the position of the second *Eve* stripe. Enzyme activity is assayed by the addition of X-gal, a modified sugar that when cleaved by β-galactosidase generates an insoluble blue product. (B and D, courtesy of Stephen Small and Michael Levine.)

by some unique combination of transcription regulators in the embryo and expresses *Eve* on the basis of this information. The entire regulatory region is strung out over 20,000 nucleotide pairs of DNA and, altogether, binds more than 20 transcription regulators. This large regulatory region is built from a series of smaller regulatory segments, each of which consists of a unique arrangement of regulatory DNA sequences recognized by specific transcription regulators. In this way, the Eve gene can respond to an enormous combination of inputs.

The Eve protein is itself a transcription regulator, and it—in combination with many other regulatory proteins—controls key events in the development of the fly. This complex organization of a discrete number of regulatory elements begins to explain how the development of an entire organism can be orchestrated by repeated applications of a few basic principles.

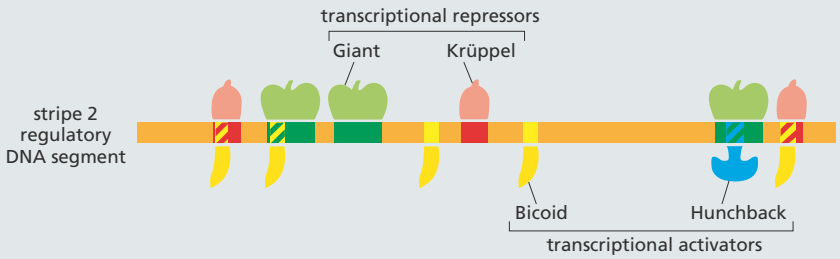


Figure 8–14 The regulatory segment that specifies Eve stripe 2 contains binding sites for four different transcription regulators. All four regulators are responsible for the proper expression of *Eve* in stripe 2. Flies that are deficient in the two activators, called Bicoid and Hunchback, fail to form stripe 2 efficiently; in flies deficient in either of the two repressors, called Giant and Krüppel, stripe 2 expands and covers an abnormally broad region of the embryo. As indicated in the diagram, in some cases the binding sites for the transcription regulators overlap, and the proteins compete for binding to the DNA. For example, the binding of Bicoid and Krüppel to the site at the far right is thought to be mutually exclusive. The regulatory segment is 480 base pairs in length.

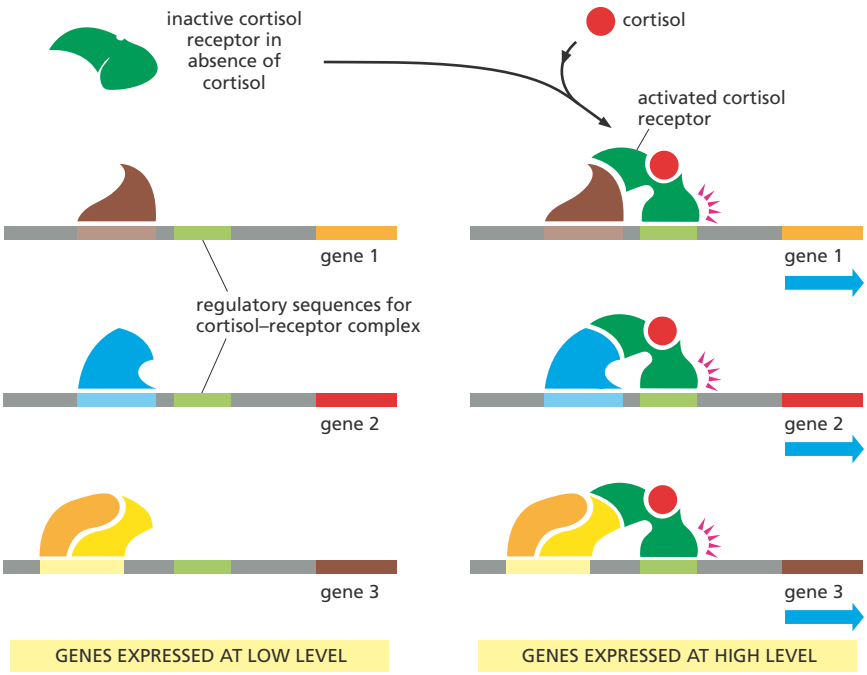


Figure 8–15 A single transcription regulator can coordinate the expression of many different genes. The action of the cortisol receptor is illustrated. On the left is a series of genes, each of which has a different gene activator protein bound to its respective regulatory DNA sequences. However, these bound proteins are not sufficient on their own to activate transcription efficiently. On the right is shown the effect of adding an additional transcription regulator—the cortisol–receptor complex—that can bind to the same regulatory DNA sequence in each gene. The activated cortisol receptor completes the combination of transcription regulators required for efficient initiation of transcription, and the genes are now switched on as a set.

transcription regulator must first form a complex with a molecule of cortisol (see Table 16–1, p. 529). In response to cortisol, liver cells increase the expression of many genes, one of which encodes the enzyme tyrosine aminotransferase, as discussed earlier. All these genes are regulated by the binding of the cortisol–receptor complex to a regulatory sequence in the DNA of each gene. When the cortisol concentration decreases again, the expression of all of these genes drops to its normal level. In this way, a single transcription regulator can coordinate the expression of many different genes (**Figure 8–15**).

Combinatorial Control Can Also Generate Different Cell Types

The ability to switch many different genes on or off using a limited number of transcription regulators is not only useful in the day-to-day regulation of cell function. It is also one of the means by which eukaryotic cells diversify into particular types of cells during embryonic development. A striking example is the development of muscle cells. A mammalian skeletal muscle cell is distinguished from other cells by the production of a large number of characteristic proteins, such as the muscle-specific forms of actin and myosin that make up the contractile apparatus (discussed in Chapter 17), as well as the receptor proteins and ion channel proteins in the plasma membrane that make the muscle cell sensitive to nerve stimulation. The genes encoding these muscle-specific proteins are all switched on coordinately as the muscle cell differentiates. Studies of developing muscle cells in culture have identified a small number of key transcription regulators, expressed only in potential muscle cells, that coordinate muscle-specific gene expression and are thus crucial for muscle-cell differentiation. This set of regulators activates the transcription of the genes that code for muscle-specific proteins by binding to specific DNA sequences present in their regulatory regions.

