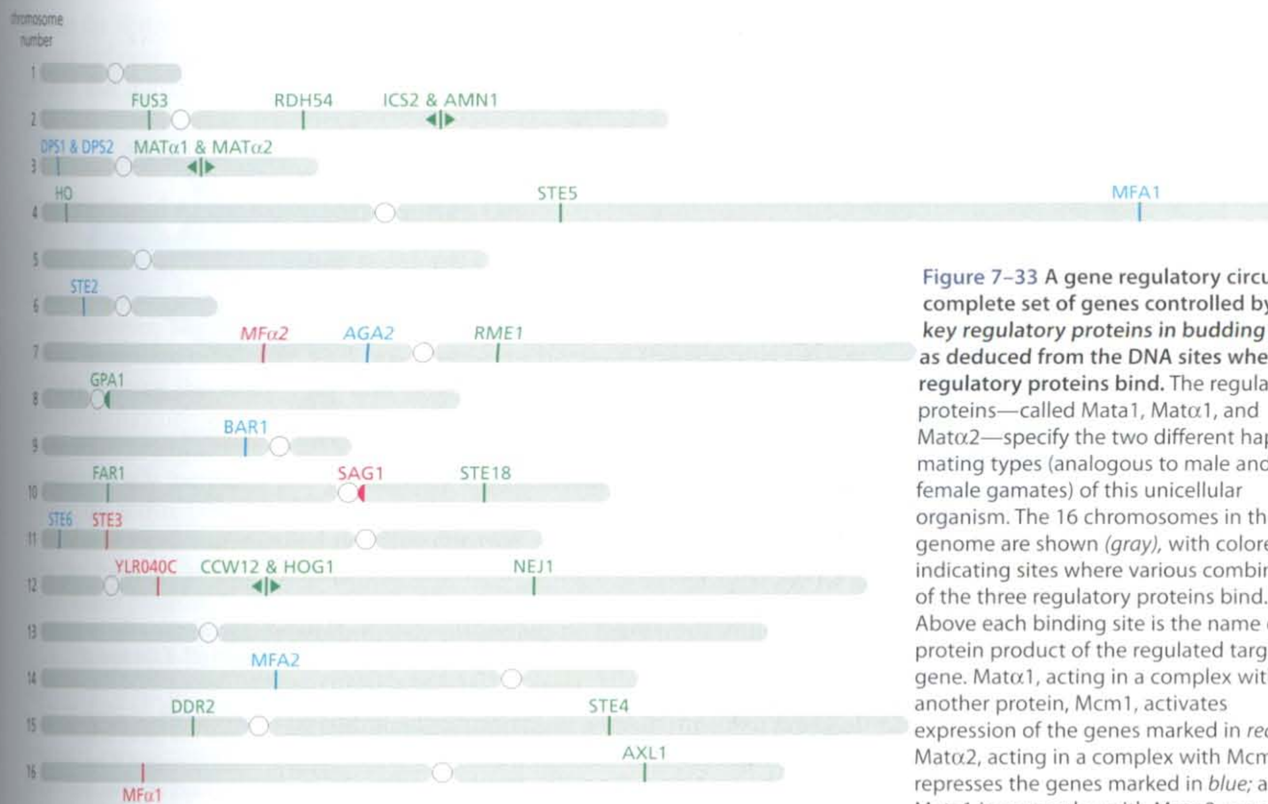


Figure 7–33 A gene regulatory circuit: the complete set of genes controlled by three key regulatory proteins in budding yeast, as deduced from the DNA sites where the regulatory proteins bind. The regulatory proteins—called Mata1, Matα1, and Matα2—specify the two different haploid mating types (analogous to male and female gamates) of this unicellular organism. The 16 chromosomes in the yeast genome are shown *(gray)*, with colored bars indicating sites where various combinations of the three regulatory proteins bind. Above each binding site is the name of the protein product of the regulated target gene. Matα1, acting in a complex with another protein, Mcm1, activates expression of the genes marked in *red;* Matα2, acting in a complex with Mcm1, represses the genes marked in *blue;* and Mata1 in a complex with Matα2 represses the genes marked in *green* (see Figures 7–21 and 7–65). *Double arrowheads* represent divergently transcribed genes, which are controlled by the indicated gene regulatory proteins. This complete map of bound regulatory proteins was determined using a combination of genome-wide chromatin immunoprecipitation (see Figure 7–32) and phylogenetic footprinting (see Figure 7–29). Such determinations of complete transcriptional circuits show that transcriptional networks are not infinitely complex, although they may appear that way initially. This type of study also helps to reveal the overall logic of the transcriptional circuits used by modern cells. (From D.J. Galgoczy et al., *Proc. Natl Acad. Sci. U.S.A.* 101:18069–18074, 2004. With permission from National Academy of Sciences.)

of their DNA into chromatin, creates special challenges and some novel opportunities for control—as we shall see. We begin with the simplest example—an on-off switch in bacteria that responds to a single signal.

The Tryptophan Repressor Is a Simple Switch That Turns Genes On and Off in Bacteria

The chromosome of the bacterium *E. coli*, a single-celled organism, is a single circular DNA molecule of about 4.6×10^6 nucleotide pairs. This DNA encodes approximately 4300 proteins, although the cell makes only a fraction of these at any one time. The expression of many genes is regulated according to the available food in the environment. This is illustrated by the five *E. coli* genes that code for enzymes that manufacture the amino acid tryptophan. These genes are arranged as a single **operon**; that is, they are adjacent to one another on the chromosome and are transcribed from a single *promoter* as one long mRNA molecule (**Figure 7–34**). But when tryptophan is present in the growth medium and enters the cell (when the bacterium is in the gut of a mammal that has just eaten a meal of protein, for example), the cell no longer needs these enzymes and shuts off their production.

The molecular basis for this switch is understood in considerable detail. As described in Chapter 6, a promoter is a specific DNA sequence that directs RNA polymerase to bind to DNA, to open the DNA double helix, and to begin

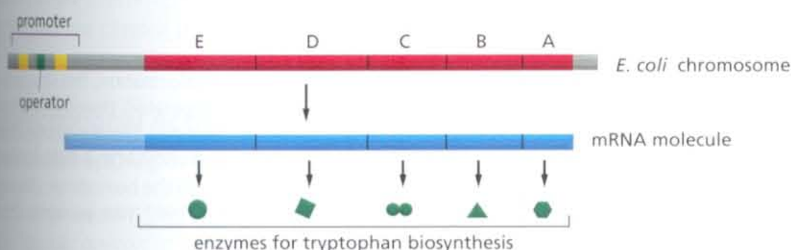


Figure 7–34 The clustered genes in *E. coli* that code for enzymes that manufacture the amino acid tryptophan. These five genes of the *Trp* operon—denoted as *TrpA, B, C, D,* and *E*—are transcribed as a single mRNA molecule, which allows their expression to be controlled coordinately. Clusters of genes transcribed as a single mRNA molecule are common in bacteria. Each such cluster is called an operon.

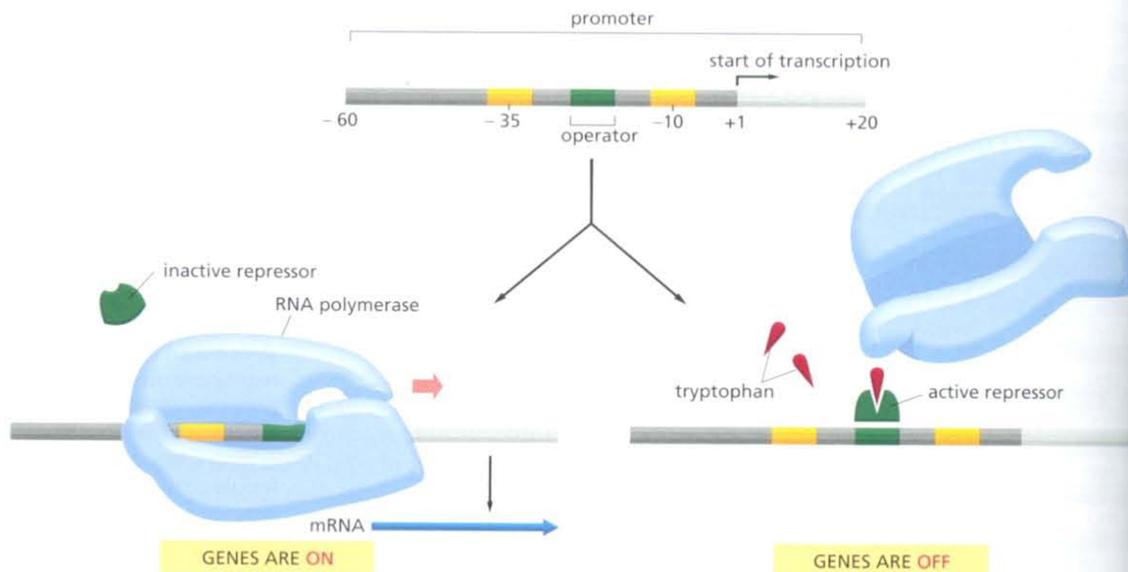


Figure 7–35 Switching the tryptophan genes on and off. If the level of tryptophan inside the cell is low, RNA polymerase binds to the promoter and transcribes the five genes of the tryptophan (*Trp*) operon. If the level of tryptophan is high, however, the tryptophan repressor is activated to bind to the operator, where it blocks the binding of RNA polymerase to the promoter. Whenever the level of intracellular tryptophan drops, the repressor releases its tryptophan and becomes inactive, allowing the polymerase to begin transcribing these genes. The promoter includes two key blocks of DNA sequence information, the –35 and –10 regions highlighted in *yellow* (see Figure 6–12).

synthesizing an RNA molecule. Within the promoter that directs transcription of the tryptophan biosynthetic genes lies a regulator element called an **operator** (see Figure 7–34). This is simply a short region of regulatory DNA of defined nucleotide sequence that is recognized by a repressor protein, in this case the **tryptophan repressor**, a member of the helix–turn–helix family (see Figure 7–11). The promoter and operator are arranged so that when the tryptophan repressor occupies the operator, it blocks access to the promoter by RNA polymerase, thereby preventing expression of the tryptophan-producing enzymes (**Figure 7–35**).

The block to gene expression is regulated in an ingenious way: to bind to its operator DNA, the repressor protein has to have two molecules of the amino acid tryptophan bound to it. As shown in **Figure 7–36**, tryptophan binding tilts the helix–turn–helix motif of the repressor so that it is presented properly to the DNA major groove; without tryptophan, the motif swings inward and the protein is unable to bind to the operator. Thus, the tryptophan repressor and operator form a simple device that switches production of the tryptophan biosynthetic enzymes on and off according to the availability of free tryptophan.

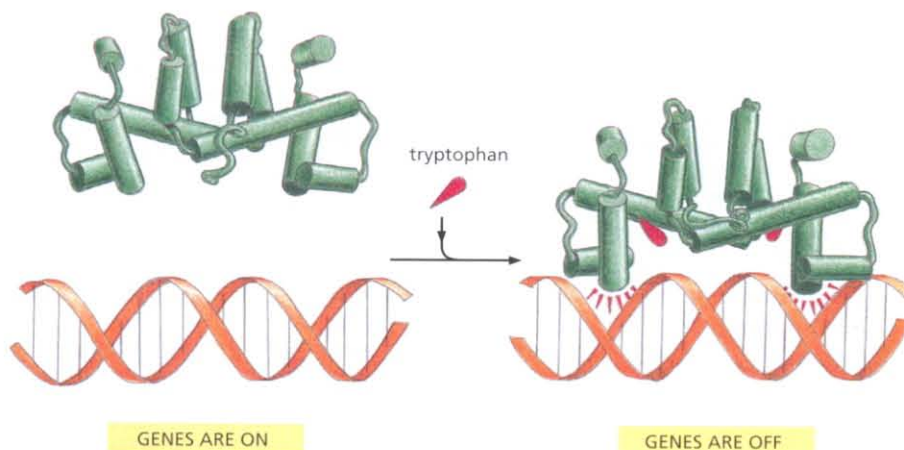


Figure 7–36 The binding of tryptophan to the tryptophan repressor protein changes its conformation. This structural change enables this gene regulatory protein to bind tightly to a specific DNA sequence (the operator), thereby blocking transcription of the genes encoding the enzymes required to produce tryptophan (the *Trp* operon). The three-dimensional structure of this bacterial helix–turn–helix protein, as determined by x-ray diffraction with and without tryptophan bound, is illustrated. Tryptophan binding increases the distance between the two recognition helices in the homodimer, allowing the repressor to fit snugly on the operator. (Adapted from R. Zhang et al., *Nature* 327:591–597, 1987. With permission from Macmillan Publishers Ltd.)

Because the active, DNA-binding form of the protein serves to turn genes off, this mode of gene regulation is called **negative control**, and the gene regulatory proteins that function in this way are called *transcriptional repressors* or *gene repressor proteins*.

Transcriptional Activators Turn Genes On

We saw in Chapter 6 that purified *E. coli* RNA polymerase (including its σ subunit) can bind to a promoter and initiate DNA transcription. Many bacterial promoters, however, are only marginally functional on their own, either because they are recognized poorly by RNA polymerase or because the polymerase has difficulty opening the DNA helix and beginning transcription. In either case these poorly functioning promoters can be rescued by gene regulatory proteins that bind to a nearby site on the DNA and contact the RNA polymerase in a way that dramatically increases the probability that a transcript will be initiated. Because the active, DNA-binding form of such a protein turns genes on, this mode of gene regulation is called **positive control**, and the gene regulatory proteins that function in this manner are known as *transcriptional activators* or *gene activator proteins*. In some cases, bacterial gene activator proteins aid RNA polymerase in binding to the promoter by providing an additional contact surface for the polymerase. In other cases, they contact RNA polymerase and facilitate its transition from the initial DNA-bound conformation of polymerase to the actively transcribing form by stabilizing a transition state of the enzyme. Like repressors, gene activator proteins must be bound to DNA to exert their effects. In this way, each regulatory protein acts selectively, controlling only those genes that bear a DNA sequence recognized by it.

