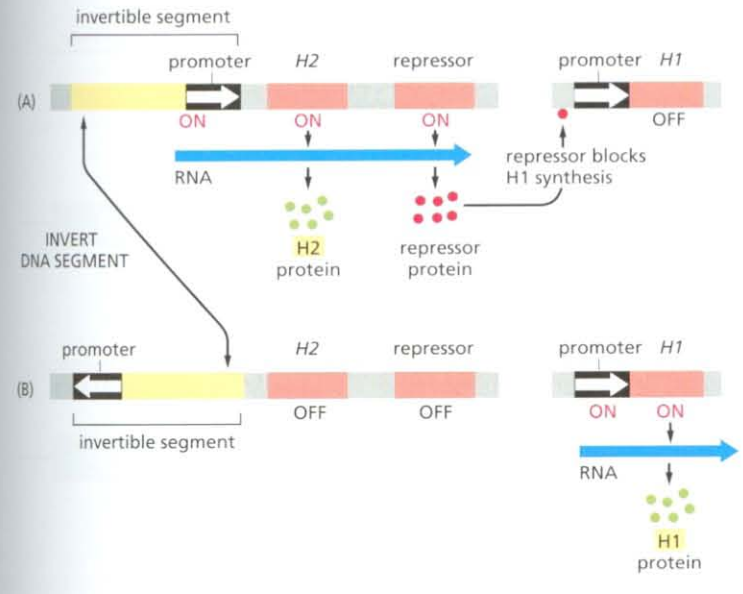


Figure 7–64 Switching gene expression by DNA inversion in bacteria. Alternating transcription of two flagellin genes in a *Salmonella* bacterium is caused by a simple site-specific recombination event that inverts a small DNA segment containing a promoter. (A) In one orientation, the promoter activates transcription of the *H2* flagellin gene as well as that of a repressor protein that blocks the expression of the *H1* flagellin gene. (B) When the promoter is inverted, it no longer turns on *H2* or the repressor, and the *H1* gene, which is thereby released from repression, is expressed instead. The recombination mechanism is activated only rarely (about once in every 10^5 cell divisions). Therefore, the production of one or other flagellin tends to be faithfully inherited in each clone of cells.

with the promoter in the other orientation, they synthesize the other type. Because inversions occur only rarely, entire clones of bacteria will have one type of flagellin or the other.

Phase variation almost certainly evolved because it protects the bacterial population against the immune response of its vertebrate host. If the host makes antibodies against one type of flagellin, a few bacteria whose flagellin has been altered by gene inversion will still be able to survive and multiply.

Bacteria isolated from the wild very often exhibit phase variation for one or more phenotypic traits. Standard laboratory strains of bacteria lose these "instabilities" over time, and underlying mechanisms have been studied in only a few cases. Not all involve DNA inversion. A bacterium that causes a common sexually transmitted human disease *(Neisseria gonorrhoeae)*, for example, avoids immune attack by means of a heritable change in its surface properties that arises from gene conversion (discussed in Chapter 5) rather than by inversion. This mechanism transfers DNA sequences from a library of silent "gene cassettes" to a site in the genome where the genes are expressed; it has the advantage of creating many variants of the major bacterial surface protein.

A Set of Gene Regulatory Proteins Determines Cell Type in a Budding Yeast

Because they are so easy to grow and to manipulate genetically, yeasts have served as model organisms for studying the mechanisms of gene control in eucaryotic cells. The common baker's yeast, *Saccharomyces cerevisiae*, has attracted special interest because of its ability to differentiate into three distinct cell types. *S. cerevisiae* is a single-celled eucaryote that exists in either a haploid or a diploid state. Diploid cells form by a process known as **mating**, in which two haploid cells fuse. In order for two haploid cells to mate, they must differ in *mating type* (sex). In *S. cerevisiae* there are two mating types, α and **a**, which are specialized for mating with each other. Each produces a specific diffusible signaling molecule (mating factor) and a specific cell-surface receptor protein. These jointly enable a cell to recognize and be recognized by its opposite cell type, with which it then fuses. The resulting diploid cells, called **a**/α, are distinct from either parent: they are unable to mate but can form spores (sporulate) when they run out of food, giving rise to haploid cells by the process of meiosis (discussed in Chapter 21).

The mechanisms by which these three cell types are established and maintained illustrate several of the strategies we have discussed for changing the pattern of gene expression. The mating type of the haploid cell is determined by a

single locus, the **mating-type (Mat) locus**, which in an **a**-type cell encodes a single gene regulatory protein, Mata1, and in an α cell encodes two gene regulatory proteins, Matα1 and Matα2. The Mata1 protein has no effect in the **a**-type haploid cell that produces it, but becomes important later in the diploid cell that results from mating. In contrast, the Matα2 protein acts in the α cell as a transcriptional repressor that turns off the **a**-specific genes, while the Matα1 protein acts as a transcriptional activator that turns on the α-specific genes. Once cells of the two mating types have fused, the combination of the Mata1 and Matα2 regulatory proteins generates a completely new pattern of gene expression, unlike that of either parent cell. **Figure 7–65** illustrates the mechanism by which the mating-type-specific genes are expressed in different patterns in the three cell types. This was among the first examples of combinatorial gene control to be identified, and it remains one of the best understood at the molecular level.

Although in most laboratory strains of *S. cerevisiae*, the **a** and α cell types are stably maintained through many cell divisions, some strains isolated from the wild can switch repeatedly between the **a** and α cell types by a mechanism of gene rearrangement whose effects are reminiscent of the DNA rearrangements in *N. gonorrhoeae*, although the exact mechanism seems to be peculiar to yeast. On either side of the *Mat* locus in the yeast chromosome, there is a silent locus encoding the mating-type gene regulatory proteins: the silent locus on one side encodes Matα1 and Matα2; the silent locus on the other side encodes Mata1. In approximately every other cell division, the active gene in the *Mat* locus is excised and replaced by a newly synthesized copy of the silent locus determining the opposite mating type. Because the change removes one gene from the active "slot" and replaces it by another, this mechanism is called the *cassette mechanism*. The change is reversible because, although the original gene at the *Mat* locus is discarded, a silent copy remains in the genome. New DNA copies made from the silent genes function as disposable cassettes that will be inserted in alternation into the *Mat* locus, which serves as the "playing head" (**Figure 7–66**).

The silent cassettes are packaged into a specialized form of chromatin and maintained in a transcriptionally inactive form. The study of these cassettes—

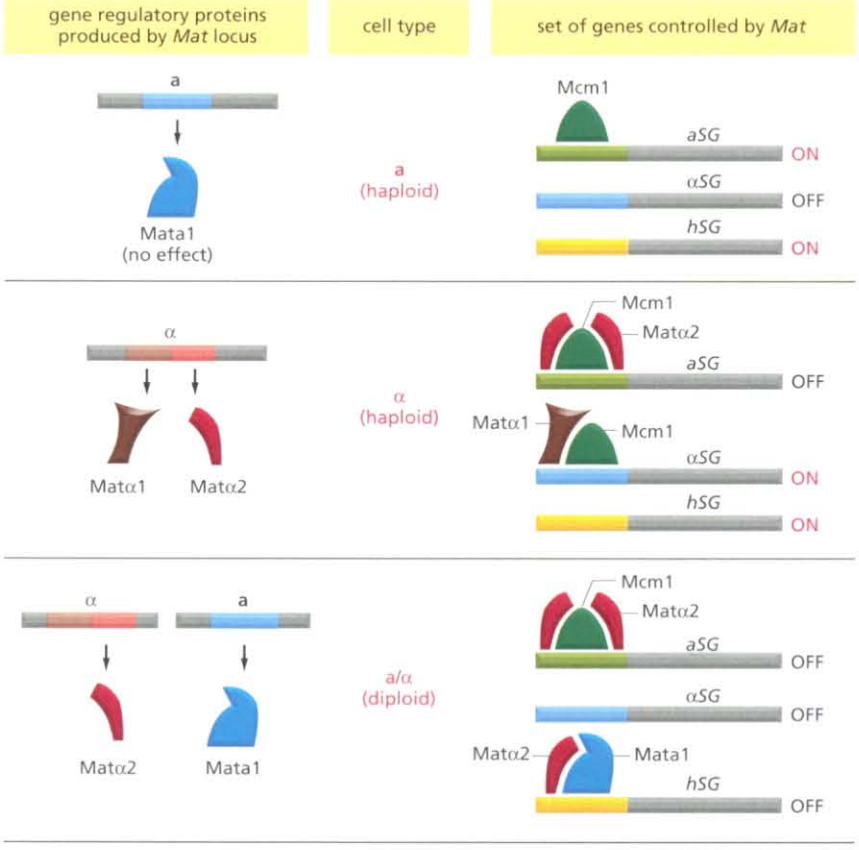


Figure 7–65 Control of cell type in yeasts. Three gene regulatory proteins (Matα1, Matα2, and Mata1) produced by the *Mat* locus determine yeast cell type. Different sets of genes are transcribed in haploid cells of type a, in haploid cells of type α, and in diploid cells (type a/α). The haploid cells express a set of haploid-specific genes *(hSG)* and either a set of α-specific genes *(αSG)* or a set of a-specific genes *(aSG)*. The diploid cells express none of these genes. The Mat regulatory proteins control many target genes in each type of cell by binding in various combinations, to specific regulatory sequences upstream of these genes. Note that the Matα1 protein is a gene activator protein, whereas the Matα2 protein is a gene repressor protein. Both work in combination with a gene regulatory protein called Mcm1 that is present in all three cell types. In the diploid cell type, Matα2 and Mata1 form a heterodimer (shown in detail in Figure 7–21) that turns off a set of genes (including the gene encoding the Matα1 activator protein) different from that turned off by the Matα2 and Mcm1 proteins. This relatively simple system of gene regulatory proteins is an example of combinatorial control of gene expression.

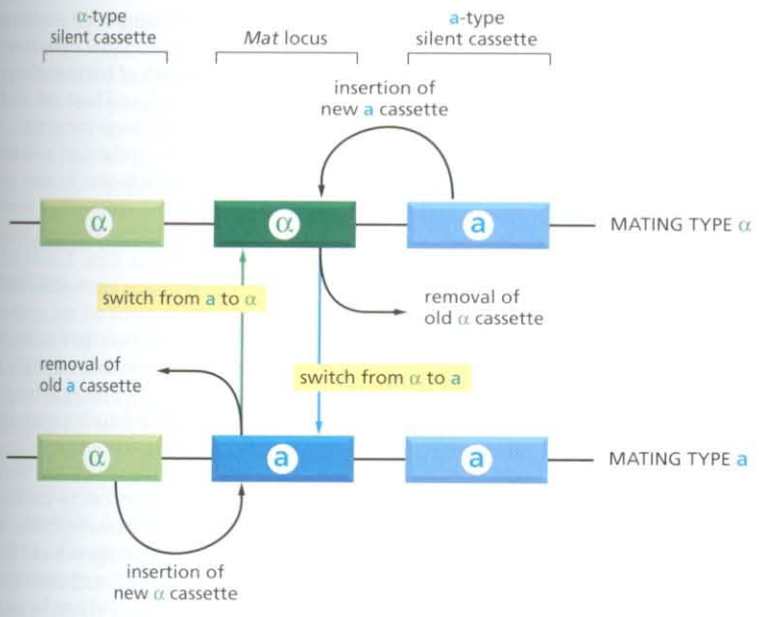


Figure 7–66 Cassette model of yeast mating-type switching. Cassette switching occurs by a gene-conversion process that involves a specialized enzyme (the HO endonuclease) that makes a double-stranded cut at a specific DNA sequence in the *Mat* locus. The DNA near the cut is then excised and replaced by a copy of the silent cassette of opposite mating type. The mechanism of this specialized form of gene conversion is similar to the repair of double-stranded breaks discussed in Chapter 5 (pp. 308–309).

which has been ongoing for nearly 40 years—has provided many of the key insights into the role of chromatin structure in gene regulation.

Two Proteins That Repress Each Other's Synthesis Determine the Heritable State of Bacteriophage Lambda

As we saw at the beginning of the present chapter, the nucleus of a single differentiated cell contains all the genetic information needed to construct a whole vertebrate or plant. This observation eliminates the possibility that an irreversible change in DNA sequence is a major mechanism of cell differentiation in these higher eucaryotes although such changes do occur in lymphocyte differentiation (discussed in Chapter 25). *Reversible* DNA sequence changes, resembling those just described for *Salmonella* and yeasts, in principle could still be responsible for some of the inherited changes in gene expression observed in higher organisms, but there is currently no evidence that such mechanisms are widely used.

Other mechanisms that we have already touched upon in this chapter, however, are also capable of producing patterns of gene regulation that can be inherited by subsequent cell generations. One of the simplest examples is found in the bacterial virus (bacteriophage) lambda where a switch causes the virus to flip-flop between two stable self-maintaining states. This type of switch can be viewed as a prototype for similar, but more complex, switches that operate in the development of higher eucaryotes.

We mentioned earlier that bacteriophage lambda can in favorable conditions become integrated into the *E. coli* cell DNA, to be replicated automatically each time the bacterium divides. Alternatively, the virus can multiply in the cytoplasm, killing its host (see Figure 5–78). Proteins encoded by the bacteriophage genome mediate the switch between these two states. The genome contains a total of about 50 genes, which are transcribed in very different patterns in the two states. A virus destined to integrate, for example, must produce the lambda *integrase* protein, which is needed to insert the lambda DNA into the bacterial chromosome, but must repress the production of the viral proteins responsible for virus multiplication. Once one transcriptional pattern or the other has been established, it is stably maintained.