Some transcription regulators can even convert one specialized cell type to another. For example, when the gene encoding the transcription regulator MyoD is artificially introduced into fibroblasts cultured from skin

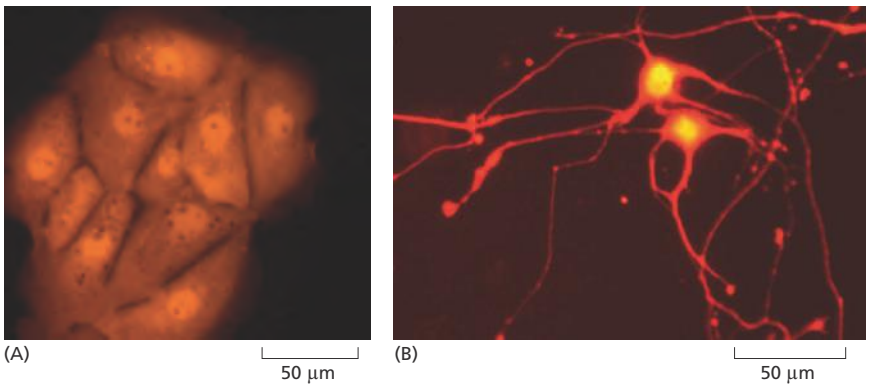


Figure 8–16 A small number of transcription regulators can convert one differentiated cell type directly into another. In this experiment, liver cells grown in culture (A) were converted into neuronal cells (B) via the artificial introduction of three nerve-specific transcription regulators. The cells are labeled with a fluorescent dye. (From S. Marro et al., *Cell Stem Cell* 9:374–378, 2011. With permission from Elsevier.)

connective tissue, the fibroblasts form musclelike cells. It appears that the fibroblasts, which are derived from the same broad class of embryonic cells as muscle cells, have already accumulated many of the other necessary transcription regulators required for the combinatorial control of the muscle-specific genes, and that addition of MyoD completes the unique combination required to direct the cells to become muscle.

This type of reprogramming can produce even more dramatic effects. For example, a set of nerve-specific transcription regulators, when artificially expressed in cultured liver cells, can convert them into functional neurons (**Figure 8–16**). Such dramatic results suggest that it may someday be possible to produce in the laboratory any cell type for which the correct combination of transcription regulators can be identified. How these transcription regulators can then lead to the generation of different cell types is illustrated schematically in **Figure 8–17**.

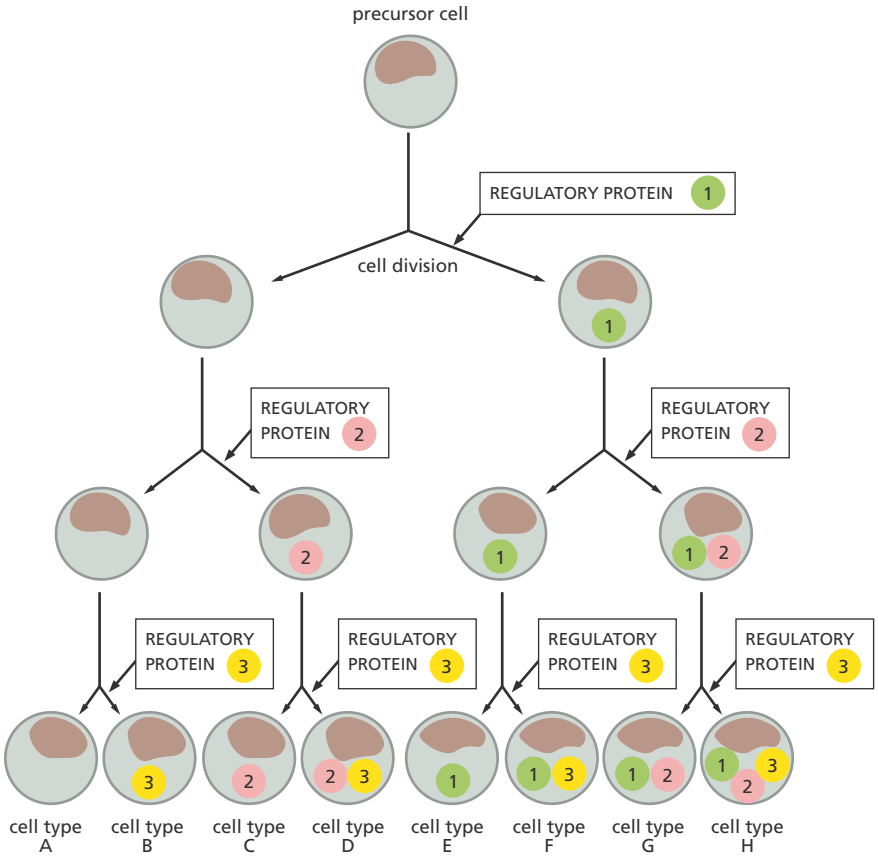


Figure 8–17 Combinations of a few transcription regulators can generate many cell types during development. In this simple scheme, a "decision" to make a new transcription regulator (shown as a *numbered* circle) is made after each cell division. Repetition of this simple rule can generate eight cell types (A through H), using only three transcription regulators. Each of these hypothetical cell types would then express many different genes, as dictated by the combination of transcription regulators that each cell type produces.

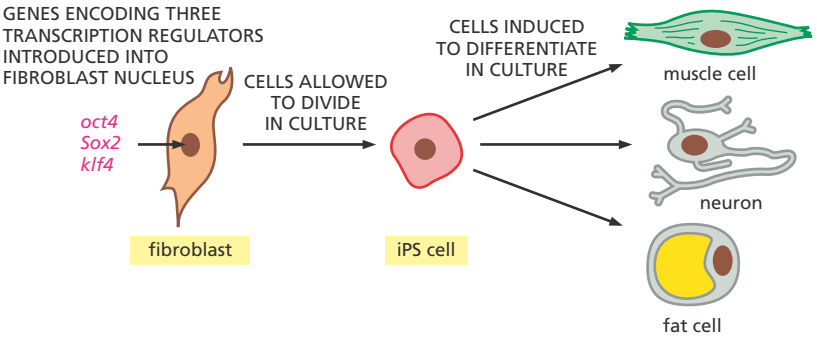


Figure 8–18 A combination of transcription regulators can induce a differentiated cell to de-differentiate into a pluripotent cell. The artificial expression of a set of four genes, each of which encodes a transcription regulator, can reprogram a fibroblast into a pluripotent cell with ES cell-like properties. Like ES cells, such *iPS cells* can proliferate indefinitely in culture and can be stimulated by appropriate extracellular signal molecules to differentiate into almost any cell type in the body.

Specialized Cell Types Can Be Experimentally Reprogrammed to Become Pluripotent Stem Cells

We have seen that, in some cases, one type of differentiated cell can be experimentally converted into another type by the artificial expression of specific transcription regulators (see Figure 8–16). Even more surprising, transcription regulators can coax various differentiated cells to *de-differentiate* into **pluripotent stem cells** that are capable of giving rise to all the specialized cell types in the body, much like the embryonic stem (ES) cells discussed in Chapter 20 (see pp. 708–711).