DNA-bound activator proteins can increase the rate of transcription initiation up to 1000-fold, a value consistent with a relatively weak and nonspecific interaction between the activator and RNA polymerase. For example, a 1000-fold change in the affinity of RNA polymerase for its promoter corresponds to a change in ΔG of ~4 kcal/mole, which could be accounted for by just a few weak, noncovalent bonds. Thus gene activator proteins can work simply by providing a few favorable interactions that help to attract RNA polymerase to the promoter.

As in negative control by a transcriptional repressor, a transcriptional activator can operate as part of a simple on–off genetic switch. The bacterial activator protein *CAP (catabolite activator protein)*, for example, activates genes that enable *E. coli* to use alternative carbon sources when glucose, its preferred carbon source, is unavailable. Falling levels of glucose cause an increase in the intracellular signaling molecule cyclic AMP, which binds to the CAP protein, enabling it to bind to its specific DNA sequence near target promoters and thereby turn on the appropriate genes. In this way the expression of a target gene is switched on or off, depending on whether cyclic AMP levels in the cell are high or low, respectively. **Figure 7–37** summarizes the different ways that positive and negative control can be used to regulate genes.

Transcriptional activators and transcriptional repressors are similar in design. The tryptophan repressor and the transcriptional activator CAP, for example, both use a helix–turn–helix motif (see Figure 7–11) and both require a small cofactor in order to bind DNA. In fact, some bacterial proteins (including CAP and the bacteriophage lambda repressor) can act as either activators or repressors, depending on the exact placement of the DNA sequence they recognize in relation to the promoter: if the binding site for the protein overlaps the promoter, the polymerase cannot bind and the protein acts as a repressor (**Figure 7–38**).

16th

A Transcriptional Activator and a Transcriptional Repressor Control the *Lac* Operon

More complicated types of genetic switches combine positive and negative controls. The *Lac operon* in *E. coli*, for example, unlike the *Trp operon*, is under both

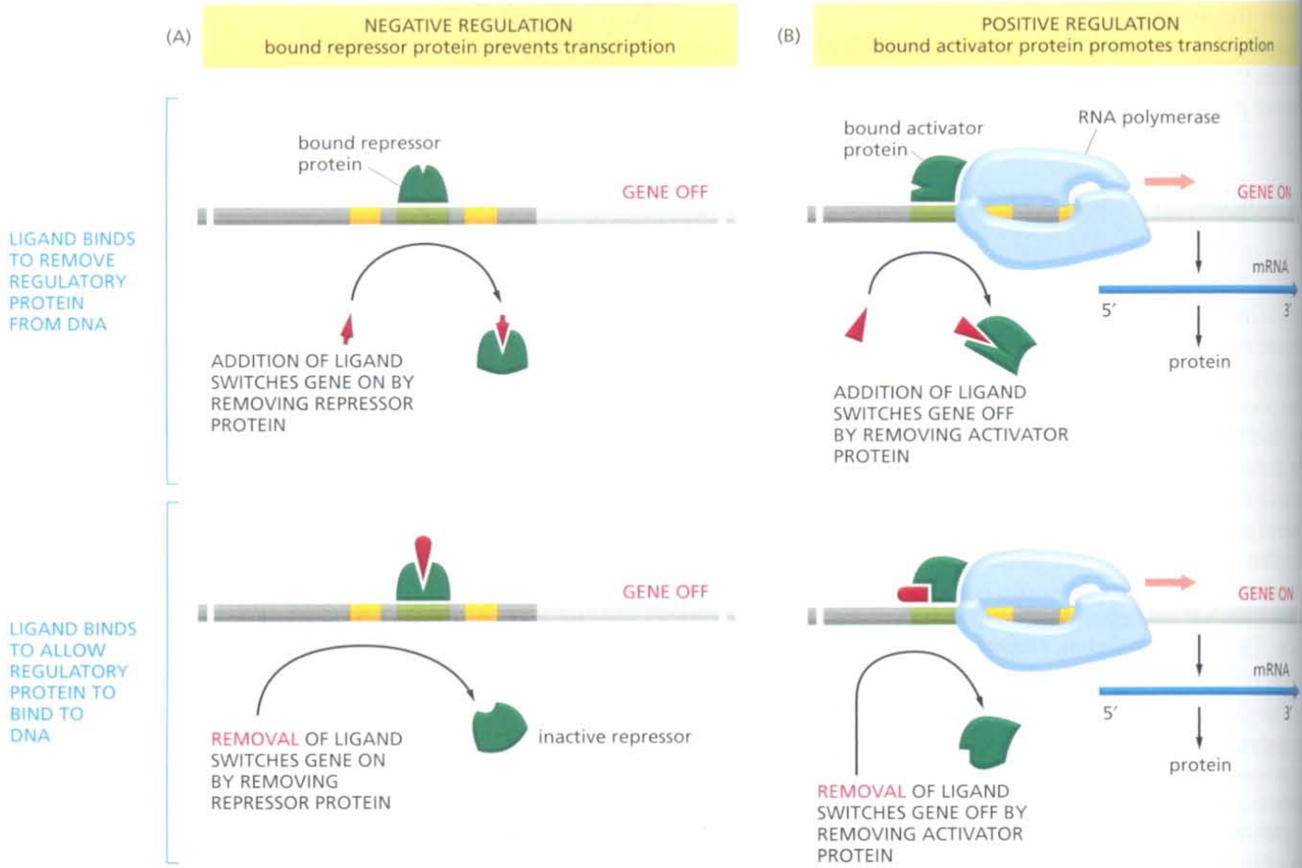


Figure 7–37 Summary of the mechanisms by which specific gene regulatory proteins control gene transcription in procaryotes. (A) Negative regulation; (B) positive regulation. Note that the addition of an "inducing" ligand can turn on a gene either by removing a gene repressor protein from the DNA *(upper left panel)* or by causing a gene activator protein to bind *(lower right panel)*. Likewise, the addition of an "inhibitory" ligand can turn off a gene either by removing a gene activator protein from the DNA *(upper right panel)* or by causing a gene repressor protein to bind *(lower left panel)*.

negative and positive transcriptional controls by the Lac repressor protein and CAP, respectively. The *Lac* operon codes for proteins required to transport the disaccharide lactose into the cell and to break it down. CAP, as we have seen, enables bacteria to use alternative carbon sources such as lactose in the absence of glucose. It would be wasteful, however, for CAP to induce expression of the *Lac* operon if lactose is not present, and the Lac repressor ensures that the *Lac* operon is shut off in the absence of lactose. This arrangement enables the control region of the *Lac* operon to respond to and integrate two different signals, so that the operon is highly expressed only when two conditions are met: lactose must be present and glucose must be absent. In any of the other three possible signal combinations, the cluster of genes is held in the off state (**Figure 7–39**).

The simple logic of this genetic switch first attracted the attention of biologists over 50 years ago. As explained above, the molecular basis of the switch was uncovered by a combination of genetics and biochemistry, providing the first glimpse into how gene expression is controlled.

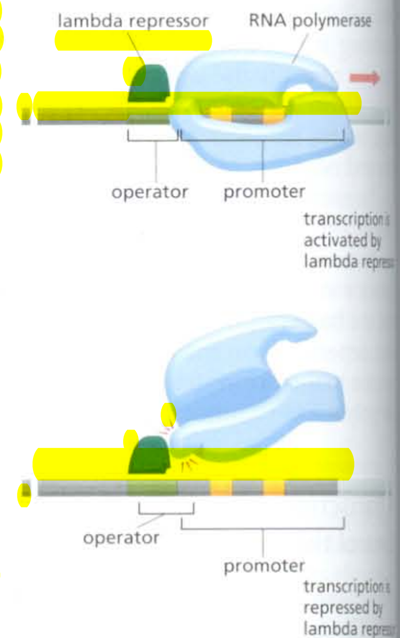


Figure 7–38 Some bacterial gene regulatory proteins can act as either a transcriptional activator or a repressor, depending on the precise placement of their DNA-binding sites. An example is the bacteriophage lambda repressor. For some genes, the protein acts as a transcriptional activator by providing a favorable contact for RNA polymerase *(top)*. At other genes *(bottom)*, the operator is located one base pair closer to the promoter, and, instead of helping polymerase, the repressor now competes with it for binding to the DNA. The lambda repressor recognizes its operator by a helix–turn–helix motif, as shown in Figure 7–11.

+ GLUCOSE
+ LACTOSE
+ GLUCOSE
– LACTOSE
– GLUCOSE
– LACTOSE
– GLUCOSE
+ LACTOSE

DNA Lo…

The cont… cal, but th… tion revea… operon w… revealed … molecule… out the i… strengthe… leads to g…
DNA … helix to c… ing one … nucleotic… 7–41). W…

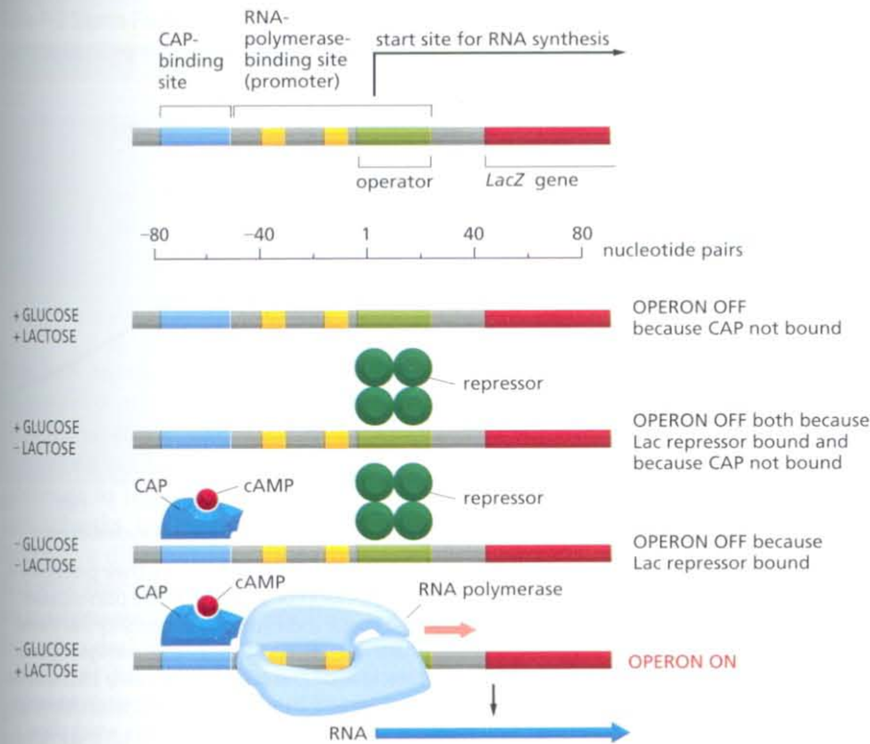


Figure 7–39 Dual control of the *Lac* operon. Glucose and lactose levels control the initiation of transcription of the *Lac* operon through their effects on CAP and the *Lac* repressor protein, respectively. *LacZ*, the first gene of the *Lac* operon, encodes the enzyme β-galactosidase, which breaks down the disaccharide lactose to galactose and glucose. Lactose addition increases the concentration of allolactose, an isomer of lactose, which binds to the repressor protein and removes it from the DNA. Glucose addition decreases the concentration of cyclic AMP; because cyclic AMP no longer binds to CAP, this gene activator protein dissociates from the DNA, turning off the operon.

This figure summarizes the essential features of the *Lac* operon, but in reality the situation is more complex. There are several Lac repressor binding sites located at different positions along the DNA. Although the one illustrated exerts the greatest effect, the others are required for full repression (see Figure 7–40). In addition, expression of the *Lac* operon never completely shuts down. A small amount of the enzyme β-galactosidase is required to convert lactose to allolactose, thereby permitting the Lac repressor to be inactivated when lactose is added to the growth medium.

DNA Looping Occurs During Bacterial Gene Regulation

The control of the Lac operon as shown in Figure 7–39 is simple and economical, but the continued study of this and other examples of bacterial gene regulation revealed a new feature of gene regulation, known as *DNA looping*. The *Lac* operon was originally thought to contain a single operator, but subsequent work revealed additional, secondary operators located nearby. A single tetrameric molecule of the Lac repressor can bind two operators simultaneously, looping out the intervening DNA. The ability to bind simultaneously to two operators strengthens the overall interaction of the Lac repressor with DNA and thereby leads to greater levels of repression in the cell (**Figure 7–40**).

DNA looping also allows two different proteins bound along a DNA double helix to contact one another readily. The DNA can be thought of as a tether, helping one DNA-bound protein interact with another even though thousands of nucleotide pairs may separate the binding sites for the two proteins (**Figure 7-41**). We shall see below that DNA looping is especially important in eucaryotic

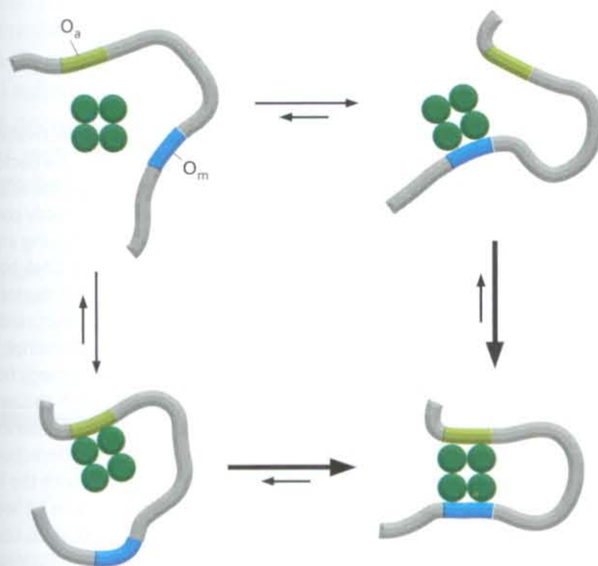


Figure 7–40 DNA looping can stabilize protein–DNA interactions. The Lac repressor, a tetramer, can simultaneously bind to two operators. The *Lac* operon has a total of three operators, but for simplicity, only two are shown here, the main operator (O_m) and an auxiliary operator (O_a). The figure shows all the possible states of the Lac repressor bound to these two operators. At the concentrations of Lac repressor in the cell, and in the absence of lactose, the state in the lower right is the most stable, and to dissociate completely from the DNA, the Lac repressor must first pass through an intermediate where it is bound to only a single operator. In these states, the local concentration of the repressor is very high in relation to the free operator, and the reaction to the double-bound form is favored over the dissociation reaction. In this way, even a low-affinity site (O_a) can increase the occupancy of a high-affinity site (O_m) and give higher levels of gene repression in the cell. (Adapted from J.M.G. Vilar and S. Leibler, *J. Mol. Biol.* 331:981–989, 2003. With permission from Academic Press.)