At the heart of this complex gene regulatory switching mechanism are two gene regulatory proteins synthesized by the virus: the **lambda repressor protein** (cI protein), which we have already encountered, and the **Cro protein**. These proteins repress each other's synthesis, an arrangement that gives rise to just two stable states (**Figure 7–67**). In state 1 (the *prophage state)* the lambda

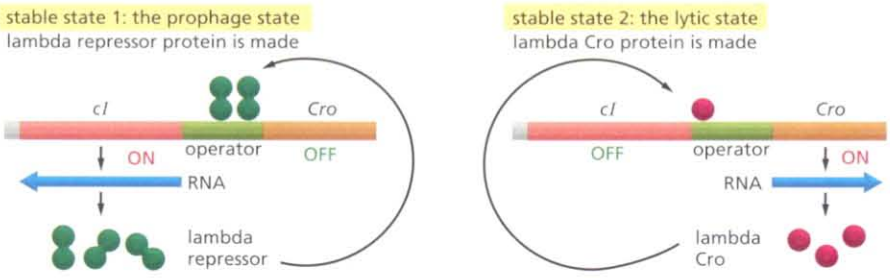


Figure 7–67 A simplified version of the regulatory system that determines the mode of growth of bacteriophage lambda in the *E. coli* host cell. In stable state 1 (the prophage state) the bacteriophage synthesizes a repressor protein, which activates its own synthesis and turns off the synthesis of several other bacteriophage proteins, including the Cro protein. In state 2 (the lytic state) the bacteriophage synthesizes the Cro protein, which turns off the synthesis of the repressor protein, so that many bacteriophage proteins are made and the viral DNA replicates freely in the *E. coli* cell, eventually producing many new bacteriophage particles and killing the cell. This example shows how two gene regulatory proteins can be combined in a circuit to produce two heritable states. Both the lambda repressor and the Cro protein recognize the operator through a helix–turn–helix motif (see Figure 7–11).

repressor occupies the operator, blocking the synthesis of Cro and also activating its own synthesis. In state 2 (the *lytic state)* the Cro protein occupies a different site in the operator, blocking the synthesis of repressor but allowing its own synthesis. In the prophage state most of the DNA of the stably integrated bacteriophage is not transcribed; in the lytic state, this DNA is extensively transcribed, replicated, packaged into new bacteriophage, and released by host cell lysis.

When the host bacteria are growing well, an infecting virus tends to adopt state 1, allowing the DNA of the virus to multiply along with the host chromosome. When the host cell is damaged, an integrated virus converts from state 1 to state 2 in order to multiply in the cell cytoplasm and make a quick exit. This conversion is triggered by the host response to DNA damage, which inactivates the repressor protein. In the absence of such interference, however, the lambda repressor both turns off production of the Cro protein and turns on its own synthesis, and this *positive feedback loop* helps to maintain the prophage state.

Simple Gene Regulatory Circuits Can Be Used to Make Memory Devices

Positive feedback loops provide a simple general strategy for cell memory—that is, for the establishment and maintenance of heritable patterns of gene transcription. **Figure 7–68** shows the basic principle, stripped to its barest essentials. Eucaryotic cells use many variations of this simple strategy. Several gene regulatory proteins that are involved in establishing the *Drosophila* body plan (discussed in Chapter 22), for example, stimulate their own transcription, thereby creating a positive feedback loop that promotes their continued synthesis; at the same time many of these proteins repress the transcription of genes encoding other important gene regulatory proteins. In this way, a few gene regulatory proteins that reciprocally affect one another's synthesis and activities can specify a sophisticated pattern of inherited behavior.

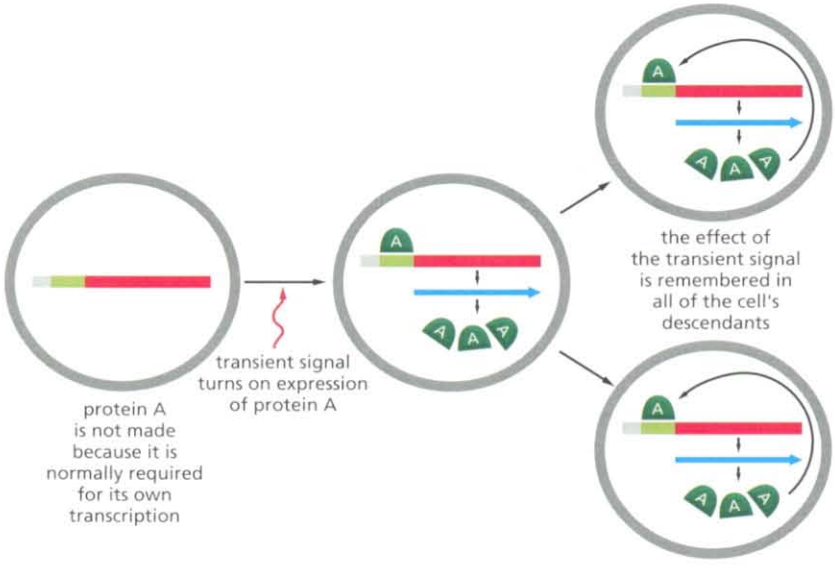


Figure 7–68 Schematic diagram showing how a positive feedback loop can create cell memory. Protein A is a gene regulatory protein that activates its own transcription. All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of the protein.

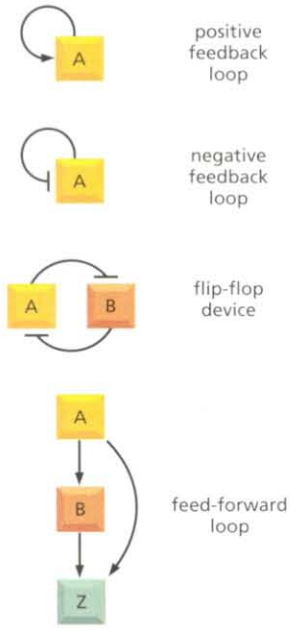


Figure 7–69 Common types of network motifs in transcriptional circuits. A and B represent gene regulatory proteins, arrows indicate positive transcriptional control, and lines with bars depict negative transcriptional control. More detailed descriptions of positive feedback loops and flip-flop devices are given in Figures 7–70 and 7–71, respectively. In feed-forward loops, A and B represent regulatory proteins that both activate the transcription of a target gene, Z.

Transcription Circuits Allow the Cell to Carry Out Logic Operations

Simple gene regulatory switches can be combined to create all sorts of control devices, just as simple electronic switching elements in a computer can be linked to perform many types of operations. The analysis of gene regulatory circuits has revealed that certain simple types of arrangements are found over and over again in cells from widely different species. For example, positive and *negative feedback loops* are especially common in all cells (**Figure 7–69**). As we have seen, the former provides a simple memory device; the latter is often used to keep the expression of a gene close to a standard level regardless of variations in biochemical conditions inside a cell. Suppose, for example, that a transcriptional repressor protein binds to the regulatory region of its own gene and exerts a strong negative feedback, such that transcription occurs at a very low rate when the concentration of repressor protein is above some critical value (determined by its affinity for its DNA binding site), and at a very high rate when it is below the critical value. The concentration of the protein will then be held close to the critical value, since any circumstance that causes a fall below that value will lead to a steep increase in synthesis, and any rise above that value will cause synthesis to switch off. Such adjustments will, however, take time, so that an abrupt change of conditions will cause a disturbance of gene expression that is strong but transient. As we discuss in Chapter 15, the negative feedback system can thus function as a detector of sudden change. Alternatively, if there is a delay in the feedback loop, the result may be spontaneous oscillations in the expression of the gene (see Figure 15–28). The quantitative details of the negative feedback loop determine which of these possible behaviors will occur.

With two or more genes, the possible range of control circuits and circuit behaviors rapidly becomes more complex. Bacteriophage lambda, as we have seen, exemplifies a common type of two-gene circuit that can flip-flop between expression of one gene and expression of the other. Another common circuit arrangement is called a *feed-forward* loop (see Figure 7–69); among other things, this can serve as a filter, responding to input signals that are prolonged but disregarding those that are brief (**Figure 7–70**). A cell can use these various network motifs as miniature logic devices to process information in surprisingly sophisticated ways.

The simple types of devices just illustrated are combined in a typical eucaryotic cell to create exceedingly complex circuits (**Figure 7–71**). Each cell in a developing multicellular organism is equipped with this control machinery, and must, in effect, use the intricate system of interlocking transcriptional switches to compute how it should behave at each time point in response to the many different past and present inputs received. We are only beginning to understand

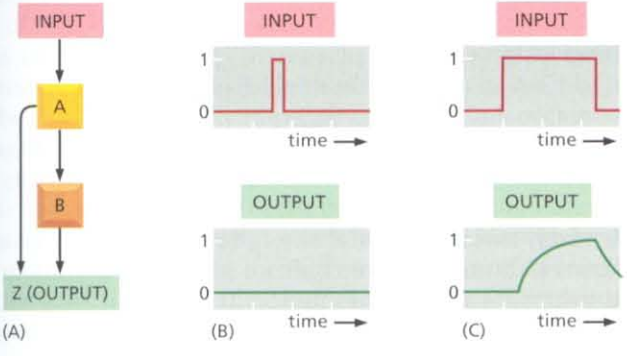


Figure 7–70 How a feed-forward loop can measure the duration of a signal. (A) In this theoretical example the gene activator proteins A and B are both required for transcription of Z, and A becomes active only when an input signal is present. (B) If the input signal to A is brief, A does not stay active long enough for B to accumulate, and the Z gene is not transcribed. (C) If the signal to A persists, B accumulates, A remains active, and Z is transcribed. This arrangement allows the cell to ignore rapid fluctuations of the input signal and respond only to persistent levels. This strategy could be used, for example, to distinguish between random noise and a true signal.

The behavior shown here was computed for one particular set of parameter values describing the quantitative properties of A, B, and Z and their syntheses. With different values of these parameters, feed-forward loops can in principle perform other types of "calculations." Many feed-forward loops have been discovered in cells, and theoretical analysis helps researchers to appreciate and subsequently test the different ways in which they may function. (Adapted from S.S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.* 31:64–68, 2002. With permission from Macmillan Publishers Ltd.)

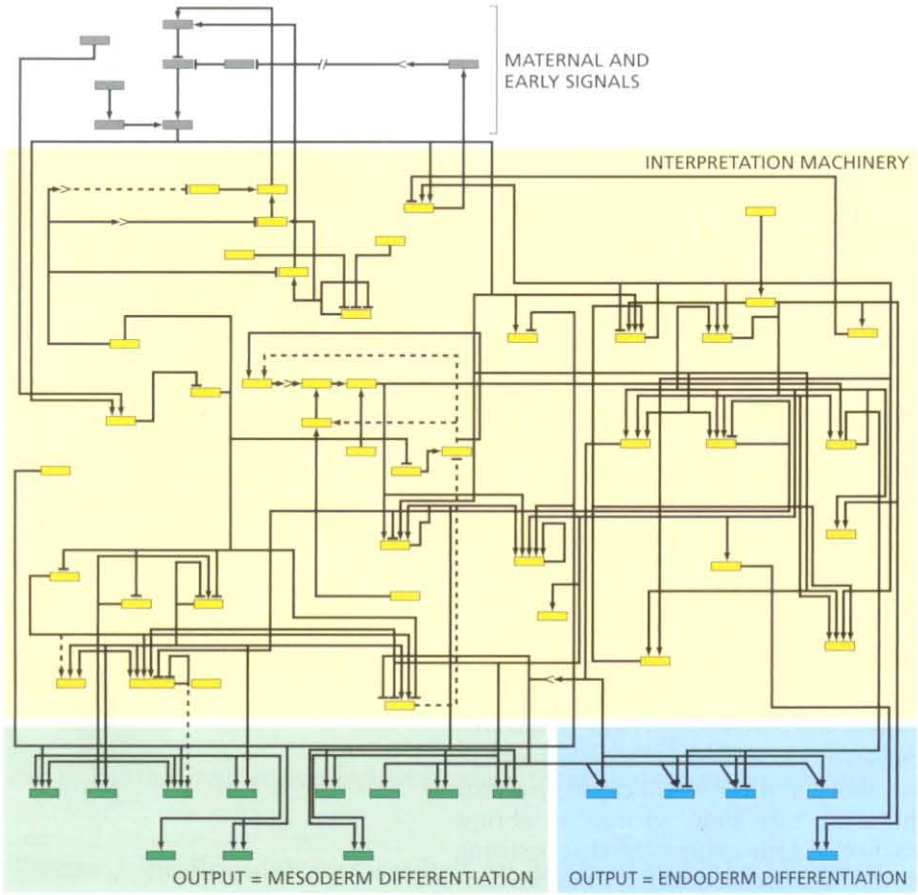


Figure 7–71 The exceedingly complex gene circuit that specifies a portion of the developing sea urchin embryo. Each *colored small box* represents a different gene. Those in *yellow* code for gene regulatory proteins and those in *green* and *blue* code for proteins that give cells of the mesoderm and endoderm, respectively, their specialized characteristics. Genes depicted in *gray* are largely active in the mother and provide the egg with cues needed for proper development. *Arrows* depict instances in which a gene regulatory protein activates the transcription of another gene. Lines ending in *bars* indicate examples of gene repression.

how to study such complex intracellular control networks. Indeed, without quantitative information far more precise and complete than we yet have, it is impossible to predict the behavior of a system such as that shown in Figure 7–71: the circuit diagram by itself is not enough.

