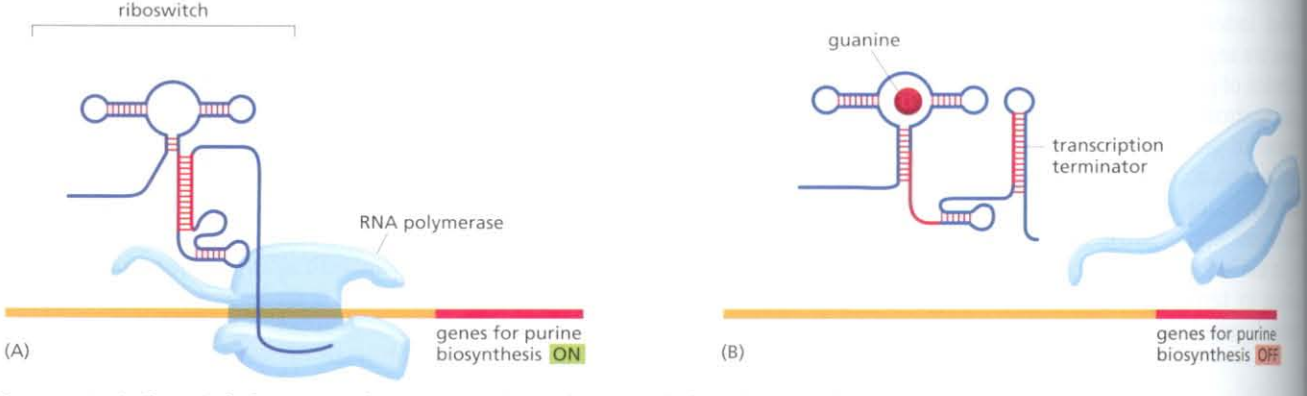
Transcription attenuation also operates in eucaryotes. A well-studied example occurs during the life cycle of HIV, the human immunodeficiency virus, causative agent of acquired immune deficiency syndrome, or AIDS. Once it has been integrated into the host genome, the viral DNA is transcribed by the cell's RNA polymerase II (see Figure 5–71). However, the host polymerase usually terminates transcription after synthesizing transcripts of several hundred nucleotides and therefore does not efficiently transcribe the entire viral genome. When conditions for viral growth are optimal, a virus-encoded protein called Tat, which binds to a specific stem-loop structure in the nascent RNA that contains a "bulged base," prevents this premature termination. Once bound to this specific RNA structure (called Tar), Tat assembles several cell proteins that allow the RNA polymerase to continue transcribing. The normal role of at least some of these proteins is to prevent pausing and premature termination by RNA polymerase when it transcribes normal cell genes. Eucaryotic genes often contain long introns; to transcribe a gene efficiently, RNA polymerase II cannot afford to linger at nucleotide sequences that happen to promote pausing. Thus, a normal cell mechanism has apparently been adapted by HIV to permit efficient transcription of its genome to be controlled by a single viral protein.

Riboswitches Might Represent Ancient Forms of Gene Control

In Chapter 6, we discussed the idea that, before modern cells arose on Earth, RNA both stored hereditary information and catalyzed chemical reactions. The recent discovery of *riboswitches* shows that RNA can also form control devices that regulate gene expression. Riboswitches are short sequences of RNA that change their conformation on binding small molecules, such as metabolites. Each riboswitch recognizes a specific small molecule and the resulting conformational change is used to regulate gene expression. Riboswitches are often located near the 5′ end of mRNAs, and they fold while the mRNA is being synthesized blocking or permitting progress of the RNA polymerase according to whether the regulatory small molecule is bound (**Figure 7–93**).

Riboswitches are particularly common in bacteria, in which they sense key small metabolites in the cell and adjust gene expression accordingly. Perhaps their most remarkable feature is the high specificity and affinity with which each recognizes only the appropriate small molecule; in many cases, every chemical feature of the small molecule is read by the RNA (Figure 7–93C). Moreover, the



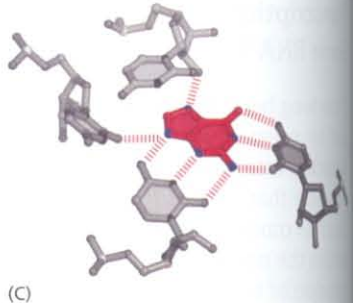

(C)

Figure 7–93 A riboswitch that responds to guanine. (A) In this example from bacteria, the riboswitch controls expression of the purine biosynthetic genes. When guanine levels in cells are low, an elongating RNA polymerase transcribes the purine biosynthetic genes, and the enzymes needed for guanine synthesis are therefore expressed. (B) When guanine is abundant, it binds the riboswitch, causing it to undergo a conformational change that forces the RNA polymerase to terminate transcription (see Figure 6–11). (C) Guanine *(red)* bound to the riboswitch. Only those nucleotides that form the guanine-binding pocket are shown. Many other riboswitches exist, including those that recognize *S*-adenosyl methione, coenzyme B_{12}, flavin mononucleotide, adenine, lysine, and glycine. (Adapted from M. Mandal and R.R. Breaker, *Nat. Rev. Mol. Cell Biol.* 5:451–63, 2004, with permission from Macmillan Publishers Ltd., and C.K. Vanderpool and S. Gottesman, *Mol. Microbiol.* 54:1076–1089, 2004, with permission from Blackwell Publishing.)

binding affinities observed are as tight as those typically observed between small molecules and proteins.

Riboswitches are perhaps the most economical examples of gene control devices, inasmuch as they bypass the need for regulatory proteins altogether. In the example shown in Figure 7–93, the riboswitch controls transcription elongation, but they also regulate other steps in gene expression, as we shall see later in this chapter. Clearly, highly sophisticated gene control devices can be made from short sequences of RNA.

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

As discussed in Chapter 6, RNA splicing shortens the transcripts of many eucaryotic genes by removing the intron sequences from the mRNA precursor. We also saw that a cell can splice an RNA transcript differently and thereby make different polypeptide chains from the same gene—a process called **alternative RNA splicing** (see Figure 6–27 and **Figure 7–94**). A substantial proportion of animal genes (estimated at 40% in flies and 75% in humans) produce multiple proteins in this way.

When different splicing possibilities exist at several positions in the transcript, a single gene can produce dozens of different proteins. In one extreme case, a *Drosophila* gene may produce as many as 38,000 different proteins from a single gene through alternative splicing (**Figure 7–95**), although only a fraction of these forms have thus far been experimentally observed. Considering that the *Drosophila* genome has approximately 14,000 identified genes, it is clear that the protein complexity of an organism can greatly exceed the number of its genes. This example also illustrates the perils in equating gene number with an organism's complexity. For example, alternative splicing is relatively rare in single-celled budding yeasts but very common in flies. Budding yeast has ~6200 genes, only about 300 of which are subject to splicing, and nearly all of these have only a single intron. To say that flies have only 2–3 times as many genes as yeasts greatly underestimates the difference in complexity of these two genomes.

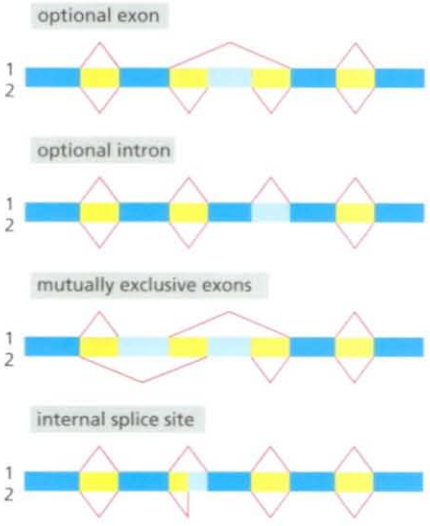


Figure 7–94 Four patterns of alternative RNA splicing. In each case a single type of RNA transcript is spliced in two alternative ways to produce two distinct mRNAs (1 and 2). The *dark blue boxes* mark exon sequences that are retained in both mRNAs. The *light blue boxes* mark possible exon sequences that are included in only one of the mRNAs. The boxes are joined by *red lines* to indicate where intron sequences *(yellow)* are removed. (Adapted with permission from A. Andreadis, M.E. Gallego and B. Nadal-Ginard, *Annu. Rev. Cell Biol.* 3:207–242, 1987. With permission from Annual Reviews.)

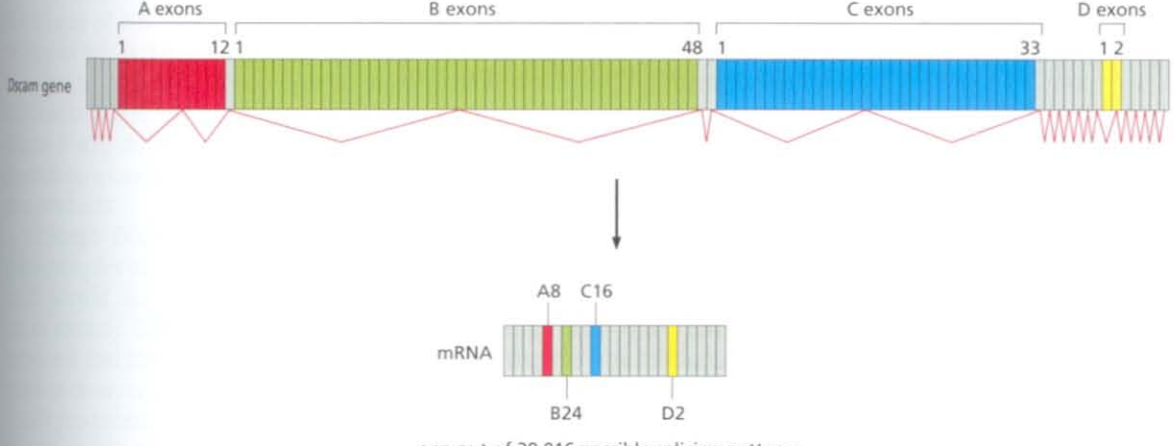


Figure 7–95 Alternative splicing of RNA transcripts of the *Drosophila Dscam* gene. DSCAM proteins are axon guidance receptors that help to direct growth cones to their appropriate targets in the developing nervous system. The final mRNA contains 24 exons, four of which (denoted A, B, C, and D) are present in the *Dscam* gene as arrays of alternative exons. Each RNA contains 1 of 12 alternatives for exon A *(red)*, 1 of 48 alternatives for exon B *(green)*, 1 of 33 alternatives for exon C *(blue)*, and 1 of 2 alternatives for exon D *(yellow)*. If all possible splicing combinations are used, 38,016 different proteins could in principle be produced from the *Dscam* gene. This figure shows only one of the many possible splicing patterns (indicated by the red line and by the mature mRNA below it). Each variant Dscam protein would fold into roughly the same structure [predominantly a series of extracellular immunoglobulin-like domains linked to a membrane-spanning region (see Figure 25–74)], but the amino acid sequence of the domains would vary according to the splicing pattern. It is suspected that this receptor diversity contributes to the formation of complex neural circuits, but the precise properties and functions of the many Dscam variants are not yet understood. (Adapted from D.L. Black, *Cell* 103:367–370, 2000. With permission from Elsevier.)

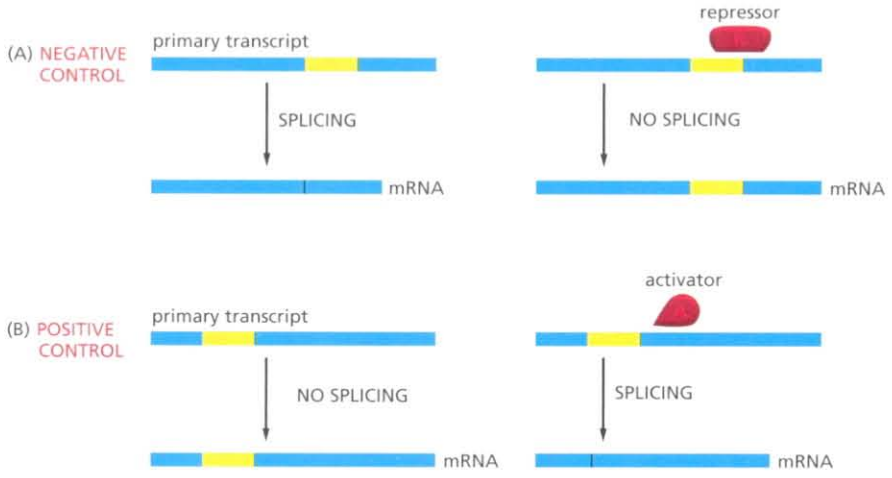


Figure 7–96 Negative and positive control of alternative RNA splicing. (A) In negative control, a repressor protein binds to the pre-mRNA transcript and blocks access of the splicing machinery to a splice junction. This often results in the use of a cryptic splice site, thereby producing an altered pattern of splicing (not shown). (B) In positive control, the splicing machinery is unable to remove a particular intron sequence efficiently without assistance from an activator protein. Because the nucleotide sequences that bind these activators can be located many nucleotide pairs from the splice junctions they control, they are often called *splicing enhancers*.

