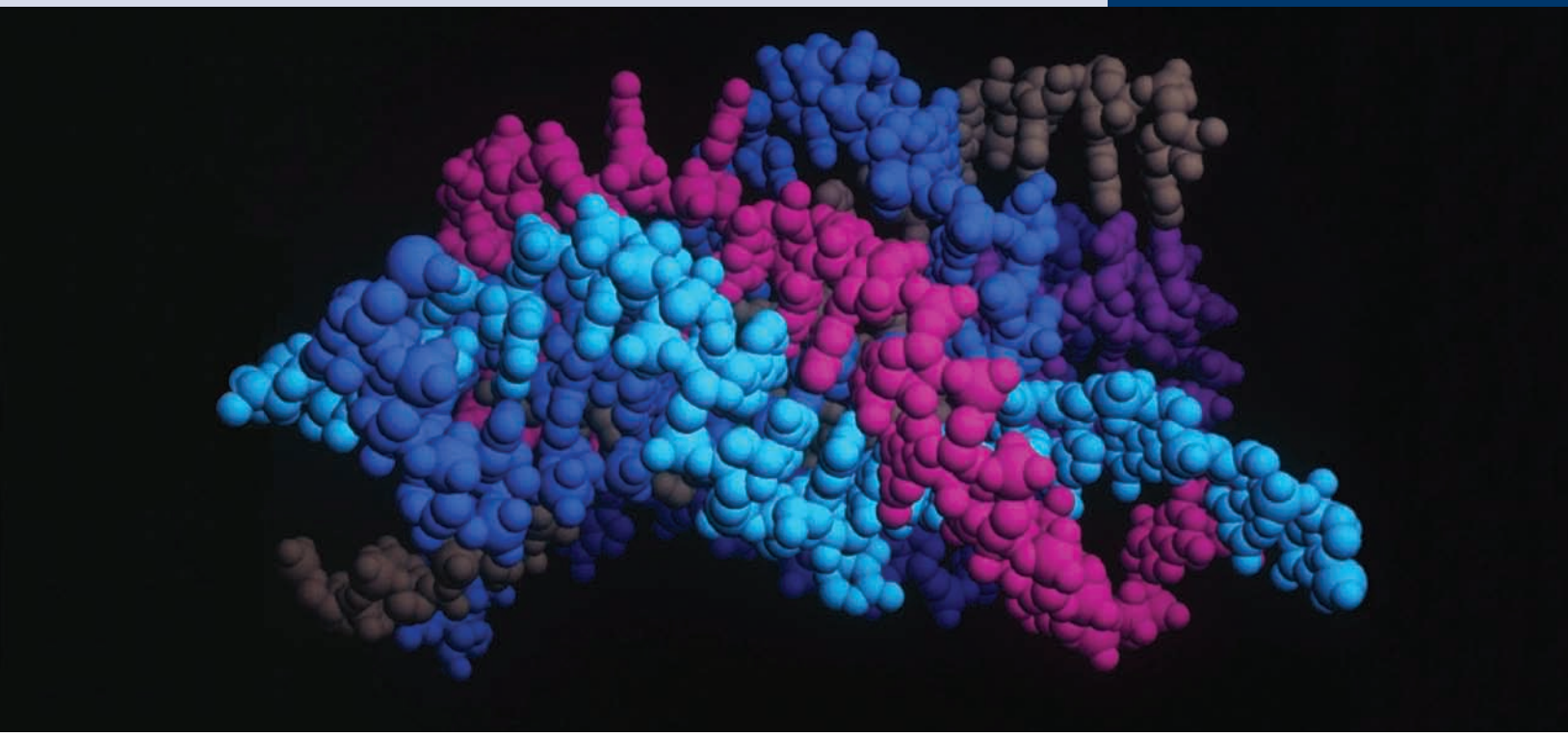
4



A self-splicing intron in an rRNA of the large ribosomal subunit. © Kenneth Eward/Photo Researchers, Inc.

The Interrupted Gene

CHAPTER OUTLINE

4.1 Introduction

The simplest form of a gene is a length of DNA that directly corresponds to its polypeptide product. Bacterial genes are almost always of this type, in which a continuous coding sequence of $3N$ base pairs encodes a polypeptide of N amino acids. In eukaryotes, however, a gene may include additional sequences that lie within the coding region and interrupt the sequence that encodes the polypeptide. These sequences are removed from the RNA product following transcription, generating an mRNA that includes a nucleotide sequence exactly corresponding to the polypeptide product according to the rules of the genetic code.

- **interrupted gene** A gene in which the coding sequence is not continuous due to the presence of introns.
- **exon** Any segment of an interrupted gene that is represented in the mature RNA product.
- **intron** A segment of DNA that is transcribed but later removed from within the transcript by splicing together the sequences (exons) on either side of it.

The sequences of DNA comprising an **interrupted gene** are divided into the two categories depicted in FIGURE 4.1.

- **Exons** are the sequences retained in the mature RNA product. By definition, a gene begins and ends with exons that correspond to the 5′ and 3′ ends of the RNA.
- **Introns** are the intervening sequences that are removed when the primary transcript is processed to give the mature RNA product.

The exon sequences are in the same order in the gene and in the RNA, but an interrupted gene is longer than its final RNA product because of the presence of the introns.

- **RNA splicing** The process of excising introns from RNA and connecting the exons into a continuous mRNA.

The processing of interrupted genes requires an additional step that is not needed for uninterrupted genes. The DNA of an interrupted gene is transcribed to an RNA copy (a transcript) that is exactly complementary to the original gene sequence. This RNA is only a precursor, though; it cannot yet be used to produce a polypeptide. First, the introns must be removed from the RNA to give a messenger RNA that consists only of the series of exons. This process is called **RNA splicing** (see *Section 2.10, Several Processes Are Required to Express the Product of a Gene*) and involves precisely deleting the introns from the primary transcript and then joining the ends of the RNA on either side to form a covalently intact molecule (see *Chapter 21, RNA Splicing and Processing*).

The original gene comprises the region in the genome between the points corresponding to the 5′ and 3′ terminal bases of the mature mRNA. We know that transcription starts at the DNA template corresponding to the 5′ end of the mRNA and usually extends beyond the complement to the 3′ end of the mature mRNA, which is generated by cleavage of the primary RNA transcript (see *Section 21.14, The 3′ Ends of mRNAs Are Generated by Cleavage and Polyadenylation*). The gene is also considered to include the regulatory regions on both sides of the gene that are required for initiating and (sometimes) terminating transcription.