Synthetic Biology Creates New Devices from Existing Biological Parts

Our discussion has focused on naturally occurring transcriptional circuits, but it is also instructive to design and construct artificial circuits in the laboratory and introduce them into cells to examine their behavior. **Figure 7–72** shows, for example, how an engineered bacterial cell can switch between three states in a prescribed order, thus functioning as an oscillator or simple clock. The construction of such new devices from existing parts is often termed *synthetic biology*. Scientists use this approach to test whether they truly understand the properties of the component parts; if so, they should be able to combine these parts in novel ways and accurately predict the characteristics of the new device. The fact that these predictions usually fail illustrates how far we are from truly understanding the detailed workings of the cell. There are many large gaps in our knowledge that will require skillful application of the quantitative approaches of the physical sciences to complex biological systems.

Circadian Clocks Are Based on Feedback Loops in Gene Regulation

Life on Earth evolved in the presence of a daily cycle of day and night, and many present-day organisms (ranging from archaea to plants to humans) have come to possess an internal rhythm that dictates different behaviors at different times of day. These behaviors range from the cyclical change in metabolic enzyme

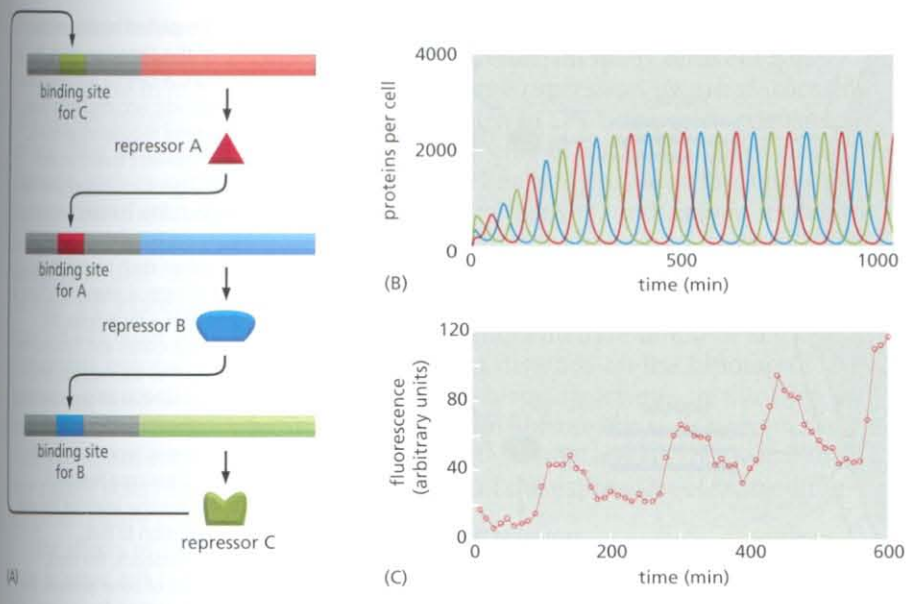


Figure 7–72 A simple gene oscillator or "clock" designed in the laboratory. (A) Recombinant DNA techniques were used to make three artificial genes, each coding for a different bacterial repressor protein, and each controlled by the product of another gene in the set, so as to create a regulatory circuit as shown. These repressors (denoted A, B, and C in the figure) are the Lac repressor (see Figure 7–39), the Tet repressor, which regulates genes in response to tetracycline, and the Lambda repressor (see Figure 7–67). When introduced into a bacterial cell, the three genes form a delayed negative feedback circuit: the product of gene A, for example, acts via genes B and C to indirectly inhibit its own expression. The delayed negative feedback gives rise to oscillations. (B) Computer prediction of the oscillatory behavior. The cell cycles repetitively through a series of states, expressing A, then B, then C, then A again, and so on, as each gene product in turn escapes from inhibition by the previous one and represses the next. (C) Actual oscillations observed in a cell containing the three artificial genes in (A), demonstrated with a fluorescent reporter of the expression of one of these genes. The increasing amplitude of the fluorescence signal reflects the growth of the bacterial cell. (Adapted from M.B. Elowitz and S. Leibler, *Nature* 403:335–338, 2000. With permission from Macmillan Publishers Ltd.)

activities of a fungus to the elaborate sleep–wake cycles of humans. The internal oscillators that control such diurnal rhythms are called circadian clocks.

By carrying its own circadian clock, an organism can anticipate the regular daily changes in its environment and take appropriate action in advance. Of course, the internal clock cannot be perfectly accurate, and so it must be capable of being reset by external cues such as the light of day. Thus, circadian clocks keep running even when the environmental cues (changes in light and dark) are removed, but the period of this free-running rhythm is generally a little less or a little more than 24 hours. External signals indicating the time of day cause small adjustments in the running of the clock, so as to keep the organism in synchrony with its environment. Following more drastic shifts, circadian cycles become gradually reset (entrained) by the new cycle of light and dark, as anyone who has experienced jet lag can attest.

We might expect that the circadian clock in a creature such as a human would itself be a complex multicellular device, with different groups of cells responsible for different parts of the oscillation mechanism. Remarkably, however, it turns out that in almost all organisms, including humans, the timekeepers are individual cells. Thus, a clock that operates in each member of a specialized group of brain cells (the SCN cells in the suprachiasmatic nucleus of the hypothalamus) controls our diurnal cycles of sleeping and waking, body temperature, and hormone release. Even if these cells are removed from the brain and dispersed in a culture dish, they will continue to oscillate individually, showing a cyclic pattern of gene expression with a period of approximately 24 hours. In the intact body, the SCN cells receive neural cues from the retina, entraining them to the daily cycle of light and dark, and they send information about the time of day to another brain area, the pineal gland, which relays the time signal to the rest of the body by releasing the hormone melatonin in time with the clock.

Although the SCN cells have a central role as timekeepers in mammals, they are not the only cells in the mammalian body that have an internal circadian rhythm or an ability to reset it in response to light. Similarly, in *Drosophila*, many different types of cells, including those of the thorax, abdomen, antenna, leg, wing, and testis all continue a circadian cycle when they have been dissected away from the rest of the fly. The clocks in these isolated tissues, like those in the SCN cells, can be reset by externally imposed light and dark cycles.

The working of circadian clocks, therefore, is a fundamental problem in cell biology. Although we do not yet understand all the details, studies in a wide variety of organisms have revealed many of the basic principles and molecular components. For animals, much of what we know has come from searches in *Drosophila* for mutations that make the fly's circadian clock run fast, or slow, or not at all; and this work has led to the discovery that many of the same components are involved in the circadian clock of mammals.

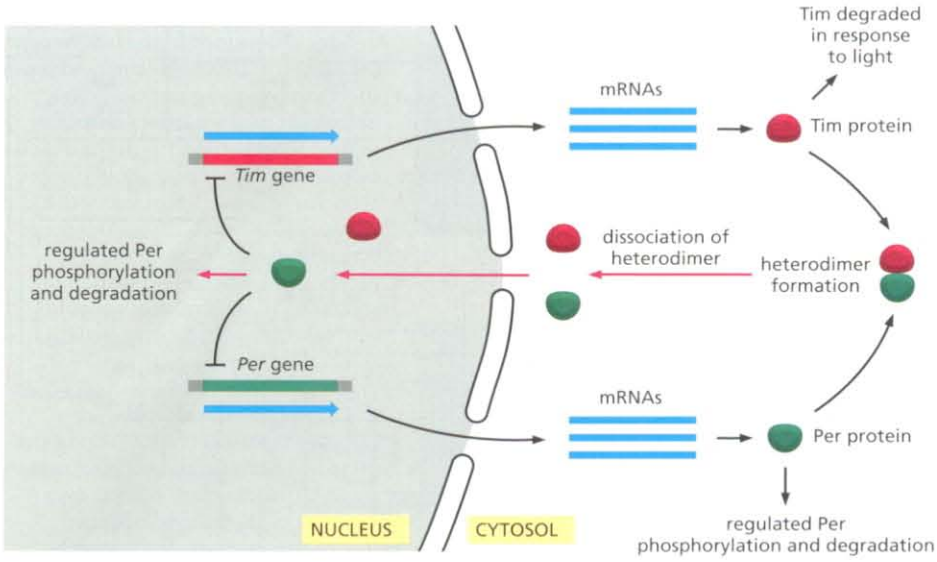


Figure 7–73 Simplified outline of the mechanism of the circadian clock in *Drosophila* cells. A central feature of the clock is the periodic accumulation and decay of two gene regulatory proteins, Tim (short for timeless, based on the phenotype of a gene mutation) and Per (short for period). The mRNAs encoding these proteins are translated in the cytosol, and, when each protein has accumulated to critical levels, they form a heterodimer. After a time delay, the heterodimer dissociates and Tim and Per are transported into the nucleus, where they regulate a number of gene products that mediate effects of the clock. Once in the nucleus, Per also represses the *Tim* and *Per* genes, creating a feedback system that causes the levels of Tim and Per to fall. In addition to this transcriptional feedback, the clock depends on a set of other proteins. For example, the controlled degradation of Per indicated in the diagram imposes delays in the periodic accumulation of Tim and Per, which are crucial to the functioning of the clock. Steps at which specific delays are imposed are shown in *red*.

Entrainment (or resetting) of the clock occurs in response to new light–dark cycles. Although most *Drosophila* cells do not have true photoreceptors, light is sensed by intracellular flavoproteins, also called cryptochromes. In the presence of light, these proteins associate with the Tim protein and cause its degradation, thereby resetting the clock. (Adapted from J.C. Dunlap, *Science* 311:184–185, 2006. With permission from AAAS.)

The mechanism of the clock in *Drosophila* is briefly outlined in **Figure 7–73**. At the heart of the oscillator is a transcriptional feedback loop that has a time delay built into it: accumulation of certain key gene products switches off the transcription of their genes, but with a delay, so that—crudely speaking—the cell oscillates between a state in which the products are present and transcription is switched off, and one in which the products are absent and transcription is switched on.

Despite the relative simplicity of the basic principle behind circadian clocks, the details are complex. One reason for this complexity is that clocks must be buffered against changes in temperature, which typically speed up or slow down macromolecular association. They must also run accurately but be capable of being reset. Although it is not yet understood how biological clocks run at a constant speed despite changes in temperature, the mechanism for resetting the *Drosophila* clock is the light-induced destruction of one of the key gene regulatory proteins (see Figure 7–73).

A Single Gene Regulatory Protein Can Coordinate the Expression of a Set of Genes

Cells need to be able to switch genes on and off individually but they also need to coordinate the expression of large groups of different genes. For example, when a quiescent eucaryotic cell receives a signal to divide, many hitherto unexpressed genes are turned on together to set in motion the events that lead eventually to cell division (discussed in Chapter 17). One way bacteria coordinate the expression of a set of genes is to cluster them together in an *operon* under control of a single promoter (see Figure 7–34). In eucaryotes, however, each gene is transcribed from a separate promoter.

How, then, do eucaryotes coordinate gene expression? This is an especially important question because, as we have seen, most eucaryotic gene regulatory proteins act as part of a regulatory protein committee, all of whose members are necessary to express the gene in the right cell, at the right time, in response to the proper signals, and to the proper level. How, then, can a eucaryotic cell 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 gene regulatory protein can still be decisive in switching any particular gene on or off, simply by completing the combination needed to maximally activate or repress that gene. This situation is analogous to dialing in the final number of a combination lock: the lock will spring open with only this simple addition if all of the other numbers have been previously entered. Moreover,

the same number can complete the combination for many different locks. Analogously, the addition of a particular protein can turn on many different genes.

An example in humans is the control of gene expression by the *glucocorticoid receptor protein*. To bind to regulatory sites in DNA, this gene regulatory protein must first form a complex with a molecule of a glucocorticoid steroid hormone, such as cortisol (see Figure 15–13). The body releases this hormone during times of starvation and intense physical activity, and among its other activities, it stimulates liver cells to increase the production of glucose from amino acids and other small molecules. To respond in this way, liver cells increase the expression of many different genes that code for metabolic enzymes and other products. Although these genes all have different and complex control regions, their maximal expression depends on the binding of the hormone–glucocorticoid receptor complex to a regulatory site in the DNA of each gene. When the body has recovered and the hormone is no longer present, the expression of each of these genes drops to its normal level in the liver. In this way a single gene regulatory protein can control the expression of many different genes (**Figure 7–74**).

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

Expression of a Critical Gene Regulatory Protein Can Trigger the Expression of a Whole Battery of Downstream Genes

The ability to switch many genes on or off coordinately is important not only in the day-to-day regulation of cell function. It is also the means by which eucaryotic cells differentiate into specialized cell types during embryonic development. The development of muscle cells provides a striking example.