Using a defined set of transcription regulators, cultured mouse fibroblasts have been reprogrammed to become **induced pluripotent stem (iPS) cells**—cells that look and behave like the pluripotent ES cells that are derived from embryos (**Figure 8–18**). The approach was quickly adapted to produce iPS cells from a variety of specialized cell types, including cells taken from humans. Such human iPS cells can then be directed to generate a population of differentiated cells for use in the study or treatment of disease, as we discuss in Chapter 20.

The Formation of an Entire Organ Can Be Triggered by a Single Transcription Regulator

We have seen that a small number of transcription regulators can control the expression of whole sets of genes and can even convert one cell type into another. But an even more stunning example of the power of transcriptional control comes from studies of eye development in *Drosophila*. In this case, a single "master" transcription regulator called Ey could be used to trigger the formation of not just a single cell type but a whole organ. In the laboratory, the *Ey* gene can be artificially expressed in fruit fly embryos in cells that would normally give rise to a leg. When these modified embryos develop into adult flies, some have an eye in the middle of a leg (**Figure 8–19**).

Figure 8–19 Artificially induced expression of the *Drosophila Ey* gene in the precursor cells of the leg triggers the misplaced development of an eye on a fly's leg. (Courtesy of Walter Gehring.)

How the Ey protein coordinates the specification of each type of cell found in the eye—and directs their proper organization in three-dimensional space—is an actively studied topic in developmental biology. In essence, however, Ey functions like any other transcription regulator, controlling the expression of multiple genes by binding to DNA sequences in their regulatory regions. Some of the genes controlled by Ey encode additional transcription regulators that, in turn, control the expression of other genes. In this way, the action of a single transcription regulator can produce a cascade of regulators that, working in combination, lead to the formation of an organized group of many different types of cells. One can begin to imagine how, by repeated applications of this principle, a complex organism self-assembles, piece by piece.

Epigenetic Mechanisms Allow Differentiated Cells to Maintain Their Identity

Once a cell has become differentiated into a particular cell type, it will generally remain differentiated, and all its progeny cells will remain that same cell type. Some highly specialized cells, including skeletal muscle cells and neurons, never divide again once they have differentiated—that is, they are *terminally differentiated* (as discussed in Chapter 18). But many other differentiated cells—such as fibroblasts, smooth muscle cells, and liver cells—will divide many times in the life of an individual. When they do, these specialized cell types give rise only to cells like themselves: smooth muscle cells do not give rise to liver cells, nor liver cells to fibroblasts.

For a proliferating cell to maintain its identity—a property called **cell memory**—the patterns of gene expression responsible for that identity must be remembered and passed on to its daughter cells through all subsequent cell divisions. Thus, in the model illustrated in Figure 8–17, the production of each transcription regulator, once begun, has to be continued in the daughter cells of each cell division. How is such perpetuation accomplished?

Cells have several ways of ensuring that their daughters "remember" what kind of cells they are. One of the simplest and most important is through a **positive feedback loop**, where a master transcription regulator activates transcription of its own gene, in addition to that of other cell-type–specific genes. Each time a cell divides the regulator is distributed to both daughter cells, where it continues to stimulate the positive feedback loop. The continued stimulation ensures that the regulator will continue to be produced in subsequent cell generations. The Ey protein discussed earlier functions in such a positive feedback loop. Positive feedback is crucial for establishing the "self-sustaining" circuits of gene expression that allow a cell to commit to a particular fate—and then to transmit that information to its progeny (**Figure 8–20**).

Although positive feedback loops are probably the most prevalent way of ensuring that daughter cells remember what kind of cells they are meant to be, there are other ways of reinforcing cell identity. One involves the methylation of DNA. In vertebrate cells, **DNA methylation** occurs on certain cytosine bases (**Figure 8–21**). This covalent modification generally

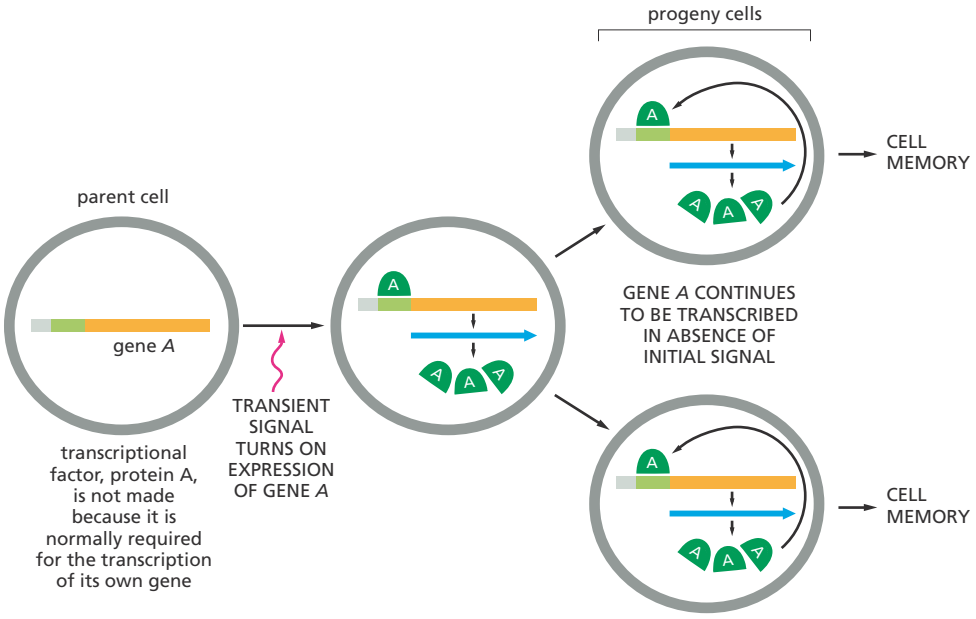


Figure 8–20 A positive feedback loop can create cell memory. Protein A is a master transcription regulator that activates the transcription of its own gene—as well as other cell-type-specific genes (not shown). All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of protein A.

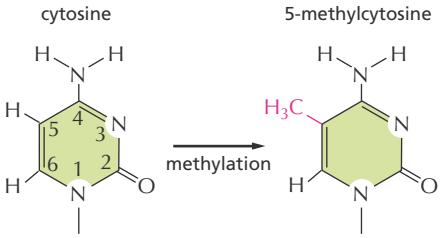


Figure 8–21 Formation of 5-methylcytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates, this modification is confined to selected cytosine (C) nucleotides that fall next to a guanine (G) in the sequence CG.

turns off genes by attracting proteins that bind to methylated cytosines and block gene transcription. DNA methylation patterns are passed on to progeny cells by the action of an enzyme that copies the methylation pattern on the parent DNA strand to the daughter DNA strand as it is synthesized (**Figure 8–22**).

