CHAPTER 7

Control of Gene Expression

An organism's DNA encodes all of the RNA and protein molecules required to construct 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 of a human—no more enables us to reconstruct the organism than a list of English words enables us to reconstruct a play by Shakespeare. In both cases, the problem is to know how the elements in the DNA sequence or the words on the list are used. Under what conditions is each gene product made, and, once made, what does it do?

In this chapter, we focus on the first half of this problem—the rules and mechanisms that enable a subset of genes to be selectively expressed in each cell and also determine the amount of each gene product. These mechanisms operate at many levels, and we shall discuss each level in turn. But first we present some of the basic principles involved.

IN THIS CHAPTER

AN OVERVIEW OF GENE CONTROL

The different cell types in a multicellular organism differ dramatically in both structure and function. If we compare a mammalian neuron with a liver cell, for example, the differences are so extreme that it is difficult to imagine that the two cells contain the same genome (**Figure 7–1**). For this reason, and because cell differentiation often seemed irreversible, biologists originally suspected that genes might be selectively lost when a cell differentiates. We now know, however, that cell differentiation generally occurs without changes in the nucleotide sequence of a cell's genome.

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

The cell types in a multicellular organism become different from one another because they synthesize and accumulate different sets of RNA and protein molecules. The initial evidence that they do this without altering the sequence of their DNA came from a classic set of experiments in frogs. When the nucleus of a fully differentiated frog cell is injected into a frog egg whose nucleus has been

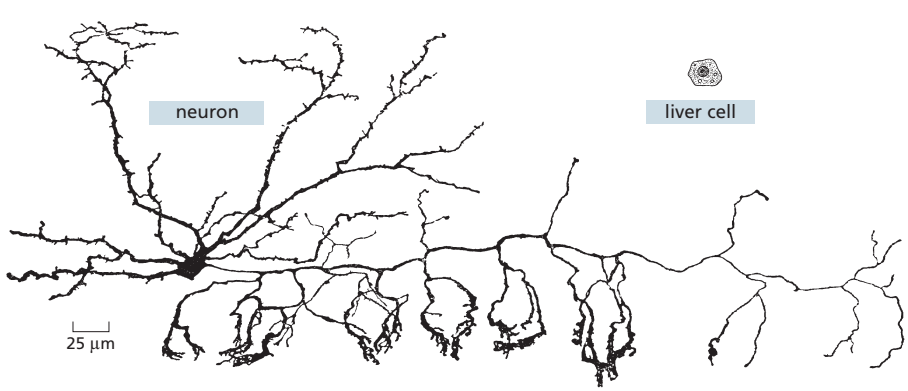


Figure 7–1 A neuron and a liver cell share the same genome. The long branches of this neuron from the retina enable it to receive electrical signals from many other neurons and convey them to neighboring neurons. The liver cell, which is drawn to the same scale, is involved in many metabolic processes, including digestion and the detoxification of alcohol and other drugs. Both of these mammalian cells contain the same genome, but they express different sets of RNAs and proteins. (Neuron adapted from S. Ramón y Cajal, Histologie du Système Nerveux de l'Homme et de Vertébrés, 1909–1911. Paris: A. Maloine Éditeur; reprinted, Madrid: C.S.I.C., 1972.)

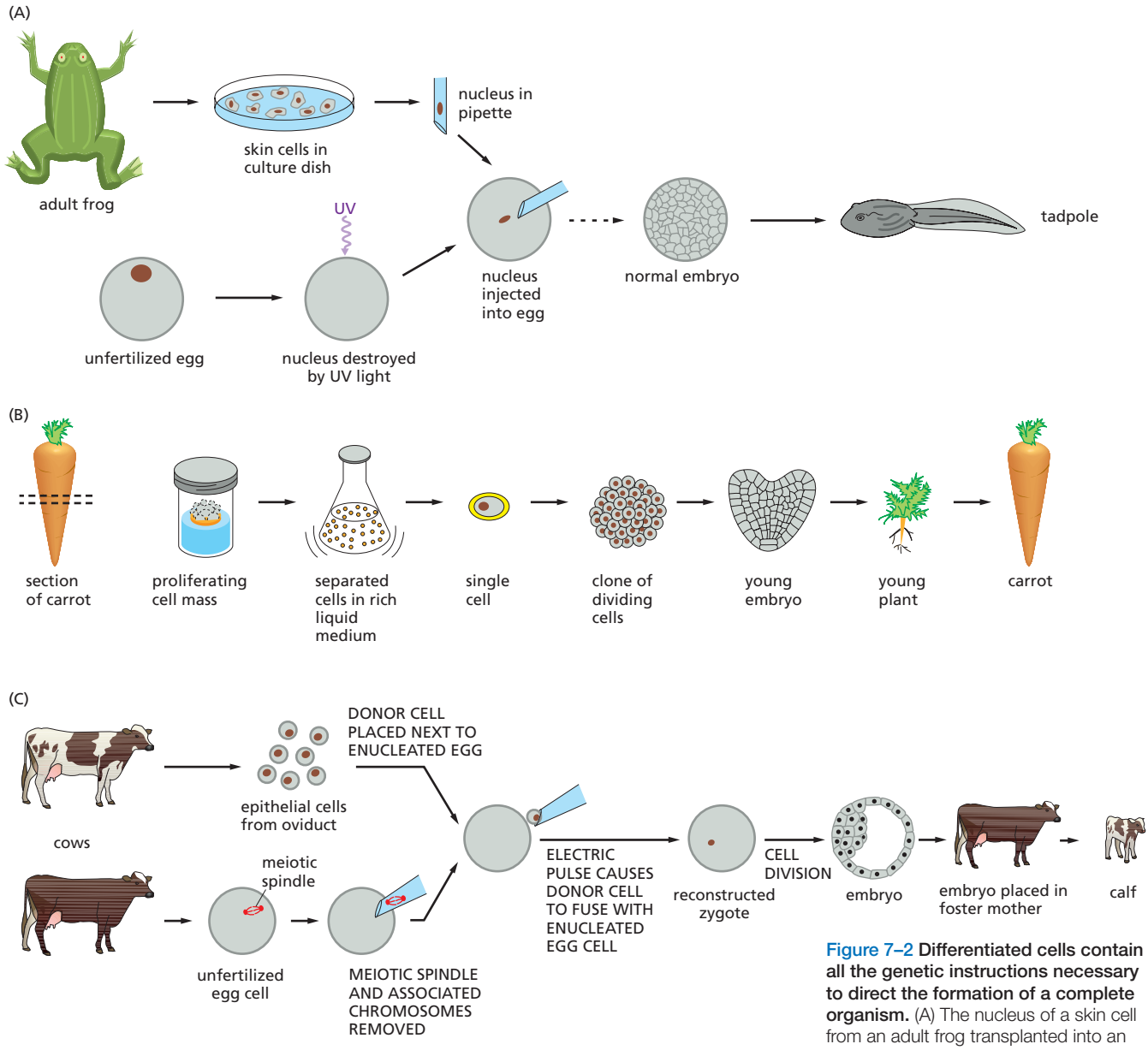


Figure 7–2 Differentiated cells contain all the genetic instructions necessary to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an enucleated egg can give rise to an entire tadpole. The *broken arrow* indicates that, to give the transplanted genome time to adjust to an embryonic environment, a further transfer step is required in which one of the nuclei is taken from an 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 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 and introduced into an enucleated egg from a different cow can 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.)

removed, the injected donor nucleus is capable of directing the recipient egg to produce a normal tadpole (Figure 7–2A). Because this tadpole contains a full range of differentiated cells, each of which derived their DNA sequences from the nucleus of the original donor skin cell, that differentiated cell cannot have lost any important DNA sequences. Experiments performed with plants produced a similar conclusion. When differentiated pieces of plant tissue are placed in culture and then dissociated into single cells, often one of these individual cells can regenerate an entire adult plant (Figure 7–2B). More recently, the same principle has been demonstrated for mammals that include sheep, mice, pigs, goats, dogs, and cattle (Figure 7–2C).

Detailed DNA sequencing of genomes present in different tissues also shows that the changes in gene expression that underlie the normal development of multicellular organisms do not generally involve changes in the DNA sequence of the genome.

Different Cell Types Synthesize Different Sets of RNAs and Proteins

As a first step in understanding cell differentiation, we would like to know how many differences there are between any one cell type and another.

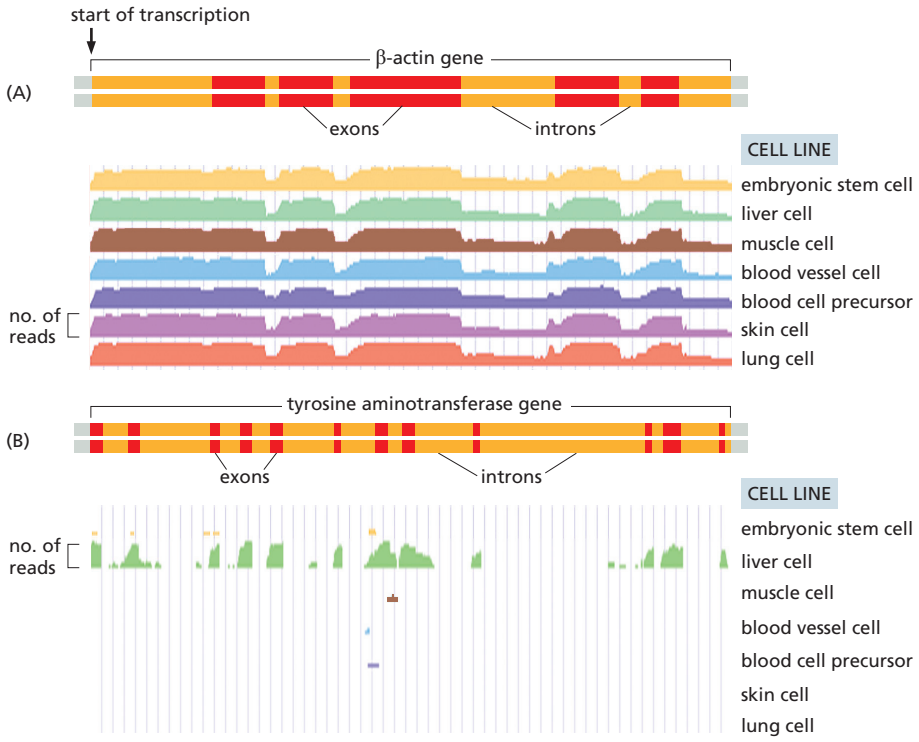


Figure 7–3 Differences in RNA levels for two human genes in seven different tissues. To obtain RNA data by the technique known as RNA-seq (see pp. 514–516), RNA was collected from seven different human cell lines grown in culture, each derived from a different tissue. Millions of "sequence reads" were obtained for each RNA sample and mapped by matching RNA sequences to the DNA sequence of the human genome. At each position along the genome, the height of the colored trace is proportional to the number of sequence reads that match the genome sequence at that point. As seen in the figure, the exon sequences in transcribed genes are present at high levels, reflecting their presence in mature mRNAs. Intron sequences are present at much lower levels and reflect pre-mRNA molecules that have not yet been spliced, plus intron sequences that have been spliced out but not yet degraded. (A) The data for one of the genes coding for actin, a major component of the cytoskeleton in all cells. Note that the left-hand end of the mature β-actin mRNA is not translated into protein. As explained later in this chapter, many mRNAs have 5′ untranslated regions that regulate their translation into protein. (B) The same type of data displayed for the enzyme tyrosine aminotransferase, which is highly expressed in liver cells but not in the other cell types tested. [Information for both panels from the University of California, Santa Cruz, Genome Browser (https://genome.ucsc.edu), which provides this type of information for every human gene. See also S. Djebali et al., *Nature* 489:101–108, 2012.]

Although we still do not have an exact answer for each cell type, we can make several general statements.

1. Many processes are common to all cells, and any two cells in a single organism therefore have many gene products in common. These include the structural proteins of chromosomes, RNA and DNA polymerases, DNA repair enzymes, ribosomal proteins and RNAs, the enzymes that catalyze the central reactions of metabolism, and many of the proteins that form the cytoskeleton such as actin (**Figure 7–3A**).

2. Some RNAs and proteins are abundant in the specialized cells in which they function and cannot be detected elsewhere, even by sensitive tests. Hemoglobin, for example, is expressed specifically in red blood cells, where it carries oxygen, whereas the enzyme tyrosine aminotransferase (which breaks down tyrosine in food) is expressed in liver but not in most other tissues (**Figure 7–3B**).

3. Analyses of RNAs reveal that, at any one time, a typical human cell expresses 30–60% of its approximately 25,000 genes at some meaningful level. There are about 20,000 protein-coding genes and an estimated 5000 noncoding RNA genes in humans. When the patterns of RNA expression in different human cell lines are compared, the level of expression of almost every gene is found to vary from one cell type to another. A few of these differences are striking, like those of hemoglobin and tyrosine aminotransferase noted above, but most are much more subtle. But even those genes that are expressed in all cell types usually vary in their *level* of expression from one cell type to the next.

4. Although there are striking differences in the protein-coding RNAs (mRNAs) in specialized cell types, they underestimate the full range of differences in the final pattern of protein production. As we shall discuss later in this chapter, there are many steps after RNA production at which gene expression can be regulated. And, as we saw in Chapter 3, proteins are often covalently modified after they are synthesized. The differences in gene expression between cell types are therefore most fully revealed through methods that directly display the levels of proteins, along with their post-translational modifications (**Figure 7–4**).

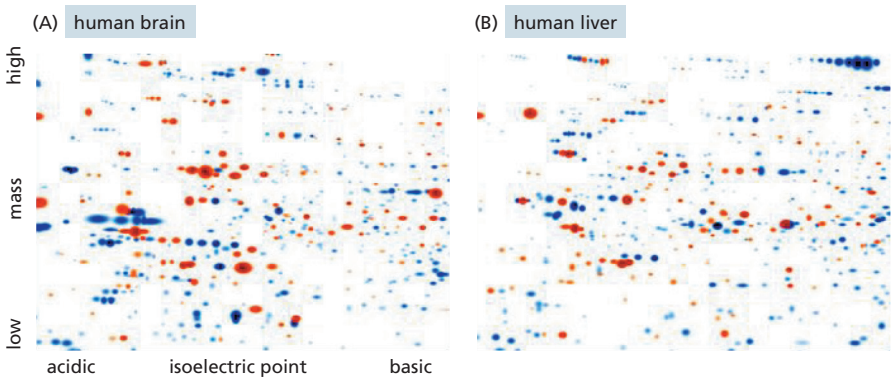


Figure 7–4 Differences in the proteins expressed by two human tissues, (A) brain and (B) liver. The proteins have been separated by size (*top* to *bottom*) and isoelectric point, the pH at which the protein has no net charge (*right* to *left*). The protein spots artificially colored *red* are common to both samples; those in *blue* are specific to that tissue. The differences between the two tissue samples vastly outweigh their similarities: even for proteins that are shared between the two tissues, their relative abundances are usually different. Note that this technique separates proteins by both size and charge; therefore, a protein that has several different phosphorylation states will appear as a series of *horizontal spots* (see *upper right-hand* portion of *right* panel). Only a small portion of the complete protein spectrum is shown for each sample.

The method used to display proteins in these panels is known as *two-dimensional gel electrophoresis* (see Figure 8–16). Although it is useful for easily visualizing the extent of protein differences between the two cell types, newer methods based on mass spectrometry (see pp. 491–492) provide much more detailed information, including the identity of each protein, the position of each modification, and the nature of the modification. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

The Spectrum of mRNAs Present in a Cell Can Be Used to Accurately Identify the Cell Type

We have seen that each cell type produces a characteristic set of mRNAs. Therefore, if all the mRNAs present in a cell are known, the cell type can be unambiguously identified, using prior knowledge from cell lines or analyses of tissues. This approach is made possible by the ability to determine the nucleotide sequence of all the mRNAs produced by a single cell (see pp. 537–538). Thus, for example, because human cells have approximately 20,000 mRNA-producing genes, this strategy provides very fine resolution of the differences among our different individual cells.

In general, the mRNA approach agrees well with the traditional categorization of cell types that is based on staining and microscopy, but the mRNA strategy has also revealed that many cells that "look" the same can differ significantly in their mRNA content and therefore in their function. This strategy has thereby identified many new cell types, most of which are subdivisions of cell types that had been classically defined (**Figure 7–5**). The ability to determine the mRNA content of individual cells also provides a new appreciation for how cells present in a tissue (liver, for example) differ according to their positions in the tissue.

External Signals Can Cause a Cell to Change the Expression of Its Genes

Although the specialized cells in a multicellular organism have characteristic patterns of gene expression, each cell is capable of altering its pattern of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of a set of proteins is dramatically increased, and once the hormone is no longer present, the production of these proteins drops back to its normal, unstimulated level. Glucocorticoids are released in the body during periods of starvation or intense exercise, and they signal the liver to increase the production of energy from amino acids and other

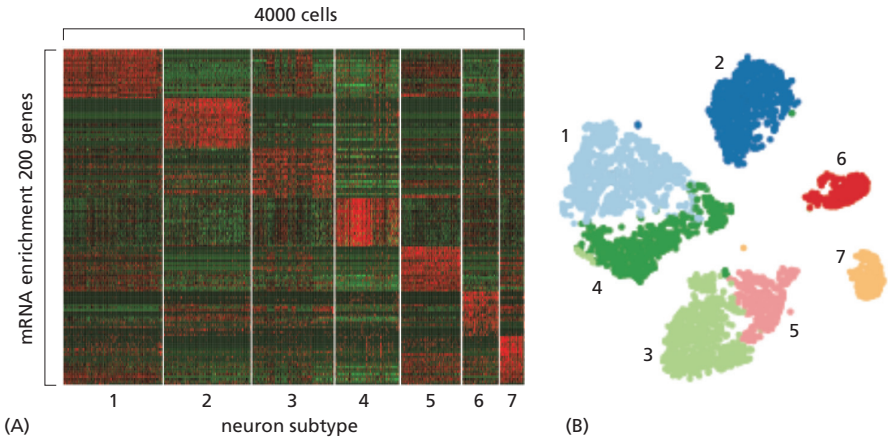


Figure 7–5 Classification of a group of neurons in the mouse brain into seven different subtypes by single-cell mRNA sequencing. For this experiment, approximately 4000 individual neurons (which were activated in response to a particular stimulus) were dissected from the brain and separated from each other. The mRNAs produced by each cell were isolated and their sequences determined by the methods described in Chapter 8. On the basis of the spectrum of the mRNAs produced by each cell, the 4000 different cells could be grouped into seven distinctive subtypes. Within each subtype, the mRNAs were similar from cell to cell, but between subtypes they differed significantly. (A) Here, the level of the mRNAs detected for approximately 200 different genes is plotted for each cell as a tiny rectangle, whose color intensity is proportional to the amount of that mRNA in that cell, with *red* indicating increased expression and *green* decreased expression, relative to all the samples. These data are plotted for each of the 4000 cells along the X axis. The cells have been arranged so that similar cells are located next to each other. In this way it is possible, using mRNA sequence data alone, to recognize seven distinctive types of neurons, as indicated. To highlight similarities, the data for the 25 mRNAs specifically enriched in each of the seven subtypes is indicated by *red blocks*. (B) By analyzing the mRNA data using a mathematical method known as *unsupervised clustering* (see Figure 8–66), the seven different subtypes can readily be distinguished on a two-dimensional "cluster diagram," with each dot representing a single cell. In addition, information regarding the extent of differences among the subgroups can also be ascertained. For example, subtypes 1 and 4 are more closely related to each other in the mRNAs they make than are subtypes 1 and 7. This type of analysis helps us to understand how the brain processes sensory information and indicates that, even though neurons may look the same under the microscope, they can differ significantly in gene expression patterns and therefore in their functions. (A and B, from M.B. Chen et al., *Nature* 587:437–442, 2020, doi 10.1038/s41586-020-2905-5. With permission from Springer Nature.)

small molecules. The set of induced proteins includes the enzyme tyrosine aminotransferase, mentioned earlier.

Other cell types respond to glucocorticoids differently. Fat cells, for example, reduce the production of tyrosine aminotransferase, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization: different cell types can respond very differently to the same extracellular signal. Other features of the gene expression pattern do not change and give each cell type its permanently distinctive character.

Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein

We have seen that differences among the various cell types of an organism depend on the particular genes that the cells express. But at what level does this control of gene expression occur? As we saw in the previous chapter, there are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus, as illustrated in **Figure 7–6**, a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling the splicing and processing of RNA transcripts (**RNA-processing control**), (3) selecting which completed mRNAs are exported from the nucleus to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), (6) selectively degrading specific protein molecules (**protein degradation control**), and (7) activating, inactivating, or localizing specific protein molecules (**protein activity control**).

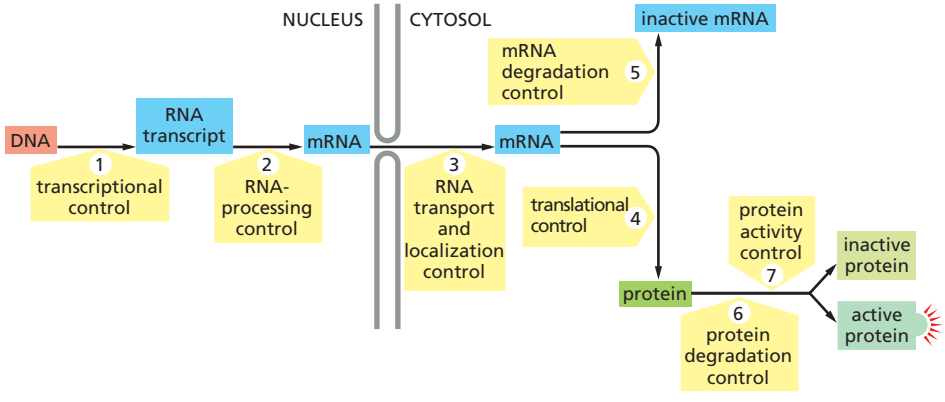


Figure 7–6 Seven steps at which eukaryotic gene expression can be controlled. Controls that operate at steps 1 through 5 are discussed in this chapter. Step 7, the regulation of protein activity, occurs largely through covalent post-translational modifications including phosphorylation, acetylation, and ubiquitylation (see Table 3–4, p. 175). Steps 6 and 7 were introduced in Chapters 3 and 6 and will be subsequently discussed in other chapters throughout the book.

For many genes, transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 7–6, only transcriptional control ensures that the cell will not synthesize superfluous intermediates. In the following sections, we discuss the DNA and protein components that regulate the initiation of gene transcription, before moving on to discuss other types of controls.

Summary

The genome of a cell contains in its entire DNA sequence the information to make many thousands of different protein and RNA molecules. But a cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. All cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. The regulation of gene expression is thus crucial for life. Although all of the many steps involved in expressing a gene can in principle be regulated, for most genes it is the initiation of RNA transcription that provides the most important point of control.

CONTROL OF TRANSCRIPTION BY SEQUENCE-SPECIFIC DNA-BINDING PROTEINS

How does a cell determine which of its thousands of genes to transcribe? Perhaps the most important concept, one that applies to all species on Earth, is based on a group of proteins known as **transcription regulators**. These proteins recognize specific sequences of DNA (typically 5–12 nucleotide pairs in length) that are often called ***cis*-regulatory sequences**, because they must be on the same chromosome (that is, *in cis*) to the genes they control. Transcription regulators bind to these sequences, which are dispersed throughout genomes, and this binding puts into motion a series of reactions that ultimately specify which genes are to be transcribed and at what rate. Approximately 10% of the protein-coding genes of most organisms are devoted to transcription regulators, making them one of the largest classes of proteins in the cell. A given transcription regulator typically recognizes a specific *cis*-regulatory sequence that is different from those recognized by the other transcription regulators in the cell.

The transcription of each gene is, in turn, controlled by its unique collection of *cis*-regulatory DNA sequences, which thus constitute a crucial part of the information coded in genomes. These sequences typically lie near the gene, often in the intergenic region directly upstream from the transcription start point of the gene. Although a few genes are controlled by a single *cis*-regulatory sequence that is recognized by a single transcription regulator, the majority have complex arrangements of *cis*-regulatory sequences, each of which is recognized by a different transcription regulator. It is therefore the positions, identity, and arrangement of *cis*-regulatory sequences that ultimately determine the time and place that each gene is transcribed. We begin our discussion by describing how transcription regulators "read" the information present in *cis*-regulatory sequences; later in the chapter, we shall discuss how they carry out their functions.

The Sequence of Nucleotides in the DNA Double Helix Can Be Read by Proteins

As discussed in Chapter 4, the DNA in a chromosome consists of a very long double helix that has both a major and a minor groove (**Figure 7–7**). Transcription regulators must recognize short, specific *cis*-regulatory sequences within this structure. When first discovered in the 1960s, it was thought that these proteins might require direct access to the interior of the double helix to distinguish between one DNA sequence and another, analogous to complementary base-pairing. It is now clear, however, that the outside of the double helix is studded with DNA sequence information that transcription regulators can recognize directly: the outside edges of each base pair display distinctive patterns of hydrogen-bond

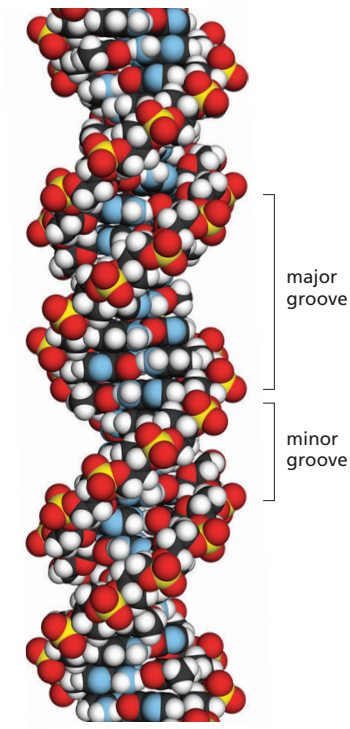
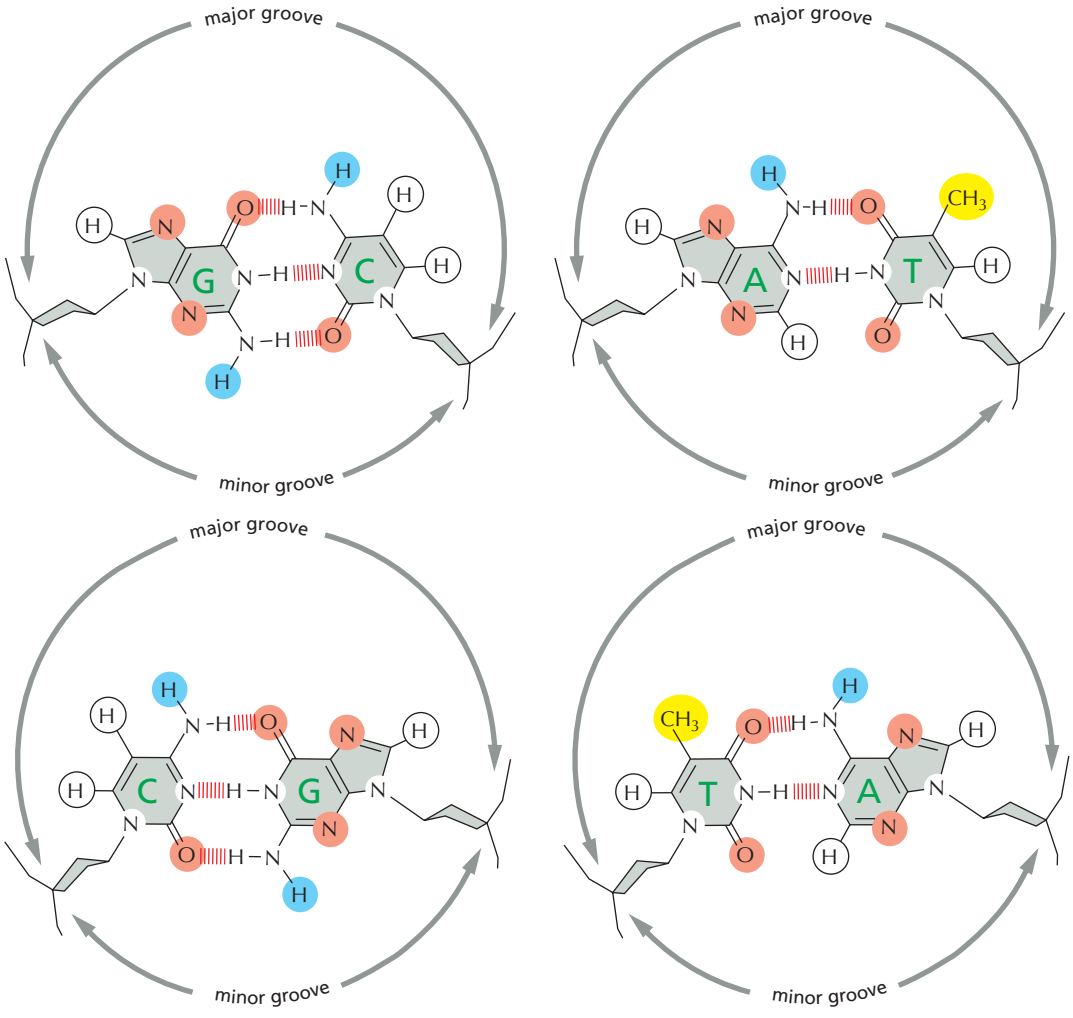


Figure 7–7 Double-helical structure of DNA. A space-filling model of DNA showing the major and minor grooves on the outside of the double helix (see Movie 4.1). The atoms are colored conventionally as follows: carbon, *black*; nitrogen, *blue*; hydrogen, *white*; oxygen, *red*; phosphorus, *yellow*.



donors, hydrogen-bond acceptors, and hydrophobic patches in both the major and minor grooves, allowing each base to be distinguished from the other three (**Figure 7–8**). Because the major groove is wider and displays more molecular features than does the minor groove, nearly all transcription regulators make the majority of their contacts with the major groove—as we shall see.

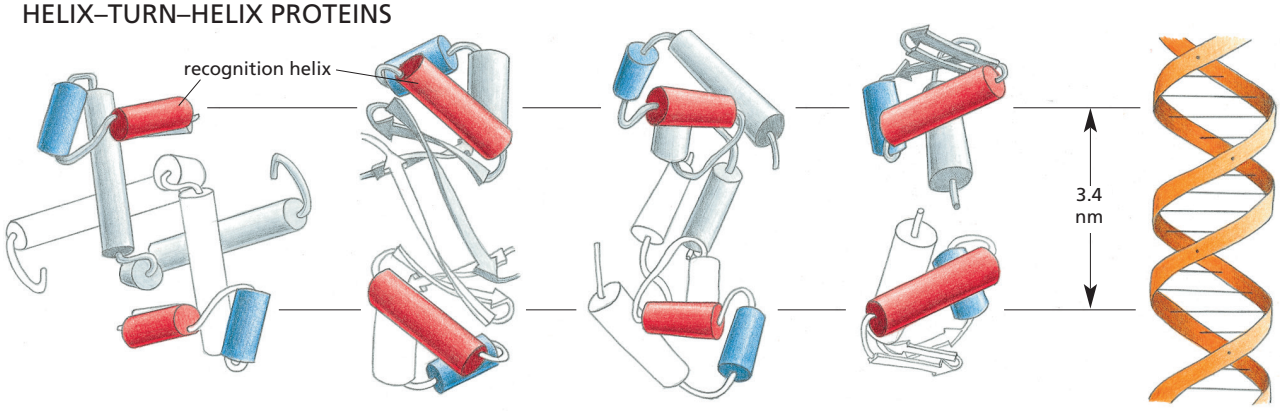
Transcription Regulators Contain Structural Motifs That Can Read DNA Sequences

Molecular recognition in biology generally relies on an exact fit between the surfaces of two molecules, and the study of transcription regulators provides some of the clearest examples of this principle. Thus, a transcription regulator recognizes its specific *cis*-regulatory sequence because the surface of the protein is complementary to surface features of the double helix that displays that sequence. Each transcription regulator makes a series of contacts with the DNA, involving hydrogen bonds, ionic bonds, and hydrophobic interactions. Although each individual contact is weak, the 20 or so contacts that are typically formed at the protein–DNA interface add together to ensure that the interaction is both highly specific and very strong (**Figure 7–9**). In fact, DNA–protein interactions include some of the tightest and most specific molecular interactions known in biology.

Although each example of protein–DNA recognition is unique in detail, x-ray crystallographic and nuclear magnetic resonance (NMR) spectroscopic studies of hundreds of transcription regulators reveal that many contain one or another of a small set of DNA-binding structural motifs (**Panel 7–1**). These motifs generally

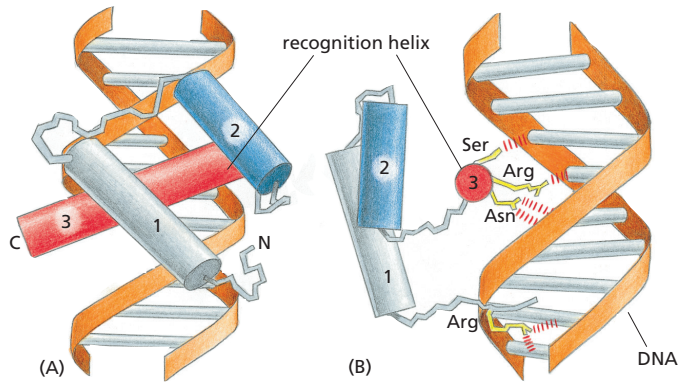
Figure 7–8 How the different base pairs in DNA can be recognized from their edges without the need to open the double helix. The four possible configurations of base pairs are shown, with potential hydrogen-bond donors indicated in *blue*, potential hydrogen-bond acceptors in *red*, and hydrogen bonds of the base pairs themselves as a series of short, parallel *red* lines. Methyl groups, which form hydrophobic protuberances, are shown in *yellow*, and hydrogen atoms that are attached to carbons, and are therefore unavailable for hydrogen-bonding, are *white*. From the major groove, each of the four base-pair configurations projects a unique pattern of features. (From C. Branden and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Publishing, 1999. With permission from Taylor & Francis.)

HELIX–TURN–HELIX PROTEINS



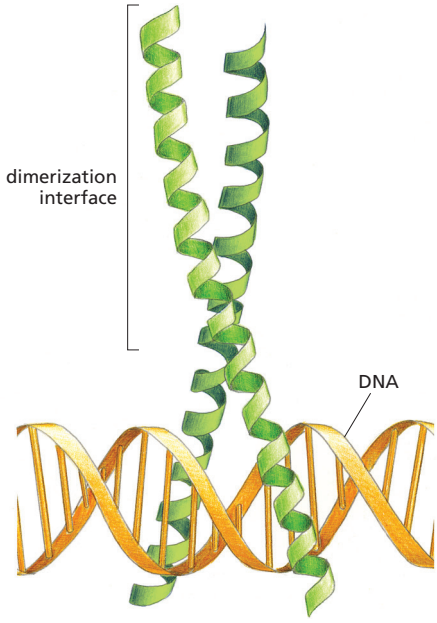
Originally identified in bacterial transcription regulators, this motif has since been found in many hundreds of DNA-binding proteins from eukaryotes, bacteria, and archaea. It is constructed from two α helices (*blue* and *red*) connected by a short extended chain of amino acids, which constitutes the "turn." The two helices are held at a fixed angle, primarily through interactions between the two helices. The more C-terminal helix (in *red*) is called the *recognition helix* because it fits into the major groove of DNA; its amino acid side chains, which differ from protein to protein, play an important part in recognizing the specific DNA sequence to which the protein binds. All of the proteins shown here bind DNA as dimers in which the two copies of the recognition helix (in *red*) are separated by exactly one turn of the DNA helix (3.4 nm); thus both recognition helices of the dimer can fit into the major groove of DNA.

HOMEODOMAIN PROTEINS



Not long after the first transcription regulators were discovered in bacteria, genetic analyses of the fruit fly *Drosophila* led to the characterization of an important class of genes, the *homeotic selector genes*, that play a critical part in orchestrating fly development (discussed in Chapter 21). It was later shown that these genes coded for transcription regulators that bound DNA through a structural motif named the homeodomain. Two different views of the same structure are shown. (A) The homeodomain is folded into three α helices, which are packed tightly together by hydrophobic interactions. The part containing helices 2 and 3 closely resembles the bacterial helix–turn–helix motif. (B) The recognition helix (helix 3, *red*) forms important contacts with the major groove of DNA. The asparagine (Asn) of helix 3, for example, contacts an adenine, as shown in Figure 7–9. A flexible arm attached to helix 1 forms contacts with nucleotide pairs in the minor groove (**Movie 7.1**).

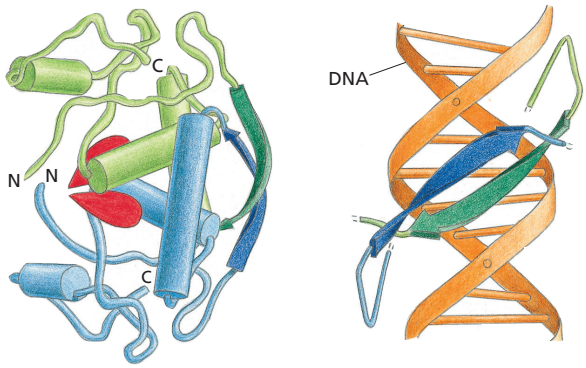
LEUCINE ZIPPER PROTEINS



The *leucine zipper* motif is named because of the way the two α helices, one from each monomer, are joined together to form a short coiled-coil. These proteins bind DNA as dimers where the two long α helices are held together by interactions between hydrophobic amino acid side chains (often on leucines) that extend from one side of each helix. Just beyond the dimerization interface, the two α helices separate from each other to form a Y-shaped structure, which allows their side chains to contact the major groove of DNA. The dimer thus grips the double helix like a clothespin on a clothesline (**Movie 7.2**).

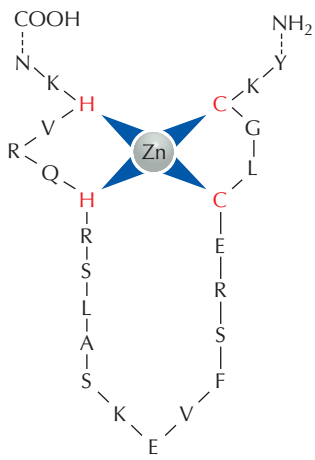
β-SHEET DNA RECOGNITION PROTEINS

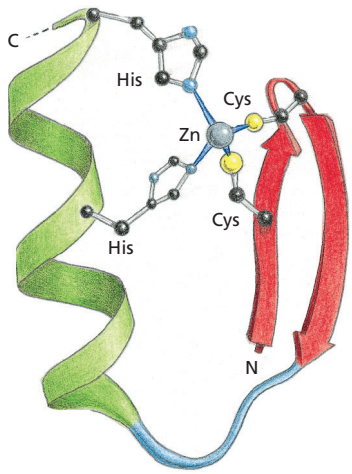
In the other DNA-binding motifs displayed in this panel, α helices are the primary mechanism used to recognize specific DNA sequences. In one group of transcription regulators, however, a two-stranded β sheet, with amino acid side chains extending from the sheet toward the DNA, reads the information on the surface of the major groove. As in the case of a recognition α helix, this β-sheet motif can be used to recognize many different DNA sequences; the exact DNA sequence recognized depends on the sequence of amino acids that make up the β sheet. Shown is a transcription regulator that binds two molecules of *S*-adenosyl methionine *(red)*. On the left is a dimer of the protein; on the right is a simplified diagram showing just the two-stranded β sheet bound to the major groove of DNA. *S*-adenosyl methionine is needed for this protein to bind DNA. Thus, the small molecule regulates the activity of the DNA-binding protein.

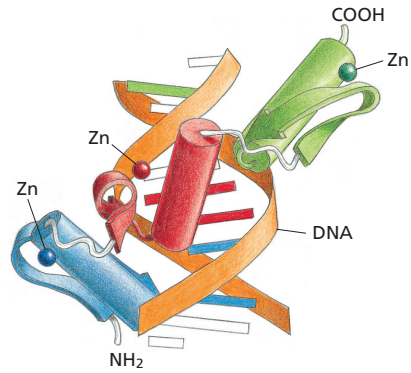


ZINC FINGER PROTEINS

This group of DNA-binding motifs includes one or more zinc atoms as structural components. All such zinc-coordinated DNA-binding motifs are called zinc fingers, referring to their appearance in early schematic drawings *(left)*. They fall into several distinct structural groups, only one of which we consider here. It has a simple structure, in which the zinc atom holds an α helix and a β sheet together *(middle)*. This type of zinc finger is often found in clusters with the α helix of each finger contacting the major groove of the DNA, forming a nearly continuous stretch of α helices along that groove (**Movie 7.3**). In this way, a strong and specific DNA–protein interaction is built up through a repeating basic structural unit. Three such fingers are shown on the *right.*

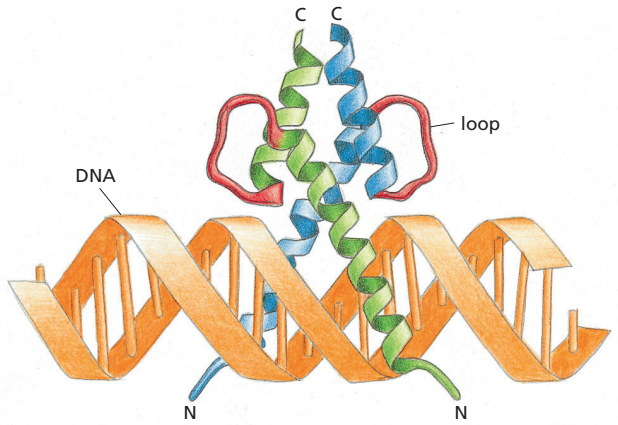


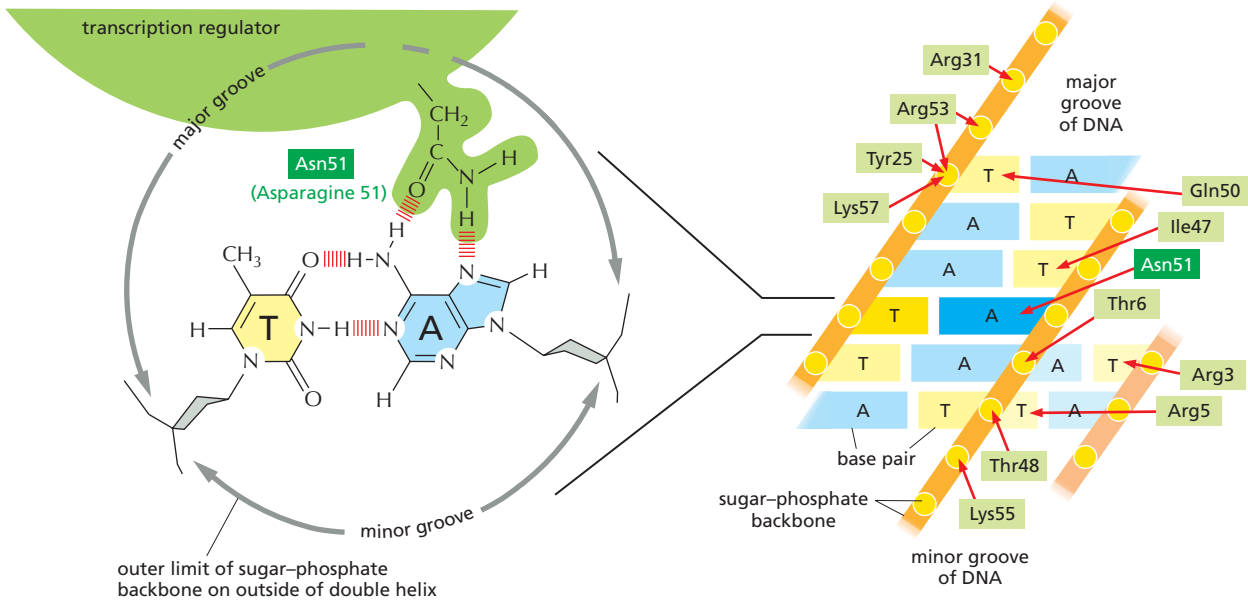




HELIX–LOOP–HELIX PROTEINS

Related to the leucine zipper, the helix–loop–helix motif consists of a short α helix connected by a loop *(red)* to a second, longer α helix. The flexibility of the loop allows one helix to fold back and park against the other thereby forming the dimerization surface. As shown, this two-helix structure binds both to DNA and to the two-helix structure of a second protein to create either a homodimer or a heterodimer. Two α helices that extend from the dimerization interface make specific contacts with the major groove of DNA.





use either α helices or β sheets to bind to the major groove of DNA, with amino acid side chains that extend from these motifs making their specific DNA contacts. Thus, a given structural motif can be used to recognize many different *cis*-regulatory sequences depending on the specific side chains that extend from it.