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

In many cases, however, alternative RNA splicing is regulated rather than constitutive. In the simplest examples, regulated splicing is used to switch from the production of a nonfunctional protein to the production of a functional one. The transposase that catalyzes the transposition of the *Drosophila* P element, for example, is produced in a functional form in germ cells and a nonfunctional form in somatic cells of the fly, allowing the P element to spread throughout the genome of the fly without causing damage in somatic cells (see Figure 5–69). The difference in transposon activity has been traced to the presence of an intron sequence in the transposase RNA that is removed only in germ cells.

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

RNA splicing can be regulated either negatively, by a regulatory molecule that prevents the splicing machinery from gaining access to a particular splice site on the RNA, or positively, by a regulatory molecule that helps direct the splicing machinery to an otherwise overlooked splice site (**Figure 7–96**).

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

The Definition of a Gene Has Had to Be Modified Since the Discovery of Alternative RNA Splicing

The discovery that eucaryotic genes usually contain introns and that their coding sequences can be assembled in more than one way raised new questions about the definition of a gene. A gene was first clearly defined in molecular terms in the early 1940s from work on the biochemical genetics of the fungus *Neurospora*. Until then, a gene had been defined operationally as a region of the genome that segregates as a single unit during meiosis and gives rise to a definable phenotypic trait, such as a red or a white eye in *Drosophila* or a round

or wrinkled seed in peas. The work on *Neurospora* showed that most genes correspond to a region of the genome that directs the synthesis of a single enzyme. This led to the hypothesis that one gene encodes one polypeptide chain. The hypothesis proved fruitful for subsequent research; as more was learned about the mechanism of gene expression in the 1960s, a gene became identified as that stretch of DNA that was transcribed into the RNA coding for a single polypeptide chain (or a single structural RNA such as a tRNA or an rRNA molecule). The discovery of split genes and introns in the late 1970s could be readily accommodated by the original definition of a gene, provided that a single polypeptide chain was specified by the RNA transcribed from any one DNA sequence. But it is now clear that many DNA sequences in higher eucaryotic cells can produce a set of distinct (but related) proteins by means of alternative RNA splicing. How, then, is a gene to be defined?

In those relatively rare cases in which a single transcription unit produces two very different eucaryotic proteins, the two proteins are considered to be produced by distinct genes that overlap on the chromosome. It seems unnecessarily complex, however, to consider most of the protein variants produced by alternative RNA splicing as being derived from overlapping genes. A more sensible alternative is to modify the original definition to count as a gene any DNA sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains (protein isoforms). This definition of a gene also accommodates those DNA sequences that encode protein variants produced by post-transcriptional processes other than RNA splicing, such as translational frameshifting (see Figure 6–78), regulated poly-A addition, and RNA editing (to be discussed below).

Sex Determination in *Drosophila* Depends on a Regulated Series of RNA Splicing Events

We now turn to one of the best-understood examples of regulated RNA splicing. In *Drosophila* the primary signal for determining whether the fly develops as a male or female is the ratio of the number of X chromosomes (X) to the number of autosomal sets (A). Individuals with an X/A ratio of 1 (normally two X chromosomes and two sets of autosomes) develop as females, whereas those with an X/A a ratio of 0.5 (normally one X chromosome and two sets of autosomes) develop as males. This ratio is assessed early in development and is remembered thereafter by each cell. Three crucial gene products transmit information about this ratio to the many other genes that specify male and female characteristics (**Figure 7–97**). As explained in **Figure 7–98**, sex determination in *Drosophila* depends on a cascade of regulated RNA splicing events that involves these three gene products.

Although *Drosophila* sex determination provides one of the best-understood examples of a regulatory cascade based on RNA splicing, it is not clear why the fly should use this strategy. Other organisms (the nematode, for example) use an entirely different scheme for sex determination—one based on transcriptional and translational controls. Moreover, the *Drosophila* male-determination pathway requires that a number of nonfunctional RNA molecules be continually produced, which seems unnecessarily wasteful. One speculation is that this RNA-splicing cascade, like the ribsoswitches discussed above, represents an ancient control strategy, left over from the early stage of evolution in which RNA

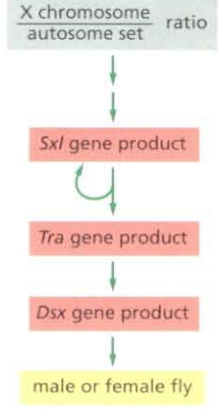


Figure 7–97 Sex determination in *Drosophila*. The gene products shown act in a sequential cascade to determine the sex of the fly according to the X-chromosome/autosome set ratio (X/A). The genes are called *Sex-lethal (Sxl), Transformer (Tra),* and *Doublesex (Dsx)* because of the phenotypes that result when the gene is inactivated by mutation. The function of these gene products is to transmit the information about the X/A ratio to the many other genes that create the sex-related phenotypes. These other genes function as two alternative sets: those that specify female features and those that specify male features (see Figure 7–98).

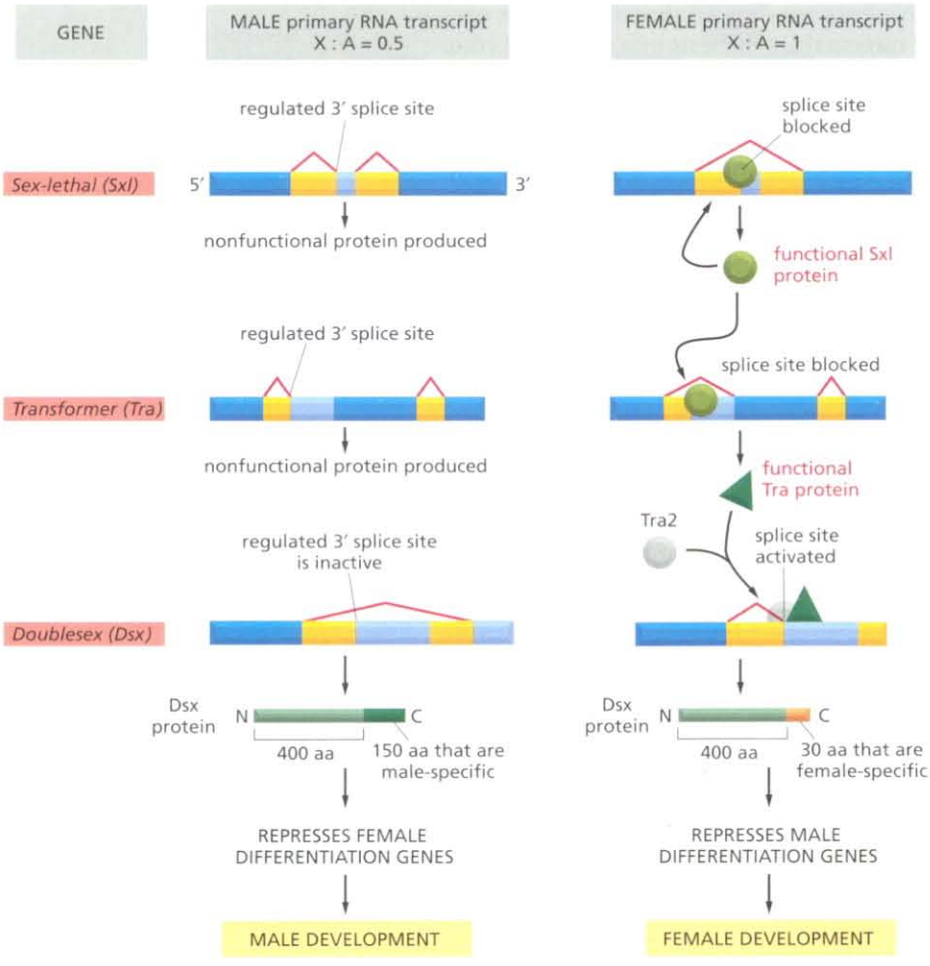


Figure 7–98 The cascade of changes in gene expression that determines the sex of a fly through alternative RNA splicing. An X-chromosome/autosome set ratio of 0.5 results in male development. Male is the "default" pathway in which the *Sxl* and *Tra* genes are both transcribed, but the RNAs are spliced constitutively to produce only nonfunctional RNA molecules, and the *Dsx* transcript is spliced to produce a protein that turns off the genes that specify female characteristics. An X/A ratio of 1 triggers the female differentiation pathway in the embryo by transiently activating a promoter within the *Sxl* gene that causes the synthesis of a special class of *Sxl* transcripts that are constitutively spliced to give functional Sxl protein. Sxl is a splicing regulatory protein with two sites of action: (1) it binds to the constitutively produced *Sxl* RNA transcript, causing a female-specific splice that continues the production of functional Sxl protein, and (2) it binds to the constitutively produced *Tra* RNA and causes an alternative splice of this transcript, which now produces an active Tra regulatory protein. The Tra protein acts with the constitutively produced Tra2 protein to produce the female-specific spliced form of the *Dsx* transcript; this encodes the female form of the Dsx protein, which turns off the genes that specify male features.

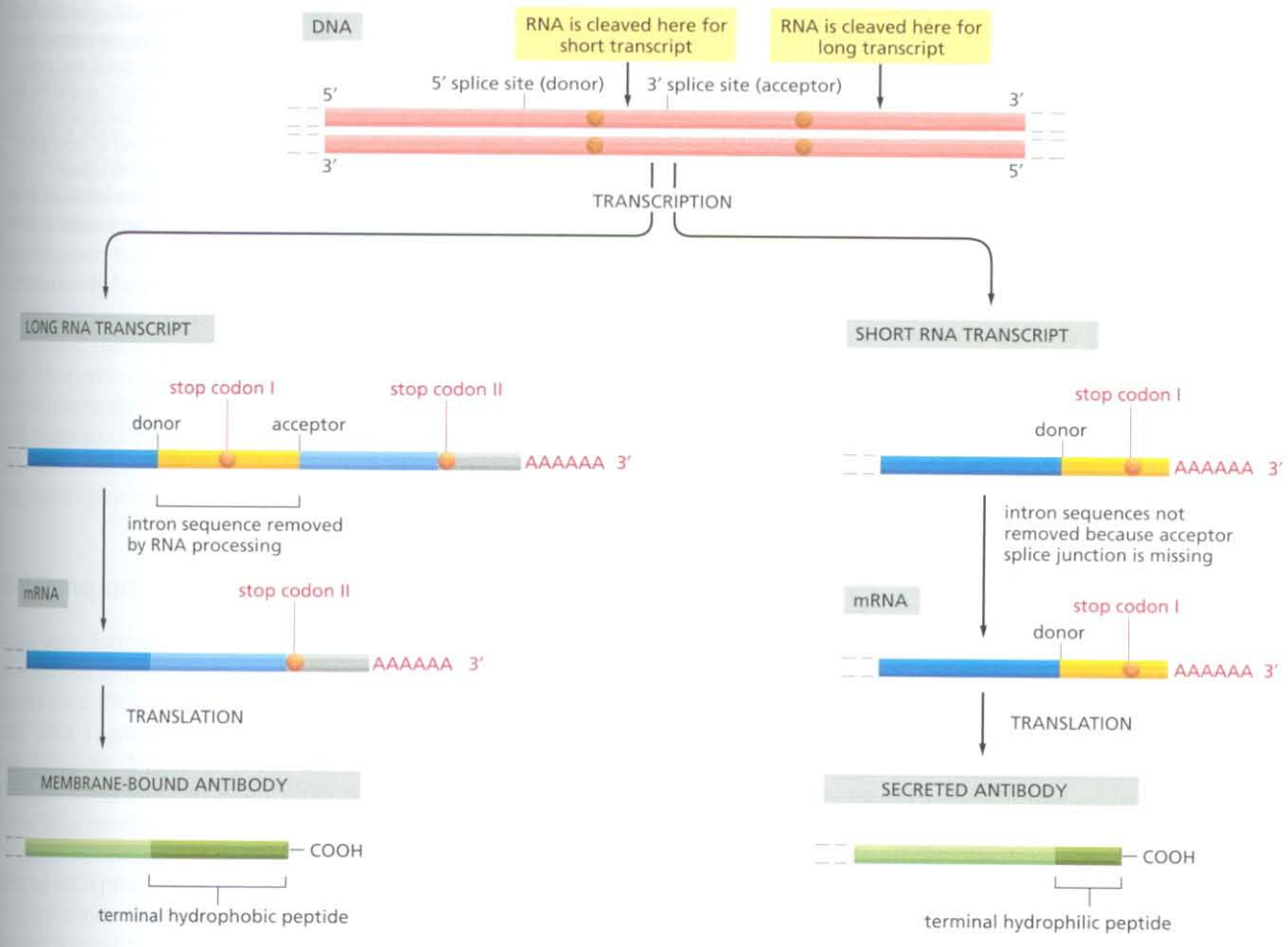
The components in this pathway were all initially identified through the study of *Drosophila* mutants that are altered in their sexual development. The *Dsx* gene, for example, derives its name (*Doublesex*) from the observation that a fly lacking this gene product expresses both male-specific and female-specific features. Note that this pathway includes both negative and positive control of splicing (see Figure 7–96). Sxl binds to the pyrimidine-rich stretch of nucleotides that is part of the standard splicing consensus sequence and blocks access by the normal splicing factor, U2AF (see Figure 6–29). Tra binds to specific RNA sequences in an exon and with Tra2 activates a normally suboptimal splicing signal by the binding of U2AF.

was the predominant biological molecule, and controls of gene expression would have had to be based almost entirely on RNA–RNA interactions.