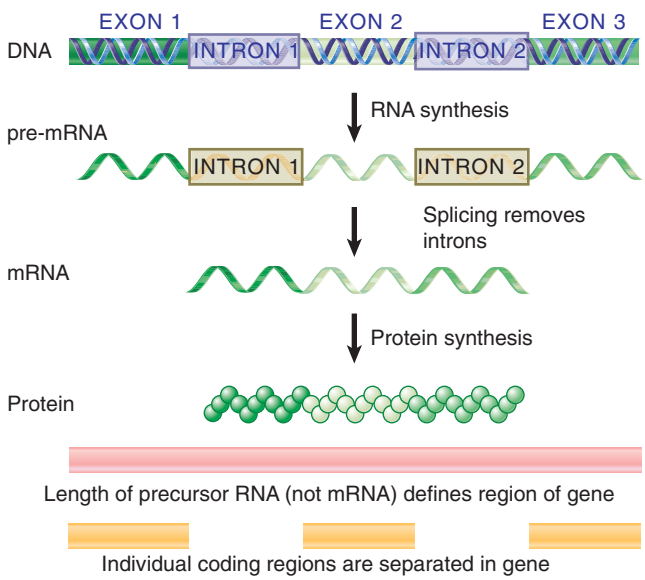


FIGURE 4.1 Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mature mRNA has only the sequences of the exons.

CONCEPT AND REASONING CHECK

Why might it be difficult to express an intact human gene inserted into a bacterial cell?

4.2 An Interrupted Gene Consists of Exons and Introns

How does the existence of introns change our view of the gene? During splicing, the exons are always joined together in the same order they are found in the original DNA, so the correspondence between the gene and polypeptide sequences is maintained. **FIGURE 4.2** shows that the order of exons in the gene remains the same as the order of exons in the processed mRNA, but the distances between sites in the gene (as determined by recombination analysis) do not correspond to the distances between sites in the processed mRNA. The length of the gene is defined by the length of the primary RNA transcript instead of the length of the processed mRNA.

All the exons are present in the primary RNA transcript, and their splicing together occurs only as an intramolecular reaction. There is usually no joining of exons carried by different RNA transcripts, so the splicing mechanism excludes any splicing together of sequences from different alleles. (However, in a phenomenon known as *trans*-splicing, sequences from different mRNAs are ligated together into a single molecule for translation.) Mutations located in different exons of a gene cannot complement one another, so they continue to be defined as members of the same complementation group.

Mutations that directly affect the sequence of a polypeptide must occur in exons. What are the effects of mutations in the introns? The introns are not part of the processed messenger RNA, so mutations in them cannot directly affect the polypeptide sequence. However, they may affect the processing of the messenger RNA by inhibiting the splicing together of exons. A mutation of this sort acts only on the allele that carries it. As a result, it fails to complement any other mutation in that allele and is part of the same complementation group as the exons.

Mutations that affect splicing are usually deleterious. The majority are single-base substitutions at the junctions between introns and exons. They may cause an exon to be left out of the product, cause an intron to be included, or make splicing occur at a different site. The most common outcome is a termination codon that truncates the polypeptide sequence. About 15% of the point mutations that cause human diseases disrupt splicing.

Some eukaryotic genes are not interrupted, and, like prokaryotic genes, correspond directly with the polypeptide product. In yeast, most genes are uninterrupted. In multicellular eukaryotes, most genes are interrupted, and introns are usually much longer than exons, so that genes are considerably larger than their coding regions. (See Figure 4.6 for the example of the mammalian β-globin genes, in which exons can be ~37% of the total length of the gene.)

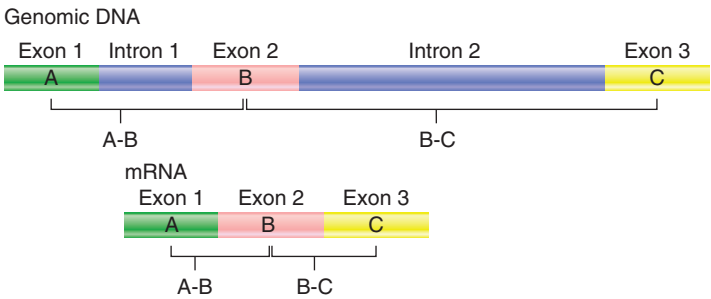


FIGURE 4.2 Exons remain in the same order in mRNA as in DNA, but distances along the gene do not correspond to distances along the mRNA or polypeptide products. The distance A-B in the gene is smaller than the distance B-C; but the distance A-B in the mRNA (and polypeptide) is greater than the distance B-C.

KEY CONCEPTS

- Introns are removed by the process of RNA splicing.
- Only mutations in exons can affect polypeptide sequence; however, mutations in introns can affect processing of the RNA and therefore prevent production of polypeptide.

CONCEPT AND REASONING CHECK

Describe the types of mutations that lead to abnormal splicing and their specific effects.

4.3 Organization of Interrupted Genes May Be Conserved

The characterization of eukaryotic genes was first made possible by the development of techniques for physically mapping DNA. When an mRNA is compared with the DNA sequence from which it was transcribed, the DNA sequence turns out to have extra regions that are not represented in the mRNA.

One technique for comparing mRNA with genomic DNA is to hybridize the mRNA with the complementary strand of the DNA. If the two sequences are colinear, a duplex is formed. **FIGURE 4.3** shows a typical result when an RNA transcribed from an uninterrupted gene is hybridized with a DNA that includes the gene. The sequences on either side of the gene are not represented in the RNA, but the DNA sequence of the gene hybridizes with the RNA to form a continuous duplex region.

Suppose we now perform the same experiment with RNA transcribed from an interrupted gene. The difference is that the sequences represented in the mRNA lie on either side of a sequence that is not in the mRNA. **FIGURE 4.4** shows that the RNA-DNA hybrid forms a duplex, but the unhybridized DNA sequence in the middle remains single stranded, forming a loop that extrudes from the duplex. The hybridizing regions correspond to the exons, and the extruded loop corresponds to the intron.

The structure of the mRNA-DNA hybrid can be visualized by electron microscopy. One of the very first examples of the visualization of an interrupted gene is

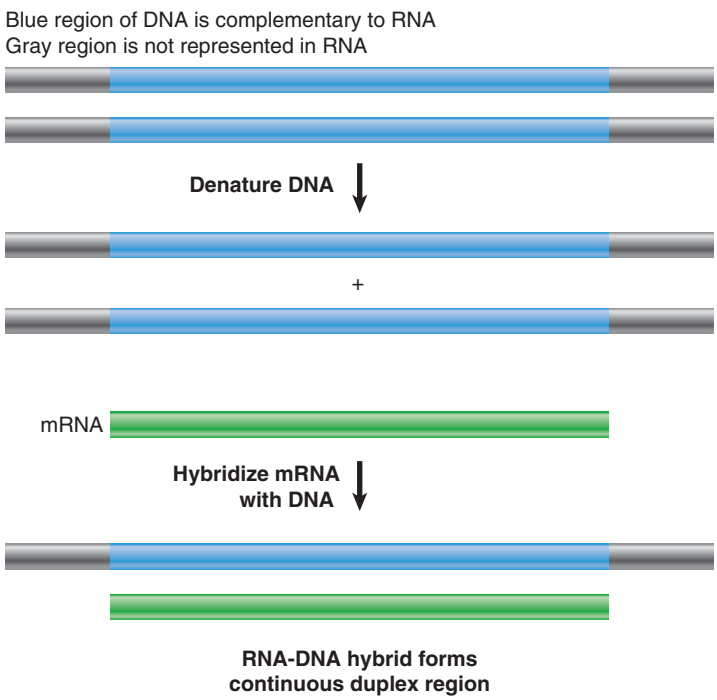


FIGURE 4.3 Hybridizing an mRNA from an uninterrupted gene with the DNA of the gene generates a duplex region corresponding to the gene.