Dimerization of Transcription Regulators Increases Their Affinity and Specificity for DNA

A monomer of a typical transcription regulator recognizes about 4–8 nucleotide pairs of DNA. These proteins do not bind tightly to a single DNA sequence and reject all others; rather, each regulator recognizes a range of closely related sequences, with the affinity of the protein for the DNA varying according to how closely the DNA matches its optimal sequence. For this reason, the *cis*-regulatory sequence for a regulator is often depicted by a "logo" that displays the range of sequences recognized by that transcription regulator (**Figure 7–10**). In Chapter 6, this same type of representation was used to depict the DNA sequences recognized by bacterial RNA polymerase (see Figure 6–12).

The DNA sequence recognized by a monomer does not usually contain sufficient information to be picked out from the background of such sequences that would occur at random across the genome. For example, an exact six-nucleotide DNA sequence would be expected to occur by chance approximately once every

Figure 7–9 The binding of a transcription regulator to a specific DNA sequence. On the *left,* a single contact is shown between a transcription regulator and DNA; such contacts allow the protein to "read" the DNA sequence from the outside of the DNA double helix. On the *right*, the complete set of contacts between a transcription regulator (a member of the homeodomain family—see Panel 7–1) and its *cis*-regulatory sequence is shown. The DNA-binding portion of the protein is 60 amino acids long, and the amino acids that directly contact DNA are numbered beginning with the amino terminus. Although the interactions in the major groove are the most important, the protein also contacts both the minor groove and phosphates in the sugar–phosphate DNA backbone, as shown. (See C. Wolberger et al., *Cell* 67:517–528, 1991.)

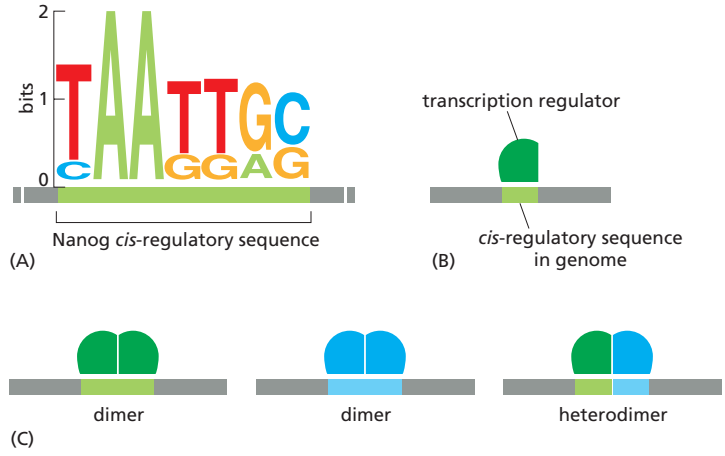


Figure 7–10 Transcription regulators and *cis*-regulatory sequences. (A) Depiction of the *cis*-regulatory sequence for Nanog, a homeodomain family member that is a key transcription regulator in embryonic stem cells. This "logo" form (see Figure 6–12) shows that the protein can recognize a collection of closely related DNA sequences and gives the preferred nucleotide pair at each position. *Cis*-regulatory sequences are almost always "read" as double-stranded DNA, but only one strand typically is shown in a logo. (B) Representation of the *cis*-regulatory sequence as a *green box* embedded in a longer DNA molecule *(gray)*. (C) Many transcription regulators form dimers (homodimers and heterodimers). In the example shown, three different DNA-binding specificities are formed from two transcription regulators.

4096 nucleotides (4^6), and the range of six-nucleotide sequences described by a typical logo would be expected to occur by chance much more often, perhaps every 1000 nucleotides. Clearly, for a bacterial genome of 4.6×10^6 nucleotide pairs, not to mention a mammalian genome of 3×10^9 nucleotide pairs, this is insufficient information to accurately control the transcription of individual genes. Additional contributions to DNA-binding specificity must therefore be present.

Many transcription regulators form dimers, with both monomers making nearly identical contacts with DNA (see Figure 7–10C). This arrangement doubles the length of the *cis*-regulatory sequence recognized and greatly increases both the affinity and the specificity of transcription regulator binding. Because the DNA sequence recognized by the protein has increased from approximately 6 nucleotide pairs to 12 nucleotide pairs, there are many fewer random occurrences of matching sequences. In many cases, heterodimers can form between two different transcription regulators, and this configuration also increases both affinity and specificity by expanding the DNA sequence recognized. Some transcription regulators can form heterodimers with more than one partner protein; in this way, the same transcription regulator can be "reused" to create several distinct DNA-binding specificities (see Figure 7–10C).

Many Transcription Regulators Bind Cooperatively to DNA

In the simplest case, the collection of noncovalent bonds that holds dimers or heterodimers together is so extensive that these structures form obligatorily and virtually never fall apart. In this case, the unit of binding is the dimer or heterodimer, and the binding curve for the transcription regulator (the fraction of DNA bound as a function of protein concentration) has a standard exponential shape (**Figure 7–11A**).

In many cases, however, the dimers and heterodimers are held together very weakly; they exist predominantly as monomers in solution, and yet dimers are observed on the appropriate DNA sequence. In this case, the proteins are said to bind to DNA cooperatively, and the curve describing their binding is S-shaped (**Figure 7–11B**). *Cooperative binding* means that, over a range of concentrations of the transcription regulator, binding is more of an all-or-none phenomenon than for noncooperative binding; that is, at most protein concentrations, the *cis*-regulatory sequence is either nearly empty or nearly fully occupied and is rarely somewhere in between. A discussion of the mathematics behind cooperative binding is given in Chapter 8 (see Figure 8–81).

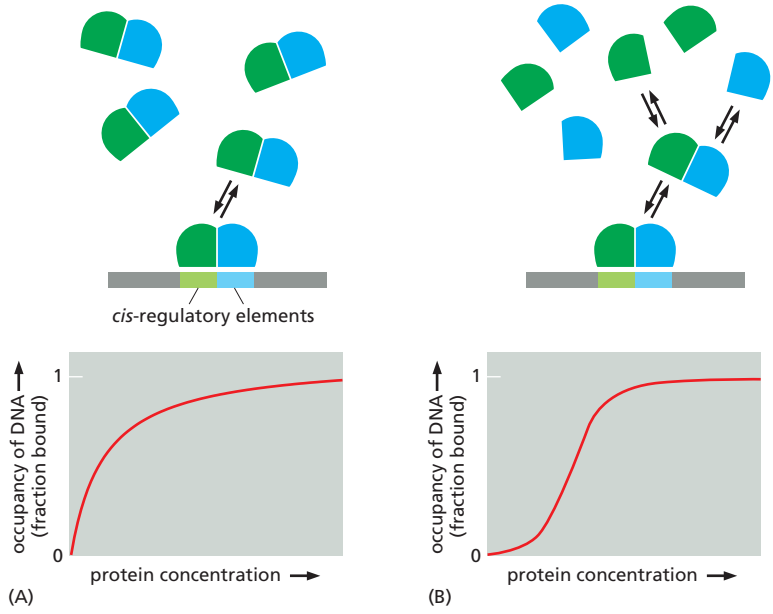


Figure 7–11 Occupancy of a *cis*-regulatory sequence by a transcription regulator. (A) Noncooperative binding by a stable heterodimer. (B) Cooperative binding by components of a heterodimer that are predominantly monomers in solution. The shape of the curve differs from that of panel A because the fraction of protein in a form competent to bind DNA (the heterodimer) increases with increasing protein concentration.

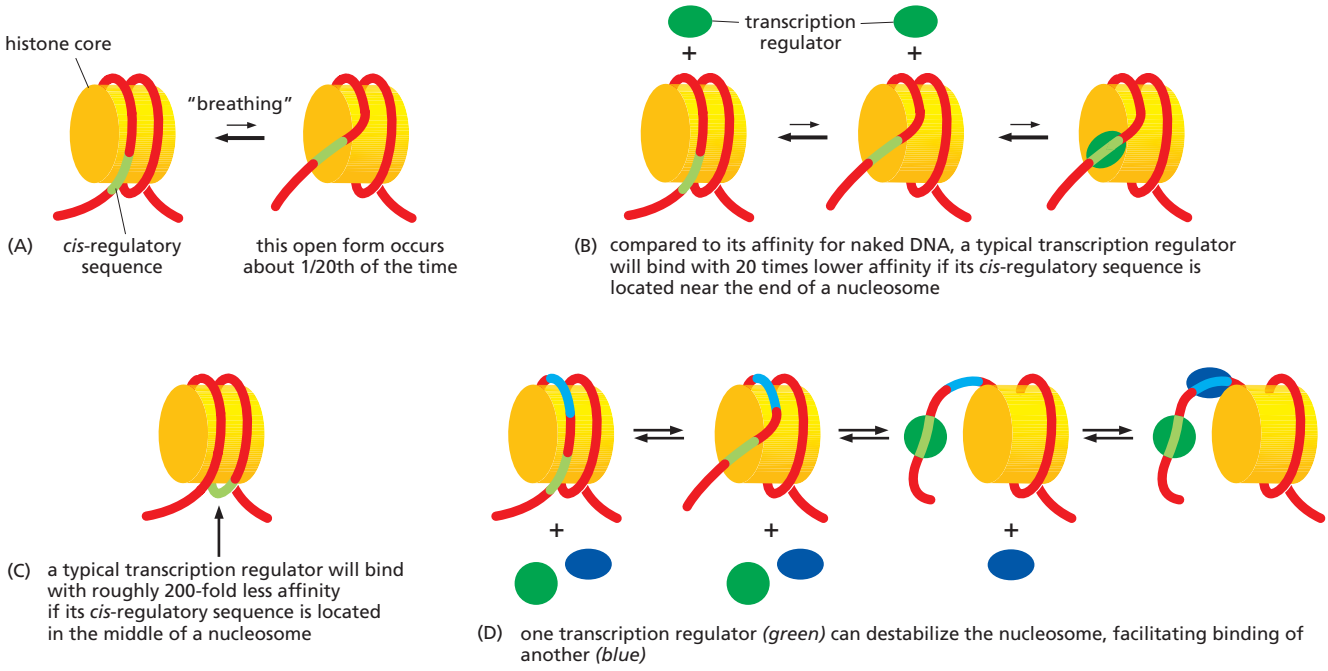


Figure 7–12 How nucleosomes affect the binding of transcription regulators.

Nucleosome Structure Promotes Cooperative Binding of Transcription Regulators

As we have just seen, cooperative binding of transcription regulators to DNA often occurs because the proteins involved have only a weak affinity for each other. However, there is a second, indirect mechanism for cooperative binding in eukaryotes, one that arises from the nucleosome structure of their chromosomes.

In general, transcription regulators bind to DNA in nucleosomes with lower affinity than they do to naked DNA. There are two reasons for this difference. First, the surface of the *cis*-regulatory sequence recognized by the transcription regulator may be facing inward on the nucleosome, toward the histone core, and therefore not be readily available to the regulatory protein. Second, even if the face of the *cis*-regulatory sequence is exposed on the outside of the nucleosome, many transcription regulators subtly alter the conformation of the DNA when they bind, and these changes are generally opposed by the tight wrapping of the DNA around the histone core. For example, many transcription regulators induce a bend or kink in the DNA when they bind.

We saw in Chapter 4 that nucleosome remodeling can alter the structure of the nucleosome, allowing transcription regulators access to the DNA. Even without remodeling, however, transcription regulators can still gain limited access to DNA in a nucleosome. The DNA at the end of a nucleosome "breathes," transiently exposing the DNA and allowing regulators to bind. This breathing occurs at a much lower rate in the middle of the nucleosome; therefore, the positions where the DNA exits the nucleosome are much easier to occupy than those in the middle of the nucleosome (**Figure 7–12**).

These properties of the nucleosome promote cooperative DNA binding by transcription regulators. If a transcription regulator seizes a "window of opportunity" provided by nucleosome breathing, it can enter the nucleosome by binding to the exposed DNA and prevent the DNA from tightly rewrapping around the nucleosome core. When this happens, the affinity of a second transcription regulator for a nearby *cis*-regulatory sequence can be increased simply by this loosening of the DNA from the histone core. If the two transcription regulators also interact with each other (as described earlier), the cooperative effect can be even greater. In some cases, the combined action of the regulatory proteins can eventually displace the histone core of the nucleosome altogether. Many transcription regulators, when their affinities for DNA and their concentrations are sufficiently high, can take advantage of nucleosome breathing and thereby

"invade" nucleosomes. Moreover, as we saw in Chapter 5, passing replication forks, which transiently displace histones, offer additional windows of opportunity for transcription regulators to bind to DNA.

Although nucleosomes generally inhibit the DNA binding of transcription regulators, some regulators—if their *cis*-regulatory sequences are exposed on the nucleosome surface—can bind with nearly the same affinity as they do on naked DNA, occupying their binding sites while the DNA is still tightly wrapped around the histone core (**Figure 7–13**). Transcription regulators with this property are sometimes called *pioneer factors*, because they are often the first proteins to bind DNA when a previously silent gene becomes transcriptionally active. Although their binding typically destabilizes the nucleosome, pioneer factors probably exert their major effects by attracting additional proteins that alter chromatin structure, such as nucleosome remodeling complexes. If one transcription regulator binds its *cis*-regulatory sequence on a nucleosome and attracts a chromatin remodeling complex, the localized action of the remodeling complex can allow a second transcription regulator to efficiently bind nearby.

Our discussion has emphasized how transcription regulators can work together in pairs. But in reality, larger numbers often cooperate by repeated use of the same principles. It is the cooperative formation of clusters of transcription regulators on DNA that probably explains why many key regulatory sequences in eukaryotic genomes are found to be "nucleosome free."

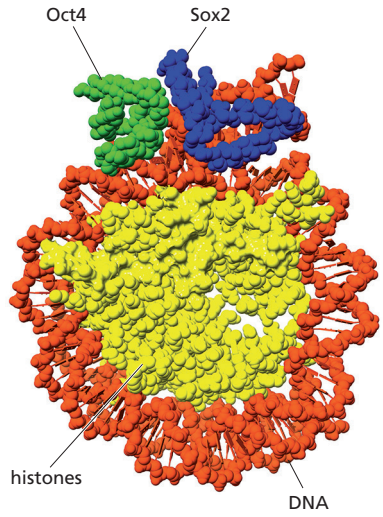


Figure 7–13 Two cooperating transcription regulators, Oct4 *(green)* and Sox2 *(blue)*, bound to a nucleosome. These two transcription regulators work together and play key roles in maintaining embryonic stem cells (see Figures 7–36 and 7–37). Only the DNA-binding portion of each regulator protein is shown. (Courtesy of Nicolas H. Thomä and Alicia K. Michael. PDB code: 6T90.)

DNA-Binding by Transcription Regulators Is Dynamic

Thus far, we have treated transcription regulators as static—we have considered them as either bound to DNA or free in solution. But in reality, the situation is highly dynamic, with transcription regulator molecules in constant motion, rapidly binding and dissociating from DNA. In most cases a given transcription regulator molecule stays on its *cis*-regulatory sequence for only a short time, but it is rapidly replaced by other molecules of the same regulator. Thus, when we consider a *cis*-regulatory sequence being fully bound by its matching transcription regulator, this state is an average, over time, of many individual association and dissociation events.

By attaching a transcription regulator to a bright fluorescent tag, it is possible to follow single regulator molecules in live cells, as they diffuse randomly within the nucleus, bind to their *cis*-regulatory sequences, and then dissociate from them. In these *single-molecule tracking experiments*, different states for the regulator can be distinguished on the basis of the tagged protein's mobility over short time periods. A high-mobility regulator state is observed for the free protein diffusing in the nucleoplasm. At the other extreme, a very low-mobility state is attributed to the regulator bound to DNA, inasmuch as its restrained motions are similar to that of a histone molecule that has been labeled in the same way (**Figure 7–14**).

Whereas a histone remains stably bound in a nucleosome, transcription regulators remain in a low-mobility, DNA-bound state only transiently. Individual regulator molecules are observed to leave their DNA-bound state at a wide variety of rates—some molecules persist for only a fraction of a second, while others remain for minutes. How can we explain these differences? We saw earlier in the chapter (see Figure 7-10) that each transcription regulator has a preferred *cis*-regulatory sequence, but that it can also bind—albeit with lower affinity—to related

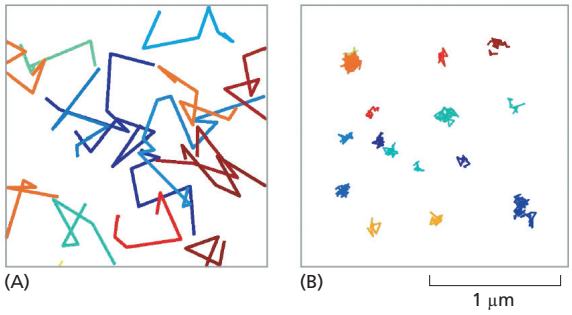


Figure 7–14 Tracking single molecules of a transcription regulator in the nucleus of a living cell. By conjugating a fluorescent tag to the glucocorticoid receptor (see pp. 573–575), the behavior of this transcription regulator can be followed in living cells, using a microscope that follows its fluorescence. Computational methods then allow the observed behavior of such molecules to be classified into sets of distinct mobility groups, two of which are shown here. (A) Sample tracks observed for individual molecules of the glucocorticoid receptor in the freely diffusing mobility group. The positions illustrated were determined for a total of 10 seconds. (B) Tracks of individual molecules bound to DNA, with positions determined over a 120-second interval. (A and B, courtesy of D.A. Garcia and G.L. Hager.)

DNA sequences. Because the forward rates at which regulatory proteins "find" their *cis*-regulatory sequences are largely independent of the exact nucleotide sequence of that DNA, affinity differences are reflected in how long a protein remains bound on the DNA—the higher the affinity, the longer the protein stays bound.

Any protein, such as a transcription regulator, that binds tightly to a specific set of DNA sequences will also bind, albeit much more weakly, to any DNA sequence. This weak binding is useful because it allows a regulator to search for its target by "scanning" the DNA in the vicinity of the initial chromosomal site that it binds. Most such regulators will fail to find a matching *cis*-regulatory DNA sequence, and it is these that are thought to dissociate within seconds. The minority that persist for minutes are likely to have engaged with a matching *cis*-regulatory sequence. But because even these regulators do not remain on DNA for long periods, they need to be constantly replaced by another such molecule. Thus, as always, the static pictures in this textbook fail to do justice to the frantic state of motion that exists inside a cell (see pp. 65–66).

Summary

Transcription regulators recognize short stretches of double-helical DNA of defined sequence called cis-regulatory sequences, and they thereby determine which of the thousands of genes in a cell will be transcribed. Transcription regulators determine many cell properties, and their importance is reflected by the fact that approximately 10% of the protein-coding genes in most organisms produce them. Although each transcription regulator has unique features, most bind to DNA as homodimers or heterodimers and recognize DNA through one of a small number of structural motifs. Transcription regulators typically work in groups and bind to DNA cooperatively, a feature that is explained by several underlying mechanisms, some of which exploit the packaging of DNA in nucleosomes.

TRANSCRIPTION REGULATORS SWITCH GENES ON AND OFF

Having seen how transcription regulators bind to *cis*-regulatory sequences embedded in the genome, we can now discuss how, once bound, these proteins influence the transcription of genes. The situation in bacteria is simpler than in eukaryotes (for one thing, chromatin structure is not an issue), and we therefore discuss bacterial mechanisms before proceeding to the more complex situation in eukaryotes.

The Tryptophan Repressor Switches Genes Off

The genome of the bacterium *Escherichia coli* consists of a single, circular DNA molecule of about 4.6×10^6 nucleotide pairs that encodes approximately 4300 proteins. Only a fraction of these proteins are made at any one time. For example, all bacteria regulate the expression of many of their genes according to the food sources that are available in the environment. Thus 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 7–15**). Such operons are common in bacteria but rare in

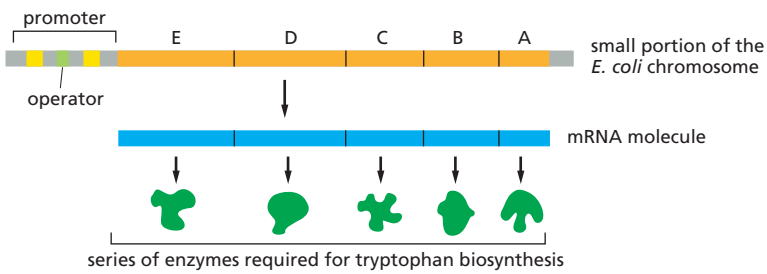


Figure 7–15 A cluster of bacterial genes can be transcribed from a single promoter. Each of these five genes encodes a different enzyme, and all of these enzymes are needed to synthesize the amino acid tryptophan from simpler molecules. The genes are transcribed as a single mRNA molecule, a feature that 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 *cis*-regulatory sequence called the *operator (green)*, situated within the promoter. (In this and subsequent figures, the *yellow blocks* in the promoter represent DNA sequences that bind RNA polymerase; see Figure 6–12).

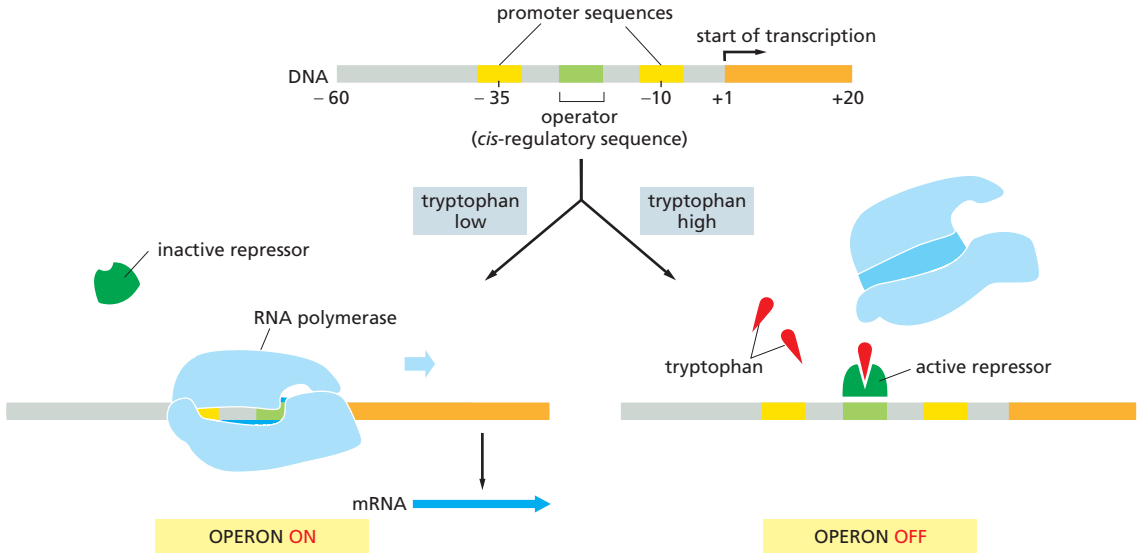


Figure 7–16 Genes can be switched off by repressor proteins. If the concentration of tryptophan inside a bacterium is low *(left)*, RNA polymerase *(blue)* binds to the promoter and transcribes the five genes of the tryptophan operon. However, if the concentration of tryptophan is high *(right)*, the tryptophan 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, this transcription regulator falls off the DNA, allowing the polymerase to again transcribe the operon. Although not shown in the figure, the tryptophan repressor exists as a stable protein dimer.

eukaryotes, where genes are typically transcribed and regulated individually (see Figures 6–75 and 6–90).

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 from much simpler molecules. 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 exactly how this repression of the tryptophan operon comes about. Within the operon's promoter is a *cis*-regulatory sequence that is recognized by a transcription regulator. When this regulator binds to this sequence, it blocks access of RNA polymerase to the promoter, thereby preventing transcription of the operon (and thus production of the tryptophan-producing enzymes). The transcription regulator is known as the *tryptophan repressor*, and its *cis*-regulatory sequence is called the *tryptophan operator*. These components are controlled in a simple way: the repressor can bind to DNA only if it has also bound several molecules of tryptophan (**Figure 7–16**).

The tryptophan repressor is an allosteric protein, and the binding of tryptophan causes a subtle change in its three-dimensional structure so that the protein can bind tightly to the operator sequence. Whenever the concentration of free tryptophan in the bacterium drops, tryptophan dissociates from the repressor, 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 continually 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 or fall in tryptophan concentration.

Repressors Turn Genes Off and Activators Turn Them On

The tryptophan repressor, as its name suggests, is a *transcription 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 *transcription 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 to nearby

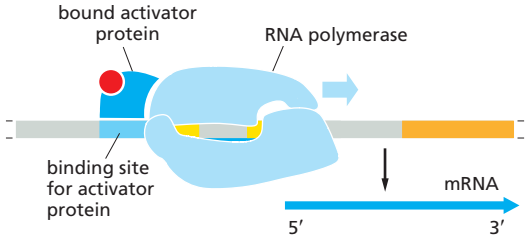


Figure 7–17 Genes can be switched on by activator proteins. An activator protein binds to its *cis*-regulatory sequence on the DNA and interacts with the RNA polymerase to help it initiate transcription. Without the activator, the promoter fails to initiate transcription efficiently. In bacteria, the binding of the activator to DNA is often controlled by the interaction of a metabolite or other small molecule *(red circle)* with the activator protein.

cis-regulatory sequences and contact the RNA polymerase to help it initiate transcription (**Figure 7–17**).

DNA-bound activator proteins can increase the rate of transcription initiation as much as 1000-fold, a value consistent with a relatively weak and nonspecific interaction between the transcription regulator and RNA polymerase. For example, a 1000-fold change in the affinity of RNA polymerase for its promoter corresponds to a change in ΔG of ~18 kJ/mole, which could be accounted for by just a few weak, noncovalent bonds. Thus, many activator proteins work simply by providing a few favorable interactions that help to attract RNA polymerase to the promoter. To provide this assistance, however, the activator protein must be bound to its *cis*-regulatory sequence, and this sequence must be positioned precisely so that these favorable interactions can occur with an RNA polymerase molecule at its promoter.

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. 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. CAP then drives the production of enzymes that allow the bacterium to digest other sugars.

Both an Activator and a Repressor Control the *Lac* Operon

The activity of a single bacterial promoter is often controlled by several different transcription regulators. The *Lac* operon in *E. coli*, for example, is controlled by both the Lac repressor and the CAP activator just discussed. The *Lac* operon encodes proteins required to import and digest the disaccharide lactose, a key nutrient in milk. In the absence of glucose (the cell's favorite energy source), 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 7–18**). 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.

All transcription regulators, whether they are repressors or activators, must be bound to DNA to exert their effects. In this way, each regulatory protein acts selectively, controlling only those genes that bear a *cis*-regulatory sequence recognized by it. The logic of the *Lac* operon first attracted the attention of biologists more than 60 years ago. The way it works was uncovered by a combination of genetics and biochemistry, providing some of the first insights into how transcription is controlled in any organism.

DNA Looping Can Occur During Bacterial Gene Regulation

We have seen that transcription activators help RNA polymerase to initiate transcription and repressors hinder it. Otherwise the two types of transcription regulators are similar: both the tryptophan repressor and the CAP activator

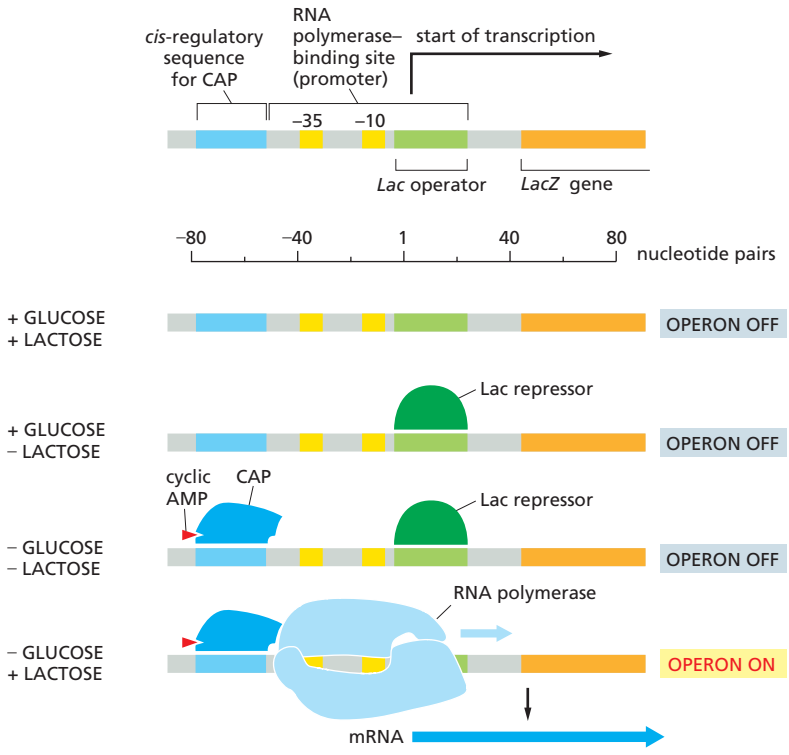


Figure 7–18 How the *Lac* operon is controlled by two transcription regulators, causing it to be expressed only when needed. *LacZ*, the first gene of the operon, encodes the enzyme β-galactosidase, which breaks down lactose to galactose and glucose. When lactose is absent, the Lac repressor binds to a *cis*-regulatory sequence, called the *Lac* operator, and shuts off expression of the operon (Movie 7.4). 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). This removes a block to expression of the *Lac* operon, but the operon can turn on only if the sugar glucose, a preferred carbon source, is absent. This is because cyclic AMP *(red triangle)* is produced by the cell in the absence of glucose, and this small molecule is required for CAP to bind to DNA and activate transcription.

protein must bind a small molecule to occupy their *cis*-regulatory sequences, and both recognize these DNA sequences using the same structural motif (the helix–turn–helix shown in Panel 7–1). Some proteins (for example, the CAP protein) can act either as a repressor or an activator, depending on the exact placement of a binding site relative to the promoter: if this site overlaps the promoter, CAP binding can prevent the assembly of RNA polymerase at the promoter, thus serving as a repressor.

Most bacteria have small, compact genomes, and the *cis*-regulatory sequences that control the transcription of a gene are typically located very near to the start point of transcription. But there are some exceptions to this generalization—*cis*-regulatory sequences can be located hundreds and even thousands of nucleotide pairs from the bacterial genes they control. In these cases, the intervening DNA loops out, allowing a transcription regulator bound at a distant site along the DNA to contact RNA polymerase (Figure 7–19). Here, the DNA is serving as a tether, enormously increasing the probability that the regulator will collide with a

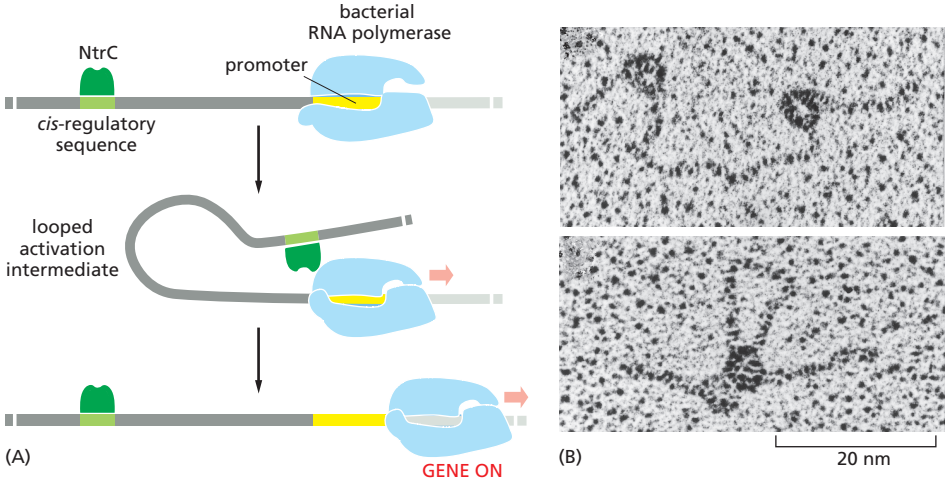


Figure 7–19 Transcriptional activation by DNA looping in bacteria. (A) The NtrC protein is a bacterial transcription regulator that activates transcription by directly contacting RNA polymerase. (B) The interaction of NtrC and RNA polymerase, with the intervening DNA looped out, can be seen in the electron microscope. (B, courtesy of Harrison Echols and Sydney Kustu.)

promoter-bound polymerase, compared with the situation where the regulator is free in solution. We will see shortly that, although the exception in bacteria, DNA looping is thought to occur in the regulation of nearly every eukaryotic gene. It has been proposed that the compact, simple genetic switches found in bacteria evolved in response to a severe competition for growth that put strong selective pressure on bacteria to maintain small genome sizes. In contrast, there appears to have been little selective pressure to "streamline" the genomes of multicellular organisms.

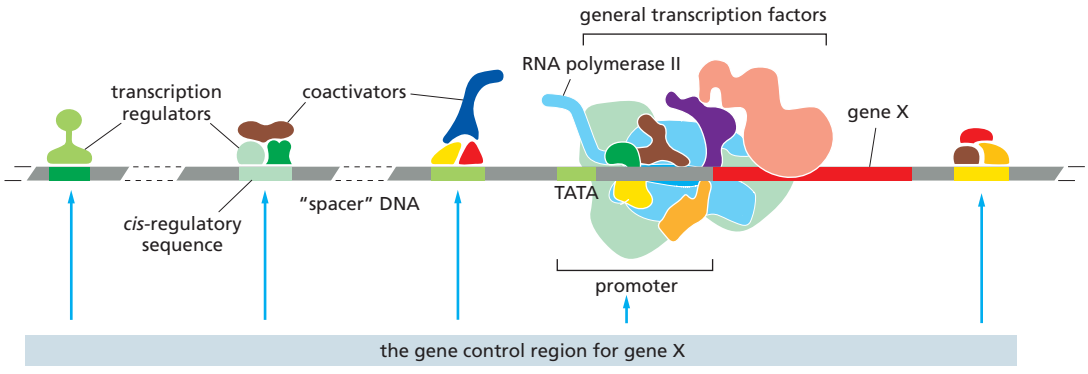
Complex Switches Control Gene Transcription in Eukaryotes

When compared to the situation in bacteria, transcription regulation in eukaryotes involves many more proteins and much longer stretches of DNA—and it often seems bewilderingly complex. Yet many of the same principles apply. As in bacteria, the time and place that each gene is to be transcribed are specified by its *cis*-regulatory sequences, which are "read" by the transcription regulators that bind to them. Once bound to DNA, positive transcription regulators (activators) help RNA polymerase to begin transcribing genes, and negative regulators (repressors) block this from happening. But in bacteria, most of the interactions between DNA-bound transcription regulators and RNA polymerases (whether they activate or repress transcription) are direct; that is, they contact each other. In contrast, these interactions are almost always indirect in eukaryotes: many intermediate proteins, including the histones and a large protein complex known as *Mediator*, act between DNA-bound transcription regulators and RNA polymerase. Moreover, in multicellular organisms, it is common for dozens of transcription regulators to control a single gene and for *cis*-regulatory sequences to be spread over tens of thousands of nucleotide pairs. DNA looping allows the DNA-bound regulatory proteins to interact with each other and ultimately to control RNA polymerase at the promoter. Many of the protein–protein interactions involved are of low affinity and are thought to trigger the formation of biomolecular condensates, which can facilitate reactions requiring such a large number of different components (see pp. 171–173). Finally, because nearly all of the DNA in eukaryotic organisms is organized in nucleosomes and higher-order chromatin structures, transcription initiation in eukaryotes must overcome this inherent block. In the next sections, we discuss each of these features of transcription initiation in eukaryotes, emphasizing how they provide extra levels of control not found in bacteria.

A Eukaryotic Gene Control Region Includes Many *cis*-Regulatory Sequences

In eukaryotes, RNA polymerase II transcribes all the protein-coding genes and many noncoding RNA genes. This polymerase requires five general transcription factors (with 27 subunits *in toto*; see Table 6–3, p. 333, and Figure 6–15), in contrast to bacterial RNA polymerase, which needs only a single general transcription factor (the σ subunit). As we saw in Chapter 6, the stepwise assembly of the general transcription factors at a eukaryotic promoter provides, in principle, multiple steps at which the cell can speed up or slow down the rate of transcription initiation in response to transcription regulators.

Because the many *cis*-regulatory sequences that control the expression of a typical gene are often spread over long stretches of DNA, we use the term **gene control region** to describe the whole expanse of DNA involved in regulating and initiating transcription of a eukaryotic gene. This includes the **promoter**, where the general transcription factors and the polymerase assemble, plus all of the *cis*-regulatory sequences to which transcription regulators bind to control the rate of the gene activation processes at the promoter (**Figure 7–20**). In animals and plants, it is not unusual to find the regulatory sequences of a gene dotted over stretches of DNA as large as 100,000 nucleotide pairs. For now, we can regard much of this DNA as "spacer" sequences that transcription regulators do not



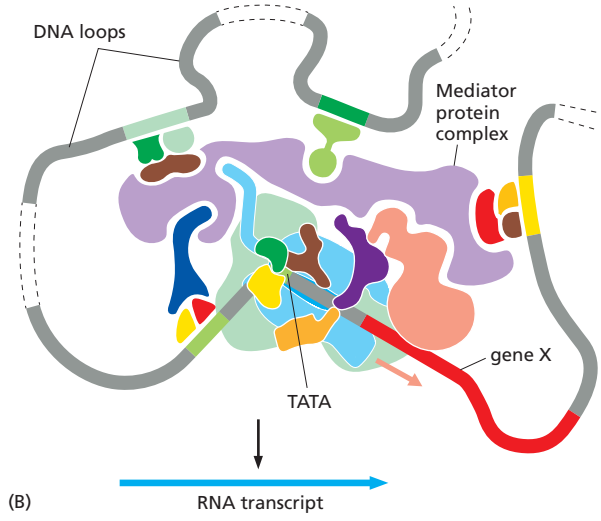


Figure 7–20 Transcription is controlled by gene control regions. (A) The gene control region of a typical eukaryotic gene depicted with the DNA arranged in a straight line. The *promoter* is the DNA sequence where the general transcription factors and the polymerase assemble (see Figure 6–15). The *cis-regulatory sequences* are binding sites for transcription regulators, whose presence on the DNA ultimately affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or entirely downstream of the gene. The broken stretches of DNA signify that the length of DNA between the *cis*-regulatory sequences 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 general transcription factor TFIID (see Figures 6–15 and 6–17). (B) DNA looping allows transcription regulators bound at many positions to "communicate" with the proteins that assemble at the promoter. As shown in this schematic diagram, many transcription regulators act through Mediator (described in Chapter 6), while some interact with the general transcription factors and RNA polymerase directly. Transcription regulators also act by recruiting proteins that alter the chromatin structure of the promoter (not shown here but discussed later in the chapter). Whereas Mediator and the general transcription factors are the same for all RNA polymerase II–transcribed genes, the transcription regulators and the locations of their binding sites relative to the promoter differ for each gene. At especially complex gene control regions, the many proteins that assemble can, by virtue of large numbers of low-specificity interactions, undergo phase transitions that further coalesce the protein and DNA components needed to initiate transcription—presumably accelerating the process.

directly recognize. We will see later in this chapter that some of this DNA is transcribed (but not translated) into long noncoding RNAs (lncRNAs), which have diverse functions in the cell.

In this chapter, we shall loosely use the term **gene** to refer to a segment of DNA that is transcribed into a functional RNA molecule, one that either codes for a protein or has a different role in the cell (see Table 6–1, p. 327). However, the classical view of a gene includes the gene control region as well, because mutations in it can produce an altered phenotype. Alternative RNA splicing further complicates the definition of a gene—a point we shall return to later.

In contrast to the small number of general transcription factors, which are abundant proteins that assemble on the promoters of all genes transcribed by RNA polymerase II, there are thousands of different transcription regulators devoted to turning individual genes on and off. As we have seen, each eukaryotic gene is usually transcribed individually. Not surprisingly, the regulation of each eukaryotic gene is different in detail from that of every other gene, and it is difficult to formulate simple rules for gene regulation that apply in every case. We can, however, make some generalizations about how transcription regulators, once bound to gene control regions on DNA, set in motion the series of events that lead to gene activation or repression.

Eukaryotic Transcription Regulators Work in Groups

In bacteria, we saw that proteins such as the tryptophan repressor, the Lac repressor, and the CAP protein bind to DNA on their own and directly affect RNA polymerase at the promoter. Eukaryotic transcription regulators, in contrast, usually assemble together in groups at their *cis*-regulatory sequences. Often two

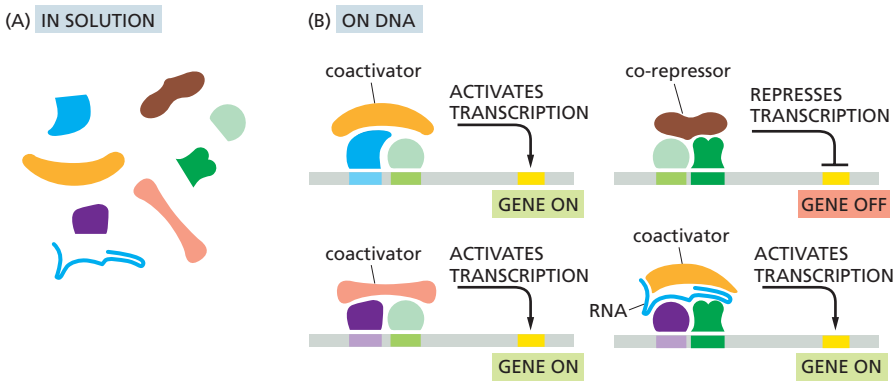


Figure 7–21 Eukaryotic transcription regulators assemble into complexes on DNA. (A) Seven different proteins and an RNA molecule are shown. The nature and function of the complex they form depend on the specific *cis*-regulatory sequences that seed their assembly. (B) Some assembled complexes activate gene transcription, while another represses transcription. Note that the *light green* and *dark green* proteins are shared by both activating and repressing complexes. Proteins that do not themselves bind DNA but assemble on other DNA-bound transcription regulators are termed coactivators or co-repressors. In some cases *(lower right)*, long, noncoding RNA molecules are also found in these assemblies. As described later in this chapter, these RNAs often act as scaffolds to hold groups of proteins together.

or more regulators bind cooperatively, as discussed earlier in the chapter (see Figure 7–10). In some especially complex gene control regions, tens and even hundreds of such proteins may coassemble on DNA. In addition, a broad class of multisubunit proteins termed *coactivators* and *co-repressors* join with them. Typically, these coactivators and co-repressors do not recognize specific DNA sequences themselves; they are brought to those sequences by specific interactions with the DNA-bound transcription regulators. As their names imply, coactivators are typically involved in activating transcription and co-repressors in repressing it. In the following sections, we will see that coactivators and co-repressors can act in a variety of different ways to influence transcription once they have been localized on the genome by transcription regulators.

As shown in **Figure 7–21**, an individual transcription regulator can often participate in more than one type of regulatory complex. A protein might function, for example, in one case as part of a complex that activates transcription and in another case as part of a complex that represses transcription. Thus, individual eukaryotic transcription regulators function as regulatory parts that are used to build complexes whose function depends on the final assembly of all of the individual components. Each eukaryotic gene is therefore regulated by a "committee" of proteins, all of which must be present to express the gene at its proper level. Often the protein–protein interactions between transcription regulators and between regulators and coactivators are too weak for them to assemble in solution; however, the appropriate combination of *cis*-regulatory sequences can "crystallize" the assembly of these complexes on DNA. In very large and complex gene control regions, this assembly may be accompanied by a phase transition to form a biomolecular condensate, whereby all the components are held together even more efficiently by keeping them in rough proximity even when individual proteins disassociate from DNA.

Activator Proteins Promote the Assembly of RNA Polymerase at the Start Point of Transcription

The *cis*-regulatory sequences to which eukaryotic transcription activator proteins bind were originally called *enhancers* because their presence "enhanced" the rate of transcription initiation. It initially came as a surprise when it was discovered that these sequences could be found tens of thousands of nucleotide pairs away from the promoter; as we have seen, DNA looping, which was not widely appreciated at the time, can now explain this initially puzzling observation.

Once bound to DNA, how do assemblies of activator proteins increase the rate of transcription initiation? At most genes, several mechanisms work in concert. Their ultimate function is to attract and position RNA polymerase II at the promoter and to release it so that transcription can begin.

Some activator proteins bind directly to one or more of the general transcription factors, accelerating their assembly on a promoter that has been brought in proximity—through DNA looping—to that activator. Most transcription activators, however, attract coactivators that then perform the biochemical tasks needed to initiate transcription. As we have seen, one of the most prevalent coactivators is

the large *Mediator* protein complex, composed of more than 30 subunits. About the same size as RNA polymerase itself, Mediator serves as a bridge between DNA-bound transcription activators, RNA polymerase, and the general transcription factors, facilitating their assembly at the promoter (see Figure 7–20).