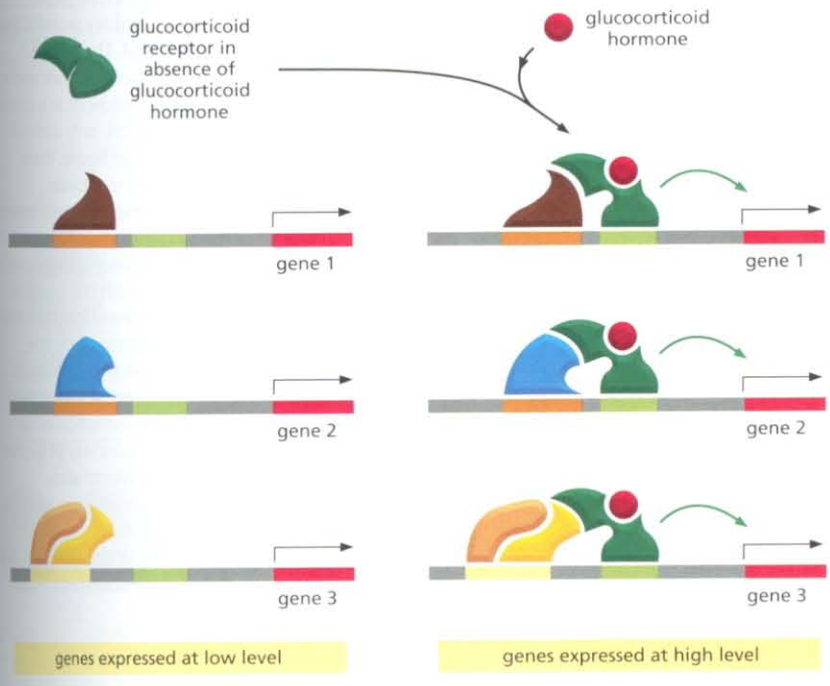


Figure 7–74 A single gene regulatory protein can coordinate the expression of several different genes. The action of the glucocorticoid receptor is illustrated schematically. On the *left* is a series of genes, each of which has various gene activator proteins bound to its regulatory region. However, these bound proteins are not sufficient on their own to fully activate transcription. On the *right* is shown the effect of adding an additional gene regulatory protein—the glucocorticoid receptor in a complex with glucocorticoid hormone—that can bind to the regulatory region of each gene. The glucocorticoid receptor completes the combination of gene regulatory proteins required for maximal initiation of transcription, and the genes are now switched on as a set. In the absence of the hormone, the glucocorticoid receptor is unavailable to bind to DNA.

In addition to activating gene expression, the hormone-bound form of the glucocorticoid receptor represses transcription of certain genes, depending on the gene regulatory proteins already present on their control regions. The effect of the glucocorticoid receptor on any given gene therefore depends upon the type of cell, the gene regulatory proteins contained within it, and the regulatory region of the gene. The structure of the DNA-binding portion of the glucocorticoid receptor is shown in Figure 7–16.

As described in Chapter 16, a mammalian skeletal muscle cell is a highly distinctive giant cell, formed by the fusion of many muscle precursor cells called *myoblasts*, and therefore containing many nuclei. The mature muscle cell synthesizes a large number of characteristic proteins, including specific types of actin, myosin, tropomyosin, and troponin (all part of the contractile apparatus), creatine phosphokinase (for the specialized metabolism of muscle cells), and acetylcholine receptors (to make the membrane sensitive to nerve stimulation). In proliferating myoblasts, these muscle-specific proteins and their mRNAs are absent or are present in very low concentrations. As myoblasts begin to fuse with one another, the corresponding genes are all switched on coordinately as part of a general transformation of the pattern of gene expression.

This entire program of muscle differentiation can be triggered in cultured skin fibroblasts and certain other cell types by introducing any one of a family of helix–loop–helix proteins—the so-called myogenic proteins (MyoD, Myf5, MyoG, and Mrf4)—that are normally expressed only in muscle cells (**Figure 7–75**A). Binding sites for these regulatory proteins are present in the regulatory DNA sequences adjacent to many muscle-specific genes, and the myogenic proteins thereby directly activate the transcription of these genes. In addition, the myogenic proteins stimulate their own transcription as well as that of various other gene regulatory proteins involved in muscle development, creating an elaborate series of positive feedback and feed-forward loops that amplify and maintain the muscle developmental program, even after the initiating signal has disappeared (Figure 7–75B; see also Chapter 22).

It is probable that those cell types that are converted to muscle cells by the addition of myogenic proteins have already accumulated a number of gene regulatory proteins that can cooperate with the myogenic proteins to switch on muscle-specific genes. Other cell types fail to be converted to muscle by myogenin or its relatives; these cells presumably have not accumulated the other gene regulatory proteins required.

The conversion of one cell type (fibroblast) to another (skeletal muscle) by a single gene regulatory protein reemphasizes one of the most important principles discussed in this chapter: differences in gene expression can produce dramatic differences between cell types—in size, shape, chemistry, and function.

20th

Combinatorial Gene Control Creates Many Different Cell Types in Eucaryotes

We have already discussed how multiple gene regulatory proteins can act in combination to regulate the expression of an individual gene. But, as the example of the myogenic proteins shows, combinatorial gene control means more than this: not only does each gene respond to many gene regulatory proteins that control it, but each regulatory protein contributes to the control of many genes. Moreover, although some gene regulatory proteins are specific to a single cell type, most are switched on in a variety of cell types, at several sites in the body, and at several times in development. This point is illustrated schematically in **Figure 7–76**, which shows how combinatorial gene control makes it possible to generate a great deal of biological complexity even with relatively few gene regulatory proteins.

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

One requirement of combinatorial control is that many gene regulatory proteins must be able to work together to influence the final rate of transcription. Experiments demonstrate that even unrelated gene regulatory proteins from widely different eucaryotic species can cooperate when introduced into the same cell. This situation reflects the high degree of conservation of the transcription

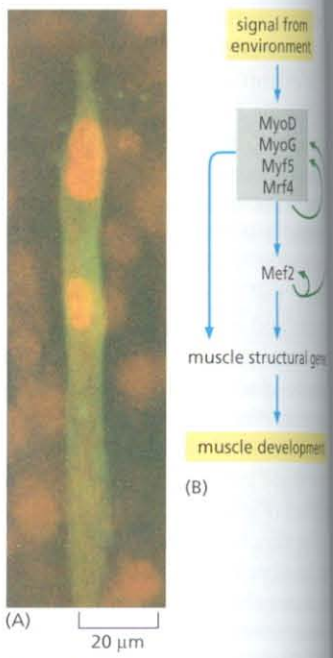


Figure 7–75 Role of the myogenic regulatory proteins in muscle development. (A) The effect of expressing the MyoD protein in fibroblasts. As shown in this immunofluorescence micrograph, fibroblasts from the skin of a chick embryo have been converted to muscle cells by the experimentally induced expression of the *MyoD* gene. The fibroblasts that have been induced to express the *MyoD* gene have fused to form elongated multinucleate muscle-like cells, which are stained green with an antibody that detects a muscle-specific protein. Fibroblasts that do not express the *MyoD* gene are barely visible above the background. (B) Simplified scheme showing some of the gene regulatory proteins involved in skeletal muscle development. External signals result in the synthesis of the four closely related myogenic gene regulatory proteins, MyoD, Myf5, MyoG, and Mrf4. These gene regulatory proteins activate their own as well as each other's synthesis in a complex series of feedback loops, only some of which are shown in the figure. These proteins in turn directly activate transcription of muscle structural genes as well as the *Mef2* gene, which encodes an additional gene regulatory protein. Mef2 acts in combination with the myogenic proteins in a feed-forward loop to further activate the transcription of muscle structural genes, as well as forming an additional positive feedback loop that helps to maintain transcription of the myogenic genes. (A, courtesy of Stephen Tapscott and Harold Weintraub; B, adapted from J.D. Molkentin and E.N. Olson, *Proc. Natl Acad. Sci. U.S.A.* 93:9366–9373, 1996. With permission from National Academy of Sciences.)

machinery. It seems that the multifunctional, combinatorial mode of action of gene regulatory proteins has put a tight constraint on their evolution: they must interlock with other gene regulatory proteins, the general transcription factors, Mediator, RNA polymerase, and the chromatin-modifying enzymes.

An important consequence of combinatorial gene control is that the effect of adding a new gene regulatory protein to a cell will depend on the cell's past history, since this history will determine which gene regulatory proteins are already present. Thus, during development a cell can accumulate a series of gene regulatory proteins that need not initially alter gene expression. The addition of the final members of the requisite combination of gene regulatory proteins completes the regulatory message, and can lead to large changes in gene expression. Such a scheme, as we have seen, helps to explain how the addition of a single regulatory protein to a fibroblast can produce the dramatic transformation of the fibroblast into a muscle cell. It also can account for the important difference, discussed in Chapter 22, between the process of *cell determination*—where a cell becomes committed to a particular developmental fate—and the process of *cell differentiation*, in which a committed cell expresses its specialized character.

A Single Gene Regulatory Protein Can Trigger the Formation of an Entire Organ

We have seen that even though combinatorial control is the norm for eucaryotic genes, a single gene regulatory protein, if it completes the appropriate combination, can be decisive in switching a whole set of genes on or off, and we have

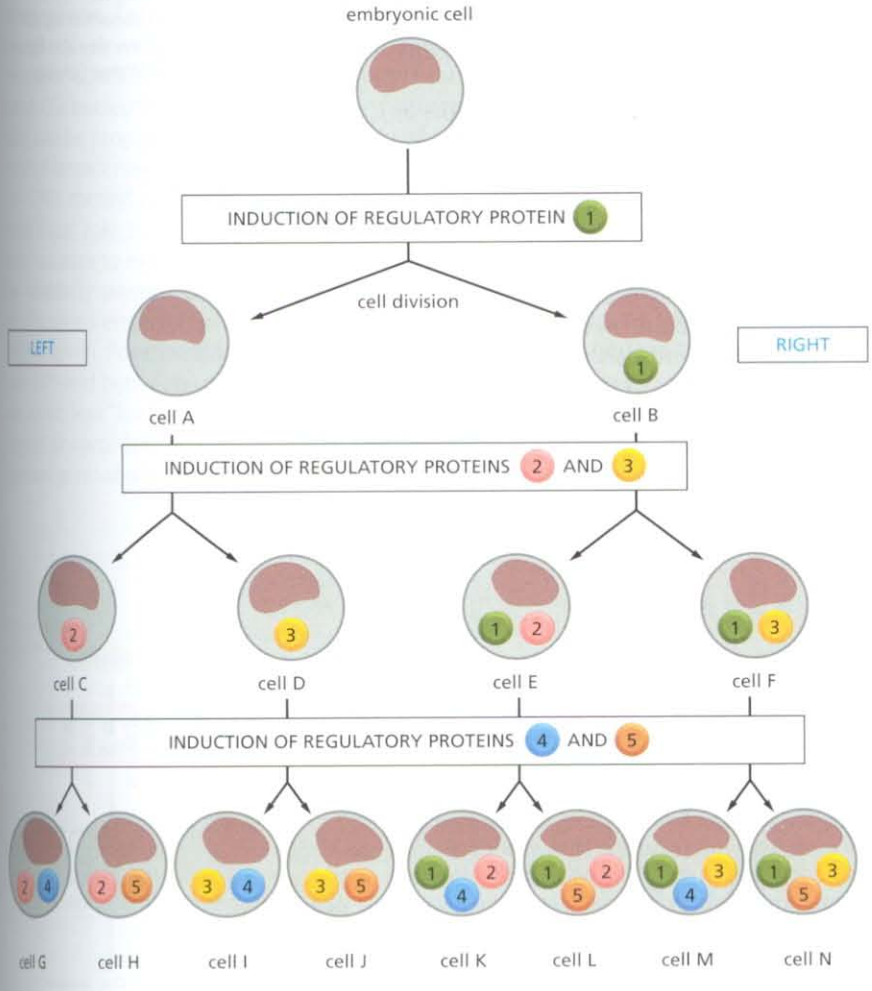
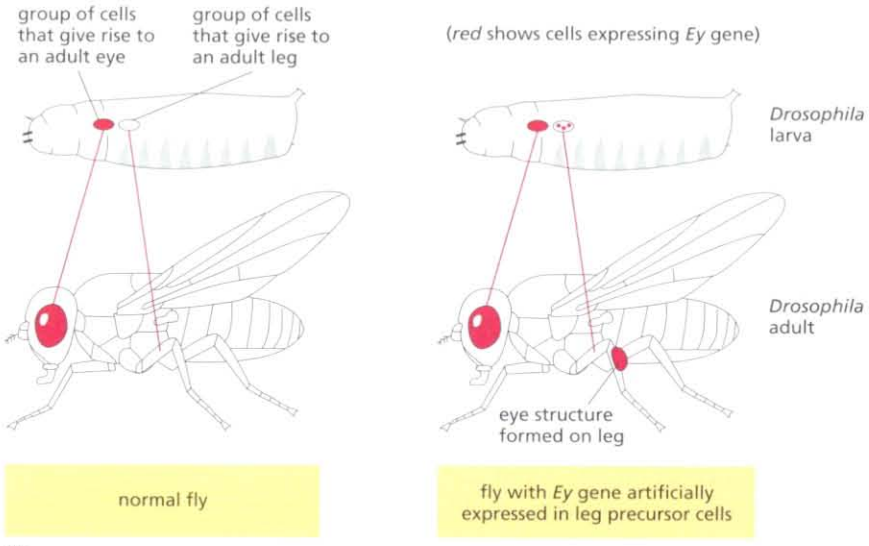


Figure 7–76 The importance of combinatorial gene control for development. Combinations of a few gene regulatory proteins can generate many cell types during development. In this simple, idealized scheme a "decision" to make one of a pair of different gene regulatory proteins (shown as numbered *circles*) is made after each cell division. Sensing its relative position in the embryo, the daughter cell toward the *left side* of the embryo is always induced to synthesize the even-numbered protein of each pair, while the daughter cell toward the *right side* of the embryo is induced to synthesize the odd-numbered protein. The production of each gene regulatory protein is assumed to be self-perpetuating once it has become initiated (see Figure 7–68). In this way, through cell memory, the final combinatorial specification is built up step by step. In this purely hypothetical example, five different gene regulatory proteins have created eight final cell types (G–N).



seen how this can convert one cell type into another. A dramatic extension of the principle comes from studies of eye development in *Drosophila*, mice, and humans. Here, a gene regulatory protein, called Ey (short for Eyeless) in flies and Pax6 in vertebrates, is crucial. When expressed in the proper context, Ey can trigger the formation of not just a single cell type but a whole organ (an eye), composed of different types of cells, all properly organized in three-dimensional space.