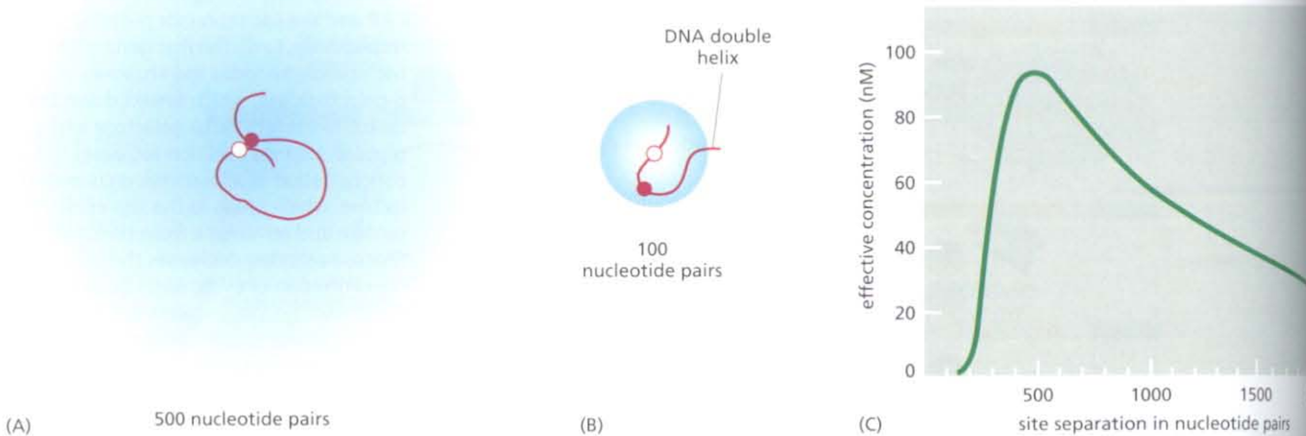


Figure 7–41 Binding of two proteins to separate sites on the DNA double helix can greatly increase their probability of interacting. (A) The tethering of one protein to the other via an intervening DNA loop of 500 nucleotide pairs increases their frequency of collision. The intensity of the *blue* coloring at each point in space indicates the probability that the red protein will be located at that distance from the *white* protein. (B) The flexibility of DNA is such that an average sequence makes a smoothly graded 90° bend (a curved turn) about once every 200 nucleotide pairs. Thus, when only 100 nucleotide pairs tethers two proteins, the contact between those proteins is relatively restricted. In such cases the protein interaction is facilitated when the two protein-binding sites are separated by a multiple of about 10 nucleotide pairs, which places both proteins on the same side of the DNA helix (which has about 10 nucleotides per turn) and thus on the inside of the DNA loop, where they can best reach each other. (C) The theoretical effective concentration of the *red* protein at the site where the *white* protein is bound, as a function of their separation. Experiments suggest that the actual effective concentrations at short distances are greater than those predicted here. (C, courtesy of Gregory Bellomy, modified from M.C. Mossing and M.T. Record, *Science* 233:889–892, 1986. With permission from AAAS.)

gene regulation. However, it also plays crucial roles in many examples of bacterial gene regulation in addition to that of the *Lac* operon. For example, DNA looping readily allows the bacterial gene activator protein NtrC to contact RNA polymerase directly even though the two proteins are bound several hundred nucleotide pairs apart (**Figure 7–42**).

Bacteria Use Interchangeable RNA Polymerase Subunits to Help Regulate Gene Transcription

We have seen the importance of gene regulatory proteins that bind to sequences of DNA and signal to RNA polymerase whether or not to start the synthesis of an

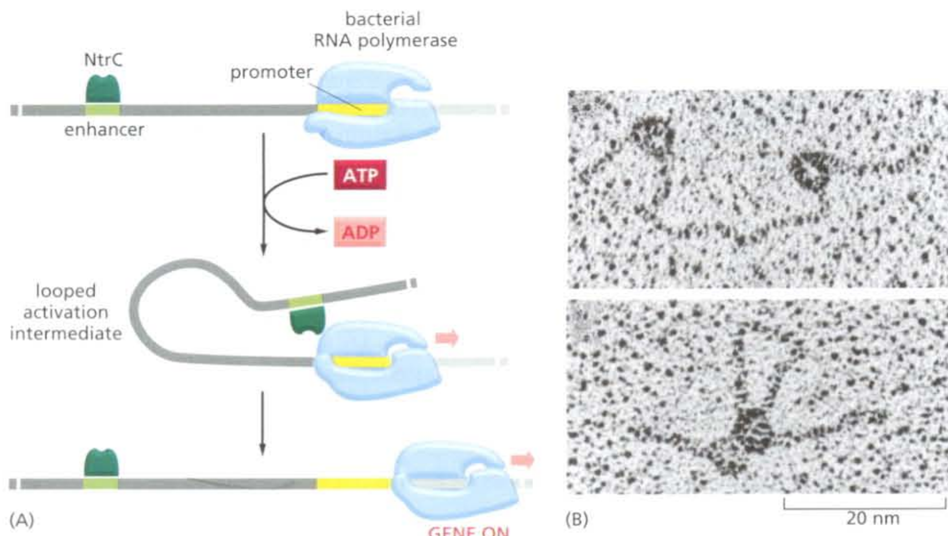


Figure 7–42 Gene activation at a distance. (A) NtrC is a bacterial gene regulatory protein that activates transcription by directly contacting RNA polymerase and causing a transition between the initial DNA-bound form of the polymerase and the transcriptionally competent form (discussed in Chapter 6). As indicated, the transition stimulated by NtrC requires the energy from ATP hydrolysis, although this requirement is unusual for bacterial transcription initiation. (B) The interaction of NtrC and RNA polymerase, with the intervening DNA looped out, can be seen in the electron microscope. (B, courtesy of Harrison Echols and Sydney Kustu.)

Table 7–2 Sigma Factors of *E. coli*

SIGMA FACTOR	PROMOTERS RECOGNIZED
σ^{70}	most genes
σ^{32}	genes induced by heat shock
σ^{38}	genes for stationary phase and stress response
σ^{28}	genes involved in motility and chemotaxis
σ^{54}	genes for nitrogen metabolism
σ^{24}	genes dealing with misfolded proteins in the periplasm

The sigma factor designations refer to their approximate molecular weights, in kilodaltons.

RNA chain. Although this is one of the main ways in which both eucaryotes and procaryotes control transcription initiation, some bacteria and their viruses use an additional strategy based on interchangeable subunits of RNA polymerase. As described in Chapter 6, a sigma (σ) subunit is required for the bacterial RNA polymerase to recognize a promoter. Most bacteria produce a whole range of sigma subunits, each of which can interact with the RNA polymerase core and direct it to a different set of promoters (**Table 7–2**). This scheme permits one large set of genes to be turned off and a new set to be turned on simply by replacing one sigma subunit with another; the strategy is efficient because it bypasses the need to deal with genes one by one. Indeed, some bacteria code for nearly one hundred different sigma subunits and therefore rely heavily on this form of gene regulation. Bacterial viruses often use it subversively to take over the host polymerase and activate several sets of viral genes rapidly and sequentially (**Figure 7–43**).

Complex Switches Have Evolved to Control Gene Transcription in Eucaryotes

Bacteria and eucaryotes share many principles of gene regulation, including the key role played by gene regulatory proteins that bind tightly to short stretches of DNA, the importance of weak protein–protein actions in gene activation, and the versatility afforded by DNA looping. However, by comparison, gene regulation in eucaryotes involves many more proteins, much longer stretches of DNA, and often seems bewilderingly complex. This increased complexity provides the eucaryotic cell with an important advantage. Genetic switches in bacteria, as we have seen, typically respond to one or a few signals. But in eucaryotes it is common for dozens of signals to converge on a single promoter, with the transcription machinery integrating all these different signals to produce the appropriate level of mRNA. We begin our description of eucaryotic gene regulation by outlining the main features that distinguish it from gene regulation in bacteria.

- As discussed in Chapter 6, eucaryotic RNA polymerase II, which transcribes all the protein-coding genes, requires five general transcription factors (27 subunits *in toto*, see Table 6–3, p. 341), whereas bacterial RNA polymerase needs only a single general transcription factor, the σ subunit. As we have seen, the stepwise assembly of the general transcription factors at a eucaryotic promoter provides, in principle, multiple steps at which the

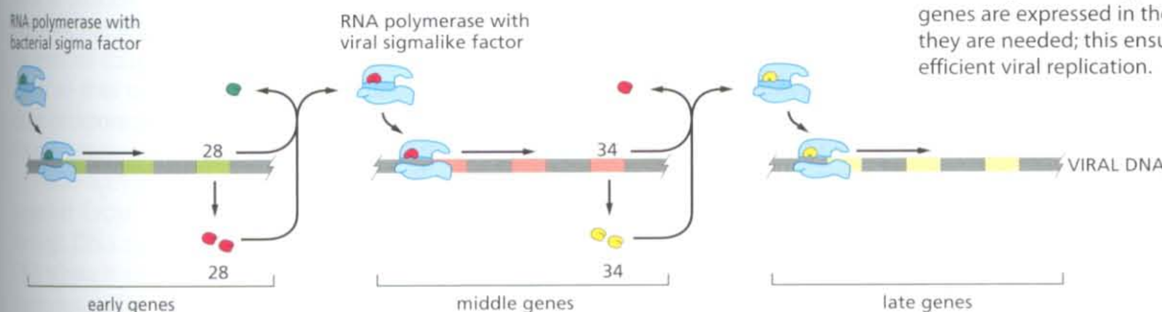


Figure 7–43 Interchangeable RNA polymerase subunits as a strategy to control gene expression in a bacterial virus. The bacterial virus SPO1, which infects the bacterium *B. subtilis,* uses the bacterial polymerase to transcribe its early genes immediately after the viral DNA enters the cell. One of the early genes, called 28, encodes a sigmalike factor that binds to RNA polymerase and displaces the bacterial sigma factor. This new form of polymerase specifically initiates transcription of the SPO1 "middle" genes. One of the middle genes encodes a second sigmalike factor, 34, that displaces the 28 product and directs RNA polymerase to transcribe the "late" genes. This last set of genes produces the proteins that package the virus chromosome into a virus coat and lyse the cell. By this strategy, sets of virus genes are expressed in the order in which they are needed; this ensures a rapid and efficient viral replication.

cell can speed up or slow down the rate of transcription initiation in response to gene regulatory proteins.

- Eucaryotic cells lack operons—sets of related genes transcribed as a unit—and therefore must regulate each gene individually.
- Each bacterial gene is typically controlled by one or only a few gene regulatory proteins, but it is common in eucaryotes for genes to be controlled by many (sometimes hundreds) of different regulatory proteins. This complexity is possible because, as we shall see, many eucaryotic gene regulatory proteins can act over very large distances (tens of thousands of nucleotide pairs) along DNA, allowing an almost unlimited number of them to influence the expression of a single gene.
- A central component of gene regulation in eucaryotes is *Mediator,* a 24-subunit complex, which serves as an intermediary between gene regulatory proteins and RNA polymerase (see Figure 6–19). Mediator provides an extended contact area for gene regulatory proteins compared to that provided by RNA polymerase alone, as in bacteria.
- The packaging of eucaryotic DNA into chromatin provides many opportunities for transcriptional regulation not available to bacteria.

Having discussed the general transcription factors for RNA polymerase II in Chapter 6 (see pp. 340–343), we focus here on the last four of these features and how they are used to control eucaryotic gene expression.

17th

A Eucaryotic Gene Control Region Consists of a Promoter Plus Regulatory DNA Sequences

Because the typical eucaryotic gene regulatory protein controls transcription when bound to DNA far away from the promoter, the DNA sequences that control the expression of a gene are often spread over long stretches of DNA. We use the term **gene control region** to describe the whole expanse of DNA involved in regulating and initiating transcription of a gene, including the **promoter**, where the general transcription factors and the polymerase assemble, and all of the **regulatory sequences** to which gene regulatory proteins bind to control the rate of the assembly processes at the promoter (**Figure 7–44**). In animals and plants, it is not unusual to find the regulatory sequences of a gene dotted over distances as great as 50,000 nucleotide pairs. Much of this DNA serves as "spacer" sequences that gene regulatory proteins do not directly recognize, but this DNA may provide the flexibility needed for efficient DNA looping. In this context, it is important to remember that, like other regions of eucaryotic chromosomes, most of the DNA in gene control regions is packaged into nucleosomes and higher-order forms of chromatin, thereby compacting its length and altering its properties.

In this chapter, we shall loosely use the term **gene** to refer only to a segment of DNA that is transcribed into RNA (see Figure 7–44). However, the classical view of a gene includes the gene control region as well, making most eucaryotic genes considerably larger. The discovery of alternative RNA splicing has further complicated the definition of a gene—a point we discussed briefly in Chapter 6 and will return to later in this chapter.