Eukaryotic Transcription Activators Direct the Modification of Local Chromatin Structure

The eukaryotic general transcription factors and RNA polymerase are unable, on their own, to assemble on a promoter that is packaged in nucleosomes. Thus, in addition to directing the assembly of the transcription machinery at the promoter, eukaryotic transcription activators—once bound to their *cis*-regulatory sequences—promote transcription by triggering changes to the chromatin structure of the promoters, rendering the underlying DNA more accessible. The enzymes that alter chromatin structure are usually carried as subunits of coactivators, which are typically multiprotein complexes, with different subunits carrying out different functions. For example, such a coactivator might carry one subunit that associates with specific DNA-bound transcription regulators, another that associates with one of the general transcription factors, and several more that alter chromatin structure in different ways.

The most important ways of locally altering chromatin are through covalent histone modifications, nucleosome remodeling, nucleosome removal, and histone replacement (all discussed in Chapter 4). Eukaryotic transcription activators use all four of these mechanisms: thus they attract coactivators that include histone modification enzymes, ATP-dependent chromatin remodeling complexes, and histone chaperones. These proteins often act cooperatively to alter the chromatin structure of promoters, providing greater access to the DNA (**Figure 7–22**).

Often a series of individual events, ultimately directed by transcription regulators, must occur before RNA polymerase can be assembled onto a promoter,

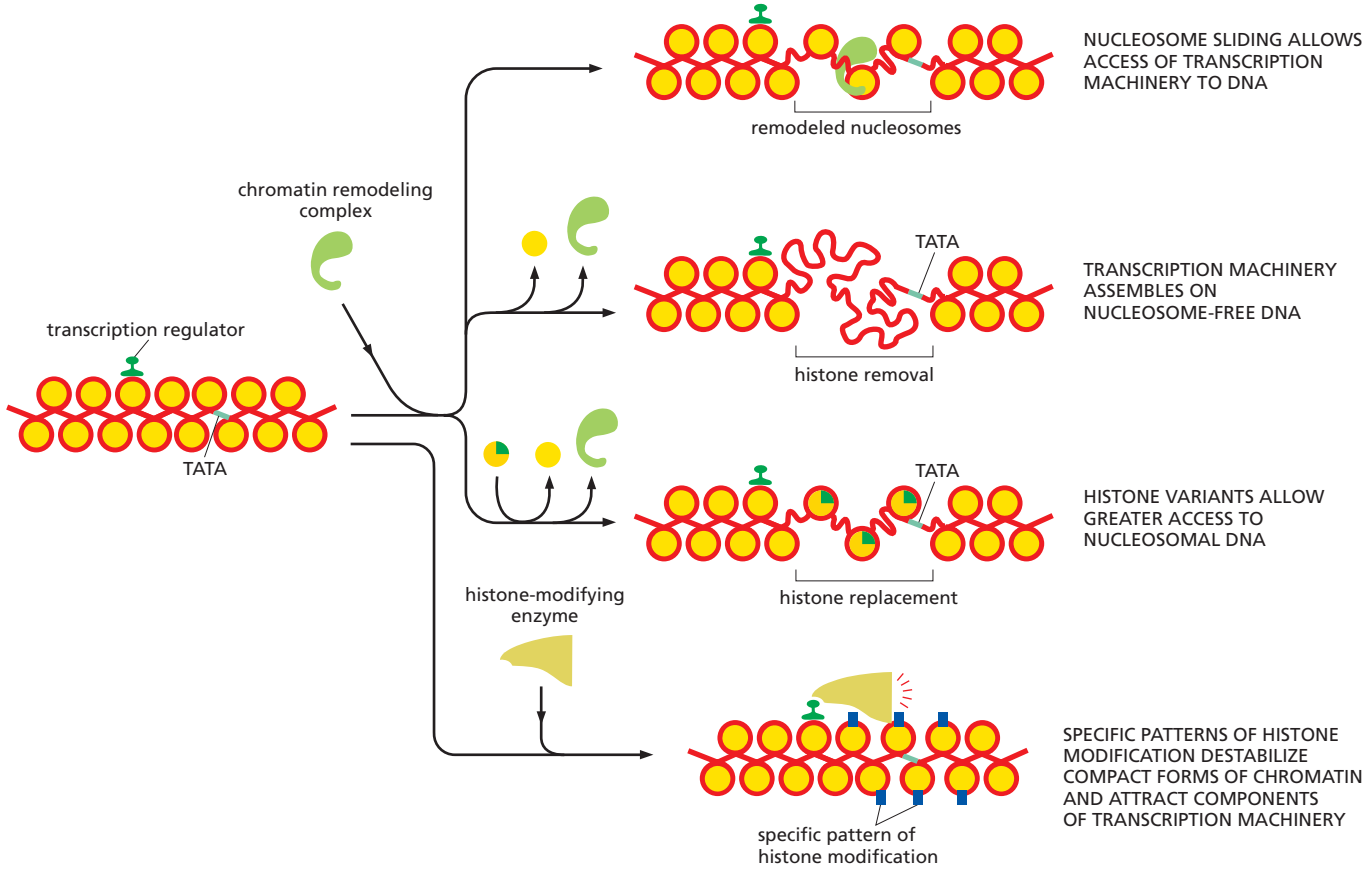


Figure 7–22 Eukaryotic transcription activator proteins direct local alterations in chromatin structure. Nucleosome remodeling, nucleosome removal, histone replacement, and certain types of histone modifications favor transcription initiation (see Table 4–2, p. 210). As illustrated, some of these changes are driven by different types of ATP-dependent chromatin remodeling complexes (see Figures 4–26 and 4–27); most also involve histone chaperones (not shown). Such alterations increase the accessibility of DNA and facilitate the binding of RNA polymerase and the general transcription factors.

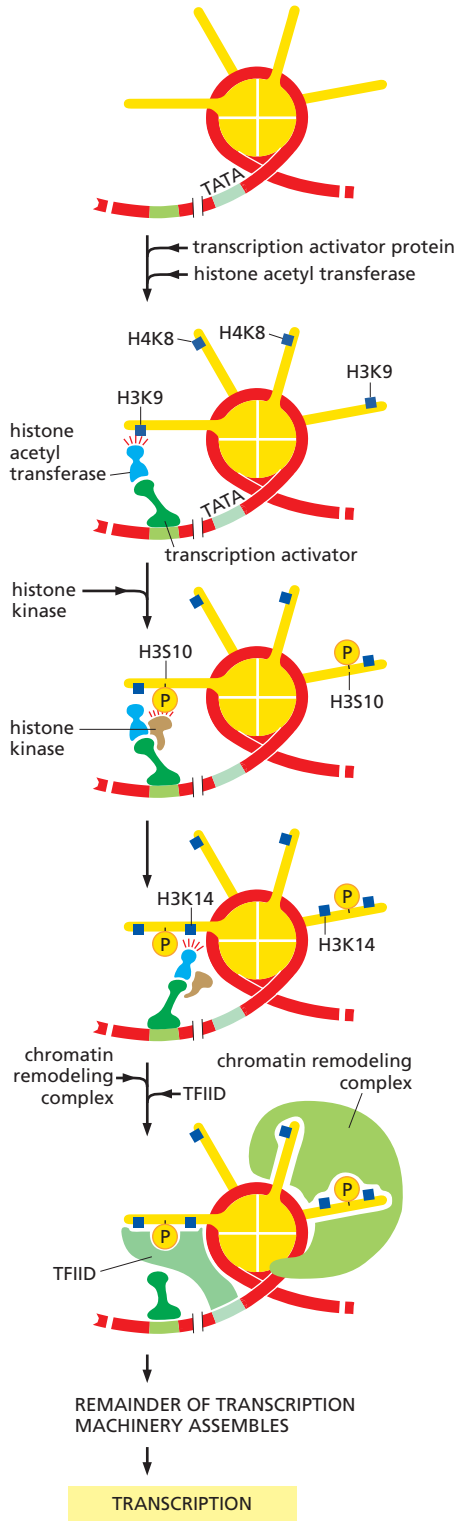


Figure 7–23 Successive histone modifications during transcription initiation. In this example, taken from the human interferon-β gene promoter, a transcription activator binds to DNA packaged into chromatin and attracts a histone acetyl transferase that acetylates lysine 9 of histone H3 and lysine 8 of histone H4 (see Figure 4–35). Next, a histone kinase, part of a different coactivator attracted by the same transcription activator, phosphorylates serine 10 of histone H3, but it can only do so after lysine 9 has been acetylated. This serine modification signals the original histone acetyl transferase to acetylate position K14 of histone H3. Next, the general transcription factor TFIID and a chromatin remodeling complex come into play to promote the subsequent steps of transcription initiation. TFIID and the remodeling complex both recognize acetylated histone tails through a *bromodomain*, a protein domain specialized to read this particular mark on histones; a bromodomain is carried in a subunit of each protein complex. Binding of TFIID causes a sharp bend in the DNA (not shown but see Figure 6–17), which facilitates sliding of the nucleosome to a new position, thereby freeing the start site of transcription for binding by RNA polymerase II.

The histone acetyl transferase, the histone kinase, and the chromatin remodeling complex are all subunits of coactivators. The order of events shown applies to a specific promoter; at other genes, the steps may occur in a different order or individual steps may be omitted altogether. (Adapted from T. Agalioti et al., *Cell* 111:381–392, 2002.)

with details that depend on the gene being regulated. In the example illustrated in **Figure 7–23**, a series of specific histone tail modifications is triggered by a transcription activator; these modifications then attract additional proteins to the promoter, including both a chromatin remodeling complex and a general transcription factor. Those proteins can in turn recruit additional proteins to the promoter, while also destabilizing adjacent nucleosomes.

Because the local chromatin changes directed by one transcription regulator often allow the binding of additional proteins—both directly (see Figure 7–12) and indirectly as just described—a cascade of events typically takes place on the control regions of eukaryotic genes to regulate their transcription.

As RNA polymerase II transcribes through a gene a different type of chromatin modification occurs. The histones just ahead of the polymerase are acetylated by enzymes carried by the polymerase, removed by histone chaperones, and deposited behind the moving polymerase. These histones are then rapidly deacetylated and methylated, also by complexes that are carried by the polymerase, leaving behind nucleosomes that are especially resistant to transcription. This remarkable process seems to prevent spurious transcription reinitiation behind a moving polymerase, which, in essence, must clear a path through chromatin as it transcribes. Later in this chapter, when we discuss *RNA interference*, the potential dangers to the cell of such inappropriate transcription will become especially obvious.

Some Transcription Activators Work by Releasing Paused RNA Polymerase

Thus far, we have emphasized how transcription regulators—once bound to DNA—can assemble multiple components and stimulate transcription initiation. But for some genes, a key regulatory step occurs after this point (**Figure 7–24**). In the most common of these cases, the RNA polymerase halts after transcribing about 50 nucleotides of RNA, and further elongation requires a new transcription activator to bind to the gene's control region (see Figure 7–24C).

The release of a paused RNA polymerase can occur in several ways. In some cases, the new activator brings in a chromatin remodeling complex that removes a nucleosome block to the elongating RNA polymerase. In other cases, the activator communicates with RNA polymerase (typically through a coactivator), signaling it to forge ahead. Finally, as we saw in Chapter 6, RNA polymerase requires *elongation factors* to effectively transcribe through chromatin (Figure 6–19). In some cases, the key step in gene activation is the delayed loading of these factors onto RNA polymerase, directed by DNA-bound transcription activators. Once loaded, these factors allow the polymerase to move through blocks imposed by chromatin structure to begin transcribing the gene effectively.

Paused polymerases are common in humans, where a significant fraction of genes that are not being transcribed have a paused polymerase located just

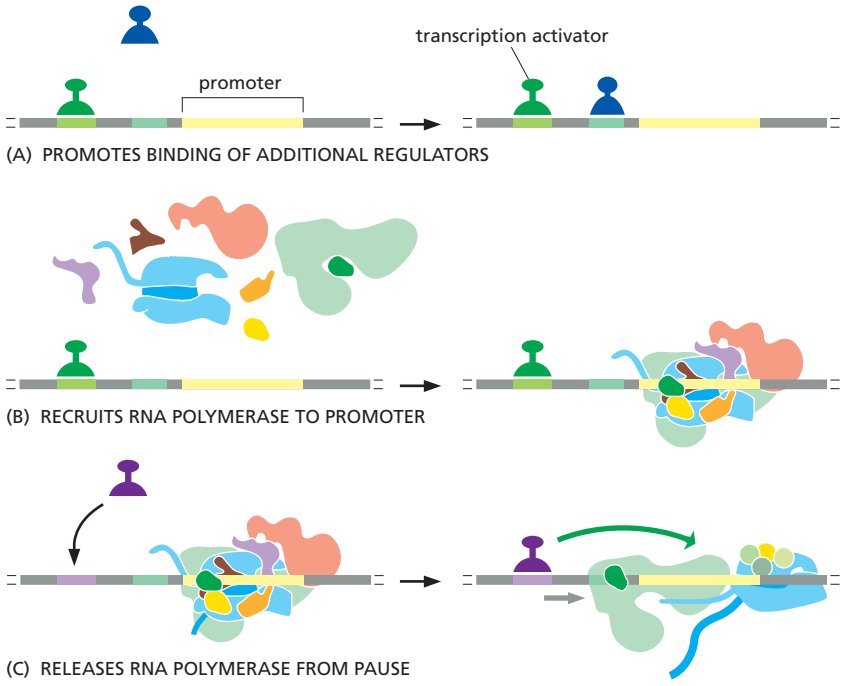


Figure 7–24 Different transcription regulators can act at different steps. (A) As described earlier in this chapter (see Figure 7–12), a DNA-bound transcription activator can promote DNA binding by additional transcription regulators. (B) As shown in more detail in Figures 7–20 and 7–22, most transcription activators direct assembly of RNA polymerase at promoters; this can occur by a variety of mechanisms. (C) Some other transcription activators, once bound to DNA, release RNA polymerase molecules that are paused after transcribing about 50 nucleotides of RNA. For simplicity, many of the additional proteins required for transcription initiation are not shown.

downstream from the promoter. Having RNA polymerase already poised on a promoter in the beginning stages of transcription bypasses the step of assembling many components at the promoter, which is often slow. This mechanism is therefore thought to allow cells to begin transcribing a gene in rapid response to an extracellular signal.

Transcription Activators Work Synergistically

We have seen that complexes of transcription activators and coactivators assemble cooperatively on DNA. We have also seen that these assemblies can promote different steps in transcription initiation. In general, where several factors work together to enhance a reaction rate, the joint effect is not merely the sum of the enhancements that each factor alone contributes, but the product of them. If, for example, factor A lowers the free-energy barrier for a reaction by a certain amount and thereby speeds up the reaction 100-fold, and factor B, by acting on that reaction, does likewise, then A and B acting in parallel can lower the energy barrier by a double amount and speed up the reaction 10,000-fold. Even if A and B work simply by attracting the same protein, the affinity of that protein for the reaction site increases multiplicatively. Thus, transcription activators often exhibit *transcriptional synergy*, where several DNA-bound activator proteins working together produce a transcription rate that is much higher than the sum of their transcription rates working alone (**Figure 7–25**).

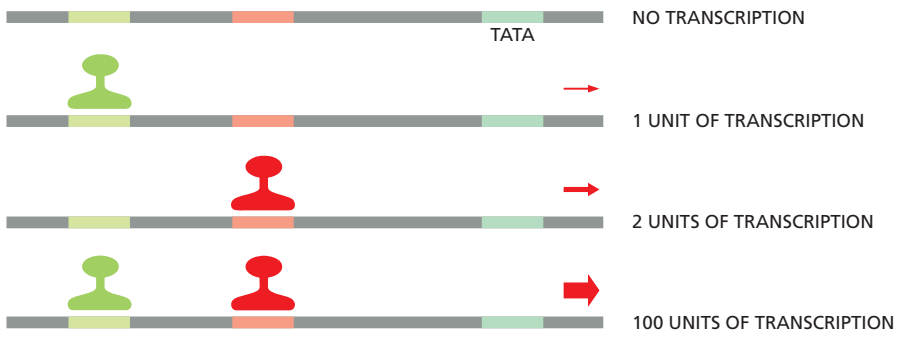


Figure 7–25 Transcriptional synergy. This experiment compares the rate of transcription produced by three experimentally constructed regulatory regions in a eukaryotic cell and reveals transcriptional synergy, a greater than additive effect of multiple activators working together. Such transcriptional synergy is not only observed between different transcription activators from the same organism; it is also seen between activator proteins from different eukaryotic species when they are experimentally introduced into the same cell. This last observation reflects the high degree of conservation of the machinery responsible for eukaryotic transcription initiation.

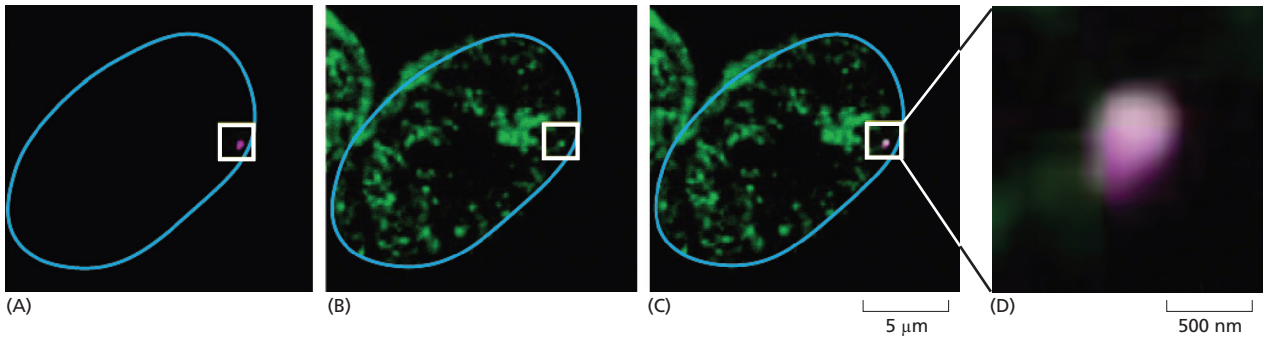


Figure 7–26 Condensate formation at the transcription control region of the *Nanog* gene in a mouse embryonic stem cell. The cell was fixed, and in (A) the *Nanog* gene was identified by hybridizing a complementary nucleotide sequence attached to a *red* fluorophore, according to a procedure known as FISH (see Figure 8–32). Nanog is a key transcription regulator in embryonic stem cells (see Figure 7–10), and its own regulatory region is one of the most complex in the mouse genome. The nucleus is indicated by the *blue oval*. (B) A subunit of Mediator fused to a green fluorescent protein (see Figure 9–16) was visualized. (C) The two preceding images have been merged, and in (D) the portion of the image in the *white square* is magnified tenfold. The size and diffuse nature of the "blob" suggest a condensate containing a large number of proteins. Note that additional condensates of Mediator are visible throughout the nucleus and may represent condensates at other enhancers. These condensates are much smaller than those of the nuclear "organelles," such as the nucleolus, discussed in Chapter 6. (From B.R. Sabari et al., *Science* 361:eaar3958, 2018. With permission from AAAS.)

As a result, the rate of transcription of a gene ultimately depends on the spectrum of regulatory proteins that are bound upstream and downstream of its transcription start site, along with the coactivator proteins they bring to the DNA.

Condensate Formation Likely Increases the Efficiency of Transcription Initiation

We have discussed in broad, conceptual terms the many different types of proteins that must assemble for transcription of a typical gene to begin. For especially complex gene control regions, such as those of key human genes that orchestrate development, several hundred individual subunits are involved and, as they begin to assemble on DNA, they become involved in networks that create phase transitions, forming small biomolecular condensates. As described in Chapter 3, such condensates hold their proteins in loose proximity, such that, when one disassociates from the assembly, it can be retained nearby by a network of fluctuating weak interactions (see pp. 171–173). Consistent with this idea, many transcription regulators, coactivators, and co-repressors contain the type of low-complexity, unstructured regions that help to drive condensate formation.

How might this aid transcription? At least some of these transcription condensates contain additional copies of key proteins, including the Mediator complex (**Figure 7–26**). The presence of these extra copies in the same condensate is proposed to make transcription initiation an efficient but highly dynamic process, with proteins within the condensate rapidly exchanging on and off DNA. According to this view, Figure 7–20B represents only a frozen moment in transcription initiation. Whether such condensates form on most eukaryotic genes that are being transcribed—or on just those whose regulation is especially complex—remains to be determined.

Eukaryotic Transcription Repressors Can Inhibit Transcription in Several Ways

Although the "default" state of eukaryotic DNA packaged into nucleosomes is resistant to transcription, eukaryotes nonetheless use transcription regulators to repress the transcription of individual genes. These transcription repressors can rapidly turn off a gene that is being actively transcribed, and they can depress the rate of transcription even below that of the very low default value. Like the transcription activators discussed earlier, transcription repressors often work on a gene-by-gene basis. But unlike the bacterial repressors discussed earlier in this chapter, eukaryotic repressors do not directly compete with the RNA polymerase for access to the DNA. Instead, they use a variety of other mechanisms, some of which are illustrated in **Figure 7–27**. Like transcription activation, transcription repression can act through more than one mechanism at a given target gene, thereby ensuring especially efficient repression.

The different mechanisms of repression depicted in Figure 7–27 have different consequences for the ease with which a repressed gene can be reactivated. For most of the strategies, the repressed state is relatively easy to rapidly reverse, for

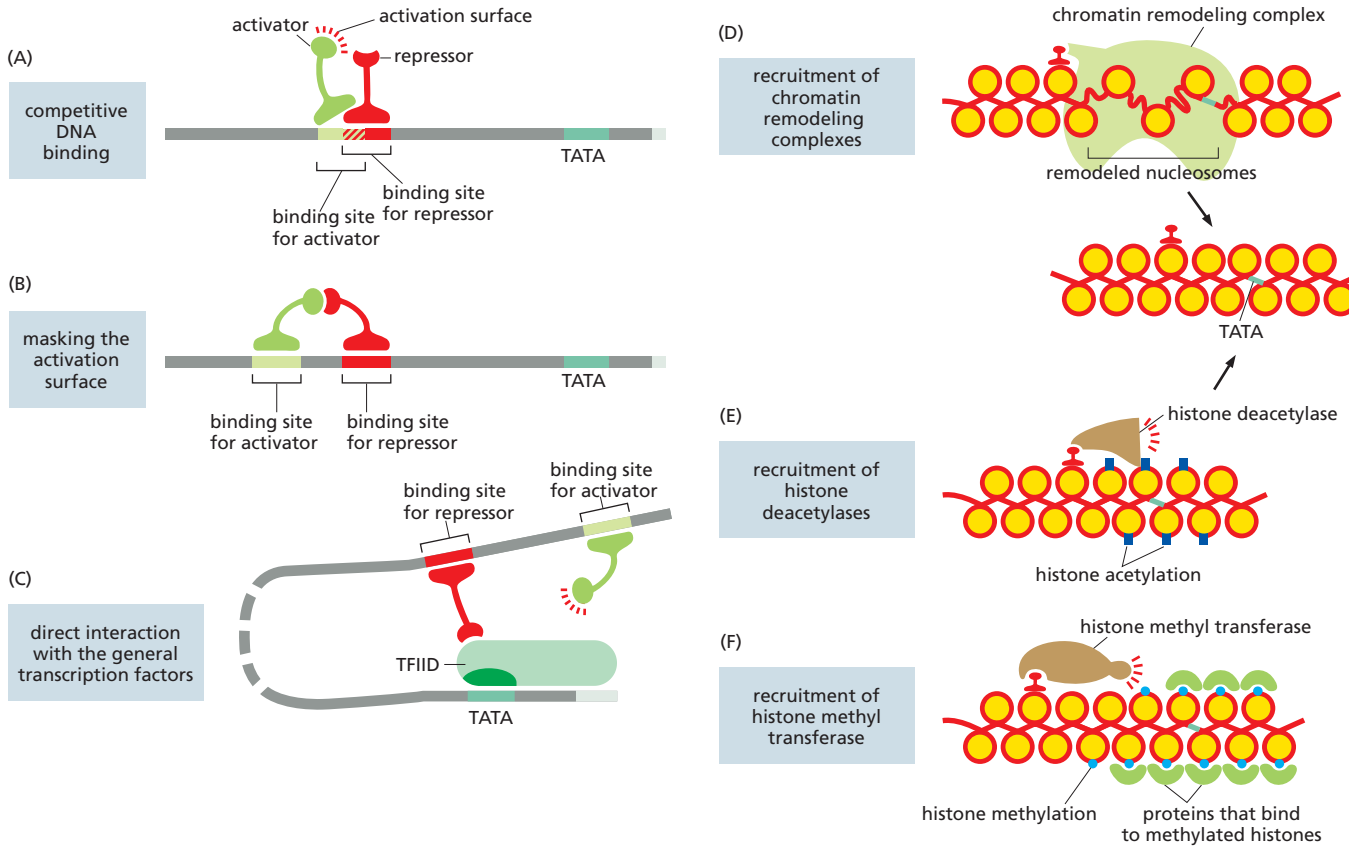


Figure 7–27 Six of the ways in which eukaryotic repressor proteins can operate. (A) A repressor protein outcompetes activator proteins for binding to the same regulatory DNA sequence. (B) Both activator and repressor proteins bind close to each other on DNA, and the repressor "quenches" the activator, preventing it from functioning (for example, by blocking the recruitment of its coactivators). (C) The repressor "poisons" assembly of the general transcription factors by binding to and stabilizing an intermediate. (D) The repressor recruits a chromatin remodeling complex that restores the nucleosomal state of the promoter region to its pre-transcriptional, default form. (E) The repressor attracts a histone deacetylase to the promoter, removing the histone acetylation needed for transcription initiation (see Figure 7–23). (F) Heterochromatin formation is triggered when a repressor attracts a specific histone methyl transferase that trimethylates either lysine 9 or lysine 27 on histone H3, thereby creating either H3K9me3- or H3K27me3-marked nucleosomes. "Read–write" mechanisms then spread each type of methylated nucleosome for thousands of nucleotide pairs along the DNA; they also help the methylation pattern to be inherited across cell divisions (see Figures 4–40 and 4–44). The final step in heterochromatin formation occurs when each type of modified nucleosome attracts additional proteins that condense the DNA and maintain it in a transcriptionally silent form.

example, by simply inactivating the repressor. But, the last mechanism—a directed methylation of specific histone amino acids that creates an unusually highly condensed form of chromatin, known as heterochromatin—is self-reinforcing and can propagate even when the initiating signal is no longer present (see Figure 4–44). As discussed in Chapter 4, chromatin that is marked by H3K9me3 (trimethylation of the lysine at position 9 of histone H3) appears to be the most difficult to transcribe. Typically located around centromeres and repeated DNA sequences such as inactive transposons, this type of heterochromatin strongly suppresses both genetic recombination and transcription. A different histone H3 modification (H3K27me3) is associated with a second form of heterochromatin that is also resistant to transcription. Although apparently easier to activate than the H3K9me3 form, this form of chromatin is also self-propagating and can persist across cell divisions, after the initiating signal has disappeared.

These two types of heterochromatin are used to tightly repress genes active in early development, presumably to make sure that these genes are not expressed in the mature organism. Tight, heritable gene repression is especially important to animals and plants whose growth depends on elaborate and complex developmental programs. Misexpression of a single gene at a critical time can have

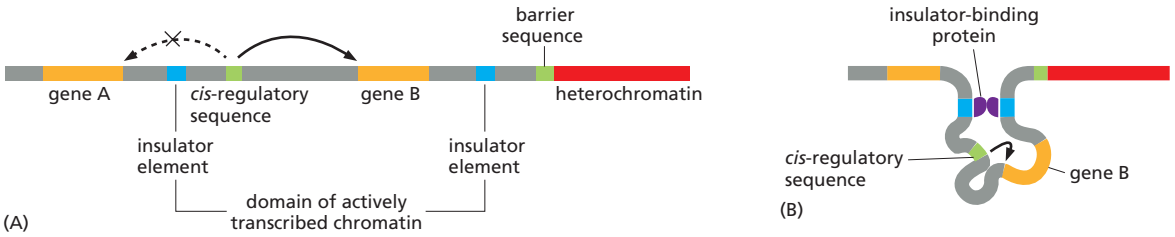


Figure 7–28 Schematic diagram summarizing the properties of insulators and barrier sequences. (A) Insulators directionally block the action of enhancers, whereas barrier sequences prevent the spread of heterochromatin. How barrier sequences likely function is depicted in Figure 4–41. (B) Insulator-binding proteins *(purple)* hold chromatin in loops that favor "correct" enhancer–promoter associations. Thus, gene B is properly regulated, and gene B's *cis*-regulatory sequences can be prevented from influencing the transcription of gene A. The major insulator-binding protein in mammals is denoted CTCF.

disastrous consequences for the individual. For this reason, many of the genes encoding the most important developmental regulatory proteins are kept tightly repressed, often by multiple mechanisms.

Insulator DNA Sequences Prevent Eukaryotic Transcription Regulators from Influencing Distant Genes

We have seen that all genes have control regions, which dictate at which times, under what conditions, and in what tissues the gene will be expressed. We have also seen that eukaryotic transcription regulators can act across very long stretches of DNA, with the intervening DNA looped out. How, then, are control regions of different genes kept from interfering with one another? For example, what keeps a transcription regulator bound on the control region of one gene from looping in the wrong direction and inappropriately influencing the transcription of an adjacent gene? And, if complex regulatory regions form biomolecular condensates, what keeps all of the control regions from forming a giant condensate where the regulatory information would become scrambled?

To avoid such cross-talk between control regions, several types of DNA elements compartmentalize the genome into discrete regulatory domains. In Chapter 4, we discussed *barrier sequences* that prevent the spread of heterochromatin into genes that need to be expressed (see Figure 4–41). A second type of DNA element, called an *insulator*, prevents *cis*-regulatory sequences from running amok and activating inappropriate genes (**Figure 7–28**). As we saw in Chapter 4, insulator sequences function by forming loops of chromatin, an effect mediated by specialized proteins that recognize them (see Figures 4–57 and 7–28B). The loops are thought to keep a gene and its control region in rough proximity and help to prevent the control region from "spilling over" to adjacent genes. More generally, the distribution of insulators and barrier sequences in a genome helps to divide it into independent domains of gene regulation and chromatin structure (see pp. 223–225).

The distribution of the more than 10,000 loops on the collection of mammalian chromosomes can change as cells differentiate or as they respond to changes in their environment. In addition, these loops formed by insulators are not static; rather, they undergo a continual process of loop extrusion and release that is driven by cohesion protein rings (see Figure 4–57). It has been proposed that the extrusion process itself helps to juxtapose enhancers with their matching promoters by sliding them past one another, while helping to break up inappropriate enhancer–promoter connections by physically separating them.

Although chromosomes are dynamically organized into domains that discourage control regions from acting indiscriminately, there are special circumstances where a control region located on one chromosome has been found to deliberately activate a gene located on a different chromosome. Although there is much we do not understand about this mechanism, it reflects the extreme versatility of transcription regulation strategies.

Summary

Transcription regulators switch the transcription of individual genes on and off in cells. In prokaryotes, these proteins typically bind to specific DNA sequences close to the RNA polymerase start site and, depending on the nature of the

regulatory protein and the precise location of its binding site relative to the start site, either activate or repress transcription of the gene. The flexibility of the DNA helix, however, also allows transcription regulators bound at distant sites to affect the RNA polymerase at the promoter by the looping out of the intervening DNA. The regulation of higher eukaryotic genes is much more complex, commensurate with a larger genome size and the large variety of cell types that are formed. A single eukaryotic gene is typically controlled by many transcription regulators bound to sequences that can be tens or even hundreds of thousands of nucleotide pairs from the promoter that directs transcription of the gene. Eukaryotic activators and repressors act by a wide variety of mechanisms—generally both altering chromatin structure and controlling the assembly of the general transcription factors and RNA polymerase at the promoter. They do this by attracting coactivators and co-repressors, protein complexes that perform the necessary biochemical reactions. The time and place that each gene is transcribed, as well as its rates of transcription under different conditions, are determined by the particular spectrum of transcription regulators present in the cell that bind to the control region of the gene.

MOLECULAR GENETIC MECHANISMS THAT CREATE AND MAINTAIN SPECIALIZED CELL TYPES

Although all cells must be able to switch genes on and off in response to changes in their environments, the cells of multicellular organisms have evolved this capacity to an extreme degree. In particular, once a cell in a multicellular organism becomes committed to differentiate into a specific cell type, the cell maintains this choice through many subsequent cell generations, which means that it remembers the changes in gene expression involved in the choice. 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, other changes in gene expression in eukaryotes, as well as most such changes in bacteria, are only transient. The tryptophan repressor, for example, switches off the tryptophan genes in bacteria only in the presence of tryptophan; as soon as tryptophan is removed from the medium, the genes are switched back on, and the descendants of the cell will have no memory that their ancestors had been exposed to tryptophan.

In this section, we shall examine some specific examples that illustrate how cell types are specified and maintained and how simple gene regulatory devices can be combined to create the "logic circuits" through which cells integrate signals and remember events in their past. We begin by considering one such complex gene control region that has been studied in great detail.

Complex Genetic Switches That Regulate *Drosophila* Development Are Built Up from Smaller Modules

We have seen that transcription regulators can be positioned at multiple sites along long stretches of DNA and that these proteins can bring into play coactivators and co-repressors that ultimately position and activate RNA polymerase to begin transcription. Here, we discuss how the numerous transcription regulators that bind to the control region of a gene can integrate external information, so as to cause the gene to be transcribed at the proper place and time.

The expression of the *Drosophila Even-skipped* (*Eve*) gene 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 embryo dies early in development. At the stage of development when *Eve* begins to be expressed, the embryo is a single giant cell containing multiple nuclei in a common cytoplasm. This cytoplasm contains a mixture of transcription regulators that are distributed unevenly along the length of the embryo, thus providing *positional information* that distinguishes one part of the embryo from another

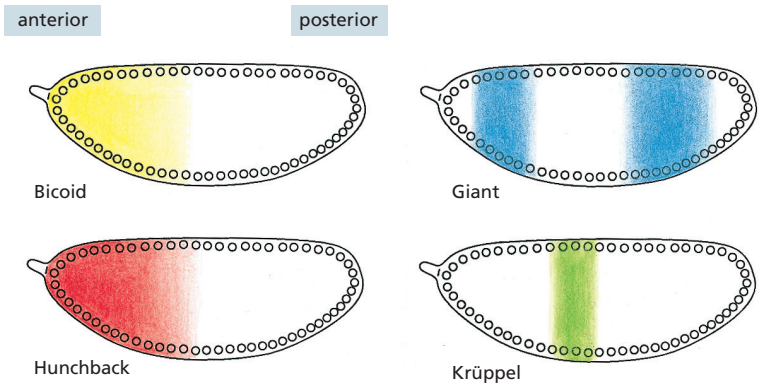


Figure 7–29 The nonuniform distribution of transcription regulators in an early *Drosophila* embryo. At this stage, the embryo is a syncytium; that is, multiple nuclei are contained in a common cytoplasm. Although the nuclei are shown in only a slice of the embryo, in reality, they are arranged in three dimensions around the inner surface of the giant cell.

(**Figure 7–29**). Although the nuclei are initially identical, they rapidly begin to express different genes because they are exposed to different transcription regulators: the nuclei near the anterior end of the developing embryo are exposed to a set of transcription regulators that is different from the set present at the middle and that present at the posterior end of the embryo.

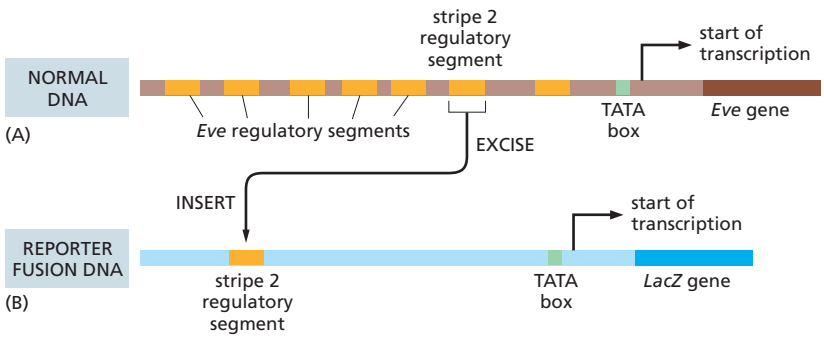
The regulatory DNA sequences that control the *Eve* gene have evolved to "read" the concentrations of transcription regulators at each position along the length of the embryo, so as to cause the *Eve* gene to be expressed in seven precisely positioned stripes, each initially five to six nuclei wide. How is this remarkable feat of information processing carried out? Although there is still much to learn, several general principles have emerged from studies of *Eve* and other genes that are similarly regulated.

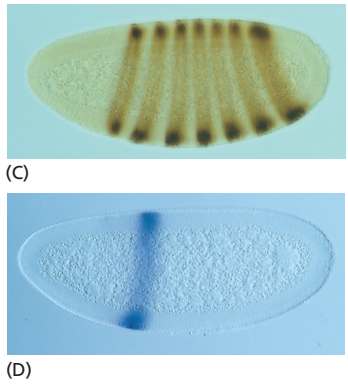
The control region of the *Eve* gene is very large (approximately 20,000 nucleotide pairs). It is formed from a series of relatively simple regulatory modules, each of which contains multiple *cis*-regulatory sequences and is responsible for specifying a particular stripe of *Eve* expression along the embryo. This modular organization of the *Eve* gene control region was revealed by experiments in which a particular regulatory module (say, that specifying stripe 2) is removed from its normal setting upstream of the *Eve* gene, placed in front of a reporter gene, and reintroduced into the *Drosophila* genome. When developing embryos derived from flies carrying this genetic construct are examined, the reporter gene is found to be expressed in precisely the position of stripe 2 but not in the other normal stripe positions (**Figure 7–30**). Similar experiments reveal the existence of other regulatory modules, which specify other stripes.

The *Drosophila Eve* Gene Is Regulated by Combinatorial Controls

A detailed study of the stripe 2 regulatory module has provided insights into how it reads and interprets positional information. The module contains recognition sequences for two transcription regulators that activate *Eve* transcription (Bicoid and Hunchback) and for two that repress it (Krüppel and Giant) (**Figure 7–31**).

Figure 7–30 Experiment demonstrating the modular construction of the *Eve* gene regulatory region. (A) A 480-nucleotide-pair section of the *Eve* regulatory region was removed and (B) inserted upstream of a test promoter that directs the synthesis of the enzyme β-galactosidase (the product of the *E. coli LacZ* gene—see Figure 7–18). (C, D) When this artificial construct was reintroduced into the genome of *Drosophila* embryos, the embryos (D) expressed β-galactosidase (detectable by histochemical staining) precisely in the position of the second of the seven *Eve* stripes. (C) The complete set of *Eve* stripes was detected using antibodies directed against the Eve protein. β-Galactosidase is simple to detect and thus provides a convenient way to monitor the expression specified by a gene control region. As used here, β-galactosidase is said to serve as a reporter, because it "reports" the activity of a gene control region. (C and D, courtesy of Stephen Small and Michael Levine.)





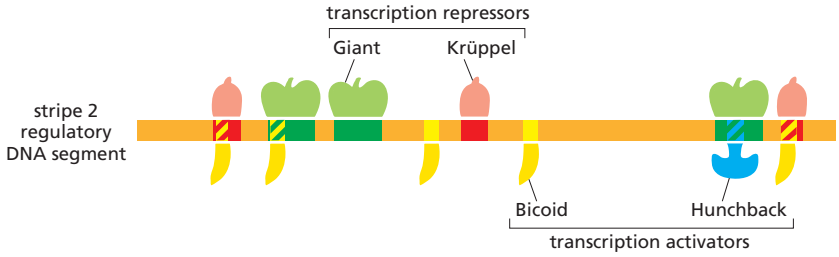


Figure 7–31 The *Eve* stripe 2 unit. The segment of the *Eve* gene control region identified in Figure 7–30 contains *cis*-regulatory sequences for four transcription regulators. It is known from genetic experiments that these four regulatory proteins are responsible for the proper expression of *Eve* in stripe 2. Flies that are deficient in the two gene activators Bicoid and Hunchback, for example, fail to efficiently express *Eve* in stripe 2. In flies deficient in either of the two gene repressors, Giant and Krüppel, stripe 2 expands and covers an abnormally broad region of the embryo. As indicated, in some cases the binding sites for the transcription regulators overlap, and the proteins can compete for binding to the DNA. For example, binding of Krüppel and binding of Bicoid to the site at the far right is mutually exclusive.

The relative concentrations of these four proteins determine whether the protein complexes that form at the stripe 2 module activate transcription of the *Eve* gene. Figure 7–32 shows the distributions of the four transcription regulators across the region of a *Drosophila* embryo where stripe 2 forms. It is thought that either of the two repressor proteins, when bound to the DNA, will turn off the stripe 2 module, whereas both Bicoid and Hunchback must bind for this module's maximal activation. This simple regulatory scheme suffices to turn on the stripe 2 module (and therefore the expression of the *Eve* gene) only in those nuclei located where the levels of both Bicoid and Hunchback are high and both Krüppel and Giant are absent—a combination that occurs in only one region of the early embryo. It is not known exactly how these four transcription regulators interact with coactivators and co-repressors to specify the final level of transcription across the stripe, but the outcome very likely relies on competition between activators and repressors that act by the mechanisms outlined in Figures 7–21, 7–22, and 7–27.

The stripe 2 element is autonomous, inasmuch as it specifies stripe 2 when isolated from its normal context (see Figure 7–30). The other stripe regulatory modules are thought to be constructed similarly, reading positional information provided by other combinations of transcription regulators. The entire *Eve* gene control region binds more than 20 different transcription regulators. Seven combinations of regulators—one combination for each stripe—specify *Eve* expression, while many other combinations (all those found in the interstripe regions of the embryo) keep all the stripe elements silent. A large and complex control region is thereby built from a series of smaller modules, each of which consists of a unique arrangement of short *cis*-regulatory sequences recognized by specific transcription regulators.

The *Eve* gene itself encodes a transcription regulator, which, after its pattern of expression is set up in seven stripes, controls the expression of other *Drosophila* genes. As development proceeds, the embryo is thus subdivided into finer and finer regions that eventually give rise to the different body parts of the adult fly, as discussed in Chapter 21.

Eve exemplifies the complexity of transcription control regions in plants and animals. As this example shows, control regions can respond to many different inputs, integrate this information, and produce a complex spatial and temporal output as

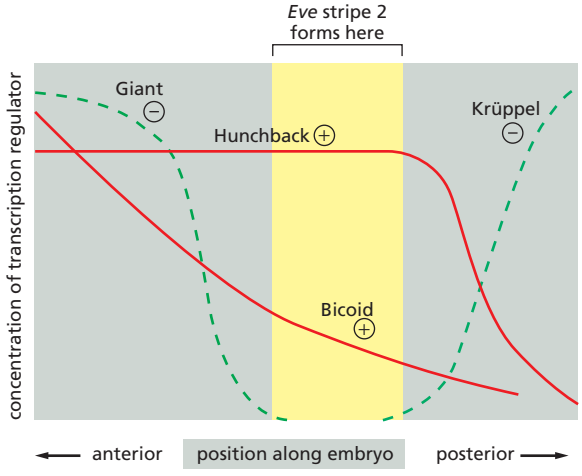


Figure 7–32 Distribution of the transcription regulators responsible for ensuring that *Eve* is expressed in stripe 2. The distributions of these proteins were visualized by staining a developing *Drosophila* embryo with antibodies directed against each of the four proteins, and a graph of the staining intensities is shown. The expression of *Eve* in stripe 2 occurs only at the position where the two activators (Bicoid and Hunchback) are present and the two repressors (Giant and Krüppel) are absent. In fly embryos that lack Krüppel, for example, stripe 2 expands posteriorly. Likewise, stripe 2 expands posteriorly if the DNA-binding sites for Krüppel in the stripe 2 module are inactivated by mutation (see also Figure 7–31).