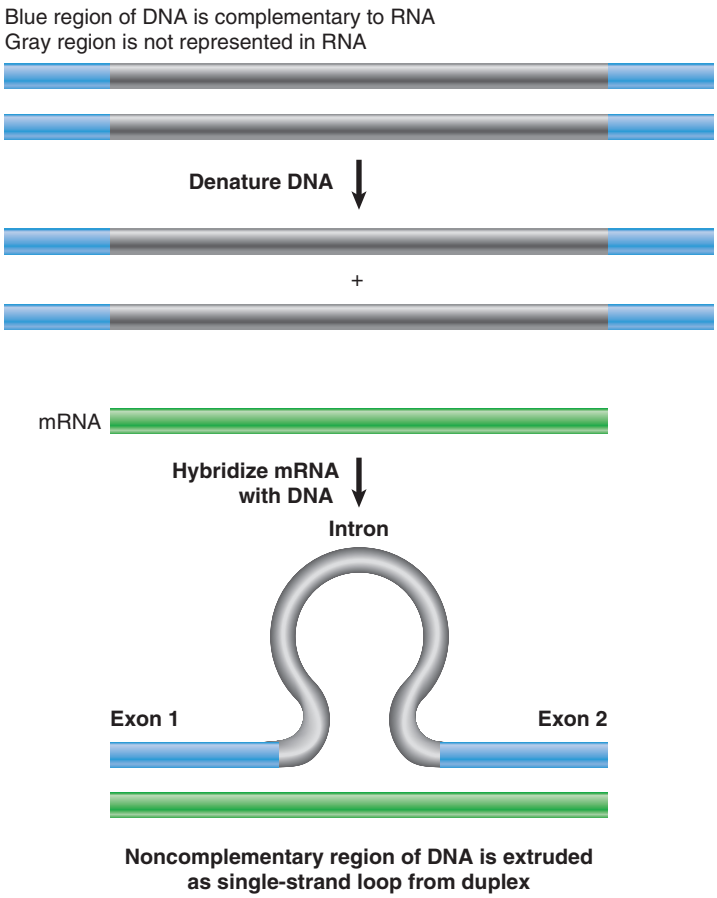


FIGURE 4.4 RNA hybridizing with the DNA from an interrupted gene produces a duplex corresponding to the exons, with an intron excluded as a single-stranded loop between the exons.

shown in **FIGURE 4.5**. On the right side of the figure, tracing the structure shows that three introns are located close to the very beginning of the gene.

When a gene is uninterrupted, the restriction map of its DNA corresponds exactly with the map of its mRNA. When a gene has introns, the maps of the gene and its mRNA are different except at each end, corresponding to the first and last exon. The gene map will have additional regions due to the presence of introns.

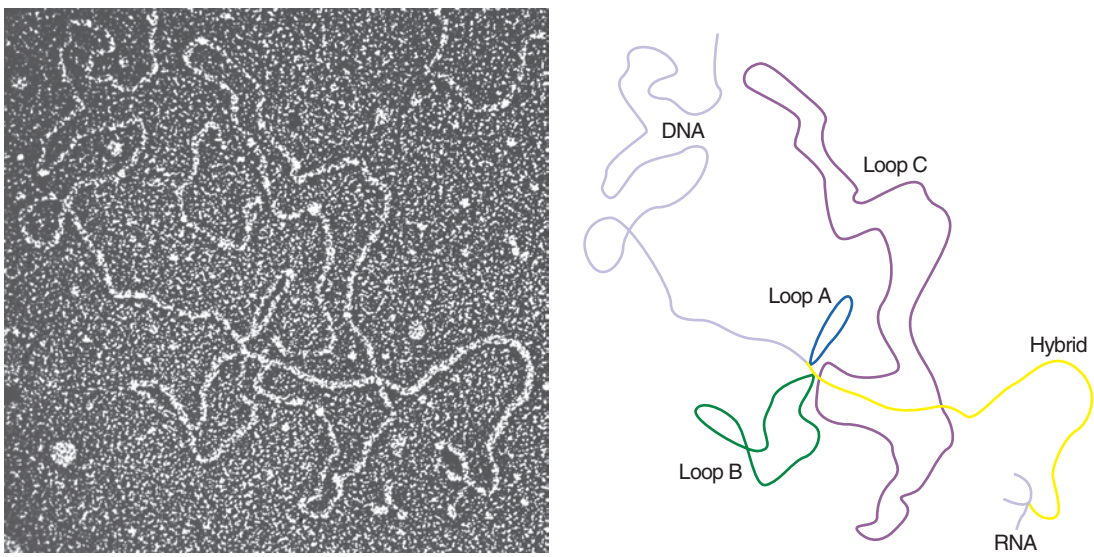


FIGURE 4.5 Hybridization between an adenovirus mRNA and its DNA identifies three loops corresponding to introns that are located at the beginning of the gene. Photo reproduced from S. M. Berget, C. Moore, and P. A. Sharp, *Proc. Natl. Acad. Sci. USA* 74 (1977): 3171–3175. Used with permission of Philip Sharp, Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology.

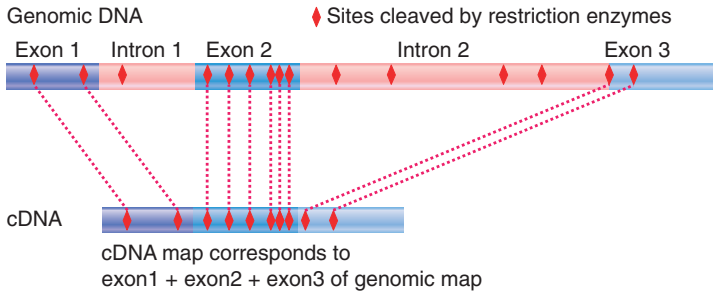


FIGURE 4.6 Comparison of the restriction maps of cDNA and genomic DNA for mouse β globin shows that the gene has two introns that are not present in the cDNA. The exons can be aligned exactly between cDNA and gene.

cDNA A single-stranded DNA complementary to an RNA and synthesized from it by reverse transcription *in vitro*.