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

We saw in Chapter 6 that the 3′ end of a eucaryotic mRNA molecule is not formed by the termination of RNA synthesis by the RNA polymerase. Instead, it results from an RNA cleavage reaction that is catalyzed by additional factors while the transcript is elongating (see Figure 6–37). A cell can control the site of this cleavage so as to change the C-terminus of the resultant protein.

A well-studied example is the switch from the synthesis of membrane-bound to secreted antibody molecules that occurs during the development of B lymphocytes (see Figure 25–17). Early in the life history of a B lymphocyte, the antibody it produces is anchored in the plasma membrane, where it serves as a receptor for antigen. Antigen stimulation causes B lymphocytes to multiply and to begin secreting their antibody. The secreted form of the antibody is identical to the membrane-bound form except at the extreme C-terminus. In this part of the protein, the membrane-bound form has a long string of hydrophobic amino acids that traverses the lipid bilayer of the membrane, whereas the secreted form has a much shorter string of hydrophilic amino acids. The switch from membrane-bound to secreted antibody therefore requires a different nucleotide sequence at the 3′ end of the mRNA; this difference is generated through a change in the length of the primary RNA transcript, caused by a change in the site of RNA cleavage, as shown in **Figure 7–99**. This change is caused by an increase in the concentration of a subunit of CstF, the protein that binds to the G/U-rich sequences of RNA cleavage and poly-A addition sites and promotes



RNA cleavage (see Figures 6–37 and 6–38). The first cleavage-poly-A addition site that an RNA polymerase transcribing the antibody gene encounters is suboptimal and is usually skipped in unstimulated B lymphocytes, leading to production of the longer RNA transcript. When activated to produce antibodies, the B lymphocyte increases its CstF concentration; as a result, cleavage now occurs at the suboptimal site, and the shorter transcript is produced. In this way, a change in concentration of a general RNA processing factor can have a dramatic effect on the expression of a particular gene.

Figure 7–99 Regulation of the site of RNA cleavage and poly-A addition determines whether an antibody molecule is secreted or remains membrane-bound. In unstimulated B lymphocytes *(left)*, a long RNA transcript is produced, and the intron sequence near its 3′ end is removed by RNA splicing to give rise to an mRNA molecule that codes for a membrane-bound antibody molecule. In contrast, after antigen stimulation *(right)*, the primary RNA transcript is cleaved upstream from the splice site in front of the last exon sequence. As a result, some of the intron sequence that is removed from the long transcript remains as a coding sequence in the short transcript. These are the nucleotide sequences that encode the hydrophilic C-terminal portion of the secreted antibody molecule.

RNA Editing Can Change the Meaning of the RNA Message

The molecular mechanisms used by cells are a continual source of surprises. An example is the process of **RNA editing**, which alters the nucleotide sequences of RNA transcripts once they are synthesized and thereby changes the coded message they carry. The most dramatic form of RNA editing was discovered in RNA transcripts that code for proteins in the mitochondria of trypanosomes. Here, one or more U nucleotides are inserted (or, less frequently, removed) from selected regions of a transcript, altering both the original reading frame and the sequence and thereby changing the meaning of the message. For some genes the editing is so extensive that over half of the nucleotides in the mature mRNA are U nucleotides that were inserted during the editing process. A set of 40- to 80-nucleotide-long RNA molecules that are transcribed separately contains the information that specifies exactly how the initial RNA transcript is to be altered. These so-called *guide RNAs* have a 5′ end that is complementary in sequence to one end of the region of the transcript to be edited, followed by a sequence that specifies the set of nucleotides to be inserted into the transcript (**Figure 7–100**). The editing mechanism is remarkably complex: at each edited position, the RNA is broken, U nucleotides are added to the broken 3′ end, and the RNA is ligated.

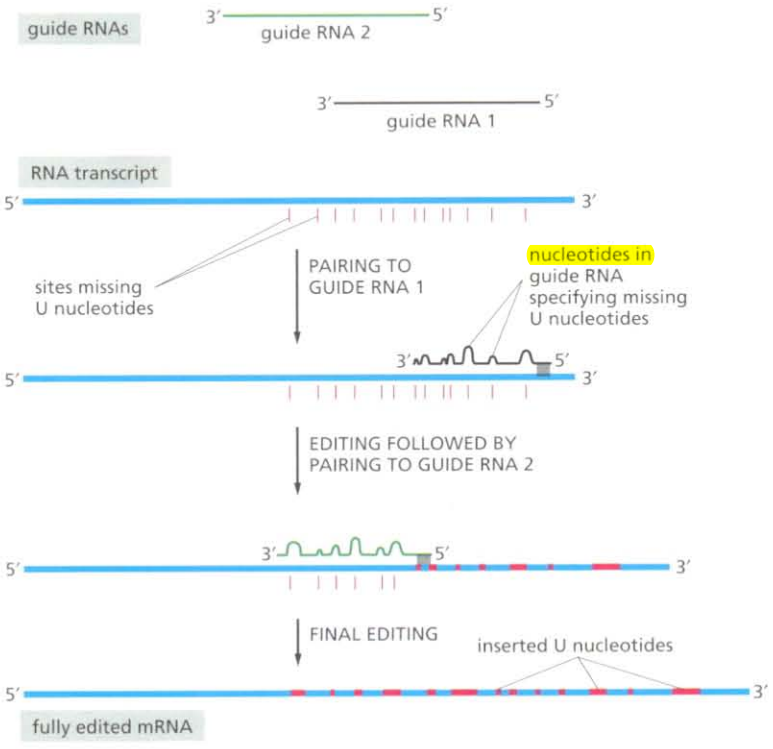


Figure 7–100 RNA editing in the mitochondria of trypanosomes. Editing generally starts near the 3′ end and progresses toward the 5′ end of the RNA transcript, as shown, because the "anchor sequence" at the 5′ end of most guide RNAs can pair only with edited sequences. The U nucleotides are added by a specialized enzyme called a uridylyl transferase.

RNA editing of a more refined type occurs in mammals. Here, two principal types of editing occur, the deamination of adenine to produce inosine (A-to-I editing) and the deamination of cytosine to produce uracil (C-to-U editing; see Figure 5–50). Because these chemical modifications alter the pairing properties of the bases (I pairs with C, and U pairs with A), they can have profound effects on the meaning of the RNA. If the edit occurs in a coding region, it can change the amino acid sequence of the protein or produce a truncated protein. Edits that occur outside coding sequences can affect the pattern of pre-mRNA splicing, the transport of mRNA from the nucleus to the cytosol, or the efficiency with which the RNA is translated.

The process of A-to-I editing is particularly prevalent in humans, where it is estimated to affect over 1000 genes. Enzymes called *ADARs* (adenosine deaminases acting on RNA) perform this type of editing; these enzymes recognize a double-stranded RNA structure that is formed through base pairing between the site to be edited and a complementary sequence located elsewhere on the same RNA molecule, typically in a 3′ intron (**Figure 7–101**). These complementary sequences specify whether the mRNA is to be edited, and if so, exactly where the edit should be made. An especially important example of A-to-I editing takes place in the mRNA that codes for a transmitter-gated ion channel in the brain. A single edit changes a glutamine to an arginine; the affected amino acid lies on the inner wall of the channel, and the editing change alters the Ca^{2+} permeability of the channel. The importance of this edit in mice has been demonstrated by deleting the relevant ADAR gene. The mutant mice are prone to epileptic seizures and die during or shortly after weaning. If the gene for the gated ion channel is mutated to produce the edited form of the protein directly, mice lacking the ADAR develop normally, showing that editing of the ion channel RNA is normally crucial for proper brain development.

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

Why RNA editing exists at all is a mystery. One idea is that it arose in evolution to correct "mistakes" in the genome. Another is that it arose as a somewhat

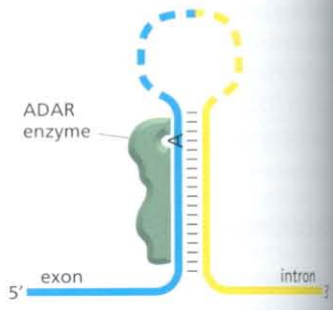


Figure 7–101 Mechanism of A-to-I RNA editing in mammals. RNA sequences carried on the same RNA molecule signal the position of an edit. Typically, a sequence complementary to the position of the edit is present in an intron, and the resulting double-stranded RNA structure attracts the A-to-I editing enzyme ADAR. This type of editing takes place in the nucleus, before the pre-mRNA has been fully processed. Mice and humans have three ADAR enzymes: ADR1 is required in the liver for proper red blood cell development, ADR2 is required for proper brain development (as described in the text), and the role of ADR3 is uncertain.

slapdash way for the cell to produce subtly different proteins from the same gene. A third possibility is that RNA editing originally evolved as a defense mechanism against retroviruses and retrotransposons and was later adapted by the cell to change the meanings of certain mRNAs. Indeed, RNA editing still plays important roles in cell defense. Some retroviruses, including HIV (see Figure 5–71), are extensively edited after they infect cells. This hyperediting creates many harmful mutations in the viral RNA genome and also causes viral mRNAs to be retained in the nucleus, where they are eventually degraded. Although some modern retroviruses protect themselves against this defense mechanism, it presumably helps to hold many viruses in check.

Primates have much higher levels of A-to-I editing than do other mammals, and most of this takes place on RNAs that are transcribed from the highly abundant Alu elements. It has been proposed that A-to-I editing has stopped these mobile elements from completely overtaking our genomes by inactivating the RNA transcripts they require to proliferate (see Figure 5–74). If this idea is correct, RNA editing may have had a profound impact on the shaping of the modern human genome.

23th

RNA Transport from the Nucleus Can Be Regulated

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

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

As we saw in Chapter 5, HIV, once inside the cell, directs the formation of a double-stranded DNA copy of its genome, which is then inserted into the genome of the host (see Figure 5–71). Once inserted, the viral DNA is transcribed as one long RNA molecule by the host cell's RNA polymerase II. This transcript is then spliced in many different ways to produce over 30 different species of mRNA, which in turn are translated into a variety of different proteins (**Figure 7–102**). In order to make progeny virus, entire, unspliced viral transcripts must be exported from the nucleus to the cytosol, where they are packaged into viral capsids and serve as the viral genome (see Figure 5–71). This large transcript, as well as alternatively spliced HIV mRNAs that the virus needs to move to the cytoplasm for protein synthesis, still carries complete introns. The host cell's normal block to the nuclear export of unspliced RNAs therefore presents a special problem for HIV.

The block is overcome in an ingenious way. The virus encodes a protein called Rev) that binds to a specific RNA sequence (called the Rev responsive element, RRE) located within a viral intron. The Rev protein interacts with a nuclear export receptor (exportin 1), which directs the movement of viral RNAs through nuclear pores into the cytosol despite the presence of intron sequences. We discuss in detail the way in which export receptors function in Chapter 12.