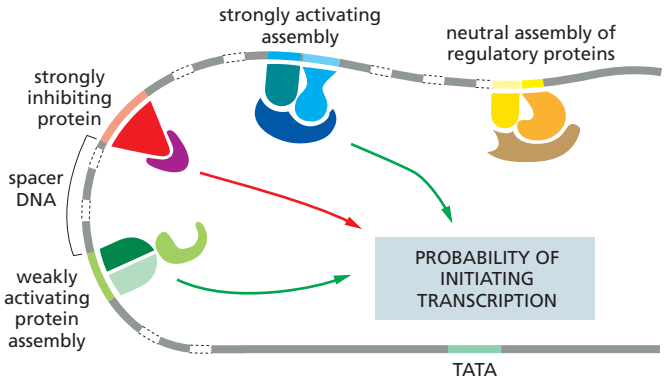


Figure 7–33 The integration of multiple inputs at a promoter. Multiple sets of transcription regulators, coactivators, and co-repressors can work together to influence transcription initiation at a promoter, as they do in the *Eve* stripe 2 module illustrated in Figure 7–31. It is not yet understood in detail how the cell achieves integration of multiple inputs, but it is likely that the final transcriptional activity of the gene results from competitions between activators and repressors that act by the mechanisms summarized in Figures 7–21, 7–22, and 7–27. As we saw earlier, for especially complex gene control regions, it has been proposed that these competitions take place and are "summed up" in localized biomolecular condensates formed by networks of weak interactions.

development proceeds. However, exactly how all these mechanisms work together to produce the final output is understood only in broad outline (**Figure 7–33**).

Transcription Regulators Are Brought into Play by Extracellular Signals

The above example from *Drosophila* clearly illustrates the power of combinatorial control, but this case is unusual in that the nuclei are exposed directly to positional cues in the form of concentrations of transcription regulators. In embryos of most other organisms and in all adults, individual nuclei are in separate cells, and extracellular information (including positional cues) must be passed across the plasma membrane so as to generate signals in the cytosol that cause different transcription regulators to become active in different cell types. Some of the different mechanisms that are known to be used to activate transcription regulators are diagrammed in **Figure 7–34**; in Chapter 15, we discuss how extracellular signals trigger these changes.

Like the fly example discussed earlier, mammalian enhancers are also modular. An example is the control region responsible for regulating the α-globin gene, which codes for one of the subunits of hemoglobin (see Figure 3–20). Here, five

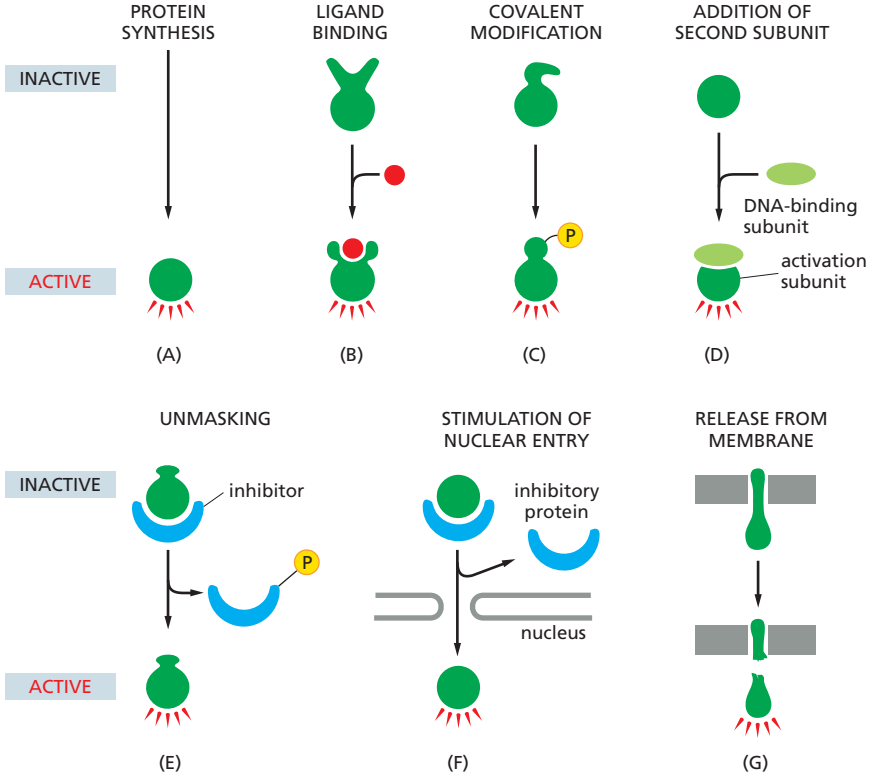


Figure 7–34 Some ways in which the activity of transcription regulators is controlled inside eukaryotic cells. (A) The protein is synthesized only when needed. (B) Activation by ligand binding. (C) Activation by covalent modification; phosphorylation is shown here, but many other modifications are possible (see Table 3–4, p. 175). (D) Formation of a complex between a DNA-binding protein and a separate protein with a transcription-activating domain. (E) Unmasking of an activation domain by the phosphorylation of an inhibitor protein. (F) Stimulation of nuclear entry by removal of an inhibitory protein that otherwise keeps the regulatory protein from entering the nucleus. (G) Release of a transcription regulator from a membrane bilayer by regulated proteolysis.

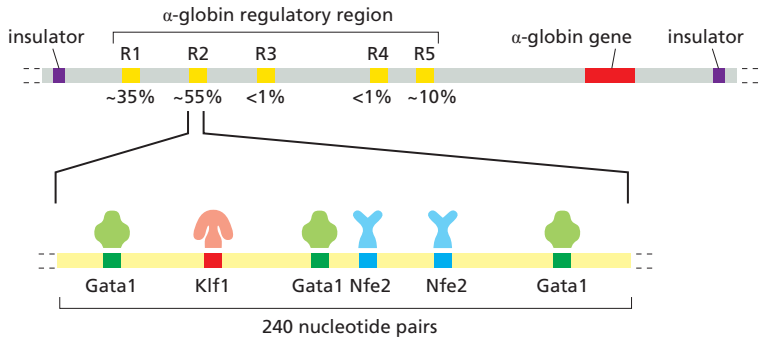


Figure 7–35 Modular structure of the control region for the mouse α-globin gene. Each of the five modules (R1–R5) can independently act as an enhancer, that is, they can each activate transcription of a reporter construct (see Figure 7–30B). However, the patterns of expression in a developing embryo are somewhat different for different modules. As indicated by the percentage designations, each module differs in the quantitative contributions it makes to the overall transcription rate in erythroid cells, with the total amount of mRNA being roughly equal to that of the sum of that produced by the individual modules. The additive properties of this control region suggest that the modules all affect the same step in transcription.

The combination of transcription regulators that recognize the R2 module, the most active of the five, is shown in the expanded view. These three transcription regulators are made in erythroid cells and are absent in most other cell types, explaining why expression of the globin gene occurs only in erythroid cells. Most of these same proteins also bind to the other α-globin regulatory modules, consistent with the modules working additively. As shown, insulator sequences flank the gene (including its control region), allowing the α-globin gene to be regulated independently of other genes on the same chromosome (see Figure 7–28). It is thought that modules R3 and R4 make no significant contribution to the overall transcription of the α-globin gene, but are once-functional modules that are in the slow evolutionary process of disappearing due to a gradual accumulation of mutations. (Courtesy of Helena Francis and Douglas Higgs.)

different modules are spread out over about 25,000 nucleotide pairs (**Figure 7–35**). Each of the five modules, when experimentally separated from the other four, can act as an independent enhancer to specify production of α-globin; but they do so only in erythroid cells, the precursors to red blood cells, because only erythroid cells express the appropriate transcription regulators. Red blood cells, which contain high concentrations of hemoglobin, are unusual in that they lack DNA and rely on their precursor cells to synthesize this protein.

Combinatorial Gene Control Creates Many Different Cell Types

We have seen that transcription regulators usually act in combination to control the expression of an individual gene. It is also generally true that each transcription regulator in an organism contributes to the control of many genes. This point is illustrated schematically in **Figure 7–36**, which shows how combinatorial gene

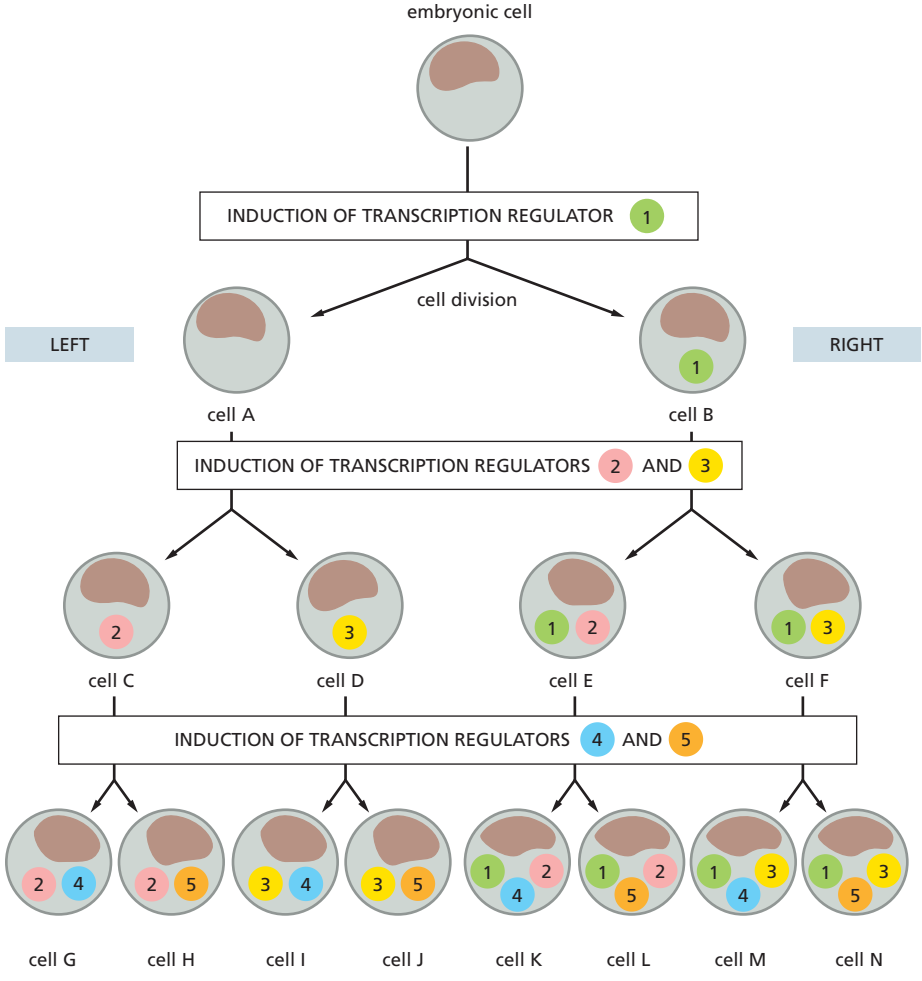


Figure 7–36 The importance of combinatorial gene control for development. Combinations of a few transcription regulators can generate many cell types during development. In this simple, idealized scheme, a "decision" to make one of a pair of different transcription regulators (shown as numbered circles) is made after each cell division. Sensing its relative position in the embryo, the daughter cell toward the *left side* of the embryo is always induced to synthesize the even-numbered protein of each pair, while the daughter cell toward the *right side* of the embryo is induced to synthesize the odd-numbered protein. The production of each transcription regulator is assumed to be self-perpetuating once it has become initiated (see Figure 7–42). In this way, through cell memory, the final combinatorial specification is built up step by step. In this purely hypothetical example, five different transcription regulators have created eight final cell types (G–N).

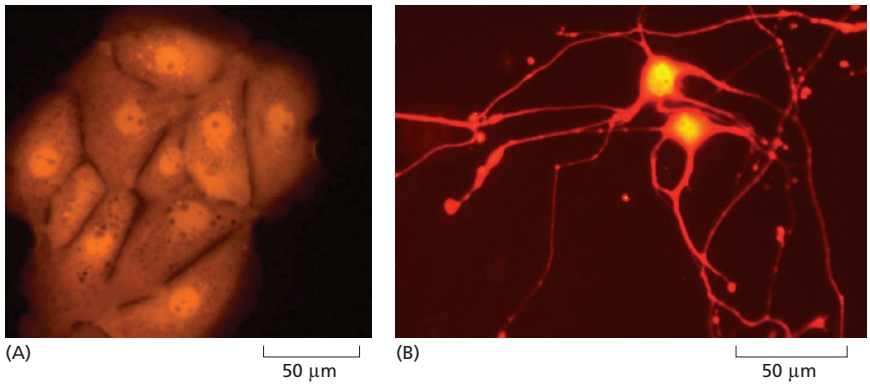


Figure 7–37 A small set of transcription regulators can convert one differentiated cell type into another. In this experiment, liver cells grown in culture (A) were converted into neuronal cells (B) by the artificial expression of three neuron-specific transcription regulators. (Both types of cells express a red fluorescent protein, which helps to visualize them.) This conversion involves the activation of many neuron-specific genes as well as the repression of many liver-specific genes. (From S. Marro et al., *Cell Stem Cell* 9:374–382, 2011. With permission from Elsevier.)

control makes it possible to generate a great deal of biological complexity even with relatively few transcription regulators.

Because of such combinatorial control, a given transcription regulator need not have a single, simply definable function as commander of a particular battery of genes or specifier of a particular cell type. Rather, transcription regulators can be likened to the words of a language: they are used with different meanings in a variety of contexts and rarely alone; it is the well-chosen combination that conveys the information that specifies a gene regulatory event.

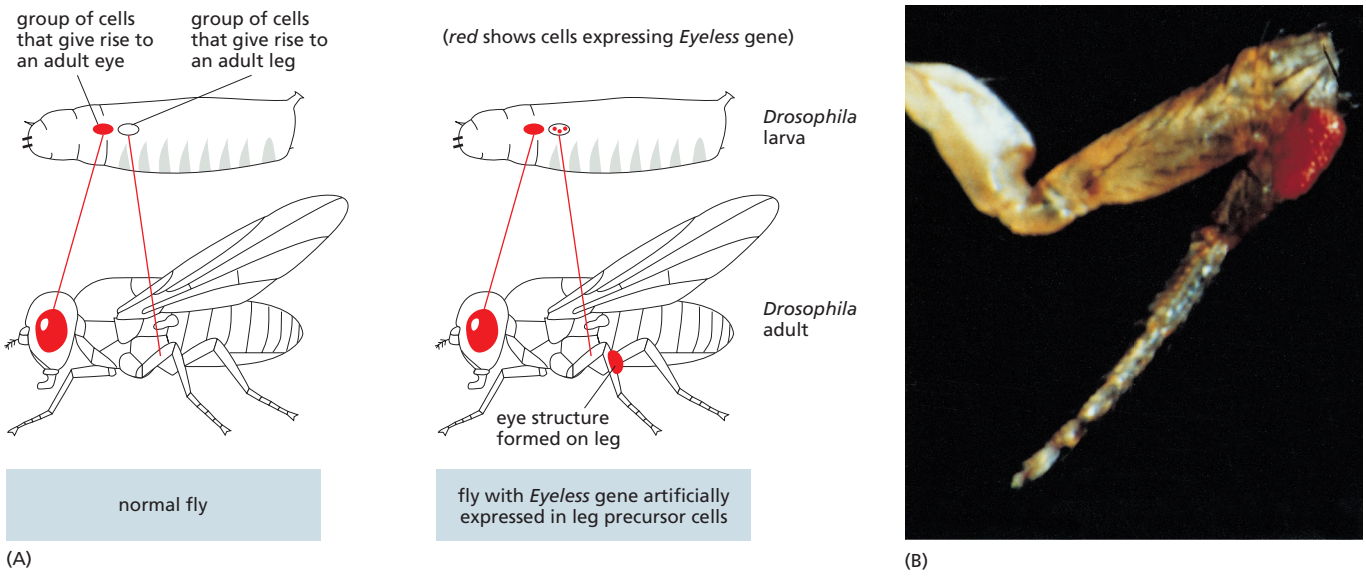
Because of combinatorial gene control, the effect of adding a new transcription regulator to a cell will depend on that cell's past history, inasmuch as this history determines the transcription regulators already present. Thus, during embryonic development, a cell can accumulate a series of transcription regulators that may not initially alter gene expression. Only the addition of the final members of a requisite combination of transcription regulators will complete the regulatory message, leading to large changes in gene expression.

The importance of a combination of transcription regulators for the specification of cell types is most easily demonstrated by their ability—when expressed artificially in a specific combination—to convert one type of cell to another. For example, the artificial expression of three neuron-specific transcription regulators in liver cells can convert the liver cells into functional nerve cells (**Figure 7–37**). In some cases, expression of even a single transcription regulator is sufficient to convert one cell type to another: when the gene encoding the transcription regulator MyoD is artificially introduced into fibroblasts cultured from skin connective tissue, the fibroblasts form muscle-like cells. As discussed in Chapter 22, 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 the addition of MyoD completes the unique combination required to direct the cells to become muscle.

An even more striking example is seen by artificially expressing, early in development, a single *Drosophila* transcription regulator (Eyeless) in groups of cells that would normally go on to form leg parts. Here, this abnormal gene expression change causes eye-like structures to develop in the legs (**Figure 7–38**).

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

Artificial manipulation of transcription regulators can also coax various differentiated cells to *de-differentiate* into pluripotent stem cells that are capable of giving rise to the different cell types in the body, as discussed in Chapter 22. Thus, when three specific transcription regulators are artificially expressed in cultured mouse fibroblasts, a number of cells become **induced pluripotent stem cells (iPS cells)**—cells that look and behave like the pluripotent embryonic stem



(ES) cells that are derived from embryos (**Figure 7–39**). This approach has been 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, a topic discussed in detail in Chapter 22.

Although it was once thought that cell differentiation was irreversible, it is now clear that by manipulating combinations of transcription regulators, cell types and differentiation pathways can be readily reversed and otherwise altered.

Combinations of Master Transcription Regulators Specify Cell Types by Controlling the Expression of Many Genes

As we saw in the introduction to this chapter, different cell types of multicellular organisms differ enormously in the proteins and RNAs they express. For example, only muscle cells express special types of actin and myosin that form the contractile apparatus, while nerve cells must make and assemble all the proteins needed to form dendrites and synapses. We have seen that these patterns of cell-type-specific expression are orchestrated by a combination of so-called **master transcription regulators**. In many cases, these proteins bind directly to *cis*-regulatory sequences of the genes particular to that cell type. Thus, MyoD binds directly to *cis*-regulatory sequences located in the control regions of the muscle-specific genes. In other cases, the master regulators control the

Figure 7–38 Expression of the *Drosophila Eyeless* gene in precursor cells of the fly leg triggers the development of an eye on the leg. (A) Simplified diagrams showing the result when a fruit fly larva contains either the normally expressed *Eyeless* gene *(left)* or an *Eyeless* gene that is additionally expressed artificially in cells that normally give rise to leg tissue *(right).* (B) Photograph of an abnormal leg that contains a misplaced eye (see also Figure 21–2). The transcription regulator was named Eyeless because its inactivation in otherwise normal flies causes the loss of eyes (see Figure 21–32). (B, courtesy of Walter Gehring.)

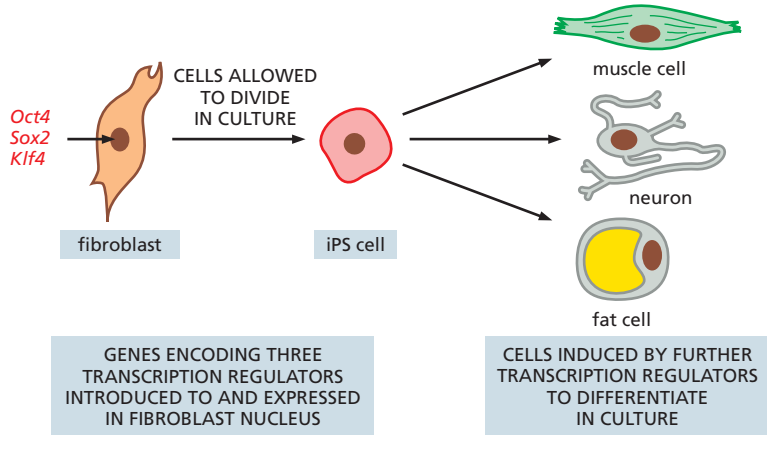


Figure 7–39 A combination of transcription regulators can induce a differentiated cell to de-differentiate into a pluripotent cell. The artificially induced expression of a set of three genes, each of which encodes a transcription regulator, can reprogram a fibroblast into a pluripotent cell with embryonic stem (ES) cell–like properties. Like ES cells, such induced pluripotent stem (iPS) cells can proliferate indefinitely in culture and can be stimulated by appropriate extracellular signal molecules to differentiate into almost any cell type found in the body. Transcription regulators such as Oct4, Sox2, and Klf4 are often called *master transcription regulators* because their expression is sufficient to trigger a change in cell identity. How two of these transcription regulators interact with DNA in a nucleosome is shown in Figure 7–13.

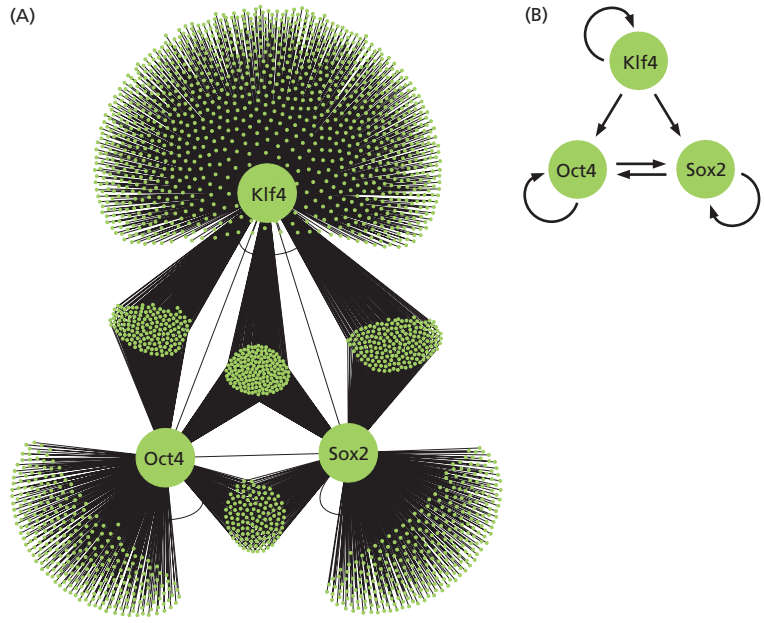


Figure 7–40 A portion of the transcription network specifying embryonic stem cells. (A) The three master transcription regulators in Figure 7–39 are shown as *large circles*. Genes whose *cis*-regulatory sequences are bound by each regulator in embryonic stem cells are indicated by a small *green dot* (representing the gene) connected by a thin line (representing the binding interaction). Note that many of the target genes are bound by more than one of the regulators. (B) The master regulators control their own expression. As shown here, the three transcription regulators bind to their own control regions (indicated by feedback loops), as well as those of the other master regulators (indicated by *straight arrows*). (Courtesy of Trevor Sorrells, based on data from J. Kim et al., *Cell* 132:1049–1061, 2008.)

expression of "downstream" transcription regulators that, in turn, bind to the control regions of other cell-type-specific genes and control their synthesis.

The specification of a particular cell type typically involves changes in the expression of several thousand genes. Genes whose protein products are required in the cell type are expressed at high levels, while those not needed are typically down-regulated. As might be imagined, the pattern of binding between the master regulators and all of the regulated genes can be extremely elaborate (**Figure 7–40**). When we consider that many of these regulated genes have control regions that span tens of thousands of nucleotide pairs, commensurate with the *Eve* example discussed earlier, we can begin to appreciate the enormous complexity of cell-type specification.

An outstanding question in biology is how the information in a genome is used to specify a multicellular organism. Although we have the general outline of the answer, we are far from understanding how a single cell type is completely specified, let alone a whole organism.

Specialized Cells Must Rapidly Turn Some Genes On and Off

Although they generally maintain their identities, specialized cells must constantly respond to changes in their environment. Among the most important changes are signals from other cells that coordinate the behavior of the whole organism. Many of these signals induce transient changes in gene transcription, and we discuss the nature of these signals in detail in Chapter 15. Here, we consider how specialized cell types rapidly and decisively switch groups of genes on and off in response to their environment. 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 maximally activate or repress that gene. This situation is analogous to dialing in the final number of a combination lock: the lock will spring open with only this simple addition if all of the other numbers have been previously entered. And just as the same number can complete the combination for many different locks, the addition of a particular protein can turn on many different genes.

An example is the rapid control of gene expression by the human glucocorticoid receptor protein. To bind to its *cis*-regulatory sequences in the genome, this transcription regulator must first form a complex with a molecule of a glucocorticoid steroid hormone, such as cortisol (see Figures 15–65 and 15–66). The body releases this hormone during times of starvation and intense physical activity,

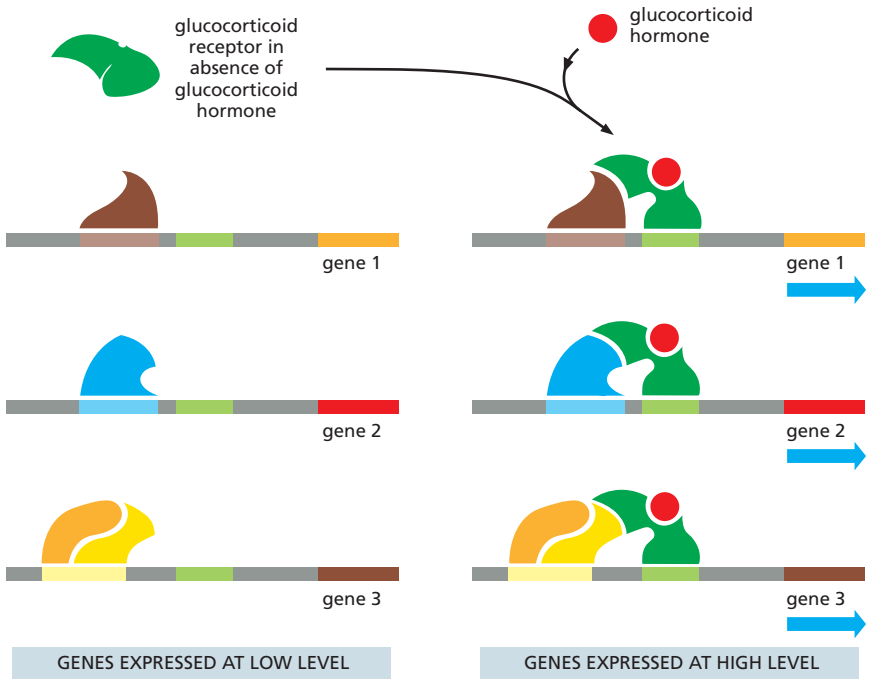


Figure 7–41 A single transcription regulator can coordinate the expression of many different genes. The action of the glucocorticoid receptor is illustrated schematically. On the *left* is a series of genes, each of which has various transcription regulators bound to its regulatory region. However, these bound proteins are not sufficient on their own to fully activate transcription. On the *right* is shown the effect of adding an additional transcription regulator—the glucocorticoid receptor in a complex with glucocorticoid hormone—that has a *cis*-regulatory sequence in the control region of each gene. The glucocorticoid receptor completes the combination of transcription regulators required for maximal initiation of transcription, and the genes are now maximally switched on as a set. When the hormone is no longer present, the glucocorticoid receptor dissociates from DNA, and the genes return to their prestimulated levels.

and among its other activities, it stimulates liver cells to increase the production of glucose from amino acids and other small molecules. To respond in this way, liver cells increase the expression of many different genes that code for metabolic enzymes, such as tyrosine aminotransferase, as we discussed earlier in this chapter (see Figure 7–3). Although these genes all have different and complex control regions, their maximal expression depends on the binding of the hormone–glucocorticoid receptor complex to its *cis*-regulatory sequence, which is present in the control region of each gene. When the body has recovered and the hormone is no longer present, the expression of each of these genes drops to its normal level in the liver. In this way, a single transcription regulator can rapidly control the expression of many different genes (**Figure 7–41**).

The effects of the glucocorticoid receptor are not confined to cells of the liver. In other cell types, activation of this transcription regulator by hormone also causes changes in the expression levels of many genes; the genes affected, however, are usually different from those affected in liver cells. As we have seen, each cell type has an individualized set of transcription regulators, and because of combinatorial control, these critically influence the action of the glucocorticoid receptor. Because the receptor is able to assemble with different sets of cell-type-specific transcription regulators, switching it on with hormone produces a different spectrum of effects in each cell type.

Differentiated Cells 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 17). 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 subsequent cell divisions. Thus, in the model we discussed in Figure 7–36, the production of each

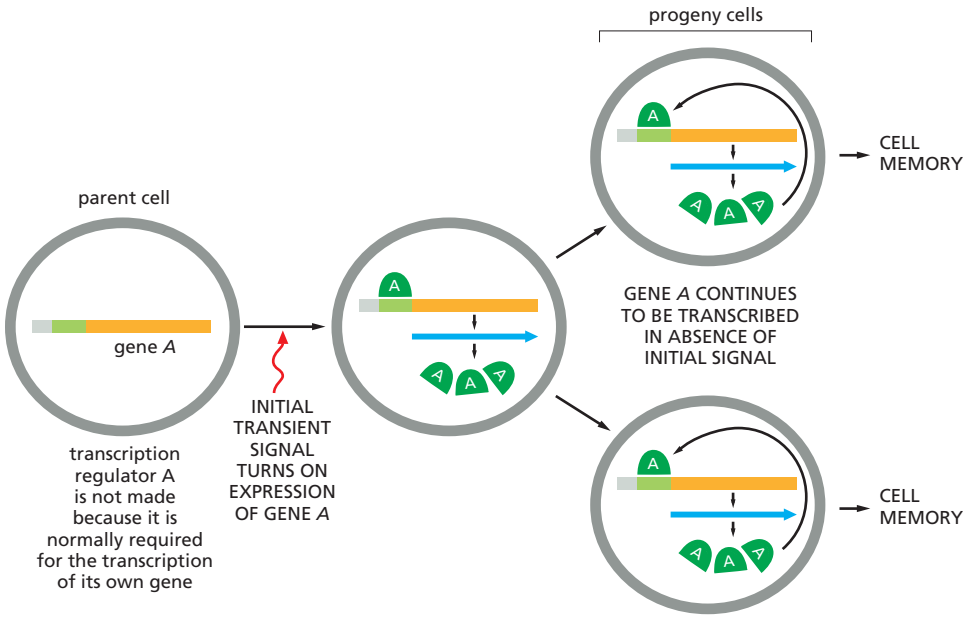


Figure 7–42 A positive feedback loop can create cell memory. Protein A is a master transcription regulator that activates the transcription of its own gene—as well as other cell-type-specific genes (not shown). All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of protein A.

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 cell-type transcription regulator activates transcription of its own gene, in addition to that of the other cell-type-specific genes needed to maintain the cell type. Each time a cell divides, the regulator is distributed to both daughter cells, where it continues to stimulate the positive feedback loop, making more of itself and the cell-type proteins it controls each division. Positive feedback is crucial for establishing "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 7–42**).

As was previously indicated in Figure 7–40B, the master regulators needed to maintain the pluripotency of iPS cells bind to *cis*-regulatory sequences in their own control regions, providing examples of this type of positive feedback loop. In addition, most of these pluripotent stem cell regulators also activate transcription of other master regulators, resulting in a complex series of indirect feedback loops. For example, if A activates B, and B activates A, this forms a positive feedback loop where A activates its own expression, albeit indirectly. The series of direct and indirect feedback loops observed in the iPS circuit is typical of other specialized cell circuits. Such a network structure strengthens cell memory, increasing the probability that a particular pattern of gene expression is transmitted through successive generations. For example, if the level of A drops below the critical threshold to stimulate its own synthesis, regulator B can rescue it. By successive application of this mechanism, a complex series of positive feedback loops among multiple transcription regulators can stably maintain a differentiated state through many cell divisions.

Positive feedback loops formed by transcription regulators are probably the most prevalent way of ensuring that daughter cells remember what kind of cells they are meant to be, and they are found in all species on Earth. For example, many bacteria and single-cell eukaryotes form different types of cells, and positive feedback loops lie at the heart of mechanisms that maintain their cell types through many rounds of cell division. Plants and animals also make extensive use of transcription feedback loops; but as we saw in Chapter 4 and shall discuss again later in the chapter, they have additional, more specialized mechanisms for making cell memory even stronger (see, for example, Figure 4–44). We will return

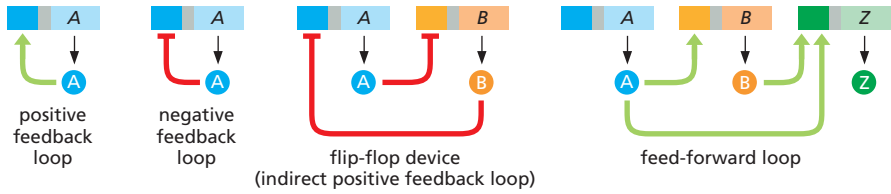


Figure 7–43 Common types of network motifs in transcription circuits. A and B represent transcription regulators, *green arrows* indicate positive transcription control, while *red lines with bars* depict negative transcription control. In the feed-forward loop, A and B represent transcription regulators that *both* activate the transcription of target gene *Z* (see also Figure 8–88).

to these additional mechanisms later in the chapter, but first, we consider how combinations of transcription regulators and *cis*-regulatory sequences can be combined to create other useful logic devices for the cell.

Transcription Circuits Allow the Cell to Carry Out Logic Operations

Simple gene regulatory switches can be combined to create all sorts of control devices, just as simple electronic switching elements in a computer can be linked to perform different types of operations. An analysis of gene regulatory circuits reveals that certain simple types of arrangements (called *network motifs*) are found over and over again in cells from widely different species. For example, positive and negative feedback loops are common in all cells (**Figure 7–43**). Whereas the former provides a simple memory device (see Figure 7–42), the latter is often used to keep the expression of a gene close to a standard level despite the variations in biochemical conditions inside a cell. Suppose, for example, that a transcription repressor protein binds to the regulatory region of its own gene and exerts a strong negative feedback, such that transcription falls to a very low rate when the concentration of the repressor protein is above some critical value (determined by its affinity for its DNA-binding site). The concentration of the protein can then be held close to the critical value, because any circumstance that causes a fall below that value can lead to a steep increase in synthesis, and any that causes a rise above that value will lead to synthesis being switched off. Such adjustments will, however, take time, so that an abrupt change of conditions will cause a disturbance of gene expression that is strong but transient. If there is a delay in the feedback loop, the result may be spontaneous oscillations in the expression of the gene (see Figure 15–18). The different types of behavior produced by a feedback loop will depend on the details of the system; for example, how tightly the transcription regulator binds to its *cis*-regulatory sequence, its rate of synthesis, and its rate of decay. We discuss these issues in quantitative terms and in more detail in Chapter 8.

With two or more transcription regulators, the possible range of circuit behaviors becomes more complex. Some bacterial viruses contain a common type of two-gene circuit that can flip-flop between expression of one gene and expression of the other (see Figure 7–43). Another common circuit arrangement is called a *feed-forward* loop; such a loop can serve as a filter, responding to input signals that are prolonged but disregarding those that are brief (**Figure 7–44**). Although

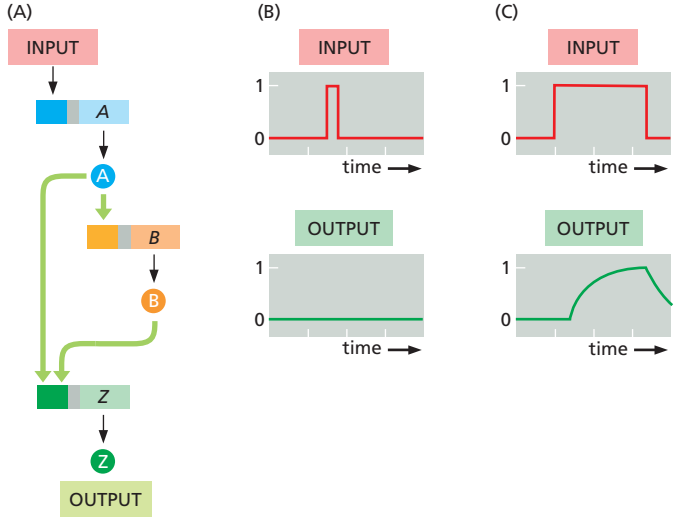


Figure 7–44 How a feed-forward loop can measure the duration of a signal. (A) In this theoretical example, transcription regulators A and B must both be present on the DNA for transcription of gene *Z*, and gene *A* becomes active only when an input signal is present. (B) If the input signal to gene *A* is brief, it does not stay active long enough for transcription regulator B to accumulate, and gene *Z* is not transcribed. (C) If the signal to gene *A* persists, transcription regulators A and B both accumulate, and gene *Z* is transcribed. This arrangement allows the cell to ignore rapid fluctuations of the input signal and respond only to persistent levels. This strategy could be used, for example, to distinguish between random noise and a true signal.

The behavior shown here was computed for one particular set of parameter values describing the quantitative properties of transcription regulators A and B, as well as the product of gene *Z*, along with their syntheses. With different values of these parameters, feed-forward loops can in principle perform other types of "calculations." Many feed-forward loops have been discovered in cells. As explained in Chapter 8, theoretical analyses are needed to help researchers to discern—and subsequently test—the different ways in which these circuits function (see Figures 8–87 and 8–88). (Adapted from S.S. Shen-Orr et al., *Nat. Genet.* 31:64–68, 2002.)

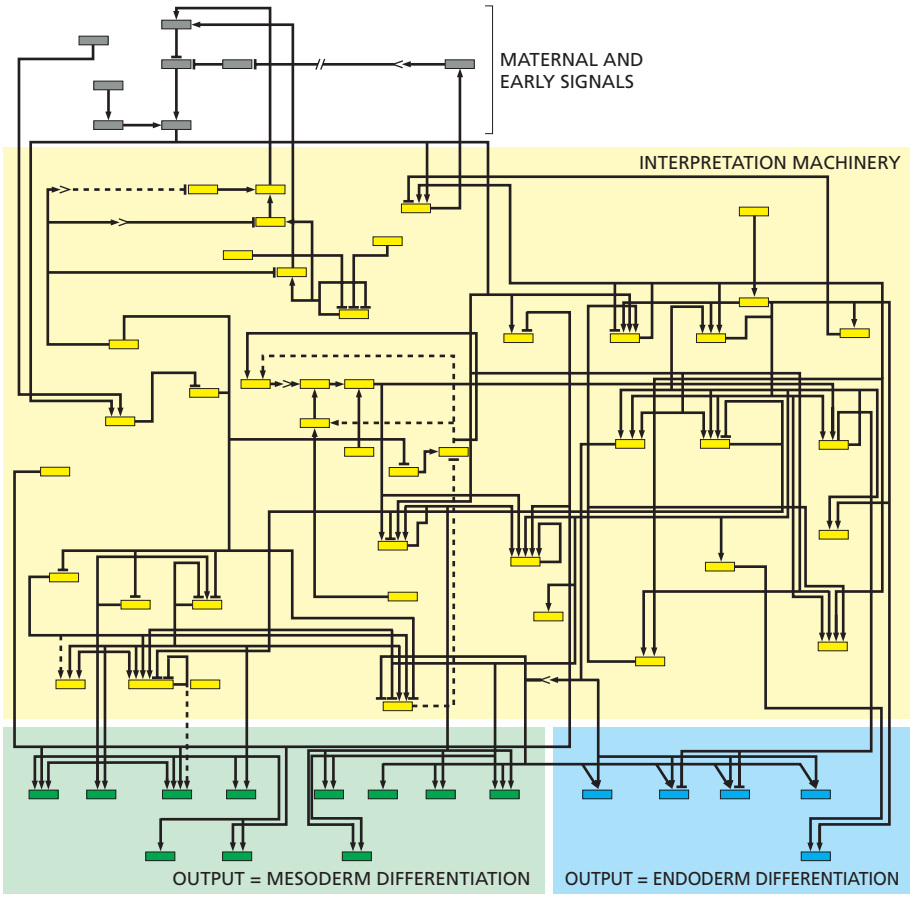


Figure 7–45 The exceedingly complex gene circuit that specifies a portion of the developing sea urchin embryo. Each *colored small box* represents a different gene. Those in *yellow* code for transcription regulators, and those in *green* and *blue* code for proteins that give cells of the mesoderm and endoderm, respectively, their specialized characteristics. Genes depicted in *gray* are largely active in the mother and provide the egg with cues needed for proper development. As in Figure 7–43, arrows depict instances in which a transcription regulator activates the transcription of another gene, and lines ending in bars indicate examples of gene repression. (From I.S. Peter and E.H. Davidson, *Nature* 474:635–639, 2011. With permission from Springer Nature.)

they arose as products of evolution, without advance planning or design, these various network motifs resemble some of the miniature logic devices found in electronic circuits. And, like circuits designed by humans, they can process information in surprisingly sophisticated ways.

The simple types of devices just illustrated are often found joined together, creating exceedingly complex circuits (**Figure 7–45**). Each cell in a developing multicellular organism is equipped with similarly complex control machinery, and it must, in effect, use its intricate system of interlocking transcription switches to "compute" how it should behave at each time point in response to the many different past and present inputs received. We are only beginning to understand how to study such complex intracellular control networks. Indeed, without new approaches, coupled with quantitative information that is far more precise and complete than we now possess, it will be impossible to predict the behavior of a system such as that shown in Figure 7–45. As explained in Chapter 8, a circuit diagram by itself is insufficient to deeply understand biological mechanisms.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different sets of genes to be transcribed in different cells. The transcription of any particular gene is generally controlled by a combination of transcription regulator proteins. Each type of cell in a higher eukaryotic organism contains a specific set of transcription regulators that ensures the expression of only those genes appropriate to that type of cell. A given transcription regulator may be active in a variety of circumstances, and it is typically involved in the regulation of many different genes.

Because specialized animal cells can maintain their unique character through many cell-division cycles, and even when grown in culture, there must exist

mechanisms to ensure this cell memory. Direct or indirect positive feedback loops, which enable transcription regulators to perpetuate their own synthesis, provide one of the simplest mechanisms for producing a cell memory. Transcription circuits also provide the cell with the means to carry out many other types of logic operations. Simple transcription circuits combined into large regulatory networks drive highly sophisticated programs of embryonic development that will require new approaches to fully decipher.

MECHANISMS THAT REINFORCE CELL MEMORY IN PLANTS AND ANIMALS

Thus far in this chapter, we have emphasized the regulation of gene transcription by proteins that associate either directly or indirectly with DNA. However, DNA itself can be covalently modified, and, as we saw in Chapter 4, certain types of chromatin states can be inherited. In this section, we shall see how these phenomena provide additional opportunities for the regulation of gene expression, particularly in mammals. Near the end of this section, we discuss how a whole chromosome can be transcriptionally shut down using such mechanisms, and how this state can be maintained through many cell divisions.

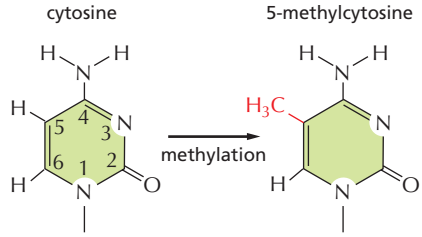


Figure 7–46 Formation of 5-methylcytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates, this event is largely confined to selected cytosine (C) nucleotides located in the sequence CG. CG sequences are sometimes denoted as CpG sequences, where the p indicates a phosphate linkage to distinguish it from a CG base pair. In this chapter, we will continue to use the simpler nomenclature CG to indicate this dinucleotide.

Patterns of DNA Methylation Can Be Inherited When Vertebrate Cells Divide

In vertebrate cells, the methylation of cytosine provides one mechanism through which gene expression patterns can be passed on to progeny cells. The methylated form of cytosine, 5-methylcytosine (5-methyl C), has the same relation to cytosine that thymine has to uracil, and the modification likewise has no effect on base-pairing (**Figure 7–46**). **DNA methylation** in vertebrate DNA occurs on cytosine (C) nucleotides largely in the sequence CG, which is base-paired to exactly the same sequence (in opposite orientation) on the other strand of the DNA helix. Consequently, a simple mechanism permits the existing pattern of DNA methylation to be inherited directly by the daughter DNA strands. An enzyme called *maintenance methyl transferase* acts preferentially on those CG sequences that are base-paired with a CG sequence that is already methylated. As a result, the pattern of DNA methylation on the parent DNA strand serves as a template for the methylation of the daughter DNA strand, causing this pattern to be inherited directly after DNA replication (**Figure 7–47**).