FIGURE 4.6 compares the restriction maps of a β-globin gene and a complementary DNA (**cDNA**) copy of its mRNA. The gene has two introns, each of which contains a series of restriction sites that are absent from the cDNA. The pattern of restriction sites in the exons is the same in both the cDNA and the gene.

Ultimately, a comparison of the nucleotide sequences of the gene and mRNA sequences precisely identifies the introns. As indicated in **FIGURE 4.7**, an intron usually has no open reading frame. An intact reading frame in the mRNA results from the removal of the introns.

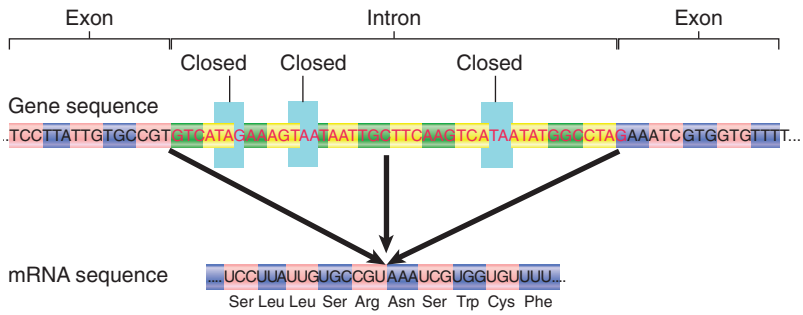


FIGURE 4.7 An intron is a sequence present in the gene but absent from the mRNA (here shown in terms of the cDNA sequence). The reading frame is indicated by the alternating open and shaded blocks; note that all three possible reading frames are closed by termination codons in the intron.

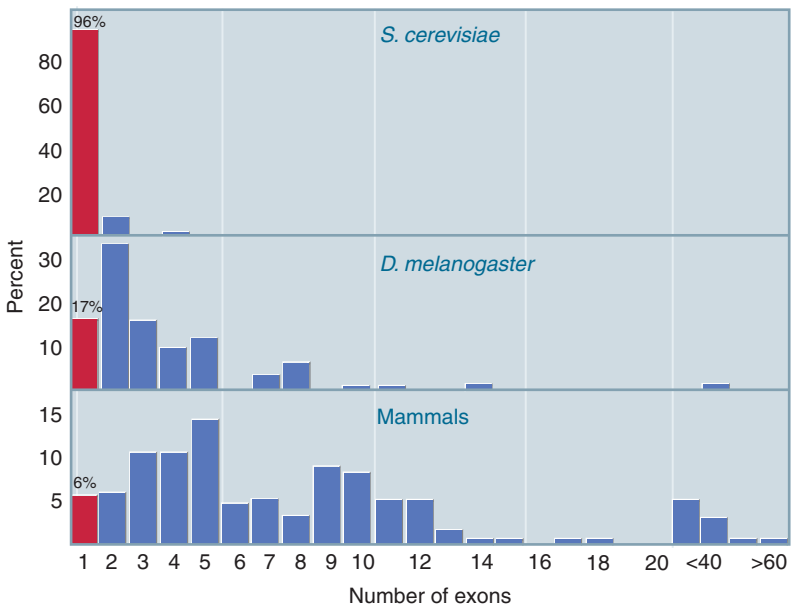


FIGURE 4.8 Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only one exon and are shown in the leftmost column in red.)

The structures of eukaryotic genes vary extensively. Some genes are uninterrupted, so that the gene sequence corresponds to the mRNA sequence. Most multicellular eukaryotic genes are interrupted, but among genes the introns vary enormously in both number and size. (See **FIGURE 4.8** for distributions of intron number variation among genes and **FIGURE 4.9** for distributions of intron-size variation.)

Genes encoding polypeptides, rRNA, or tRNA may all have introns. Introns are also found in mitochondrial genes of plants, fungi, protists, and one metazoan (a sea anemone), and in chloroplast genes. Genes with introns have been found in every class of eukaryotes, archaea, bacteria, and bacteriophages, although they are extremely rare in prokaryotic genomes.

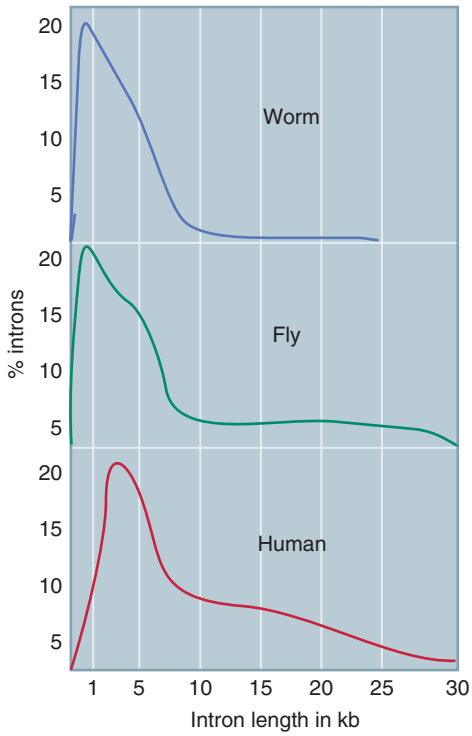


FIGURE 4.9 Introns range from very short to very long.

Some interrupted genes have only one or a few introns. The globin genes provide an extensively studied example (see *Section 4.8, Members of a Gene Family Have a Common Organization*). The two general classes of globin gene, α and β, share a common organization. They originated from an ancient gene duplication event and are described as **paralogous genes,** or **paralogs**. The consistent structure of mammalian globin genes is evident from the "generic" globin gene presented in **FIGURE 4.10**.

paralogous genes (paralogs) Genes that share a common ancestry due to gene duplication.

Introns are found at homologous positions (relative to the coding sequence) in all known active globin genes, including those of mammals, birds, and frogs. Although intron lengths vary, the first intron is always fairly short and the second is usually longer. Most of the variation in the lengths of different globin genes results from length variation in the second intron. For example, the second intron in the mouse β-globin gene is 150 bp of the total 850 bp of the gene, whereas the homologous intron in the mouse major β-globin gene is 585 bp of the total 1382 bp. The difference in length of the genes is much greater than that of their mRNAs (β-globin mRNA = 585 bases; β-globin mRNA = 620 bases).

The globin genes are examples of a general phenomenon: genes that share a common ancestry have similar organizations with conservation of the positions (of at least some) of the introns. Variations in the lengths of the genes are primarily due to intron length variation.

