CHAPTER **EIGHT** **8**

Control of Gene Expression

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

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

In mammals, hundreds of different cell types carry out a range of specialized functions that depend upon genes that are switched on in that cell type but not in most others: for example, the β cells of the pancreas

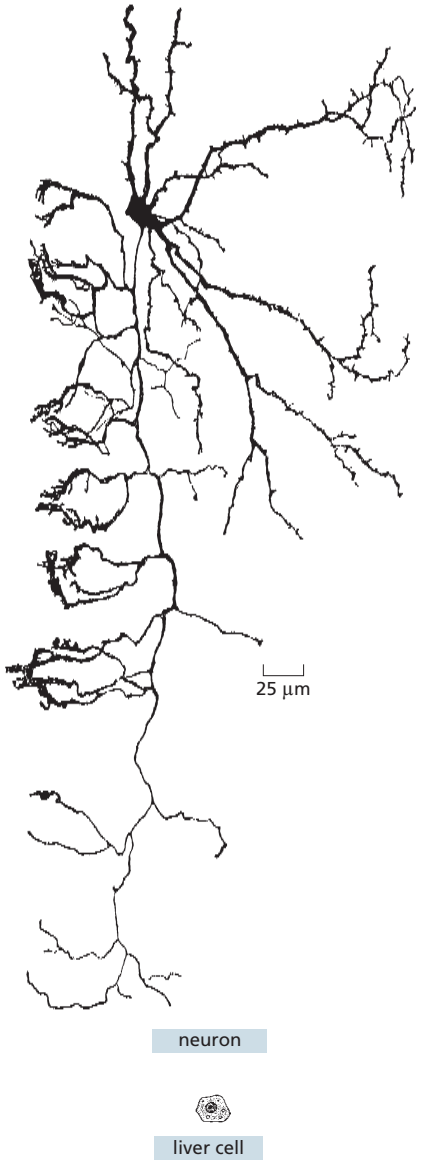

Figure 8–1 A neuron and a liver cell share the same genome. The long branches of this neuron from the retina enable it to receive electrical signals from numerous other neurons and pass these signals along to many neighboring neurons. The liver cell, which is drawn to the same scale, is involved in many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

make the protein hormone insulin, while the α cells of the pancreas make the hormone glucagon; the B lymphocytes of the immune system make antibodies, while developing red blood cells make the oxygen-transport protein hemoglobin. The differences between a neuron, a white blood cell, a pancreatic β cell, and a red blood cell depend on the precise control of gene expression. A typical differentiated cell expresses only about half the genes in its total repertoire. This selection, which differs from one cell type to the next, is the basis for the specialized properties of each cell type.

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

AN OVERVIEW OF GENE EXPRESSION

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

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

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

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

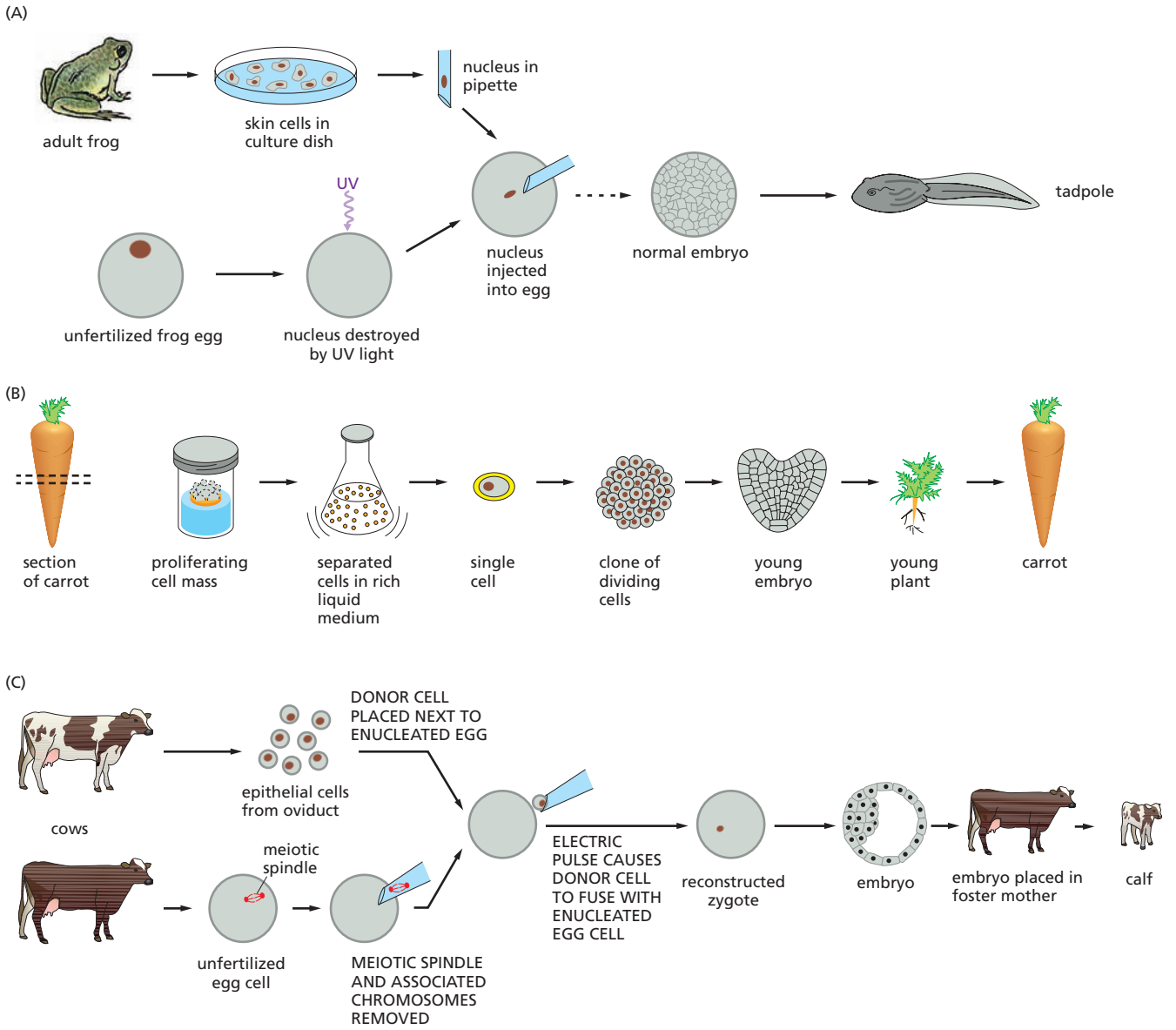

Figure 8–2 Differentiated cells contain all the genetic instructions needed to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an "enucleated" egg—one whose nucleus has been destroyed—can give rise to an entire tadpole. The broken arrow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

These experiments all demonstrate that the DNA in specialized cell types of multicellular organisms still contains the entire set of instructions needed to form a whole organism. The various cell types of an organism therefore differ not because they contain different genes, but because they express them differently.

Different Cell Types Produce Different Sets of Proteins

The extent of the differences in gene expression between different cell types may be roughly gauged by comparing the protein composition of cells in liver, heart, brain, and so on. In the past, such analysis

was performed by two-dimensional gel electrophoresis (see Panel 4–5, p. 167). Nowadays, the total protein content of a cell can be rapidly analyzed by a method called mass spectrometry (see Figure 4–56). This technique is much more sensitive than electrophoresis and it enables the detection of proteins that are produced even in minor quantities.

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

Gene expression can also be studied by cataloging a cell's RNA molecules, including the mRNAs that encode protein. The most comprehensive methods for such analyses involve determining the nucleotide sequence of all RNAs made by the cell, an approach that can also reveal the relative abundance of each. Estimates of the number of different mRNA sequences in human cells suggest that, at any one time, a typical differentiated human cell expresses perhaps 5000–15,000 protein-coding genes from a total of about 19,000. And studies of a variety of tissue types confirm that the collection of expressed mRNAs differs from one cell type to the next.

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

Although each cell type in a multicellular organism expresses its own group of genes, these collections are not static. Specialized cells are capable of altering their patterns of gene expression in response to extracellular cues. For example, if a liver cell is exposed to the steroid hormone cortisol, the production of several proteins is dramatically increased. Released by the adrenal gland during periods of starvation, intense exercise, or prolonged stress, cortisol signals liver cells to boost the production of glucose from amino acids and other small molecules. The set of proteins whose production is induced by cortisol includes enzymes such as tyrosine aminotransferase, which helps convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins returns to its resting level.

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

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

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

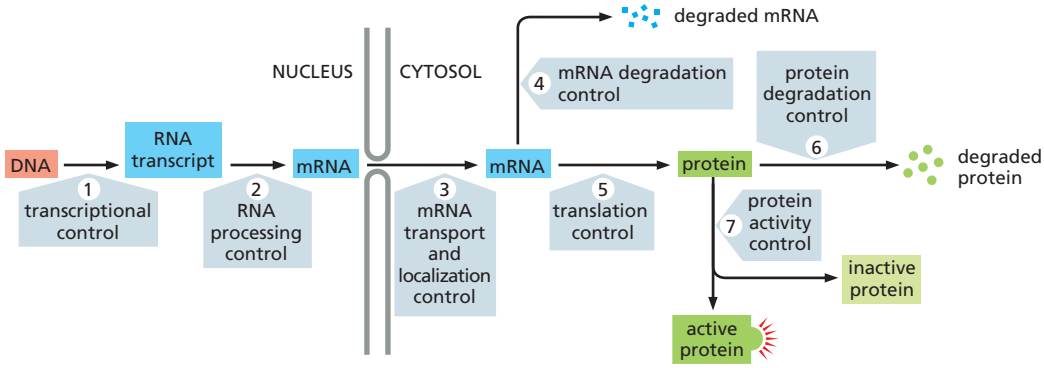

Figure 8–3 Gene expression in eukaryotic cells can be controlled at various steps. Examples of regulation at each of these steps are known, although for most genes the main site of control is step 1: transcription of a DNA sequence into RNA.

degraded, (5) selecting which mRNAs are translated into protein by ribosomes, or (6) regulating how rapidly specific proteins are destroyed after they have been made; in addition, the activity of individual proteins, once they have been synthesized, can be further regulated in a variety of ways.

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

HOW TRANSCRIPTION IS REGULATED

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

Transcription Regulators Bind to Regulatory DNA Sequences

Nearly all genes, whether bacterial or eukaryotic, contain sequences that direct and control their transcription. In Chapter 7, we saw that the **promoter** region of a gene binds the enzyme *RNA polymerase* and correctly orients the enzyme to begin its task of making an RNA copy of the gene. The promoters of both bacterial and eukaryotic genes include a *transcription initiation site*, where RNA synthesis begins, plus nearby sequences that contain recognition sites for proteins that associate with RNA polymerase: sigma factor in bacteria (see Figure 7–9) or the general transcription factors in eukaryotes (see Figure 7–12).