Although DNA methylation patterns can be maintained in differentiated cells by the mechanism shown in Figure 7–47, methylation patterns are dynamic during mammalian development. Shortly after fertilization, there is a genome-wide wave of demethylation, when the vast majority of methyl groups are lost from the

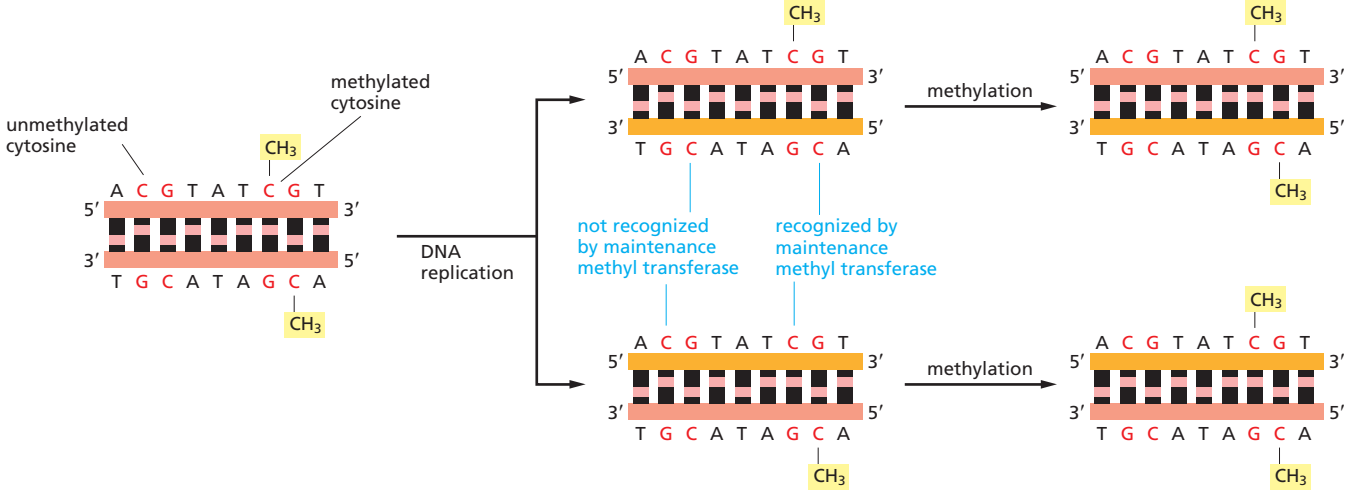


Figure 7–47 How DNA methylation patterns are faithfully inherited. In vertebrate DNA, a large fraction of the cytosine nucleotides in the sequence CG is methylated (see Figure 7–46). Because of the existence of a methyl-directed methylating enzyme (the maintenance methyl transferase), once a pattern of DNA methylation is established, that pattern of methylation is inherited in the progeny DNA, as shown.

DNA. This demethylation may occur either by suppression of maintenance DNA methyl transferase activity, resulting in the passive loss of methyl groups during each round of DNA replication, or by *DNA demethylases* that actively remove methyl groups from DNA. Later in development, several *de novo DNA methyl transferases* come into play and methylate about 70% of the CG sequences in the genome. This extensive methylation occurs largely indiscriminately, although proteins that are bound to specific sequences on the genome can block the methylation of those sequences. In addition, some sequence-specific DNA-binding proteins direct DNA methylases to specific locations in genomes, resulting in very high local densities of methylation in the neighborhoods of those DNA-bound proteins. Conversely, DNA demethylases can also be directed to certain regions of the genome, resulting in loss of methyl groups in those regions. Despite these selective mechanisms, the patterns of overall methylation across differentiated cell types are broadly similar, and many methylated positions—on their own—appear to have little or no impact on gene expression.

DNA methylation has several uses in the vertebrate cell. A very important role of dense methylation is to work in conjunction with other gene expression control mechanisms to establish a particularly efficient form of gene repression. This combination of mechanisms enables unneeded eukaryotic genes to be repressed to a very high degree. The rate at which a vertebrate gene is transcribed can vary 10^6-fold between one tissue and another, and unexpressed vertebrate genes are much less "leaky" in terms of transcription than bacterial genes, in which the largest known differences in transcription rates between expressed and unexpressed gene states are only about 1000-fold.

Dense DNA methylation helps to repress transcription in several ways. The methyl groups on methylated cytosines lie in the major groove of DNA and interfere directly with the binding of some proteins (transcription regulators as well as the general transcription factors) required for transcription initiation. In addition, the cell contains a repertoire of proteins that bind specifically to methylated DNA. The best characterized of these also associate with histone-modifying enzymes, leading to a repressive, heterochromatin state where chromatin structure and DNA methylation act synergistically (**Figure 7–48**).

Many genes that are needed only in differentiated cells are tightly repressed in this way in embryonic cells. As differentiation proceeds, they become activated, although this process typically requires many steps, often involving "pioneer factors" (see Figure 7–13), histone demethylases, and DNA demethylases. The latter enzymes convert 5-methyl C to 5-hydroxymethyl C, which is later replaced by C either through DNA repair (see Figure 5–41A) or, passively, through multiple rounds of DNA replication. In addition, many genes active in embryonic tissues become repressed during differentiation by the mechanisms shown in Figure 7-48. The reactivation of these genes is one of the key steps in converting differentiated cells back into stem cells, as explained in Chapter 22.

CG-Rich Islands Are Associated with Many Genes in Mammals

Because of the way in which DNA repair enzymes work, methylated C nucleotides in the vertebrate genome tend to be eliminated in the course of evolution. Accidental deamination of an unmethylated C gives rise to U (see Figure 5–38B), which is not normally present in DNA and thus is recognized easily by the DNA repair enzyme uracil DNA glycosylase. The deamination product is thereby excised and replaced with a C, as discussed in Chapter 5. But accidental deamination of a 5-methyl C cannot be repaired in this way, for the deamination product is a T and so is indistinguishable from the other, nonmutant T nucleotides in the DNA. Although a special repair system exists to remove some of these incorrect T nucleotides, many of the deaminations escape detection, so that those C nucleotides in the genome that are methylated tend to mutate to T over evolutionary time.

During the course of evolution, more than three out of every four CGs have been lost in this way, leaving vertebrates with a remarkable deficiency of this dinucleotide. This ratio probably reflects a balance between methylated CG loss

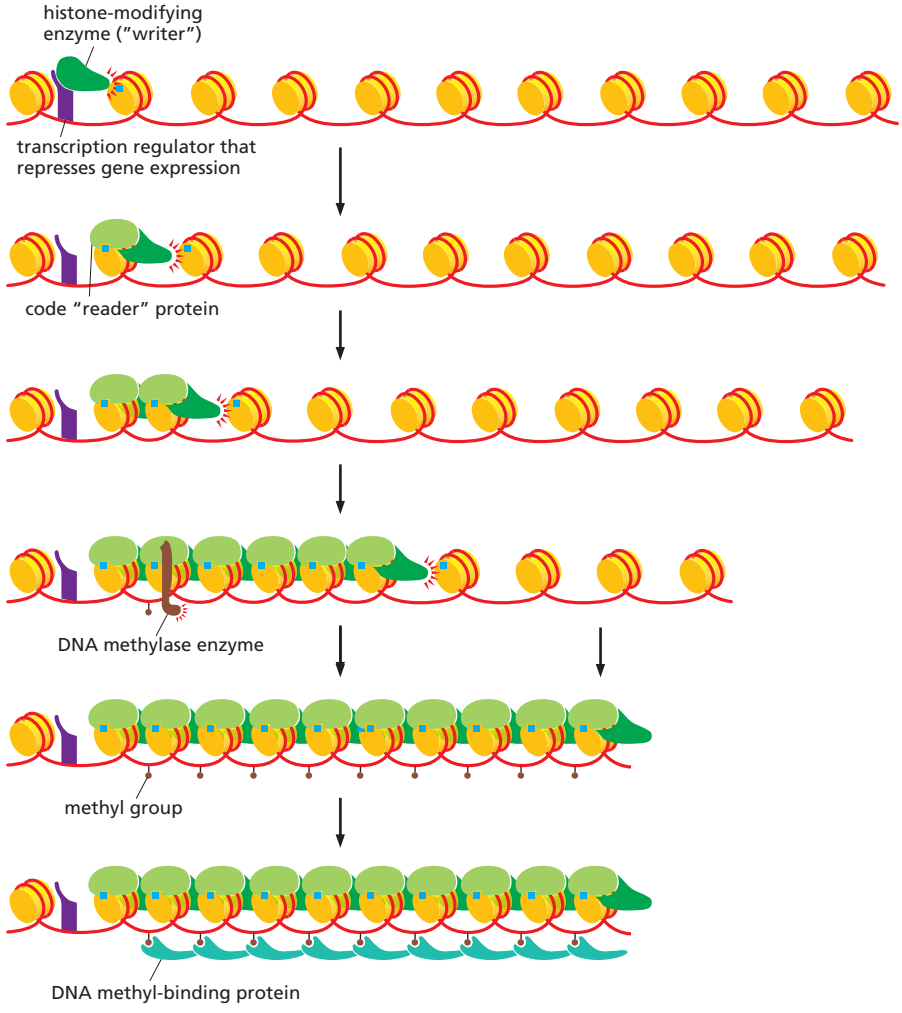


Figure 7–48 Multiple mechanisms can produce especially stable gene repression. In this schematic example, histone reader and writer proteins (discussed in Chapter 4), triggered by transcription regulators, establish a repressive form of chromatin whose nucleosomes are marked by the trimethylation of specific lysine amino acids in histones (see Figure 4–35, as well as Figure 7–27F). An additional layer of repression can occur when a *de novo* DNA methylase is attracted by the modified histones and methylates nearby cytosines in DNA; and these are, in turn, bound by DNA methyl-binding proteins. During DNA replication, some of the modified *(blue dot)* histones will be inherited by one daughter chromosome, some by the other, and in each daughter they can induce reconstruction of the same pattern of chromatin modifications (see Figure 4–44). At the same time, the mechanism shown in Figure 7–47 will cause both daughter chromosomes to inherit the same methylation pattern. This makes the two mechanisms for inheriting a repressed gene mutually reinforcing, accounting for the inheritance by daughter cells of both the histone and the DNA modifications. It can also explain the tendency of some chromatin modifications to spread along a chromosome (see Figure 4–39). This type of heterochromatin is assembled and disassembled on different genes as mammalian development proceeds, depending on whether the gene product is needed. For example, when endoderm precursor cells differentiate into the hepatocytes of the liver, an estimated 6000 genes are unpackaged from this repressive form of chromatin and become actively transcribed. At roughly the same time, about 1600 genes active in endoderm cells become packaged into this type of chromatin and are thereby tightly repressed in hepatocytes.

by DNA repair and CG gain by random mutation. The CG sequences that remain are very unevenly distributed in the genome; they are present at 10 times their average density in selected regions, called **CG islands**, which average 1000 nucleotide pairs in length. The human genome contains roughly 20,000 CG islands, and they usually include promoters of genes. For example, 60% of human protein-coding genes have promoters embedded in CG islands, and these include virtually all the promoters of the so-called *housekeeping genes*—those genes that code for the many proteins that are essential for cell viability and are therefore expressed in nearly all cells (Figure 7–49). Over evolutionary time scales, the CG

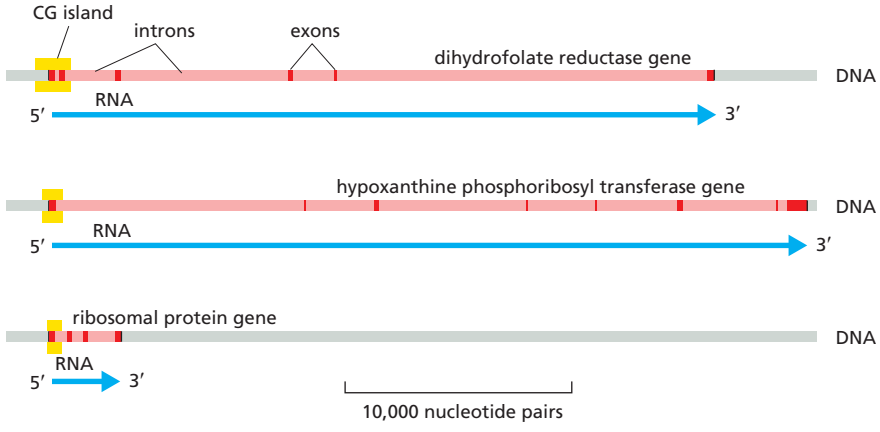


Figure 7–49 The CG islands surrounding the promoter in three mammalian housekeeping genes. The *yellow boxes* show the extent of each island. As for most genes in mammals, the exons *(dark red)* are very short relative to the introns *(light red)*. (Adapted from A.P. Bird, *Trends Genet.* 3:342–347, 1987.)

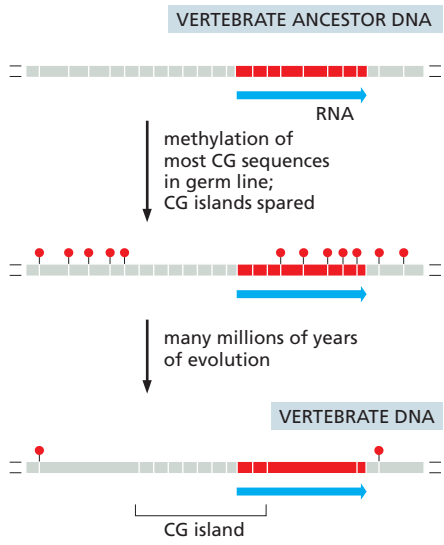


Figure 7–50 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes. The vertical *white lines* mark the location of CG dinucleotides in the DNA sequences, while *red circles* indicate the presence of a methyl group on the CG dinucleotide. CG sequences that lie in regulatory sequences of genes that are transcribed in germ cells are unmethylated and therefore tend to be retained in evolution. Methylated CG sequences, on the other hand, tend to be lost through deamination of 5-methyl C to T, unless the CG sequence is critical for survival.

islands were spared the accelerated mutation rate of bulk CG sequences because they remained unmethylated in the germ line (**Figure 7–50**).

CG islands remain unmethylated in most somatic tissues whether or not the associated gene is expressed. The unmethylated state is maintained by a group of proteins that bind specifically to unmethylated CG sequences in the genome and modify the neighboring nucleosomes by methylating histone H3 (on the lysine at position 4; see Figure 4–35). These modified nucleosomes somehow repel the *de novo* methylases, and the unmethylated state is thereby continually maintained. Unmethylated CG islands have several properties that make them particularly suitable for promoters. For example, some of the same proteins that protect them from methylation recruit additional histone-modifying enzymes that decompress the chromatin, making the islands particularly "promoter friendly." As a result, RNA polymerase is often found bound to promoters within CG islands, even when the associated gene is not being actively transcribed. At unmethylated CG islands, the competition between polymerase binding and nucleosome assembly at promoters is thus always tipped toward the former. However, additional steps are needed for the final "push" to transcribe the adjacent gene; these are directed by transcription regulators that bind to *cis*-regulatory sequences of DNA, often well upstream from the CG islands.

Genomic Imprinting Is Based on DNA Methylation

Mammalian cells are diploid, containing one set of genes inherited from the father and one set from the mother. The expression of a small minority of genes depends on which parent they came from: when the paternally inherited gene copy is active, the maternally inherited gene copy is silent, or vice versa. This phenomenon is called **genomic imprinting**.

Roughly 300 genes are imprinted in humans. Because only one copy of an imprinted gene is expressed, imprinting can "unmask" harmful mutations that would normally be covered by the other, functional copy. For example, Angelman syndrome, a disorder of the nervous system in humans that causes reduced mental ability and severe speech impairment, results from a gene deletion on one chromosomal homolog and the silencing, by imprinting, of the intact gene on the other homolog.

The *insulin-like growth factor-2* (*Igf2*) gene in the mouse provides a well-studied example of imprinting. Mice that do not express *Igf2* at all are born half the size of normal mice. However, only the paternal copy of *Igf2* is transcribed, and only this gene copy matters for the phenotype. As a result, mice with a mutated paternally derived *Igf2* gene are stunted, while mice with a mutated maternally derived *Igf2* gene are normal.

In the early embryo, genes subject to imprinting are marked by methylation according to whether they were derived from a sperm or an egg chromosome. In this way, DNA methylation is used as a mark to distinguish two copies of a gene that can be otherwise identical (**Figure 7–51**). Such imprinted genes are somehow protected from the wave of DNA demethylation that takes place shortly after fertilization (see pp. 435–436), enabling the somatic cells produced during embryonic development to "remember" the parental origin of each of the two copies of the gene and to regulate their expression accordingly. In most cases, the methyl imprint silences nearby gene expression. In some cases, however, it can activate expression of a gene. In the case of *Igf2*, for example, methylation of an insulator element on the paternally derived chromosome blocks its function and allows distant *cis*-regulatory sequences to activate transcription of the *Igf2* gene.

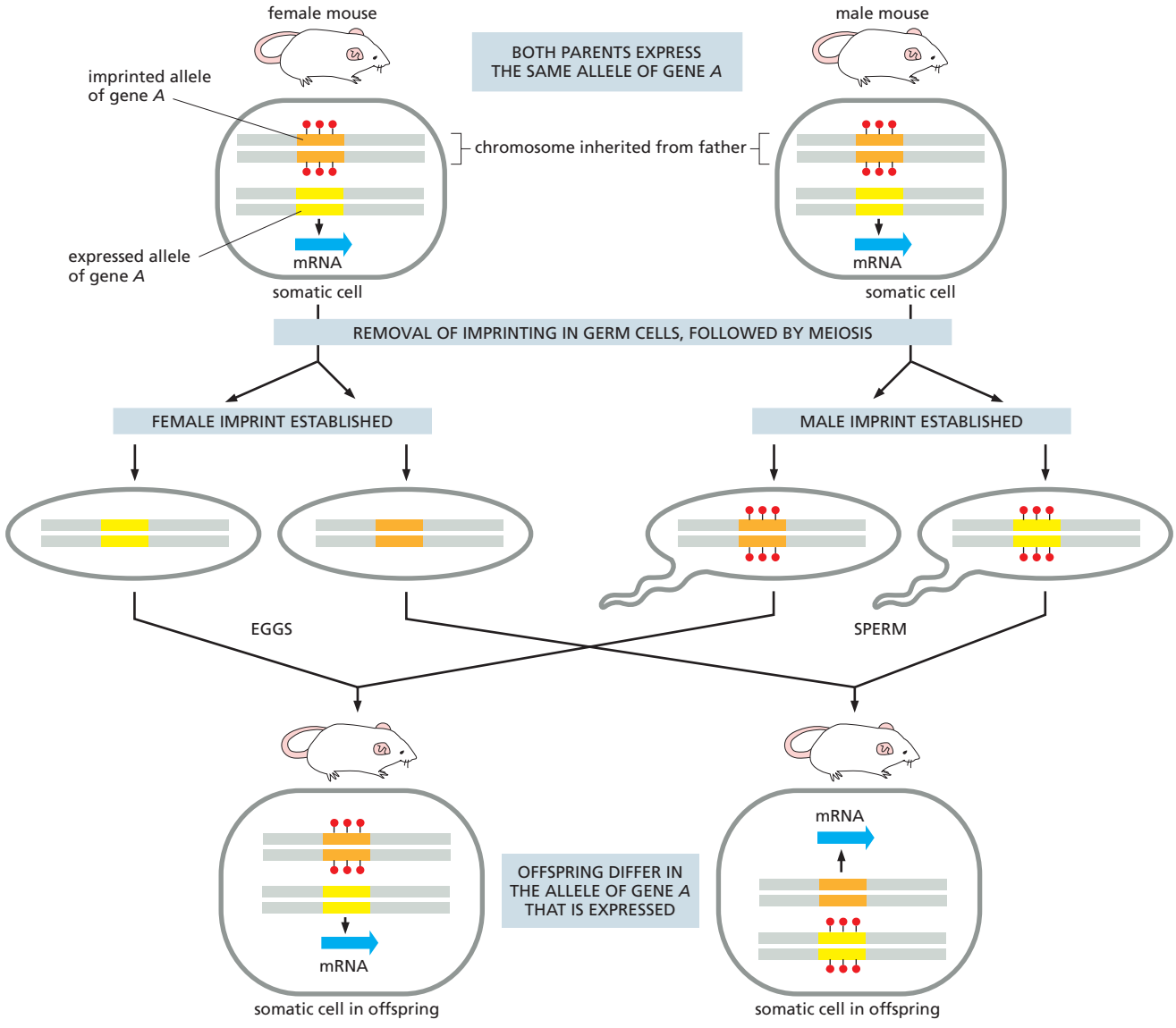
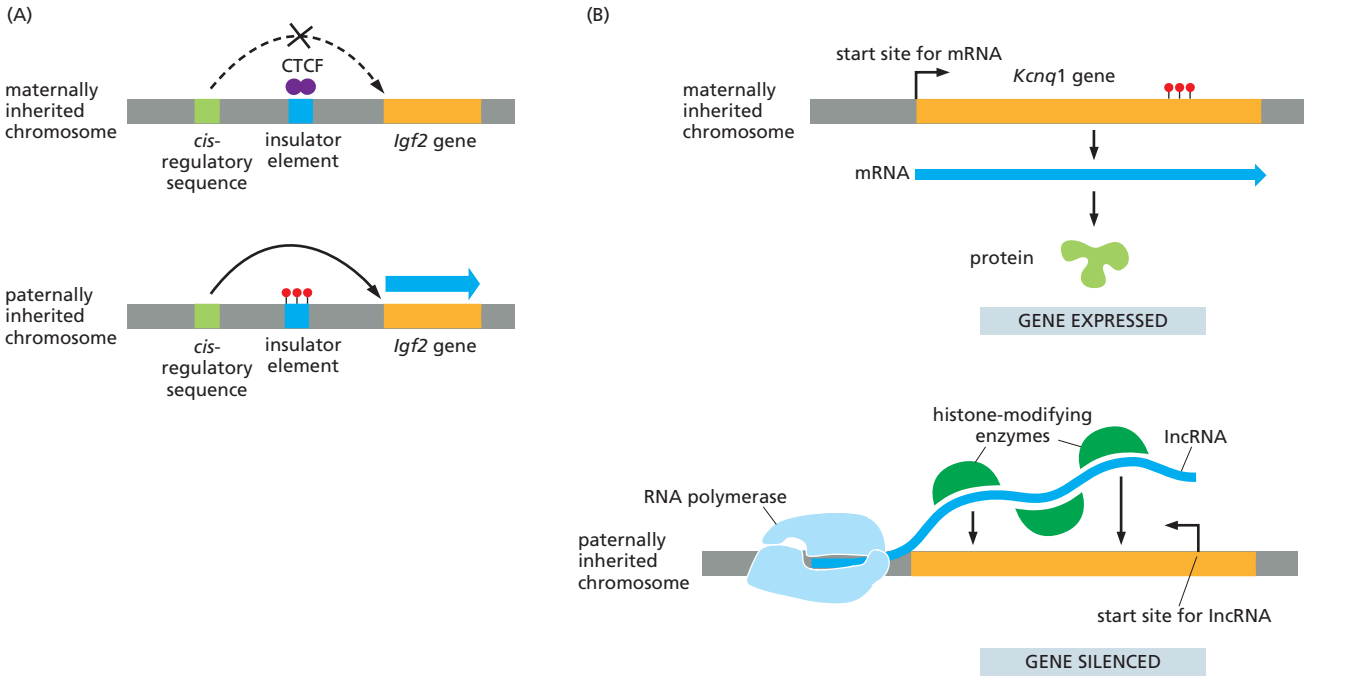


Figure 7–51 How imprinting can cause a non-Mendelian pattern of inheritance. The *top* portion of the figure shows a pair of homologous chromosomes in the somatic cells of two adult mice, one male and one female. In this example, both mice have inherited the top homolog from their father and the bottom homolog from their mother, and the paternal copy of a gene subject to imprinting (indicated in *orange*) is methylated, preventing its expression. The maternally derived copy of the same gene *(yellow)* is expressed. The remainder of the figure shows the outcome of a cross between these two mice. During germ-cell formation, but before meiosis, the imprints are erased and then, much later in germ-cell development, they are reimposed in a sex-specific pattern (*middle* portion of figure). In eggs produced from the female, neither allele of the *A* gene is methylated. In sperm from the male, both alleles of gene *A* are methylated. Shown at the *bottom* of the figure are two of the possible imprinting patterns inherited by the progeny mice; the mouse on the *left* has the same imprinting pattern as each of the parents, whereas the mouse on the *right* has the opposite pattern. If the two alleles of gene *A* are distinct (for example, if one codes for a mutant protein), the different imprinting patterns can cause phenotypic differences in the progeny mice, even though they carry exactly the same DNA sequences of the two *A* gene alleles.

Imprinting provides an important exception to classical "Mendelian" genetic behavior, and several hundred mouse genes are thought to be affected in this way. However, the majority of mouse genes are not imprinted, and therefore the rules of Mendelian inheritance apply to most of the mouse genome.

On the maternally derived chromosome, the insulator is not methylated, and the *Igf2* gene is therefore not transcribed (Figure 7–52A).

Other cases of imprinting are also based on DNA methylation, but they employ different "downstream" mechanisms. Some involve *long noncoding RNAs* (*lncRNAs*), which are defined as RNA molecules more than 200 nucleotides in length that do not code for proteins. We discuss lncRNAs broadly at the end of this chapter; here, we focus on the role of a specific lncRNA in imprinting. In the case



of the *Kcnq1* gene, which codes for a voltage-gated calcium channel needed for proper heart function, the lncRNA is made only from the paternal allele (which is unmethylated), and it is not released by the RNA polymerase, remaining instead at its site of synthesis on the DNA template. This RNA in turn recruits the histone-modifying and DNA-methylating enzymes that direct the formation of repressive chromatin, which silences the protein-coding gene associated on the paternally derived chromosome (**Figure 7–52B**). The maternally derived gene, on the other hand, is immune to these effects because its imprinted methylation blocks the synthesis of the lncRNA but allows transcription of the adjacent protein-coding gene. Thus, like *Igf2*, the specificity of *Kcnq1* imprinting arises from an inherited methylation pattern; the difference lies in the way these patterns cause the differential gene expression.

Why imprinting should exist at all is a mystery. In vertebrates, it is restricted to mammals that develop within the mother, and many of the imprinted genes are involved in fetal development. One idea is that imprinting reflects a middle ground in the evolutionary struggle between males to produce larger offspring and females to limit offspring size by "halving" the dosage of certain gene products that might accelerate growth. Whatever its purpose might be, imprinting provides startling evidence that features of DNA other than its sequence of nucleotides can be inherited.

A Chromosome-wide Alteration in Chromatin Structure Can Be Inherited

We have seen that DNA methylation and certain types of chromatin structure can be heritable, preserving patterns of gene expression across cell generations. Perhaps the most striking example of this effect occurs in mammals, in which an alteration in the chromatin structure of an entire chromosome can modulate the levels of expression of most genes on that chromosome.

Males and females differ in their *sex chromosomes*. Females have two X chromosomes, whereas males have one X and one Y chromosome. In humans, the X and Y sex chromosomes differ radically in gene content: the X chromosome is three times larger and contains about 900 protein-coding genes compared to the Y chromosome's 55 protein-coding genes. Mammals have evolved a *dosage compensation* mechanism to ensure that the same amount of most of the

Figure 7–52 Some mechanisms of imprinting. (A) On chromosomes inherited from the female, the CTCF protein binds to an insulator (see Figure 7–28), blocking communication between *cis*-regulatory sequences *(green)* and the *Igf2* gene *(orange)*. *Igf2* is therefore not expressed from the maternally inherited chromosome. Because of imprinting, the insulator on the male-derived chromosome is methylated *(red circles)*; this inactivates the insulator by blocking the binding of the CTCF protein and allows the *cis*-regulatory sequences to activate transcription of the *Igf2* gene. In other examples of imprinting, methylation simply blocks gene expression by interfering with the binding of proteins required for a gene's transcription. (B) Imprinting of the mouse *Kcnq1* gene. On the maternally derived chromosome, synthesis of the lncRNA is blocked by methylation of the DNA *(red circles)*, and the *Kcnq1* gene is expressed. On the paternally derived chromosome, the lncRNA is synthesized, remains in place, and by directing alterations in chromatin structure blocks expression of the *Kcnq1* gene. Although shown as directly binding to lncRNA, the histone-modifying enzymes are likely to be recruited indirectly, through additional proteins.

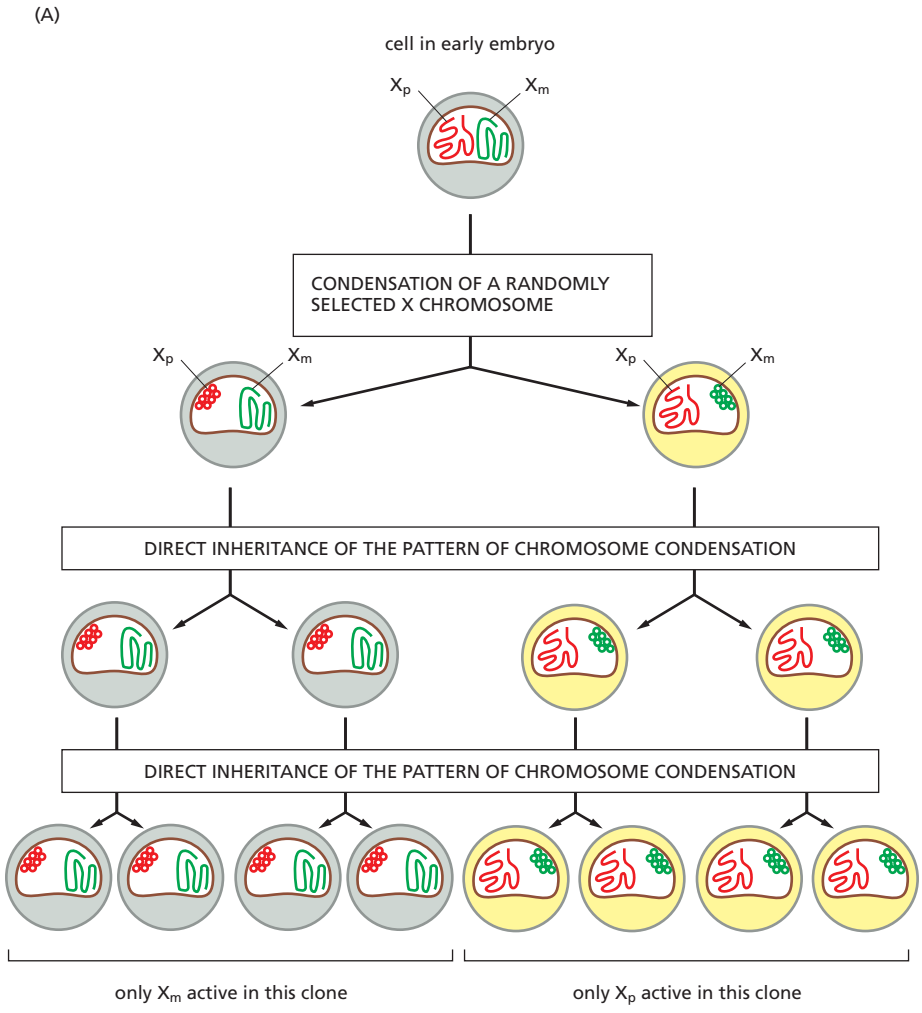


Figure 7–53 X-inactivation. (A) The clonal inheritance in female mammals of a condensed, inactive X chromosome. (B) A calico cat, whose patches of color reflect the random nature of the X-inactivation process. (B, bluecaterpillar/Deposit photos.)

X-chromosome gene products is made in both male and female cells, despite the fact that females contain twice as many X-chromosome genes. Mutations that interfere with this dosage compensation are generally lethal.

Mammals achieve dosage compensation by the transcriptional inactivation of one of the two X chromosomes in female somatic cells, a process known as **X-inactivation**. As a result of X-inactivation, two X chromosomes can coexist within the same nucleus, be exposed to the same diffusible transcription regulators, and yet differ entirely in their expression.

Early in the development of a female embryo, when it consists of a few hundred cells, one of the two X chromosomes in each cell becomes highly condensed into a type of heterochromatin. In placental mammals, the initial choice of which X chromosome to inactivate—the maternally inherited one (X_m) or the paternally inherited one (X_p)—appears to be random. And once either X_p or X_m has been inactivated, it remains silent throughout all subsequent cell divisions of that cell and its progeny, indicating that the inactive state is faithfully maintained through many cycles of DNA replication and mitosis. Because X-inactivation is random and takes place after several hundred cells have already formed in the embryo, every female is a mosaic of clonal groups of cells in which either X_p or X_m is silenced (**Figure 7–53**), distributed in small clusters in the adult animal because sister cells tend to remain close together during later stages of development (**Figure 7–54**).

X-inactivation creates the orange and black coat coloration of some female cats (see Figure 7–53B). In these "calico" cats, one X chromosome carries a gene that produces orange hair color, and the other X chromosome carries an allele

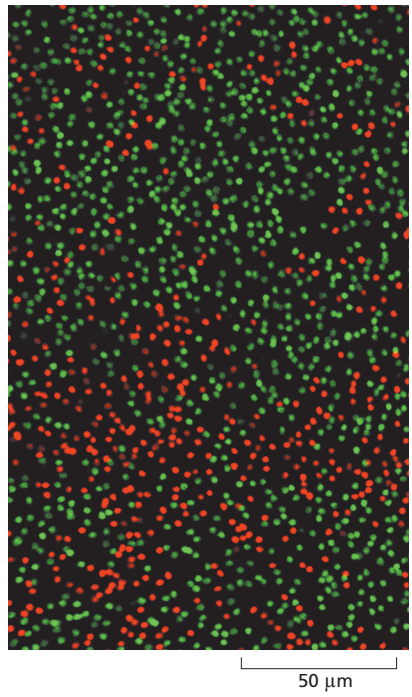


Figure 7–54 Photoreceptor cells in the retina of a female mouse showing patterns of X-inactivation. Using genetic engineering techniques (described in Chapter 8), the germ line of a mouse was modified so that one copy of the X chromosome (if active) makes a green fluorescent protein and the other (if active) a red fluorescent protein. Both proteins concentrate in the nucleus, and, in the field of cells shown here, it is clear that only one of the two X chromosomes is active in each cell. If both chromosomes were active, the nuclei would fluoresce both red and green, and therefore appear yellow. (From H. Wu et al., *Neuron* 81:103–119, 2014. With permission from Elsevier.)

of the same gene that results in black hair color; it is the random X-inactivation that produces patches of cells of two distinctive colors. In contrast, male cats of this genetic stock are either solid orange or solid black, depending on which X chromosome they have inherited from their mothers. Although X-inactivation is maintained over thousands of cell divisions, it is reversed during germ-cell formation, so that all the haploid oocytes contain an active X chromosome and can express X-linked gene products.

The Mammalian X-Inactivation in Females Is Triggered by the Synthesis of a Long Noncoding RNA

How is an entire chromosome transcriptionally inactivated? In humans, the chromosome-wide inactivation process begins with the synthesis of a long noncoding RNA, called **Xist**, whose gene lies on the X chromosome. This transcript (about 20,000 nucleotides in length) is synthesized by only one of the two X chromosomes in females, and exactly how this seemingly random choice is made remains to discovered. Once an Xist RNA molecule is synthesized, it does not leave the X chromosome from which it was made; rather, it diffuses along only that chromosome. Ultimately, about 2000 molecules of Xist are synthesized per X chromosome, and they eventually coat the chromosome that produces it. The spread of Xist across the chromosome does not itself cause transcriptional silencing; this long RNA contains binding sites for many different proteins that carry out the actual gene silencing. These include DNA methylases, histone-modifying enzymes, and structural components specific to the inactive X chromatin. As a result, extensive methylation of the inactive X occurs (including at CG islands), and the chromosome is folded into compact structures that are generally resistant to transcription (**Figure 7–55**). These multiple layers, each of which can be self-propagating (see Figure 7–48), ensure that the randomly chosen X chromosome remains inactive through multiple cell divisions.

Not every gene on the inactive X chromosome is transcriptionally silenced. Of the approximately 900 protein-coding genes on the human X chromosome, 15–20% remain actively expressed after the chromosome-wide inactivation process has been completed. And for many of these genes, both copies—one from the active X and one from the inactive X—must be expressed to obtain sufficient levels of their gene products for proper development to occur.

How do select genes escape silencing after the majority of the X chromosome is rendered transcriptionally inactive? As we saw in Chapter 4 and earlier in this chapter, transcriptionally active genes generally occur in DNA loops that are held in place by insulator proteins such as CTCF (see Figure 7–28), and this is the case for the "escapees" of the inactive X chromosome. These loops are believed to extend from the bulk of the tightly packaged chromosome. In contrast, most of the inactive genes lie in the interior of the inactive X chromosome, which is depleted for CTCF. It has been proposed that X-inactivation is accompanied by the formation of a specialized biomolecular condensate, where the proteins and RNAs needed for gene repression are kept at high local concentrations; according to this model, the loops of active genes would extend outward, beyond the boundary of the condensate.

We have described the way that placental mammals deal with dosage compensation on the X chromosome, but the details of this process differ from those in most other animals in important ways. For example, in marsupials, the choice of which X chromosome to inactivate is not random; instead, the X chromosome inherited from the father is automatically silenced. And in flies, dosage

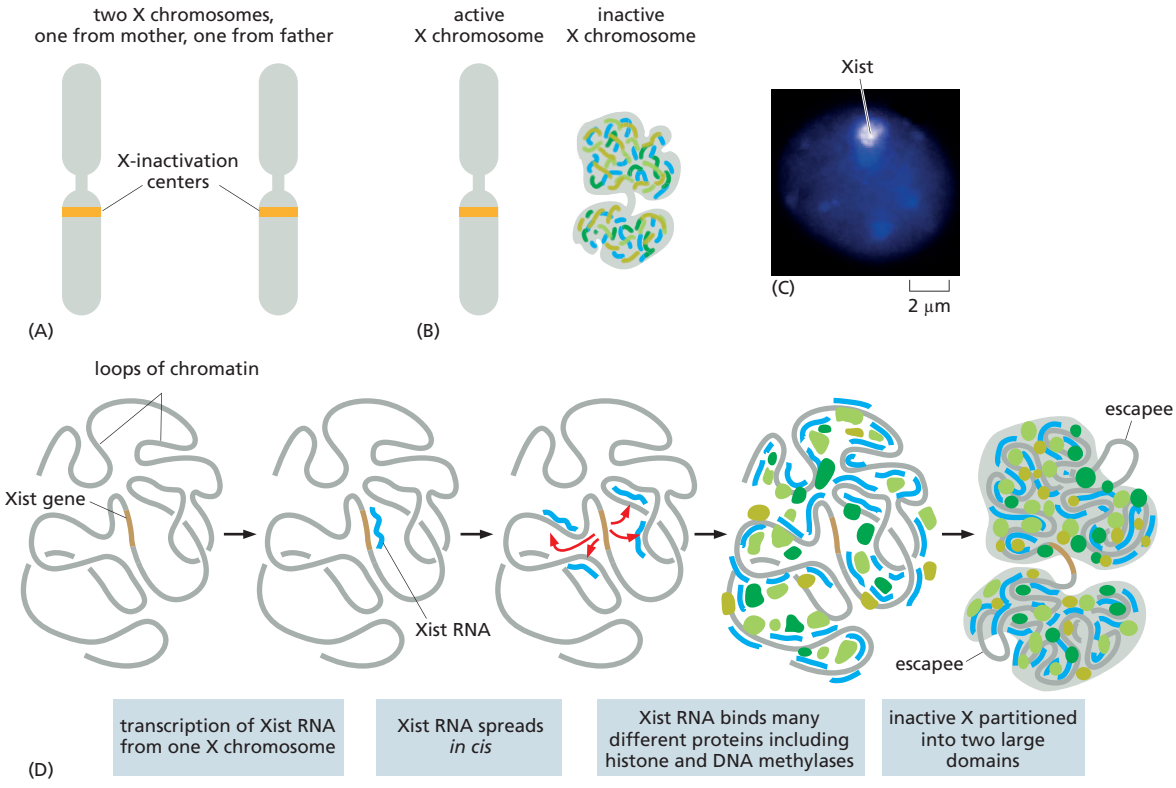


Figure 7–55 Mammalian X-inactivation. The two X chromosomes in a female mammal (A) before and (B) after X-inactivation. (C) At an early stage of X-chromosome inactivation, mouse chromosomes have been hybridized with a fluorescent probe that is complementary to the Xist RNA, which coats only the inactive X chromosome; the remaining DNA has been stained blue with a dye. (D) A schematic illustration of how the continuing synthesis of Xist RNA at the Xist locus moves Xist molecules outward, across the chromosome. As Xist molecules coat the chromosome, they begin to associate with a variety of structural proteins and enzymes that modify histones and DNA. [Although some of these proteins are bound to the chromosome prior to Xist spreading (not shown), most are brought in by direct association with Xist.] The two major chromosome domains that are created at the completion of the inactivation process have been proposed to be biomolecular condensates. Genes that escape the inactivation process are shown as loops, extending from the compact domains. (B and D, based on a figure supplied by Agnese Loda and Edith Heard; C, from L. Giorgetti et al., *Nature* 535:575–579, 2016. Reproduced with permission from SNCSC.)

compensation takes place in the male, where the single X chromosome is up-regulated approximately twofold to match the female dose. Finally, in nematode worms, the hermaphrodites reduce gene expression by roughly half on both X chromosomes to match the single X-chromosome dosage in males.

The fundamentally different mechanisms of dosage compensation among animals suggest that it has been a relatively recent evolutionary innovation. We have some clues for its origin in humans. Some of the key components in X-inactivation also function to repress the many transposons in the human genome, a process we discuss later in the chapter. It has been proposed that Xist evolved from multiple transposons that inserted into our X chromosome, eventually "tricking" the cell into inactivating the whole chromosome.

Stable Patterns of Gene Expression Can Be Transmitted to Daughter Cells

Imprinting and X-inactivation are examples of **monoallelic gene expression**, where only one of the two copies of a gene is expressed in a diploid genome. In addition to the silenced genes on the X chromosome and the 300 or so genes that are imprinted, there are another 1000–2000 human genes that exhibit monoallelic expression. Like X-inactivation (but unlike imprinting), the choice of which copy of the gene is expressed and which is silenced appears random. Yet once the choice is made, it can persist for many cell divisions. Because the choice is often made relatively late in development, cells of the same tissue in the same individual can express different copies of a given gene. In other words, somatic tissues are often mosaics, where different clones of cells have subtly different patterns of gene expression. The mechanisms responsible for this type of monoallelic expression and its memory through cell divisions are not known in detail, and its general purpose—if any—is poorly understood. However, several different mechanisms are known that may contribute to such inheritance, as we now discuss.

In considering the general question of cell memory, it is useful to return to our discussion of the different cell types in an organism. As we have seen, once a cell in an organism differentiates into a particular cell type, it generally remains

specialized in that way; if it divides, its daughters inherit the same specialized character. Perhaps the simplest way for a cell to remember its identity is through a positive feedback loop in which a key transcription regulator activates, either directly or indirectly, the transcription of its own gene (see Figure 7–42). As we discussed earlier in this chapter, interlocking positive feedback loops of the type shown in Figure 7–40B provide greater stability by buffering the circuit against fluctuations in the level of any one transcription regulator. Because transcription regulators are synthesized in the cytosol and diffuse throughout the nucleus, feedback loops based on this mechanism will affect both copies of a gene in a diploid cell. However, as discussed earlier, the expression pattern of a gene on one chromosome can differ from that of the copy of the same gene on the other chromosome (as in X-inactivation or in imprinting). Such differences can also be inherited through many cell divisions, and they cannot be explained by this type of transcription feedback loop.

The ability of a daughter cell to retain a memory of the gene expression patterns that were present in the parent cell is an example of **epigenetic inheritance**, which we define as a heritable alteration in a cell or organism's phenotype that does not result from changes in the nucleotide sequence of DNA. In **Figure 7–56**, we illustrate four mechanisms that can produce epigenetic inheritance, contrasting those self-propagating mechanisms that work *in cis*, affecting only one chromosomal copy, with self-propagating mechanisms that work *in trans*, affecting both chromosomal copies of a gene.