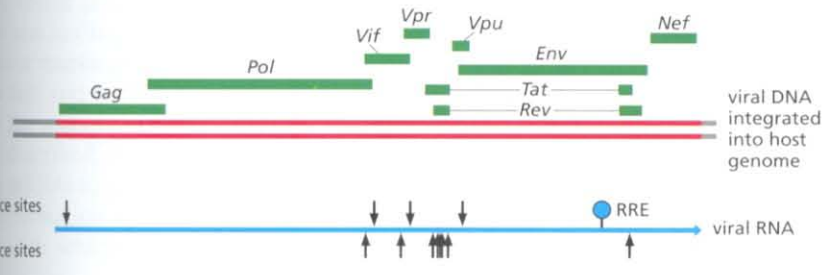


Figure 7–102 The compact genome of HIV, the human AIDS virus. The positions of the nine HIV genes are shown in *green*. The *red double line* indicates a DNA copy of the viral genome that has become integrated into the host DNA *(gray)*. Note that the coding regions of many genes overlap, and that those of *Tat* and *Rev* are split by introns. The *blue line* at the bottom of the figure represents the pre-mRNA transcript of the viral DNA and shows the locations of all the possible splice sites *(arrows)*. There are many alternative ways of splicing the viral transcript; for example the *Env* mRNAs retain the intron that has been spliced out of the *Tat* and *Rev* mRNAs. The Rev response element (RRE) is indicated by a blue ball and stick. It is a 234-nucleotide long stretch of RNA that folds into a defined structure; Rev recognizes a particular hairpin within this larger structure.

The *Gag* gene codes for a protein that is cleaved into several smaller proteins that form the viral capsid. The *Pol* gene codes for a protein that is cleaved to produce reverse transcriptase (which transcribes RNA into DNA), as well as the integrase involved in integrating the viral genome (as double-stranded DNA) into the host genome. Pol is produced by ribosomal frameshifting of translation that begins at *Gag* (see Figure 6–78). The *Env* gene codes for the envelope proteins (see Figure 5–71). Tat, Rev, Vif, Vpr, Vpu, and Nef are small proteins with a variety of functions. For example, Rev regulates nuclear export (see Figure 7–103) and Tat regulates the elongation of transcription across the integrated viral genome (see p. 478).

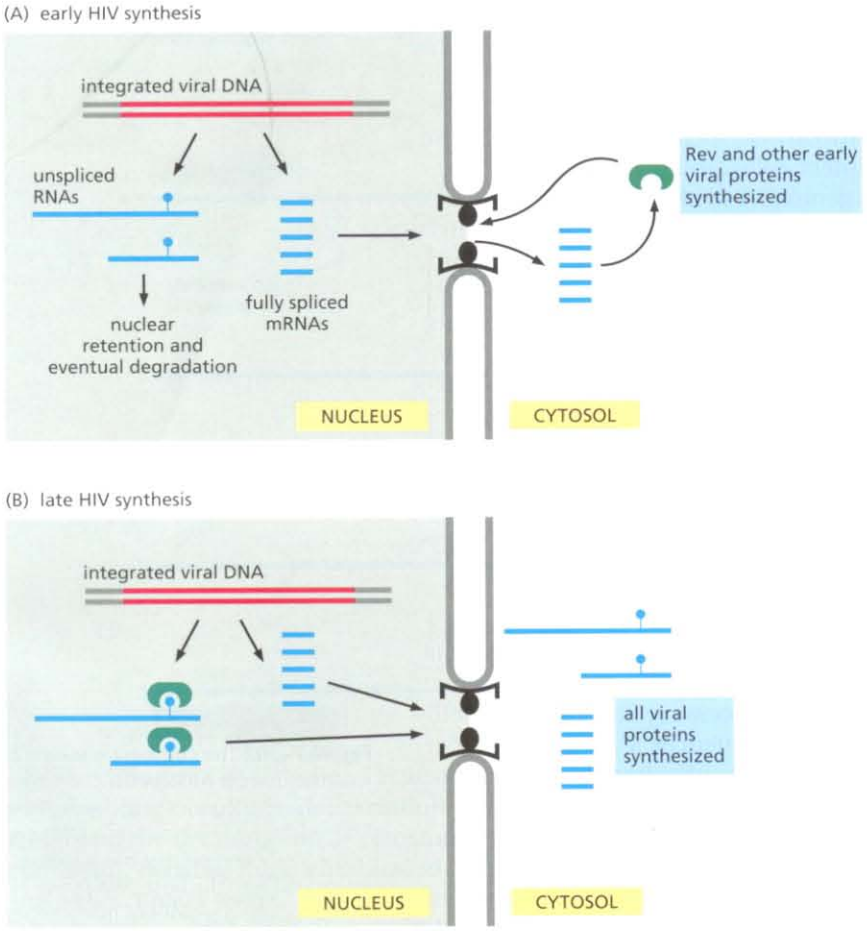


Figure 7–103 Regulation of nuclear export by the HIV Rev protein. Early in HIV infection (A), only the fully spliced RNAs (which contain the coding sequences for Rev, Tat, and Nef) are exported from the nucleus and translated. Once sufficient Rev protein has accumulated and been transported into the nucleus (B), unspliced viral RNAs can be exported from the nucleus. Many of these RNAs are translated into protein and the full-length transcripts are packaged into new viral particles.

The regulation of nuclear export by Rev has several important consequences for HIV growth and pathogenesis. In addition to ensuring the nuclear export of specific unspliced RNAs, it divides the viral infection into an early phase (in which Rev is translated from a fully spliced RNA and all of intron-containing viral RNAs are retained in the nucleus and degraded) and a late phase (in which unspliced RNAs are exported due to Rev function). This timing helps the virus replicate by providing the gene products in roughly the order in which they are needed (**Figure 7–103**). Regulation by Rev may also help the HIV virus to achieve latency, a condition in which the HIV genome has become integrated into the host cell genome but the production of viral proteins has temporarily ceased. If, after its initial entry into a host cell, conditions became unfavorable for viral transcription and replication, Rev is made at levels too low to promote export of unspliced RNA. This situation stalls the viral growth cycle until conditions improve, Rev levels increase, and the virus enters the replication cycle.

Some mRNAs Are Localized to Specific Regions of the Cytoplasm

Once a newly made eucaryotic mRNA molecule has passed through a nuclear pore and entered the cytosol, it is typically met by ribosomes, which translate it into a polypeptide chain (see Figure 6–40). Once the first round of translation "passes" the nonsense-mediated decay test (see Figure 6–80), the mRNA is usually translated in earnest. If the mRNA encodes a protein that is destined to be secreted or expressed on the cell surface, a signal sequence at the protein's amino terminus will direct it to the endoplasmic reticulum (ER); components of the cell's protein-sorting apparatus recognize the signal sequence as soon as it emerges from the ribosome and direct the entire complex of ribosome, mRNA, and nascent protein to the membrane of the ER, where the remainder of the polypeptide chain is synthesized, as discussed in Chapter 12. In other cases free

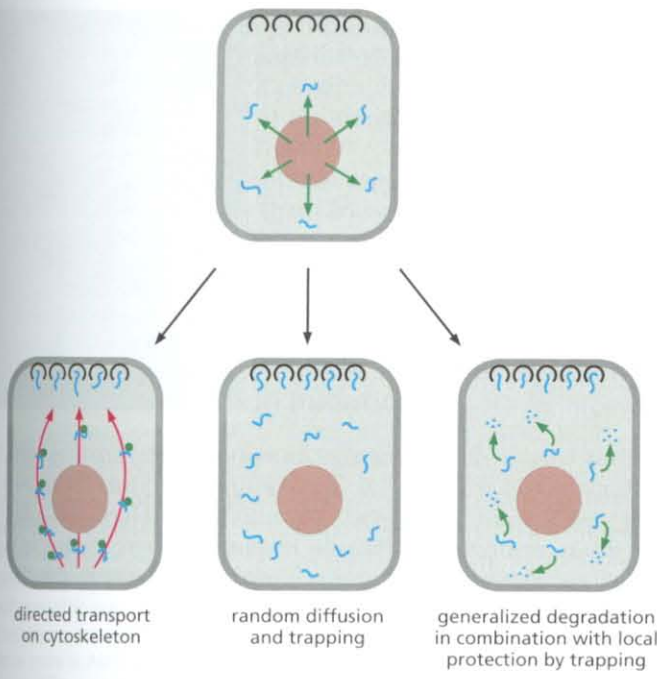


Figure 7–104 Three mechanisms for the localization of mRNAs. The mRNA to be localized leaves the nucleus through nuclear pores *(top)*. Some localized mRNAs *(left diagram)* travel to their destination by associating with cytoskeletal motors. As described in Chapter 16, these motors use the energy of ATP hydrolysis to move unidirectionally along filaments in the cytoskeleton *(red)*. At their destination, anchor proteins *(black)* hold the mRNAs in place. Other mRNAs randomly diffuse through the cytosol and are simply trapped and therefore concentrated at their sites of localization *(center diagram)*. Some of these mRNAs *(right diagram)* are degraded in the cytosol unless they have bound, through random diffusion, a localized protein complex that anchors and protects the mRNA from degradation *(black)*. Each of these mechanisms requires specific signals on the mRNA, which are typically located in the 3′ UTR (see Figure 7–105). In many cases of mRNA localization, additional mechanisms block the translation of the mRNA until it is properly localized. (Adapted from H.D. Lipshitz and C.A. Smibert, *Curr. Opin. Genet. Dev.* 10:476–488, 2000. With permission from Elsevier.)

ribosomes in the cytosol synthesize the entire protein, and signals in the completed polypeptide chain may then direct the protein to other sites in the cell.

Some mRNAs are themselves directed to specific intracellular locations before their efficient translation begins, allowing the cell to position its mRNAs close to the sites where the encoded protein is needed. This strategy provides the cell with many advantages. For example, it allows the establishment of asymmetries in the cytosol of the cell, a key step in many stages of development. Localized mRNA, coupled with translation control, also allows the cell to regulate gene expression independently in its different parts. This feature is particularly important in large, highly polarized cells such as neurons, where growth cones must respond to signals without waiting to involve the distant nucleus. RNA localization has been observed in many organisms, including unicellular fungi, plants, and animals, and it is likely to be a common mechanism that cells use to concentrate high-level production of proteins at specific sites.

Several distinct mechanisms for mRNA localization have been discovered (**Figure 7–104**), but all of them require specific signals in the mRNA itself. These signals are usually concentrated in the 3′ *untranslated region (UTR)*, the region of RNA that extends from the stop codon that terminates protein synthesis to the start of the poly-A tail (**Figure 7–105**). This mRNA localization is usually coupled with translational controls to ensure that the mRNA remains quiescent until it has been moved into place.

The *Drosophila* egg exhibits an especially striking example of mRNA localization. The mRNA encoding the bicoid gene regulatory protein is localized by attachment to the cytoskeleton at the anterior tip of the developing egg. When fertilization triggers the translation of this mRNA, it generates a gradient of the bicoid protein that plays a crucial part in directing the development of the anterior part of the embryo (shown in Figure 7–53 and discussed in more detail in Chapter 22). Many mRNAs in somatic cells are also localized in a similar way. The mRNA that encodes actin, for example, is localized to the actin-filament-rich cell cortex in mammalian fibroblasts by means of a 3′ UTR signal.

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

untranslated regions of eucaryotic mRNAs resemble the transcriptional control regions of genes: their nucleotide sequences contain information specifying the way the RNA is to be used, and proteins interpret this information by binding specifically to these sequences. Thus, over and above the specification of the amino acid sequences of proteins, mRNA molecules are rich with many additional types of information.

The 5′ and 3′ Untranslated Regions of mRNAs Control Their Translation

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

In bacterial mRNAs, a conserved stretch of six nucleotides, the *Shine-Dalgarno sequence,* is always found a few nucleotides upstream of the initiating AUG codon. This sequence forms base pairs with the 16S RNA in the small ribosomal subunit, correctly positioning the initiating AUG codon in the ribosome. Because this interaction strongly contributes to the efficiency of initiation, it provides the bacterial cell with a simple way to regulate protein synthesis through **translational control** mechanisms. These mechanisms, carried out by proteins or by RNA molecules, generally involve either exposing or blocking the Shine–Dalgarno sequence (**Figure 7–106**).