It is the gene regulatory proteins that allow the genes of an organism to be turned on or off individually. In contrast to the small number of general transcription factors, which are abundant proteins that assemble on the promoters of all genes transcribed by RNA polymerase II, there are thousands of different gene regulatory proteins. For example, of the roughly 25,000 human genes, an estimated 8% (~2000 genes) encode gene regulatory proteins. Most of these recognize DNA sequences using one of the DNA-binding motifs described previously. Not surprisingly, the eucaryotic cell regulates each of its many genes in a unique way. Given the sheer number of genes in eucaryotes and the complexity of their regulation, it has been difficult to formulate simple rules for gene regulation that apply in every case. We can, however, make some generalizations about how gene regulatory proteins, once bound to gene control regions on DNA, set in motion the train of events that lead to gene activation or repression.

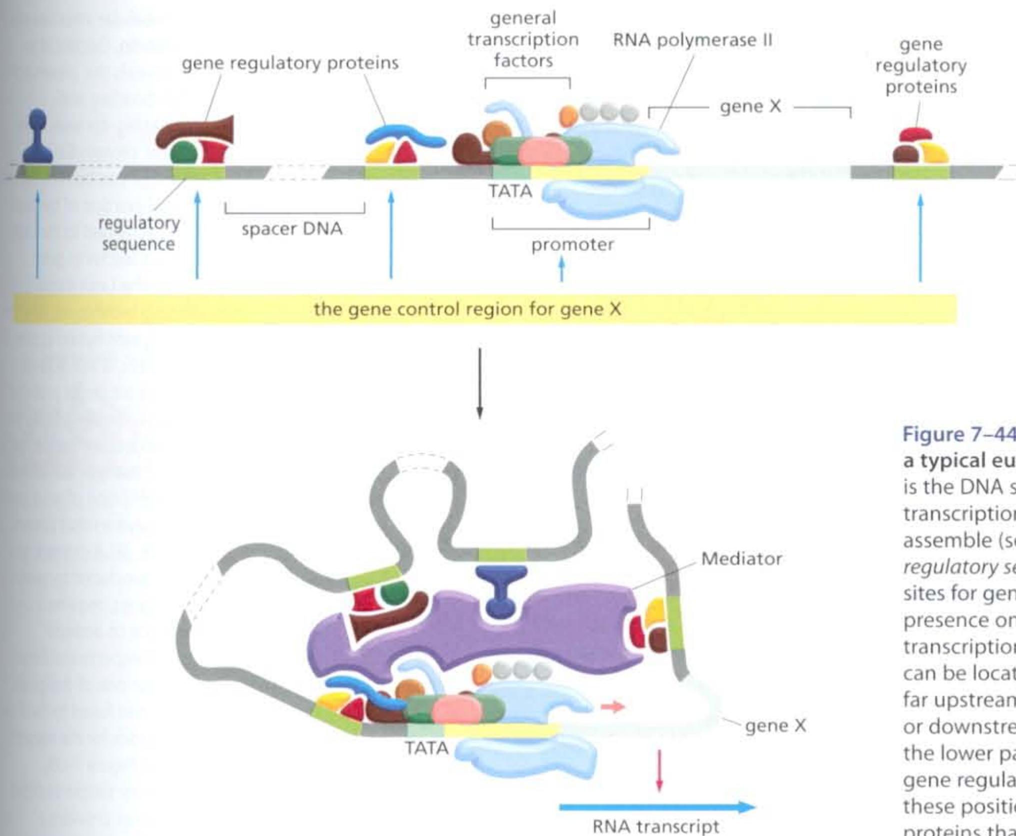


Figure 7–44 The gene control region for a typical eucaryotic gene. The *promoter* is the DNA sequence where the general transcription factors and the polymerase assemble (see Figure 6–16). The *regulatory sequences* serve as binding sites for gene regulatory proteins, whose presence on the DNA affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or downstream of the gene. As shown in the lower panel, DNA looping allows gene regulatory proteins bound at any of these positions to interact with the proteins that assemble at the promoter. Many gene regulatory proteins act through Mediator, while others influence the general transcription factors and RNA polymerase directly. Although not shown here, many gene regulatory proteins also influence the chromatin structure of the DNA control region thereby affecting transcription initiation indirectly (see Figure 4–45). As noted in the text, for simplicity, "gene X" refers here to the coding sequence within the gene.

Whereas Mediator and the general transcription factors are the same for all polymerase II transcribed genes, the gene regulatory proteins and the locations of their binding sites relative to the promoter differ for each gene.

Eucaryotic Gene Activator Proteins Promote the Assembly of RNA Polymerase and the General Transcription Factors at the Startpoint of Transcription

The DNA sites to which eucaryotic gene activator proteins bind were originally called *enhancers* because their presence "enhanced" the rate of transcription initiation. It came as a surprise when it was first discovered that these activator proteins could be bound tens of thousands of nucleotide pairs away from the promoter, but, as we have seen, DNA looping provides at least one explanation for this initially puzzling observation.

The simplest gene activator proteins have a modular design consisting of two distinct domains. One domain usually contains one of the structural motifs discussed previously that recognizes a specific DNA sequence. The second domain—sometimes called an *activation domain*—accelerates the rate of transcription initiation. This type of modular design was first revealed by experiments in which genetic engineering techniques were used to create a chimeric protein containing the activation domain of one protein fused to the DNA-binding domain of a different protein (**Figure 7–45**).

Once bound to DNA, how do eucaryotic gene activator proteins increase the rate of transcription initiation? As we will see shortly, there are several mechanisms by which this can occur, and, in many cases, these different mechanisms work in concert at a single promoter. But, regardless of the precise biochemical pathway, the ultimate function of activators is to attract, position, and modify the general transcription factors, Mediator, and RNA polymerase II at the promoter so that transcription can begin. They do this both by acting directly on these components and, indirectly, by changing the chromatin structure around the promoter.

Some activator proteins bind directly to one or more of the general transcription factors, accelerating their assembly on a promoter that is linked through DNA to that activator. Others interact with Mediator and attract it to DNA where it can then facilitate assembly of RNA polymerase and the general transcription factors at the promoter (see Figure 7–44). In this sense, eucaryotic

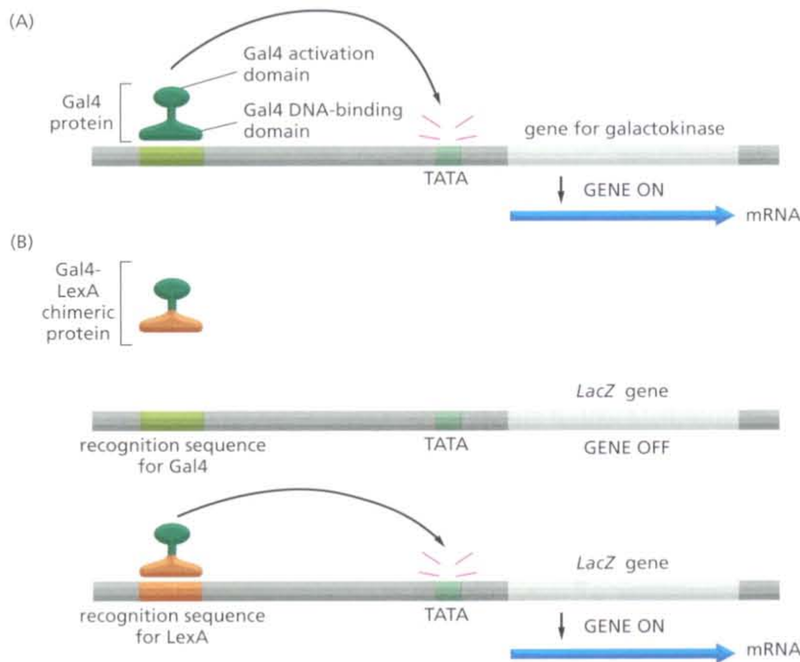


Figure 7–45 The modular structure of a gene activator protein. Outline of an experiment that reveals the presence of independent DNA-binding and transcription-activating domains in the yeast gene activator protein Gal4. A functional activator can be reconstituted from the C-terminal portion of the yeast Gal4 protein if it is attached to the DNA-binding domain of a bacterial gene regulatory protein (the LexA protein) by genetic engineering techniques. When the resulting bacterial–yeast hybrid protein is produced in yeast cells, it will activate transcription from yeast genes provided that the specific DNA-binding site for the bacterial protein has been inserted next to them. (A) Gal4 is normally responsible for activating the transcription of yeast genes that code for the enzymes that convert galactose to glucose. (B) A chimeric gene regulatory protein, produced by genetic engineering techniques, requires a LexA recognition sequence to activate transcription. In the experiment shown, the control region for one of the genes controlled by LexA was fused to the *E. coli* *LacZ* gene, which codes for the enzyme β-galactosidase (see Figure 7–39). β-Galactosidase is very simple to detect biochemically and thus provides a convenient way to monitor the expression level specified by a gene control region; used here, *LacZ* is said to serve as a *reporter gene,* since it "reports" the activity of a gene control region.

activators resemble those of bacteria in recruiting RNA polymerase to specific sites on DNA so it can begin transcribing.

Eucaryotic Gene Activator Proteins Also Modify Local Chromatin Structure

The general transcription factors, Mediator, and RNA polymerase seem unable on their own to assemble on a promoter that is packaged in standard nucleosomes. Indeed, it has been proposed such packaging may have evolved to prevent "leaky" transcription. In addition to their direct actions in assembling the transcription machinery at the promoter, gene activator proteins also promote transcription initiation by changing the chromatin structure of the regulatory sequences and promoters of genes.

As we saw in Chapter 4, four of the most important ways of locally altering chromatin are through covalent histone modifications, nucleosome remodeling, nucleosome removal, and nucleosome replacement. Gene activator proteins use all four of these mechanisms by attracting histone modification enzymes, ATP-dependent chromatin remodeling complexes, and histone chaperones to alter the chromatin structure of promoters they control (**Figure 7–46**). In general terms, these local alterations in chromatin structure are believed to make the underlying DNA more accessible, thereby facilitating the assembly of the general transcription factors, Mediator, and RNA polymerase at the promoter. Local chromatin modification also allows additional gene regulatory proteins to bind to the control region of the gene. However, the most important role of covalent histone modifications in transcription is probably not in directly changing chromatin structure; rather, as discussed in Chapter 4, these modifications provide favorable interactions for the binding of a large set of proteins that read a "histone code." For transcription initiation, these proteins include other histone-modifying enzymes (reader–writer complexes), chromatin remodeling complexes, and at least one of the general transcription factors (**Figure 7–47**).

The alterations of chromatin structure that occur during transcription initiation can persist for variable lengths of time. In some cases, as soon as the gene regulatory protein dissociates from DNA, the chromatin modifications are rapidly reversed, restoring the gene to its pre-activated state. This rapid reversal is especially important for genes that the cell must quickly switch on and off in

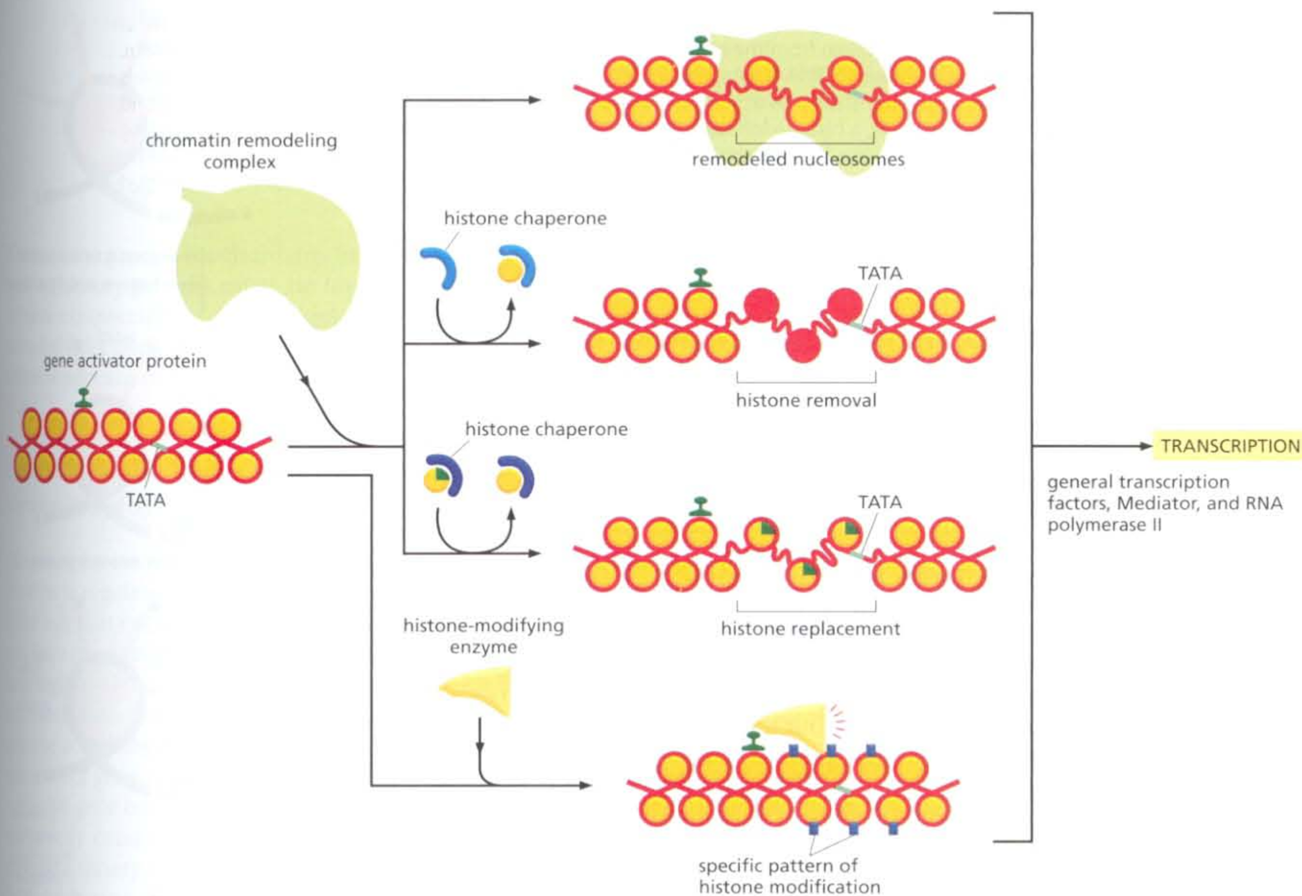


Figure 7–46 Four ways eucaryotic activator proteins can direct local alterations in chromatin structure to stimulate transcription initiation. Although shown as separate pathways, these mechanisms often work together during the activation of a gene. For example, prior acetylation of histones makes it easier for histone chaperones to remove them from nucleosomes. A few patterns of histone modification that promote transcription initiation are listed in Figure 4–44, and a specific example is given in Figure 7–47. Nucleosome remodeling and histone removal favor transcription initiation by increasing the accessibility of DNA and thereby facilitating the binding of Mediator, RNA polymerase, and the general transcription factors as well as additional activator proteins. Transcription initiation and the formation of a compact chromatin structure can be regarded as competing biochemical assembly reactions, and enzymes that increase—even transiently—the accessibility of DNA in chromatin will tend to favor transcription initiation.

response to external signals, such as the glucocorticoid hormone discussed earlier in this chapter. However, in other cases, the altered chromatin structure seems to persist, even after the gene regulatory protein that directed its establishment has dissociated from DNA. In principle, this memory can extend into the next cell generation because, as discussed in Chapter 4, chromatin structure can be self-renewing (see Figure 4–52). It is interesting to consider the possibility that different histone modifications persist for different times in order to provide the cell with a mechanism for long-, medium-, and short-term memory of gene expression patterns.