It is important to note that many of the changes in gene expression that occur in cells are transient and depend on the continued presence of a signal that is

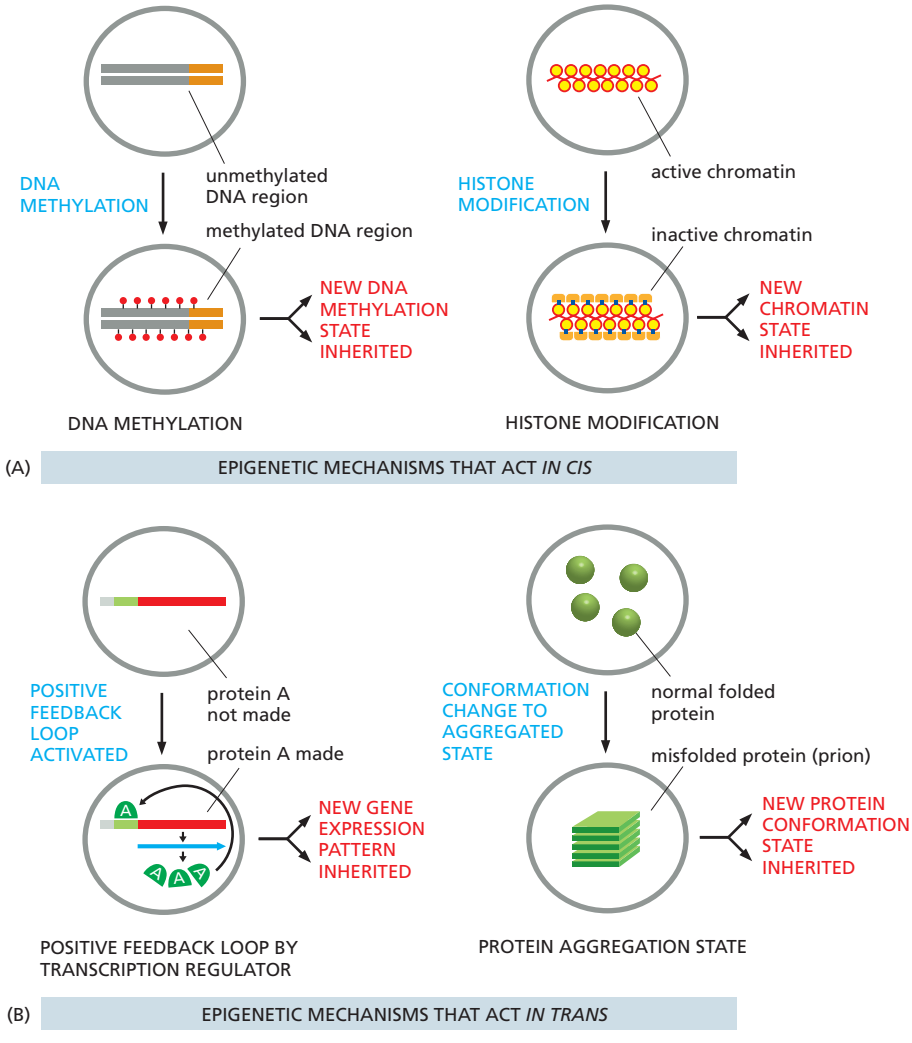


Figure 7–56 Four distinct mechanisms that can produce an epigenetic form of inheritance in an organism. (A) Two epigenetic mechanisms that act *in cis*. As discussed in this chapter, a maintenance methylase can propagate specific patterns of cytosine methylation (see Figure 7–47). Alternatively, as discussed in Chapter 4, a histone-modifying enzyme that replicates the same covalent modification that attracts it to chromatin can result in a chromatin structure being self-propagating (see Figure 4–44). Note that the term epigenetic is sometimes misused to refer to all covalent modifications of histones, whether or not they are self-propagating. But many histone modifications are erased each time a cell divides, and they therefore do not fit our definition. (B) Two epigenetic mechanisms that act *in trans*. Positive feedback loops formed by transcription regulators are found in all species and are probably the most common form of cell memory. As discussed in Chapter 3, some proteins can form self-propagating prions (see Figure 3–33). When these proteins are involved in gene expression, prions can transmit a particular pattern of gene expression to daughter cells.

external to the cell. When the signal disappears, so does the new gene expression pattern; in other words, the pattern is not directly heritable (see Chapter 15). Gene expression changes of both types—both heritable and non-heritable—are crucial for the function of all cells on earth. And the discovery more than 60 years ago that gene expression can be regulated by cells ranks as one of the fundamental principles of biology.

Summary

In addition to the positive feedback loops created by transcription regulators, eukaryotic cells can use both inherited forms of DNA methylation and inherited states of chromatin condensation as mechanisms for generating a cell memory of gene expression patterns. An especially dramatic case that involves chromatin condensation is the inactivation of an entire X chromosome in female mammals. DNA methylation underlies the phenomenon in mammals of genomic imprinting, in which the expression of a gene depends on whether that gene was inherited from the mother or the father. All of these mechanisms allow cells to pass on gene expression patterns to their progeny cells, contributing to the epigenetic inheritance that makes complex multicellular life possible.

POST-TRANSCRIPTIONAL CONTROLS

In principle, every step required for the process of gene expression can be controlled. Indeed, one can find examples of each type of regulation, and many genes are known to be regulated by multiple mechanisms. As we have seen, controls on the initiation of gene transcription are one critical form of regulation for all genes. But other, equally important controls often act later in the pathway from DNA to protein to change the amount of gene product that is made—and in some cases, even to alter the amino acid sequence of a protein product. These **post-transcriptional controls**, which operate after RNA polymerase has bound to the gene's promoter and has begun its RNA synthesis, are crucial for the regulation of many genes.

In the following sections, we consider the varieties of post-transcriptional regulation in a temporal order, following the sequence of events that an RNA molecule might experience after its transcription has begun (**Figure 7–57**).

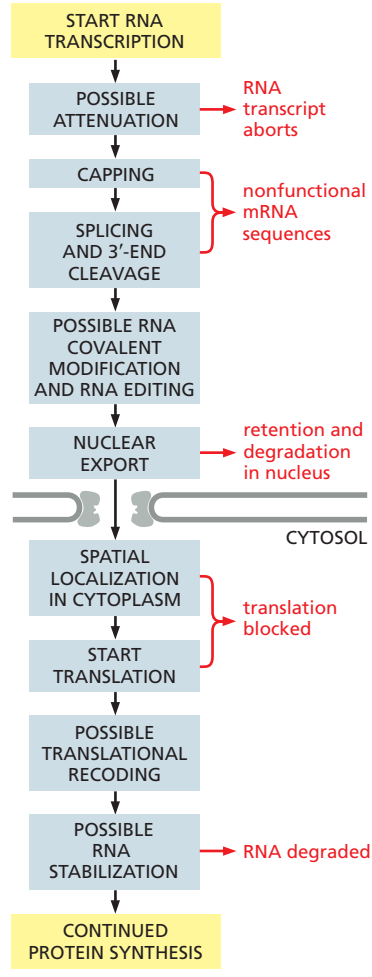
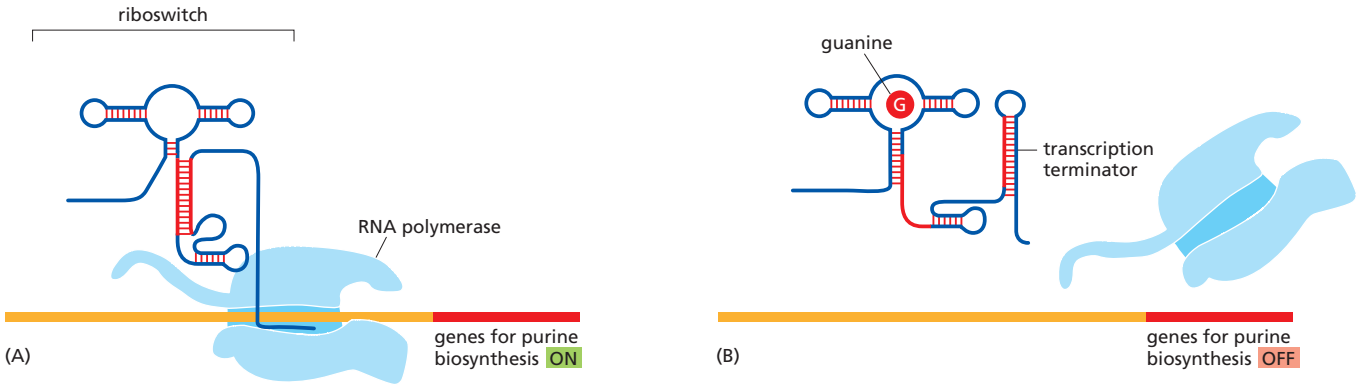


Figure 7–57 Post-transcriptional controls of gene expression. The final synthesis rate of a protein can, in principle, be controlled at any of the steps listed in capital letters, although only a few of the steps depicted here are likely to be critical for the regulation of any one particular protein. As we shall discuss, the 3′ end cleavage, splicing, editing, and translation recoding steps also make it possible for the cell to produce more than one protein variant from the same gene.

Transcription Attenuation Causes the Premature Termination of Some RNA Molecules

It has long been known that the expression of some genes is inhibited by premature termination of transcription, a phenomenon called *transcription attenuation*. In some of these cases, the nascent RNA chain adopts a structure that causes it to interact with the RNA polymerase in such a way as to abort its transcription. When the gene product is required, regulatory proteins bind to the nascent RNA chain to remove the attenuation, allowing the transcription of a complete RNA molecule.

A well-studied example of transcription attenuation occurs during the life cycle of HIV, the human immunodeficiency virus that is the causative agent of acquired immune deficiency syndrome, or AIDS. Once the HIV genome has been integrated into the host genome, the viral DNA is transcribed by the cell's RNA polymerase II (see Figure 5–61). However, this polymerase usually terminates transcription after synthesizing transcripts of several hundred nucleotides and thus fails to efficiently transcribe the entire viral genome. But when conditions for viral growth are optimal, a virus-encoded protein called Tat, which binds to a specific stem–loop structure in the nascent RNA that contains a "bulged base," prevents this premature termination (see Figure 6–92). Once bound to this specific RNA structure (called TAR), Tat assembles several host-cell proteins that allow the RNA polymerase to continue transcribing. The normal role of at least some of these proteins is to prevent pausing and premature termination by RNA polymerase when it transcribes normal cell genes. A normal cell mechanism has



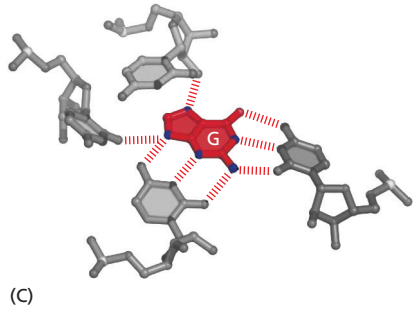


Figure 7–58 A riboswitch that responds to guanine. In this example from bacteria, the riboswitch controls expression of the purine biosynthetic genes. (A) When guanine levels in cells are low, an elongating RNA polymerase transcribes the purine biosynthetic genes, and the enzymes needed for guanine synthesis are therefore expressed. (B) When guanine is abundant, it binds the riboswitch, causing it to undergo a conformational change that forces the RNA polymerase to terminate transcription (see Figure 6–11). (C) Guanine *(red)* bound to the riboswitch. Only those nucleotides that form the guanine-binding pocket are shown. Many other riboswitches exist, including those that recognize *S*-adenosylmethionine, coenzyme B_{12}, flavin mononucleotide, adenine, lysine, and glycine. (A and B, adapted from M. Mandal and R.R. Breaker, *Nat. Rev. Mol. Cell Biol.* 5:451–463, 2004; and C.K. Vanderpool and S. Gottesman, *Mol. Microbiol.* 54:1076–1089, 2004; C, adapted from A. Serganov et al., *Chem. Biol*. 11:1729–1741, 2004. PDB code: 1Y27.)

apparently been highjacked by HIV to permit transcription of its genome to be controlled by a single viral protein.

Riboswitches Probably Represent Ancient Forms of Gene Control

In Chapter 6, we discussed the idea that, before modern cells arose on Earth, RNA played the role of both DNA and proteins, storing hereditary information and catalyzing chemical reactions (see pp. 389–393). The discovery of *riboswitches* shows that RNA can also form control devices. Riboswitches are short sequences of RNA that change their conformation when they bind a specific small molecule, such as a metabolite. Riboswitches are often located near the 5′ end of mRNAs, and they fold while the mRNA is being synthesized, blocking or permitting progress of the RNA polymerase according to whether the regulatory small molecule is bound (**Figure 7–58**).

Riboswitches are particularly common in bacteria, where they sense key small metabolites in the cell and adjust gene expression accordingly. Each recognizes only the appropriate small molecule with high specificity. In many cases, every chemical feature of the small molecule is read by the RNA, and the binding affinities observed are as tight as those typically observed between small molecules and proteins (see Figure 7–58C).

Riboswitches are perhaps the most economical examples of gene control devices, inasmuch as they completely bypass the need for regulatory proteins. In the example illustrated (see Figure 7–58), the riboswitch controls transcription elongation, but riboswitches can also regulate other steps in gene expression, as we shall see later in this chapter. The fact that highly sophisticated gene control devices can be made from short sequences of RNA provides important support for the early "RNA world" hypothesis.

Alternative RNA Splicing Can Produce Different Forms of a Protein from the Same Gene

As discussed in Chapter 6, RNA splicing shortens the transcripts of many eukaryotic genes by removing the intron sequences from mRNA precursors. A cell can splice an RNA transcript differently and thereby make different

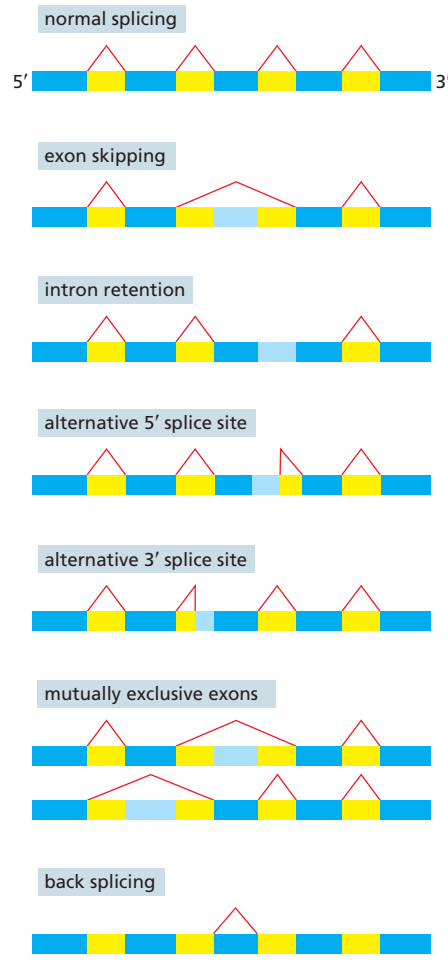


Figure 7–59 Patterns of alternative RNA splicing. *Dark blue boxes* mark exon sequences that are retained in spliced mRNAs. *Light blue boxes* mark possible exon sequences that are included only in the indicated mRNAs. The boxes are joined by *red lines* to indicate where intron sequences *(yellow)* are removed. In back splicing (discussed later in the chapter), a single exon is removed as a circular RNA molecule. (Adapted from H. Keren et al., *Nat. Rev. Genet*. 11:345–355, 2010.)

polypeptide chains from the same gene—a process called **alternative RNA splicing** (Figure 7–59; see also Figure 6–27). Many animal genes produce multiple proteins in this way.

When different splicing possibilities exist at several positions in the transcript, a single gene can produce dozens of different proteins. In one extreme case, a single *Drosophila* gene can, in principle, produce as many as 38,000 different proteins through alternative splicing (Figure 7–60), although only a fraction of these forms have thus far been experimentally observed. Considering that the *Drosophila* genome has approximately 14,000 protein-coding genes, it is clear that the protein complexity of an organism can greatly exceed the number of its genes. This example also illustrates the perils in equating gene number with an organism's complexity. For example, alternative splicing is rare in single-celled budding yeasts but very common in flies. Budding yeast has about 6200 genes, only about 300 of which are subject to splicing, and nearly all of these have only a single intron. The fact that flies have only 2–3 times as many genes as yeasts greatly underestimates the difference in complexity of these two genomes.

In some cases, alternative RNA splicing occurs because there is an *intron sequence ambiguity*: the standard spliceosome mechanism for removing intron sequences (discussed in Chapter 6) is unable to distinguish clearly between two or more alternative pairings of 5′ and 3′ splice sites, so that different choices are made by chance on different individual transcripts. Where such *constitutive* alternative splicing occurs, several versions of the protein encoded by the gene are made in all cells in which the gene is expressed.

In many cases, however, alternative RNA splicing is regulated. In the simplest examples, regulated splicing is used to switch from the production of a

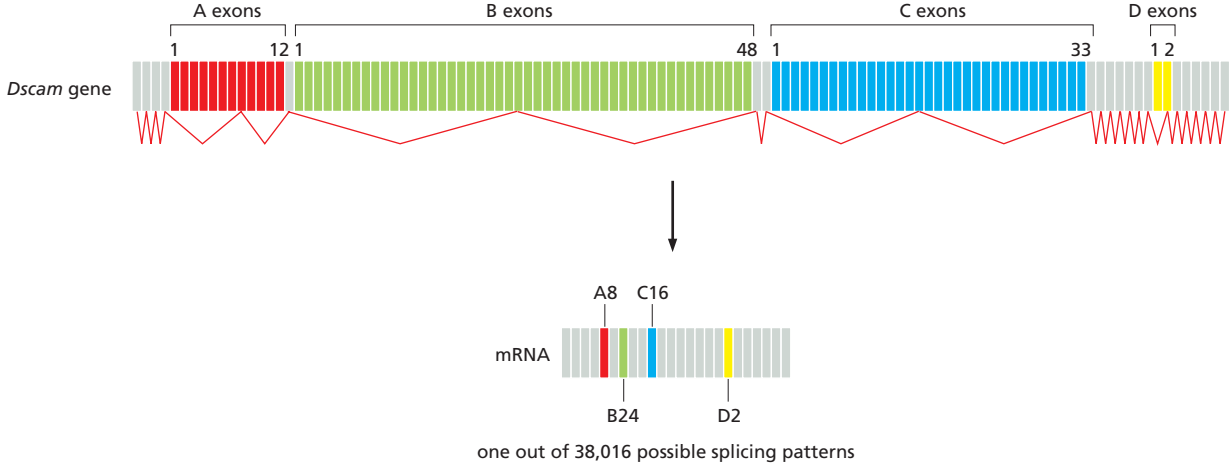


Figure 7–60 Alternative splicing of RNA transcripts of the *Drosophila Dscam* gene. Dscam proteins have several different functions. In cells of the fly immune system, they mediate the phagocytosis of bacterial pathogens. In cells of the nervous system, they are needed for proper wiring of neurons. Each mature mRNA contains 24 exons, four of which (denoted A, B, C, and D) are present in the *Dscam* gene as arrays of alternative exons. Each RNA contains 1 of 12 alternatives for exon A *(red)*, 1 of 48 alternatives for exon B *(green)*, 1 of 33 alternatives for exon C *(blue)*, and 1 of 2 alternatives for exon D *(yellow)*. This figure shows only one of the many possible splicing patterns (indicated by the *red line* and by the mature mRNA below it). Each variant Dscam protein folds into roughly the same structure (predominantly a series of extracellular immunoglobulin-like domains linked to a membrane-spanning region; see Figure 24–48), but the amino acid sequences of the domains vary according to the splicing pattern. The diversity of Dscam variants contributes to the plasticity of the immune system, as well as to the formation of complex neural circuits. (Adapted from D.L. Black, *Cell* 103:367–370, 2000. With permission from Elsevier.)

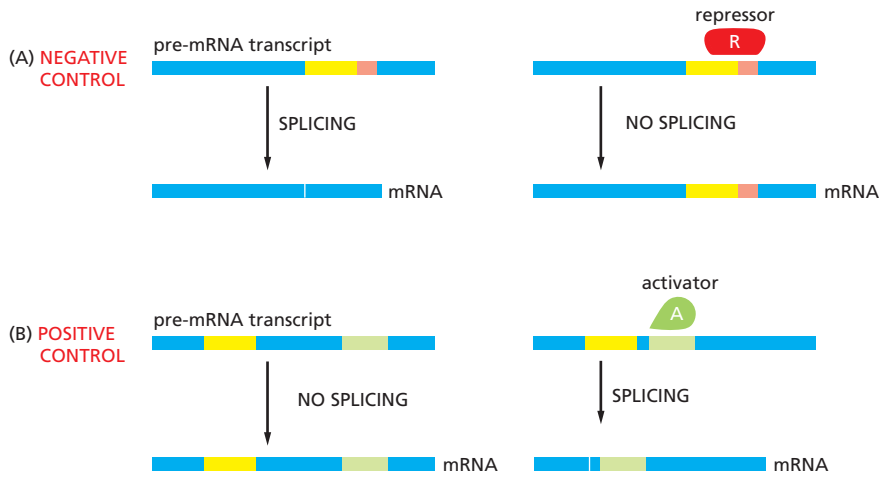


Figure 7–61 Negative and positive control of alternative RNA splicing. (A) In negative control, a repressor protein binds to a specific sequence in the pre-mRNA transcript and blocks access of the splicing machinery to a splice junction. This often results in the use of a secondary splice site, thereby producing an altered pattern of splicing (see Figure 7–59). (B) In positive control, the splicing machinery is unable to remove a particular intron sequence efficiently without assistance from an activator protein. Because an RNA molecule is flexible, the nucleotide sequences that bind these activators can be located many nucleotide pairs from the splice junctions they control, and they are often called *splicing enhancers*, by analogy with the transcription enhancers mentioned earlier in this chapter.

nonfunctional protein to the production of a functional one (or the other way around). The transposase that catalyzes the transposition of the *Drosophila* P element, for example, is produced in a functional form in germ cells and a nonfunctional form in somatic cells of the fly, allowing the P element to spread throughout the genome of the fly without causing damage in somatic cells (see Figure 5–59 and Table 5–4, p. 308). This difference in transposon activity has been traced to the presence of an intron sequence in the transposase RNA that is removed only in germ cells.

In addition to enabling switching from the production of a functional protein to the production of a nonfunctional one (or vice versa), the regulation of RNA splicing can generate different versions of a protein in different cell types, according to the needs of the cell. Tropomyosin, for example, is produced in specialized forms in different types of cells (see Figure 6–27). Cell-type-specific forms of many other proteins are produced in the same way.

RNA splicing can be regulated either negatively, by a regulatory molecule that prevents the splicing machinery from gaining access to a particular splice site on the RNA, or positively, by a regulatory molecule that helps direct the splicing machinery to an otherwise overlooked splice site (**Figure 7–61**).

Because of the plasticity of RNA splicing, the blocking of a "strong" splicing site will often expose a "weak" site and result in a different pattern of splicing. Thus, the splicing of a pre-mRNA molecule can be thought of as a delicate balance between competing splice sites—a balance that can easily be tipped by the effects on splicing of RNA-bound regulatory proteins.

The Definition of a Gene Has Been Modified Since the Discovery of Alternative RNA Splicing

The discovery that eukaryotic genes usually contain introns and that their coding sequences can be assembled in more than one way raised new questions about the definition of a gene. A gene was first clearly defined in molecular terms in the early 1940s from work on the biochemical genetics of the fungus *Neurospora*. Until then, a gene had been defined as a region of the genome that segregates as a single unit during meiosis and gives rise to a definable phenotypic trait—such as a red or a white eye in *Drosophila* or a round or wrinkled seed in peas. The *Neurospora* findings revealed that most genes correspond to a region of the genome that directs the synthesis of a single enzyme, leading to the view that each gene encodes one polypeptide chain. As more was learned about the mechanism of gene expression in the 1960s, a gene became identified as that stretch of DNA that was transcribed into the RNA coding for either a single polypeptide chain or a single structural RNA such as a tRNA or an rRNA molecule. The discovery of introns in the late 1970s could be readily accommodated by

the original definition of a gene, provided that a single polypeptide chain was specified by the RNA transcribed from any one DNA sequence. But now that it is clear that many DNA sequences in eukaryotic cells can produce a set of distinct (but related) proteins by means of alternative RNA splicing, how should a gene be defined?

It is relatively rare that a single transcription unit produces two very different eukaryotic proteins, and in those cases, the two proteins are considered to be produced by distinct genes that overlap on the chromosome. It seems unnecessarily complex, however, to consider most of the protein variants produced by alternative RNA splicing as being derived from overlapping genes. A more sensible alternative is to modify the original definition to consider any DNA sequence that is transcribed as a single unit and encodes a set of closely related polypeptide chains (protein isoforms) as a single protein-coding gene. This definition of a **gene** also accommodates those DNA sequences that encode protein variants produced by post-transcriptional processes other than RNA splicing, such as the transcript cleavage and RNA editing discussed shortly.

Back Splicing Can Produce Circular RNA Molecules

We have seen that pre-mRNA splicing is remarkably plastic, and recent discoveries have revealed a new surprise. Some pre-mRNAs undergo what is termed "back splicing" where a 3′ splice site is joined to a downstream 5′ splice site, thereby reversing the normal joining order (see Figure 7–59). This process typically releases a single exon sequence as a covalently closed, circular RNA molecule. These unusual RNAs are exported from the nucleus but are rarely translated into protein. Instead, they have been proposed to "soak up" complementary RNAs as well as RNA-binding proteins and to provide scaffolds for multisubunit RNA–protein complexes. Because they lack free ends, which are the normal substrates for RNA-degrading enzymes, these circular RNAs are much more stable than typical mRNAs. Although usually made in small amounts, their stability can allow them to accumulate to high concentrations in cells, and several specific circular RNAs are especially prominent in cells of the mammalian brain and immune systems. Although we still have much to learn about these peculiar RNAs, they attest to the many surprises that RNA biology has in store for us. We shall revisit this general issue at the end of the chapter, when we discuss the diversity of noncoding RNAs.

A Change in the Site of RNA Transcript Cleavage and Poly-A Addition Can Change the C-terminus of a Protein

We saw in Chapter 6 that the 3′ end of a eukaryotic mRNA molecule is not formed by the termination of RNA synthesis by the RNA polymerase, as it is in bacteria. Instead, it results from an RNA cleavage reaction that is catalyzed by additional proteins while the transcript is elongating (see Figure 6–36). A cell can control the site of this cleavage so as to change the C-terminus of the resultant protein. In the simplest cases *of alternate cleavage and polyadenylation*, one protein variant is simply a truncated version of the other; in many other cases, however, the alternative cleavage and polyadenylation sites lie within intron sequences, and the pattern of splicing is thereby altered. This process can produce two closely related proteins that differ only in the amino acid sequences at their C-terminal ends. An analysis of RNAs produced from the human genome in a variety of cell types indicates that as many as half of all human protein-coding genes produce mRNA species with more than one site of polyadenylation.

A well-studied example of regulated polyadenylation is the switch from the synthesis of membrane-bound to secreted antibody molecules that occurs during the development of B lymphocytes (see Figure 24–22). Early in the life history of a B lymphocyte, the antibody it produces is anchored in the plasma membrane, where it serves as a receptor for antigen. Antigen stimulation causes B lymphocytes to multiply and to begin secreting their antibody. The secreted form of the

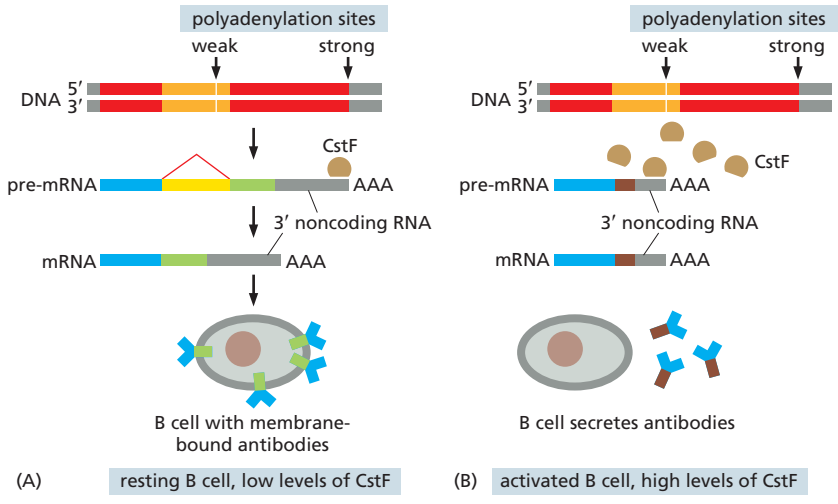


Figure 7–62 Regulation of the site of RNA cleavage and poly-A addition determines whether an antibody molecule is secreted or remains membrane-bound. (A) In unstimulated B lymphocytes, a long RNA transcript is produced, and the intron sequence *(yellow)* near its 3′ end is removed by RNA splicing to provide an mRNA molecule that codes for a membrane-bound antibody molecule. Only a portion of the antibody gene is shown in the figure; the actual gene and its mRNA would extend further to the left of the diagram. (B) After antigen stimulation, the RNA transcript is cleaved and polyadenylated upstream from the intron's 3′ splice site. As a result, some of the intron sequence remains as a coding sequence in the short transcript, specifying the hydrophilic C-terminal portion of the secreted antibody molecule *(brown)*. (Adapted from D. Di Giammartino et al., *Mol. Cell* 43:853–866, 2011.)

antibody is identical to the membrane-bound form except at the extreme C-terminus. In this part of the protein, the membrane-bound form has a long string of hydrophobic amino acids that traverses the lipid bilayer of the membrane, whereas the secreted form has a much shorter string of hydrophilic amino acids. The switch from membrane-bound to secreted antibody is generated through a change in the site of RNA cleavage and polyadenylation, as shown in **Figure 7–62**.

The change is caused by an increased concentration of a subunit of a protein (CstF) that promotes RNA cleavage (see Figure 6–36). The first cleavage/poly-A addition site that a transcribing RNA polymerase encounters is suboptimal and is usually skipped in unstimulated B lymphocytes, leading to production of the longer RNA transcript. But when activated to produce antibodies, the B lymphocyte produces more CstF; as a result, cleavage now occurs at the suboptimal site, and the shorter transcript is produced. In this way, a change in concentration of a general RNA-processing factor can have a dramatic effect on the expression of a specific gene.

Nucleotides in mRNA Can Be Covalently Modified

In the previous chapter, we saw how specialized proteins modify the 5′ and 3′ ends of eukaryotic mRNAs and how a complex assembly of proteins and RNA molecules removes intron sequences. However, mRNA molecules are subject to more than 100 additional kinds of covalent changes, predominantly chemical modifications of individual bases, a few of which are shown in **Figure 7–63**. The reasons for most of these modifications of individual mRNAs remain a mystery. We do not know what they might do or even if they are biologically meaningful, inasmuch as many of them may simply represent "spillover" from the processes that modify the highly abundant tRNA and rRNA molecules (see Figures 6–43 and 6–57).

One of the most prominent and best understood mRNA modifications is the methylation of the amino group on adenine to produce N^6-methyladenosine

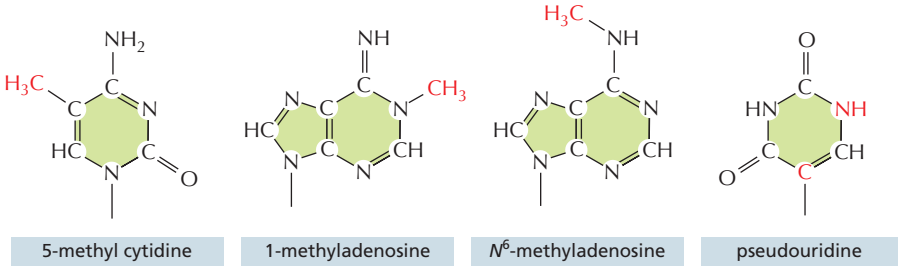


Figure 7–63 Four of the many types of covalent base modifications found in mRNA. Differences from the normal nucleosides are indicated in *red*. Each base is joined to a ribose sugar (not shown) by the indicated bond to form the nucleoside.

(see Figure 7–63). This addition is constantly being removed by protein complexes that contain demethylases, or "erasers," making the modification temporary. The methylases responsible for this modification typically act as the RNA is being transcribed; they recognize short sequences in the emerging RNA (often with the help of other proteins) and methylate the adenosines adjacent to these sequences.

In humans an average of 1–3 N^6-methyladenosine modifications occur on each mRNA molecule. What are their consequences? One effect is the destabilization of the hairpin helices that are formed by intramolecular base-pairing. This modification can thereby change the secondary structure of mRNA, which in some cases alters the splicing pattern of transcripts. In other cases, the modification promotes destruction of mRNAs through "reader" proteins that attract the RNA degradation machinery. A rapid destruction of certain mRNAs is especially important during cell differentiation, when the mRNAs produced earlier need to be cleared out. Finally, other specific N^6-methyladenosine modifications are known to promote translation of the modified mRNA. In this case, reader proteins that attract the translation machinery come into play. The many other mRNA modifications, some of which are shown in Figure 7–63, are more poorly understood, but some may likewise help to determine exactly how each mRNA is to be handled by the cell.

RNA Editing Can Change the Meaning of the RNA Message

The molecular mechanisms used by cells are providing scientists with a continual source of surprises. An example is a covalent modification of mRNA that alters its nucleotide sequence and thereby changes the coded message it carries—a process known as **RNA editing**.

In animals, two principal types of such RNA editing occur: the deamination of adenine to produce inosine (A-to-I editing) and, less frequently, the deamination of cytosine to produce uracil (C-to-U editing; see Figure 5–43). Because these chemical modifications alter the pairing properties of the bases (I pairs with C, and U pairs with A), they can have profound effects on the meaning of the RNA. If the edit occurs in a coding region, it can either change the amino acid sequence of the protein or produce a truncated protein by creating a premature stop codon. Edits that occur outside coding sequences can affect the pattern of pre-mRNA splicing, the transport of mRNA from the nucleus to the cytosol, the efficiency with which the RNA is translated, or the base-pairing between microRNAs (miRNAs) and their mRNA targets, a form of gene regulation that will be discussed later in the chapter.

The process of A-to-I editing is particularly prevalent in humans, where it occurs for approximately 1000 genes. Enzymes called *ADARs* (*adenosine deaminases acting on RNA*) perform this type of editing; these enzymes recognize a double-strand RNA structure that is formed through base-pairing between the site to be edited and a complementary sequence located elsewhere on the same RNA molecule, typically in an intron (**Figure 7–64**). The structure of

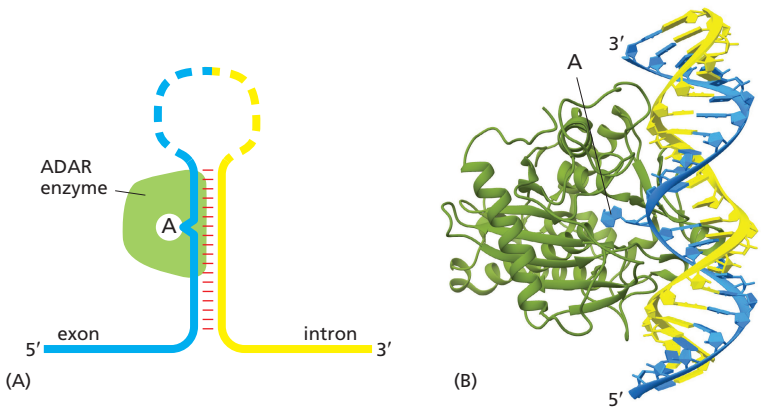


Figure 7–64 Mechanism of A-to-I RNA editing in mammals. (A) Typically, a sequence complementary to the position of the edit is present in an intron, and the resulting double-strand RNA structure attracts an A-to-I editing enzyme (ADAR). In the case illustrated, the edit is made in an exon; in most cases, however, it occurs in noncoding portions of the mRNA. Editing by ADAR takes place in the nucleus, before the pre-mRNA has been fully processed. Mice and humans have two ADAR genes: *ADR1* is expressed in many tissues and is required in the liver for proper red blood cell development; *ADR2* is expressed only in the brain, where it is required for proper brain development. (B) The human ADR2 enzyme bound to double-stranded RNA. The adenine to be edited is seen to be flipped out of the RNA double helix and buried deep in the catalytic pocket of the enzyme. Base flipping, which allows the enzyme access to the entire base, is also observed in enzymes that repair DNA (see Figure 5–42). (PDB code: 5ED1.)

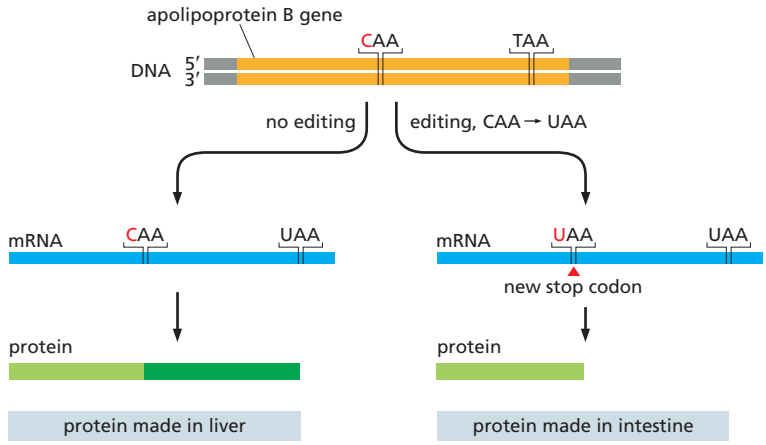


Figure 7–65 C-to-U RNA editing produces a truncated form of apolipoprotein B. As indicated, a tissue-specific edit in the middle of a coding sequence creates a truncated version of this protein in the intestine.

the double-stranded RNA specifies whether the mRNA is to be edited, and if so, where the edit should be made. An especially important example of A-to-I editing takes place in the mRNA that codes for a transmitter-gated ion channel in the brain. A single edit changes a glutamine to an arginine; the affected amino acid lies on the inner wall of the channel, and the editing change alters the Ca^{2+} permeability of the channel. Mutant mice that cannot make this edit are prone to epileptic seizures and die during or shortly after weaning, showing that editing of the ion channel RNA is normally crucial for proper brain development.

C-to-U editing, which is carried out by a different set of enzymes, is also important in mammals. For example, in certain cells of the gut, the mRNA for apolipoprotein B undergoes a C-to-U edit that creates a premature stop codon and therefore produces a shorter form of the protein. In cells of the liver, the editing enzyme is not expressed, and the full-length apolipoprotein B is produced. The two protein isoforms have different properties, and each plays a role in lipid metabolism that is specific to the organ that produces it (**Figure 7–65**).

Why RNA editing exists at all is a mystery. One idea is that it arose in evolution to correct "mistakes" in the genome. Another is that it arose as a somewhat slapdash way for the cell to produce subtly different proteins from the same gene. A third possibility is that RNA editing originally evolved as a defense mechanism against retroviruses and retrotransposons and was only later adapted by the cell to change the meanings of certain mRNAs. The last explanation receives support from the fact that RNA editing plays important roles in cell defense. The RNA genomes of some retroviruses, including HIV, are extensively edited after they infect cells. This hyperediting creates many harmful mutations in the viral RNA genome and also causes viral mRNAs to be retained in the nucleus, where they are eventually degraded. Although some modern retroviruses can protect themselves against this defense mechanism, RNA editing presumably helps to hold many viruses in check.

The Human AIDS Virus Illustrates How RNA Transport from the Nucleus Can Be Regulated

It has been estimated that in mammals only about one-twentieth of the total mass of RNA that is synthesized ever leaves the nucleus. We saw in Chapter 6 that most mammalian RNA molecules undergo extensive processing and that the "leftover" RNA fragments (excised introns and RNA sequences 3′ to the cleavage/poly-A site) are degraded in the nucleus. Incompletely processed and otherwise damaged RNAs are also eventually degraded as part of the quality-control system that acts on RNA production.

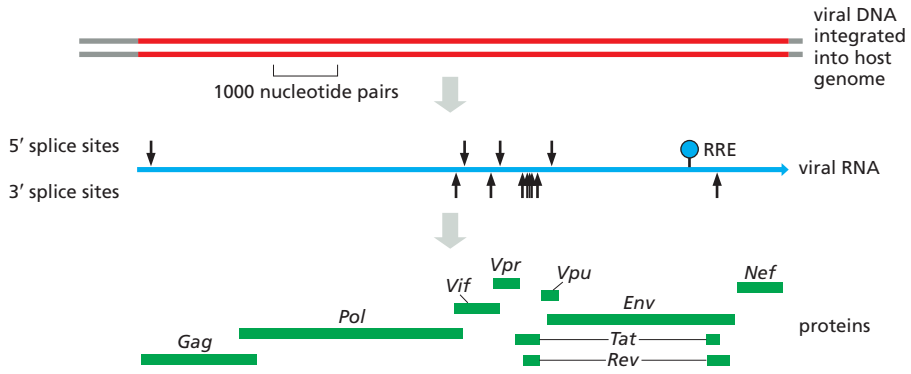


Figure 7–66 The compact genome of HIV, the human AIDS virus. The positions of the nine HIV genes are shown in *green.* The *red double line* indicates a DNA copy of the viral genome that has become integrated into the host DNA *(gray).* Note that the coding regions of many HIV genes overlap, and that those for *Tat* and *Rev* are split by introns. The *blue line* in the middle of the figure represents the pre-mRNA transcript of the viral DNA and shows the locations of all the possible splice sites *(arrows).* There are many alternative ways of splicing the viral transcript; for example, the *Env* mRNAs retain the intron that has been spliced out of the *Tat* and *Rev* mRNAs. The Rev response element (RRE) is indicated by a *blue* ball and stick. It is a 234-nucleotide-long stretch of RNA that folds into a defined structure; Rev recognizes a particular hairpin within this larger structure.

The *Gag* gene codes for a protein that is cleaved into several smaller proteins that form the viral capsid. The *Pol* gene codes for a protein that is cleaved to produce reverse transcriptase (which transcribes RNA into DNA), as well as the integrase involved in integrating the viral genome (as double-stranded DNA) into the host genome. The *Env* gene codes for the envelope proteins (see Figure 5–61). Tat, Rev, Vif, Vpr, Vpu, and Nef are small proteins with a variety of functions. As discussed in the text, Rev regulates nuclear export (see Figure 7–67), and Tat regulates the elongation of transcription across the integrated viral genome (see pp. 445–446).

As described in Chapter 6, the export of RNA molecules from the nucleus is delayed until processing has been completed. However, mechanisms that deliberately override this control point can be used to regulate gene expression. This strategy forms the basis for one of the best-understood examples of regulated nuclear transport of mRNA, which occurs in the human AIDS virus, HIV.

As we saw in Chapter 5, HIV, once inside the cell, directs the formation of a double-strand DNA copy of its single-strand RNA genome, which is then inserted into the genome of the host (see Figure 5–61). Once inserted, the viral DNA can be transcribed as one long RNA molecule by the host cell's RNA polymerase II. This transcript is then spliced in many different ways to produce more than 30 different species of mRNA, which in turn are translated into a variety of different proteins (**Figure 7–66**). In order to make progeny virus, entire unspliced viral transcripts must be exported from the nucleus to the cytosol, where they are packaged into viral capsids and serve as the viral genome. This large transcript, as well as certain alternatively spliced HIV mRNAs that are needed to produce viral proteins, still carry complete introns. The host cell's normal block to the nuclear export of unspliced RNAs therefore presents a special problem for HIV.

The block is overcome by a viral-coded protein (called Rev) that binds to a specific RNA sequence (called the Rev response element; RRE) located within a viral intron. The Rev protein interacts with a nuclear export receptor (Crm1), which directs the movement of viral RNAs through nuclear pores into the cytosol despite the presence of intron sequences. (How export receptors function is discussed in detail in Chapter 12.) The regulation of nuclear export by Rev has several important consequences for HIV growth and pathogenesis. In addition to ensuring the nuclear export of specific unspliced RNAs, it divides the viral infection into an early phase (in which Rev is translated from a fully spliced RNA, and all of the intron-containing viral RNAs are retained in the nucleus and degraded) and a late phase (in which unspliced RNAs are exported because of Rev function). This timing helps the virus replicate by providing the gene products in roughly the order in which they are needed (**Figure 7–67**).