The most striking evidence for the role of Ey comes from experiments in fruit flies in which the *Ey* gene is artificially expressed early in development in groups of cells that normally will go on to form leg parts. This abnormal gene expression causes eyes to develop in the legs (**Figure 7–77**).

The *Drosophila* eye is composed of thousands of cells, and the question of how a regulatory protein coordinates the construction of a whole organ is a central topic in *developmental biology*. As discussed in Chapter 22, it involves cell–cell interactions as well as intracellular gene regulatory proteins. Here, we note that Ey directly controls the expression of many other genes by binding to their regulatory regions. Some of the genes controlled by Ey themselves code for gene regulatory proteins that, in turn, control the expression of other genes. Moreover, some of these regulatory gene products act back on *Ey* itself to create a positive feedback loop that ensures the continued synthesis of the Ey protein as the cells divide and further differentiate (**Figure 7–78**). In this way, the action of just one regulatory protein can permanently turn on a cascade of gene regulatory proteins and cell–cell interaction mechanisms, whose actions result in an organized group of many different types of cells. One can begin to imagine how, by repeated applications of this principle, a complex organism is assembled piece by piece.

Figure 7–77 Expression of the *Drosophila* *Ey* gene in precursor cells of the leg triggers the development of an eye on the leg. (A) Simplified diagrams showing the result when a fruit fly larva contains either the normally expressed *Ey* gene *(left)* or an *Ey* gene that is additionally expressed artificially in cells that normally give rise to leg tissue *(right)*. (B) Photograph of an abnormal leg that contains a misplaced eye (see also Figure 22–2). (B, courtesy of Walter Gehring.)

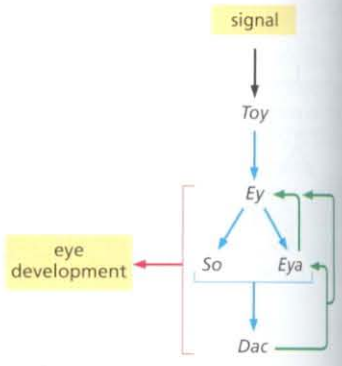


Figure 7–78 Gene regulatory proteins that specify eye development in *Drosophila*. *Toy (Twin of eyeless)* and *Ey (Eyeless)* encode similar gene regulatory proteins, Toy and Ey, either of which, when ectopically expressed, can trigger eye development. In normal eye development, expression of *Ey* requires the *Toy* gene. Once its transcription is activated by Toy, Ey activates the transcription of *So (Sine oculis)* and *Eya (Eyes absent)*, which act together to switch on the *Dac (Dachshund)* gene. As indicated by the *green arrows*, some of the gene regulatory proteins form a series of interlocking positive feedback loops that reinforce the initial commitment to eye development. The Ey protein is known to bind directly to numerous target genes for eye development, including those encoding lens crystallins, rhodopsins, and other photoreceptor proteins. (Adapted from T. Czerny et al., *Mol. Cell* 3:297–307, 1999. With permission from Elsevier.)

The Pattern of DNA Methylation Can Be Inherited When Vertebrate Cells Divide

Thus far, we have emphasized the regulation of gene transcription by proteins that associate with DNA. However, DNA itself can be covalently modified, and in the following sections we shall see that this, too, provides opportunities for the regulation of gene expression. In vertebrate cells, the methylation of cytosine provides a powerful mechanism through which gene expression patterns are passed on to progeny cells. The methylated form of cytosine, 5-methylcytosine (5-methyl C), has the same relation to cytosine that thymine has to uracil, and the modification likewise has no effect on base-pairing (**Figure 7–79**). **DNA methylation** in vertebrate DNA is restricted to cytosine (C) nucleotides in the sequence CG, which is base-paired to exactly the same sequence (in opposite orientation) on the other strand of the DNA helix. Consequently, a simple mechanism permits the existing pattern of DNA methylation to be inherited directly by the daughter DNA strands. An enzyme called *maintenance methyltransferase* acts preferentially on those CG sequences that are base-paired with a CG sequence that is already methylated. As a result, the pattern of DNA methylation on the parental DNA strand serves as a template for the methylation of the daughter DNA strand, causing this pattern to be inherited directly following DNA replication (**Figure 7–80**).

The stable inheritance of DNA methylation patterns can be explained by maintenance DNA methyltransferases. DNA methylation patterns, however, are dynamic during vertebrate development. Shortly after fertilization there is a genome-wide wave of demethylation, when the vast majority of methyl groups are lost from the DNA. This demethylation may occur either by suppression of maintenance DNA methyltransferase activity, resulting in the passive loss of methyl groups during each round of DNA replication, or by a specific demethylating enzyme. Later in development, new methylation patterns are established by several *de novo DNA methyltransferases* that are directed to DNA by sequence-specific DNA-binding proteins where they modify adjacent unmethylated CG nucleotides. Once the new patterns of methylation are established, they can be propagated through rounds of DNA replication by the maintenance methyl transferases.

DNA methylation has several uses in the vertebrate cell. Perhaps its most important role is to work in conjunction with other gene expression control mechanisms to establish a particularly efficient form of gene repression that can be faithfully passed on to progeny cells (**Figure 7–81**). This combination of mechanisms ensures that unneeded eucaryotic genes can be repressed to very high degrees. For example, the rate at which a vertebrate gene is transcribed can vary 10^6-fold between one tissue and another. The unexpressed vertebrate genes are much less "leaky" in terms of transcription than bacterial genes, in which the largest known differences in transcription rates between expressed and unexpressed gene states are about 1000-fold.

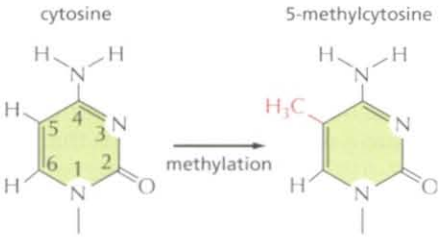
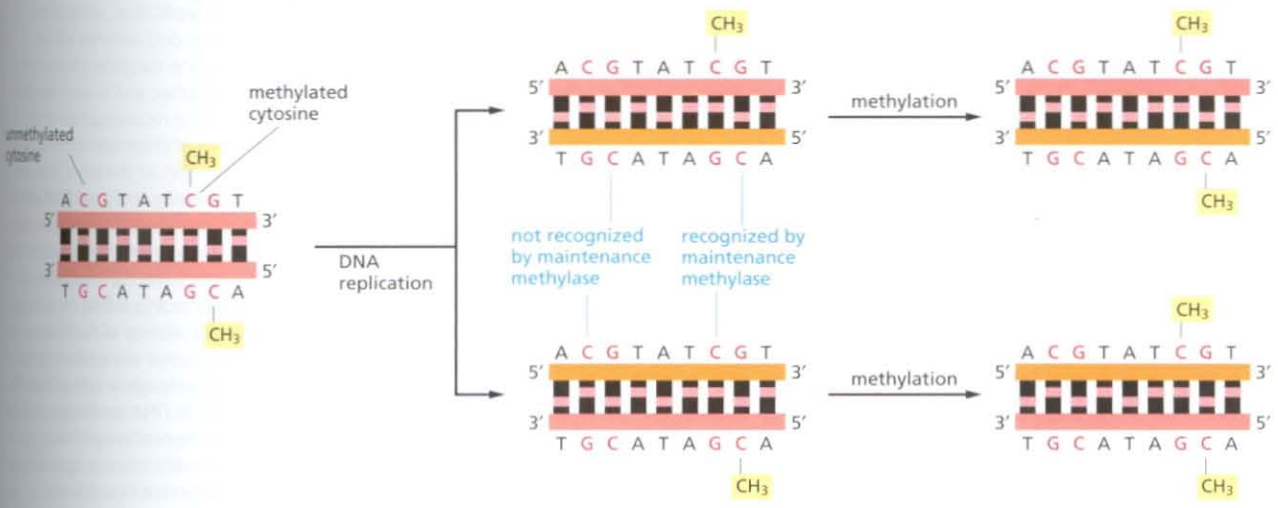


Figure 7–79 Formation of 5-methylcytosine occurs by methylation of a cytosine base in the DNA double helix. In vertebrates this event is confined to selected cytosine (C) nucleotides located in the sequence CG.

Figure 7–80 How DNA methylation patterns are faithfully inherited. In vertebrate DNAs a large fraction of the cytosine nucleotides in the sequence CG are methylated (see Figure 7–79). Because of the existence of a methyl-directed methylating enzyme (the maintenance methyltransferase), once a pattern of DNA methylation is established, that pattern of methylation is inherited in the progeny DNA, as shown.



How DNA methylation helps to repress gene expression is not understood in detail, but two general mechanisms have emerged. DNA methylation of the promoter region of a gene or of its regulatory sequences can interfere directly with the binding of proteins required for transcription initiation. In addition, the cell has a repertoire of proteins that specifically bind to methylated DNA (see Figure 7–81), thereby blocking access of other proteins. One reflection of the importance of DNA methylation to humans is the widespread involvement of errors in this mechanism in cancer progression (see Chapter 20).

We shall return to the topic of gene silencing by DNA methylation later in this chapter, when we discuss X-chromosome inactivation and other examples of large-scale gene silencing. First, however, we describe some of the other ways in which DNA methylation affects our genomes.

Genomic Imprinting Is Based on DNA Methylation

Mammalian cells are diploid, containing one set of genes inherited from the father and one set from the mother. The expression of a small minority of genes depends on whether they have been inherited from the mother or the father: while the paternally inherited gene copy is active, the maternally inherited gene copy is silent, or vice-versa. This phenomenon is called **genomic imprinting**. The gene for *insulin-like growth factor-2* (*Igf2*) is a well-studied example of an

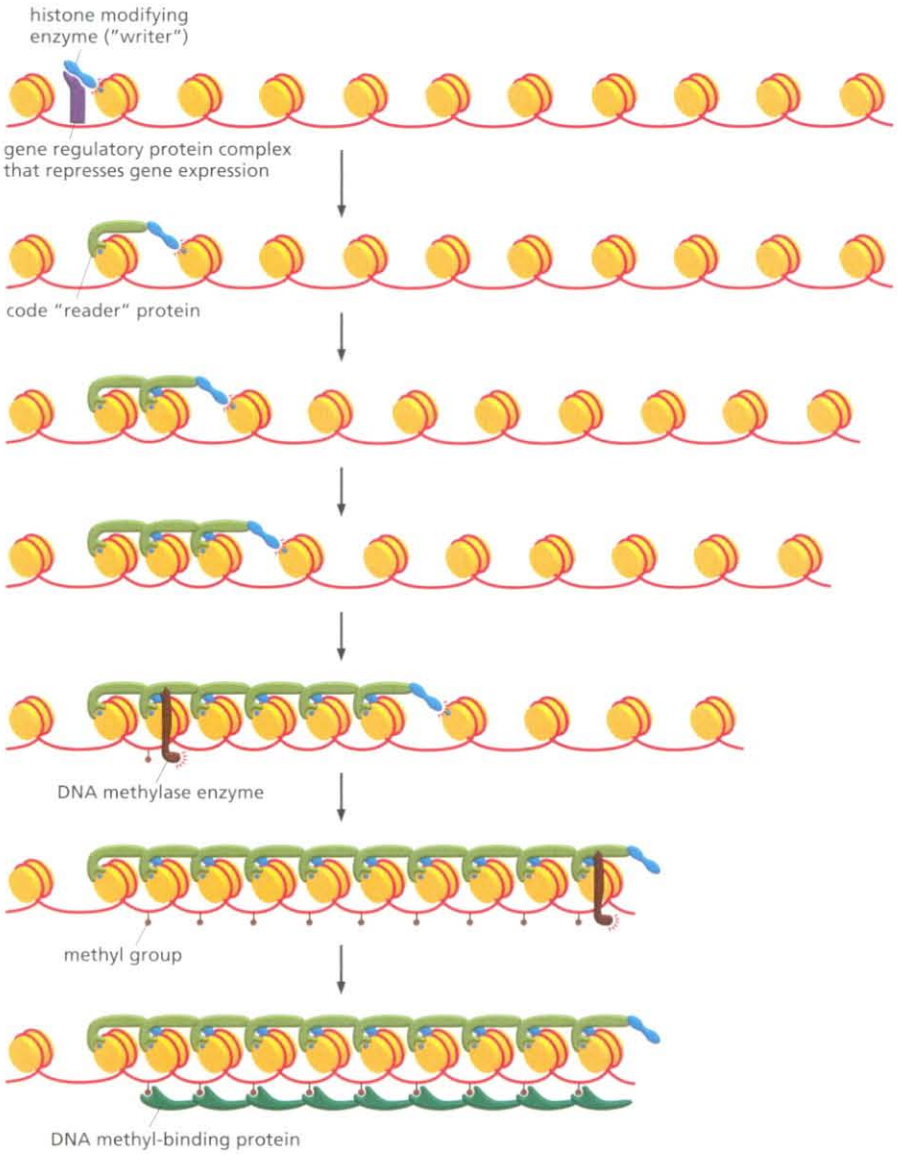


Figure 7–81 Multiple mechanisms contribute to stable gene repression. In this schematic example, histone reader and writer proteins, under the direction of gene regulatory proteins, establish a repressive form of chromatin. A *de novo* DNA methylase is attracted by the histone reader and methylates nearby cytosines in DNA, which are, in turn, bound by DNA methyl-binding proteins. During DNA replication, some of the modified (*blue* dot) histones will be inherited by one daughter chromosome, some by the other, and in each daughter they can induce reconstruction of the same pattern of chromatin modification (see Figure 5–39). At the same time, the mechanism shown in Figure 7–80 will cause both daughter chromosomes to inherit the same methylation pattern. The two inheritance mechanisms will be mutually reinforcing, if DNA methylation stimulates the activity of the histone writer. This scheme can account for the inheritance by daughter cells of both the histone and the DNA modifications. It can also explain the tendency of some chromatin modifications to spread along a chromosome (see Figure 4–45).

imprinted gene. *Igf2* is required for prenatal growth, and mice that do not express *Igf2* at all are born half the size of normal mice. However, only the paternal copy of *Igf2* is transcribed, and only this gene copy matters for the phenotype. As a result, mice with a mutated paternally derived *Igf2* gene are stunted, while mice with a mutated maternally derived *Igf2* gene are normal.