In addition to the promoter, the vast majority of genes include **regulatory DNA sequences** that are used to switch the gene on or off. Some regulatory DNA sequences are as short as 10 nucleotide pairs and act as simple switches that respond to a single signal; such simple regulatory switches predominate in bacteria. Other regulatory DNA sequences, especially those in eukaryotes, are very long (sometimes spanning more than 100,000 nucleotide pairs) and act as molecular microprocessors,

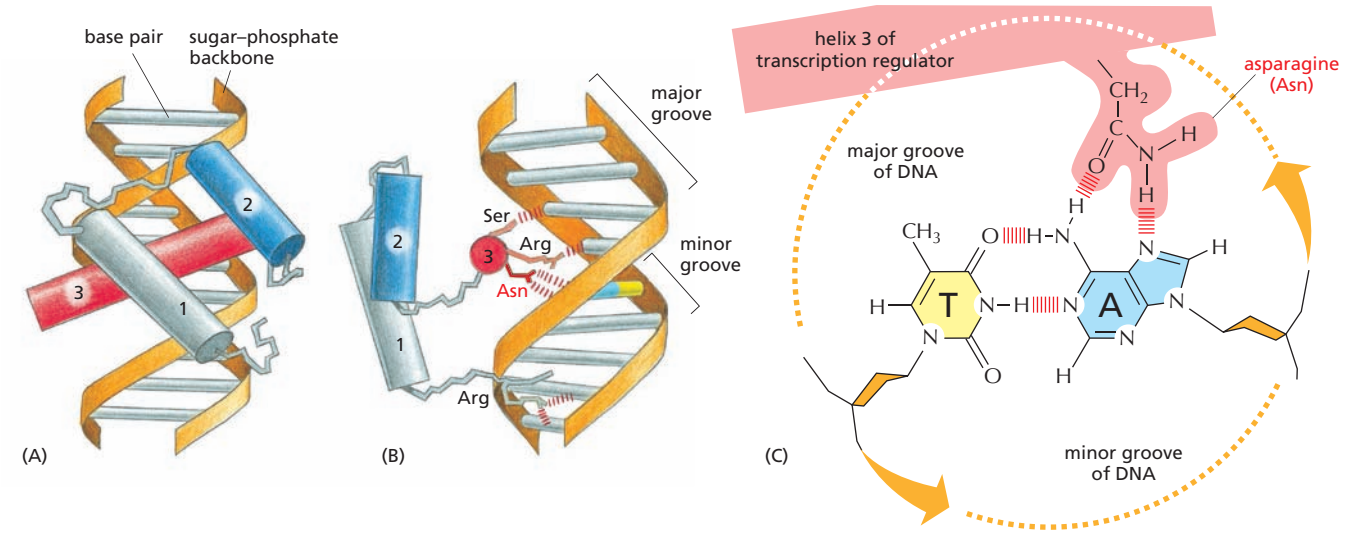

Figure 8–4 A transcription regulator interacts with the DNA double helix. (A) The regulator shown recognizes DNA via three α helices, drawn as numbered cylinders, which allow the protein to fit into the major groove and form tight associations with the base pairs in a short stretch of DNA. This particular structural motif, called a *homeodomain*, is found in many eukaryotic DNA-binding proteins (**Movie 8.1**). (B) Most of the contacts with the DNA bases are made by helix 3 (*red*), which is shown here end-on. (C) An asparagine side chain from helix 3 forms two hydrogen bonds with the adenine in an A-T base pair. The view is end-on, looking down the center of the DNA double helix, and the protein contacts the base pair from the major-groove side. Note that the interactions between the protein and DNA take place along the edges of the nucleotide base and do not disrupt the hydrogen bonds that hold the base pairs together. For simplicity, only one amino acid–base contact is shown; in reality, transcription regulators form hydrogen bonds (as shown here), ionic bonds, and hydrophobic interactions with multiple bases. Most of these contacts occur in the major groove, but some proteins also interact with bases in the minor groove, as shown in (B). Typically, the protein–DNA interface would consist of 10–20 such contacts, each involving a different amino acid and each contributing to the overall strength of the protein–DNA interaction.

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

Regulatory DNA sequences do not work by themselves. To have any effect, these sequences must be recognized by proteins called **transcription regulators**. It is the binding of a transcription regulator to a regulatory DNA sequence that acts as the switch to control transcription. The simplest bacterium produces several hundred different transcription regulators, each of which recognizes a different DNA sequence and thereby regulates a distinct set of genes. Humans make many more—2000 or so—indicating the importance and complexity of this form of gene regulation in the development and function of a complex organism.

Proteins that recognize a specific nucleotide sequence do so because the surface of the protein fits tightly against the surface features of the DNA double helix in that region. Because these surface features will vary depending on the nucleotide sequence, different DNA-binding proteins will recognize different nucleotide sequences. In most cases, the protein inserts into the major groove of the DNA double helix and makes a series of intimate, noncovalent molecular contacts with the nucleotide pairs within the groove (**Figure 8–4, Movie 8.2**). Although each individual contact is weak, the 10 to 20 contacts that typically form at the protein–DNA interface combine to ensure that the interaction is both highly specific and very strong; indeed, protein–DNA interactions are among the tightest and most specific molecular interactions known in biology.

Many transcription regulators bind to the DNA helix as dimers. Such dimerization roughly doubles the area of contact with the DNA, thereby greatly increasing the potential strength and specificity of the protein–DNA interaction (**Figure 8–5, Movie 8.3**).

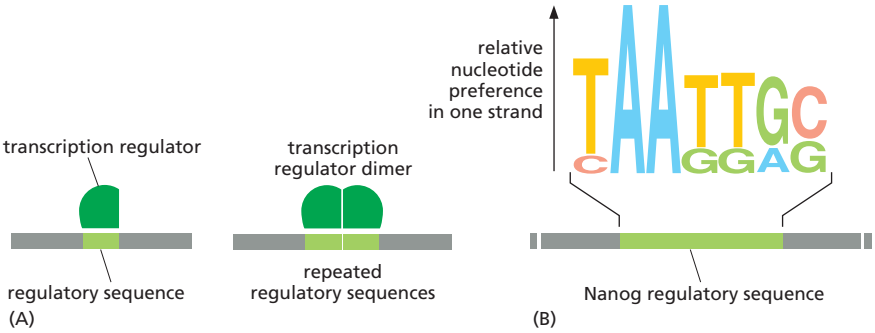

Figure 8–5 Many transcription regulators bind to DNA as dimers. (A) As shown, such dimerization doubles the number of protein–DNA contacts. Here, and throughout the book, regulatory sequences are represented by colored bars; each bar represents a double-helical segment of DNA, as in Figure 8–4. (B) Shown here is a regulatory sequence recognized by Nanog, a homeodomain family member that is a key regulator in embryonic stem cells. This diagram, called a "logo," represents the preferred nucleotide at each position of the sequence; the height of each letter is proportional to the frequency with which this base is found at that position in the regulatory sequence. In the first position, for example, T is found more often than C, while A is the only nucleotide found in the second and third position of the sequence. Although regulatory sequences in the cell are double-stranded, a logo typically shows the sequence of only one DNA strand; the other strand is simply the complementary sequence. Logos are useful because they reveal at a glance the range of DNA sequences to which a given transcription regulator will bind.

Transcription Switches Allow Cells to Respond to Changes in Their Environment

The simplest and best-understood examples of gene regulation occur in bacteria. The genome of the bacterium *E. coli* consists of a single, circular DNA molecule of about 4.6×10^6 nucleotide pairs. This DNA encodes approximately 4300 proteins, although only a fraction of these are made at any one time. Bacteria regulate the expression of many of their genes according to the food sources that are available in the environment. In *E. coli*, for example, five genes code for enzymes that manufacture tryptophan when this amino acid is scarce. These genes are arranged in a cluster on the chromosome and are transcribed from a single promoter as one long mRNA molecule; such coordinately transcribed clusters are called *operons* (**Figure 8–6**). Although operons are common in bacteria (see Figure 7–40), they are rare in eukaryotes, where genes are transcribed and regulated individually.

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

The tryptophan repressor is an allosteric protein (see Figure 4–44): the binding of tryptophan causes a subtle change in its three-dimensional

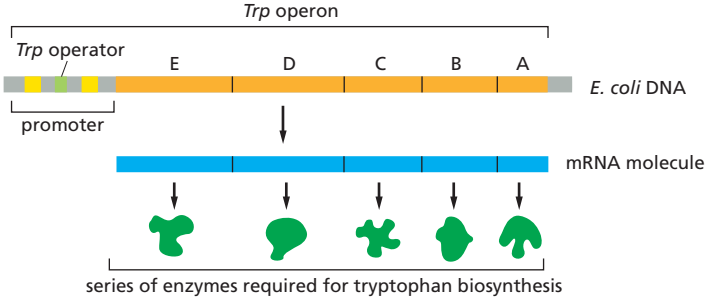

Figure 8–6 A cluster of bacterial genes can be transcribed from a single promoter. Each of these five genes encodes a different enzyme; all of the enzymes are needed to synthesize the amino acid tryptophan from simpler molecular building blocks. The genes are transcribed as a single mRNA molecule, a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

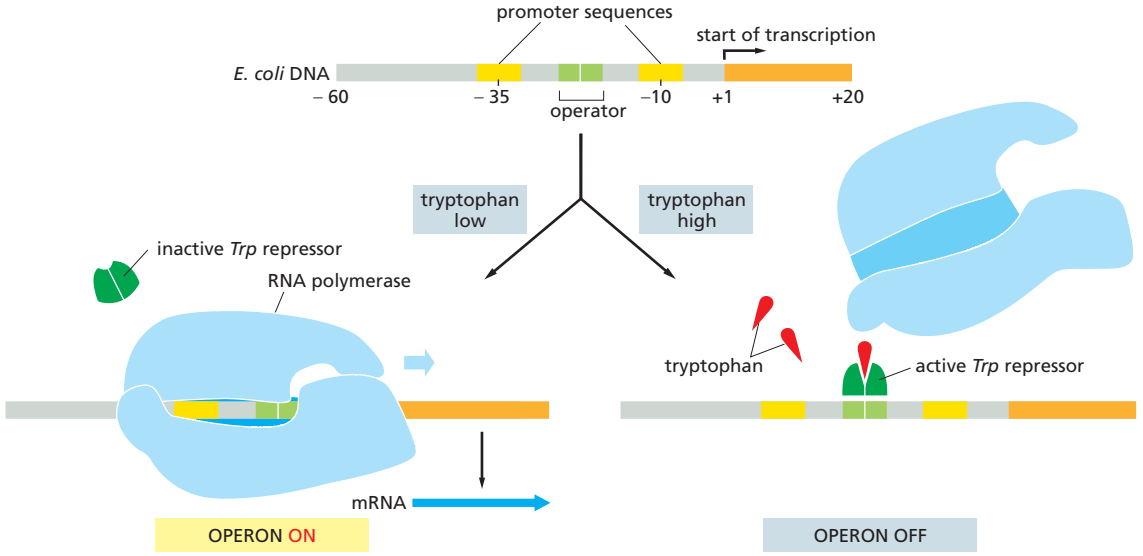

Figure 8–7 Genes can be switched off by repressor proteins. If the concentration of tryptophan inside a bacterium is low (*left*), RNA polymerase (*blue*) binds to the promoter and transcribes the five genes of the tryptophan operon. However, if the concentration of tryptophan is high (*right*), the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

structure so that the protein can bind to the operator sequence. When the concentration of free tryptophan in the bacterium drops, the repressor no longer binds to DNA, and the tryptophan operon is transcribed. The repressor is thus a simple device that switches production of a set of biosynthetic enzymes on and off according to the availability of tryptophan—a form of feedback inhibition (see Figure 4–42).

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

Repressors Turn Genes Off and Activators Turn Them On

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

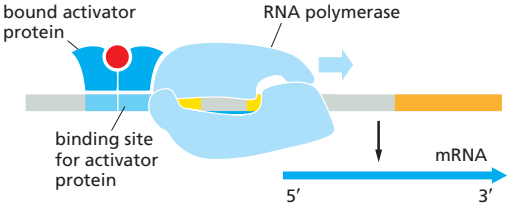

Figure 8–8 Genes can be switched on by activator proteins. An activator protein binds to a regulatory sequence on the DNA and then interacts with the RNA polymerase to help it initiate transcription. Without the activator, the promoter fails to initiate transcription efficiently. In bacteria, the binding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

Like the tryptophan repressor, activator proteins often have to interact with a second molecule to be able to bind DNA. For example, the bacterial activator protein *CAP* has to bind cyclic AMP (cAMP) before it can bind to DNA (see Figure 4–20). Genes activated by CAP are switched on in response to an increase in intracellular cAMP concentration, which occurs when glucose, the bacterium's preferred carbon source, is no longer available; as a result, CAP drives the production of enzymes that allow the bacterium to digest other sugars.

The Lac Operon Is Controlled by an Activator and a Repressor

In many instances, the activity of a single promoter is controlled by two different transcription regulators. The *Lac operon* in *E. coli*, for example, is controlled by both the *Lac repressor* and the CAP activator that we just discussed. The *Lac* operon encodes proteins required to import and digest the disaccharide lactose. In the absence of glucose, the bacterium makes cAMP, which activates CAP to switch on genes that allow the cell to utilize alternative sources of carbon—including lactose. It would be wasteful, however, for CAP to induce expression of the *Lac* operon if lactose itself were not present. Thus the *Lac* repressor shuts off the operon in the absence of lactose. This arrangement enables the control region of the *Lac* operon to integrate two different signals, so that the operon is highly expressed only when two conditions are met: glucose must be absent and lactose must be present (**Figure 8–9**). This 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. None of the other combinations of conditions produce this result.