Regulation by Rev and by Tat, the HIV protein that counteracts premature transcription termination (see pp. 445–446), allows the virus to achieve latency, a condition in which the HIV genome has become integrated into the host-cell genome, but the production of viral proteins has temporarily ceased. If, after the virus's initial entry into a host cell, conditions are unfavorable for viral replication, Rev and Tat are made at levels too low to promote transcription and export of unspliced RNA. This stalls the viral growth cycle until conditions improve, whereupon Rev and Tat levels increase and the virus enters the replication cycle.

mRNAs Can Be Localized to Specific Regions of the Cytosol

Once a newly made eukaryotic mRNA molecule has passed through a nuclear pore and entered the cytosol, it is typically met by ribosomes, which translate it into a polypeptide chain. Once the first round of translation "passes" the

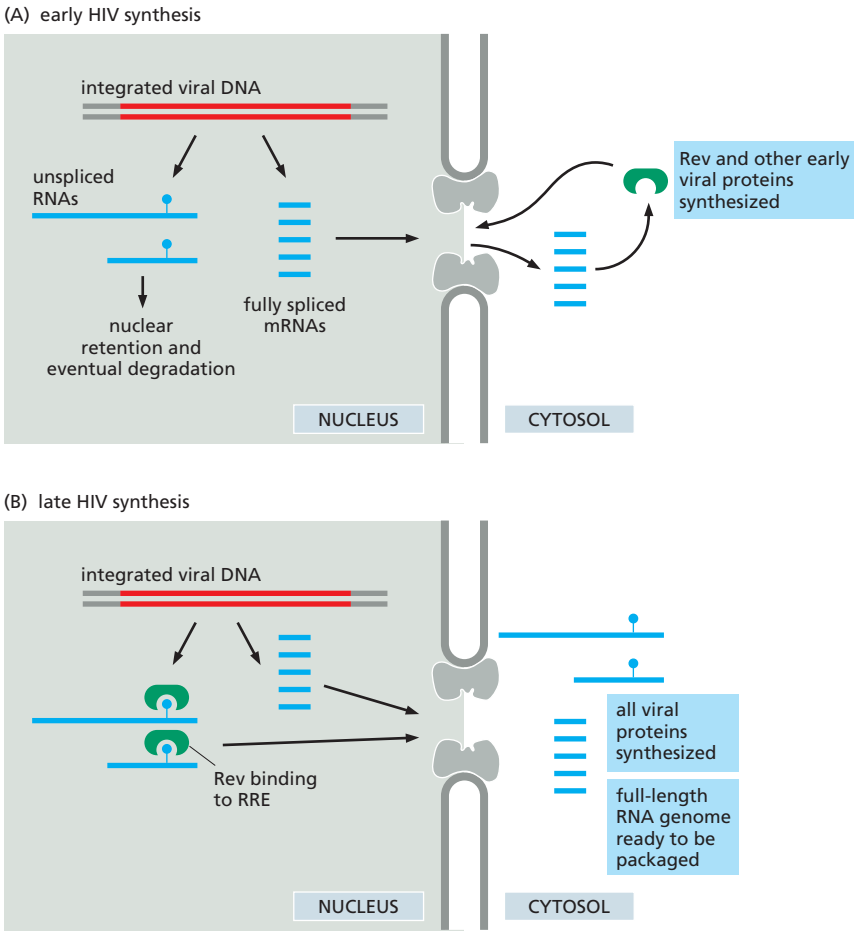


Figure 7–67 Regulation of nuclear export by the HIV Rev protein. (A) Early in HIV infection, only the fully spliced RNAs (which contain the coding sequences for Rev, Tat, and Nef) are exported from the nucleus and translated. (B) Once sufficient Rev protein has accumulated and been transported into the nucleus, unspliced viral RNAs can be exported from the nucleus. Many of these RNAs are translated into protein, and the full-length transcripts are packaged into new viral particles.

nonsense-mediated decay test (see Figure 6–80), the mRNA is usually translated in earnest. If the mRNA encodes a protein that is destined to be secreted or expressed on the cell surface, a signal sequence at the protein's N-terminus will direct it to the endoplasmic reticulum (ER). In this case, as discussed in Chapter 12, components of the cell's protein-sorting apparatus recognize the signal sequence as soon as it emerges from the ribosome and direct the entire complex of ribosome, mRNA, and nascent protein to the membrane of the ER, where the remainder of the polypeptide chain is synthesized. In other cases, free ribosomes in the cytosol synthesize the entire protein, and signals in the completed polypeptide chain may then direct the protein to other sites in the cell.

Many mRNAs are themselves directed to specific intracellular locations before their efficient translation begins, allowing the cell to position its mRNAs close to the sites where the encoded protein is needed. RNA localization has been observed in many organisms, including unicellular fungi, plants, and animals, and it appears to be a common mechanism that cells use to concentrate high-level production of proteins at specific sites. This strategy also provides the cell with other advantages. For example, it allows the establishment of asymmetries in the cytosol of the cell, a key step in many stages of development.

The localization of mRNA, coupled with translational control, also allows the cell to regulate gene expression independently in different regions. This feature is particularly important in large, highly polarized cells such as neurons; in those cells, specific mRNAs are transported for long distances along axons and dendrites to synapses, and the translation of the mRNAs that become localized there is often controlled by synaptic activity.

The mechanisms for mRNA localization that have been discovered all require specific signals in the mRNA itself (**Figure 7–68**). These signals are usually

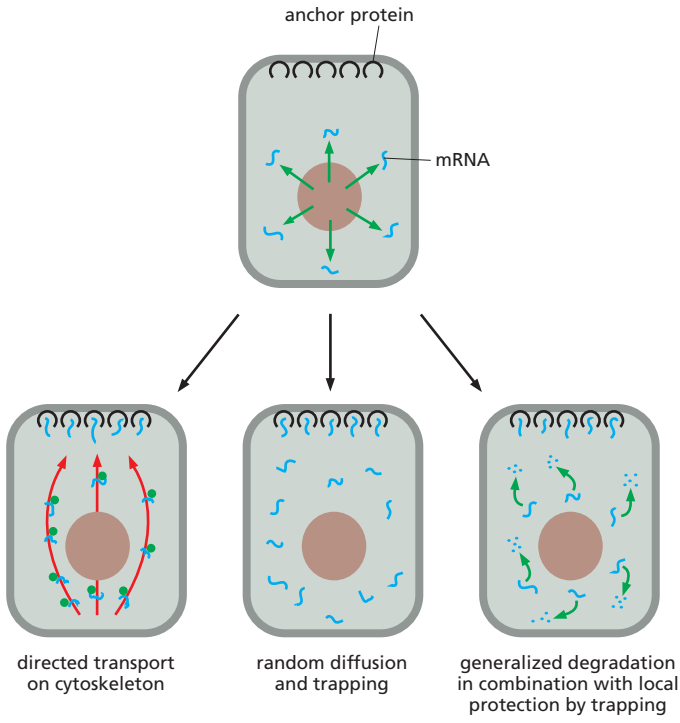


Figure 7–68 Mechanisms for the localization of mRNAs. The mRNA to be localized leaves the nucleus through nuclear pores *(top)*. Some localized mRNAs *(left diagram)* travel to their destination by associating with cytoskeletal motors, which use the energy of ATP hydrolysis to move the mRNAs unidirectionally along filaments in the cytoskeleton *(red)* (see Chapter 16). At their destination, the mRNAs are held in place by anchor proteins *(black)*. Other mRNAs randomly diffuse through the cytosol and are simply trapped by anchor proteins at their sites of localization *(center diagram)*. As an additional feature, many mRNAs *(right diagram)* are degraded in the cytosol unless they have bound, through random diffusion, a localized anchor protein complex that protects the mRNA from degradation *(black)*. These mechanisms require specific signals on the mRNA, which are typically located in the 3′ UTR. In all cases, other RNA-bound components block the translation of the mRNA until it is properly localized. Even then, additional signals are often needed to begin translation. (Adapted from H.D. Lipshitz and C.A. Smibert, *Curr. Opin. Genet. Dev.* 10:476–488, 2000.)

concentrated in the 3′ *untranslated region* (*UTR*), the region of RNA that extends from the stop codon that terminates protein synthesis to the start of the poly-A tail (**Figure 7–69**). As in neurons, mRNA localization is usually coupled with translational controls to ensure that the localized mRNA remains quiescent until it is needed.

The *Drosophila* egg provides an especially striking example of mRNA localization. The mRNA encoding the Bicoid transcription regulator is localized by attachment to the cytoskeleton at the anterior tip of the developing egg. When fertilization triggers the translation of this mRNA, it generates a gradient of the Bicoid protein that plays a crucial part in directing the development of the anterior part of the embryo (see Figures 7–29 and 21–19). Many mRNAs in somatic cells are also localized in a similar way. The mRNA that encodes actin, for example, is localized to the actin-filament-rich cell cortex in mammalian fibroblasts by means of a 3′ UTR signal.

We saw in Chapter 6 that mRNA molecules exit from the nucleus bearing numerous markings in the form of RNA modifications (the 5′ cap and the 3′ poly-A tail) and bound proteins (exon junction complexes, for example) that signify the successful completion of the different pre-mRNA processing steps. As just described, the 3′ UTR of an mRNA can be thought of as a "ZIP code" that directs mRNAs to different places in the cell. Shortly, we will see that mRNAs also carry information specifying their average lifetime in the cytosol and the efficiency with

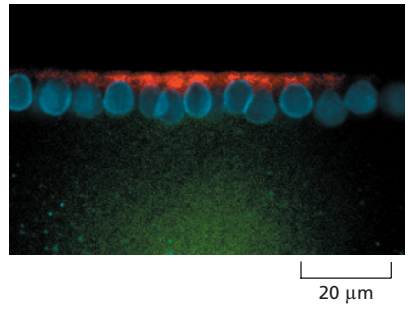


Figure 7–69 An experiment demonstrating the importance of the 3′ UTR in localizing mRNAs to specific regions of the cytoplasm. For this experiment, two different fluorescently labeled RNAs were prepared by transcribing DNA *in vitro* in the presence of fluorescently labeled derivatives of ribonucleoside triphosphates. One RNA (labeled with a *red* fluorochrome) contains the coding region for the *Drosophila* Hairy protein and includes the adjacent 3′ UTR. The other RNA (labeled *green*) contains the Hairy coding region with the 3′ UTR deleted. The two RNAs were mixed and injected into a *Drosophila* embryo at a stage of development when multiple nuclei reside in a common cytoplasm (see Figure 7–29). When the fluorescent RNAs were visualized 10 minutes later, the full-length *hairy* RNA *(red)* was localized to the apical side of nuclei *(blue)*, whereas the transcript missing the 3′ UTR *(green)* failed to localize and is seen as a diffuse cloud. Hairy is one of many transcription regulators that specify positional information in the developing *Drosophila* embryo, and the localization of its mRNA (shown in this experiment to depend on its 3′ UTR) is critical for proper fly development. (Courtesy of Simon Bullock and David Ish-Horowicz.)

which they are translated into protein. In a broad sense, the untranslated regions of eukaryotic mRNAs resemble the transcriptional control regions of genes: their nucleotide sequences contain information specifying the way the RNA is to be used, and proteins interpret this information by binding specifically to these sequences. Thus, in addition to the specification of amino acid sequences, mRNA molecules are rich with other types of information.

Untranslated Regions of mRNAs Control Their Translation

Once an mRNA has been synthesized, one of the most common ways of regulating the levels of its protein product is to control the step that initiates translation. Even though the details of translation initiation differ between eukaryotes and bacteria (as we saw in Chapter 6), they each use some of the same basic regulatory strategies.

In bacterial mRNAs, a conserved stretch of nucleotides—the *Shine–Dalgarno sequence*—is always found a few nucleotides upstream of the initiating AUG codon, and it is required to start protein synthesis (see Figure 6–75). The control of bacterial translation generally involves either exposing or blocking this critical sequence (**Figure 7–70**).

Eukaryotic mRNAs do not contain such a sequence. Instead, as discussed in Chapter 6, the selection of an AUG codon as a translation start site is largely determined by its proximity to the cap at the 5′ end of the mRNA molecule, which is the site at which the small ribosomal subunit binds to the mRNA and begins

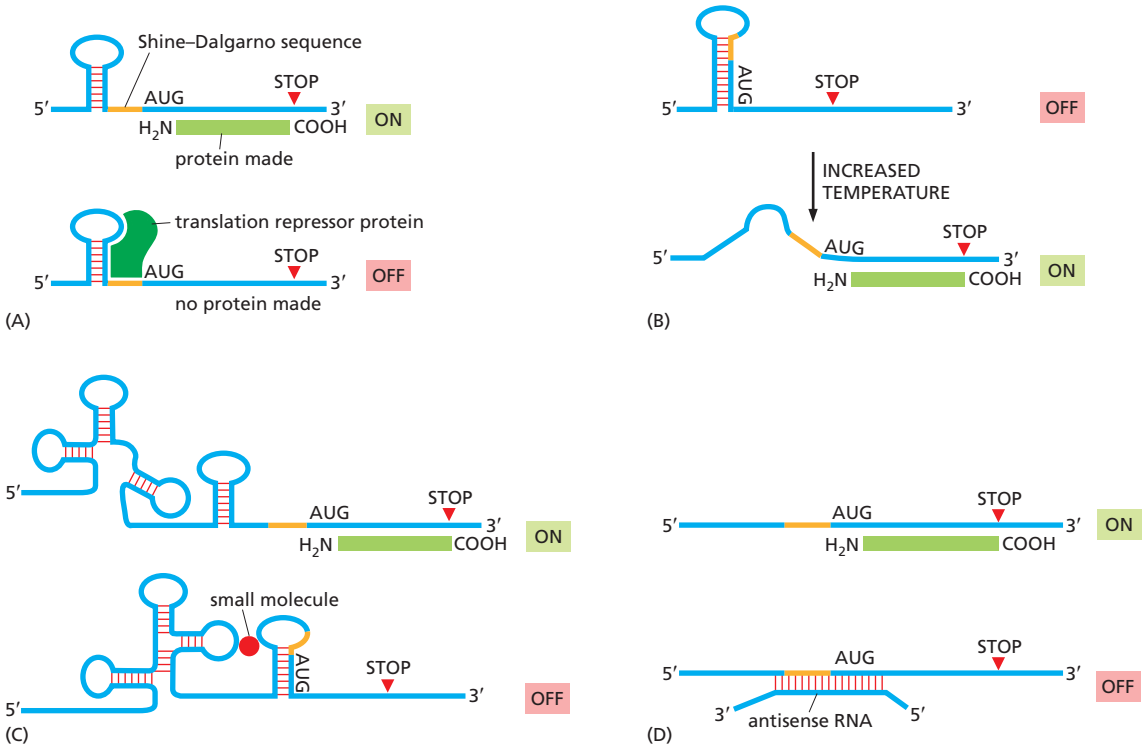


Figure 7–70 Mechanisms of translational control. Although these examples are from bacteria, many of the same principles operate in eukaryotes. (A) Sequence-specific RNA-binding proteins repress translation of specific mRNAs by blocking access of the ribosome to the Shine–Dalgarno sequence *(orange)*. For example, some ribosomal proteins repress translation of their own mRNA. This negative feedback mechanism allows the cell to maintain balanced quantities of the various components needed to form ribosomes. (B) An RNA "thermosensor" permits efficient translation initiation only at elevated temperatures at which the stem–loop structure has been melted. An example occurs in the human pathogen *Listeria monocytogenes*, in which the translation of its virulence genes increases at 37°C, the temperature of the host. (C) Binding of a small molecule to a riboswitch causes a major rearrangement of RNA structure, creating a different set of stem–loop structures. In the bound structure, the Shine–Dalgarno sequence *(orange)* is sequestered, and translation initiation is thereby blocked. In many bacteria, *S*-adenosylmethionine acts in this manner to block production of the enzymes that synthesize it. (D) An "antisense" RNA produced from elsewhere in the genome base-pairs with a specific mRNA and blocks its translation. Many bacteria regulate expression of iron-storage proteins in this way.

scanning for an initiating AUG codon. In eukaryotes, translational repressors can bind to the 5′ end of the mRNA and thereby inhibit translation initiation (see Figure 7–74). A particularly important type of translational control in eukaryotes relies on small RNAs (termed *microRNAs*, or *miRNAs*) that bind to mRNAs and reduce protein output, as described later in this chapter.

The Phosphorylation of an Initiation Factor Regulates Protein Synthesis Globally

Eukaryotic cells decrease their overall rate of protein synthesis in response to a variety of situations, including deprivation of growth factors or nutrients, infection by viruses, and sudden increases in temperature. (This response is coordinated by the TOR signaling pathway, which is described in Chapter 17; see Figure 17–61.) Much of the decrease in translation is caused by the phosphorylation of the translation initiation factor eIF2 by specific protein kinases that respond to the changes in conditions.

The normal function of eIF2 was outlined in Chapter 6 (see Figure 6–74). It forms a complex with GTP and mediates the binding of the methionyl initiator tRNA to the small ribosomal subunit, which then binds to the 5′ end of the mRNA and begins scanning along the mRNA. When an AUG codon is recognized, the eIF2 protein hydrolyzes the bound GTP to GDP, causing a conformational change in the protein and releasing it from the small ribosomal subunit. The large ribosomal subunit then joins the small one to form a complete ribosome that begins protein synthesis.

Because eIF2 binds very tightly to GDP, a guanine nucleotide exchange factor (see p. 880) designated eIF2B is required to release the GDP from eIF2 so that a new GTP molecule can bind—as required for eIF2 reuse (**Figure 7–71A**). When eIF2 is phosphorylated, it binds to eIF2B unusually tightly, inactivating this exchange factor. Because there is more eIF2 than eIF2B in cells, even a fraction of phosphorylated eIF2 can trap nearly all of the eIF2B. Without this exchange factor, GDP remains bound to nearly all of the nonphosphorylated eIF2, greatly slowing protein synthesis (**Figure 7–71B**).

Regulation of the level of active eIF2 is especially important in mammalian cells. As described in Chapter 17, eIF2 down-regulation is part of the mechanism that allows these cells to enter a nonproliferating, resting state (called G_0) in which the rate of total protein synthesis is reduced to about one-fifth the rate in proliferating cells.

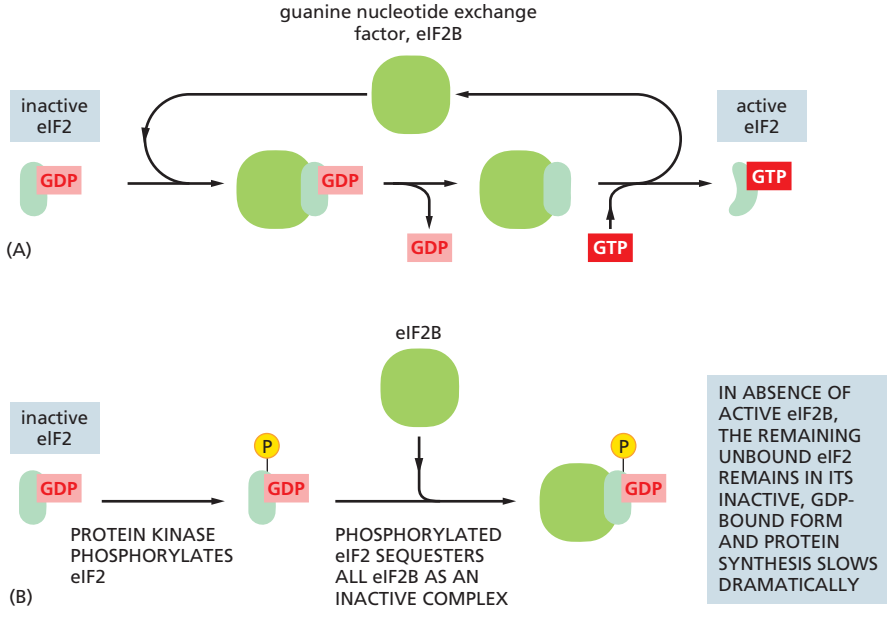


Figure 7–71 The eIF2 cycle. (A) The recycling of used eIF2 by a guanine nucleotide exchange factor (eIF2B). (B) How eIF2 phosphorylation controls protein synthesis rates by sequestering eIF2B.

Initiation at AUG Codons Upstream of the Translation Start Can Regulate Eukaryotic Translation Initiation

We saw in Chapter 6 that eukaryotic translation typically begins at the first AUG downstream of the 5′ end of the mRNA, which is the first AUG encountered by a scanning small ribosomal subunit. But, as we have seen, the nucleotides immediately surrounding the AUG also influence the efficiency of translation initiation. If the recognition site is poor enough, scanning ribosomal subunits will sometimes ignore the first AUG codon in the mRNA and skip to the second or third AUG codon instead. This phenomenon, known as "leaky scanning," is a strategy frequently used to produce two or more closely related proteins, differing only in their N-termini, from the same mRNA. A particularly important use of this mechanism is the production of the same protein with and without a signal sequence attached at its N-terminus. This allows the protein to be directed to two different locations in the cell (for example, to both mitochondria and the cytosol). Cells can regulate the relative abundance of the protein isoforms produced by leaky scanning; for example, a cell-type-specific increase in the abundance of the initiation factor eIF4F favors the use of the AUG closest to the 5′ end of the mRNA, even if it is surrounded by nonoptimal nucleotides.

Another type of control found in eukaryotes uses one or more short *open reading frames*—short stretches of DNA that begin with a start codon (ATG) and end with a stop codon, with no stop codons in between—that lie between the 5′ end of the mRNA and the beginning of a gene. Often, the amino acid sequences coded by these upstream open reading frames (uORFs) are not important; instead, the uORFs serve a purely regulatory function. A uORF present on an mRNA molecule will generally decrease translation of the downstream gene by trapping a scanning ribosome initiation complex and causing the ribosome to translate the uORF and dissociate from the mRNA before it reaches the bona fide protein-coding sequence.

When the activity of a general translation factor (such as eIF2 discussed earlier) is reduced, one might expect that the translation of all mRNAs would be reduced equally. Contrary to this expectation, however, the phosphorylation of eIF2 can have selective effects, even enhancing the translation of specific mRNAs that contain uORFs. This can enable cells, for example, to adapt to starvation for specific nutrients by shutting down the synthesis of all proteins except those that are required for synthesis of the missing nutrients. The details of this mechanism have been worked out for the yeast mRNA that encodes a protein called Gcn4, a transcription regulator that activates many genes that encode proteins that are important for amino acid synthesis.

The *Gcn4* mRNA encodes several short uORFs, and when amino acids are abundant, ribosomes translate the uORFs and generally dissociate before they reach the *Gcn4* coding region. But a global decrease in eIF2 activity brought about by amino acid starvation makes it more likely that a scanning small ribosomal subunit will move across the uORFs (without translating them) before it acquires a molecule of eIF2. This ribosomal subunit is then free to initiate translation on the actual *Gcn4* sequences, and the increased level of this transcription regulator increases the production of amino acid biosynthetic enzymes. Thus, when cells encounter "hard times," phosphorylation of eIF2 globally decreases translation while increasing synthesis of those proteins most needed by the cell to cope with the new conditions.

Internal Ribosome Entry Sites Also Provide Opportunities for Translational Control

Although most eukaryotic mRNAs are translated beginning with the first AUG downstream from the 5′ cap, certain AUGs, as we just saw, can be skipped over during the scanning process. There is a second way that cells can initiate translation at positions distant from the 5′ end of the mRNA, using a specialized

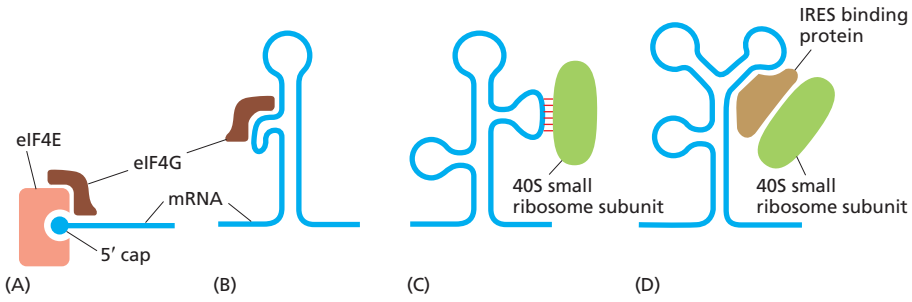


Figure 7–72 Internal ribosome entry sites (IRESs) can promote translation initiation by a variety of mechanisms. (A) The normal cap-dependent mechanism requires eIF4G binding to the cap to begin assembly of the other translation components (see Figure 6–74). (B) The cap and eIF4E are bypassed by direct binding of eIF4G to a specific RNA structure formed by the IRES. (C) The small ribosome subunit binds directly to the IRES through base-pairing between sequences in the IRES and the l8S rRNA, positioning it to begin translation. (D) Specialized proteins bind to an IRES and then attract the small ribosome subunit.

type of RNA sequence called an **internal ribosome entry site (IRES)**. In some cases, two distinct protein-coding sequences are carried in tandem on the same eukaryotic mRNA; translation of the first occurs by the usual scanning mechanism utilizing the first AUG encountered, and translation of the second occurs by means of an IRES located much further into the mRNA. IRESs are typically several hundred nucleotides in length, and they fold into specific structures that bypass the need for a 5′ cap and the translation factor that recognizes it, eIF4E (**Figure 7–72**).

It is estimated that 10% of all mammalian mRNAs contain an IRES. Some of these protein synthesis start sites are specifically activated by external signals such as stress. But the best-understood examples occur with viruses, which use IRESs as part of a strategy to get their own mRNA molecules translated while blocking the normal 5′ cap–dependent translation of host mRNAs. On infection, these viruses produce a protease (encoded in the viral genome) that cleaves the host-cell translation factor eIF4G, rendering it unable to bind to eIF4E, the cap-binding complex (see Figure 6–74). This shuts down most of the host cell's translation and effectively diverts the translation machinery to the IRES sequences present on the viral mRNAs. (The truncated eIF4G remains competent to initiate translation at these internal sites.)

The many ways in which viruses manipulate their host's protein-synthesis machinery for their own advantage continue to surprise cell biologists. Studying the many results of this evolutionary "arms race" between humans and pathogens has led to many fundamental insights into the workings of the cell, and we revisit this topic in Chapter 23.

Changes in mRNA Stability Can Control Gene Expression

Most mRNAs in a bacterial cell are very unstable, having half-lives of less than a couple of minutes. Exonucleases, which degrade in the 3′-to-5′ direction, are usually responsible for the rapid destruction of these mRNAs. Because its mRNAs are both rapidly synthesized and rapidly degraded, a bacterium can adapt quickly to environmental changes.

As a general rule, the mRNAs in eukaryotic cells are more stable. Some, such as that encoding β-globin, have half-lives of more than 10 hours. But most are considerably less stable, with half-lives of less than 30 minutes. The mRNAs that code for proteins such as growth factors and transcription regulators, whose production rates need to change rapidly in cells, are especially short-lived.

We saw in Chapter 6 that the cell has several mechanisms that rapidly destroy incorrectly processed RNAs. But now we focus on the ultimate fate of a typical "normal" eukaryotic mRNA molecule. Two general mechanisms exist for eventually destroying it, both of which begin with a gradual shortening of the poly-A tail by an exonuclease, a process that starts as soon as the mRNA reaches the cytosol. In a broad sense, this poly-A shortening acts as a timer that counts down the lifetime of each mRNA. Once the poly-A tail is reduced to a critical length (about 25 nucleotides in humans), the two destruction pathways converge. In one, the 5′ cap is removed (a process called decapping), and the "exposed" mRNA is rapidly degraded from its 5′ end. In the other, the mRNA continues to

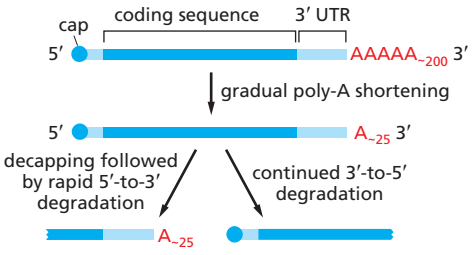


Figure 7–73 Two mechanisms of eukaryotic mRNA decay. Once a mature mRNA is exported to the cytosol, enzymes known as deadenylases gradually shorten its poly-A tail. When a critical threshold of poly-A tail length occurs, the two degradation mechanisms shown are triggered, probably by loss of the poly-A–binding proteins. Although 5′-to-3′ and 3′-to-5′ degradation are shown here on separate RNA molecules, these two processes can occur together on the same molecule. (Adapted from C.A. Beelman and R. Parker, *Cell* 81:179–183, 1995.)

be degraded from the 3′ end, through the poly-A tail into the coding sequences (Figure 7–73).

Nearly all mRNAs are subject to both types of decay, which can occur simultaneously on the same mRNA molecule. Specific nucleotide sequences determine how fast each step occurs and therefore how long each mRNA will persist in the cell and be able to produce protein. The 3′ UTR sequences are especially important in controlling mRNA lifetimes, and they often carry binding sites for specific proteins that increase or decrease the rates of poly-A shortening, decapping, or 3′-to-5′ degradation. The half-life of an mRNA is also affected by how efficiently it is translated. Poly-A shortening and decapping compete directly with the machinery that translates the mRNA; therefore, any factors that increase the translation efficiency for an mRNA will tend to reduce its degradation.

Although poly-A shortening controls the half-life of most eukaryotic mRNAs, some mRNAs can be degraded by a specialized mechanism that bypasses this step altogether. In these cases, specific endonucleases cleave the mRNA internally, effectively decapping one end and removing the poly-A tail from the other, so that both halves are rapidly degraded. The mRNAs that are destroyed in this way carry specific nucleotide sequences—often in their 3′ UTRs—that serve as recognition sequences for these endonucleases. This strategy makes it simple to tightly regulate the stability of these mRNAs by blocking or exposing the endonuclease site in response to extracellular signals. For example, the addition of iron to cells decreases the stability of the mRNA that encodes the receptor protein that binds the iron-transporting protein transferrin, causing less of this receptor to be made. This effect is mediated by the iron-sensitive RNA-binding protein aconitase. During iron starvation, aconitase binds the 3′ UTR of the transferrin receptor mRNA and increases receptor production by blocking endonucleolytic cleavage of the mRNA (Figure 7–74A). On the addition of iron, aconitase is released from

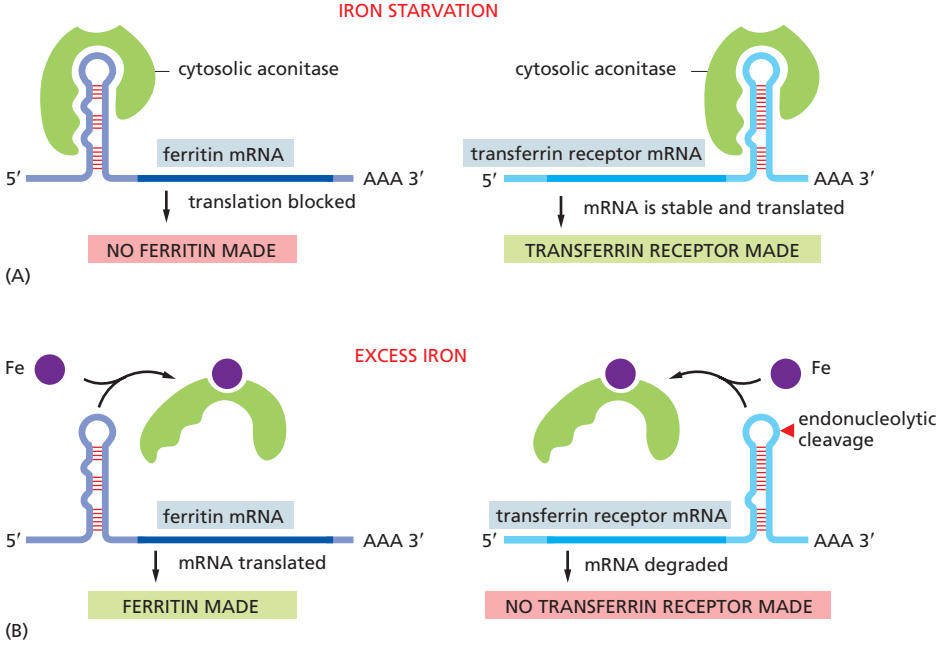


Figure 7–74 Two post-translational controls mediated by iron. (A) During iron starvation, the binding of aconitase to the 5′ UTR of the ferritin mRNA blocks translation initiation; its binding to the 3′ UTR of the transferrin receptor mRNA blocks an endonuclease cleavage site and thereby stabilizes the mRNA. (B) In response to an increase in iron concentration in the cytosol, a cell increases its synthesis of ferritin in order to bind the extra iron and decreases its synthesis of transferrin receptors in order to import less iron across the plasma membrane. Both responses are mediated by the same iron-responsive regulatory protein, aconitase, which recognizes common features in a stem–loop structure in the mRNAs encoding ferritin and the transferrin receptor. Aconitase dissociates from the mRNA when it binds iron. But because the transferrin receptor and ferritin are regulated by different types of mechanisms, their levels respond oppositely to iron concentrations even though they are regulated by the same iron-responsive regulatory protein. (Adapted from M.W. Hentze et al., *Science* 238:1570–1573, 1987; and J.L. Casey et al., *Science* 240:924–928, 1988.)

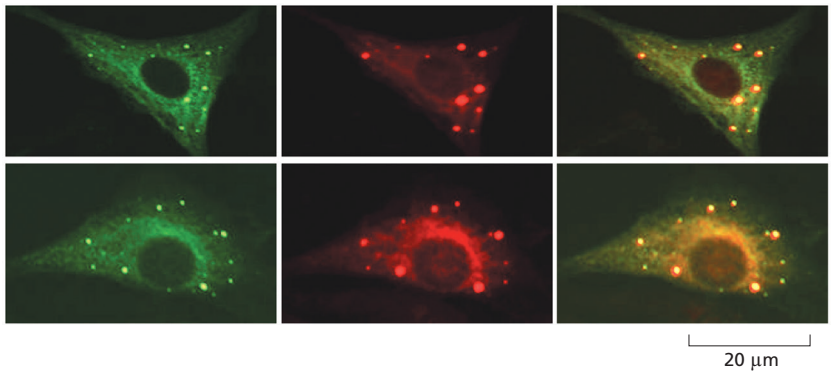


Figure 7–75 Visualization of P-bodies. Human cells were stained with antibodies to a component of the mRNA decapping enzyme Dcp1a *(left panels)* and to the Argonaute protein *(middle panels).* As described later in this chapter, Argonaute is a key component of RNA interference pathways and both it and the decapping enzyme destabilize mRNAs. The merged images *(right panels)* show that the two proteins co-localize to P-bodies in the cytoplasm. (Adapted from J. Liu et al., *Nat. Cell Biol.* 7:719–723, 2005. Reproduced with permission from SNCSC.)

the mRNA, exposing the cleavage site and thereby decreasing mRNA stability (**Figure 7–74B**).

Regulation of mRNA Stability Involves P-bodies and Stress Granules

We saw in Chapter 6 that large aggregates of RNA and protein can form membraneless compartments in the nucleus, such as nucleoli and Cajal bodies. The cytosol also contains such biomolecular condensates, and here we discuss two of them, *Processing* or *P-bodies* and *stress granules*, each of which has a role in handling mRNAs (**Figure 7–75**). When an mRNA in the cytosol is no longer actively translated, it often moves to P-bodies where several fates are possible. P-bodies are rich in mRNA-degrading enzymes, and mRNAs that have already undergone significant poly-A shortening can continue to be degraded within P-bodies. Alternatively, some intact mRNAs can be stored in P-bodies in a translationally repressed form. According to the needs of the cell, these mRNAs can then be moved back to the cytosol and "reactivated" to begin translation again (**Figure 7–76**). mRNAs stored in this way often code for proteins that the cell needs quickly, and this strategy bypasses the time-consuming steps of *de novo* mRNA production.

Stress granules are dynamic membraneless organelles that form when the cell undergoes a sudden block to translation, whether by starvation, small-molecule inhibitors, or genetic manipulation. These treatments allow ongoing translation to be completed but block new translation initiation. The resulting ribosome-free mRNAs accumulate in stress granules that grow in size as more and more mRNAs enter them. As the stressful conditions are relieved, the stress granules shrink along with the release of the stored mRNAs to the cytosol where they resume being translated. Clearly, once a cell has made the large investment in producing a properly processed mRNA molecule, it carefully controls its subsequent fate.

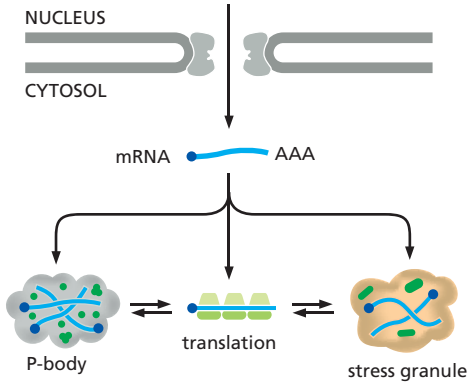


Figure 7–76 Possible fates of an intact mRNA molecule. An mRNA molecule released from the nucleus can be actively translated *(center)*, stored in P-bodies *(left),* or, if the cell is stressed, moved into stress granules *(right)*. As the needs of the cell change, stored mRNAs can be reactivated and returned to the cytosol to be translated into protein. Although not shown, all mRNA molecules are eventually degraded, and some of the final steps take place in P-bodies.

Summary

Many steps in the pathway from RNA to protein are regulated by cells in order to control gene expression. Most genes are regulated at multiple levels, in addition to being controlled at the initiation stage of transcription. The regulatory mechanisms include (1) attenuation of the RNA transcript by its premature termination, (2) alternative RNA splice-site selection, (3) control of 3′-end formation by cleavage and poly-A addition, (4) RNA covalent modifications including editing, (5) control of transport from the nucleus to the cytosol, (6) localization of mRNAs to particular parts of the cytoplasm, (7) control of translation initiation, and (8) regulated mRNA degradation. Most of these control processes require the recognition of specific sequences or structures in the RNA molecule being regulated, a task performed by either regulatory proteins or regulatory RNA molecules.

REGULATION OF GENE EXPRESSION BY NONCODING RNAs

In the previous chapter, we introduced the *central dogma*, according to which the flow of genetic information proceeds from DNA through RNA to protein (see Figure 6–1). But we have seen throughout this book that RNA molecules perform many critical tasks in the cell besides serving as intermediate carriers of genetic information. Among these noncoding RNAs are the rRNA and tRNA molecules, which are responsible for reading the genetic code and synthesizing proteins. The RNA molecule in telomerase serves as a template for the replication of chromosome ends, snoRNAs modify ribosomal RNA, and snRNAs direct RNA splicing. And earlier in this chapter we saw that Xist RNA has an important role in inactivating one copy of the X chromosome in female mammals. In this section, we introduce several additional classes of noncoding RNAs that have important roles in regulating gene expression and in protecting the genome from viruses and transposable elements. These RNAs also make possible powerful new experimental techniques in genome editing.

Small Noncoding RNA Transcripts Regulate Many Animal and Plant Genes Through RNA Interference

We begin our discussion with a group of short RNAs that carry out **RNA interference**, or **RNAi**. Here, short single-stranded RNAs (20–30 nucleotides) serve as guide RNAs that selectively bind—through complementary base-pairing—other RNAs in the cell. When the target is a mature mRNA, the small noncoding RNAs can inhibit its translation or catalyze its rapid destruction. If the target RNA molecule is in the process of being transcribed, the small noncoding RNA can bind to it and direct the formation of repressive chromatin on its attached DNA template to block further transcription (**Figure 7–77**).

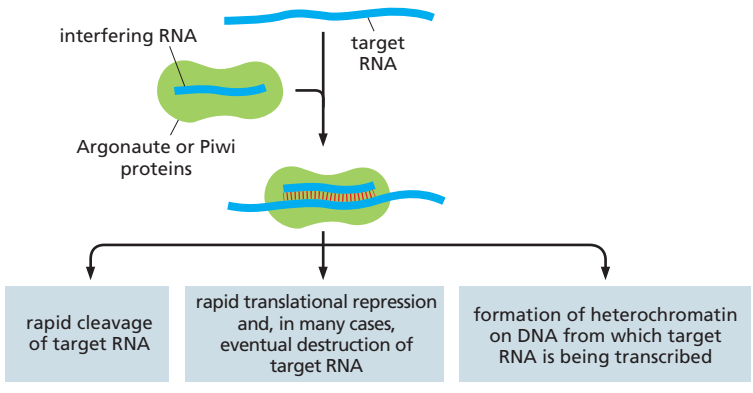


Figure 7–77 RNA interference in eukaryotes. Single-strand interfering RNAs locate target RNAs through complementary base-pairing, and, at this point, several fates are possible, as shown. As described in the text, there are several types of RNA interference; the way that interfering RNA is produced and the ultimate fate of the target RNA depend on the particular system.

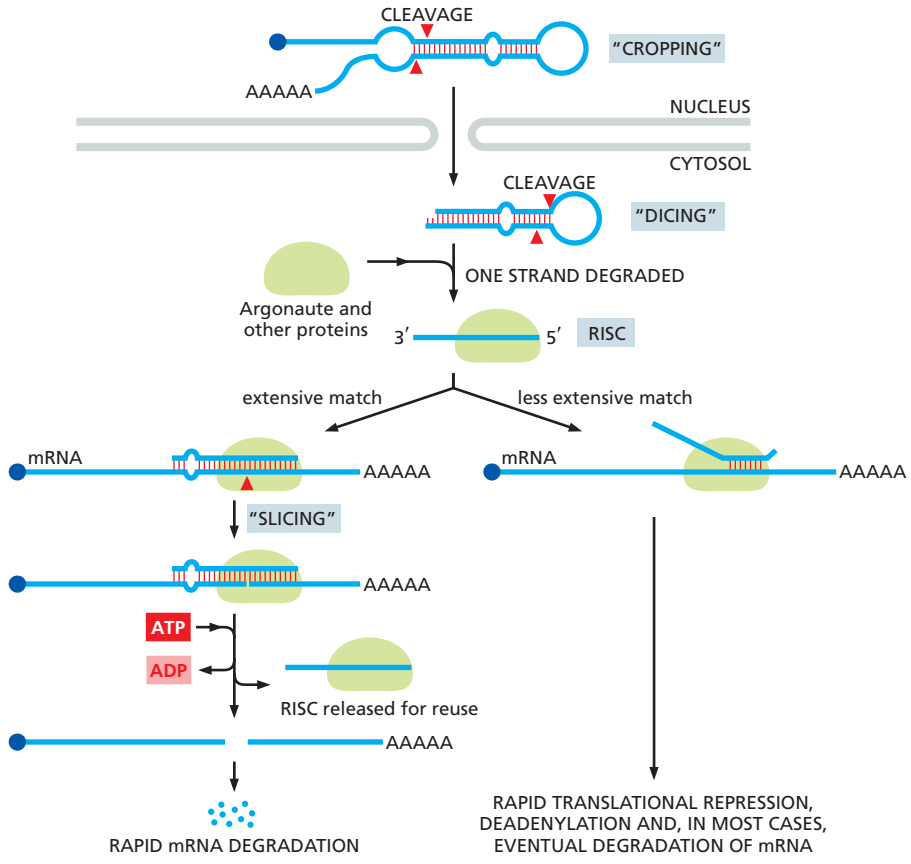


Figure 7–78 miRNA processing and mechanism of action. The precursor miRNA, through complementary base-pairing between one part of its sequence and another, forms a double-strand structure. This RNA is "cropped" while still in the nucleus and then exported to the cytosol, where it is further cleaved ("diced") by the Dicer enzyme to form the miRNA proper. Argonaute, in conjunction with other components of RISC, initially associates with both strands of the miRNA and then cleaves and discards one of them. The other strand guides RISC to specific mRNAs through base-pairing. If the RNA–RNA match is extensive, as is commonly seen in plants, Argonaute cleaves the target mRNA ("slicing"), causing its rapid degradation. In mammals, the miRNA–mRNA match often does not extend beyond a short seven-nucleotide "seed" region near the 5′ end of the miRNA. This less extensive base-pairing leads to a rapid inhibition of translation and, in most cases, eventual destruction of the mRNA.

Three classes of small noncoding RNAs work in this way—*microRNAs* (*miRNAs*), *small interfering RNAs* (*siRNAs*), and *piwi-interacting RNAs* (*piRNAs*)—and we discuss them in turn in the next sections. Although they differ in both the way the short pieces of single-stranded RNA are generated and in their ultimate functions, all three types of RNAs locate their targets through RNA–RNA base-pairing, and they generally cause reductions in gene expression.

miRNAs Regulate mRNA Translation and Stability

More than 1000 different **microRNAs (miRNAs)** are produced from the human genome, and these appear to regulate at least one-half of all human protein-coding genes. Once made, miRNAs base-pair with specific mRNAs and fine-tune their translation and stability. The miRNA precursors are synthesized by RNA polymerase II and are capped and polyadenylated. They then undergo a special type of processing, after which the miRNA (typically 23 nucleotides in length) is assembled with a set of proteins to form an *RNA-induced silencing complex*, or *RISC*. Once formed, the RISC seeks out its target mRNAs by searching for complementary nucleotide sequences (**Figure 7–78**). This search is greatly facilitated by the Argonaute protein, a component of RISC, which holds the 5′ region of the miRNA so that it is optimally positioned for base-pairing to another RNA molecule (**Figure 7–79**). In animals, the extent of base-pairing is typically at least seven nucleotide pairs, and this pairing most often occurs in the 3′ UTR of the target mRNA.