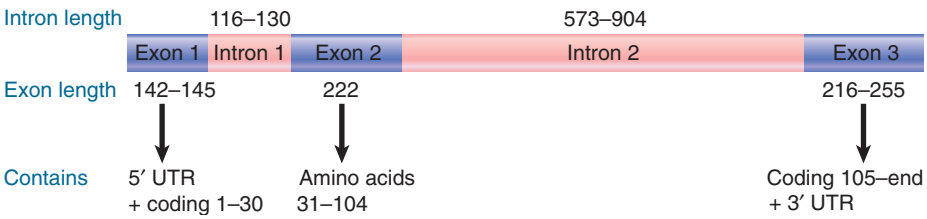


FIGURE 4.10 All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian β-globin genes.

METHODS AND TECHNIQUES

The Discovery of Introns by DNA-RNA Hybridization

The discovery of introns (intervening sequences) in eukaryotic genes was unexpectedly made in 1977. Two different experimental methods, both taking advantage of DNA-RNA hybridization, led to this startling finding. One avenue of research was the mapping of the adenovirus genome using RNA displacement loops to determine the position of mRNA transcripts relative to the viral genome. Adenovirus is a double-stranded virus that infects human epithelial cells, causing respiratory cold symptoms and stomach flu. This virus was an early model system for studying eukaryotic transcription.

In 1977, two independent groups, the labs of Richard Roberts at Cold Spring Harbor Laboratories and Phillip Sharp at the Massachusetts Institute of Technology, used the method of *R loop mapping* to determine the regions of the adenovirus genome that are transcribed. Using this technique, double-stranded genomic DNA is mixed with individual mRNA transcripts under conditions that promote DNA-RNA hybrids between complementary sequences. Since the mRNA will anneal to only one strand of the DNA, the other strand of DNA is displaced, resulting in a loop (R loop) of single-stranded DNA visible under the electron microscope (see Figure 4.4). When mapping the R loops formed between adenovirus DNA and viral mRNA transcripts expressed late in the infection cycle, it was observed that the 5′ and 3′ ends of the RNA did not anneal with the DNA. This was not surprising for the 3′ end of the mRNA transcript, which was known to be modified following transcription by the addition of a polyadenylate tail; however, the discontinuity of the 150 to 200 nucleotides at the 5′ end was unexpected. Hybridization of the 5′ terminal mRNA and discrete single-stranded fragments of genomic DNA demonstrated that this mRNA sequence is actually composed of sequences derived from three distinct and dispersed regions of the adenovirus genome (see Figure 4.5). This was the first evidence of the post-transcriptional process of mRNA splicing, and Drs. Roberts and Sharp were recognized for their discovery of "interrupted genes" in 1993 with a Nobel Prize in physiology or Medicine.

A second line of research, construction of physical maps of eukaryotic genes, revealed that interrupted genes are not unique to viral genomes. Prior to the advent of whole genome sequencing, physical maps of genomes were constructed by mapping restriction endonuclease cleavage sites in genomic DNA. Restriction endonucleases are bacterial enzymes that cleave double-stranded DNA in a site-specific fashion; different restriction endonucleases recognize and cleave distinct sequences (see *Section 3.2, Nucleases*). Whole-genomic DNA can be fragmented by cleavage with restriction endonucleases, and the resulting fragments are separated based on size through agarose gel electrophoresis. It is impossible to directly visualize the fragments for a specific gene because of the complexity and size of a typical eukaryotic genome; however, using the method of Southern blotting and DNA hybridization the fragments derived from a specific gene can be identified. The separated fragments of genomic DNA are denatured and transferred to a membrane in a process referred to as Southern blotting (named for its inventor Ed Southern; see *Section 3.9, Blotting Methods*). The membrane with the bound DNA is then immersed in a solution containing a specific radio-labeled DNA probe. The DNA probe will anneal only to the complementary sequences bound to the membrane, and when this membrane is washed to remove unbound probe, dried, and exposed to X-ray film, the fragments of DNA annealing to the probe are revealed. Typically, a physical map of a gene is constructed by using a cDNA (copied DNA from an mRNA transcript) to probe the genomic DNA. One of the first genes to be mapped in this fashion was the rabbit β-globin gene, performed by Jeffreys and Flavell in 1977. This led to the revelation that there is a sequence of 600 base pairs in the coding region of the genomic DNA that is not present in the cDNA. Another interrupted gene discovered at the same time is the chicken ovalbumin gene, which is also interrupted in the coding region. With time, it became clear that most eukaryotic genes are interrupted by intervening sequences that are transcribed and then removed by mRNA splicing. In 1978 Walter Gilbert proposed that these intervening sequences be referred to as *introns*.

KEY CONCEPTS

- Introns can be detected by the presence of additional regions when genes are compared with their RNA products by restriction mapping or electron microscopy. The ultimate determination, though, is based on comparison of sequences.
- The positions of introns are usually conserved when homologous genes are compared between different organisms. The lengths of the corresponding introns may vary greatly.

CONCEPT AND REASONING CHECK

Why would the genes of viruses with DNA genomes that infect eukaryotic cells be expected to have introns?

4.4 Exon Sequences Are Usually Conserved, but Introns Vary

Is a single-copy structural gene completely unique among other genes in a genome? The answer depends on how "completely unique" is defined. Considered as a whole, the gene is unique, but its exons may be related to those of other genes. As a general rule, when two genes are related, the relationship between their exons is closer than the relationship between their introns. In an extreme case, the exons of two different genes may encode the same polypeptide sequence while the introns are different. This situation can result from the duplication of a common ancestral gene followed by unique base substitutions in both copies, with substitutions restricted in the exons by the need to encode a functional polypeptide.

As we will see in *Chapter 8, Genome Evolution*, when we consider the evolution of the genome, exons can be considered basic building blocks that may be assembled in various combinations. It is possible for a gene to have some exons related to those of another gene, with the remaining exons unrelated. Usually, in such cases, the introns are not related at all. Such homologies between genes may result from duplication and translocation of individual exons.

The homology between two genes can be plotted in the form of a dot matrix comparison, as in **FIGURE 4.11**. A dot is placed in each position that is identical in both genes. The dots form a solid line on the diagonal of the matrix if the two sequences are completely identical. If they are not identical, the line is broken by gaps that lack homology and is displaced laterally or vertically by nucleotide deletions or insertions in one or the other sequence.

When the two mouse β-globin genes are compared in this way, a line of homology extends through the three exons and the small intron. The line disappears in the flanking UTRs and in the large intron. This is a typical pattern in related genes; the coding sequences and areas of introns adjacent to exons retain their similarity, but there is greater divergence in longer introns and the regions on either side of the coding sequence.

The overall degree of divergence between two homologous exons in related genes corresponds to the differences between the polypeptides. It is mostly a result of base substitutions. In the translated regions, changes in exon sequences are constrained by selection against mutations that alter or destroy the function of the polypeptide. Many of the preserved changes do not affect codon meanings because they change one codon into another for the same amino acid (i.e., they are synonymous substitutions). Similarly, there are higher rates of changes in nontranslated regions of the gene (specifically, those that are transcribed to the 5′ UTR [leader] and 3′ UTR [trailer] of the mRNA).