The elegant logic of the *Lac* operon first attracted the attention of biologists more than 50 years ago. The molecular basis of the switch in *E. coli* was uncovered by a combination of genetics and biochemistry, providing the first insight into how transcription is controlled. In a eukaryotic

QUESTION 8–1

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

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

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

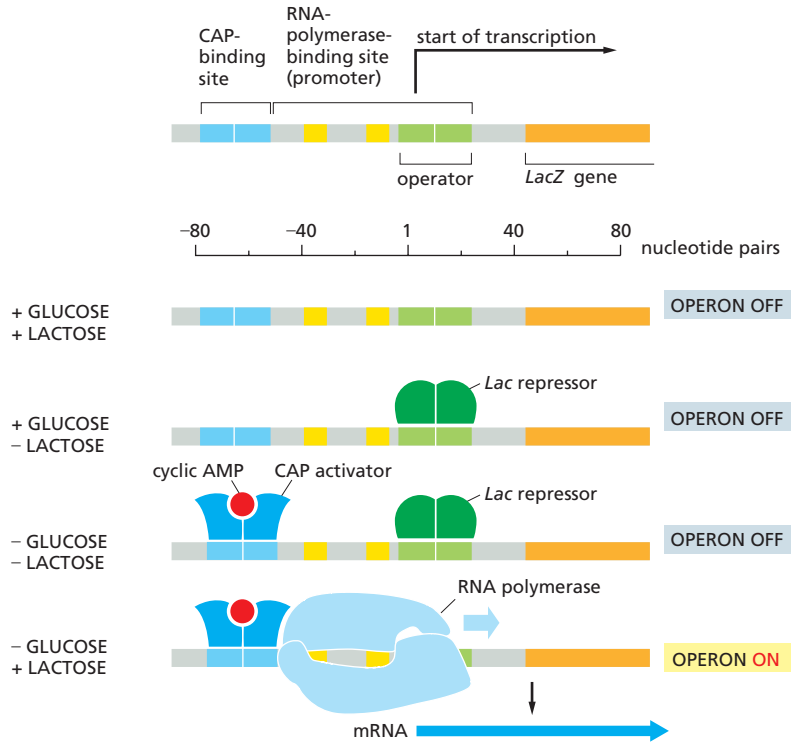

Figure 8–9 The *Lac* operon is controlled by two transcription regulators, the *Lac* repressor and CAP. When lactose is absent, the *Lac* repressor binds to the *Lac* operator and shuts off expression of the operon. Addition of lactose increases the intracellular concentration of a related compound, allolactose; allolactose binds to the *Lac* repressor, causing it to undergo a conformational change that releases its grip on the operator DNA (not shown). When glucose is absent, cyclic AMP (*red* circle) is produced by the cell, and CAP binds to DNA. For the operon to be transcribed, glucose must be absent (allowing the CAP activator to bind) and lactose must be present (releasing the *Lac* repressor). *LacZ*, the first gene of the operon, encodes the enzyme β-galactosidase, which breaks down lactose to galactose and glucose (**Movie 8.4**).

cell, similar transcription regulatory devices are combined to generate increasingly complex circuits, including those that enable a fertilized egg to form the tissues and organs of a multicellular organism.

Eukaryotic Transcription Regulators Control Gene Expression from a Distance

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

Many models for this "action at a distance" have been proposed, but the simplest of these seems to apply in most cases. The DNA between the enhancer and the promoter loops out, bringing the activator protein into close proximity with the promoter (**Figure 8–10**). The DNA thus acts as a tether, allowing a protein that is bound to an enhancer—even one that is thousands of nucleotide pairs away—to interact with the proteins in the vicinity of the promoter (see Figure 7–12). Often, additional proteins serve as adaptors to close the loop; the most important of these is a large complex of proteins known as *Mediator*. Together, all of these proteins ultimately attract and position the general transcription factors and RNA polymerase at the promoter, forming a *transcription initiation complex* (see Figure 8–10). Eukaryotic repressor proteins do the opposite: they decrease transcription by preventing the assembly of this complex.

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

In a eukaryotic cell, the proteins that guide the formation of the transcription initiation complex must also deal with the problem of DNA packaging. As discussed in Chapter 5, eukaryotic DNA is wound around clusters of histone proteins to form nucleosomes, which, in turn, are

QUESTION 8–2

Explain how DNA-binding proteins can make sequence-specific contacts to a double-stranded DNA molecule without breaking the hydrogen bonds that hold the bases together. Indicate how, through such contacts, a protein can distinguish a T-A from a C-G pair. Indicate the parts of the nucleotide base pairs that could form noncovalent interactions—hydrogen bonds, electrostatic attractions, or hydrophobic interactions (see Panel 2–3, pp. 70–71)—with a DNA-binding protein. The structures of all the base pairs in DNA are given in Figure 5–4.

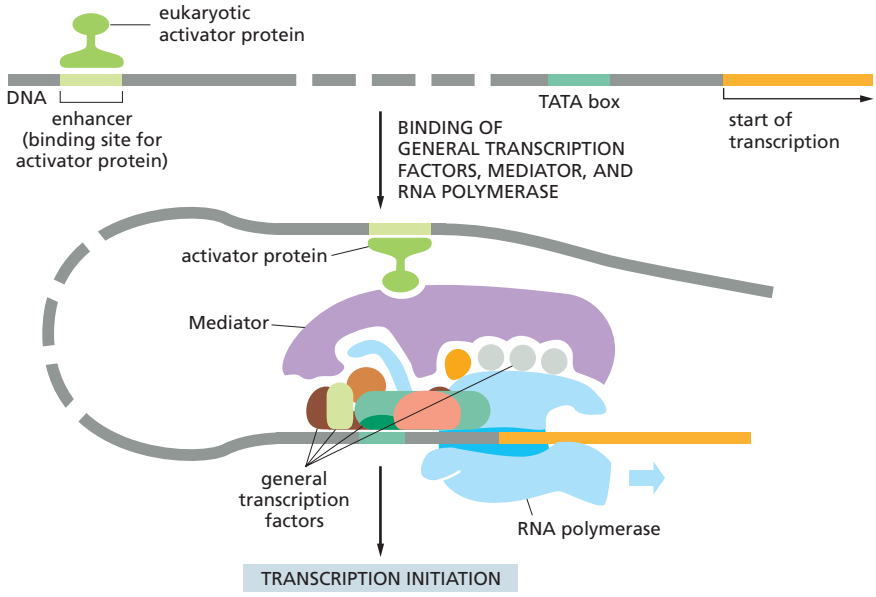

Figure 8–10 In eukaryotes, gene activation can occur at a distance. An activator protein bound to a distant enhancer attracts RNA polymerase and the general transcription factors to the promoter. Looping of the intervening DNA permits contact between the activator and the transcription initiation complex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

folded into higher-order structures. How do transcription regulators, general transcription factors, and RNA polymerase gain access to the underlying DNA? Although some of these proteins can bind efficiently to DNA that is wrapped up in nucleosomes, others are thwarted by these compact structures. More critically, nucleosomes that are positioned over a promoter can inhibit the initiation of transcription by physically blocking the assembly of the general transcription factors and RNA polymerase on the promoter. Such packaging may have evolved in part to prevent leaky gene expression by blocking the initiation of transcription in the absence of the proper activator proteins.

In eukaryotic cells, activator and repressor proteins can exploit the mechanisms used to package DNA to help turn genes on and off. As we saw in Chapter 5, chromatin structure can be altered by *chromatin-remodeling complexes* and by enzymes that covalently modify the histone proteins that form the core of the nucleosome (see Figures 5–24 and 5–25). Many gene activators take advantage of these mechanisms by attracting such chromatin-modifying proteins to promoters. For example, the recruitment of *histone acetyltransferases* promotes the attachment of acetyl groups to selected lysines in the tail of histone proteins; these acetyl groups themselves attract proteins that promote transcription, including some of the general transcription factors (**Figure 8–11**). And the recruitment of chromatin-remodeling complexes makes nearby DNA more accessible. These actions enhance the efficiency of transcription initiation.

In a similar way, gene repressor proteins can modify chromatin in ways that reduce the efficiency of transcription initiation. For example, many repressors attract *histone deacetylases*—enzymes that remove the acetyl groups from histone tails, thereby reversing the positive effects that acetylation has on transcription initiation. Although some eukaryotic repressor proteins work on a gene-by-gene basis, others can orchestrate the formation of large swathes of transcriptionally inactive chromatin. As discussed in Chapter 5, these transcription-resistant regions of DNA include the heterochromatin found in interphase chromosomes and the inactive X chromosome in the cells of female mammals.

QUESTION 8–3

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

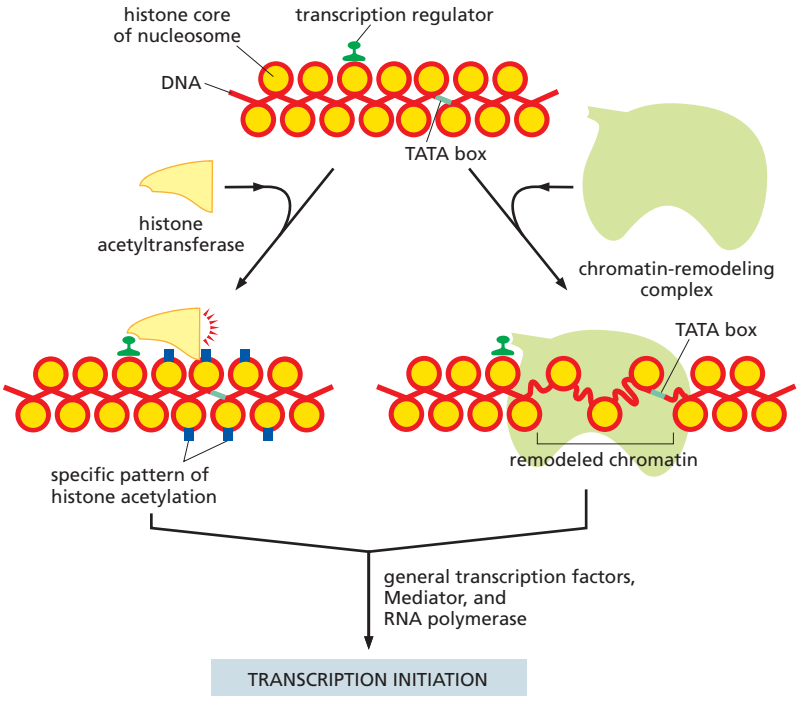

Figure 8–11 Eukaryotic transcriptional activators can recruit chromatin-modifying proteins to help initiate gene transcription. On the left, the recruitment of histone-modifying enzymes such as histone acetyltransferases adds acetyl groups to specific histones, which can then serve as binding sites for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

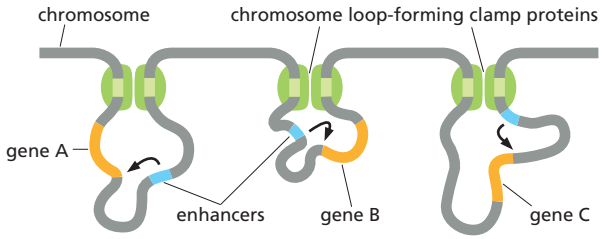

Figure 8–12 Animal and plant chromosomes are arranged in DNA loops. In this schematic diagram, specialized proteins (*green*) hold chromosomal DNA in loops, thereby favoring the association of each gene with its proper enhancer. The loops, sometimes called *topological associated domains* (TADs), range in size between thousands and millions of nucleotide pairs and are typically much larger than the loops that form between regulatory sequences and promoters (see Figure 8–10).

The Arrangement of Chromosomes into Looped Domains Keeps Enhancers in Check

We have seen that all genes have regulatory 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. What, then, prevents a transcripton regulator—bound to the control region of one gene—from looping in the wrong direction and inappropriately influencing the transcription of a neighboring gene?

To avoid such unwanted cross-talk, the chromosomal DNA of plants and animals is arranged in a series of loops that hold individual genes and their regulatory regions in rough proximity. This localization restricts the action of enhancers, preventing them from wandering across to adjacent genes. The chromosomal loops are formed by specialized proteins that bind to sequences that are then drawn together to form the base of the loop (**Figure 8–12**).