Eucaryotic mRNAs do not contain a Shine–Dalgarno sequence. Instead, as discussed in Chapter 6, the selection of an AUG codon as a translation start site is largely determined by its proximity to the cap at the 5′ end of the mRNA molecule, which is the site at which the small ribosomal subunit binds to the mRNA and begins scanning for an initiating AUG codon. Despite the differences in translation initiation, eucaryotes use similar strategies to regulate translation. For example, translational repressors bind to the 5′ end of the mRNA and thereby inhibit translation initiation. Other repressors recognize nucleotide sequences in the 3′ UTR of specific mRNAs and decrease translation initiation by interfering with the communication between the 5′ cap and 3′ poly-A tail, a step required for efficient translation (see Figure 6–72). A particularly important type of translational control in eucaryotes relies on small RNAs (termed *microRNAs* or *miRNAs*) that bind to mRNAs and reduce protein output. The miRNAs are synthesized and processed in a specialized way, and we shall return to them later in the chapter.

The Phosphorylation of an Initiation Factor Regulates Protein Synthesis Globally

Eucaryotic cells decrease their overall rate of protein synthesis in response to a variety of stressful situations, including deprivation of growth factors or nutrients, infection by viruses, and sudden increases in temperature. Much of this decrease is caused by the phosphorylation of the translation initiation factor eIF2 by specific protein kinases that respond to the changes in conditions.

The normal function of eIF2 was outlined in Chapter 6. It forms a complex with GTP and mediates the binding of the methionyl initiator tRNA to the small ribosomal subunit, which then binds to the 5′ end of the mRNA and begins scanning along the mRNA. When an AUG codon is recognized, the eIF2 protein hydrolyzes the bound GTP to GDP, causing a conformational change in the protein and releasing it from the small ribosomal subunit. The large ribosomal subunit then joins the small one to form a complete ribosome that begins protein synthesis (see Figure 6–71).

Because eIF2 binds very tightly to GDP, a guanine nucleotide exchange factor (see Figure 3–73), designated eIF2B, is required to cause GDP release so that a new GTP molecule can bind and eIF2 can be reused (**Figure 7–107**A). The

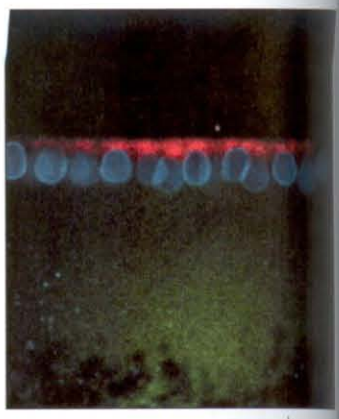


Figure 7–105 An experiment demonstrating the importance of the 3′ UTR in localizing mRNAs to specific regions of the cytoplasm. For this experiment, two different fluorescently labeled RNAs were prepared by transcribing DNA *in vitro* in the presence of fluorescently labeled derivatives of UTP. One RNA (labeled with a *red* fluorochrome) contains the coding region for the *Drosophila* hairy protein and includes the adjacent 3′ UTR (see Figure 6–22). The other RNA (labeled *green*) contains the hairy coding region with the 3′ UTR deleted. The two RNAs were mixed and injected into a *Drosophila* embryo at a stage of development when multiple nuclei reside in a common cytoplasm (see Figure 7–53). When the fluorescent RNAs were visualized 10 minutes later, the full-length hairy RNA *(red)* was localized to the apical side of nuclei (blue), but the transcript missing the 3′ UTR *(green)* failed to localize. Hairy is one of many gene regulatory proteins that specifies positional information in the developing *Drosophila* embryo (discussed in Chapter 22). The localization of its mRNA (shown in this experiment to depend on its 3′ UTR) is thought to be critical for proper fly development. (Courtesy of Simon Bullock and David Ish-Horowicz.)

reuse of eIF2 is inhibited when it is phosphorylated—the phosphorylated eIF2 binds to eIF2B unusually tightly, inactivating eIF2B. There is more eIF2 than eIF2B in cells, and even a fraction of phosphorylated eIF2 can trap nearly all of the eIF2B. This prevents the reuse of the nonphosphorylated eIF2 and greatly slows protein synthesis (Figure 7–107B).

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

24th

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

We saw in Chapter 6 that eucaryotic translation typically begins at the first AUG downstream of the 5′ end of the mRNA, which is the first AUG encountered by a scanning small ribosomal subunit. But the nucleotides immediately surrounding the AUG also influence the efficiency of translation initiation. If the recognition site is poor enough, scanning ribosomal subunits will sometimes ignore the first

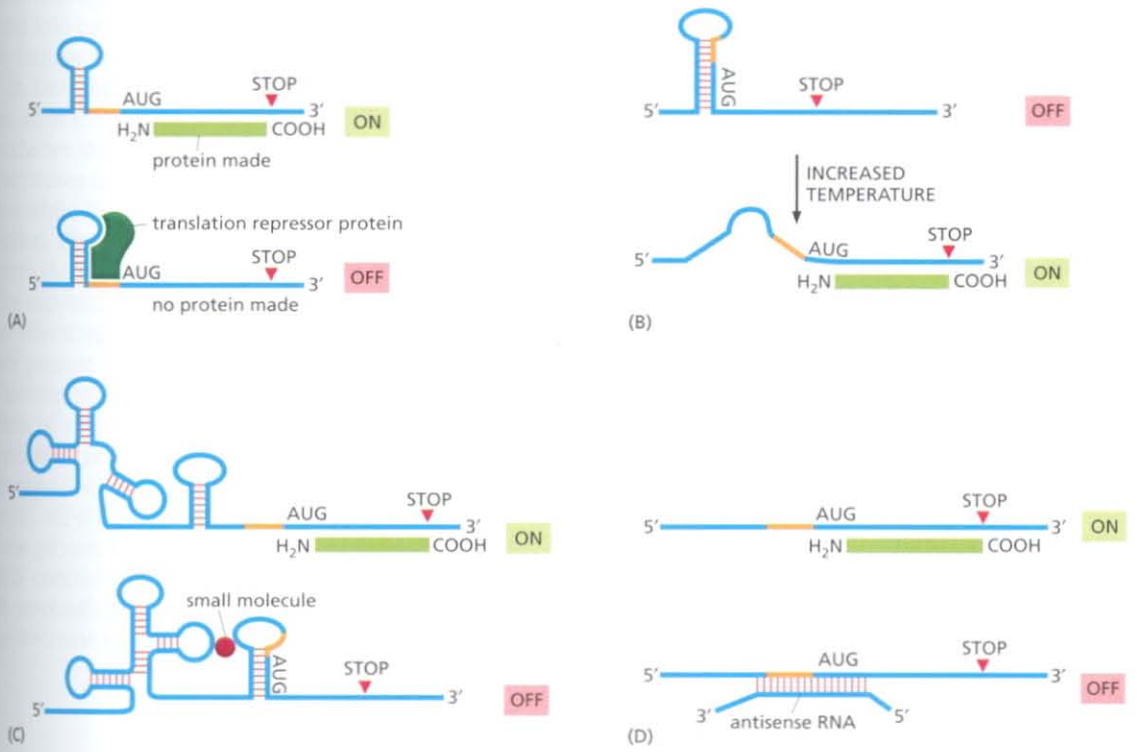


Figure 7–106 Mechanisms of translational control. Although these examples are from bacteria, many of the same principles operate in eucaryotes. (A) Sequence-specific RNA binding proteins repress translation of specific mRNAs by blocking access of the ribosome to the Shine–Dalgarno sequence (*orange*). For example, some ribosomal proteins repress translation of their own RNA. This mechanism comes into play only when the ribosomal proteins are produced in excess over ribosomal RNA and are therefore not incorporated into ribosomes, and it allows the cell to maintain correctly balanced quantities of the various components needed to form ribosomes. In these cases, the regulatory RNA sequence present on the mRNA often matches the RNA sequence that the protein recognizes during ribosome assembly. (B) An RNA "thermosensor" permits efficient translation initiation only at elevated temperatures in which the stem-loop structure has been melted. An example occurs in the human pathogen *Listeria monocytogenes,* in which the translation of its virulence genes increases at 37°C, the temperature of the host. (C) Binding of a small molecule to a riboswitch causes a structural rearrangement of the RNA, sequestering the Shine–Dalgarno sequence *(orange)* and blocking translation initiation. In many bacteria, *S*-adenosyl methionine acts in this manner to block production of the enzymes that synthesize it. (D) An "antisense" RNA produced elsewhere from the genome base-pairs with a specific mRNA, and blocks its translation. Many bacteria regulate expression of iron-storage proteins in this way. When iron is abundant, an antisense transcript is down-regulated, thereby allowing efficient translation of genes encoding the storage proteins. Antisense RNAs are used extensively by eucaryotic cells to regulate gene expression. The mechanism is somewhat different from that shown here and is discussed in detail later in this chapter.

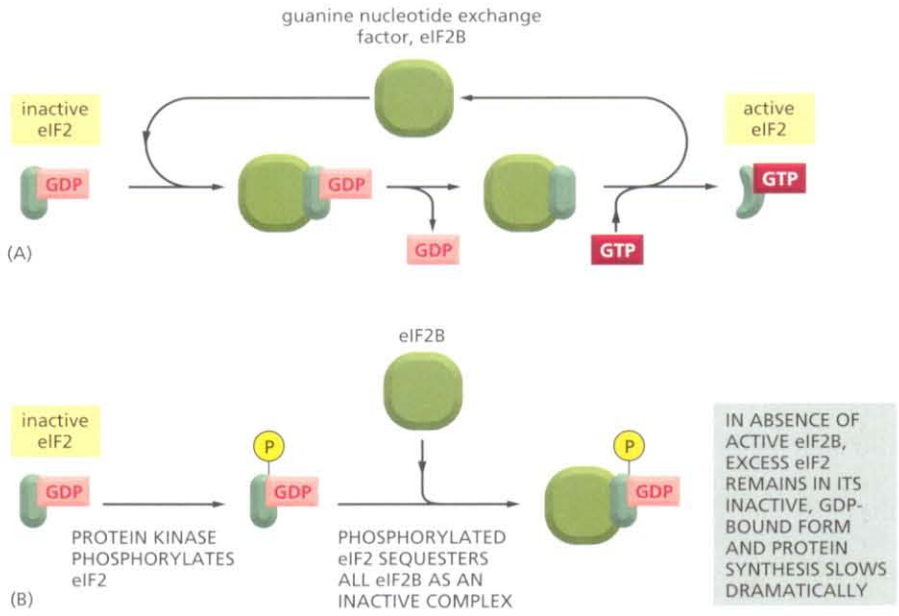


Figure 7–107 The eIF2 cycle. (A) The recycling of used eIF2 by a guanine nucleotide exchange factor (eIF2B). (B) eIF2 phosphorylation controls protein synthesis rates by tying up eIF2B.

AUG codon in the mRNA and skip to the second or third AUG codon instead. This phenomenon, known as "leaky scanning," is a strategy frequently used to produce two or more closely related proteins, differing only in their amino termini, from the same mRNA. Very importantly it allows some genes to produce the same protein with and without a signal sequence attached at its amino terminus so that the protein is directed to two different locations in the cell (for example, to both mitochondria and the cytosol). In some cases, the cell can regulate the relative abundance of the protein isoforms produced by leaky scanning; for example, a cell-type-specific increase in the abundance of the initiation factor eIF4F favors the use of the AUG closest to the 5′ end of the mRNA.

Another type of control found in eucaryotes uses one or more short open reading frames (nucleotide sequences free from stop codons) that lie between the 5′ end of the mRNA and the beginning of the gene. Open reading frames (ORFs) will be discussed more fully in Chapter 8: for present purposes an ORF can be considered a stretch of DNA that begins with a start codon (ATG) and ends with a stop codon, with no stop codons in between, and thus could in principle encode a polypeptide. Often, the amino acid sequences coded by these upstream open reading frames (uORFs) are not important; rather the uORFs serve a purely regulatory function. An uORF present on an mRNA molecule will generally decrease translation of the downstream gene by trapping a scanning ribosome initiation complex and causing the ribosome to translate the uORF and dissociate from the mRNA before it reaches the protein-coding sequences.