In the early embryo, genes subject to imprinting are marked by methylation according to whether they were derived from a sperm or an egg chromosome. In this way, DNA methylation is used as a mark to distinguish two copies of a gene that may be otherwise identical (**Figure 7–82**). Because imprinted genes are somehow protected from the wave of demethylation that takes place shortly after fertilization (see p. 467), this mark enables somatic cells to "remember" the

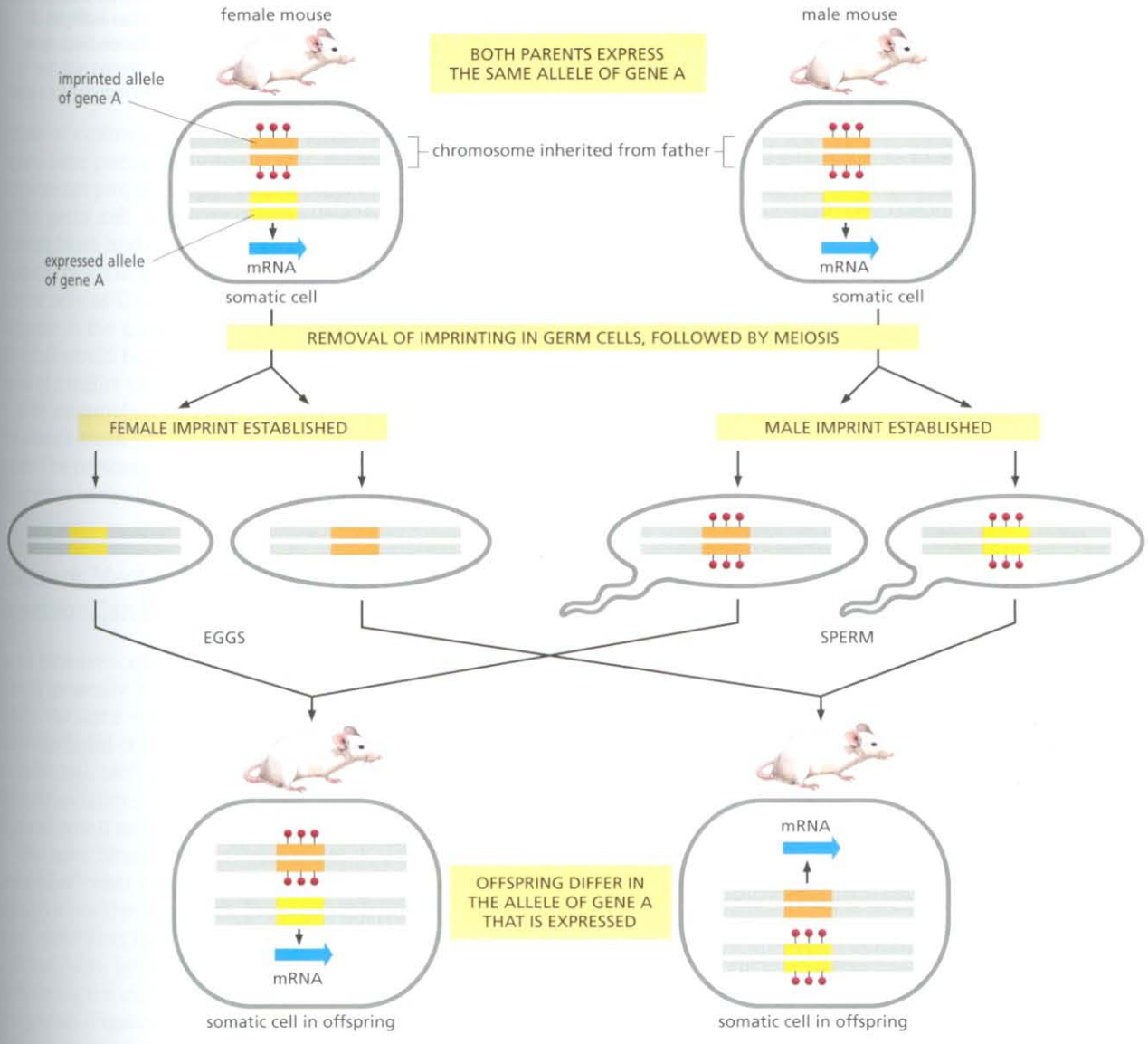


Figure 7–82 Imprinting in the mouse. The *top* portion of the figure shows a pair of homologous chromosomes in the somatic cells of two adult mice, one male and one female. In this example, both mice have inherited the top homolog from their father and the bottom homolog from their mother, and the paternal copy of a gene subject to imprinting (indicated in *orange*) is methylated, preventing its expression. The maternally derived copy of the same gene *(yellow)* is expressed. The remainder of the figure shows the outcome of a cross between these two mice. During germ cell formation, but before meiosis, the imprints are erased and then, much later in germ cell development, they are reimposed in a sex-specific pattern (*middle* portion of figure). In eggs produced from the female, neither allele of the A gene is methylated. In sperm from the male, both alleles of gene A are methylated. Shown at the *bottom* of the figure are two of the possible imprinting patterns inherited by the progeny mice; the mouse on the *left* has the same imprinting pattern as each of the parents, whereas the mouse on the *right* has the opposite pattern. If the two alleles of A gene are distinct, these different imprinting patterns can cause phenotypic differences in the progeny mice, even though they carry exactly the same DNA sequences of the two A gene alleles. Imprinting provides an important exception to classical genetic behavior, and several hundred mouse genes are thought to be affected in this way. However, the majority of mouse genes are not imprinted, and therefore the rules of Mendelian inheritance apply to most of the mouse genome.

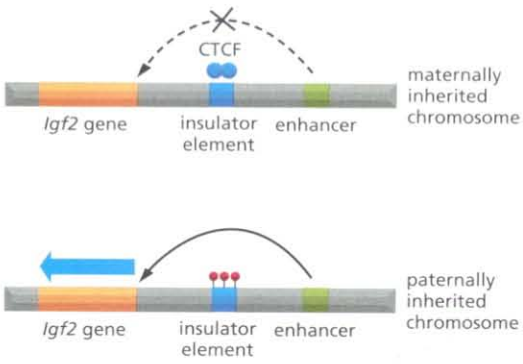


Figure 7–83 Mechanism of imprinting the mouse *Igf2* gene. On chromosome inherited from the female, a protein called CTCF binds to an insulator (see Figure 7–62), blocking communication between the enhancer *(green)* and the *Igf2* gene *(orange)*. IGF2 is therefore expressed from the maternally inherited chromosome. Because of imprinting, the insulator on the male-derived chromosome is methylated; this inactivates the insulator, by blocking binding of the CTCF protein, and allows the enhancer to activate transcription of the *Igf2* gene. In other examples of imprinting, methylation blocks gene expression by interfering with the binding of proteins required for a gene transcription.

parental origin of each of the two copies of the gene and to regulate their expression accordingly. In most cases, the methyl imprint silences nearby gene expression. In some cases, however, the methyl imprint can activate expression of a gene. In the case of *Igf2*, for example, methylation of an insulator element (see Figure 7–62) on the paternally derived chromosome blocks its function and allows a distant enhancer to activate transcription of the *Igf2* gene. On the maternally derived chromosome, the insulator is not methylated and the *Igf2* gene is therefore not transcribed (**Figure 7–83**).

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

21th

CG-Rich Islands Are Associated with Many Genes in Mammals

Because of the way in which DNA repair enzymes work, methylated C nucleotides in the genome tend to be eliminated in the course of evolution. Accidental deamination of an unmethylated C gives rise to U (see Figure 5–45), which is not normally present in DNA and thus is recognized easily by the DNA repair enzyme uracil DNA glycosylase, excised, and then replaced with a C (as discussed in Chapter 5). But accidental deamination of a 5-methyl C cannot be repaired in this way, for the deamination product is a T and so is indistinguishable from the other, nonmutant T nucleotides in the DNA. Although a special repair system exists to remove these mutant T nucleotides, many of the deaminations escape detection, so that those C nucleotides in the genome that are methylated tend to mutate to T over evolutionary time.

During the course of evolution, more than three out of every four CGs have been lost in this way, leaving vertebrates with a remarkable deficiency of this dinucleotide. The CG sequences that remain are very unevenly distributed in the genome; they are present at 10–20 times their average density in selected regions, called **CG islands**, which are 1000–2000 nucleotide pairs long. These islands, with some important exceptions, seem to remain unmethylated in all cell types. They often surround the promoters of the so-called *housekeeping genes*—those genes that code for the many proteins that are essential for cell viability and are therefore expressed in most cells (**Figure 7–84**).

The distribution of CG islands (also called CpG islands to distinguish the CG dinucleotides from the CG base pair) can be explained if we assume that CG methylation was adopted in vertebrates primarily as a way of maintaining DNA in a transcriptionally inactive state (see Figure 7–81). In vertebrates, new methyl-C to T mutations can be transmitted to the next generation only if they occur in the germ line, the cell lineage that gives rise to sperm or eggs. Most of the DNA in vertebrate germ cells is inactive and highly methylated. Over long

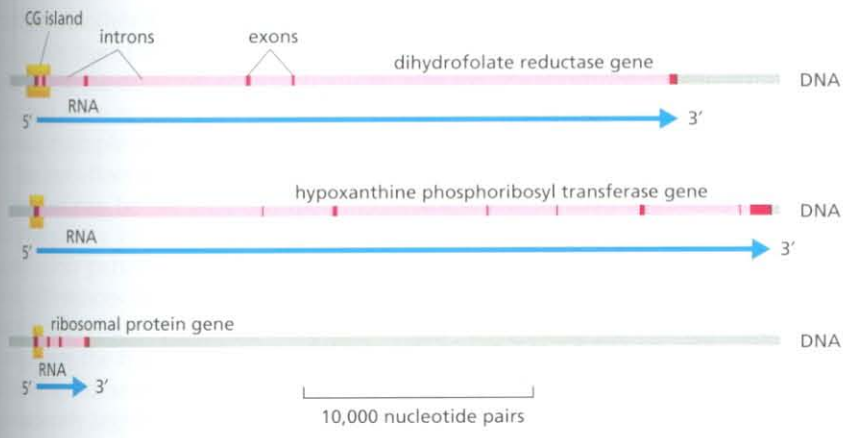


Figure 7–84 The CG islands surrounding the promoter in three mammalian housekeeping genes. The *yellow boxes* show the extent of each island. As for most genes in mammals (see Figure 6–25), the exons *(dark red)* are very short relative to the introns *(light red)*. (Adapted from A.P. Bird, *Trends Genet.* 3:342–347, 1987. With permission from Elsevier.)

periods of evolutionary time, the methylated CG sequences in these inactive regions have presumably been lost through spontaneous deamination events that were not properly repaired. However, promoters of genes that remain active in the germ cell lineages (including most housekeeping genes) are kept unmethylated, and therefore spontaneous deaminations of Cs that occur within them can be accurately repaired. Such regions are preserved in modern-day vertebrate cells as CG islands (**Figure 7–85**). In addition, any mutation of a CG sequence in the genome that destroyed the function or regulation of a gene in the adult would be selected against, and some CG islands are presumably the result of a higher than normal density of critical CG sequences for these genes.

The mammalian genome contains an estimated 20,000 CG islands. Most of the islands mark the 5′ ends of transcription units and thus, presumably, of genes. The presence of CG islands thereby provides a convenient way of identifying genes in the DNA sequences of vertebrate genomes.

Epigenetic Mechanisms Ensure That Stable Patterns of Gene Expression Can Be Transmitted to Daughter Cells

As we have seen, once a cell in an organism differentiates into a particular cell type, it generally remains specialized in that way; if it divides, its daughters inherit the same specialized character. For example, liver cells, pigment cells, and endothelial cells (discussed in Chapter 23) divide many times in the life of an individual, each of them faithfully producing daughter cells of the same type. Such differentiated cells must remember their specific pattern of gene expression and pass it on to their progeny through all subsequent cell divisions.