The importance of these loops is highlighted by the effects of mutations that prevent the loops from properly forming. Such mutations, which lead to genes being expressed at the wrong time and place, are found in numerous cancers and inherited diseases.

GENERATING SPECIALIZED CELL TYPES

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

In this section, we discuss some of the special features of transcriptional regulation that allow multicellular organisms to create and maintain specialized cell types. These cell types ultimately produce the tissues and organs that give worms, flies, and even humans their distinctive characteristics.

Eukaryotic Genes Are Controlled by Combinations of Transcription Regulators

The genes we have examined thus far have all been controlled by a small number of transcription regulators. While this is true for many simple bacterial systems, most eukaryotic transcription regulators work as part of a large "committee" of regulatory proteins, all of which cooperate to express the gene in the right cell type, in response to the right conditions, at the right time, and in the required amount.

The term **combinatorial control** refers to the process by which groups of transcription regulators work together to determine the expression of a single gene. The bacterial *Lac* operon we discussed earlier provides a simple example of the use of multiple regulators to control transcription (see Figure 8–9). In eukaryotes, such regulatory inputs have been amplified, so that a typical gene is controlled by dozens of transcription regulators that bind to regulatory sequences that may be spread over tens of thousands of nucleotide pairs. Together, these regulators direct the assembly of the Mediator, chromatin-remodeling complexes, histone-modifying enzymes, general transcripton factors, and, ultimately, RNA polymerase (**Figure 8–13**). In many cases, multiple repressors and activators are bound to the DNA that controls transcription of a given gene; how the cell integrates the effects of all of these proteins to determine the final level of gene expression is only now beginning to be understood. An example of such a complex regulatory system—one that participates in the development of a fruit fly from a fertilized egg—is described in **How We Know**, pp. 280–281.

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

In addition to being able to switch individual genes on and off, all cells—whether prokaryote or eukaryote—need to coordinate the expression of different genes. When a eukaryotic cell receives a signal to proliferate, for example, a number of hitherto unexpressed genes are turned on together to set in motion the events that lead eventually to cell division (discussed in Chapter 18). As discussed earlier, bacteria often coordinate the expression of a set of genes by having them clustered together in an operon under the control of a single promoter (see Figure 8–6). Such clustering is

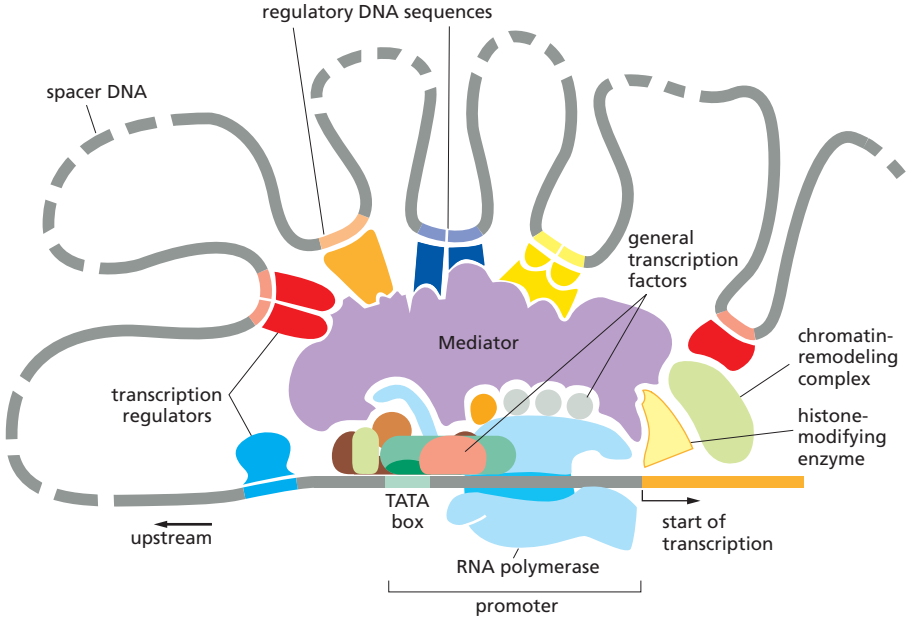

Figure 8–13 Transcription regulators work together as a "committee" to control the expression of a eukaryotic gene. Whereas the general transcription factors that assemble at the promoter are the same for all genes transcribed by RNA polymerase (see Figure 7–12), the transcription regulators and the locations of their DNA binding sites relative to the promoters are different for different genes. These regulators, along with chromatin-modifying proteins, are assembled at the promoter by the Mediator. The effects of multiple transcription regulators combine to determine the final rate of transcription initiation. The "spacer" DNA sequences that separate the regulatory DNA sequences are not recognized by any transcription regulators.

HOW WE KNOW

GENE REGULATION—THE STORY OF *EVE*

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

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

Seeing *Eve*

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

This pattern of expression never varies: the Eve protein can be found in the very same places in every *Drosophila* embryo (see Figure 8–14B). How can the expression of a gene be regulated with such spatial precision—such that one cell will produce a protein while a neighboring cell does not? To find out, researchers took a trip upstream.

Dissecting the DNA

As we have seen in this chapter, regulatory DNA sequences control which cells in an organism will express a particular gene, and at what point during development that gene will be turned on. In eukaryotes, these regulatory sequences are frequently located upstream of the gene itself. One way to locate a regulatory DNA sequence—and study how it operates—is to remove a piece of DNA from the region upstream of a gene of interest and insert that DNA upstream of a **reporter gene**—one that encodes a protein with an activity that is easy to monitor experimentally. If the piece of DNA contains a regulatory sequence, it will drive the expression of the reporter gene. When this patchwork piece of DNA is subsequently introduced into a cell or organism, the reporter gene will be expressed in the same cells and tissues that normally express the gene from which the regulatory sequence was derived (see Figure 10–24).

By excising various segments of the DNA sequences upstream of *Eve*, and coupling them to a reporter gene, researchers found that the expression of the gene is controlled by a series of seven regulatory modules—each of which specifies a single stripe of *Eve* expression. In this way, researchers identified, for example, a single segment of regulatory DNA that specifies stripe 2. They could excise this regulatory segment, link it to a reporter gene, and introduce the resulting DNA segment into the fly. When they examined embryos that carried this engineered DNA, they found that the reporter gene is expressed in the precise position of stripe 2 (**Figure 8–14**). Similar experiments revealed the existence of six other regulatory modules, one for each of the other Eve stripes.

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

The regulatory segment that defines stripe 2, for example, contains regulatory DNA sequences for four transcription regulators: two that activate *Eve* transcription and two that repress it (**Figure 8–15**). In the narrow band of tissue that constitutes stripe 2, it just so happens that the repressor proteins are not present—so the *Eve* gene is expressed; in the bands of tissue on either side of the stripe, where the repressors are present, *Eve* is kept quiet. And so a stripe is formed.

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

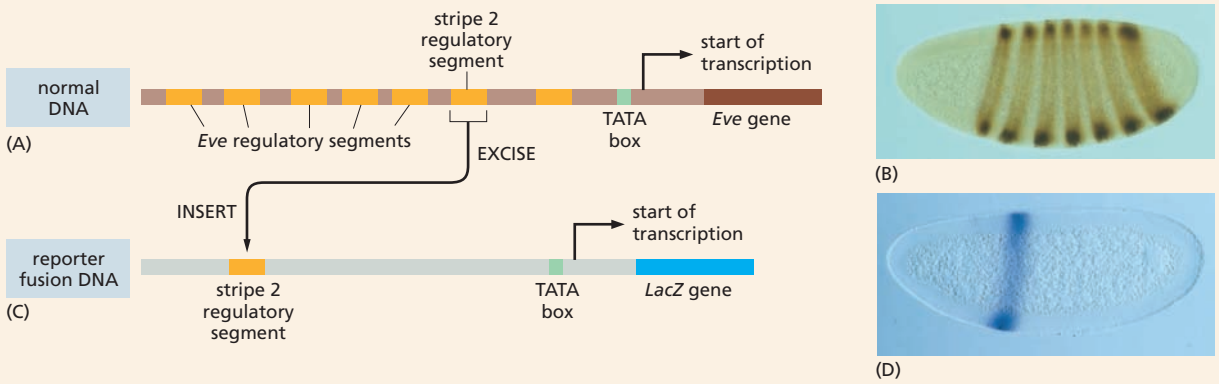

Figure 8–14 An experimental approach using a reporter gene reveals the modular construction of the *Eve* gene regulatory region. (A) Expression of the *Eve* gene is controlled by a series of regulatory segments (*orange*) that direct the production of Eve protein in stripes along the embryo. (B) Embryos stained with antibodies to the Eve protein show the seven characteristic stripes of *Eve* expression. (C) In the laboratory, the regulatory segment that directs the formation of stripe 2 can be excised from the DNA shown in part (A) and inserted upstream of the *E. coli LacZ* gene, which encodes the enzyme β-galactosidase (see Figure 8–9). (D) When the engineered DNA containing the stripe 2 regulatory segment is introduced into the genome of a fly, the resulting embryo expresses β-galactosidase mRNA precisely in the position of the second *Eve* stripe. This mRNA is visualized by *in situ* hybridization (see p. 352) using a labeled RNA probe that base pairs only with the *lacZ* mRNA. (B and D, courtesy of Stephen Small and Michael Levine.)

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

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

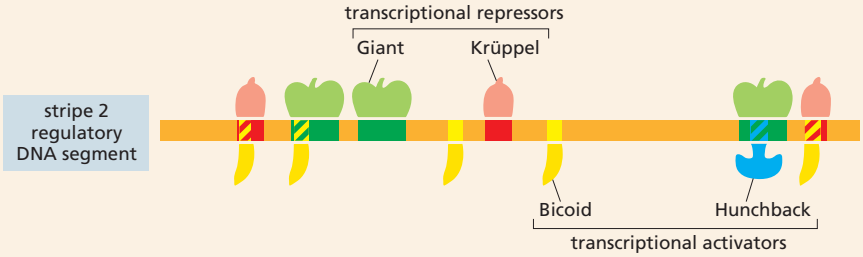

Figure 8–15 The regulatory segment that specifies *Eve* stripe 2 contains binding sites for four different transcription regulators. All four regulators are responsible for the proper expression of *Eve* in stripe 2. Flies that are deficient in the two activators, called Bicoid and Hunchback, fail to form stripe 2 efficiently; in flies deficient in either of the two repressors, called Giant and Krüppel, stripe 2 expands and covers an abnormally broad region of the embryo. As indicated in the diagram, in some cases the binding sites for the transcription regulators overlap, and the proteins compete for binding to the DNA. For example, the binding of Bicoid and Krüppel to the site at the far right is thought to be mutually exclusive. The regulatory segment is 480 base pairs in length.

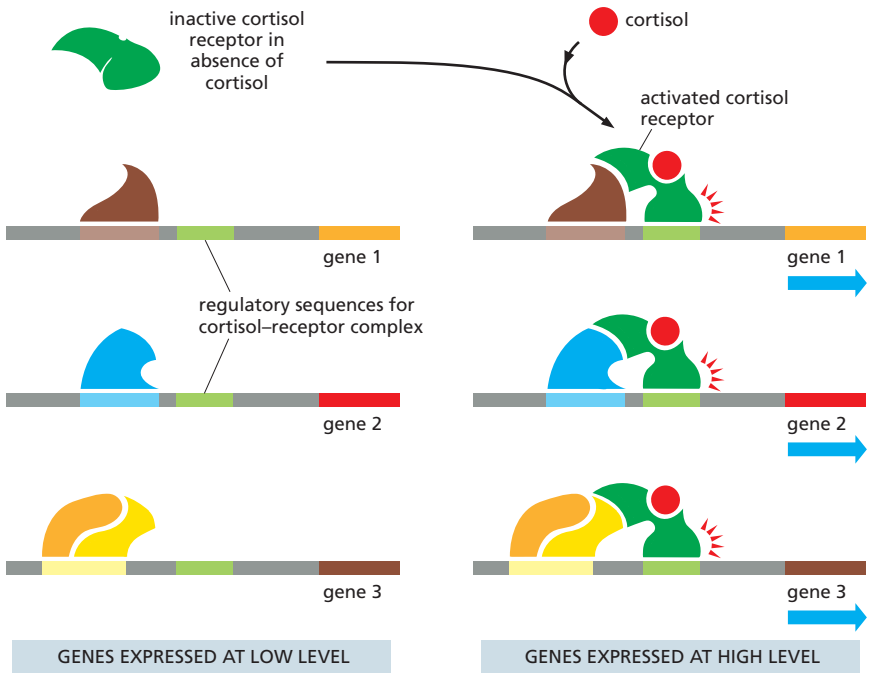

Figure 8–16 A single transcription regulator can coordinate the expression of many different genes. The action of the cortisol receptor is illustrated. On the left is a series of genes, each of which has a different activator protein bound to its respective regulatory DNA sequences. However, these bound proteins are not sufficient on their own to activate transcription efficiently. On the right is shown the effect of adding an additional transcription regulator—the cortisol–receptor complex—that binds to the same regulatory DNA sequence in each gene. The activated cortisol receptor completes the combination of transcription regulators required for efficient initiation of transcription, and all three genes are now switched on as a set.

rarely seen in eukaryotic cells, where each gene is transcribed and regulated individually. So how do eukaryotic cells coordinate the expression of multiple genes? In particular, given that a eukaryotic cell uses a committee of transcription regulators to control each of its genes, how can it rapidly and decisively switch whole groups of genes on or off?