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

The *Gcn4* mRNA contains four short uORFs, and these are responsible for selectively increasing the translation of *Gcn4* in response to the eIF2 phosphorylation provoked by amino acid starvation. The mechanism by which *Gcn4* translation is increased is complex. In outline, the small subunit of the ribosome moves along the mRNA, encountering each of the uORFs but directing translation of only a subset of them; if the last uORF is translated, as is the case in normal unstarved cells, the ribosomes dissociate at the end of the uORF, and translation of *Gcn4* is inefficient. The global decrease in eIF2 activity brought about

by nutrient starvation (see Figure 7–108) makes it more likely that a scanning small ribosomal subunit will move through the fourth uORF before it acquires a molecule of eIF2 (see Figure 6–72). Such a ribosomal subunit is free to initiate translation on the actual *Gcn4* sequences, and the increased amount of this gene regulatory protein that results leads to the production of a set of proteins that increase amino acid synthesis inside the cell.

Internal Ribosome Entry Sites Provide Opportunities for Translation Control

Although approximately 90% of eucaryotic mRNAs are translated beginning with the first AUG downstream from the 5′ cap, certain AUGs, as we saw in the previous section, can be skipped over during the scanning process. In this section, we discuss yet another way that cells can initiate translation at positions distant from the 5′ end of the mRNA, using a specialized type of RNA sequence called an **internal ribosome entry site** (**IRES**). An IRES can occur in many different places in an mRNA and, in some unusual cases, two distinct protein-coding sequences are carried in tandem on the same eucaryotic mRNA; translation of the first occurs by the usual scanning mechanism, and translation of the second occurs through an IRES. IRESs are typically several hundred nucleotides in length and fold into specific structures that bind many, but not all, of the same proteins that are used to initiate normal 5′ cap-dependent translation (**Figure 7–108**). In fact, different IRESs require different subsets of initiation factors. However, all of them bypass the need for a 5′ cap structure and the translation initiation factor that recognizes it, eIF4E.

Some viruses use IRESs as part of a strategy to get their own mRNA molecules translated while blocking normal 5′-cap-dependent translation of host mRNAs. On infection, these viruses produce a protease (encoded in the viral genome) that cleaves the host cell translation factor eIF4G, rendering it unable to bind to eIF4E, the cap-binding complex. This shuts down most of the host cell's translation and effectively diverts the translation machinery to the IRES sequences, which are present on many viral mRNAs. The truncated eIF4G remains competent to initiate translation at these internal sites and may even stimulate the translation of certain IRES-containing viral mRNAs. A selective activation of IRES-mediated translation can also occur on host cell mRNAs. For example, when mammalian cells enter the programmed cell death pathway (discussed in Chapter 18), eIF4G is cleaved, and a general decrease in translation ensues. However, some proteins critical for the control of cell death seem to be translated from IRES-containing mRNAs, allowing their continued synthesis. In this way, the IRES mechanism allows translation of selected mRNAs at a high rate despite a general decrease in the cell's overall capacity to initiate protein synthesis.

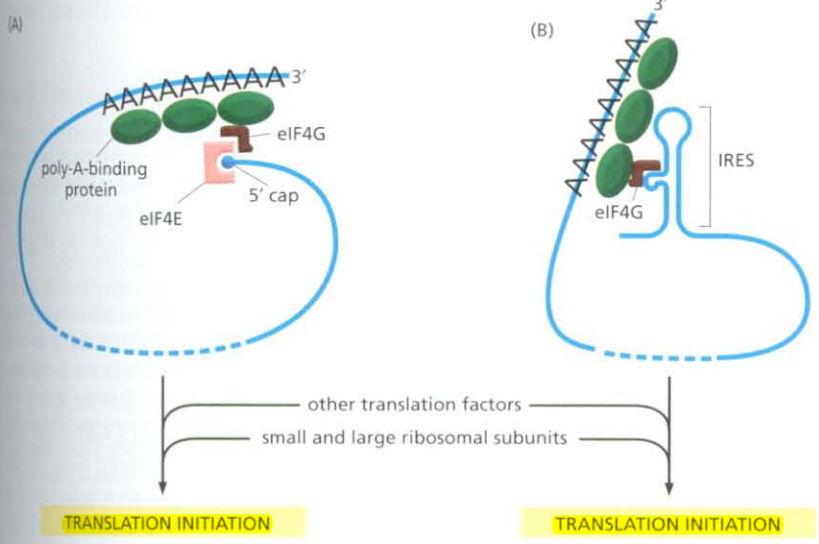


Figure 7–108 Two mechanisms of translation initiation. (A) The normal, cap-dependent mechanism requires a set of initiation factors whose assembly on the mRNA is stimulated by the presence of a 5′ cap and a poly-A tail (see also Figure 6–72). (B) The IRES-dependent mechanism seen mainly in viruses, requires only a subset of the normal translation initiating factors, and these assemble directly on the folded IRES. (Adapted from A. Sachs, *Cell* 101:243–245, 2000. With permission from Elsevier.)

Changes in mRNA Stability Can Regulate Gene Expression

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

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

Two general mechanisms exist for destroying eucaryotic mRNAs. Both begin with the gradual shortening of the poly-A tail by an exonuclease, a process that starts as soon as the mRNA reaches the cytoplasm. In a broad sense, this poly-A shortening acts as a timer that counts down the lifetime of each mRNA. Once the poly-A tail is reduced to a critical length (about 25 nucleotides in humans), the two pathways diverge. In one, the 5′ cap is removed (a process called decapping) and the "exposed" mRNA is rapidly degraded from its 5′ end. In the other, the mRNA continues to be degraded from the 3′ end, through the poly-A tail into the coding sequences (**Figure 7–109**). Most eucaryotic mRNAs are degraded by both mechanisms.

Nearly all mRNAs are subject to these two types of decay, and the specific sequences of each mRNA determine how fast each step occurs and therefore how long each mRNA will persist in the cell and be able to produce protein. The 3′ UTR sequences are especially important in controlling mRNA lifetimes, and they often carry binding sites for specific proteins that increase or decrease the rate of poly-A shortening, decapping, or 3′-to-5′ degradation. The half-life of an mRNA is also affected by how efficiently it is translated. Poly-A shortening and decapping compete directly with the machinery that translates the mRNA; therefore any factors that affect the translation efficiency of an mRNA will tend to have the opposite effect on its degradation (**Figure 7–110**).

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

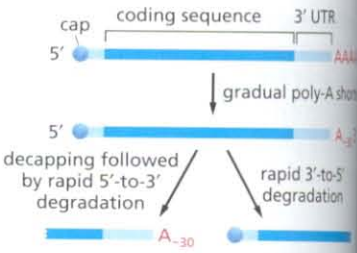


Figure 7–109 Two mechanisms of eucaryotic mRNA decay. A critical threshold of poly-A tail length induces 3′-to-5′ degradation, which may be triggered by the loss of the poly-A binding proteins (see Figure 6–40). As shown in Figure 7–110, the deadenylase associates with both the 3′ poly-A tail and the 5′ cap, and this arrangement may signal decapping after poly-A shortening. Although 5′ to 3′ and 3′ to 5′ degradation are shown here on separate RNA molecules, these two processes can occur together on the same molecule. (Adapted from C.A. Beelman and R. Parker, *Cell* 81:179–183, 1995. With permission from Elsevier.)

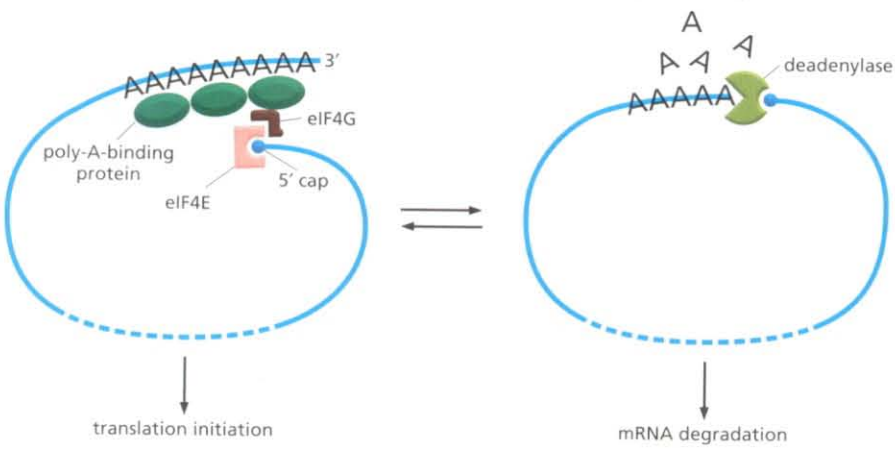


Figure 7–110 The competition between mRNA translation and mRNA decay. The same two features of an mRNA molecule, its 5′ cap and the 3′ poly-A tail, are used in both translation initiation and deadenylation-dependent mRNA decay (see Figure 7–109). The deadenylase that shortens the poly-A tail in the 3′-to-5′ direction associates with the 5′ cap as described in Chapter 6 (see Figure 6–[illegible]); the translation initiation machinery associates with both the 5′ cap and the poly-A tail. (Adapted from M. Gao et al., *Mol. Cell* 5:479–488, 2000. With permission from Elsevier.)

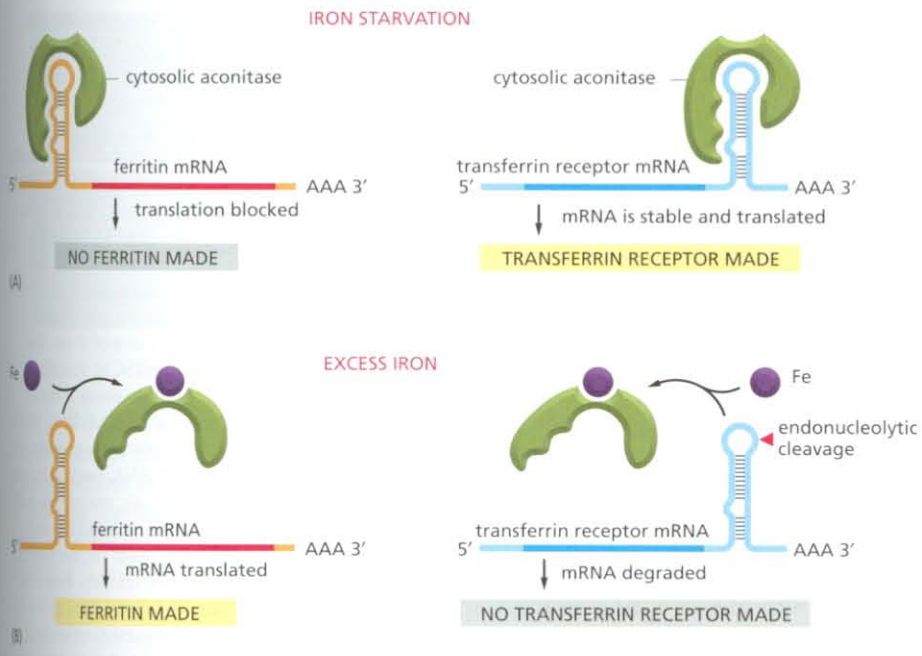


Figure 7–111 Two post-translational controls mediated by iron. (A) During iron starvation, the binding of aconitase to the 5′ UTR of the ferritin mRNA blocks translation initiation; its binding to the 3′ UTR of the transferrin receptor mRNA blocks an endonuclease cleavage site and thereby stabilizes the mRNA. (B) In response to an increase in iron concentration in the cytosol, a cell increases its synthesis of ferritin in order to bind the extra iron and decreases its synthesis of transferrin receptors in order to import less iron across the plasma membrane. Both responses are mediated by the same iron-responsive regulatory protein, aconitase, which recognizes common features in a stem-and-loop structure in the mRNAs encoding ferritin and transferrin receptor. Aconitase dissociates from the mRNA when it binds iron. But because the transferrin receptor and ferritin are regulated by different types of mechanisms, their levels respond oppositely to iron concentrations even though they are regulated by the same iron-responsive regulatory protein. (Adapted from M.W. Hentze et al., *Science* 238:1570–1573, 1987 and J.L. Casey et al., *Science* 240:924–928, 1988. With permission from AAAS.)

blocking endonucleolytic cleavage of the mRNA. On the addition of iron, aconitase is released from the mRNA, exposing the cleavage site and thereby decreasing the stability of the mRNA (**Figure 7–111**).

Cytoplasmic Poly-A Addition Can Regulate Translation

The initial polyadenylation of an RNA molecule (discussed in Chapter 6) occurs in the nucleus, apparently automatically for nearly all eucaryotic mRNA precursors. As we have just seen, the poly-A tails on most mRNAs gradually shorten in the cytosol, and the RNAs are eventually degraded. In some cases, however, the poly-A tails of specific mRNAs are lengthened in the cytosol, and this mechanism provides an additional form of translation regulation.