Another mechanism for inheriting gene expression patterns involves the modification of histones. When a cell replicates its DNA, each daughter double helix receives half of its parent's histone proteins, which contain the covalent modifications of the parent chromosome. Enzymes responsible for these modifications may bind to the parental histones and confer the same modifications to the new histones nearby. This cycle of modification reestablishes the pattern of chromatin structure found in the parent chromosome (**Figure 8–23**).

Because all of these cell-memory mechanisms transmit patterns of gene expression from parent to daughter cell without altering the actual nucleotide sequence of the DNA, they are considered to be forms of **epigenetic inheritance**. Such epigenetic changes play an important part in controlling patterns of gene expression, allowing transient signals from the environment to be permanently recorded by our cells—a fact that has important implications for understanding how cells operate and how they malfunction in disease.

POST-TRANSCRIPTIONAL CONTROLS

We have seen that transcription regulators control gene expression by promoting or hindering the transcription of specific genes. The vast majority of genes in all organisms are regulated in this way. But many additional points of control can come into play later in the pathway from DNA to protein, giving cells a further opportunity to regulate the amount or activity of the gene products that they make (see Figure 8–3). These **post-transcriptional controls**, which operate after transcription has begun, play a crucial part in regulating the expression of almost all genes.

We have already encountered a few examples of such post-transcriptional control. We have seen how alternative RNA splicing allows different

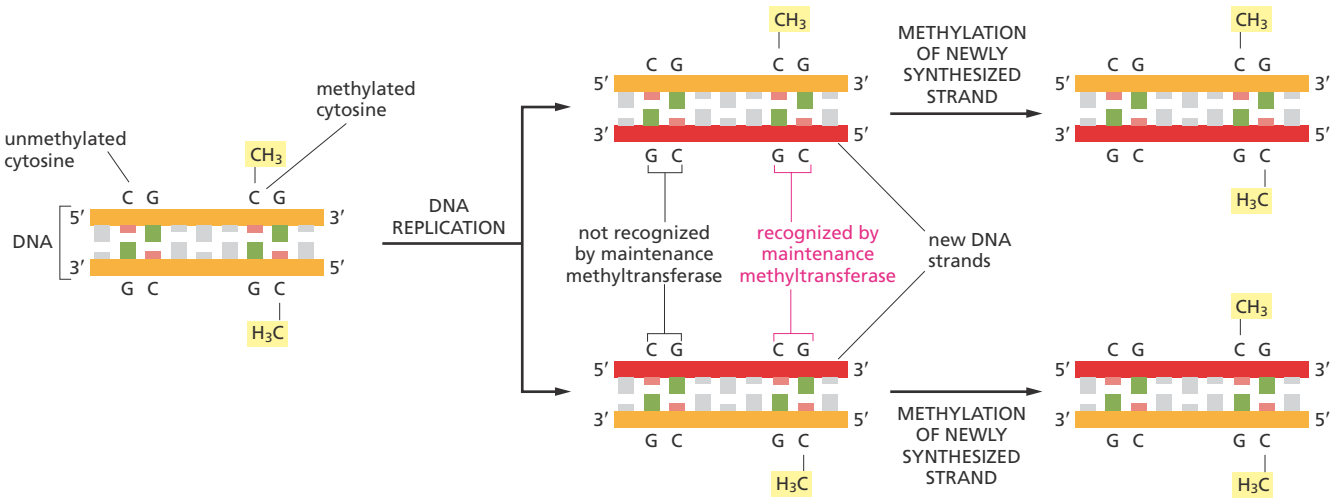


Figure 8–22 DNA methylation patterns can be faithfully inherited when a cell divides. An enzyme called a maintenance methyltransferase guarantees that once a pattern of DNA methylation has been established, it is inherited by newly made DNA. Immediately after DNA replication, each daughter double helix will contain one methylated DNA strand—inherited from the parent double helix—and one unmethylated, newly synthesized strand. The maintenance methyltransferase interacts with these hybrid double helices and methylates only those CG sequences that are base-paired with a CG sequence that is already methylated.

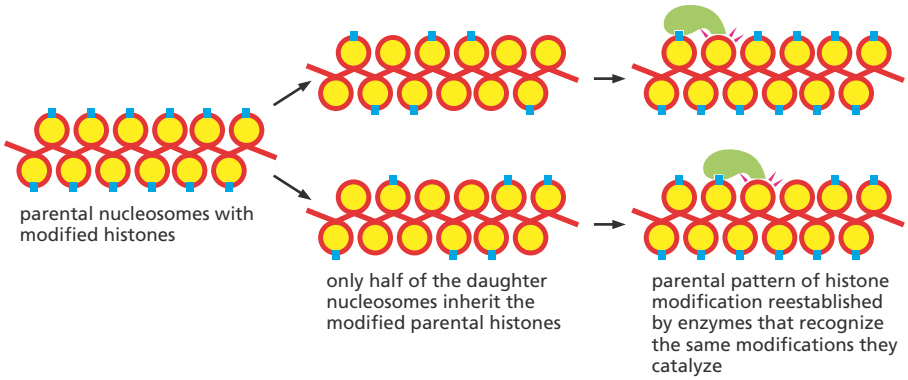


Figure 8–23 Histone modifications may be inherited by daughter chromosomes. When a chromosome is replicated, its resident histones are distributed more or less randomly to each of the two daughter DNA double helices. Thus, each daughter chromosome will inherit about half of its parent's collection of modified histones. The remaining stretches of DNA receive newly synthesized, not-yet-modified histones. If the enzymes responsible for each type of modification bind to the specific modification they create, they can catalyze the spread of this modification on the new histones. This cycle of modification and recognition can restore the parental histone modification pattern and, ultimately, allow the inheritance of the parental chromatin structure. This mechanism may apply to some but not all types of histone modifications.

forms of a protein, encoded by the same gene, to be made in different tissues (Figure 7–22). And we have discussed how various post-translational modifications of a protein can regulate its concentration and activity (see Figure 4–43). In the remainder of this chapter, we consider several other examples—some only recently discovered—of the many ways in which cells can manipulate the expression of a gene after transcription has commenced.