We have already described several ways of enabling daughter cells to "remember" what kind of cells they are supposed to be. One of the simplest is through a positive feedback loop in which a key gene regulatory protein activates, either directly or indirectly, the transcription of its own gene (see Figures 7–68 and 7–69). Interlocking positive feedback loops provide even more stability by buffering the circuit against fluctuations in the level of any one gene regulatory protein (Figures 7–75B and 7–78). We also saw above that DNA methylation can serve as a means for propagating gene expression patterns to descendants (see Figure 7–80).

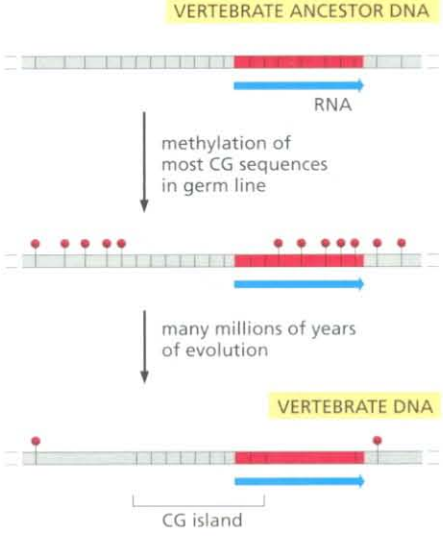


Figure 7–85 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes. A *black line* marks the location of a CG dinucleotide in the DNA sequence, while a *red "lollipop"* indicates the presence of a methyl group on the CG dinucleotide. CG sequences that lie in regulatory sequences of genes that are transcribed in germ cells are unmethylated and therefore tend to be retained in evolution. Methylated CG sequences, on the other hand, tend to be lost through deamination of 5-methyl C to T, unless the CG sequence is critical for survival.

Positive feedback loops and DNA methylation are common to both bacteria and eucaryotes; but eucaryotes also have available to them another means of maintaining a differentiated state through many cell generations. As we saw in Chapter 4, chromatin structure itself can be faithfully propagated from parent to daughter cell. There are several mechanisms to bring this about, but the simplest is based on the covalent modifications of histones. As we have seen, these modifications form a "histone code," with different patterns of modification serving as binding sites for different reader proteins. If these proteins, in turn, serve as (or attract) writer enzymes that replicate the very modification patterns that attracted them in the first place, then the distribution of active and silent regions of chromatin can be faithfully propagated (see Figure 5–39). In a sense, self-sustaining modification of histones is a form of positive feedback loop that is tied to the DNA but does not require the participation of the underlying DNA sequences.

The ability of a daughter cell to retain a memory of the gene expression patterns that were present in the parent cell is an example of **epigenetic inheritance**. This term has subtly different meanings in different branches of biology, but we will use it in its broadest sense to cover any heritable difference in the phenotype of a cell or an organism that does not result from changes in the nucleotide sequence of DNA (see Figure 4–35). We have just discussed three of the most important mechanisms underlying epigenetic changes, but others also exist (**Figure 7–86**). Cells often combine these mechanisms to ensure that patterns of gene expression are maintained and inherited accurately and reliably—over a period of up to a hundred years or more, in our own case.

For more than half a century, biologists have been preoccupied with DNA as the carrier of heritable information—and rightly so. However, it has become clear that human chromosomes also carry a great deal of information that is epigenetic, and not contained in the sequence of the DNA itself. Imprinting is one

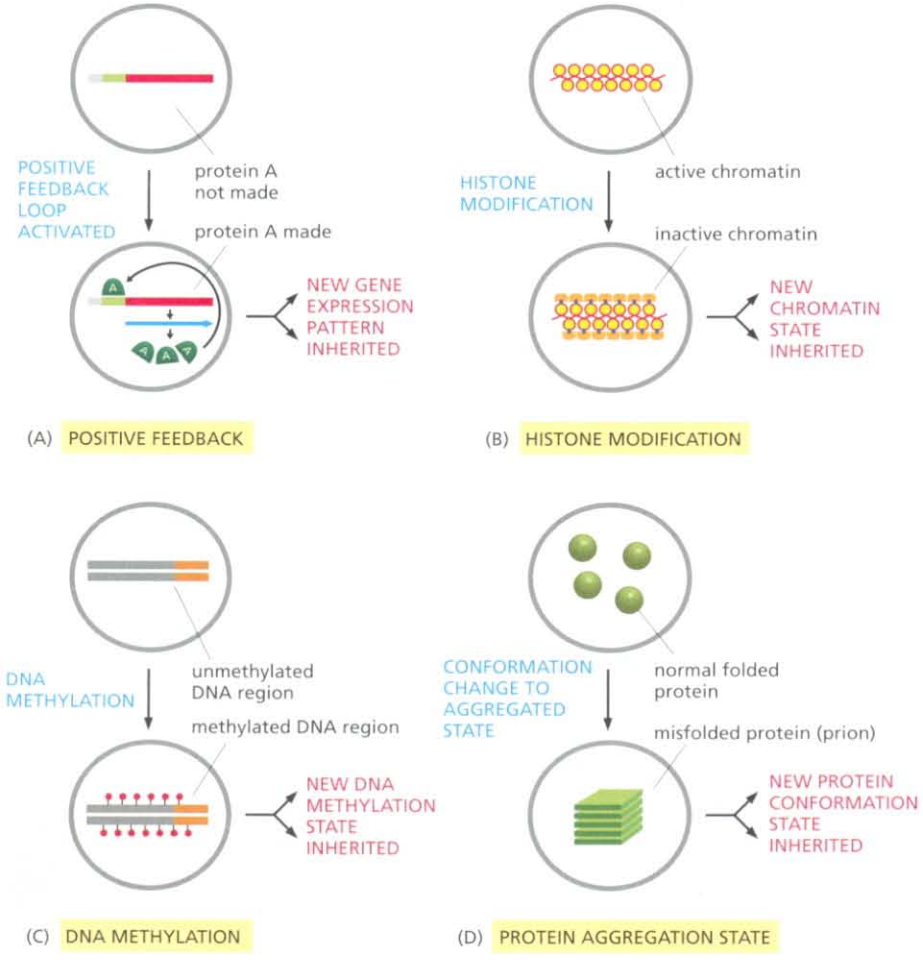


Figure 7–86 Four distinct mechanisms that can produce an epigenetic form of inheritance in an organism. (For the inheritance of histone modifications, see Figure 4–52; for the inheritance of protein conformations, see Figure 6–95.)

example. Another is seen in the phenomenon of *mono-allelic expression*, in which only one of the two copies of certain human genes is expressed. For many such genes, the decision of which allele to express and which to silence is random, but once made, it is passed on to progeny cells. Below, we will see an extreme example of this phenomenon in X-chromosome inactivation.

The net effect of random and environmentally triggered epigenetic changes in humans can be seen by comparing identical twins: their genomes have the same sequence of nucleotides, but when their histone modification and DNA methylation patterns are compared, many differences are observed. Because these differences are roughly correlated not only with age but also with the time that the twins have spent apart from each other, it is believed that some of these changes are the result of environmental factors (**Figure 7–87**). Although studies of the *epigenome* are in early stages, the idea that environmental events can be permanently registered by our cells is a fascinating one that presents an important challenge to the next generation of biological scientists.

Figure 7–87 Identical twins raised apart from one another. (Courtesy of Nancy L. Segal.)

Chromosome-Wide Alterations in Chromatin Structure Can Be Inherited

We have seen that chromatin states and DNA methylation can be heritable, serving to establish and preserve patterns of gene expression for many cell generations. Perhaps the most striking example of this effect occurs in mammals, in which an alteration in the chromatin structure of an entire chromosome can modulate the levels of expression of all genes on that chromosome.

Males and females differ in their *sex chromosomes*. Females have two X chromosomes, whereas males have one X and one Y chromosome. As a result, female cells contain twice as many copies of X-chromosome genes as do male cells. In mammals, the X and Y sex chromosomes differ radically in gene content: the X chromosome is large and contains more than a thousand genes, whereas the Y chromosome is small and contains less than 100 genes. Mammals have evolved a *dosage compensation* mechanism to equalize the dosage of X-chromosome gene products between males and females. Mutations that interfere with dosage compensation are lethal: the correct ratio of X chromosome to *autosome* (non-sex chromosome) gene products is critical for survival.

Mammals achieve dosage compensation by the transcriptional inactivation of one of the two X chromosomes in female somatic cells, a process known as **X-inactivation**. Early in the development of a female embryo, when it consists of a few thousand cells, one of the two X chromosomes in each cell becomes highly condensed into a type of heterochromatin. The condensed X chromosome can be easily seen under the light microscope in interphase cells; it was originally called a *Barr body* and is located near the nuclear membrane (**Figure 7–88**). As a result of X-inactivation, two X chromosomes can coexist within the same nucleus, exposed to the same diffusible gene regulatory proteins, yet differ entirely in their expression.

The initial choice of which X chromosome to inactivate, the maternally inherited one (X_m) or the paternally inherited one (X_p), is random. Once either X_p or X_m has been inactivated, it remains silent throughout all subsequent cell divisions of that cell and its progeny, indicating that the inactive state is faithfully maintained through many cycles of DNA replication and mitosis. Because X-inactivation is random and takes place after several thousand cells have already formed in the embryo, every female is a mosaic of clonal groups of cells

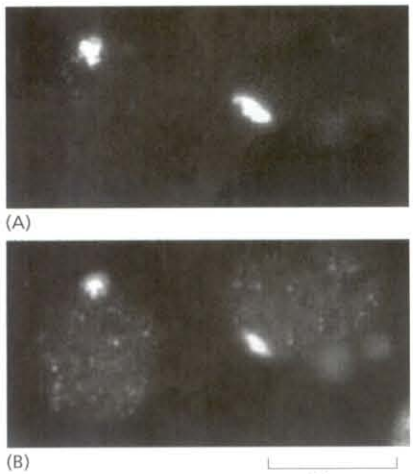


Figure 7–88 X-chromosome inactivation in female cells. (A) Only the inactive X chromosome is coated with XIST RNA, visualized here by *in situ* hybridization to fluorescently labeled RNAs of complementary nucleotide sequence. The panel shows the nuclei of two adjacent cells. (B) The same sample, stained with antibodies against a component of the Polycomb group complex, which coats the X chromosome and helps to silence expression of its genes. (From B. Panning, *Methods Enzymol.* 376:419–428, 2004. With permission from Academic Press.)

in which either X_p or X_m is silenced (**Figure 7–89**). These clonal groups are distributed in small clusters in the adult animal because sister cells tend to remain close together during later stages of development. For example, X-chromosome inactivation causes the red and black "tortoise-shell" coat coloration of some female cats. In these cats, one X chromosome carries a gene that produces red hair color, and the other X chromosome carries an allele of the same gene that results in black hair color; it is the random X-inactivation that produces patches of cells of two distinctive colors. In contrast to the females, male cats of this genetic stock are either solid red or solid black, depending on which X chromosome they inherit from their mothers.

Although X-chromosome inactivation is maintained over thousands of cell divisions, it is not always permanent. In particular, it is reversed during germ-cell formation, so that all haploid oocytes contain an active X chromosome and can express X-linked gene products.

How is an entire chromosome transcriptionally inactivated? X-chromosome inactivation is initiated and spreads from a single site in the middle of the X chromosome, the **X-inactivation center** (**XIC**). Encoded within the XIC is an unusual RNA molecule, *XIST RNA*, which is expressed solely from the inactive X chromosome and whose expression is necessary for X-inactivation. The XIST RNA is not translated into protein and remains in the nucleus, where it eventually coats the entire inactive X chromosome. The spread of XIST RNA from the XIC over the entire chromosome correlates with the spread of gene silencing, indicating that XIST RNA drives the formation and spread of heterochromatin (**Figure 7–90**). Curiously, about 10% of the genes on the X-chromosome escape this silencing and remain active.