In homologous introns, the pattern of divergence involves both changes in length (due to deletions and insertions) and base substitutions. Introns evolve much more rapidly than exons. When a

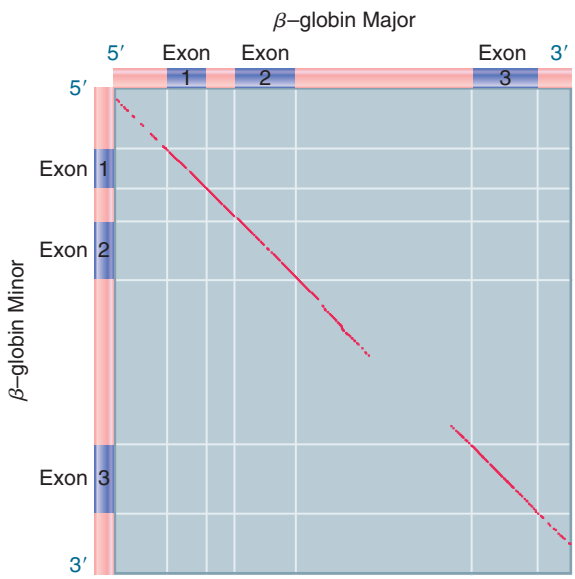


FIGURE 4.11 The sequences of the mouse β^{maj}- and β^{min}-globin genes are closely related in coding regions but differ in the flanking UTRs and the long intron. Data provided by Philip Leder, Harvard Medical School.

gene is compared among different species, there are instances where its exons are homologous but its introns have diverged so much that very little homology is retained. Although mutations in certain intron sequences (branch site, splicing junctions, and perhaps other sequences influencing splicing) will be subject to selection, most intron mutations are expected to be selectively neutral.

In general, mutations occur at the same rate in both exons and introns, but exon mutations are eliminated more effectively by selection. However, because of the low level of functional constraints, introns may more freely accumulate point substitutions and other changes. The empirical observation of faster evolution in introns implies that introns have fewer sequence-specific functions, but that does not necessarily mean that their presence is not required for normal gene function (see *Section 4.6, Some DNA Sequences Encode More Than One Polypeptide*).

KEY CONCEPTS

- Comparisons of related genes in different species show that the sequences of the corresponding exons are usually conserved but the sequences of the introns are much less similar.
- Introns evolve much more rapidly than exons because of the lack of selective pressure to produce a polypeptide with a useful sequence.

CONCEPT AND REASONING CHECK

In comparing two genes in different species, consider whether they are homologous given the following differences: (1) different exons; (2) the same exons, but completely different introns; (3) the same exons, and introns in the same positions but of different sizes.

4.5 Genes Show a Wide Distribution of Sizes Due Primarily to Intron Size and Number Variation

Figure 4.8 compares the organization of genes in a yeast, an insect, and mammals. In the yeast *Saccharomyces cerevisiae*, the majority of genes (~96%) are not interrupted, and those that have introns generally have three or fewer. There are virtually no *S. cerevisiae* genes with more than four exons.

In insects and mammals, the situation is reversed. Only a few genes have uninterrupted coding sequences (6% in mammals). Insect genes tend to have a fairly small number of exons—typically fewer than 10. Mammalian genes are split into more pieces, and some have more than 60 exons. About half of mammalian genes have more than 10 introns.

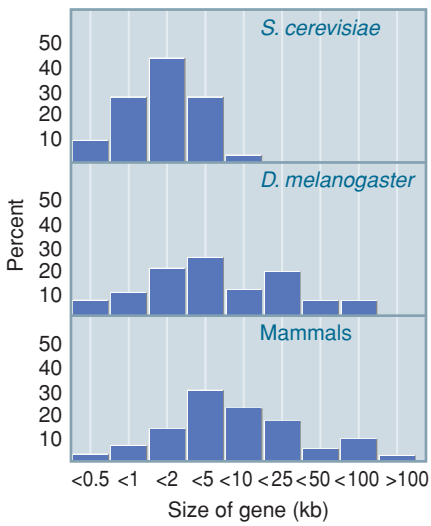


FIGURE 4.12 Yeast genes are short, but genes in flies and mammals have a dispersed distribution extending to very long sizes.

If we examine the effect of intron number variation on the total size of genes, we see in **FIGURE 4.12** that there is a striking difference between yeast and multicellular eukaryotes. The average yeast gene is 1.4 kb long, and very few are longer than 5 kb. The predominance of interrupted genes in multicellular eukaryotes, however, means that the gene can be much larger than the sum total of the exon lengths. Only a small percentage of genes in flies or mammals are shorter than 2 kb, and most have lengths between 5 kb and 100 kb. The average human gene is 27 kb long (see Figure 6.11). The dystrophin gene, with a length of 2000 kb, is the longest known human gene.

The switch from largely uninterrupted to largely interrupted genes seems to have occurred with the evolution of multicellular eukaryotes. In fungi other than *S. cerevisiae*, the majority of genes are interrupted, but they have a relatively small number of exons (<6) and are fairly short (<5 kb). In the fruit fly, gene sizes have a bimodal distribution—many are short but some are quite long. With this increase in the length of the gene due to the increased number of introns, the correlation between genome size and organism complexity becomes weak (see Figure 8.7).

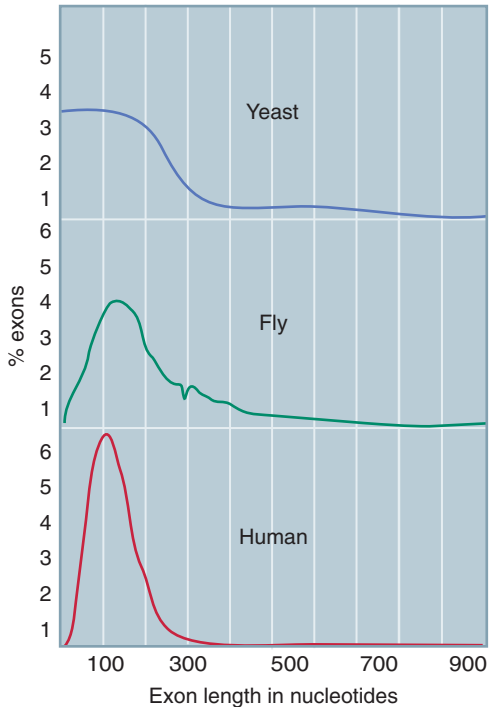


FIGURE 4.13 Exons encoding polypeptides are usually short.