Each mRNA Controls Its Own Degradation and Translation

The more time an mRNA persists in the cell before it is degraded, the more protein it will produce. In bacteria, most mRNAs last only a few minutes before being destroyed. This instability allows a bacterium to adapt quickly to environmental changes. Eukaryotic mRNAs are generally more stable. The mRNA that encodes β-globin, for example, has a half-life of more than 10 hours. Most eukaryotic mRNAs, however, have half-lives of less than 30 minutes, and the most short-lived are those that encode proteins whose concentrations need to change rapidly based on the cell's needs, such as transcription regulators. Whether bacterial or eukaryotic, an mRNA's lifetime is dictated by specific nucleotide sequences within the untranslated regions that lie both upstream and downstream of the protein-coding sequence. These sequences often harbor binding sites for proteins that are involved in RNA degradation.

In addition to the nucleotide sequences that regulate its half-life, each mRNA possesses sequences that help control how often or how efficiently it will be translated into protein. These sequences control translation initiation. Although the details differ between eukaryotes and bacteria, the general strategy is similar for both.

Bacterial mRNAs contain a short ribosome-binding sequence located a few nucleotide pairs upstream of the AUG codon where translation begins (see Figure 7–37). This binding sequence forms base pairs with the RNA in the small ribosomal subunit, correctly positioning the initiating AUG codon within the ribosome. Because this interaction is needed for efficient translation initiation, it provides an ideal target for translational control. By blocking—or exposing—the ribosome-binding sequence, the bacterium can either inhibit—or promote—the translation of an mRNA (**Figure 8–24**).

Eukaryotic mRNAs possess a 5′ cap that helps guide the ribosome to the first AUG, the codon where translation will start (see Figure 7–36). Eukaryotic repressor proteins can inhibit translation initiation by binding to specific nucleotide sequences in the 5′ untranslated region of the mRNA, thereby preventing the ribosome from finding the first AUG—a mechanism similar to that in bacteria. When conditions change, the cell can inactivate the repressor to initiate translation of the mRNA.

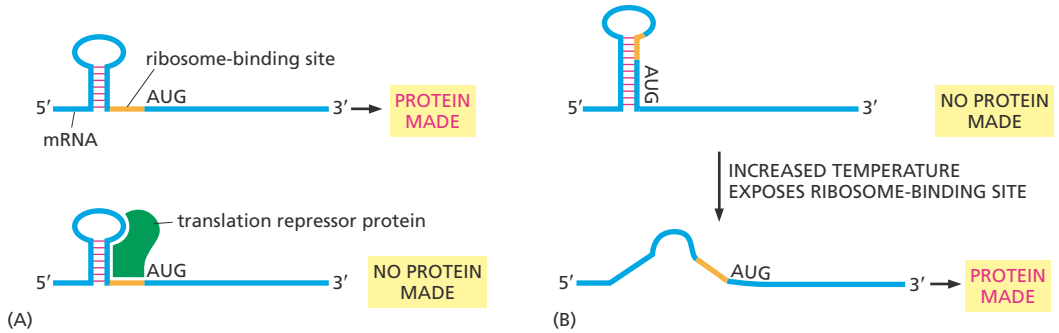


Figure 8–24 A bacterial gene's expression can be controlled by regulating translation of its mRNA. (A) Sequence-specific RNA-binding proteins can repress the translation of specific mRNAs by keeping the ribosome from binding to the ribosome-binding sequence (*orange*) in the mRNA. Some ribosomal proteins exploit this mechanism to inhibit the translation of their own mRNA. In this way, "extra" ribosomal proteins—those not incorporated into ribosomes—serve as a signal to halt their synthesis. (B) An mRNA from the pathogen *Listeria monocytogenes* contains a "thermosensor" RNA sequence that controls the translation of a set of mRNAs produced by virulence genes. At the warmer temperature that the bacterium encounters inside its human host, the thermosensor sequence denatures, exposing the ribosome-binding sequence, so the virulence proteins are made.

Regulatory RNAs Control the Expression of Thousands of Genes

As we saw in Chapter 7, RNAs perform many critical tasks in cells. In addition to the mRNAs, which code for proteins, *noncoding RNAs* have various functions. It has long been known that some have key structural and catalytic roles, particularly in protein synthesis by ribosomes (see pp. 246–247). But a recent series of surprising discoveries has revealed several new classes of noncoding RNAs and shown that these RNAs are far more prevalent than previously suspected.

What, then, are all these newly discovered noncoding RNAs doing? Many have unanticipated but important roles in regulating gene expression and are therefore referred to as **regulatory RNAs**. There are at least three major types of regulatory RNAs—*microRNAs*, *small interfering RNAs*, and *long noncoding RNAs*. We discuss each one in turn.

MicroRNAs Direct the Destruction of Target mRNAs

MicroRNAs, or **miRNAs**, are tiny RNA molecules that control gene expression by base-pairing with specific mRNAs and reducing both their stability and their translation into protein. In humans, miRNAs are thought to regulate the expression of at least one-third of all protein-coding genes.

Like other noncoding RNAs, such as tRNA and rRNA, a precursor miRNA transcript undergoes a special type of processing to yield the mature, functional miRNA molecule, which is only about 22 nucleotides in length. This small but mature miRNA is packaged with specialized proteins to form an *RNA-induced silencing complex* (*RISC*), which patrols the cytoplasm in search of mRNAs that are complementary to the bound miRNA molecule (**Figure 8–25**). Once a target mRNA forms base pairs with an miRNA, it is either destroyed immediately by a nuclease present within the RISC or its translation is blocked. In the latter case, the bound mRNA molecule is delivered to a region of the cytoplasm where other nucleases eventually degrade it. Destruction of the mRNA releases the RISC and allows it to seek out additional mRNA targets. Thus, a single miRNA—as part of a RISC—can eliminate one mRNA molecule after another, thereby efficiently blocking production of the protein that the mRNAs encode.