A special type of chromatin modification occurs as RNA polymerase II transcribes through a gene. In most cases, the nucleosomes just ahead of the polymerase are acetylated by writer complexes carried by the polymerase, removed by histone chaperones, and deposited behind the moving polymerase. They are then rapidly deacetylated and methylated, also by reader–writer complexes that are carried by the polymerase, leaving behind nucleosomes that are especially resistant to transcription. Although this remarkable process may seem counter-intuitive, it likely evolved to prevent spurious transcription re-initiation behind a moving polymerase, which is, in essence, clearing a path through chromatin. Later in this chapter, when we discuss *RNA interference*, the potential dangers to the cell of such inappropriate transcription will become especially obvious.

We have just seen that gene activator proteins can profoundly influence chromatin structure. However, even before these activator proteins are brought into play, many genes are "poised" to become rapidly activated. For example, the regulatory regions for many genes are "marked" by a short, nucleosome-free region flanked by nucleosomes that contain the histone variant H2AZ. This arrangement, which is specified by DNA sequence, allows free access of gene regulatory proteins to the nucleosome-free region; in addition, the H2AZ-containing nucleosomes are thought to be easily disassembled, thus further facilitating transcription initiation.

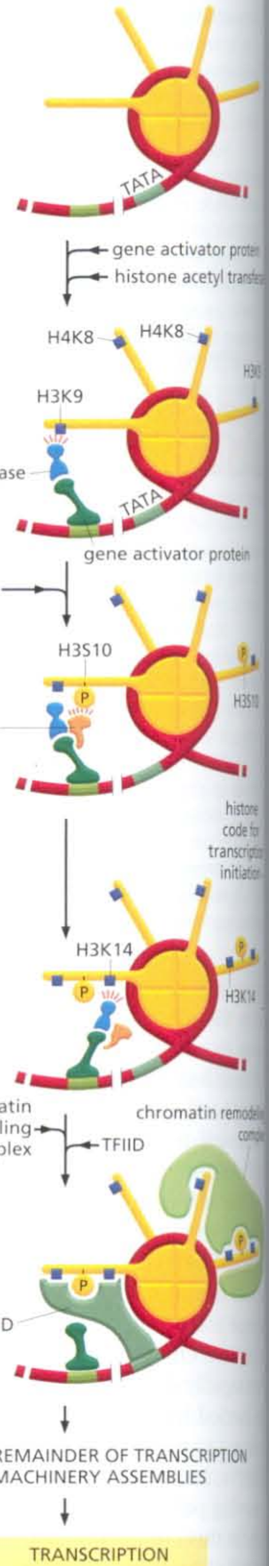


Figure 7–47 Writing and reading the histone code during transcription initiation. In this example, taken from the human interferon gene promoter, a gene activator protein binds to DNA packaged into chromatin and first attracts a histone acetyl transferase to acetylate lysine 9 of histone H3 and lysine 8 of histone H4. Next, a histone kinase, attracted by the gene activator protein, phosphorylates serine 10 of histone H3, but can only do so after lysine 9 has been acetylated. The serine modification then signals the histone acetyl transferase to acetylate position K14 of histone H3. At this point the histone code for transcription initiation, set into motion by the binding of the gene activator protein, has been written. Note that the writing is sequential, with each histone modification depending on a prior modification.

The final reading of the code occurs when the general transcription factor TFIID and the chromatin remodeling complex SWI/SNF bind, both of which strongly promote the subsequent steps of transcription initiation. TFIID and SWI/SNF both recognize acetylated histone tails through a *bromodomain*, a protein domain specialized to read this particular mark on histones; a bromodomain is carried in a subunit of each protein complex. (Adapted from T. Agalioti, G. Chen and D. Thanos, *Cell* 111:381–392, 2002. With permission from Elsevier.)

Gene Activator Proteins Work Synergistically

We have seen that eucaryotic gene activator proteins can influence different steps in transcription initiation. In general, where several factors work together to enhance a reaction rate, the joint effect is not merely the sum of the enhancements that each factor alone contributes, but the product. If, for example, factor A lowers the free-energy barrier for a reaction by a certain amount and thereby speeds up the reaction 100-fold, and factor B, by acting on another aspect of the reaction, does likewise, then A and B acting in parallel will lower the barrier by a double amount and speed up the reaction 10,000-fold. Even if A and B work simply by attracting the same protein, the affinity of that protein for the reaction site increases multiplicatively. Thus, gene activator proteins often exhibit *transcriptional synergy*, where several activator proteins working together produce a transcription rate that is much higher than that of the sum of the activators working alone (**Figure 7–48**). It is not difficult to see how multiple gene regulatory proteins, each binding to a different regulatory DNA sequence, work together to control the final rate of transcription of a eucaryotic gene.

Since gene activator proteins can influence many different steps on the pathway to transcriptional activation, it is worth considering whether these steps always occur in a prescribed order. For example, does histone modification always precede chromatin remodeling, as in the example of Figure 7–47? Does Mediator enter before or after RNA polymerase? The answers to these questions appear to be different for different genes—and even for the same gene under the influence of different gene regulatory proteins (**Figure 7–49**).

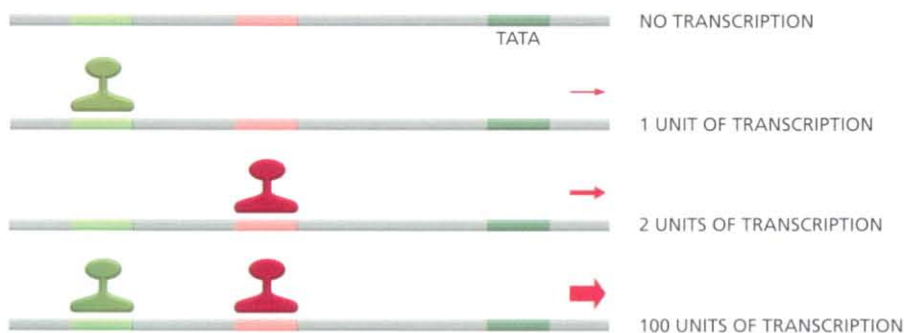


Figure 7–48 Transcriptional synergy. This experiment compares the rate of transcription produced by three experimentally constructed regulatory regions in a eucaryotic cell and reveals transcriptional synergy, the greater than additive effect of multiple activators. Transcriptional synergy is typically observed between different gene activator proteins from the same organism and even between activator proteins from different eucaryotic species when they are experimentally introduced into the same cell. This last observation reflects the high degree of conservation of the elaborate machinery responsible for eucaryotic transcription initiation.

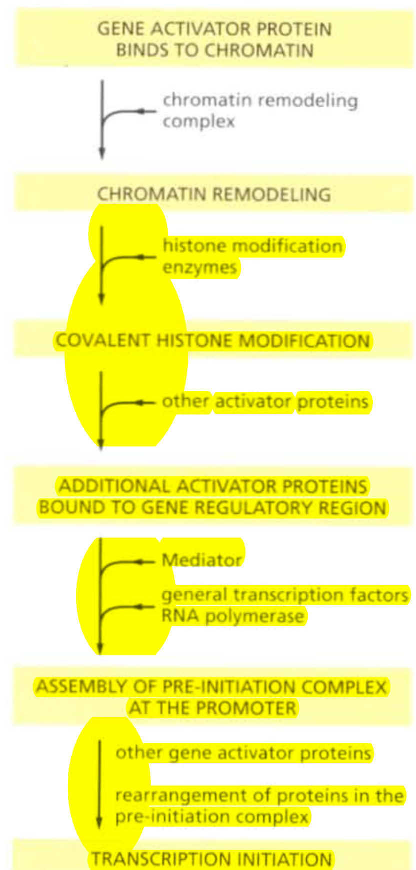


Figure 7–49 An order of events leading to transcription initiation of a specific gene. In this well-studied example from the budding yeast *S. cerevisiae,* the steps toward transcription initiation occur in a particular order; however, this order differs from one gene to the next. For example, at another gene, histone modification occurs first, followed by RNA polymerase recruitment, followed by chromatin remodeling complex recruitment. Figure 7–47 illustrates yet another possible order of events.

Whatever the precise mechanisms and the order in which they are carried out, a gene regulatory protein must be bound to DNA either directly or indirectly to influence transcription of its target gene, and the rate of transcription of a gene ultimately depends upon the spectrum of regulatory proteins bound upstream and downstream of its transcription start site.

Eucaryotic Gene Repressor Proteins Can Inhibit Transcription in Various Ways

Like bacteria, eucaryotes use **gene repressor proteins** in addition to activator proteins to regulate transcription of their genes. However, because of differences in the way that eucaryotes and bacteria initiate transcription, eucaryotic repressors have many more possible mechanisms of action. We saw in Chapter 4 that large regions of the genome can be shut down by the packaging of DNA into heterochromatin. However, eucaryotic genes are rarely organized along the genome according to function, so this strategy is not generally useful for most examples of gene regulation. Instead, most eucaryotic repressors must work on a gene-by-gene basis. Unlike bacterial repressors, most eucaryotic repressors do not directly compete with the RNA polymerase for access to the DNA; rather they use a variety of other mechanisms, some of which are illustrated in **Figure 7–50**. Like gene activator proteins, many eucaryotic repressor proteins act through more than one mechanism at a given target gene, thereby ensuring robust and efficient repression.

Gene repression is especially important to animals and plants whose growth depends on elaborate and complex developmental programs. Misexpression of a single gene at a critical time can have disastrous consequences for the individual. For this reason, many of the genes encoding the most important developmental regulatory proteins are kept tightly repressed when they are not needed.

Eucaryotic Gene Regulatory Proteins Often Bind DNA Cooperatively

So far we have seen that when eucaryotic activator and repressor proteins bind to specific DNA sequences, they set in motion a complex series of events that culminate in transcription initiation or its opposite, repression. However, these proteins rarely recognize DNA as individual polypeptides. In reality, efficient DNA binding in the eucaryotic cell typically requires several sequence-specific DNA proteins acting together. For example, two gene regulatory proteins with a weak affinity for each other might cooperate to bind to a DNA sequence, neither protein having a sufficient affinity for DNA to bind to the DNA site on its own. In one well-studied case, the DNA-bound protein dimer creates a distinct surface that is recognized by a third protein that carries an activator domain that stimulates transcription. This example illustrates an important general point: protein–protein interactions that are too weak to form complexes in solution can do so on DNA, with the DNA sequence acting as a "crystallization" site or seed for the assembly of a protein complex.

As shown in **Figure 7–51**, an individual gene regulatory protein can often participate in more than one type of regulatory complex. A protein might function, for example, in one case as part of a complex that activates transcription and in another case as part of a complex that represses transcription. Thus,

individual eucaryotic gene regulatory proteins are not necessarily dedicated activators or repressors; instead, they function as regulatory parts that are used to build complexes whose function depends on the final assembly of all of the individual components. This final assembly, in turn, depends both on the arrangement of control region DNA sequences and on the particular gene regulatory proteins present in active form in the cell. Each eucaryotic gene is therefore regulated by a "committee" of proteins, all of which must be present to express the gene at its proper level.

In some cases, the precise DNA sequence to which a regulatory protein binds directly can affect the conformation of this protein and thereby influence its subsequent transcriptional activity. When bound to one type of DNA sequence, for example, a steroid hormone receptor protein interacts with a co-repressor and ultimately turns off transcription. When bound to a slightly different DNA sequence, it assumes a different conformation and interacts with a coactivator, thereby stimulating transcription.