Once an mRNA has been bound by an miRNA, several outcomes are possible. If the base-pairing is extensive (which is unusual in humans but common in many plants), the mRNA is cleaved (*sliced*) by the Argonaute protein, effectively removing the mRNA's poly-A tail and exposing it to exonucleases (see Figure 7–73). After cleavage of the mRNA, the RISC with its associated miRNA is released, and it can seek out additional mRNAs (see Figure 7–78). Thus, a single miRNA can act

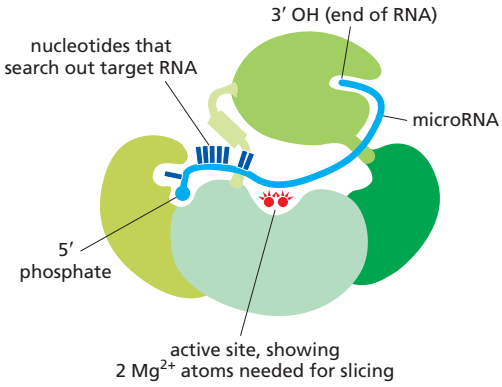


Figure 7–79 Human Argonaute protein carrying an miRNA. The protein is folded into four structural domains, each indicated by a different color. The miRNA is held in an extended form that is optimal for forming RNA–RNA base pairs. The active site of Argonaute that slices a target RNA, when it is extensively base-paired with the miRNA, is indicated in *red*. Many Argonaute proteins (three out of the four human proteins, for example) lack the catalytic site and therefore bind target RNAs without slicing them. (Adapted from C.D. Kuhn and L. Joshua-Tor, *Trends Biochem. Sci.* 38:263–271, 2013.)

catalytically to destroy many complementary mRNAs. These miRNAs can thus be thought of as guide sequences that repeatedly bring destructive nucleases into contact with specific mRNAs.

If the base-pairing between the miRNA and the mRNA is less extensive (as observed for most human miRNAs), Argonaute does not slice the mRNA; rather, translation of the mRNA is repressed by the recruitment of deadenylase enzymes—which shorten the poly-A tail—and other proteins that directly block access of the mRNA to the proteins needed to translate it. In many cases, the "blocked" mRNAs are shuttled to P-bodies (see Figure 7–76) where, sequestered from ribosomes, they are either degraded or, at a later time, released to the cytosol to be translated again.

Several features make miRNAs especially useful regulators of gene expression. First, a single miRNA can regulate a whole set of different mRNAs, so long as the mRNAs carry a short complementary sequence in their UTRs. This situation is common in humans, where a single miRNA can control hundreds of different mRNAs. Second, regulation by miRNAs can be combinatorial. As discussed earlier for transcription regulators, combinatorial control greatly expands the possibilities available to the cell by linking gene expression to a combination of different regulators rather than a single regulator. Like many transcription regulators, different miRNAs can bind cooperatively to their target mRNAs if their recognition sites are spaced appropriately. The basis for the cooperative binding is a scaffold protein that weakly holds two different RISCs together at a fixed spacing, thereby coupling their individual miRNA–mRNA binding energies. Third, an miRNA occupies relatively little space in the genome when compared with a protein. Indeed, their small size is one reason that miRNAs were discovered only recently. Although we are only beginning to appreciate the full impact of miRNAs, it is clear that they represent an important part of the cell's repertoire for regulating the expression of genes. We shall discuss specific examples of miRNAs with key roles in development in Chapter 21.

RNA Interference Also Serves as a Cell Defense Mechanism

Many of the proteins that participate in the miRNA regulatory mechanisms just described also serve a second function as a defense mechanism: they orchestrate the degradation of foreign RNA molecules, specifically those that occur in double-strand form. Many transposable elements and viruses produce double-stranded RNA at least transiently in their life cycles, and RNA interference helps to keep these potentially dangerous invaders in check. As we shall see, this form of RNAi also provides scientists with a powerful experimental technique to turn off the expression of individual genes.

The presence of double-stranded RNA in the cell triggers RNAi by attracting a protein complex containing *Dicer*, the same nuclease that processes miRNAs (see Figure 7–78). This protein cleaves the double-stranded RNA into small fragments (of approximately 23 nucleotide pairs) called **small interfering RNAs (siRNAs)**.

These double-stranded siRNAs are then bound by Argonaute and other components of RISC. As we saw earlier for miRNAs, one strand of the duplex RNA is then cleaved by Argonaute and discarded. The single-strand siRNA molecule that remains directs RISC back to complementary RNA molecules produced by the virus or transposable element. Because the match is usually exact, Argonaute also cleaves these molecules, leading to their rapid destruction.

Each time RISC cleaves a new RNA molecule, the RISC is released; thus, as we saw for miRNAs, a single RNA molecule can act catalytically to destroy many complementary RNAs. Some organisms employ an additional mechanism that amplifies the RNAi response even further. In these organisms, RNA-dependent RNA polymerases use siRNAs as primers to produce additional copies of double-stranded RNAs that are then cleaved into siRNAs. This amplification ensures that, once initiated, RNA interference can continue even after all the initiating double-stranded RNA has been degraded or diluted out. For example, it permits progeny cells to continue carrying out the specific RNA interference that was provoked in the parent cells.

In some organisms, the RNA interference activity can be spread by the transfer of RNA fragments from cell to cell. This is particularly important in plants (whose cells are linked by fine connecting channels, as discussed in Chapter 19), because it allows an entire plant to become resistant to an RNA virus after only a few of its cells have been infected. In a broad sense, the RNAi response resembles certain aspects of the animal immune system; in both, an invading organism elicits a customized response, and—through amplification of the "attack" molecules—the host becomes systemically protected.

We have seen that although miRNAs and siRNAs are generated in different ways, they rely on some of the same proteins and seek out their targets in a fundamentally similar manner. Because siRNAs are found in widespread species, they are believed to be the most ancient form of RNA interference, with miRNAs being a later evolutionary refinement. The siRNA-mediated defense mechanisms are especially crucial for plants, worms, and insects. In mammals, a protein-based immune system (described in Chapter 24) has largely taken over the task of fighting off viruses.

RNA Interference Can Direct Heterochromatin Formation

The siRNA interference pathway just described does not necessarily stop with the inactivation of target RNA molecules. In some cases, the RNA interference machinery can also selectively shut off the *synthesis* of the target RNAs. For this to occur, the short siRNAs produced by the Dicer protein are assembled with a group of proteins (including Argonaute) to form an RITS (RNA-induced transcriptional silencing) complex. Using single-stranded siRNA as a guide sequence, this complex binds complementary RNA transcripts as they emerge from a transcribing RNA polymerase II (**Figure 7–80**). Positioned on the genome in this way, the RITS complex then attracts enzymes that covalently modify nearby histones and DNA causing the formation of a "constitutive" form of heterochromatin. As described in Chapter 4, this form of heterochromatin is distinguished by the H3K9me3 mark, and, in many cases, it also includes DNA methylation (see Figure 7–48). Although low levels of transcription probably persist (and may be important to continually signal where the heterochromatin should be formed), this form of heterochromatin, as we have seen, is generally resistant to transcription and effectively shuts off the genes that lie within it. In some cases, an RNA-dependent RNA polymerase and a Dicer enzyme are also recruited by the RITS complex to continually generate additional siRNAs *in situ*. This positive feedback loop ensures continued repression of the target gene even after the original, initiating siRNA molecules have disappeared.

RNAi-directed heterochromatin formation is an especially important cell defense mechanism; it limits the spread of transposable elements in genomes by maintaining their DNA sequences in a transcriptionally silent form. However, this same mechanism is also used in some normal processes in the cell. For example,

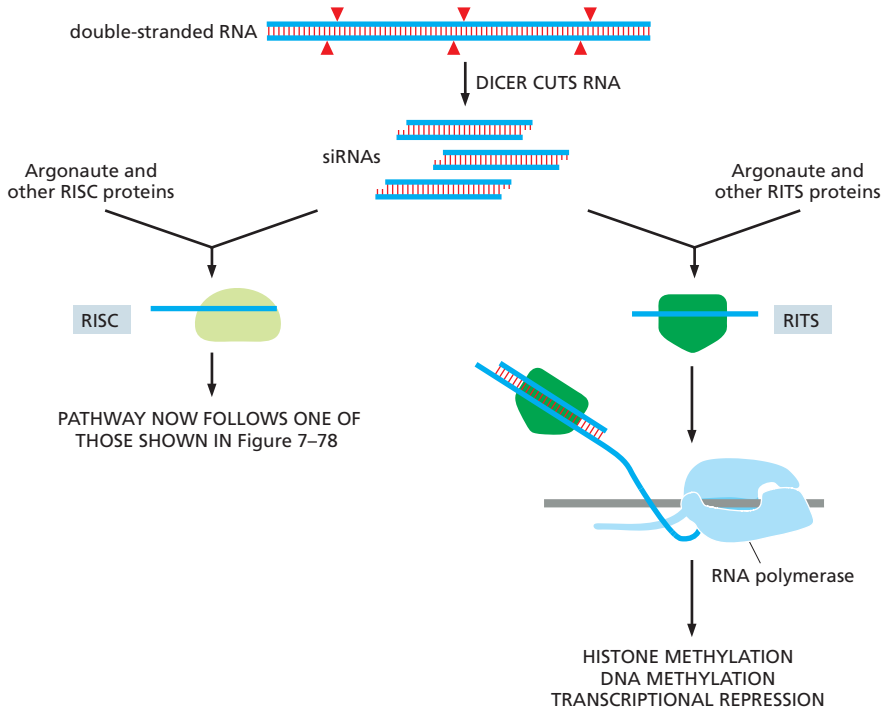


Figure 7–80 RNA interference directed by siRNAs. In many organisms, double-stranded RNA can trigger both the destruction of complementary mRNAs *(left)* and transcriptional silencing *(right).* The change in chromatin structure induced by the bound RITS (RNA-induced transcriptional silencing) complex resembles that of Figure 7–48.

in many organisms, the RNA interference machinery maintains the heterochromatin formed around centromeres. Centromeric DNA sequences are transcribed in both directions, producing complementary RNA transcripts that can base-pair to form double-stranded RNA. This double-stranded RNA triggers the RNA interference pathway and stimulates formation of the heterochromatin that surrounds centromeres, which is necessary for the centromeres to segregate chromosomes accurately during mitosis.

piRNAs Protect the Germ Line from Transposable Elements

A third system of RNA interference relies on **piRNAs** (**piwi-interacting RNAs**, named for Piwi, a class of proteins related to Argonaute). piRNAs are found in many organisms, and they carry out a diverse set of functions. Here, we describe one of their most important roles, which is to hold transposable elements (transposons) in check in the germ line of animals. The germ line is especially susceptible to transposon movement because many of the histone modifications and methylated DNA sites are "erased" during gametogenesis, temporarily releasing transposons from their normal constraints. piRNAs cover this vulnerability. Unlike miRNAs and siRNAs, they are synthesized from specialized piRNA "clusters" in the genome as long, single-strand RNA molecules that are then broken up and trimmed by specialized processing enzymes (different from the Dicer enzymes discussed earlier) into fragments that are slightly longer than miRNAs and siRNAs. These RNAs are covalently modified at their 3′ ends by a 2′-*O*-methyl group (see Figure 6–43A) and assembled with Piwi proteins. Once complexed with their proteins, piRNAs seek out RNA targets by complementary base-pairing and, much like siRNAs, they both cleave the complementary RNAs and package the DNA on which they are being transcribed into repressive forms of chromatin. The piRNA clusters in the genome are rich with sequence fragments from transposons, and the piRNAs attack any transposon whose sequence is represented in the piRNA cluster. In this way, the genome contains a "hit list" of transposons that need to be inactivated during the vulnerable period of gametogenesis. It has been proposed that piRNA clusters are unusually attractive landing sites for transposons and, for this reason, they carry a record of all past bursts of transposon activity.

Although the genome carries a linear record of transposons to be inactivated, piRNAs have an additional way to attack those transposons that are most active during gametogenesis. In brief, once a piRNA and its associated proteins cleave a complementary, transposon-coded mRNA, additional piRNAs can be created from nearby sequences in the transposon mRNA. This mechanism not only amplifies the original response but extends its breadth by incorporating additional sequence information from active transposons, information that might not be carried in the piRNA clusters themselves.

Many mysteries surround piRNAs. More than a million piRNA species are coded in the genomes of many mammals and expressed in the testes, yet only a fraction seem to be directed against the transposons present in those genomes. Are the other piRNAs remnants of past invaders? Do they cover so much "sequence space" that they are broadly protective for any foreign DNA? Another curious feature of piRNAs is that many of them (particularly if base-pairing does not have to be perfect) should, in principle, attack the normal mRNAs made by the organism, yet they do not. It has been proposed that these large numbers of piRNAs may form a system to distinguish "self" RNAs from "foreign" RNAs and attack only the latter. If this is the case, there must be a special way for the cell to spare its own RNAs. One idea is that RNAs produced in the previous generation of an organism are somehow registered and set aside from piRNA attack in subsequent generations. Another idea holds that all legitimate mRNAs carry specialized sequences that spare them from attack. Whether or not this mechanism truly exists, and, if so, how it might work, are questions that demonstrate our incomplete understanding of the full range of RNA interference.

RNA Interference Has Become a Powerful Experimental Tool

Although it likely first arose in evolution as a defense mechanism against viruses and transposable elements, RNA interference, as we have seen, has become thoroughly integrated into many aspects of normal cell biology, ranging from the control of gene expression to a fine tuning of chromosome structure. RNA interference has also been developed by scientists into a powerful experimental tool that allows almost any gene to be inactivated by evoking an RNAi response to it. This technique, which can be readily carried out in cultured cells and, in many cases, whole animals and plants, has made possible new genetic approaches in cell and molecular biology. We shall discuss it in detail in the following chapter when we cover the modern genetic methods used to study cells (see pp. 533–534). RNAi also has potential in treating human disease. Because many human disorders result from the misexpression of genes, the ability to turn these genes off by experimentally introducing complementary siRNA molecules into cells holds great medical promise. Although delivery of RNA molecules to the appropriate tissue has been a persistent problem in using RNAi as a human therapy, the strategy is currently used to treat a rare disease called transthyretin amyloidosis. This inherited disease, which affects heart and nerve function, is caused by the accumulation of a mutated protein, and siRNAs directed by complementary base-pairing to the mutated mRNA relieve its symptoms. In this case, the siRNAs are delivered to the liver (the key site of synthesis of the mutated protein) by a special combination of lipids that forms tiny vesicles to encase the siRNA.

Cells Have Additional Mechanisms to Hold Transposons and Integrated Viral Genomes in Check

From the preceding sections, it should be clear that cells are locked in an eternal "arms race" with parasitic DNA elements, such as transposons and viruses. Indeed, it seems that our own genome came close to being overrun with such elements; even with our many defense mechanisms, they still make up nearly half our DNA (see Figure 4–63). Most of these elements have accumulated mutations

that prevent them from being active, but this process likely occurred after host-cell mechanisms came into play to hold them in check.

We have seen how siRNAs and piRNAs constitute surveillance systems to monitor transcription from transposable elements, to destroy their transcripts, and to package their DNA into repressive forms of chromatin. Although these overlapping defense mechanisms may seem highly effective, cells have at least one additional strategy for recognizing transposons and integrated viruses and silencing them. In contrast to the RNA-based strategies, which utilize complementary base-pairing to recognize these genome invaders, this additional system employs a special set of sequence-specific DNA-binding proteins to monitor the genome. When these proteins recognize a DNA sequence present in a transposon or integrated viral genome, they bind directly to that sequence and recruit both histone "writers" that place H3K9me3 marks on nearby histones and DNA methylases that heavily methylate the surrounding DNA. As discussed earlier in the chapter, this repressive form of chromatin can then spread and render the underlying DNA resistant to transcription and recombination (see Figure 7–48). Our genome codes for hundreds of different sequence-specific DNA proteins that carry out this surveillance (called KRAB-ZPF proteins), and they cover a wide variety of transposable element DNA and viral sequences. Most recognize a DNA sequence that is crucial for that element to transpose (or in the case of integrated viral genomes, for the virus to multiply), making it difficult for the element to escape through a mutation. However, such escape does apparently occur, because KRAB-ZPF proteins are evolving rapidly (compared with other human genes), and they appear to be "keeping up" with mutated versions of resident transposable elements. Their rapid evolution also suggests that the KRAB-ZPF proteins can easily adapt through mutation to attack new parasitic elements that might enter the genome.

Transposable elements, if left unchecked, present many challenges to the cell: their sequences can serve as recombination sites leading to crossovers between nonhomologous chromosomes, double-strand DNA breaks are produced in the host genome when they move, and they can disrupt coding or regulatory sequences when they insert into a new position. On the other hand, their movement has provided a source of variation that is necessary for natural selection to occur. But the many different strategies host cells have evolved to neutralize these invaders suggest that the short-term dangers must far outweigh any long-term advantages.

Bacteria Use Small Noncoding RNAs to Protect Themselves from Viruses

In the previous sections, we emphasized the defense systems of animals and plants, but it is important to keep in mind that bacteria and archaea make up the vast majority of Earth's diversity. Not surprisingly, the viruses that infect these single-cell organisms greatly outnumber plant and animal viruses. Many species of bacteria (and almost all species of archaea) use a repository of small noncoding RNA molecules to seek out and destroy invading viruses. Many features of this defense mechanism, known as **CRISPR**, resemble those of miRNAs, siRNAs, and piRNAs that we saw earlier. When bacteria and archaea are first infected by a virus, short fragments of that viral DNA become integrated into their genomes by a process that is only beginning to be understood. These serve as "vaccinations," in the sense that they become the templates for producing small noncoding RNAs known as **crRNAs** (CRISPR RNAs) that will thereafter destroy the virus should it reinfect the descendants of the original cell. This aspect of the CRISPR system resembles both human adaptive immunity and piRNA-based surveillance, insofar as the cell carries a record of past exposures that is used to protect against future exposures.

In most cases, crRNAs associate with special proteins that allow them to seek out and destroy invading viral genomes, which are typically composed of double-stranded DNA. Many distinct CRISPR systems exist across different species of

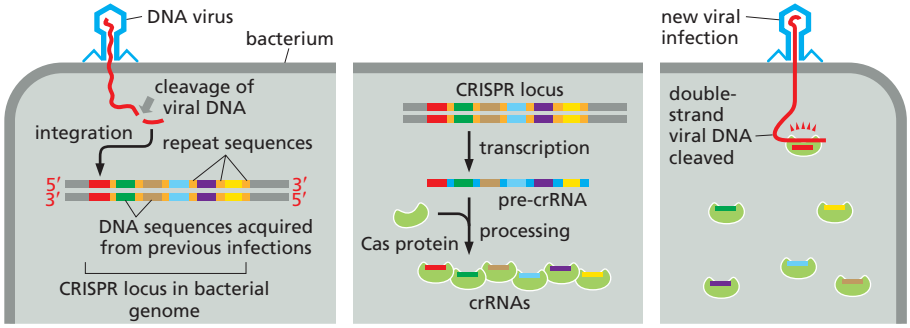


Figure 7–81 CRISPR-mediated immunity in bacteria and archaea. After infection by a virus *(left panel)*, a small bit of DNA from the viral genome is inserted into the CRISPR locus. For this to happen, a small fraction of infected cells must survive the initial viral infection. The surviving cells, or more generally their descendants, transcribe the CRISPR locus and process the transcript into crRNAs *(middle panel)*. Upon reinfection with a virus that the population has already been "vaccinated" against, the incoming viral DNA is destroyed by a complementary crRNA *(right panel)*.

For a CRISPR system to be effective, the crRNAs must not destroy the CRISPR locus itself, even though the crRNAs are complementary in sequence to it. How is this possible? In many species, there must be additional short nucleotide sequences carried by the target molecule in order for crRNAs to attack it. Because these sequences, known as PAMs (protospacer adjacent motifs), lie outside the crRNA sequences, the host CRISPR locus is spared (see Figure 8–57).

bacteria and archaea. Here we merely outline one of the most common and best understood, describing its three steps (**Figure 7–81**). In the first step, viral DNA sequences are integrated into special regions of the bacterial genome known as CRISPR (clustered regularly interspersed short palindromic repeat) loci, named for the peculiar DNA sequences that first drew the attention of scientists. In its simplest form, a CRISPR locus consists of several hundred repeats of a host DNA sequence interspersed with a large collection of DNA sequences (typically 25–70 nucleotide pairs each) derived from prior exposures to viruses and other foreign DNA. The newest viral sequence is always integrated at the 5′ end of the CRISPR locus, the end that is transcribed first. Each locus, therefore, carries a temporal, ordered record of prior infections. Many bacterial and archaeal species carry several large CRISPR loci in their genomes and are thus immune to a wide range of viruses.

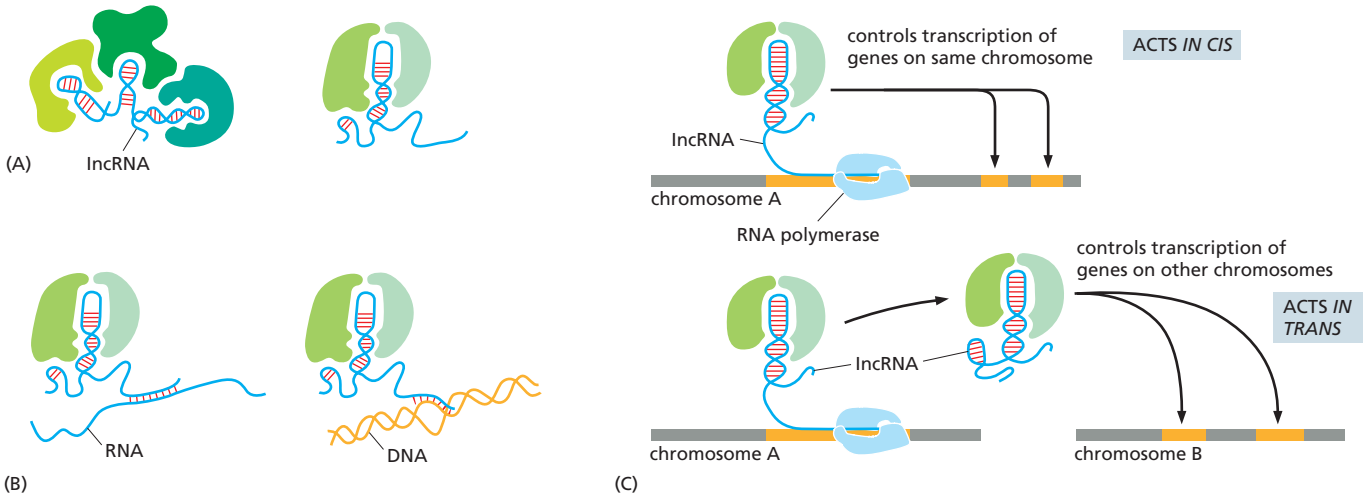
In the second step, the CRISPR locus is transcribed to produce a long RNA molecule, which is then processed into the much shorter (approximately 30 nucleotides) crRNAs. These crRNAs become complexed with *Cas* (*CRISPR-associated*) proteins, and, in the final step, they seek out complementary viral DNA sequences and direct their destruction by nucleases. Although structurally dissimilar, Cas proteins are analogous to the Argonaute and Piwi proteins discussed earlier: they hold small single-stranded RNAs in an extended configuration that is optimized, in this case, for seeking and forming complementary base pairs with double-stranded DNA.

We still have much to learn about CRISPR-based immunity in bacteria and archaea. For example, the mechanism through which viral sequences are first identified and integrated into the host genome is poorly understood. Moreover, in different species of bacteria and archaea, crRNAs are processed in different ways, and in some cases, the crRNAs can attack viral RNAs as well as DNAs. As might be predicted, many viruses have evolved anti-CRISPR systems to counteract the defense systems of their hosts. These anti-CRISPRs range from viral proteins that bind to and inactivate the Cas proteins to special coats that form around the viral DNA and protect it from CRISPR attack during replication, gene expression, and virus assembly.

In Chapter 8, we describe how bacterial CRISPR systems have been artificially "moved" into plants and animals, where they have revolutionized our ability to manipulate genomes.

Long Noncoding RNAs Have Diverse Functions in the Cell

In this and the preceding chapters, we have seen that noncoding RNA molecules have many functions in the cell. Yet there remain many noncoding RNAs whose functions are still unknown. Many of these RNAs belong to a group known as **long noncoding RNA (lncRNA)**, arbitrarily defined as RNAs longer than 200 nucleotides that do not code for protein. The sheer number of lncRNAs (an estimated 5000 for the human genome, for example) came as a surprise to scientists. Most of these lncRNAs are transcribed by RNA polymerase II and have 5′ caps and



poly-A tails, and, in many cases, they are spliced. It has been difficult to accurately annotate lncRNAs, in part because low levels of RNA are now known to be made from about 75% of the human genome. Most of these RNAs are thought to result from a background "noise" of leaky transcription, and they are rapidly degraded. According to this idea, such nonfunctional RNAs provide no fitness advantage or disadvantage to the organism and are a tolerated by-product of the complex patterns of gene expression that need to be produced in multicellular organisms. For these reasons, it is difficult to estimate the number of lncRNAs that are likely to have a function in the cell and to distinguish them from the background of transcription noise.

In terms of biological function, lncRNA should be considered a catch-all phrase encompassing a great diversity of functions. We have already encountered a few notable lncRNAs, including the RNA in telomerase (see Figure 5–33), Xist RNA (see Figure 7–55), and an RNA involved in imprinting (see Figure 7–52). Other lncRNAs have been implicated in controlling the enzymatic activity of proteins, inactivating transcription regulators, affecting splicing patterns, and blocking translation of certain mRNAs through complementary base-pairing. However, there are three unifying features of lncRNAs that can account for their many roles in the cell. The first is that they can function as *scaffold RNA molecules*, holding together groups of proteins to coordinate their functions (**Figure 7–82A**; see also Figure 7–21). We have already seen examples in telomerase, the ribosome, and X-inactivation, where an RNA molecule holds together and organizes protein components. These RNA-based scaffolds are analogous to protein scaffolds we discussed in Chapter 3 (see Figure 3–76). RNA molecules are well suited to act as scaffolds: small bits of RNA sequence, often those portions that form stem–loop structures, can serve as binding sites for proteins, and these can be strung together with random sequences of RNA in between. This property may be one reason that many lncRNAs show relatively little primary-sequence conservation across species.

A second key feature of lncRNAs is their ability to serve as guide sequences, binding to specific RNA or DNA target molecules through base-pairing. By doing so, they bring proteins that are bound to them into close proximity with the DNA and RNA sequences (**Figure 7–82B**). This behavior is similar to that of snoRNAs (see Figure 6–43), miRNAs (see Figure 7–78), siRNAs (see Figure 7–80), and crRNAs (see Figure 7–81), all of which act in this way to guide protein enzymes to specific nucleic acid sequences. A third characteristic of RNA in general is its ability to organize biomolecular condensates, the non-membrane-bound assemblies of proteins and nucleic acids discussed in this and previous chapters. For example, rRNA is crucial for formation of the nucleolus, and untranslated mRNA provides the framework for P-bodies and stress granules. The propensity of RNA to form

Figure 7–82 Roles of long noncoding RNA (lncRNA). (A) As described in Chapter 6, RNAs can fold into short, specific three-dimensional structures whose specific features can be recognized by proteins. Thus, lncRNAs can serve as scaffolds, bringing together proteins that function together in the same process and thereby facilitating their interactions and speeding the reactions that they catalyze. (B) lncRNAs can also, through formation of complementary base pairs, localize the proteins that they bind near specific nucleotide sequences on RNA or DNA molecules. (C) In some cases, lncRNAs act only *in cis* at their sites of synthesis—as, for example, when the RNA is held in place by the RNA polymerase that produced them *(top)*. But as shown, other lncRNAs diffuse from their sites of synthesis and are said to act *in trans*.

condensates derives in part from its ability to bind multiple proteins, as discussed above, but also because many RNAs can form multiple weak intramolecular interactions that, on their own, can lead to condensation. Some lncRNAs are thought to function solely by organizing and driving formation of such condensates.

In some of the simplest cases, lncRNAs work simply by base-pairing, without bringing in enzymes or other proteins. For example, a number of lncRNA genes are embedded in protein-coding genes, but they are transcribed in the "wrong direction." These *antisense RNAs* can form complementary base pairs with the mRNA (transcribed in the "correct" direction) and block its translation into protein (see Figure 7–70D). Other antisense lncRNAs base-pair with pre-mRNAs as they are synthesized and change the pattern of RNA splicing by masking the preferred splice-site sequences. Still others act as "sponges," base-pairing with miRNAs and thereby reducing their effects.

Finally, we note that some lncRNAs act only *in cis*; that is, they affect only the chromosome from which they are transcribed. This readily occurs when the transcribed RNA has not yet been released from RNA polymerases (**Figure 7–82C**) or when the completed RNA molecule does not diffuse away from the chromosome as for the case of Xist (see Figure 7–55). Many lncRNAs, however, leave their site of synthesis and act *in trans*. Although the best-understood lncRNAs work in the nucleus, many are found in the cytosol. The functions—if any—of the great majority of these cytosolic lncRNAs remain undiscovered.

Summary

RNA molecules have many uses in the cell besides carrying the information needed to specify the order of amino acids during protein synthesis. Although we have encountered noncoding RNAs in other chapters (tRNAs, rRNAs, snoRNAs, for example), the sheer number of noncoding RNAs produced by cells has surprised scientists. One well-understood use of noncoding RNAs occurs in RNA interference, where guide RNAs (miRNAs, siRNAs, piRNAs) base-pair with mRNAs. RNA interference can cause mRNAs to be either destroyed or translationally repressed. It can also cause specific genes to become packaged into heterochromatin suppressing their transcription. In bacteria and archaea, RNA interference is used as an adaptive immune response to destroy viruses that infect them. A large family of large noncoding RNAs (lncRNAs) has recently been discovered through detailed genomic analyses. Although the function (if any) of most of these RNAs is unknown, some serve as RNA scaffolds to bring specific proteins and RNA molecules together to speed up needed reactions.

PROBLEMS

Which statements are true? Explain why or why not.

7–1 When the nucleus of a fully differentiated carrot cell is injected into a frog egg whose nucleus has been removed, the injected donor nucleus is capable of programming the recipient egg to produce a normal carrot.

7–2 In terms of the way it interacts with DNA, the helix–loop–helix motif is more closely related to the leucine zipper motif than it is to the helix–turn–helix motif.

7–3 Many transcription regulators in eukaryotes can act even when they are bound to DNA thousands of nucleotide pairs away from the promoter they influence.

7–4 Once cells have differentiated to their final specialized forms, they never again alter expression of their genes.

7–5 CG islands are thought to have arisen during evolution because they were associated with portions of the genome that remained unmethylated in the germ line.

7–6 In one extreme case, a single gene in *Drosophila*—the *Dscam* gene—has the potential to produce more than 38,000 different proteins by alternative splicing; thus, the complexity of this one gene rivals the complexity of the whole human genome.

7–7 crRNAs in bacteria and piRNAs in animals serve analogous functions; they defend against foreign invaders.

Discuss the following problems.

7–8 Comparisons of the patterns of mRNA abundance across different human cell types show that the level of expression of almost every active gene is different. The patterns of mRNA abundance are so characteristic of cell type that they can be used to determine the tissue of origin of cancer cells, even though the cells may have metastasized to different parts of the body. By definition, however, cancer cells are different from their noncancerous precursor cells. How do you suppose then that patterns of mRNA expression might be used to determine the tissue source of a human cancer?

7–9 What are the two fundamental components of a genetic switch?

7–10 The nucleus of a eukaryotic cell is much larger than a bacterium, and it contains much more DNA. As a consequence, a transcription regulator in a eukaryotic cell must be able to select its specific binding site from among many more unrelated sequences than does a transcription regulator in a bacterium. Does this present any special problems for eukaryotic gene regulation?

Consider the following situation. Assume that the eukaryotic nucleus and the bacterial cell each have a single copy of the same DNA binding site. In addition, assume that the nucleus is 500 times the volume of the bacterium and has 500 times as much DNA. If the concentration of the transcription regulator that binds the site were the same in the nucleus and in the bacterium, would the regulator occupy its binding site equally as well in the eukaryotic nucleus as it does in the bacterium? Explain your answer.

7–11 The genes encoding the enzymes for arginine biosynthesis are located at several positions around the genome of *E. coli*. The ArgR transcription regulator coordinates their expression. The activity of ArgR is modulated by arginine. Upon binding arginine, ArgR dramatically changes its affinity for the *cis*-regulatory sequences in the promoters of the genes for the arginine biosynthetic enzymes. Given that ArgR is a transcription repressor, would you expect that ArgR would bind more tightly or less tightly to the regulatory sequences when arginine is abundant? If ArgR functioned instead as a transcription activator, would you expect the binding of arginine to increase or to decrease its affinity for its regulatory sequences? Explain your answers.

7–12 Some transcription regulators bind to DNA and cause the double helix to bend at a sharp angle. Such "bending proteins" can affect the initiation of transcription without directly contacting any other protein. Can you devise a plausible explanation for how such proteins might work to modulate transcription? Draw a diagram that illustrates your explanation.

7–13 How is it that protein–protein interactions that are too weak to cause proteins to assemble in solution can nevertheless allow the same proteins to assemble into complexes on DNA?

7–14 Imagine the two situations shown in **Figure Q7–1**. In cell 1, a transient signal induces the synthesis of protein A, which is a transcription activator that turns on many genes including its own. In cell 2,

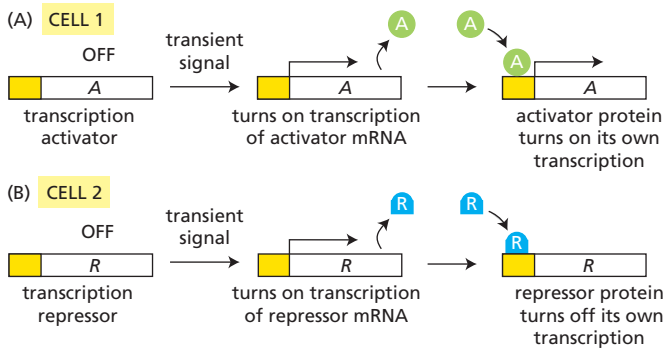


Figure Q7–1 Gene regulatory circuits and cell memory (Problem 7–14). (A) Induction of synthesis of transcription activator A by a transient signal. (B) Induction of synthesis of transcription repressor R by a transient signal.

a transient signal induces the synthesis of protein R, which is a transcription 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.

7–15 Examine the two pedigrees shown in **Figure Q7–2**. One results from deletion of a maternally imprinted autosomal gene. The other pedigree results from deletion of a paternally imprinted autosomal gene. In both pedigrees, affected individuals (*red symbols*) are heterozygous for the deletion. These individuals are affected because one copy of the chromosome carries an imprinted, inactive gene, while the other carries a deletion of the gene. *Dotted yellow symbols* indicate individuals that carry the deleted locus but do not display the mutant phenotype. Which pedigree is based on paternal imprinting and which on maternal imprinting? Explain your answer.

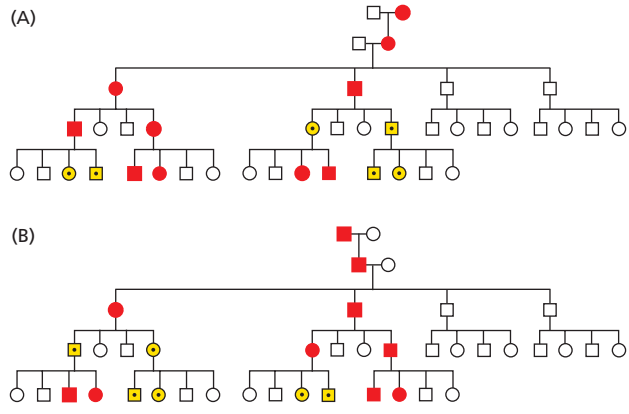


Figure Q7–2 Pedigrees reflecting maternal and paternal imprinting (Problem 7–15). In one pedigree, the gene is paternally imprinted; in the other, it is maternally imprinted. In generations 3 and 4, only one of the two parents in the indicated matings is shown; the other parent is a normal individual from outside this pedigree. Affected individuals are represented by *red circles* for females and *red squares* for males. *Dotted yellow symbols* indicate individuals that carry the deletion but do not display the phenotype.

7–16 To determine the role of the *Xist* gene in X-inactivation, scientists generated embryonic stem cells that carried one normal X chromosome and one mutant X chromosome with a nonfunctional *Xist* gene. Sequence differences allowed them to distinguish the two X chromosomes. What pattern of X-inactivation do you predict was observed in mice derived from these embryonic stem cells? Explain your reasoning.

A. Only the normal X chromosomes were inactivated.

B. Only the mutant X chromosomes were inactivated.

C. None of the X chromosomes were inactivated.

D. The X chromosomes were randomly inactivated.

7–17 The level of β-tubulin gene expression in cells is controlled by an unusual regulatory pathway, in which the intracellular concentration of free tubulin dimers (composed of one α-tubulin and one β-tubulin subunit) regulates the rate of new β-tubulin synthesis at the level of β-tubulin mRNA stability. The first 12 nucleotides of the coding portion of the mRNA were found to contain the site responsible for this autoregulatory control. Because the critical segment of the mRNA involves a coding region, it was not clear whether the regulation of mRNA stability resulted from the interaction of tubulin dimers with the RNA or with the nascent protein. Either interaction might plausibly trigger a nuclease that would destroy the mRNA.

These two possibilities were tested by mutagenizing the regulatory region on a cloned version of the gene. The mutant genes were then expressed in cells, and the stability of their mRNAs was assayed in the presence of excess free tubulin dimers. The results from a dozen mutants that affect the regulatory region of the mRNA are shown in **Figure Q7–3**. Does the regulation of β-tubulin mRNA stability result from an interaction with the RNA or from an interaction with the encoded protein? Explain your reasoning. (The genetic code is inside the back cover of this book.)

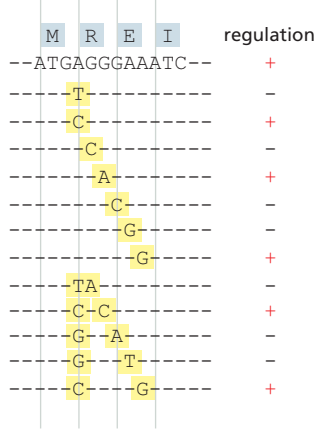


Figure Q7–3 Effects of mutations on the regulation of β-tubulin mRNA stability (Problem 7–17). The wild-type sequence for the first 12 nucleotides of the coding portion of the gene is shown at the top, and the first four amino acids beginning with methionine (M) are indicated above the codons. The nucleotide changes in the 12 mutants are shown below; only the altered nucleotides are indicated. Regulation of mRNA stability is shown on the right: + indicates wild-type response to changes in intracellular tubulin concentration, and – indicates no response to changes. Vertical lines mark the position of the first nucleotide in each codon.

7–18 If you insert a *β-galactosidase* gene lacking its own transcription control region into a cluster of piRNA genes in *Drosophila*, you find that β-galactosidase expression from a normal copy elsewhere in the genome is strongly inhibited in the fly's germ cells. If the inactive *β-galactosidase* gene is inserted outside the piRNA gene cluster, the normal gene is properly expressed. What do you suppose is the basis for this observation? How would you test your hypothesis?

REFERENCES

General

Barressi MJF & Gilbert SF (2020) Developmental Biology, 12th ed. Sunderland, MA: Sinauer Associates.

Brown TA (2017) Genomes 4. New York: Garland Science.

Craig N, Green R, Greider C, Storz G, Wolberger C & Cohen-Fix O (2021) Molecular Biology: Principles of Genome Function, 2nd ed. Oxford: Oxford University Press.

Goldberg ML, Fischer J, Hood L & Hartwell L (2021) Genetics: From Genes to Genomes, 7th ed. Boston: McGraw-Hill.

Watson J, Baker T, Bell S et al. (2013) Molecular Biology of the Gene, 7th ed. Menlo Park, CA: Benjamin Cummings.

An Overview of Gene Control

Davidson EH (2006) The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Burlington, MA: Elsevier.

Gurdon JB (1992) The generation of diversity and pattern in animal development. *Cell* 68, 185–199.

Ptashne M & Gann A (2002) Genes and Signals. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Control of Transcription by Sequence-specific DNA-binding Proteins

Gilbert W & Müller-Hill B (1967) The *lac* operator is DNA. *Proc. Natl. Acad. Sci. USA* 58, 2415.

McKnight SL (1991) Molecular zippers in gene regulation. *Sci. Am.* 264, 54–64.

Weirauch MT & Hughes TR (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In A Handbook of Transcription Factors (Hughes TR, ed.), pp. 25–74. New York: Springer.

Transcription Regulators Switch Genes On and Off

Beckwith J (1987) The operon: an historical account. In *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology (Neidhart FC, Ingraham JL, Low KB et al., eds.), Vol. 2, pp. 1439–1443. Washington, DC: ASM Press.

Goldstein I & Hager GL (2018) Dynamic enhancer function in the chromatin context. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 10(1), 10.1002/wsbm.1390.

Hahn S (2018) Phase separation, protein disorder, and enhancer function. *Cell* 175(7), 1723–1725.

Hnisz D, Shrinivas K, Young RA . . . Sharp PA (2017) A phase separation model for transcriptional control. *Cell* 169(1), 13–23.

Jacob F & Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356.

McSwiggen DT, Mir M, Darzacq X & Tjian R (2019) Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes Dev*. 33(23-24), 1619–1634.

Patel AB, Greber BJ & Nogales E (2020) Recent insights into the structure of TFIID, its assembly, and its binding to core promoter. *Curr. Opin. Struct. Biol*. 61, 17–24.

Ptashne M (2004) A Genetic Switch: Phage Lambda Revisited, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Soutourina J (2018) Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol*. 19(4), 262–274.

Molecular Genetic Mechanisms That Create and Maintain Specialized Cell Types

Alon U (2007) Network motifs: theory and experimental approaches. *Nature* 8, 450–461.

Hobert O (2011) Regulation of terminal differentiation programs in the nervous system. *Annu. Rev. Cell Dev. Biol.* 27, 681–696.

Lawrence PA (1992) The Making of a Fly: The Genetics of Animal Design. New York: Blackwell Scientific.

Rickels R & Shilatifard A (2018) Enhancer logic and mechanics in development and disease. *Trends Cell Biol*. 28(8), 608–630.

Mechanisms That Reinforce Cell Memory in Plants and Animals

Brahma S & Henikoff S (2020) Epigenome regulation by dynamic nucleosome unwrapping. *Trends Biochem. Sci*. 45(1), 13–26.

Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev*. 25(10), 1010–1022.

Galupa R & Heard E (2018) X-chromosome inactivation: a crossroads between chromosome architecture and gene regulation. *Annu. Rev. Genet.* 52, 535–566.

Klemm SL, Shipony Z & Greenleaf WJ (2019) Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* 20(4), 207–220.

Ondičová M, Oakey RJ & Walsh CP (2020) Is imprinting the result of "friendly fire" by the host defense system? *PLoS Genet.* 16(4), e1008599.

Schmitz RJ, Lewis ZA & Goll MG (2019) DNA methylation: shared and divergent features across eukaryotes. *Trends Genet.* 35(11), 818–827.

Post-transcriptional Controls

Breaker RR (2018) Riboswitches and translation control. *Cold Spring Harb. Perspect. Biol*. 10(11), a032797.

Chen LL (2020) The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat. Rev. Mol. Cell Biol*. 21(8), 475–490.

Das S, Vera M, Gandin V . . . Tutucci E (2021) Intracellular mRNA transport and localized translation. *Nat. Rev. Mol. Cell Biol.* 22, 483–504.

Gottesman S & Storz G (2011) Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* 3, a003798.

Hershey JWB, Sonenberg N & Mathews MB (2012) Principles of translational control: an overview. *Cold Spring Harb. Perspect. Biol.* 4, a011528.

Kortmann J & Narberhaus F (2012) Bacterial RNA thermometers: molecular zippers and switches. *Nat. Rev. Microbiol.* 10, 255–265.

Schwartz S (2016) Cracking the epitranscriptome. *RNA* 22(2), 169–174.

Tauber D, Tauber G & Parker R (2020) Mechanisms and regulation of RNA condensation in RNP granule formation. *Trends Biochem. Sci*. 45(9), 764–778.

Thompson SR (2012) Tricks an IRES uses to enslave ribosomes. *Trends Microbiol.* 20, 558–566.

Tian B & Manley JL (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol*. 18(1), 18–30.

Yablonovitch AL, Deng P, Jacobson D & Li JB (2017) The evolution and adaptation of A-to-I RNA editing. *PLoS Genet.* 13(11), e1007064.

Regulation of Gene Expression by Noncoding RNAs

Bartel DP (2018) Metazoan microRNAs. *Cell* 173(1), 20–51.

Czech B, Munafò M, Ciabrelli F . . . Hannon GJ (2018) piRNA-guided genome defense: from biogenesis to silencing. *Annu. Rev. Genet.* 52, 131–157.

Fire A, Xu S, Montgomery MK . . . Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391, 806–811.

Geis FK & Goff SP (2020) Silencing and transcriptional regulation of endogenous retroviruses: an overview. *Viruses* 12(8), 884.

Ozata DM, Gainetdinov I, Zoch A . . . Zamore PD (2019) PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet*. 20(2), 89–108.

Statello L, Guo CJ, Chen LL & Huarte M (2021) Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* 22(2), 96–118.

Wiedenheft B, Sternberg SH & Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482, 331–338.