Two features of miRNAs make them especially useful regulators of gene expression. First, a single miRNA can inhibit the transcription of a whole set of different mRNAs so long as all the mRNAs carry a common sequence, usually located in either their 5′ or 3′ untranslated regions. In humans, some individual miRNAs influence the transcription of hundreds of different mRNAs in this manner. Second, a gene that encodes an miRNA occupies relatively little space in the genome compared with one that encodes a transcription regulator. Indeed, their very small size is one reason that miRNAs were discovered only recently. There are thought

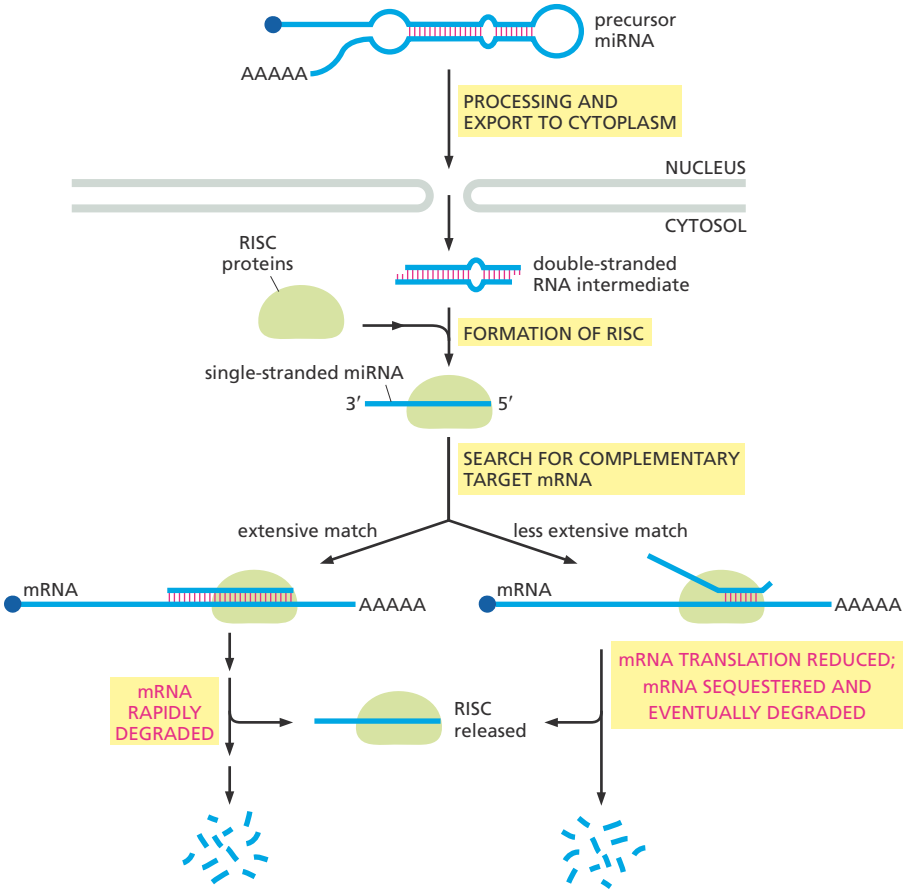


Figure 8–25 An miRNA targets a complementary mRNA molecule for destruction. Each precursor miRNA transcript is processed to form a double-stranded intermediate, which is further processed to form a mature, single-stranded miRNA. This miRNA assembles with a set of proteins into a complex called RISC, which then searches for mRNAs that have a nucleotide sequence complementary to its bound miRNA. Depending on how extensive the region of complementarity is, the target mRNA is either rapidly degraded by a nuclease within the RISC or transferred to an area of the cytoplasm where other cellular nucleases destroy it.

to be roughly 500 different miRNAs encoded by the human genome. Although we are only beginning to understand the full impact of these miRNAs, it is clear that they play a critical part in regulating gene expression and thereby influence many cell functions.

Small Interfering RNAs Are Produced From Double-Stranded, Foreign RNAs to Protect Cells From Infections

Some of the same components that process and package miRNAs also play another crucial part in the life of a cell: they serve as a powerful cell defense mechanism. In this case, the system is used to eliminate "foreign" RNA molecules—in particular, the double-stranded RNAs produced by many viruses and transposable genetic elements (discussed in Chapter 9). The process is called **RNA interference** (**RNAi**).

In the first step of RNAi, the double-stranded, foreign RNAs are cut into short fragments (approximately 22 nucleotide pairs in length) by a protein called Dicer—the same protein used to generate the double-stranded RNA intermediate in miRNA production (see Figure 8–25). The resulting double-stranded RNA fragments, called **small interfering RNAs** (**siRNAs**), are then taken up by the same RISCs that carry miRNAs. The RISC discards one strand of the siRNA duplex and uses the remaining single-stranded RNA to seek and destroy complementary foreign RNA molecules (**Figure 8–26**). In this way, the infected cell turns the foreign RNA back on itself.

RNAi operates in a wide variety of organisms, including single-celled fungi, plants, and worms, indicating that it is an evolutionarily ancient defense mechanism. In some organisms, including plants, the RNAi defense response can spread from tissue to tissue, allowing the entire organism to become resistant to a virus after only a few of its cells

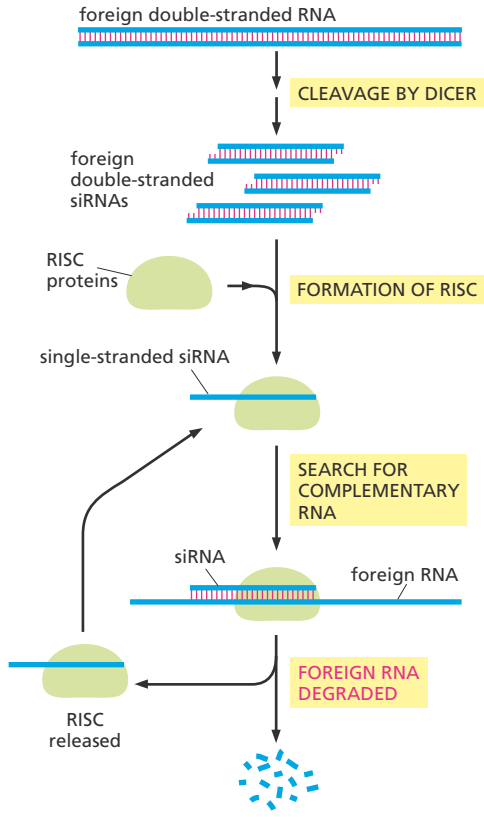


Figure 8–26 siRNAs are produced from double-stranded, foreign RNAs in the process of RNA interference. Double-stranded RNAs from a virus or transposable genetic element are first cleaved by a nuclease called Dicer. The resulting double-stranded fragments are incorporated into RISCs, which discard one strand of the foreign RNA duplex and use the other strand to locate and destroy foreign RNAs with a complementary sequence.

have been infected. In this sense, RNAi resembles certain aspects of the adaptive immune responses of vertebrates; in both cases, an invading pathogen elicits the production of molecules—either siRNAs or antibodies—that are custom-made to inactivate the specific invader and thereby protect the host.

Thousands of Long Noncoding RNAs May Also Regulate Mammalian Gene Activity

At the other end of the size spectrum are the **long noncoding RNAs**, a class of RNA molecules that are more than 200 nucleotides in length. There are thought to be upwards of 8000 of these RNAs encoded in the human and mouse genomes. Yet, with few exceptions, their roles in the biology of the organism are not entirely clear.