In addition to containing XIST RNA, the X-chromosome heterochromatin is characterized by a specific variant of histone 2A, by hypoacetylation of histones

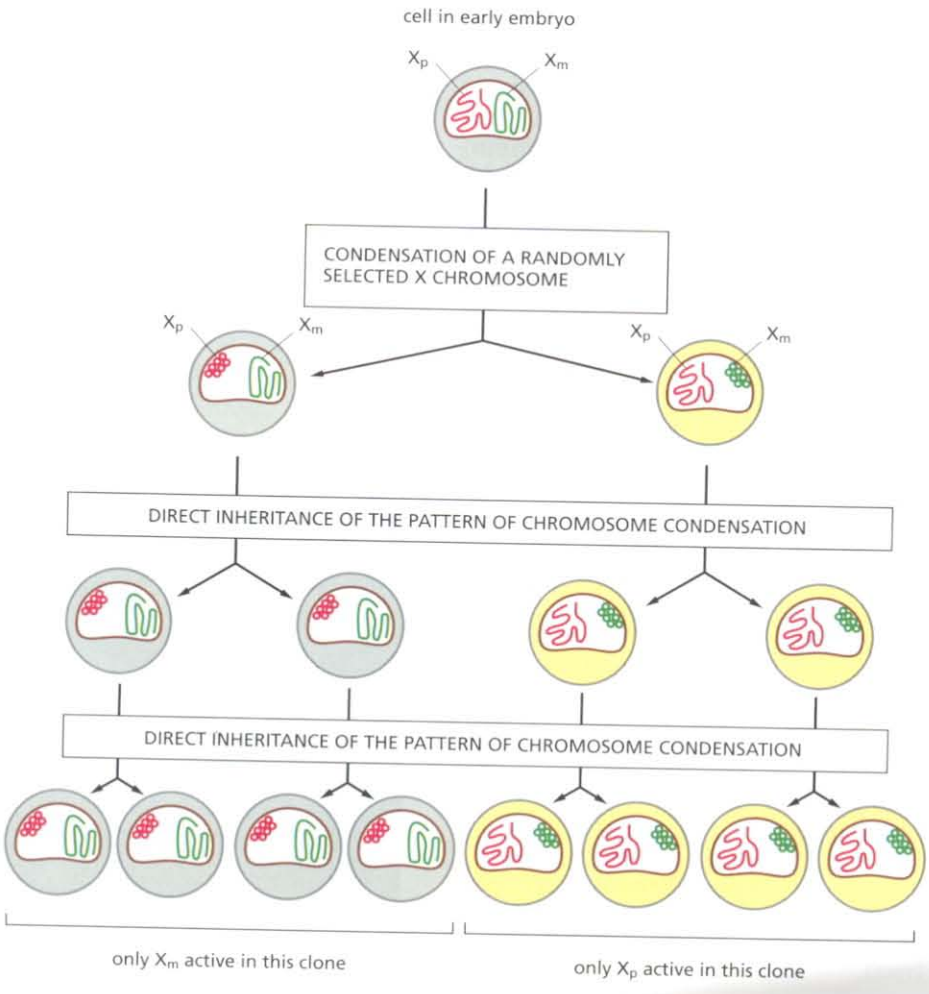


Figure 7–89 X-inactivation. The clonal inheritance of a condensed inactive X chromosome that occurs in female mammals.

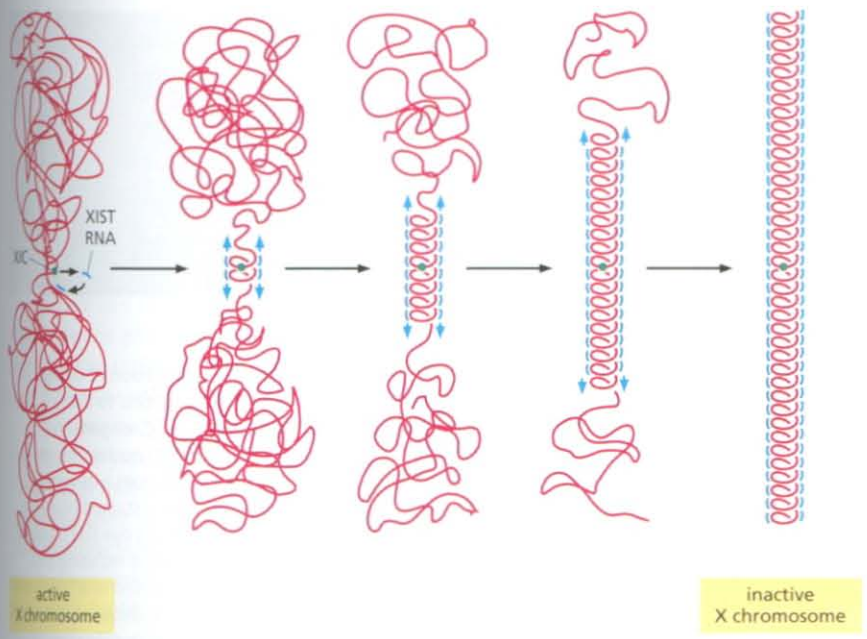


Figure 7–90 Mammalian X-chromosome inactivation. X-chromosome inactivation begins with the synthesis of XIST (X-inactivation specific transcript) RNA from the XIC (X-inactivation center) locus. The association of XIST RNA with one of a female's two X chromosomes is correlated with the condensation of that chromosome. Early in embryogenesis, both XIST association and chromosome condensation gradually move from the XIC locus outward to the chromosome ends. The details of how this occurs remain to be deciphered.

H3 and H4, by ubiquitylation of histone 2A, by methylation of a specific position on histone H3, and by specific patterns of methylation on the underlying DNA (for a suggestion of how these features may be causally linked, see Figure 7–81). The combination of such modifications presumably makes most of the inactive X chromosome unusually resistant to transcription. Because these modifications are, at least in principle, self-propagating, it is easy to see how, once formed, an inactive X chromosome can be stably maintained through many cell divisions.

Many features of mammalian X-chromosome inactivation remain to be discovered. How is the initial decision made as to which X chromosome to inactivate? What mechanism prevents the other X chromosome from also being inactivated? How does XIST RNA coordinate the formation of heterochromatin? How do some genes on the X chromosome escape inactivation? We are just beginning to understand this mechanism of gene regulation that is crucial for the survival of our own species.

X-chromosome inactivation in females is only one of the ways in which sexually reproducing organisms achieve dosage compensation. In *Drosophila*, most of the genes on the single X chromosome present in male cells are transcribed at approximately twofold higher levels than their counterparts in female cells. This male-specific "up-regulation" of transcription results from an alteration in chromatin structure over the entire male X chromosome. A dosage-compensation complex, containing several histone-modifying enzymes as well as two noncoding RNAs transcribed from the X chromosome, assembles at hundreds of positions along the X chromosome and produces patterns of histone modification that are thought to upregulate transcription—through effects on either initiation or elongation—at most genes on the male X chromosome.

The nematode worm uses a third strategy for dosage compensation. Here, the two sexes are male (with one X chromosome) and hermaphrodite (with two X chromosomes), and dosage compensation occurs by an approximately twofold "down-regulation" of transcription from each of the two X chromosomes in cells of the hermaphrodite. This is brought about through chromosome-wide structural changes in the X chromosomes of hermaphrodites (**Figure 7–91**). A dosage-compensation complex, which is completely different from that of *Drosophila* and resembles instead the *condensin* complex that compacts chromosomes during mitosis and meiosis (see Figure 17–27), assembles along each X chromosome of hermaphrodites and, by an unknown mechanism, superimposes a global twofold repression on the normal expression level of each gene.

Although the strategy and components used to cause dosage compensation differ between mammals, flies, and worms, they all involve structural alterations over the entire X chromosome. It seems likely that features of chromosome

structure that are quite general were independently adapted and harnessed during evolution to overcome a highly specific problem in gene regulation encountered by sexually reproducing animals.

The Control of Gene Expression is Intrinsically Noisy

So far in this chapter we have discussed gene expression as though it were a strictly deterministic process, so that, if only one knew the concentrations of all the relevant regulatory proteins and other control molecules, the level of gene expression would be precisely predictable. In reality, there is a large amount of random variation in the behavior of cells. In part, this is because there are random fluctuations in the environment, and these disturb the concentrations of regulatory molecules inside the cell in unpredictable ways. Another possible cause, in some cases, may be chaotic behavior of the intracellular control system: mathematical analysis shows that even quite simple control systems may be acutely sensitive to the control parameters, in such a way that, for example, a tiny difference of initial conditions may lead to a radically different long-term outcome. But in addition to these causes of unpredictability, there is a further, more fundamental reason why all cell behavior is inescapably random to some degree.

Cells are chemical systems consisting of relatively small numbers of molecules, and chemical reactions at the level of individual molecules occur in an essentially random, or *stochastic*, manner. A given molecule has a certain probability per unit time of undergoing a chemical reaction, but whether it will actually do so at any given moment is unpredictable, depending on random thermal collisions and the probabilistic rules of quantum mechanics. The smaller the number of molecules governing a process inside the cell, the more severely it will be affected by the randomness of chemical events at the single-molecule level. Thus there is some degree of randomness in every aspect of cell behavior, but certain processes are liable to be random in the extreme.

The control of transcription, in particular, depends on the precise chemical condition of the gene. Consider a simple idealized case, in which a gene is transcribed so long as it has a transcriptional activator protein bound to its regulatory region, and transcriptionally silent when this protein is not bound. The association/dissociation reaction between the regulatory DNA and the protein is stochastic: if the bound state has a half-life $t_{1/2}$ of an hour, the gene may remain activated sometimes for 30 minutes or less, sometimes for a couple of hours or more at a stretch, before the activator protein dissociates. In this way, transcription will flicker on and off in an essentially random way. The average rate of flickering, and the ratio of the average time spent in the "on" state to the average time spent in the "off" state, will be determined by the k_{off} and k_{on} values for the binding reaction and by the concentration of the activator protein in the cell. The quantity of gene transcripts accumulated in the cell will fluctuate accordingly; if the lifetime of the transcripts is long compared with $t_{1/2}$, the fluctuations will be smoothed out; if it is short, they will be severe.

One way to demonstrate such random fluctuations in the expression of individual gene copies is to genetically engineer cells in which one copy of a gene control region is linked to a sequence coding for a green fluorescent reporter protein, while another copy is linked similarly to a sequence coding for a red fluorescent reporter. Although both these gene constructs are in the same cell and experiencing the same environment, they fluctuate independently in their level of expression. As a result, in a population of cells that all carry the same pair of constructs, some cells appear green, others red, and still others a mixture of the two colors, and thus in varying shades of yellow (see Figure 8–75). More generally, cell fate decisions are often made in a stochastic manner, presumably as a result of such random fluctuations; we shall encounter an example in Chapter 23, where we discuss the genesis of the different types of white blood cells.

In some types of cells, and for some aspects of cell behavior, randomness in the control of gene transcription, such as we have just described, seems to be the major source of random variability; in other cell types, other sources of random variation predominate. Where noise in a control system would be

Figure 7–91 Localization of dosage compensation proteins to the X chromosomes of *C. elegans* hermaphrodite (XX) nuclei. This image shows many nuclei from a developing embryo. Total DNA is stained *blue* with the DNA-intercalating dye DAPI, and the Sdc2 protein is stained *red* using anti-Sdc2 antibodies coupled to a fluorescent dye. This experiment shows that the Sdc2 protein associates with only a limited set of chromosomes, identified by other experiments to be the two X chromosomes. Sdc2 is bound along the entire length of the X chromosome and recruits the dosage-compensation complex. (From H.E. Dawes et al., Science 284:1800–1804, 1999. With permission from AAAS.)

harmful, special control mechanisms have evolved to minimize its effects; the feed-forward loop discussed earlier is an example of such a device, serving to filter out the effects of rapid fluctuations in a control signal. But in all cells, some degree of randomness is inevitable. It is a fundamental feature of cell behavior.

Summary

The many types of cells in animals and plants are created largely through mechanisms that cause different sets of genes to be transcribed in different cells. Since specialized animal cells can maintain their unique character through many cell division cycles and even when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides. These features reflect the cell's memory of its developmental history. Bacteria and yeasts also exhibit cell memory and provide unusually accessible model systems in which to study gene regulatory mechanisms.

Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide the simplest mechanism for cell memory. Transcription circuits also provide the cell with the means to carry out logic operations and measure time. Simple transcription circuits combined into large regulatory networks drive highly sophisticated programs of embryonic development.

In eucaryotes, the transcription of any particular gene is generally controlled by a combination of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific set of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be active in a variety of circumstances and is typically involved in the regulation of many different genes.

Unlike bacteria, eucaryotic cells use inherited states of chromatin condensation as an additional mechanism to regulate gene expression and to create cell memory. An especially dramatic case is the inactivation of an entire X chromosome in female mammals. DNA methylation can also silence genes in a heritable manner in eucaryotes. In addition, it underlies the phenomenon of genomic imprinting in mammals, in which the expression of a gene depends on whether it was inherited from the mother or the father.

22th

POST-TRANSCRIPTIONAL CONTROLS

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

In the following sections, we consider the varieties of post-transcriptional regulation in temporal order, according to the sequence of events that an RNA molecule might experience after its transcription has begun (**Figure 7–92**).

Transcription Attenuation Causes the Premature Termination of Some RNA Molecules

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

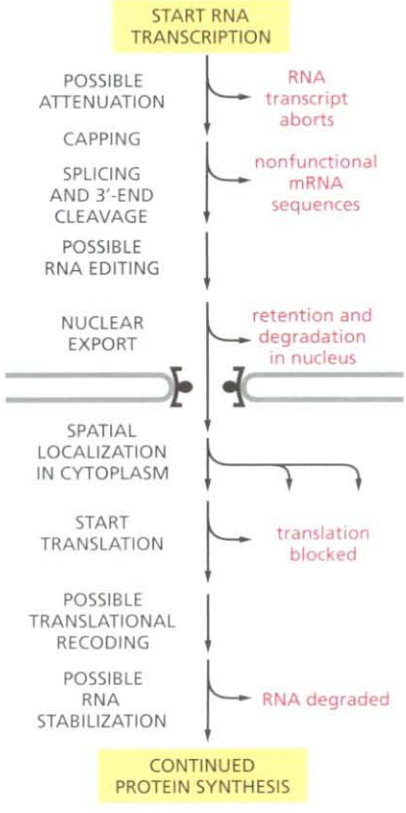


Figure 7–92 Post-transcriptional controls on gene expression. The final synthesis rate of a protein can, in principle, be controlled at any of the steps shown. RNA splicing, RNA editing, and translation recoding (described in Chapter 6) can also alter the sequence of amino acids in a protein, making it possible for the cell to produce more than one protein variant from the same gene. Only a few of the steps depicted here are likely to be critical for the regulation of any one particular protein.