Maturing oocytes and eggs provide the most striking example. Many of the normal mRNA degradation pathways seem to be disabled in these giant cells, so that the cells can build up large stores of mRNAs in preparation for fertilization. Many mRNAs are stored in the cytoplasm with only 10 to 30 As at their 3′ end, and in this form they are not translated. At specific times during oocyte maturation and just after fertilization, when the cell requires the proteins encoded by these mRNAs, poly-A is added to selected mRNAs by a cytosolic poly-A polymerase, greatly stimulating their translation.

Small Noncoding RNA Transcripts Regulate Many Animal and Plant Genes

In the previous chapter, we introduced the central dogma, according to which the flow of genetic information proceeds from DNA through RNA to protein (Figure 6–2). But we have seen that RNA molecules perform many critical tasks in the cell besides serving as intermediate carriers of genetic information. A series of recent, striking discoveries has revealed that noncoding RNAs are far more prevalent than previously imagined and play previously unanticipated, but widespread, roles in regulating gene expression.

Of special importance to animals and plants is a type of short noncoding RNA called **microRNA** (miRNA). Humans, for example, express more than 400 different miRNAs, and these appear to regulate at least one-third of all human protein-coding genes. Once made, miRNAs base-pair with specific mRNAs and regulate their stability and their translation. The miRNA precursors are synthesized by RNA

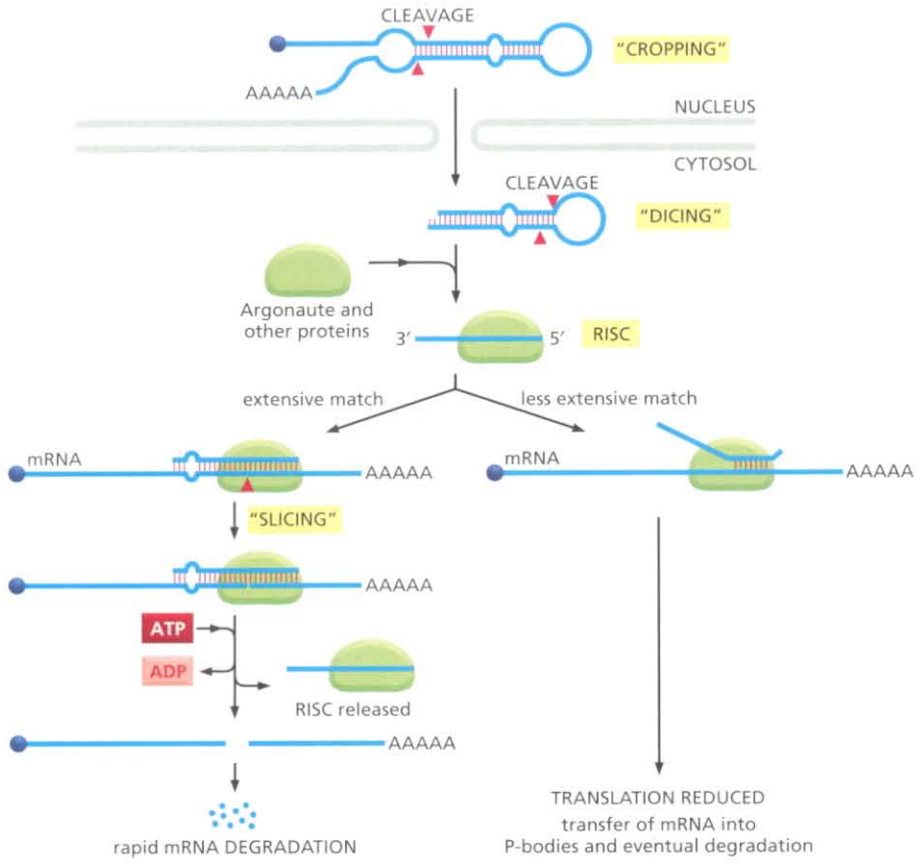


Figure 7–112 miRNA processing and mechanism of action. The precursor miRNA, through complementarity between one part of its sequence and another, forms a double-stranded structure. This is cropped while still in the nucleus, and then exported to the cytosol where it is further cleaved by the Dicer enzyme to form the miRNA proper. Argonaute, in conjunction with other components of RISC, initially associates with both strands of the miRNA and cleaves and discards one of them. The other strand guides RISC to specific mRNAs through base pairing. If the RNA:RNA match is extensive, as is commonly seen in plants, Argonaute cleaves the target mRNA, causing its rapid degradation. In animals, the miRNA-mRNA match often does not extend beyond a short 7-nucleotide "seed" region near the 5′ end of the miRNA. This less extensive base pairing leads to inhibition of translation, mRNA destabilization, and transfer of the mRNA to P-bodies, where it is eventually degraded.

polymerase II and are capped and polyadenylated. They then undergo a special type of processing, after which the miRNA is assembled with a set of proteins to form an *RNA-induced silencing complex* or *RISC*. Once formed, the RISC seeks out its target mRNAs by searching for complementary nucleotide sequences (**Figure 7–112**). This search is greatly facilitated by the Argonaute protein, a component of RISC, which displays the 5′ region of the miRNA so that it is optimally positioned for base-pairing to another RNA molecule (**Figure 7–113**). In animals, the extent of base-pairing is typically seven nucleotide pairs, and it usually takes place in the 3′ UTR of the target mRNA.

Once an mRNA has been bound by an miRNA, several outcomes are possible. If the base-pairing is extensive, the mRNA is cleaved by the Argonaute protein, effectively removing its poly-A tail and exposing it to exonucleases (see Figure 7–109). Following cleavage of the mRNA, RISC (with its associated miRNA) is released, and it can seek out additional mRNAs. Thus, a single miRNA can act catalytically to destroy many complementary mRNAs. The miRNAs can be thought of as guide sequences that bring destructive nucleases into contact with specific mRNAs.

If the base-pairing between the miRNA and the mRNA is less extensive, Argonaute does not slice the mRNA; rather, translation of the mRNA is repressed and the mRNA is destabilized. This effect is associated with shortening of the poly-A tail and the movement of the mRNA to cytosolic structures called *processing bodies (P-bodies)*. Here the mRNAs are sequestered from ribosomes and eventually decapped and degraded. P-bodies are dynamic structures composed of large assemblies of mRNAs and RNA-degrading enzymes, and they are believed to be the sites in the cell where the final destruction of most mRNAs, even those not controlled by miRNAs, takes place (**Figure 7–114**).

Several features make miRNAs especially useful regulators of gene expression. First, a single miRNA can regulate a whole set of different mRNAs so long as the mRNAs carry a common sequence in their UTRs. This situation is common in humans, where some miRNAs control hundreds of different mRNAs. Second, regulation by miRNAs can be combinatorial. When the base-pairing between the miRNA and mRNA fails to trigger cleavage, additional miRNAs binding to the

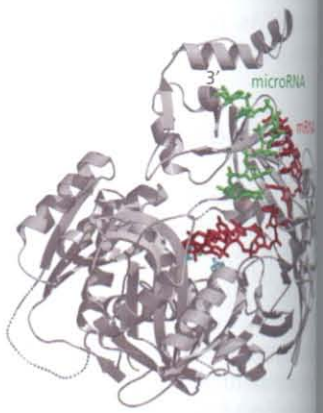


Figure 7–113 Structure of Argonaute protein bound to a perfectly base-paired miRNA and mRNA. (Adapted from N.H. Tolia and L. Joshua-Tor, *Nat. Chem. Biol.* 3:36–43, 2007. With permission from Macmillan Publishers Ltd.)

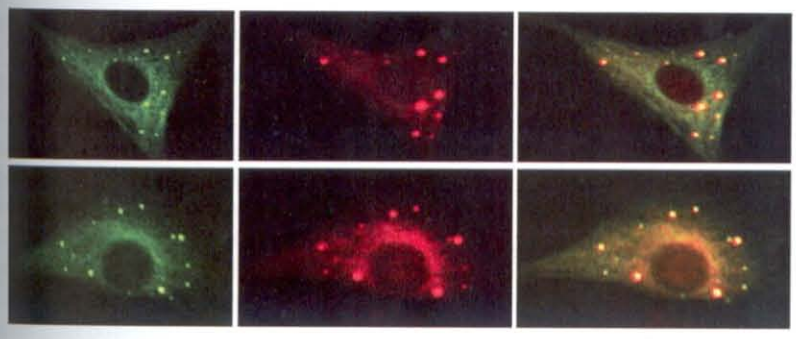


Figure 7–114 Visualization of P-bodies. Human cells were stained with antibodies to a component of the mRNA decapping enzyme Dcp1a *(left panels)* and to the Argonaute protein *(middle panels)*. The merged image *(right panels)* shows that the two proteins co-localize to foci in the cytoplasm called P-bodies. (Adapted from J. Liu et al., *Nat. Cell Biol.* 7:643-644, 2005. With permission from Macmillan Publishers Ltd.)

same mRNA lead to further reductions in its translation. As discussed earlier in this chapter for gene regulatory proteins, combinatorial control greatly expands the possibilities available to the cell by linking gene expression to a combination of different regulators rather than a single regulator. Third, an miRNA occupies relatively little space in the genome when compared with a protein. Indeed, their small size is one reason that miRNAs were discovered only recently. Although we are only beginning to understand the full impact of miRNAs, it is clear that they represent a very important part of the cell's equipment for regulating the expression of its genes.

RNA Interference Is a Cell Defense Mechanism

Many of the proteins that participate in the miRNA regulatory mechanisms just described also serve a second function as a defense mechanism: they orchestrate degradation of foreign RNA molecules, specifically those that occur in double-stranded form. Termed **RNA interference** (**RNAi**), this mechanism is found in a wide variety of organisms, including single-celled fungi, plants, and worms, suggesting that it is evolutionarily ancient. Many transposable elements and viruses produce double-stranded RNA, at least transiently, in their life cycles, and RNAi helps to keep these potentially dangerous invaders in check. As we shall see, RNAi has also provided scientists with a powerful experimental technique to turn off the expression of individual genes.

The presence of double-stranded RNA in the cell triggers RNAi by attracting a protein complex containing *Dicer*, the same nuclease that processes miRNAs (see Figure 7–112). This protein complex cleaves the double-stranded RNA into small (approximately 23- nucleotide-pair) fragments called **small interfering RNAs** (**siRNA**). These double-stranded siRNAs are then bound by Argonaute and other components of the RISC, as we saw above for miRNAs, and one strand of the duplex RNA is cleaved by Argonaute and discarded. The single-stranded siRNA molecule that remains directs RISC back to complementary RNA molecules produced by the virus or transposable element; because the match is exact, Argonaute cleaves these molecules, leading to their rapid destruction (**Figure 7–115**).

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

In some organisms, the RNA interference activity can be spread by the transfer of RNA fragments from cell to cell. This is particularly important in plants (whose cells are linked by fine connecting channels, as discussed in Chapter 19), because it allows an entire plant to become resistant to an RNA virus after only a few of its cells have been infected. In a broad sense, the RNAi

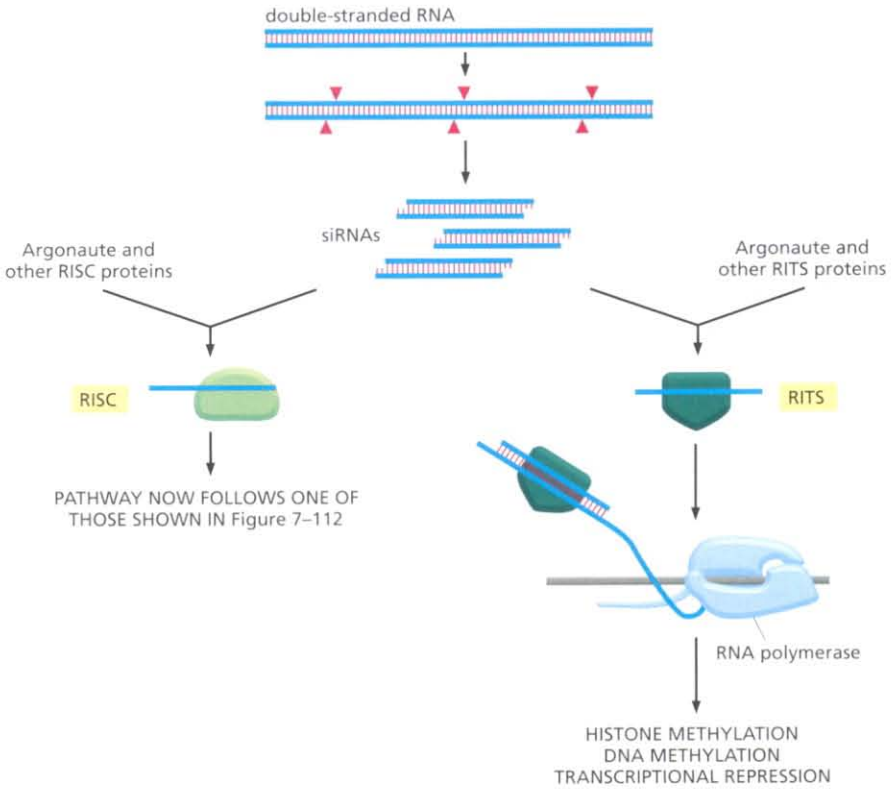


Figure 7–115 siRNA-mediated heterochromatin formation. In many organisms, double-stranded RNA can trigger both the destruction of complementary mRNAs *(left)* and transcriptional silencing *(right)*. The change in chromatin structure induced by the bound RITS (RNA-induced transcriptional silencing) complex resembles that in Figure 7–81.

response resembles certain aspects of animal immune systems; in both, an invading organism elicits a customized response, and—through amplification of the "attack" molecules—the host becomes systemically protected.