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

An example of such coordinated regulation in humans is seen in response to cortisol (see Table 16–1, p. 536). As discussed earlier in this chapter, when this hormone is present, liver cells increase the expression of many genes, including those that allow the liver to produce glucose in response to starvation or prolonged stress. To switch on such cortisol-responsive genes, the cortisol receptor—a transcription regulator—first forms a complex with a molecule of cortisol. This cortisol–receptor complex then binds to a regulatory sequence in the DNA of each cortisol-responsive gene. When the cortisol concentration decreases again, the expression of all of these genes drops to normal levels. In this way, a single transcription regulator can coordinate the expression of many different genes (**Figure 8–16**).

Combinatorial Control Can Also Generate Different Cell Types

The ability to switch many different genes on or off using a limited number of transcription regulators is not only useful in the day-to-day regulation of cell function. It is also one of the means by which eukaryotic cells diversify into particular types of cells during embryonic development.

One striking example is the development of muscle cells. A mammalian skeletal muscle cell is distinguished from other cells by the production of a large number of characteristic proteins, such as the muscle-specific forms of actin and myosin that make up the contractile apparatus, as well as the receptor proteins and ion channel proteins in the plasma membrane that allow the muscle cell to contract in response to stimulation by nerves (discussed in Chapter 17). The genes encoding this unique array of muscle-specific proteins are all switched on coordinately as the muscle cell differentiates. Studies of developing muscle cells in culture have identified a small number of key transcription regulators, expressed only in potential muscle cells, that coordinate muscle-specific gene expression and are thus crucial for muscle-cell differentiation. This set of regulators activates the transcription of the genes that code for muscle-specific proteins by binding to specific DNA sequences present in their regulatory regions.

In the same way, other sets of transcription regulators can activate the expression of genes that are specific for other cell types. How different combinations of transcription regulators can tailor the development of different cell types is illustrated schematically in **Figure 8–17**.

Still other transcription regulators can maintain cells in an undifferentiated state, like the precursor cell shown in Figure 8–17. Some undifferentiated cells are so developmentally flexible they are capable of giving rise to all the specialized cell types in the body. The *embryonic stem (ES) cells* we discuss in Chapter 20 retain this remarkable quality, a property called *pluripotency*.

The differentiation of a particular cell type involves changes in the expression of thousands of genes: genes that encode products needed by the cell are expressed at high levels, while those that are not needed are expressed at low levels or shut down completely. A given transcription regulator, therefore, often controls the expression of hundreds or even

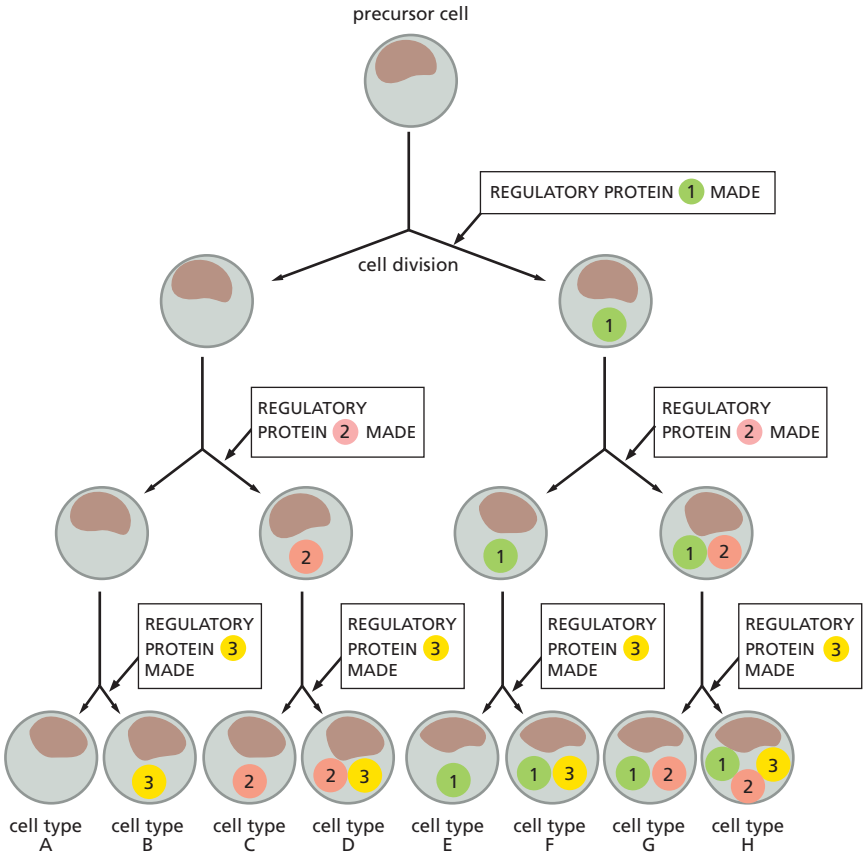

Figure 8–17 Combinations of a few transcription regulators can generate many cell types during development. In this simple scheme, a "decision" to make a new transcription regulator (shown as a numbered circle) is made after each cell division. Repetition of this simple rule can generate eight cell types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

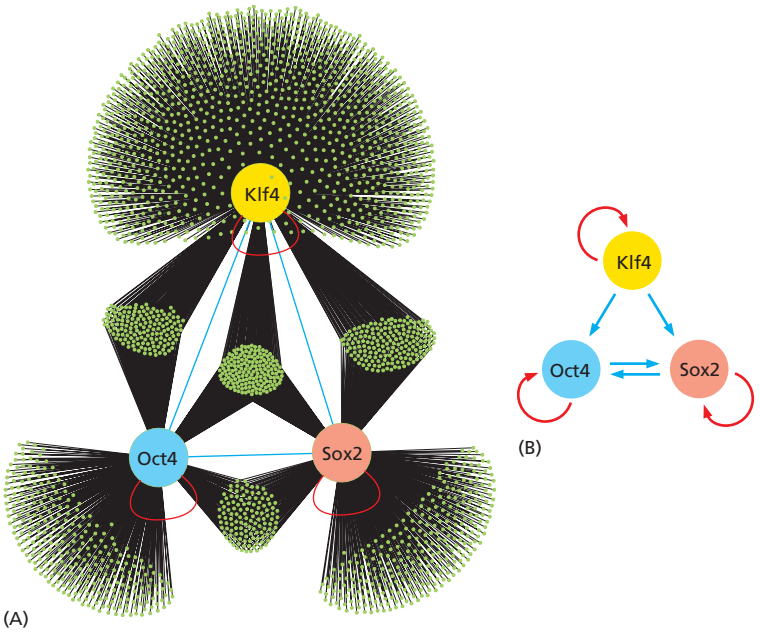

Figure 8–18 A set of three transcription regulators forms the regulatory network that specifies an embryonic stem cell. (A) The three transcription regulators—Klf4, Oct4, and Sox2—are shown in large *colored* circles. The genes whose regulatory sequences contain binding sites for each of these regulators are indicated by small *green* dots. The lines that link each regulator to a gene represent the binding of that regulator to the regulatory region of the gene. Note that although each regulator controls the expression of a unique set of genes, many of these target genes are bound by more than one transcription regulator—and a substantial set interacts with all three. (B) These three regulators also control their own expression. As shown here, each regulator binds to the regulatory region of its own gene, as indicated by the feedback loops (*red*). In addition, the regulators also bind to each other's regulatory regions (*blue*). Positive feedback loops, a common form of regulation, are discussed later in the chapter.

thousands of genes (**Figure 8–18**). Because each gene, in turn, is typically controlled by many different transcription regulators, a relatively small number of regulators acting in different combinations can form the enormously complex regulatory networks that generate specialized cell types. It is estimated that approximately 1000 transcription regulators are sufficient to control the 24,000 genes that give rise to an individual human.

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

We have seen that transcription regulators, working in combination, can control the expression of whole sets of genes and can produce a variety of cell types. But in some cases a single transcription regulator can initiate the formation of not just one cell type but a whole organ. A stunning example of such transcriptional control comes from studies of eye development in the fruit fly *Drosophila*. Here, a single transcription regulator called Ey triggers the differentiation of all of the specialized cell types that come together to form the eye. Flies with a mutation in the *Ey* gene have no eyes at all, which is how the regulator was discovered.

Figure 8–19 A master transcription regulator can direct the formation of an entire organ. Artificially induced expression of the *Drosophila Ey* gene in the precursor cells of the leg triggers the misplaced development of an eye on a fly's leg. The experimentally induced organ appears to be structurally normal, containing the various types of cells found in a typical fly eye. It does not, however, communicate with the fly's brain. (Walter Gehring, courtesy of Biozentrum, University of Basel.)

How the Ey protein coordinates the specification of each type of cell found in the eye—and directs their proper organization in three-dimensional space—is an actively studied topic in developmental biology. In essence, however, Ey functions like the transcription regulators we have already discussed, controlling the expression of multiple genes by binding to DNA sequences in their regulatory regions. Some of the genes controlled by Ey encode additional transcription regulators that, in turn, control the expression of other genes. In this way, the action of this *master transcription regulator*, which sits at the apex of a regulatory network like the one shown in Figure 8–18, produces a cascade of regulators that, working in combination, lead to the formation of an organized group of many different types of cells. One can begin to imagine how, by repeated applications of this principle, an organism as complex as a fly—or a human—progressively self-assembles, cell by cell, tissue by tissue, and organ by organ.

Master regulators such as Ey are so powerful that they can even activate their regulatory networks outside the normal location. In the laboratory, the *Ey* gene has been artificially expressed in fruit fly embryos in cells that would normally give rise to a leg. When these modified embryos develop into adult flies, some have an eye in the middle of a leg (**Figure 8–19**).

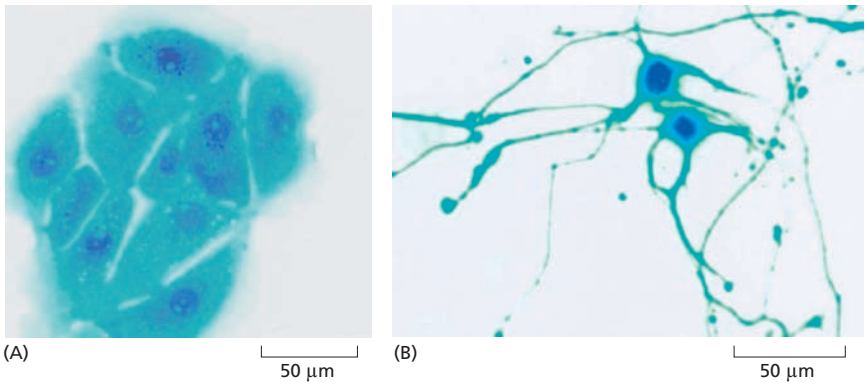

Figure 8–20 A small number of transcription regulators can convert one differentiated cell type directly into another. In this experiment, liver cells grown in culture (A) were converted into neuronal cells (B) via the artificial introduction of three nerve-specific transcription regulators. The cells are labeled with a fluorescent dye. Such interconversion would never take place during normal development. The result shown here depends on an experimenter expressing several nerve-specific regulators in liver cells, where these regulators would, during normal development, be tightly shut off. (From S. Marro et al., *Cell Stem Cell* 9:374–382, 2011. With permission from Elsevier.)