Is there an evolutionary trend in intron length as well as intron number? Yes, organisms with larger genomes tend to have larger introns, whereas exon size does not tend to increase (**FIGURE 4.13**). In multicellular eukaryotes, the average exon codes for ~50 amino acids, and the general distribution is consistent with the hypothesis that genes have evolved by the gradual addition of exon units that code for short, functionally independent protein domains (see *Section 8.6, How Did Interrupted Genes Evolve?*). There is no significant difference in the average size of exons in different multicellular eukaryotes, although the size range is smaller in vertebrates for which there are few exons longer than 200 bp. In yeast, there are some longer exons that represent uninterrupted genes for which the coding sequence is intact.

Figure 4.9 shows that introns vary widely in size among multicellular eukaryotes. In worms and flies, the average intron is not much longer than the exons. There are no very long introns in worms, but flies contain many. In vertebrates, the size distribution is much wider, extending from approximately the same length as the exons (<200 bp) up to 60 kb in extreme cases. Some fish, such as fugu, have compressed genomes with shorter introns and intergenic spaces than mammals have.

Very long genes are the result of very long introns, not the result of encoding longer products. There is no correlation between total gene size and total exon size in multicellular eukaryotes, nor is there a good correlation between gene size and number of exons. The size of a gene is, therefore, determined primarily by the lengths of its individual introns. In mammals and insects, the average gene length is approximately five times that of the total length of its exons.

KEY CONCEPTS

- Most genes are uninterrupted in *S. cerevisiae* but are interrupted in most other eukaryotes.
- Exons are usually short, typically coding for <100 amino acids.
- Introns are short in unicellular eukaryotes but can be many kilobases in length in multicellular eukaryotes.
- The overall length of a gene is determined largely by its introns.

CONCEPT AND REASONING CHECK

Would you expect there to be a general evolutionary trend for increase or decrease in the size of introns? Why or why not?

4.6 Some DNA Sequences Encode More Than One Polypeptide

Most structural genes consist of a sequence of DNA that encodes a single polypeptide, although the gene may include noncoding regions at both ends and introns within the coding region. However, there are some cases in which a single sequence of DNA encodes more than one polypeptide.

overlapping gene A gene in which part of the sequence is found within part of the sequence of another gene.

The simplest **overlapping gene** is one in which one gene is part of another. In other words, the first half (or second half) of a gene independently specifies a protein that is the first (or second) half of the protein specified by the full gene. This relationship is illustrated in FIGURE 4.14.

A more complex form of overlapping gene occurs when the same sequence of DNA encodes two nonhomologous proteins because it has more than one reading frame. Usually, a coding DNA sequence is read in only one of the three potential reading frames. In some viral and mitochondrial genes, however, there is some overlap between two adjacent genes that are read in different reading frames, as illustrated in FIGURE 4.15. The length of overlap is usually short, so that most of the DNA sequence encodes a unique protein sequence.

mature transcript A modified RNA transcript. Modification may include the removal of intron sequences and alterations to the 5′ and 3′ ends.

alternative splicing The production of different RNA products from a single product by changes in the usage of splicing junctions.

In many genes, alternative proteins result from switches in the process of connecting the exons. A single gene may generate a variety of mRNA products that differ in their exon content. Often, this is because there are pairs of exons that are treated as mutually exclusive—one or the other is included in the **mature transcript**, but not both. The alternative proteins have one part in common and one unique part. Such an example is presented in FIGURE 4.16. The 3′ half of the rat troponin T gene contains five exons, but only four are used to construct an individual mRNA. Three exons, W, X, and Z, are included in all mRNAs. However, in one **alternative splicing** pattern the α exon is included between X and Z, whereas in the other pattern it is replaced by the β exon. The α and β forms of troponin T therefore

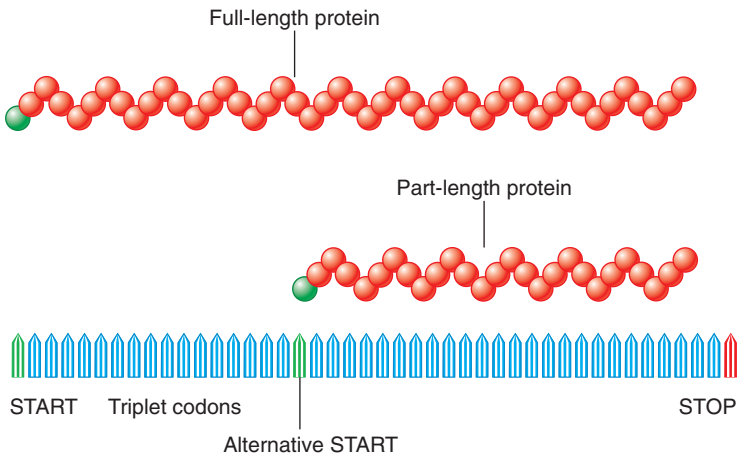


FIGURE 4.14 Two proteins can be generated from a single gene by starting (or terminating) expression at different points.

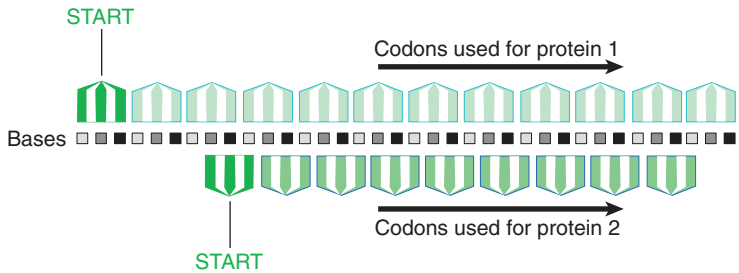


FIGURE 4.15 Two genes may overlap by reading the same DNA sequence in different frames.

differ in the sequence of the amino acids between W and Z, depending on which of the alternative exons (α or β) is used.

There are also cases of alternative splicing in which certain exons are optional; i.e., they may be included or spliced out, as in the example presented in FIGURE 4.17. There is a single **primary transcript** from the gene, but it can be spliced in either one of two ways. In the first, more standard way, the two introns are spliced out and the three exons are joined together. In the second way, the second exon is excluded along with the two introns as if a single large intron is spliced out. The two alternate proteins are the same at their ends, but one has an additional sequence in the middle. (Other types of combinations that are produced by alternative splicing are discussed in *Section 21.11, Alternative Splicing Is a Rule, Rather Than an Exception, in Multicellular Eukaryotes.*)

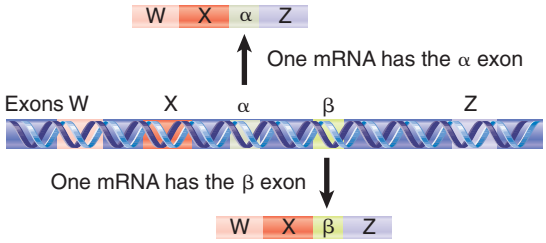


FIGURE 4.16 Alternative splicing generates the α and β variants of troponin T.

Sometimes the two alternative splicing patterns operate simultaneously, with a certain proportion of the primary mRNA transcripts being spliced in each way. However, in some genes the splicing patterns are alternatives that are expressed under different conditions, e.g., one in one cell type and one in another cell type.