RNA Interference Can Direct Heterochromatin Formation

The RNA interference pathway just described does not necessarily stop with the destruction of target RNA molecules. In some cases, the RNA interference machinery can selectively shut off synthesis of the target RNAs. For this remarkable mechanism to occur, the short siRNAs produced by the Dicer protein are assembled with a group of proteins (including Argonaute) to form the RITS (RNA-induced transcriptional silencing) complex. Using single-stranded siRNA as a guide sequence, this complex binds complementary RNA transcripts as they emerge from a transcribing RNA polymerase II (see Figure 7–115). Positioned on the genome in this manner, the RITS complex attracts proteins that covalently modify nearby histones and eventually directs the formation and spread of heterochromatin to prevent further transcription initiation. In some cases, the RITS complex also induces the methylation of DNA, which, as we have seen, can repress gene expression even further. Because heterochromatin and DNA methylation can be self-propagating, an initial RNA interference signal can continue to silence gene expression long after all the siRNA molecules have dissipated.

RNAi-directed heterochromatin formation is an important cell defense mechanism that limits the accumulation of transposable elements in the genome by maintaining them in a transcriptionally silent form. However, this same mechanism is also used in many normal processes in the cell. For example, in many organisms, the RNA interference machinery maintains the heterochromatin formed around centromeres. Centromeric DNA sequences are transcribed in both directions, producing complementary RNA transcripts that can base-pair to form double-stranded RNA. This double-stranded RNA triggers the RNA interference pathway and stimulates heterochromatin formation at centromeres. This heterochromatin, in turn, is necessary for the centromeres to segregate chromosomes accurately during mitosis (see Figure 4–50).

RNA Interference Has Become a Powerful Experimental Tool

Although it probably arose initially as a defense mechanism, RNA interference has become thoroughly integrated into many aspects of normal cell biology, ranging from the control of gene expression to the structure of chromosomes. It has also been developed by scientists into a powerful experimental tool that allows almost any gene to be inactivated by evoking an RNAi response to it. This technique, carried out in cultured cells, and in some cases, whole animals and plants, has revolutionized genetic approaches in cell and molecular biology. We shall discuss it in more detail in the following chapter (see pp. 571–572). RNAi also has great potential in treating human disease. Since many human disorders result from the misexpression of genes, the ability to turn these genes off by experimentally introducing complementary siRNA molecules holds great medical promise. Remarkably, the mechanism of RNA interference was discovered only recently, and we are still being surprised by its mechanistic details and by its broad biological implications.

Summary

Many steps in the pathway from RNA to protein are regulated by cells in order to control gene expression. Most genes are regulated at multiple levels, in addition to being controlled at the initiation stage of transcription. The regulatory mechanisms include (1) attenuation of the RNA transcript by its premature termination, (2) alternative RNA splice-site selection, (3) control of 3′-end formation by cleavage and poly-A addition, (4) RNA editing, (5) control of transport from the nucleus to the cytosol, (6) localization of mRNAs to particular parts of the cell, (7) control of translation initiation, and (8) regulated mRNA degradation. Most of these control processes require the recognition of specific sequences or structures in the RNA molecule being regulated, a task performed by either regulatory proteins or regulatory RNA molecules. A particularly widespread form of post-transcriptional control is RNA interference, where guide RNAs base-pair with mRNAs. RNA interference can cause mRNAs to be either destroyed or translationally repressed. It can also cause specific genes to be packaged into heterochromatin.

25th

PROBLEMS

Which statements are true? Explain why or why not.

7-1 In terms of its biochemical function, the helix-loop-helix motif is more closely related to the leucine zipper motif than it is to the helix-turn-helix motif.

7-2 Reversible genetic rearrangements are a common way of regulating gene expression in procaryotes and mammalian cells.

7-3 CG islands are thought to have arisen during evolution because they were associated with portions of the genome that remained active, hence unmethylated, in the germline.

Discuss the following problems.

7-4 A small portion of a two-dimensional display of proteins from human brain is shown in **Figure Q7–1**. These proteins were separated on the basis of size in one dimension and electrical charge (isoelectric point) in the other. Not all protein spots on such displays are products of different genes; some represent modified forms of a protein that migrate to different positions. Pick out a couple of sets of spots that could represent proteins that differ by the number of phosphates they carry. Explain the basis for your selection.

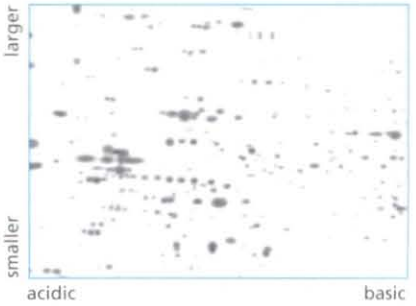


Figure Q7–1 Two-dimensional separation of proteins from the human brain (Problem 7–4). The proteins were displayed using two-dimensional gel electrophoresis. Only a small portion of the protein spectrum is shown. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

7–5 DNA microarray analysis of the patterns of mRNA abundance in different human cell types shows that the level of expression of almost every active gene is different. The patterns of mRNA abundance are so characteristic of cell type that they can be used to determine the tissue of origin of cancer cells, even though the cells may have metastasized to different parts of the body. By definition, however, cancer cells are different from their noncancerous precursor cells. How do you suppose then that patterns of mRNA expression might be used to determine the tissue source of a human cancer?

7–6 The nucleus of a eucaryotic cell is much larger than a bacterium, and it contains much more DNA. As a consequence, a DNA-binding protein in a eucaryotic cell must be able to select its specific binding site from among many more unrelated sequences than does a DNA-binding protein in a bacterium. Does this present any special problems for eucaryotic gene regulation?

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

7–7 DNA-binding proteins often find their specific sites much faster than would be anticipated by simple three-dimensional diffusion. The Lac repressor, for example, associates with the *Lac* operator—its DNA-binding site—more than 100 times faster than expected from this model. Clearly, the repressor must find the operator by mechanisms that reduce the dimensionality or volume of the search in order to hasten target acquisition.

Several techniques have been used to investigate this problem. One of the most elegant used strongly fluorescent RNA polymerase molecules that could be followed individually. An array of DNA molecules was aligned in parallel and anchored to a glass slide. Fluorescent RNA polymerase molecules were then allowed to flow across them at an oblique angle (**Figure Q7–2**A). Traces of individual RNA polymerases showed that about half flowed in the same direction as the bulk and about half deviated from the bulk flow in a characteristic manner (Figure Q7–2B). If the RNA polymerase molecules were first incubated with short DNA fragments containing a strong promoter, all the traces followed the bulk flow.

A. Offer an explanation for why some RNA polymerase molecules deviated from the bulk flow as shown in Figure Q7–2B. Why did incubation with short DNA fragments containing a strong promoter eliminate traces that deviated from the bulk flow?

B. Do these results suggest an explanation for how site-specific DNA-binding molecules manage to find their sites faster than expected by diffusion?

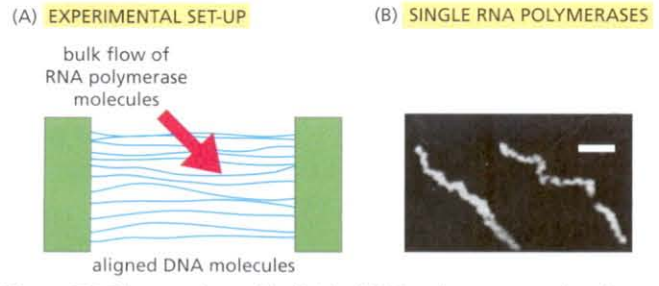


Figure Q7–2 Interactions of individual RNA polymerase molecules with DNA (Problem 7–7). (A) Experimental set-up. DNA molecules are aligned and anchored to a glass slide, and highly fluorescent RNA polymerase molecules are allowed to flow across them. (B) Traces of two individual RNA polymerase molecules. The one on the *left* has traveled with the bulk flow, and the one on the *right* has deviated from it. The scale bar is 10 μm. (B, reprinted from H. Kabata et al., *Science* 262:1561–1563, 1993. With permission from AAAS.

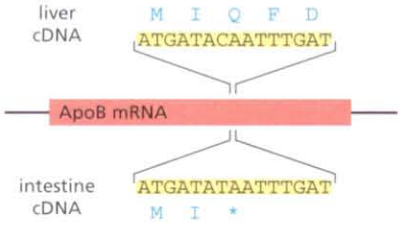


Figure Q7–3 Location of sequence differences in cDNA clones from ApoB RNA isolated from liver and intestine (Problem 7–9). The encoded amino acid sequences, in the one-letter code, are shown aligned with the cDNA sequence

C. Based on your explanation, would you expect a site-specific DNA-binding molecule to find its target site faster in a population of short DNA molecules or in a population of long DNA molecules? Assume that the concentration of target sites is identical and that there is one target site per DNA molecule.

7–8 Most people who are completely blind have circadian rhythms that are 'free-running;' that is, their rhythms are not synchronized to environmental time cues and they oscillate on a cycle of about 24.5 hours. Why do you suppose the circadian clocks of blind people are not entrained to the same 24-hour clock as the majority of the population? Can you guess what symptoms might be associated with a free-running circadian clock? Do you suppose that blind people have trouble sleeping?

7–9 In humans, two closely related forms of apolipoprotein B (ApoB) are found in blood as constituents of the plasma lipoproteins. ApoB48 (molecular mass, 48 kilodaltons) is synthesized by the intestine and is a key component of chylomicrons, the large lipoprotein particles responsible for delivery of dietary triglycerides to adipose tissue for storage. ApoB100 (molecular mass, 100,000 kilodaltons) is synthesized in the liver for formation of the much smaller, very low-density lipoprotein particles used in the distribution of triglycerides to meet energy needs. A classic set of studies defined the surprising relationship between these two proteins.

Sequences of cloned cDNA copies of the mRNAs from these two tissues revealed a single difference: cDNAs from intestinal cells had a T, as part of a stop codon, at a position where the cDNAs from liver cells had a C, as part of a glutamine codon (**Figure Q7–3**). To verify the differences in the mRNAs and to search for corresponding differences in the genome, RNA and DNA were isolated from intestinal and liver cells and then subjected to PCR amplification, using oligonucleotides that flanked the region of interest. The amplified DNA segments from the four samples were tested for the presence of the T or C by hybridization to oligonucleotides containing either the liver cDNA sequence (oligo-Q) or the intestinal cDNA sequence (oligo-STOP). The results are shown in **Table Q7–1**.

Are the two forms of ApoB produced by transcriptional control from two different genes, by a processing control of the RNA transcript from a single gene, or by differential cleavage of the protein product from a single gene? Explain your reasoning.

Table Q7–1 Hybridization of specific oligonucleotides to the amplified segments from liver and intestine RNA and DNA (Problem 7–9).

	RNA		DNA	
	LIVER	INTESTINE	LIVER	INTESTINE
Oligo-Q	+	–	+	+
Oligo-STOP	–	+	–	–

Hybridization is indicated by +; its absence by –.