One of the best understood of the long noncoding RNAs is *Xist*. This enormous RNA molecule, some 17,000 nucleotides long, is a key player in X inactivation—the process by which one of the two X chromosomes in the cells of female mammals is permanently silenced (see Figure 5–30). Early in development, Xist is produced by only one of the X chromosomes in each female nucleus. The transcript then "sticks around," coating the chromosome and presumably attracting the enzymes and chromatin-remodeling complexes that promote the formation of highly condensed heterochromatin. Other long noncoding RNAs may promote the silencing of specific genes in a similar manner.

Some long noncoding RNAs arise from protein-coding regions of the genome, but are transcribed from the "wrong" DNA strand. Some of these *antisense* transcripts are known to bind to the mRNAs produced from that DNA segment, regulating their translation and stability—in some cases by producing siRNAs (see Figure 8–26).

Regardless of how the various long noncoding RNAs operate—or what exactly they do—the discovery of this large class of RNAs reinforces the idea that a eukaryotic genome is densely packed with information that provides not only an inventory of the molecules and structures every cell must make, but a set of instructions for how and when to assemble these parts to guide the growth and development of a complete organism.

ESSENTIAL CONCEPTS

- **A typical eukaryotic cell expresses only a fraction of its genes, and the distinct types of cells in multicellular organisms arise because different sets of genes are expressed as cells differentiate.**
- **In principle, gene expression can be controlled at any of the steps between a gene and its ultimate functional product. For the majority of genes, however, the initiation of transcription is the most important point of control.**
- **The transcription of individual genes is switched on and off in cells by transcription regulator proteins, which bind to short stretches of DNA called regulatory DNA sequences.**
- **In bacteria, transcription regulators usually bind to regulatory DNA sequences close to where RNA polymerase binds. This binding can either activate or repress transcription of the gene. In eukaryotes, regulatory DNA sequences are often separated from the promoter by many thousands of nucleotide pairs.**
- **Eukaryotic transcription regulators act in two main ways: (1) they can directly affect the assembly process that requires RNA polymerase**

and the general transcription factors at the promoter, and (2) they can locally modify the chromatin structure of promoter regions.

- In eukaryotes, the expression of a gene is generally controlled by a combination of different transcription regulator proteins.
- In multicellular plants and animals, the production of different transcription regulators in different cell types ensures the expression of only those genes appropriate to the particular type of cell.
- One differentiated cell type can be converted to another by artificially expressing an appropriate set of transcription regulators. A differentiated cell can also be reprogrammed into a stem cell by artificially expressing a particular set of such regulators.
- Cells in multicellular organisms have mechanisms that enable their progeny to "remember" what type of cell they should be. A prominent mechanism for propagating cell memory relies on transcription regulators that perpetuate transcription of their own gene—a form of positive feedback.
- A master transcription regulator, if expressed in the appropriate precursor cell, can trigger the formation of a specialized cell type or even an entire organ.
- The pattern of DNA methylation can be transmitted from one cell generation to the next, producing a form of epigenetic inheritance that helps a cell remember the state of gene expression in its parent cell. There is also evidence for a form of epigenetic inheritance based on transmitted chromatin structures.
- Cells can regulate gene expression by controlling events that occur after transcription has begun. Many of these post-transcriptional mechanisms rely on RNA molecules that can influence their own stability or translation.
- MicroRNAs (miRNAs) control gene expression by base-pairing with specific mRNAs and inhibiting their stability and translation.
- Cells have a defense mechanism for destroying "foreign" double-stranded RNAs, many of which are produced by viruses. It makes use of small interfering RNAs (siRNAs) that are produced from the foreign RNAs in a process called RNA interference (RNAi).
- Scientists can take advantage of RNAi to inactivate specific genes of interest.
- The recent discovery of thousands of long noncoding RNAs in mammals has opened a new window to the roles of RNAs in gene regulation.

KEY TERMS

combinatorial control
differentiation
DNA methylation
epigenetic inheritance
gene expression
long noncoding RNA
microRNA (miRNA)
positive feedback loop
post-transcriptional control
promoter
regulatory DNA sequence
regulatory RNA
reporter gene
RNA interference (RNAi)
small interfering RNA (siRNA)
transcription regulator
transcriptional activator
transcriptional repressor

QUESTIONS

QUESTION 8–4

A virus that grows in bacteria (bacterial viruses are called bacteriophages) can replicate in one of two ways. In the prophage state, the viral DNA is inserted into the bacterial chromosome and is copied along with the bacterial genome each time the cell divides. In the lytic state, the viral DNA is released from the bacterial chromosome and replicates many times in the cell. This viral DNA then produces viral coat proteins that together with the replicated viral DNA form many new virus particles that burst out of the bacterial cell. These two forms of growth are controlled by two transcription regulators, called c1 ("c one") and Cro, that are encoded by the virus. In the prophage state, cI is expressed; in the lytic state, Cro is expressed. In addition to regulating the expression of other genes, c1 represses the *Cro* gene, and Cro represses the *c1* gene (**Figure Q8–4**). When bacteria containing a phage in the prophage state are briefly irradiated with UV light, c1 protein is degraded.

A. What will happen next?

B. Will the change in (A) be reversed when the UV light is switched off?

C. Why might this response to UV light have evolved?

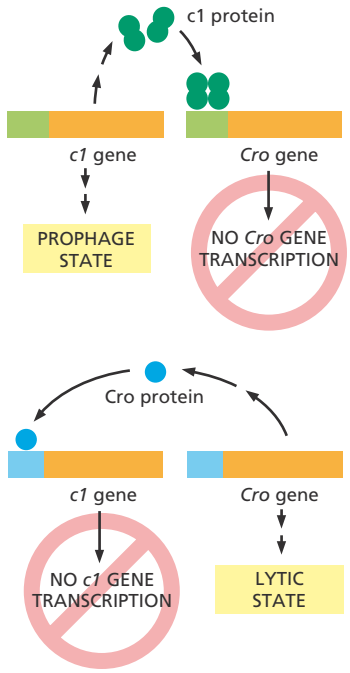


Figure Q8–4

QUESTION 8–5

Which of the following statements are correct? Explain your answers.

A. In bacteria, but not in eukaryotes, many mRNAs contain the coding region for more than one gene.

B. Most DNA-binding proteins bind to the major groove of the DNA double helix.

C. Of the major control points in gene expression (transcription, RNA processing, RNA transport, translation, and control of a protein's activity), transcription initiation is one of the most common.

QUESTION 8–6

Your task in the laboratory of Professor Quasimodo is to determine how far an enhancer (a binding site for an activator protein) could be moved from the promoter of the *straightspine* gene and still activate transcription. You systematically vary the number of nucleotide pairs between these two sites and then determine the amount of transcription by measuring the production of Straightspine mRNA. At first glance, your data look confusing (**Figure Q8–6**). What would you have expected for the results of this experiment? Can you save your reputation and explain these results to Professor Quasimodo?