Transcription Regulators Can Be Used to Experimentally Direct the Formation of Specific Cell Types in Culture

We have seen that the *Ey* gene, when introduced into a fly embryo, can produce an eye in an unnatural location; this somewhat unusual outcome is made possible by the cooperation of numerous transcription regulators in a variety of cell types—a situation that is common in a developing embryo. Perhaps even more surprising is that some transcription regulators can convert one specialized cell type to another in a culture dish. For example, when the gene encoding the transcription regulator MyoD is artificially introduced into fibroblasts cultured from skin, the fibroblasts form musclelike cells. It appears that the fibroblasts, which are derived from the same broad class of embryonic cells as muscle cells, have already accumulated many of the other necessary transcription regulators required for the combinatorial control of the muscle-specific genes, and that addition of MyoD completes the unique combination required to direct the cells to become muscle.

This same type of *reprogramming* can produce even more impressive transformations. For example, a set of nerve-specific transcription regulators, when artificially expressed in cultured liver cells, can convert them into functional neurons (**Figure 8–20**). And the combination of transcription regulators shown in Figure 8–18 can be used in the laboratory to coax differentiated cells to *de-differentiate* into **induced pluripotent stem (iPS) cells**; these reprogrammed cells behave much like naturally occurring ES cells, and they can be directed to generate a variety of specialized differentiated cells (**Figure 8–21**). This approach, initially performed using cultured fibroblasts, has been adapted to produce iPS cells from a variety of specialized cell types, including those taken from humans. Differentiated cells produced from human iPS cells are currently being used in the study or treatment of disease, as we discuss in Chapter 20. Taken together, these dramatic demonstrations suggest that it may someday be possible to produce in the laboratory any cell type for which the correct combination of transcription regulators can be identified.

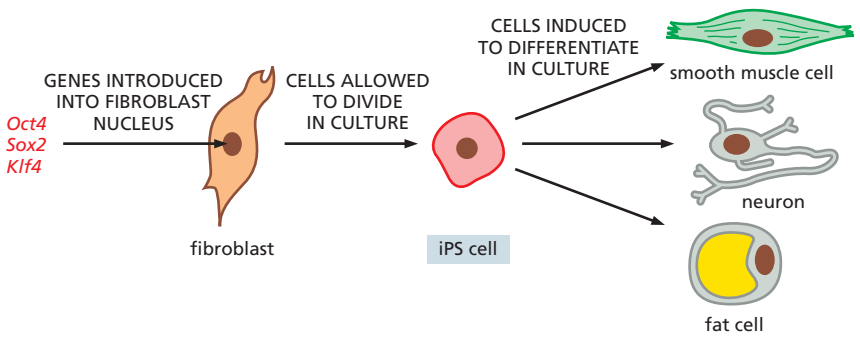

Figure 8–21 A combination of transcription regulators can induce a differentiated cell to de-differentiate into a pluripotent iPS cell. The artificial expression of a set of three genes, each of which encodes a transcription regulator, can reprogram a fibroblast into a pluripotent cell with ES cell-like properties. Like ES cells, such iPS cells can proliferate indefinitely in culture and can be stimulated by appropriate extracellular signal molecules to differentiate into almost any cell type in the body.

Differentiated Cells Maintain Their Identity

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

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

Cells have several ways of ensuring that their daughters remember what kind of cells they should be. One of the simplest and most important is through a **positive feedback loop**, where a master transcription regulator activates transcription of its own gene, in addition to that of other cell-type-specific genes. Each time a cell divides, the regulator is distributed to both daughter cells, where it continues to stimulate the positive feedback loop (**Figure 8–22**). The continued stimulation ensures that the regulator will continue to be produced in subsequent cell generations. The Ey protein and the transcription regulators involved in the generation of ES cells and iPS cells take part in such positive feedback loops (see Figure 8–18B). Positive feedback is crucial for establishing the "self-sustaining" circuits of gene expression that allow a cell to commit to a particular fate—and then to transmit that decision to its progeny.

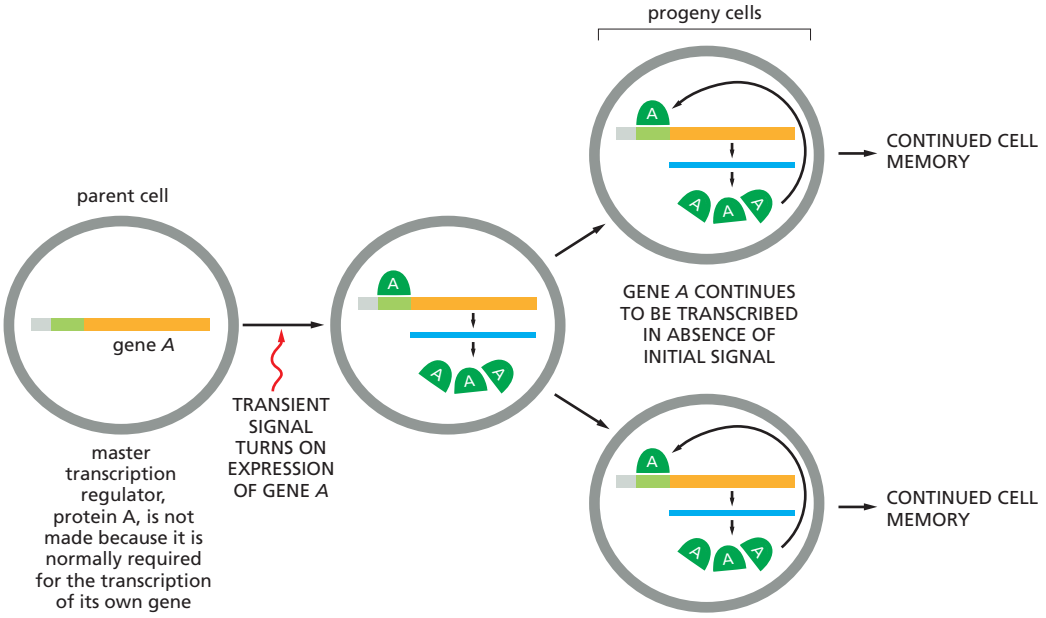

Figure 8–22 A positive feedback loop can generate cell memory. Protein A is a master transcription regulator that activates the transcription of its own gene—as well as other cell-type-specific genes (not shown). All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of protein A. As shown in Figure 8–18, each of the regulators needed to form iPS cells influences its own expression using this type of positive feedback loop.

Although positive feedback loops are probably the most prevalent way of ensuring that daughter cells remember what kind of cells they are meant to be, there are other ways of reinforcing cell identity. One involves the methylation of DNA. In vertebrate cells, **DNA methylation** occurs on certain cytosine bases (**Figure 8–23**). This covalent modification generally turns off the affected genes by attracting proteins that bind to methylated cytosines and block gene transcription. DNA methylation patterns are passed on to progeny cells by the action of an enzyme that copies the methylation pattern on the parent DNA strand to the daughter DNA strand as it is synthesized (**Figure 8–24**).

Another mechanism for inheriting gene expression patterns involves the modification of histones. When a cell replicates its DNA, each daughter double helix receives half of its parent's histone proteins, which contain the covalent modifications that were present on the parent chromosome. Enzymes responsible for these modifications may bind to the parental histones and confer the same modifications to the new histones nearby. It has been proposed that this cycle of modification helps reestablish the pattern of chromatin structure found in the parent chromosome (**Figure 8–25**).

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

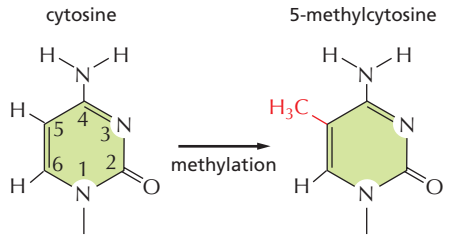

Figure 8–23 Formation of 5-methylcytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates, this modification is confined to selected cytosine (C) nucleotides that fall next to a guanine (G) in the sequence 5′-CG-3′.

POST-TRANSCRIPTIONAL CONTROLS

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

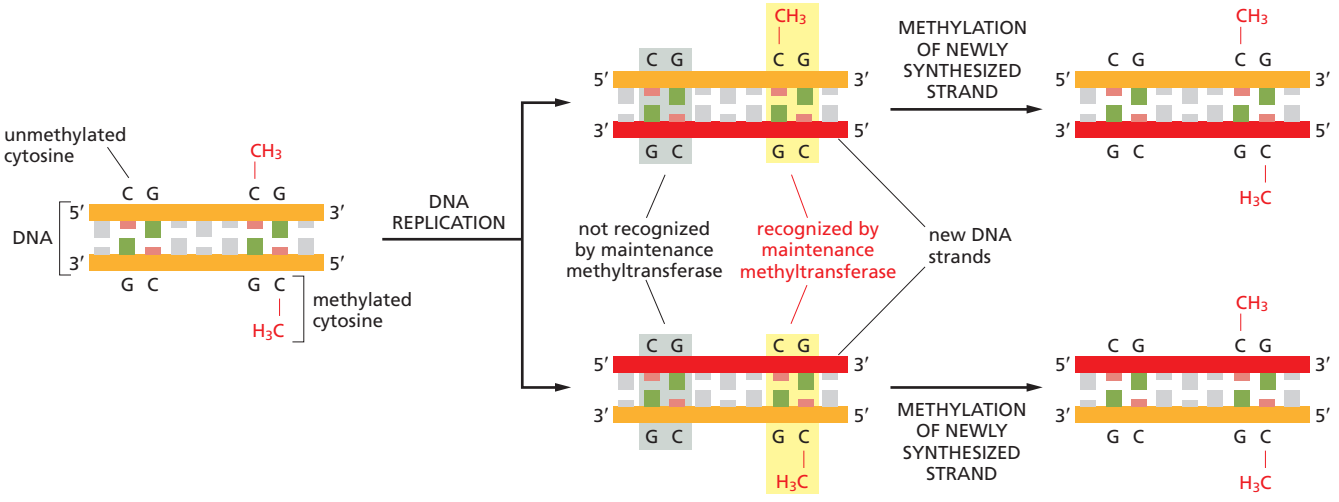

Figure 8–24 DNA methylation patterns can be faithfully inherited when a cell divides. An enzyme called a maintenance methyltransferase guarantees that once a pattern of DNA methylation has been established, it is inherited by newly made DNA. Immediately after DNA replication, each daughter double helix will contain one methylated DNA strand—inherited from the parent double helix—and one unmethylated, newly synthesized strand. The maintenance methyltransferase interacts with these hybrid double helices and methylates only those CG sequences that are base-paired with a CG sequence that is already methylated.

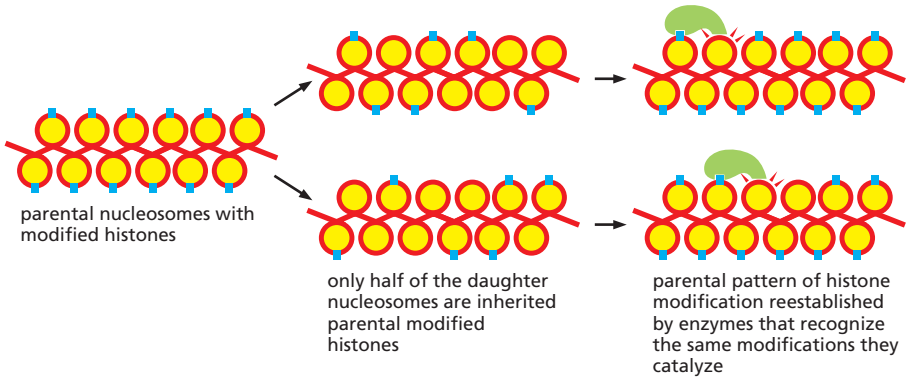

Figure 8–25 Histone modifications may be inherited by daughter chromosomes. As shown in this model, when a chromosome is replicated, its resident histones are distributed more or less randomly to each of the two daughter DNA double helices. Thus, each daughter chromosome will inherit about half of its parent's collection of modified histones. The remaining stretches of DNA receive newly synthesized, not-yet-modified histones. If the enzymes responsible for each type of modification bind to the specific modification they create, they can catalyze the "filling in" of this modification on the new histones. This cycle of modification and recognition can restore the parental histone modification pattern and, ultimately, allow the inheritance of the parental chromatin structure.

post-transcriptional controls, which operate after transcription has begun, play a crucial part in further fine-tuning the expression of almost all genes.