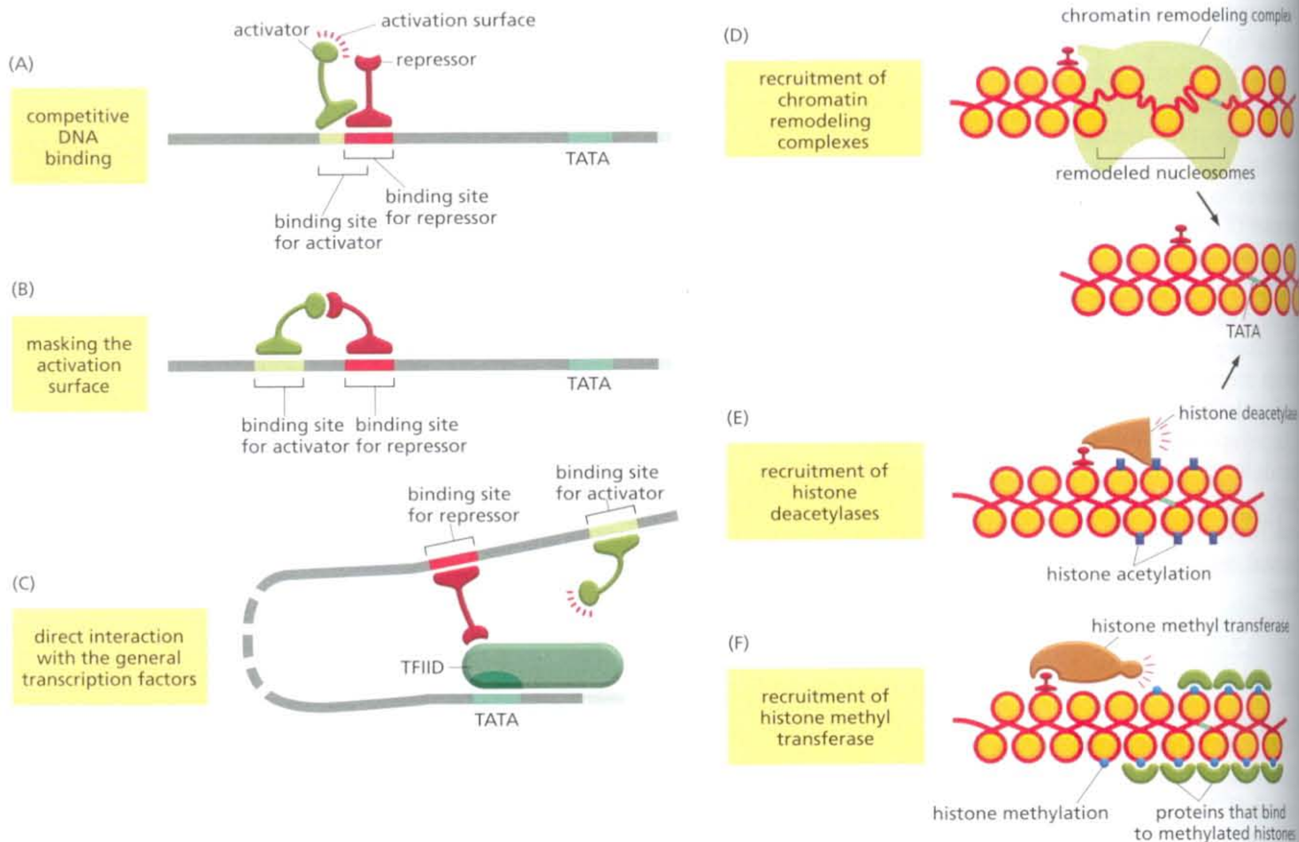


Figure 7–50 Six ways in which eucaryotic gene repressor proteins can operate. (A) Gene activator proteins and gene repressor proteins compete for binding to the same regulatory DNA sequence. (B) Both proteins can bind DNA, but the repressor binds to the activation domain of the activator protein, thereby preventing it from carrying out its activation functions. In a variation of this strategy, the repressor binds tightly to the activator without having to be bound to DNA directly. (C) The repressor blocks assembly of the general transcription factors. Some repressors also act at late stages in transcription initiation, for example, by preventing the release of the RNA polymerase from the general transcription factors. (D) The repressor recruits a chromatin remodeling complex which returns the nucleosomal state of the promoter region to its pre-transcriptional form. (E) The repressor attracts a histone deacetylase to the promoter. As we have seen, histone acetylation can stimulate transcription initiation (Figure 7–47), and the repressor simply reverses this modification. (F) The repressor attracts a histone methyl transferase which modifies certain positions on histones which, in turn, are bound by proteins that maintain the chromatin in a transcriptionally silent form. For example, in *Drosophila*, the histone methyl transferase Suv39 methylates the K9 position of histone H3, a modification that is bound by the HP1 protein. In another example, E(z) methylates the K27 position of H3, and this modification is bound by the Polycomb protein. HP1 and Polycomb recognize methylated lysines through a *chromodomain*. They can act locally to turn off specific genes, as shown here, or can occupy a whole region of a chromosome to repress a cluster of genes. A seventh mechanism of negative control—inactivation of a transcriptional activator by heterodimerization—is illustrated in Figure 7–24. For simplicity, nucleosomes have been omitted from (A)–(C), and the scales of (D)–(F) have been reduced relative to (A)–(C).

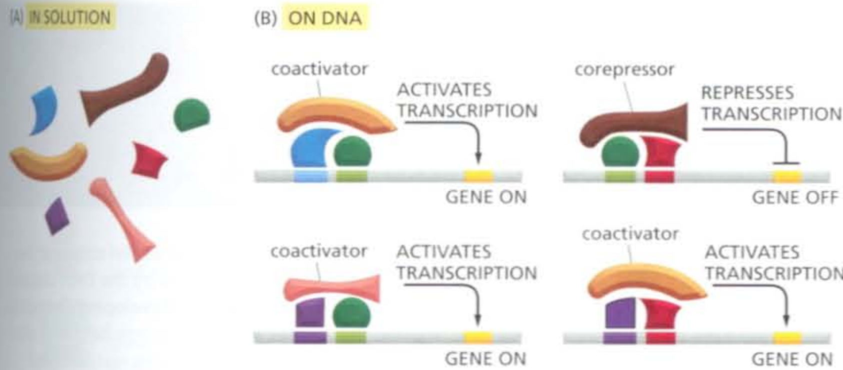


Figure 7–51 Eucaryotic gene regulatory proteins often assemble into complexes on DNA. Seven gene regulatory proteins are shown in (A). The nature and function of the complex they form depends on the specific DNA sequence that seeds their assembly. In (B), some assembled complexes activate gene transcription, while another represses transcription. Note that both the *red* and the *green* proteins are shared by both activating and repressing complexes. Proteins that do not themselves bind DNA but assemble on other DNA-bound gene regulatory proteins are often termed coactivators or co-repressors. However, these terms are somewhat confusing because they encompass an enormous variety of proteins including histone readers and writers, chromatin remodeling complexes, and many other classes of proteins. Some have no intrinsic activity themselves but simply serve as a "scaffolding" to attract those that do.

Typically, a few relatively short stretches of nucleotide sequence guide the assembly of a group of regulatory proteins on DNA (see Figure 7–51). However, in some extreme cases of regulation by committee a more elaborate protein–DNA structure is formed (**Figure 7–52**). Since the final assembly requires the presence of many gene regulatory proteins that bind DNA, it provides a simple way to ensure that a gene is expressed only when the cell contains the correct combination of these proteins. We saw earlier how the formation of heterodimers in solution provides a mechanism for the combinatorial control of gene expression. The assembly of complexes of gene regulatory proteins on DNA provides a second important mechanism for combinatorial control, one that offers far richer opportunities.

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

Given that gene regulatory proteins can be positioned at multiple sites along long stretches of DNA, that these proteins can assemble into complexes at each site, and that the complexes influence the chromatin structure as well as the recruitment and assembly of the general transcription machinery at the promoter, there would seem to be almost limitless possibilities for the elaboration of control devices to regulate eucaryotic gene transcription.

A particularly striking example of a complex, multicomponent genetic switch is that controlling the transcription of the *Drosophila Even-skipped* (*Eve*) 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 embryo dies early in development. As discussed in Chapter 22, at the stage of development when *Eve* begins to be expressed, the embryo is a single giant cell containing multiple nuclei in a common cytoplasm. This cytoplasm is not uniform, however: it contains a mixture of gene regulatory proteins that are distributed unevenly along the length of the embryo, thus providing positional information that distinguishes one part of the embryo from another (**Figure 7–53**). (The way these differences are initially set up is discussed in Chapter 22.) Although the nuclei are initially identical, they rapidly begin to express different genes because they are exposed to different gene regulatory proteins. The nuclei near the anterior end of the developing embryo, for example, are exposed to a set of gene regulatory proteins that is distinct from the set that influences nuclei at the posterior end of the embryo.

The regulatory DNA sequences controlling the *Eve* gene are designed to read the concentrations of gene regulatory proteins at each position along the length of the embryo and to interpret this information in such a way that the *Eve* gene is expressed in seven stripes, each initially five to six nuclei wide and positioned precisely along the anterior–posterior axis of the embryo (**Figure 7–54**). How is this remarkable feat of information processing carried out? Although not all of the molecular details are understood, several general principles have emerged from studies of *Eve* and other *Drosophila* genes that are similarly regulated.

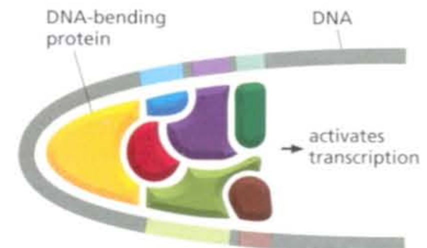


Figure 7–52 Schematic depiction of a committee of gene regulatory proteins bound to an enhancer. The protein shown in *yellow* is called an architectural protein since its main role is to bend the DNA to allow the cooperative assembly of the other components. The structure depicted here is based on that found in the control region of the gene that codes for a subunit of the T cell receptor (discussed in Chapter 25), and it activates transcription at a nearby promoter. Only certain cells of the developing immune system, which eventually give rise to mature T cells, have the complete set of proteins needed to form this structure.

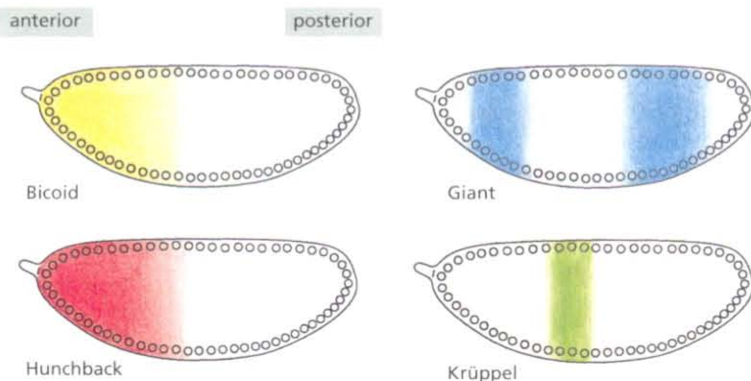


Figure 7–53 The nonuniform distribution of four gene regulatory proteins in an early *Drosophila* embryo. At this stage the embryo is a syncytium, with multiple nuclei in a common cytoplasm. Although the detail is not shown in these drawings, all of these proteins are concentrated in the nuclei.

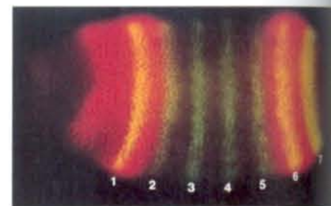
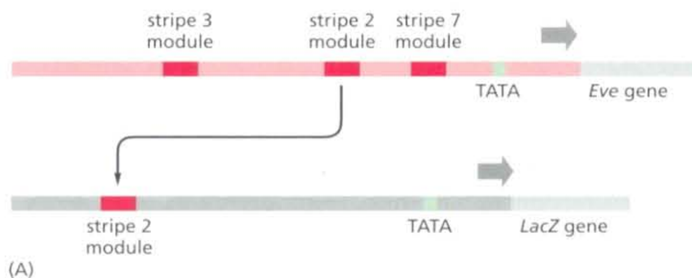


Figure 7–54 The seven stripes of the protein encoded by the *Even-skipped (Eve)* gene in a developing *Drosophila* embryo. Two and one-half hours after fertilization, the egg was fixed and stained with antibodies that recognize the Eve protein *(green)* and antibodies that recognize the Giant protein *(red)*. Where Eve and Giant proteins are both present, the staining appears *yellow*. At this stage in development, the egg contains approximately 4000 nuclei. The Eve and Giant proteins are both located in the nuclei, and the Eve stripes are about four nuclei wide. The staining pattern of the Giant protein is also shown in Figure m7-52/7-53. (Courtesy of Michael Levine.)

The regulatory region of the *Eve* gene is very large (approximately 20,000 nucleotide pairs). It is formed from a series of relatively simple regulatory modules, each of which contains multiple regulatory sequences and is responsible for specifying a particular stripe of *Eve* expression along the embryo. This modular organization of the *Eve* gene control region is revealed by experiments in which a particular regulatory module (say, that specifying stripe 2) is removed from its normal setting upstream of the *Eve* gene, placed in front of a reporter gene (see Figure 7–45), and reintroduced into the *Drosophila* genome. When developing embryos derived from flies carrying this genetic construct are examined, the reporter gene is found to be expressed in precisely the position of stripe 2 (**Figure 7–55**). Similar experiments reveal the existence of other regulatory modules, each of which specifies either one of the other six stripes or some other part of the *Eve* expression pattern normally displayed at later stages of development (see Figure 22–39).

The *Drosophila Eve* Gene Is Regulated by Combinatorial Controls

A detailed study of the stripe 2 regulatory module has provided insights into how it reads and interprets positional information. It contains recognition sequences for two gene regulatory proteins (Bicoid and Hunchback) that activate *Eve* transcription and two (Krüppel and Giant) that repress it (**Figure 7–56**). (The gene regulatory proteins of *Drosophila* often have colorful names reflecting the phenotype that results if the gene encoding the protein is inactivated by mutation.) The relative concentrations of these four proteins determine whether the protein complexes that form at the stripe 2 module activate transcription of the *Eve* gene. **Figure 7–57** shows the distributions of the four gene regulatory proteins across the region of a *Drosophila* embryo where stripe 2 forms. It is thought that either of the two repressor proteins, when bound to the DNA, will turn off the



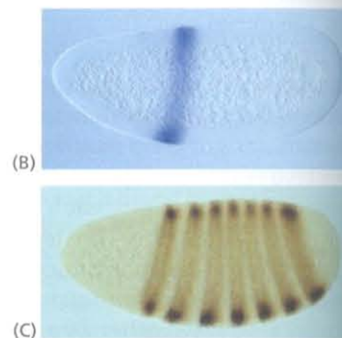


Figure 7–55 Experiment demonstrating the modular construction of the *Eve* gene regulatory region. (A) A 480-nucleotide-pair piece of the Eve regulatory region was removed and inserted upstream of a test promoter that directs the synthesis of the enzyme β-galactosidase (the product of the *E. coli LacZ* gene). (B) When this artificial construct was reintroduced into the genome of *Drosophila* embryos, the embryos expressed β-galactosidase (detectable by histochemical staining) precisely in the position of the second of the seven Eve stripes (C). (B and C, courtesy of Stephen Small and Michael Levine.)