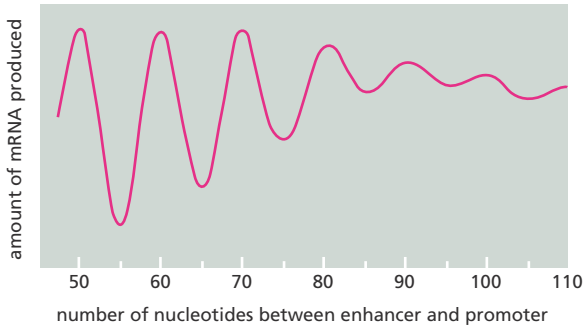


Figure Q8–6

QUESTION 8–7

The λ repressor binds as a dimer to critical sites on the λ genome to repress the virus's lytic genes. This is necessary to maintain the prophage (integrated) state. Each molecule of the repressor consists of an N-terminal DNA-binding domain and a C-terminal dimerization domain (**Figure Q8–7**). Upon induction (for example, by irradiation with UV light), the genes for lytic growth are expressed, λ progeny are produced, and the bacterial cell is lysed (see Question 8–4). Induction is initiated by cleavage of the λ repressor at a site between the DNA-binding domain and the dimerization domain, which causes the repressor to dissociate from the DNA. In the absence of bound repressor, RNA polymerase binds and initiates lytic growth. Given that the number (concentration) of DNA-binding domains is unchanged by cleavage of the repressor, why do you suppose its cleavage results in its dissociation from the DNA?

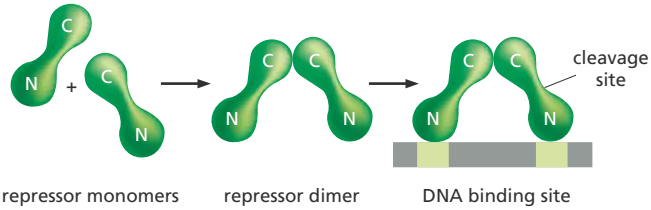


Figure Q8–7

QUESTION 8–8

The genes that encode the enzymes for arginine biosynthesis are located at several positions around the genome of *E. coli*, and they are regulated coordinately by a transcription regulator encoded by the *ArgR* gene.

The activity of the ArgR protein is modulated by arginine. Upon binding arginine, ArgR alters its conformation, dramatically changing its affinity for the DNA sequences in the promoters of the genes for the arginine biosynthetic enzymes. Given that ArgR is a repressor protein, would you expect that ArgR would bind more tightly or less tightly to the DNA sequences when arginine is abundant? If ArgR functioned instead as an activator protein, would you expect the binding of arginine to increase or to decrease its affinity for its regulatory DNA sequences? Explain your answers.

QUESTION 8–9

When enhancers were initially found to influence transcription many thousands of nucleotide pairs from the promoters they control, two principal models were invoked to explain this action at a distance. In the "DNA looping" model, direct interactions between proteins bound at enhancers and promoters were proposed to stimulate transcription initiation. In the "scanning" or "entry-site" model, RNA polymerase (or another component of the transcription machinery) was proposed to bind at the enhancer and then scan along the DNA until it reached the promoter. These two models were tested using an enhancer on one piece of DNA and a β-globin gene and promoter on a separate piece of DNA (**Figure Q8–9**). The β-globin gene was not expressed from the mixture of pieces. However, when the two segments of DNA were joined via a linker (made of a protein that binds to a small molecule called biotin), the β-globin gene was expressed.

Does this experiment distinguish between the DNA looping model and the scanning model? Explain your answer.

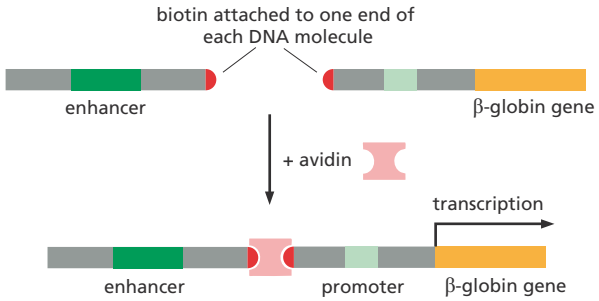


Figure Q8–9

QUESTION 8–10

Differentiated cells of an organism contain the same genes. (Among the few exceptions to this rule are the cells of the mammalian immune system, in which the formation of specialized cells is based on limited rearrangements of the genome.) Describe an experiment that substantiates the first sentence of this question, and explain why it does.

QUESTION 8–11

Figure 8–17 shows a simple scheme by which three transcription regulators are used during development to create eight different cell types. How many cell types could you create, using the same rules, with four different transcription regulators? As described in the text, MyoD is a transcription regulator that by itself is sufficient to induce muscle-specific gene expression in fibroblasts. How does this observation fit the scheme in Figure 8–17?

QUESTION 8–12

Imagine the two situations shown in **Figure Q8–12**. In cell I, a transient signal induces the synthesis of protein A, which is a transcriptional activator that turns on many genes including its own. In cell II, a transient signal induces the synthesis of protein R, which is a transcriptional repressor that turns off many genes including its own. In which, if either, of these situations will the descendants of the original cell "remember" that the progenitor cell had experienced the transient signal? Explain your reasoning.

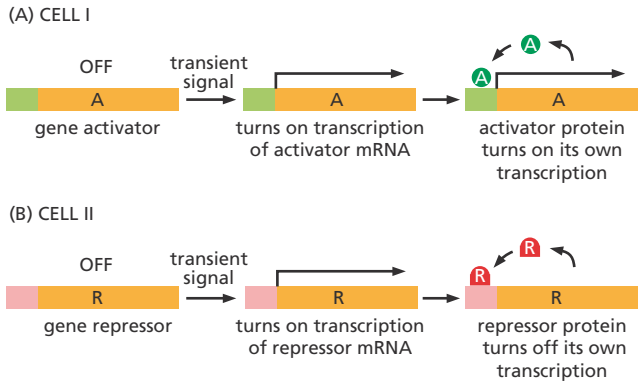


Figure Q8–12

QUESTION 8–13

Discuss the following argument: "If the expression of every gene depends on a set of transcription regulators, then the expression of these regulators must also depend on the expression of other regulators, and their expression must depend on the expression of still other regulators, and so on. Cells would therefore need an infinite number of genes, most of which would code for transcription regulators." How does the cell get by without having to achieve the impossible?

Page left intentionally blank