We have already encountered a few examples of such post-transcriptional control. For example, alternative RNA splicing allows different forms of a protein, encoded by the same gene, to be made in different tissues (Figure 7–23). And we saw that various post-translational modifications of a protein can regulate its concentration and activity (see Figure 4–47). In the remainder of this chapter, we consider several other examples—some only recently discovered—of the many ways in which cells can manipulate the expression of a gene after transcription has commenced.

mRNAs Contain Sequences That Control Their Translation

We saw in Chapter 7 that an mRNA's lifespan is dictated by specific nucleotide sequences within the untranslated regions that lie both upstream and downstream of the protein-coding sequence. These sequences often contain binding sites for proteins that are involved in RNA degradation. But they also carry information specifying whether—and how often—the mRNA is to be translated into protein.

Although the details differ between eukaryotes and bacteria, the general strategy is similar for both. Bacterial mRNAs contain a short ribosome-binding sequence located a few nucleotide pairs upstream of the AUG codon where translation begins (see Figure 7–40). This binding sequence forms base pairs with the rRNA in the small ribosomal subunit, correctly positioning the initiating AUG codon within the ribosome. Because this interaction is needed for efficient translation initiation, it provides an ideal target for translational control. By blocking—or exposing—the ribosome-binding sequence, the bacterium can either inhibit—or promote—the translation of an mRNA (**Figure 8–26**).

In eukaryotes, specialized repressor proteins can similarly inhibit translation initiation by binding to specific nucleotide sequences in the 5′ untranslated region of the mRNA, thereby preventing the ribosome from finding the first AUG. When conditions change, the cell can inactivate the repressor to initiate translation of the mRNA.

Regulatory RNAs Control the Expression of Thousands of Genes

As we saw in Chapter 7, RNAs perform many critical biological tasks. In addition to the mRNAs, which code for proteins, *noncoding RNAs* have a variety of functions. Some, such as transfer RNAs (tRNAs) and ribosomal

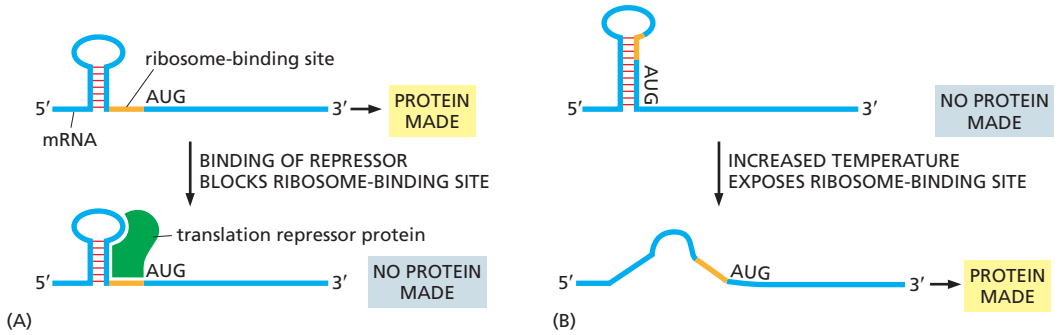

Figure 8–26 A bacterial gene's expression can be controlled by regulating translation of its mRNA. (A) Sequence-specific RNA-binding proteins can repress the translation of specific mRNAs by keeping the ribosome from binding to the ribosome-binding sequence (*orange*) in the mRNA. Some bacteria exploit this mechanism to inhibit the translation of ribosomal proteins. If a ribosomal protein is accidentally produced in excess over other ribosomal components, the free protein will inhibit translation of its own mRNA, thereby blocking its own synthesis. As new ribosomes are assembled, the levels of the free protein decrease, allowing the mRNA to again be translated and the ribosomal protein to be produced. (B) An mRNA from the pathogen *Listeria monocytogenes* contains a "thermosensor" RNA sequence that controls the translation of a set of mRNAs that code for proteins the bacterium needs to successfully infect its host. At the warmer temperatures inside a host, base pairs within the thermosensor come apart, exposing the ribosome-binding sequence, so the necessary protein is made.

RNAs (rRNAs) play key structural and catalytic roles in the cell, particularly in protein synthesis (see pp. 252–253). And the RNA component of telomerase is crucial for the complete duplication of eukaryotic chromosomes (see Figure 6–23). But we now know that many organisms, particularly animals and plants, produce thousands of additional noncoding RNAs.

Many of these noncoding RNAs have crucial roles in regulating gene expression and are therefore referred to as **regulatory RNAs**. These regulatory RNAs include *microRNAs*, *small interfering RNAs*, and *long noncoding RNAs*, and we discuss each in the remaining sections of the chapter.

MicroRNAs Direct the Destruction of Target mRNAs

MicroRNAs, or **miRNAs**, are tiny RNA molecules that control gene expression by base-pairing with specific mRNAs and reducing both their stability and their translation into protein. Like other RNAs, miRNAs also undergo processing to produce the mature, functional miRNA molecule. The mature miRNA, about 22 nucleotides in length, is packaged with specialized proteins to form an *RNA-induced silencing complex (RISC)*, which patrols the cytosol in search of mRNAs that are complementary in sequence to its bound miRNA (**Figure 8–27**). Once a target mRNA base-pairs with an miRNA, it is either destroyed immediately—by a nuclease that is part of the RISC—or its translation is blocked. In the latter case, the bound mRNA molecule is delivered to a region of the cytosol where other nucleases eventually degrade it. Destruction of the mRNA releases the miRNA-bearing RISC, allowing it to seek out additional mRNA targets. Thus, a single miRNA—as part of a RISC—can eliminate one mRNA molecule after another, thereby efficiently blocking production of the encoded protein.

There are thought to be roughly 500 different miRNAs encoded by the human genome; these RNAs may regulate as many as one-third of our protein-coding genes. Although we are only beginning to understand the full impact of these miRNAs, it is clear that they play a critical part in regulating gene expression and thereby influence many cell functions.

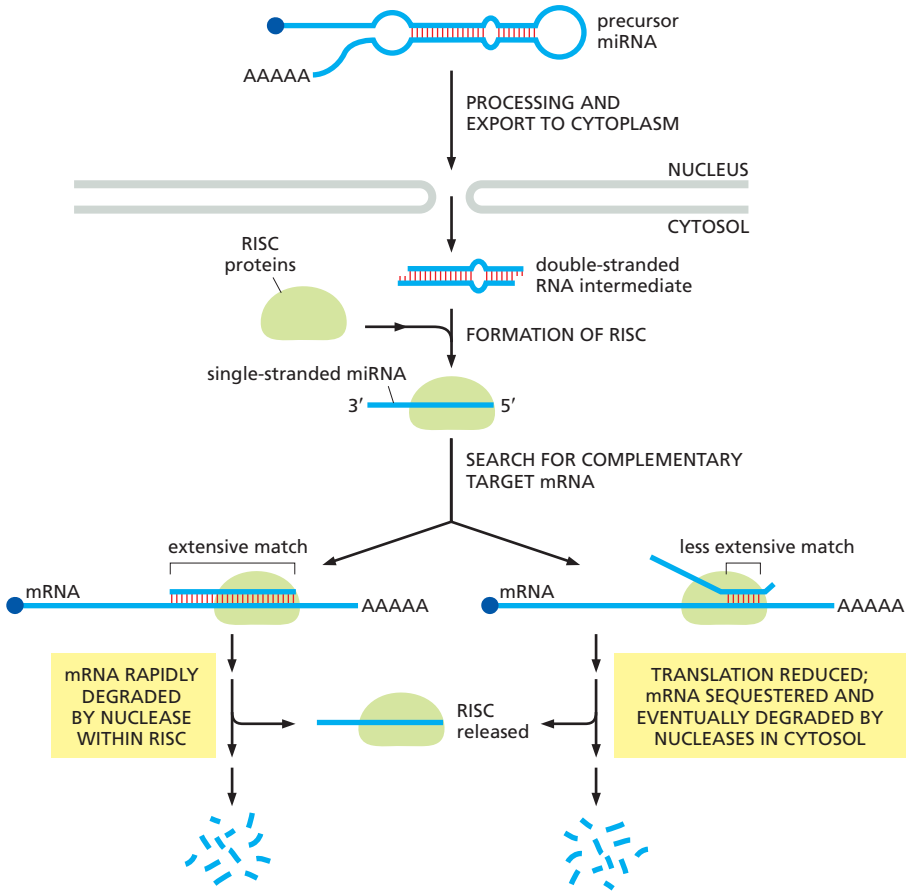

Figure 8–27 An miRNA targets a complementary mRNA molecule for destruction. Each precursor miRNA transcript is processed to form a double-stranded intermediate, which is further processed to form a mature, single-stranded miRNA. This miRNA assembles with a set of proteins into a complex called RISC, which then searches for mRNAs that have a nucleotide sequence complementary to its bound miRNA. Depending on how extensive the region of complementarity is, the target mRNA is either rapidly degraded by a nuclease within the RISC (shown on the *left*) or transferred to an area of the cytoplasm where other nucleases destroy it (shown on the *right*).

Small Interfering RNAs Protect Cells From Infections

Some of the same components that process and package miRNAs also play another crucial part in the life of a cell: they serve as a powerful cell defense mechanism. In this case, the system is used to eliminate "foreign" RNA molecules—in particular, long, double-stranded RNA molecules. Such RNAs are rarely produced by normal genes, but they often serve as intermediates in the life cycles of viruses and in the movement of some transposable genetic elements (discussed in Chapter 9). This form of RNA targeting, called **RNA interference** (**RNAi**), keeps these potentially destructive elements in check.

In the first step of RNAi, double-stranded, foreign RNAs are cut into short fragments (approximately 22 nucleotide pairs in length) in the cytosol by a protein called Dicer—the same protein used to generate the double-stranded RNA intermediate in miRNA production (see Figure 8–27). The resulting double-stranded RNA fragments, called **small interfering RNAs** (**siRNAs**), are then taken up by the same RISC proteins that carry miRNAs. The RISC discards one strand of the siRNA duplex and uses the remaining single-stranded RNA to seek and destroy complementary RNA molecules (**Figure 8–28**). In this way, the infected cell effectively turns the foreign RNA against itself.

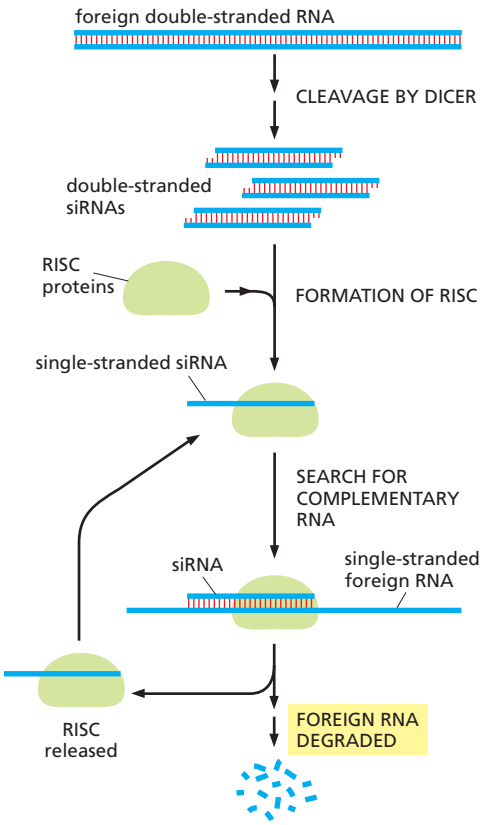

Figure 8–28 siRNAs are produced from double-stranded, foreign RNAs during the process of RNA interference. Double-stranded RNAs from a virus or transposable genetic element are first cleaved by a nuclease called Dicer. The resulting double-stranded fragments (known as siRNAs) are incorporated into RISCs, which discard one strand of the duplex and use the other strand to locate and destroy foreign RNAs that contain a complementary sequence.