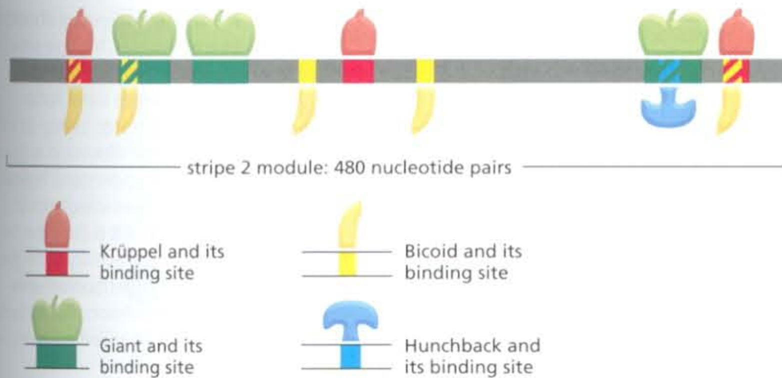


Figure 7–56 Close-up view of the *Eve* stripe 2 unit. The segment of the *Eve* gene control region identified in the previous figure contains regulatory sequences, each of which binds one or another of four gene regulatory proteins. It is known from genetic experiments that these four regulatory proteins are responsible for the proper expression of *Eve* in stripe 2. Flies that are deficient in the two gene activators Bicoid and Hunchback, for example, fail to express efficiently *Eve* in stripe 2. In flies deficient in either of the two gene repressors, Giant and Krüppel, stripe 2 expands and covers an abnormally broad region of the embryo. The DNA binding sites for these gene regulatory proteins were determined by cloning the genes encoding the proteins, overexpressing the proteins in *E. coli*, purifying them, and performing DNA-footprinting experiments (see Figure 7–29). The *top* diagram indicates that, in some cases, the binding sites for the gene regulatory proteins overlap and the proteins can compete for binding to the DNA. For example, binding of Krüppel and binding of Bicoid to the site at the far right are thought to be mutually exclusive.

stripe 2 module, whereas both Bicoid and Hunchback must bind for its maximal activation. This simple regulatory unit thereby combines these four positional signals so as to turn on the stripe 2 module (and therefore the expression of the *Eve* gene) only in those nuclei that are located where the levels of both Bicoid and Hunchback are high and both Krüppel and Giant are absent. This combination of activators and repressors occurs in only one region of the early embryo; everywhere else, therefore, the stripe 2 module is silent.

We have thus far discussed two mechanisms of combinatorial control of gene expression—heterodimerization of gene regulatory proteins in solution (see Figure 7–20) and the assembly of combinations of gene regulatory proteins into small complexes on DNA (see Figure 7–51). It is likely that both mechanisms participate in the complex regulation of *Eve* expression. In addition, the regulation of stripe 2 just described illustrates a third type of combinatorial control.

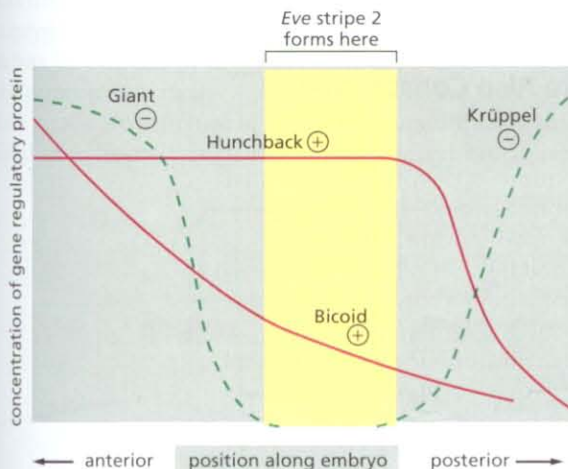


Figure 7–57 Distribution of the gene regulatory proteins responsible for ensuring that *Eve* is expressed in stripe 2. The distributions of these proteins were visualized by staining a developing *Drosophila* embryo with antibodies directed against each of the four proteins (see Figures 7–53 and 7–54). The expression of *Eve* in stripe 2 occurs only at the position where the two activators (Bicoid and Hunchback) are present and the two repressors (Giant and Krüppel) are absent. In fly embryos that lack Krüppel, for example, stripe 2 expands posteriorly. Likewise, stripe 2 expands posteriorly if the DNA-binding sites for Krüppel in the stripe 2 module (see Figure 7–56) are inactivated by mutation.

The *Eve* gene itself encodes a gene regulatory protein, which, after its pattern of expression is set up in seven stripes, regulates the expression of other *Drosophila* genes. As development proceeds, the embryo is thus subdivided into finer and finer regions that eventually give rise to the different body parts of the adult fly, as discussed in Chapter 22.

This example from *Drosophila* embryos is unusual in that the nuclei are exposed directly to positional cues in the form of concentrations of gene regulatory proteins. In embryos of most other organisms, individual nuclei are in separate cells, and extracellular positional information must either pass across the plasma membrane or, more usually, generate signals in the cytosol in order to influence the genome.

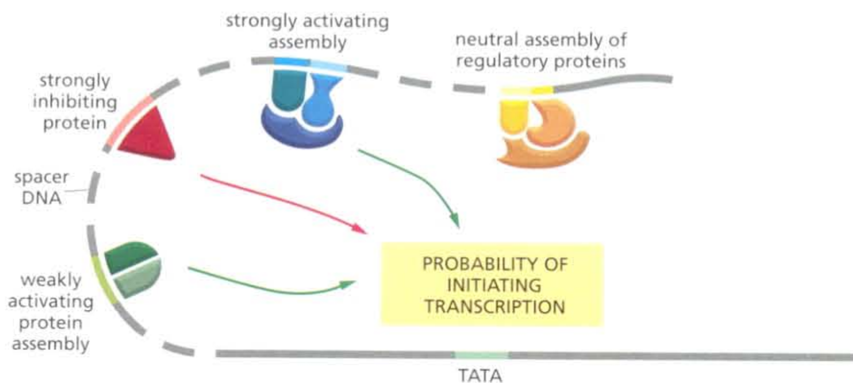


Figure 7–58 The integration of multiple inputs at a promoter. Multiple sets of gene regulatory proteins can work together to influence transcription initiation at a promoter, as they do in the *Eve* stripe 2 module illustrated in Figure 7–56. It is not yet understood in detail how the cell achieves integration of multiple inputs, but it is likely that the final transcriptional activity of the gene results from a competition between activators and repressors that act by the mechanisms summarized in Figures 7– and 7–50.

Because the individual regulatory sequences in the *Eve* stripe 2 module are strung out along the DNA, many sets of gene regulatory proteins can be bound simultaneously at separate sites and influence the promoter of a gene. The promoter integrates the transcriptional cues provided by all of the bound proteins (**Figure 7–58**).

The regulation of *Eve* expression is an impressive example of combinatorial control. Seven combinations of gene regulatory proteins—one combination for each stripe—activate *Eve* expression, while many other combinations (all those found in the interstripe regions of the embryo) keep the stripe elements silent. The other stripe regulatory modules are thought to be constructed similarly to those described for stripe 2, being designed to read positional information provided by other combinations of gene regulatory proteins. The entire gene control region, strung out over 20,000 nucleotide pairs of DNA, binds more than 20 different regulatory proteins. A large and complex control region is thereby built from a series of smaller modules, each of which consists of a unique arrangement of short DNA sequences recognized by specific gene regulatory proteins.

Complex Mammalian Gene Control Regions Are Also Constructed from Simple Regulatory Modules

Perhaps 8% of the coding capacity of a mammalian genome is devoted to the synthesis of proteins that serve as regulators of gene transcription. This large number of genes reflects the exceedingly complex network of controls governing expression of mammalian genes. Each gene is regulated by a set of gene regulatory proteins; each of those proteins is the product of a gene that is in turn regulated by a whole set of other proteins, and so on. Moreover, the regulatory protein molecules are themselves influenced by signals from outside the cell, which can make them active or inactive in a whole variety of ways (**Figure 7–59**). Thus, we can view the pattern of gene expression in a cell as the result of a complicated molecular computation that the intracellular gene control network performs in response to information from the cell's surroundings. We shall discuss these issues further in Chapters 15 and 22, which deal with cell signaling and development, but the complexity is remarkable even at the level of an individual genetic switch regulating the activity of a single gene. It is not unusual, for example, to find a mammalian gene with a control region that is 100,000 nucleotide pairs in length, in which many modules, each containing a number of regulatory sequences that bind gene regulatory proteins, are interspersed with long stretches of other noncoding DNA.

One of the best-understood examples of a complex mammalian regulatory region is found in the human β-globin gene, which is expressed exclusively in red blood cells. A complex array of gene regulatory proteins controls the expression of the gene, some acting as activators and others as repressors (**Figure 7–60**). The concentrations (or activities) of many of these gene regulatory proteins change during development, and only a particular combination of all the proteins triggers transcription of the gene. The human β-globin gene is part of a cluster of globin genes (**Figure 7–61**A) that are all transcribed exclusively in erythroid cells,

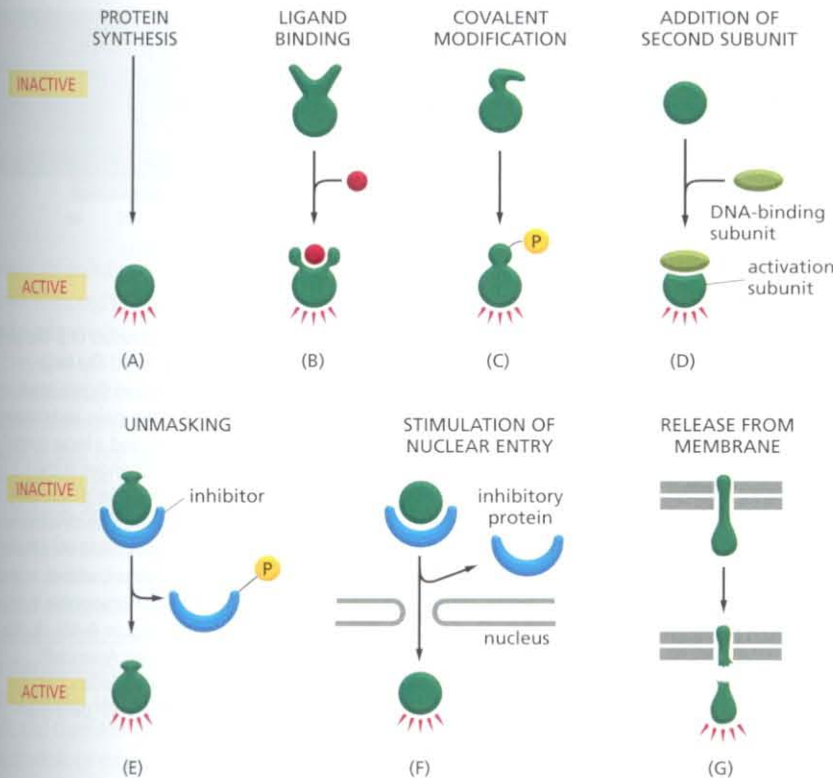


Figure 7–59 Some ways in which the activity of gene regulatory proteins is regulated in eucaryotic cells. (A) The protein is synthesized only when needed and is rapidly degraded by proteolysis so that it does not accumulate. (B) Activation by ligand binding. (C) Activation by covalent modification. Phosphorylation is shown here, but many other modifications are possible (see Table 3–3, p. 186). (D) Formation of a complex between a DNA-binding protein and a separate protein with a transcription-activating domain. (E) Unmasking of an activation domain by the phosphorylation of an inhibitor protein. (F) Stimulation of nuclear entry by removal of an inhibitory protein that otherwise keeps the regulatory protein from entering the nucleus. (G) Release of a gene regulatory protein from a membrane bilayer by regulated proteolysis.

that is, cells of the red blood cell lineage, but at different stages of mammalian development (see Figure 7–61B). The ε-globin gene is expressed in the early embryo, γ in the later embryo and fetus, and δ and β primarily in the adult. The gene products differ slightly in their oxygen-binding properties, suiting them for the different oxygenation conditions in the embryo, fetus, and adult. Each of the globin genes has its own set of regulatory proteins that are necessary to turn the gene on at the appropriate time.