At the same time, RNAi can also selectively shut off the synthesis of foreign RNAs by the host's RNA polymerase. In this case, the siRNAs produced by Dicer are packaged into a protein complex called RITS (for RNA-induced transcriptional silencing). Using its single-stranded siRNA as a guide, the RITS complex attaches itself to complementary RNA sequences as they emerge from an actively transcribing RNA polymerase (**Figure 8–29**). Positioned along a gene in this way, the RITS complex then attracts proteins that covalently modify nearby histones in a way that promotes the localized formation of heterochromatin (see Figure 5–27). This heterochromatin then blocks further transcription initiation at that site. Such RNAi-directed heterochromatin formation helps limit the spread of transposable genetic elements throughout the host genome.

RNAi operates in a wide variety of organisms, including single-celled fungi, plants, and worms, indicating that it is an evolutionarily ancient defense mechanism, particularly against viral infection. In some organisms, including many plants, the RNAi defense response can spread from tissue to tissue, allowing an entire organism to become resistant to a virus after only a few of its cells have been infected. In this sense, RNAi resembles certain aspects of the adaptive immune responses of vertebrates; in both cases, an invading pathogen elicits the production of molecules—either siRNAs or antibodies—that are custom-made to inactivate the specific invader and thereby protect the host.

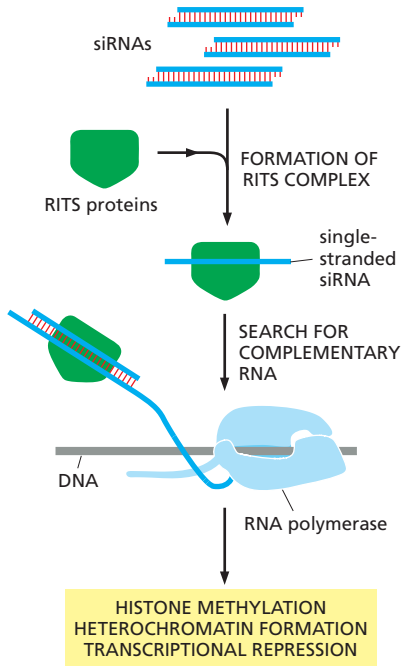

Figure 8–29 RNAi can also trigger transcriptional silencing. In this case, a single-stranded siRNA is incorporated into a RITS complex, which uses the single-stranded siRNA to search for complementary RNA sequences as they emerge from a transcribing RNA polymerase. The binding of the RITS complex attracts proteins that promote the modification of histones and the formation of tightly packed heterochromatin. This change in chromatin structure, directed by complementary base-pairing, causes transcriptional repression. Such silencing is used in plants, animals, and fungi to hold transposable elements in check.

Thousands of Long Noncoding RNAs May Also Regulate Mammalian Gene Activity

At the other end of the size spectrum are the **long noncoding RNAs**, a class of RNA molecules that are defined as being more than 200 nucleotides in length. There are thought to be upward of 5000 of these lengthy RNAs encoded in the human and mouse genomes. Yet, with few exceptions, their roles in the biology of the organism, if any, are not entirely clear.

One of the best understood of the long noncoding RNAs is *Xist*. This enormous RNA molecule, some 17,000 nucleotides long, is a key player in X-inactivation—the process by which one of the two X chromosomes in the cells of female mammals is permanently silenced (see Figure 5–28). Early in development, Xist is produced by only one of the X chromosomes in each female nucleus. The transcript then "sticks around," coating the chromosome and attracting the enzymes and chromatin-remodeling complexes that promote the formation of highly condensed heterochromatin. Other long noncoding RNAs may promote the silencing of specific genes in a similar manner.

Some long noncoding RNAs fold into specific, three-dimensional structures via complementary base pairing, as discussed in Chapter 7 (see for example Figure 7–5). These structures can serve as scaffolds, which bring together proteins that function together in a particular cell process (**Figure 8–30**). For example, one of the roles of the RNA molecule in telomerase—the enzyme that duplicates the ends of eukaryotic chromosomes (see

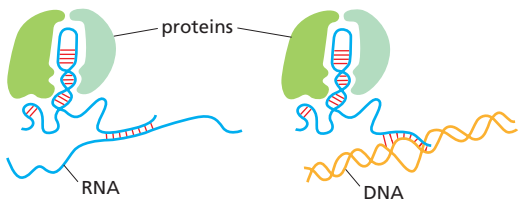

Figure 8–30 Long noncoding RNAs can serve as scaffolds, bringing together proteins that function in the same cell process. As described in Chapter 7, RNAs can fold into three-dimensional structures that can be recognized by specific proteins. By engaging in complementary base-pairing with other RNA molecules, these long noncoding RNAs can, in principle, localize proteins to specific sequences in RNA or DNA molecules, as shown.

Figure 6–23)—is to hold its different protein subunits together. By bringing together protein subunits, long noncoding RNAs can play important roles in many cell activities.

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

ESSENTIAL CONCEPTS

- A typical eukaryotic cell expresses only a fraction of its genes, and the distinct types of cells in multicellular organisms arise because different sets of genes are expressed as cells differentiate.
- In principle, gene expression can be controlled at any of the steps between a gene and its ultimate functional product. For the majority of genes, however, the initiation of transcription is the most important point of control.
- The transcription of individual genes is switched on and off in cells by transcription regulators, proteins that bind to short stretches of DNA called regulatory DNA sequences.
- In bacteria, transcription regulators usually bind to regulatory DNA sequences close to where RNA polymerase binds. This binding can either activate or repress transcription of the gene. In eukaryotes, regulatory DNA sequences are often separated from the promoter by many thousands of nucleotide pairs.
- Eukaryotic transcription regulators act in two main ways: (1) they can directly affect the assembly process that requires RNA polymerase and the general transcription factors at the promoter, and (2) they can locally modify the chromatin structure of promoter regions.
- In eukaryotes, the expression of a gene is generally controlled by a combination of different transcription regulators.
- In multicellular plants and animals, the production of different transcription regulators in different cell types ensures the expression of only those genes appropriate to the particular type of cell.
- A master transcription regulator, if expressed in the appropriate precursor cell, can trigger the formation of a specialized cell type or even an entire organ.
- One differentiated cell type can be converted to another by artificially expressing an appropriate set of transcription regulators. A differentiated cell can also be reprogrammed into a stem cell by artificially expressing a different, specific set of such regulators.
- Cells in multicellular organisms have mechanisms that enable their progeny to "remember" what type of cell they should be. A prominent mechanism for propagating cell memory relies on transcription regulators that perpetuate transcription of their own gene—a form of positive feedback.
- The pattern of DNA methylation can be transmitted from one cell generation to the next, producing a form of epigenetic inheritance that helps a cell remember the state of gene expression in its parent cell. There is also evidence for a form of epigenetic inheritance based on transmitted chromatin structures.
- Cells can regulate gene expression by controlling events that occur after transcription has begun. Many of these post-transcriptional mechanisms rely on RNA molecules that can influence their own stability or translation.

- MicroRNAs (miRNAs) control gene expression by base-pairing with specific mRNAs and inhibiting their stability and translation.
- Cells have a defense mechanism for destroying "foreign" double-stranded RNAs, many of which are produced by viruses. It makes use of small interfering RNAs (siRNAs) that are produced from the foreign RNAs in a process called RNA interference (RNAi).
- The recent discovery of thousands of long noncoding RNAs in mammals has revealed new roles for RNAs in assembling protein complexes and regulating gene expression.

KEY TERMS

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

QUESTIONS

QUESTION 8–4

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

A. What will happen next?

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

C. What advantage might this response to UV light provide to the virus?

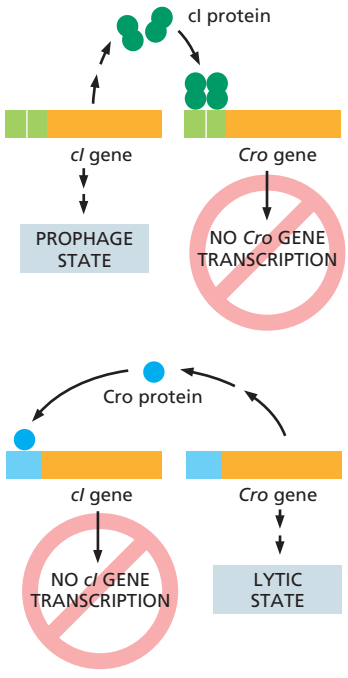

Figure Q8–4

QUESTION 8–5

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

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

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

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

QUESTION 8–6

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

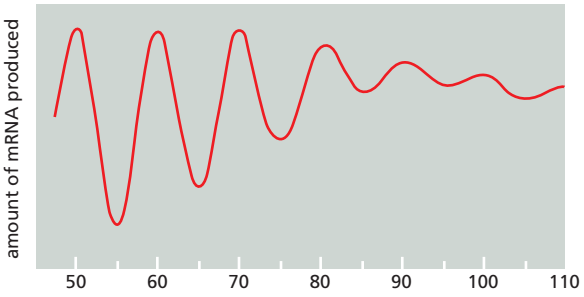

Figure Q8–6

QUESTION 8–7

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

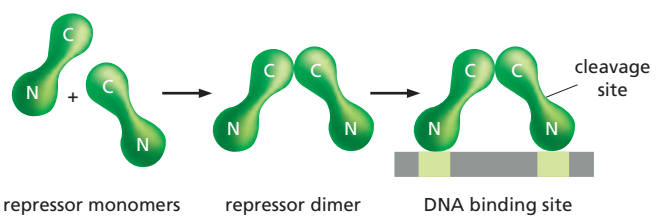

Figure Q8–7

QUESTION 8–8

The *Arg* genes that encode the enzymes for arginine biosynthesis are located at several positions around the genome of *E. coli*, and they are regulated coordinately by a transcription regulator encoded by the *ArgR* gene. The activity of the ArgR protein is modulated by arginine. Upon binding arginine, ArgR alters its conformation, dramatically changing its affinity for the DNA sequences in the promoters of the genes for the arginine biosynthetic enzymes. Given that ArgR is a repressor protein, would you expect that ArgR would bind more tightly or less tightly to the DNA sequences when arginine is abundant? If ArgR functioned instead as an activator protein, would you expect the binding of arginine to increase or to decrease its affinity for its regulatory DNA sequences? Explain your answers.

QUESTION 8–9

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

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

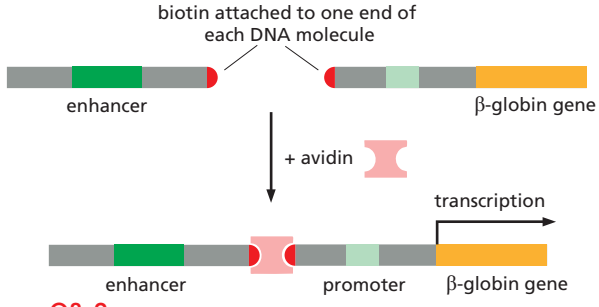

Figure Q8–9

QUESTION 8–10

Differentiated cells of an organism contain the same genes. (Among the few exceptions to this rule are the cells of the mammalian immune system, in which the formation of

specialized cells is based on limited rearrangements of the genome.) Describe an experiment that substantiates the first sentence of this question, and explain why it does.

QUESTION 8–11

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

QUESTION 8–12

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

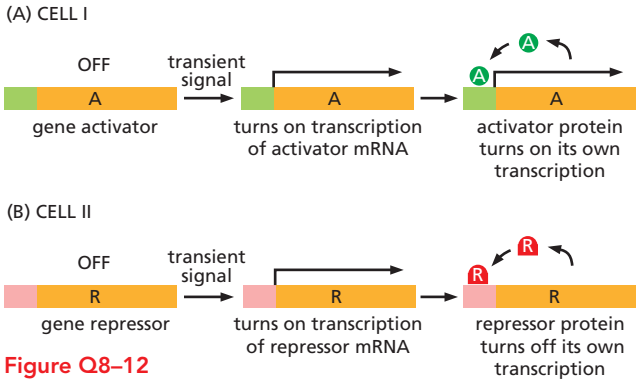

Figure Q8–12

QUESTION 8–13

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