The globin genes are unusual in that, at the appropriate time and place, they are transcribed at extremely high rates: indeed, red blood cells are little more

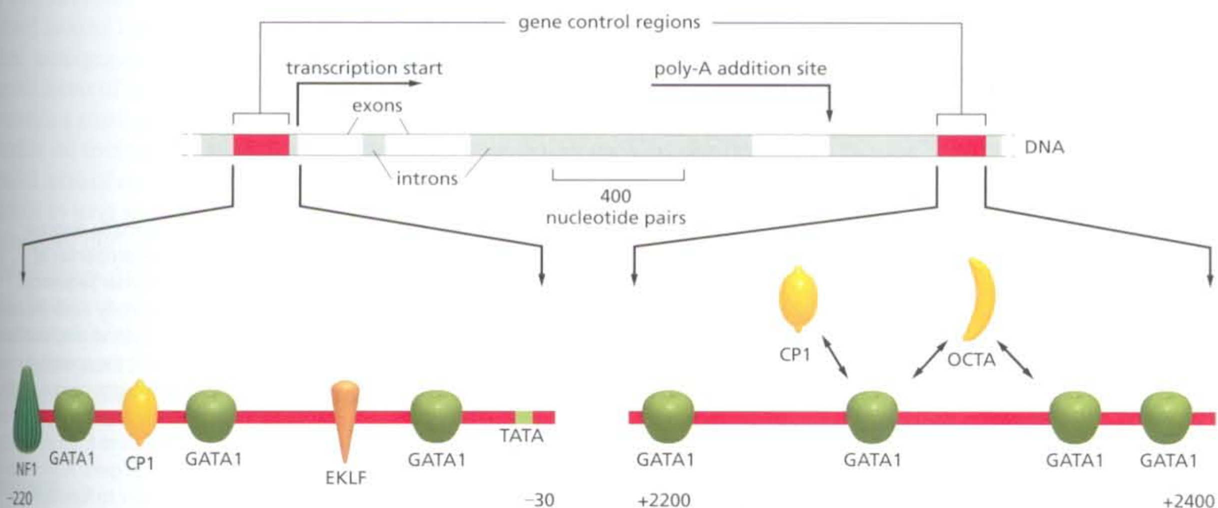
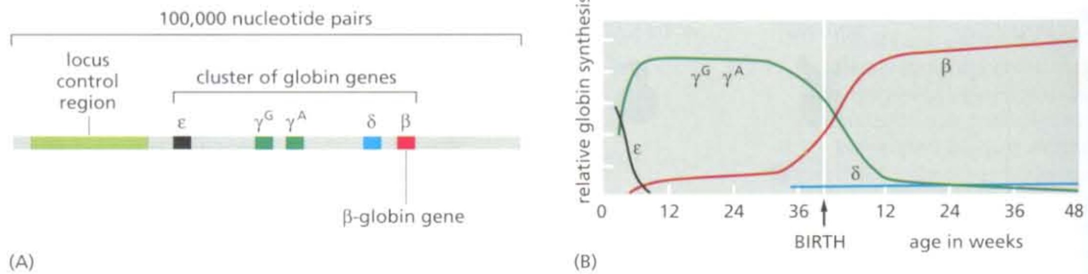


Figure 7–60 Model for the control of the human β-globin gene. The diagram shows some of the gene regulatory proteins that control expression of the gene during red blood cell development (see Figure 7–61). Some of the gene regulatory proteins shown, such as CP1, are found in many types of cells, while others, such as GATA1, are present in only a few types of cells—including red blood cells—and therefore are thought to contribute to the cell-type specificity of β-globin gene expression. As indicated by the *double-headed arrows*, several of the binding sites for GATA1 overlap those of other gene regulatory proteins; it is thought that by binding to these sites, GATA1 excludes binding of other proteins. Once bound to DNA, the gene regulatory proteins recruit chromatin remodeling complexes, histone modifying enzymes, the general transcription factors, Mediator, and RNA polymerase to the promoter. (Adapted from B. Emerson, in Gene Expression: General and Cell-Type Specific [M. Karin, ed.], pp. 116–161. Boston: Birkhauser, 1993.)



than bags of hemoglobin that was synthesized by precursor cells. To achieve this extraordinarily high level of transcription, the globin genes, in addition to their individual regulatory sequences, share a control region called the *locus control region* (LCR), which lies far upstream from the gene cluster and is needed for the proper expression of each gene in the cluster (see Figure 7–61A). The importance of the LCR can be seen in patients with a certain type of thalassemia, a severe inherited form of anemia. In these patients, the β-globin locus has suffered a deletion that removes all or part of the LCR. Although the β-globin and its nearby regulatory region are intact, the gene remains transcriptionally silent, even in erythroid cells.

The way in which the LCR functions is not understood in detail, but it is known that the gene regulatory proteins that bind the LCR interact, through DNA looping, with proteins bound to the control regions of the globin genes they regulate. In this way, the proteins bound at the LCR help attract chromatin remodeling complexes, histone-modifying enzymes, and components of the transcription machinery that act in conjunction with the specific regulatory regions of each individual globin gene. In addition, the LCR includes a *barrier sequence* (see Figure 4–47) that prevents the spread of neighboring heterochromatin into the β-globin locus, as discussed in Chapter 4. This dual feature distinguishes the globin LCR from many other types of regulatory sequences in the human genome; however, the globin genes are not alone in having an LCR, as LCRs are also present upstream of other highly transcribed, cell-type-specific genes. We should probably think of LCRs, not as unique DNA elements with specialized properties, but rather as especially powerful combinations of more fundamental types of regulatory sequences.

Figure 7–61 The cluster of β-like globin genes in humans. (A) The large chromosomal region shown spans 100,000 nucleotide pairs and contains the five globin genes and a locus control region (LCR). (B) Changes in the expression of the β-like globin genes at various stages of human development. Each of the globin chains encoded by these genes combines with an α-globin chain to form the hemoglobin in red blood cells (see Figure 4–86). (A, after F. Grosveld, G.B. van Assendelft, D.R. Greaves and G. Kollias, *Cell* 51:975–985, 1987. With permission from Elsevier.)

19th

Insulators Are DNA Sequences That Prevent Eucaryotic Gene Regulatory Proteins from Influencing Distant Genes

All genes have control regions, which dictate at which times, under what conditions, and in what tissues the gene will be expressed. We have also seen that eucaryotic gene regulatory proteins can act across very long stretches of DNA. How, then, are control regions of different genes kept from interfering with one another? In other words, what keeps a gene regulatory protein bound on the control region of one gene from inappropriately influencing the transcription of adjacent genes?

To avoid such cross-talk, several types of DNA elements function to compartmentalize the genome into discrete regulatory domains. In Chapter 4 we discussed barrier sequences that prevent the spread of heterochromatin into genes that need to be expressed. A second type of DNA element, called an insulator, prevents enhancers from running amok and activating inappropriate genes (**Figure 7–62**). An insulator can apparently block the communication

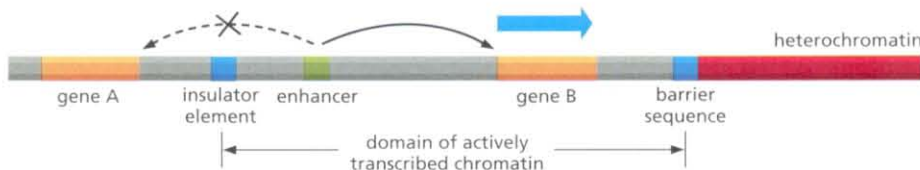


Figure 7–62 Schematic diagram summarizing the properties of insulators and barrier sequences. Insulators directionally block the action of enhancers *(left-hand side)*, and barrier sequences prevent the spread of heterochromatin *(right-hand side)*. Thus gene B is properly regulated and gene B's enhancer is prevented from influencing the transcription of gene A. How barrier sequences are likely to function is depicted in Figure 4–47. It is not yet understood how insulators exert their effects; one possibility is that they serve as "decoys," tying up the transcriptional machinery and preventing it from interacting with an authentic enhancer. Another is that they anchor DNA to the nuclear envelope, thereby interfering with DNA looping between an enhancer and an inappropriate promoter.

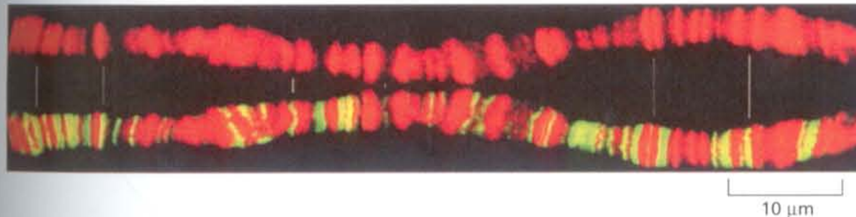


Figure 7–63 Localization of a *Drosophila* insulator-binding protein on polytene chromosomes. A polytene chromosome (discussed in Chapter 4) was stained with propidium iodide *(red)* to show its banding patterns—with bands appearing *bright red* and interbands as dark gaps in the pattern *(top)*. The positions on this polytene chromosome that are bound by a particular insulator protein (called BEAF) are stained *bright green* using antibodies directed against the protein *(bottom)*. BEAF is preferentially localized to interband regions, reflecting its role in organizing chromosomes into structural, as well as functional, domains. For convenience, these two micrographs of the same polytene chromosome are arranged as mirror images. (Courtesy of Uli Laemmli, from K. Zhao et al., *Cell* 81:879–889, 1995. With permission from Elsevier.)

between an enhancer and a promoter, but, to do so, it must be located between the two. Although proteins that bind to insulators have been identified, how they directionally neutralize enhancers is still a mystery.

Even though their mechanisms are not understood in detail, the distribution of insulators and barrier sequences in a genome is thought to divide it into independent domains of gene regulation and chromatin structure (**Figure 7–63**). Aspects of this organization can be visualized by staining whole chromosomes for the specialized proteins that bind these DNA elements.

Although chromosomes are organized into orderly domains that discourage enhancers from acting indiscriminately, there are special circumstances where an enhancer located one chromosome has been found to activate a gene located on a second chromosome. A remarkable example occurs in the regulation of the mammalian olfactory receptors. These are the proteins expressed by sensory neurons that allow mammals to discriminate accurately among many thousands of distinct smells (see p. 917). Humans, for example, have 350 olfactory receptor genes, and they are carefully regulated so that only one of these genes is expressed in each sensory neuron. The olfactory receptor genes are dispersed among many different chromosomes, but there is only a single enhancer for all of them. Once this enhancer activates a receptor gene by associating with its regulatory region, it remains stably associated thereby precluding activation of any of the other receptor genes. Although there is much we do not understand about this mechanism, it does indicate the extreme versatility of transcriptional regulation strategies.

Gene Switches Rapidly Evolve

We have seen that the control regions of eucaryotic genes are often spread out over long stretches of DNA, whereas those of procaryotic genes are typically clustered around the start point of transcription. It seems likely that the close-packed arrangement of bacterial genetic switches developed from more extended forms of switches in response to the evolutionary pressure on bacteria to maintain a small genome size. This compression comes at a price, however, as it restricts the complexity and adaptability of the control device. In contrast, the extended form of eucaryotic control regions—with discrete regulatory modules separated by long stretches of spacer DNA—facilitates the reshuffling of regulatory modules during evolution, both to create new regulatory circuits and to modify old ones. As we saw in Chapters 1 and 4, and we shall see again in Chapter 22, changes in gene regulation—rather than the acquisition of new genes—underlie much of the wide variety of life on Earth. Unraveling the history of how modern gene control regions have evolved presents a fascinating challenge to biologists, with many clues available in present-day genomes.

Summary

Gene regulatory proteins switch the transcription of individual genes on and off in cells. In procaryotes these proteins usually bind to specific DNA sequences close to the RNA polymerase start site and, depending on the nature of the regulatory protein and the precise location of its binding site relative to the start site, either activate or repress transcription of the gene. The flexibility of the DNA helix, however, also allows proteins bound at distant sites to affect the RNA polymerase at the promoter by the looping out of the intervening DNA. The regulation of higher eucaryotic genes is much more complex, commensurate with a larger genome size and the large variety of cell types that

are formed. A single eucaryotic gene is typically controlled by many gene regulatory proteins bound to sequences that can be thousands of nucleotide pairs from the promoter that directs transcription of the gene. Eucaryotic activators and repressors act by a wide variety of mechanisms—generally altering chromatin structure and controlling the assembly of the general transcription factors, Mediator, and RNA polymerase at the promoter. The time and place that each gene is transcribed, as well as its rates of transcription under different conditions, are determined by the spectrum of gene regulatory proteins that bind to the regulatory region of the gene.

THE MOLECULAR GENETIC MECHANISMS THAT CREATE SPECIALIZED CELL TYPES

Although all cells must be able to switch genes on and off in response to changes in their environments, the cells of multicellular organisms have evolved this capacity to an extreme degree and in highly specialized ways to form an organized array of differentiated cell types. In particular, once a cell in a multicellular organism becomes committed to differentiate into a specific cell type, the cell maintains this choice through many subsequent cell generations, which means that it remembers the changes in gene expression involved in the choice. This phenomenon of *cell memory* is a prerequisite for the creation of organized tissues and for the maintenance of stably differentiated cell types. In contrast, other changes in gene expression in eucaryotes, as well as most in bacteria, are only transient. The tryptophan repressor, for example, switches off the tryptophan genes in bacteria only in the presence of tryptophan; as soon as tryptophan is removed from the medium, the genes are switched back on, and the descendants of the cell will have no memory that their ancestors had been exposed to tryptophan. Even in bacteria, however, a few types of changes in gene expression can be inherited stably.

In this section we shall examine not only cell memory mechanisms, but also how gene regulatory devices can be combined to create "logic circuits" through which cells integrate signals, keep time, remember events in their past, and adjust the levels of gene expression over entire chromosomes. We begin by considering some of the best-understood genetic mechanisms of cell differentiation, which operate in bacterial and yeast cells.

DNA Rearrangements Mediate Phase Variation in Bacteria

We have seen that cell differentiation in higher eucaryotes usually occurs without detectable changes in DNA sequence. In some procaryotes, in contrast, a stably inherited pattern of gene regulation is achieved by DNA rearrangements that activate or inactivate specific genes. Since a change in DNA sequence will be copied faithfully during all subsequent DNA replication cycles, an altered state of gene activity will be inherited by all the progeny of the cell in which the rearrangement occurred. Some of these DNA rearrangements are, however, reversible so that occasional individuals can switch back to the original DNA configuration. The result is an alternating pattern of gene activity that can be detected by observations over long time periods and many generations.

A well-studied example of this differentiation mechanism occurs in *Salmonella* bacteria and is known as **phase variation**. Although this mode of differentiation has no known counterpart in higher eucaryotes, it can have considerable impact on animals because disease-causing bacteria use it to evade detection by the immune system. The switch in *Salmonella* gene expression is brought about by the occasional inversion of a specific 1000-nucleotide-pair piece of DNA. This change alters the expression of the cell-surface protein flagellin, for which the bacterium has two different genes (**Figure 7–64**). A site-specific recombination enzyme catalyzes the inversion and thereby changes the orientation of a promoter that is located within the inverted DNA segment. With the promoter in one orientation, the bacteria synthesize one type of flagellin;