So, alternative (or differential) splicing can generate different proteins with related sequences from a single stretch of DNA. It is curious that the multicellular eukaryotic genome is often extremely large, with long genes that are often widely dispersed along a chromosome, but at the same time there may be multiple products from a single locus. Due to alternative splicing, there are ~15% more proteins than genes in flies and worms, but it is estimated that the majority of human genes are alternatively spliced (see *Section 6.5, The Human Genome Has Fewer Genes Than Originally Expected*).

primary transcript The original unmodified RNA product corresponding to the transcription unit of a gene.

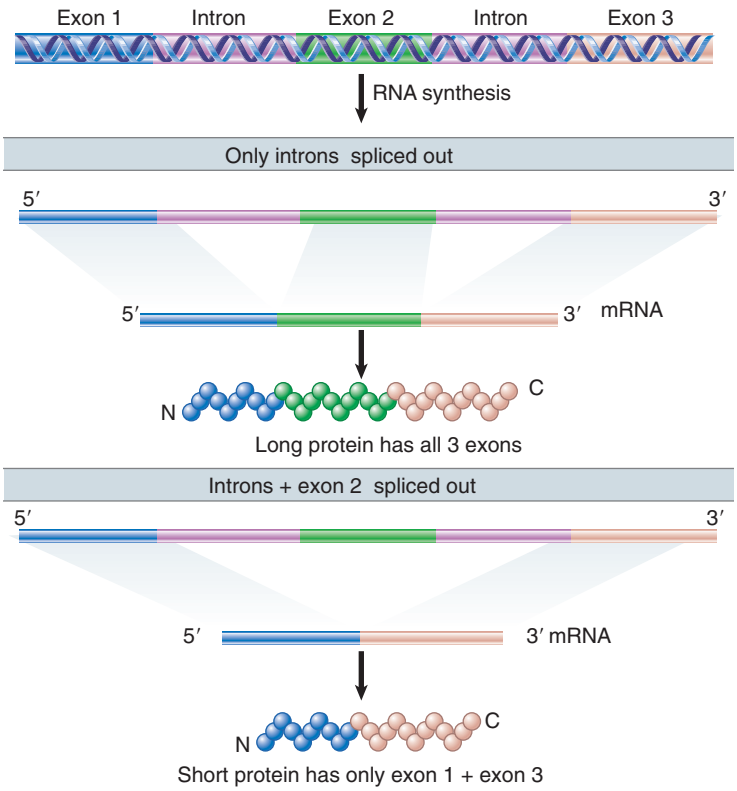


FIGURE 4.17 Alternative splicing uses the same pre-mRNA to generate mRNAs that have different combinations of exons.

KEY CONCEPTS

- The use of alternative start or stop codons allows two proteins to be generated where one is equivalent to a fragment of the other.
- Nonhomologous protein sequences can be produced from the same sequence of DNA when it is read in different reading frames by two (overlapping) genes.
- Homologous proteins that differ by the presence or absence of certain regions can be generated by differential (alternative) splicing when certain exons are included or excluded. This may take the form of including or excluding individual exons or of choosing between alternative exons.

CONCEPT AND REASONING CHECK

What are the advantages to having the same gene produce two or more related polypeptides? Why might different versions of the same polypeptide be necessary?

4.7 Some Exons Can Be Equated with Protein Functional Domains

The issue of the evolution of interrupted genes is more fully considered in *Section 8.6, How Did Interrupted Genes Evolve?* If proteins evolve by recombining parts of ancestral proteins that were originally separate, the accumulation of protein domains is likely to have occurred sequentially, with one exon added at a time. Are the current functions of these domains the same as their original functions? In other words, can we assign particular functions of current proteins to individual exons?

In some cases, there is a clear relationship between the structures of the gene and the protein. The example *par excellence* is provided by the immunoglobulin (antibody) proteins, which are encoded by genes in which every exon corresponds exactly to a known functional domain of the protein. **FIGURE 4.18** compares the structure of an immunoglobulin with its gene.

An immunoglobulin is a tetramer of two light chains and two heavy chains that covalently bond to generate a protein with several distinct domains. Light chains and heavy chains differ in structure, and there are several types of heavy chains. Each type of chain is produced from a gene that has a series of exons corresponding to the structural domains of the protein.

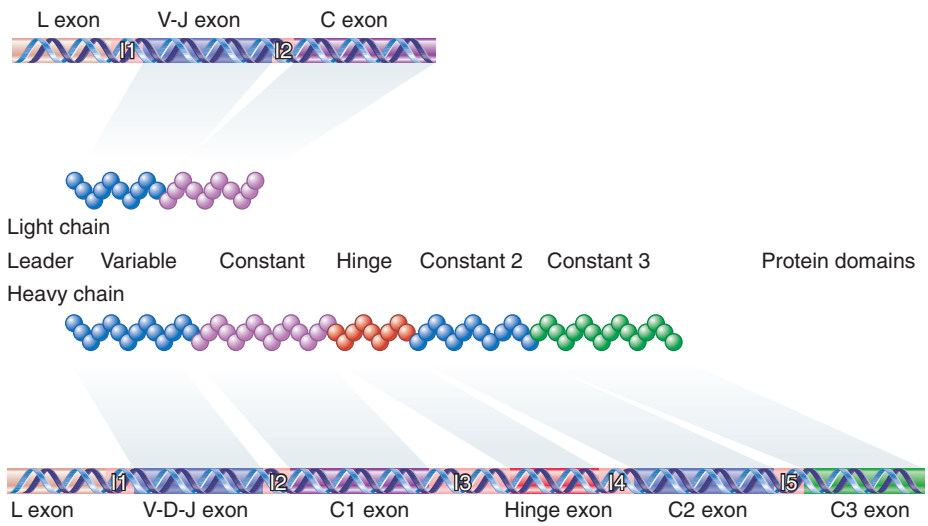


FIGURE 4.18 Immunoglobulin light chains and heavy chains are encoded by genes whose structures (in their expressed forms) correspond to the distinct domains in the protein. Each protein domain corresponds to an exon; introns are numbered I1 to I5.

In many instances, some of the exons of a gene can be identified with particular functions. In secretory proteins, such as insulin, the first exon often specifies the signal sequence involved in membrane secretion.

The view that exons are the functional building blocks of genes is supported by cases in which two genes may share some related exons but also have unique exons. **FIGURE 4.19** summarizes the relationship between the human LDL (plasma low density lipoprotein) receptor and other proteins. The LDL receptor gene has a series of exons related to the exons of the EGF (epidermal growth factor) precursor gene, and another series of exons related to those of the blood protein complement factor C9. Apparently, the LDL receptor gene evolved by the assembly of modules suitable for its various functions. These modules are also used in different combinations in other proteins.

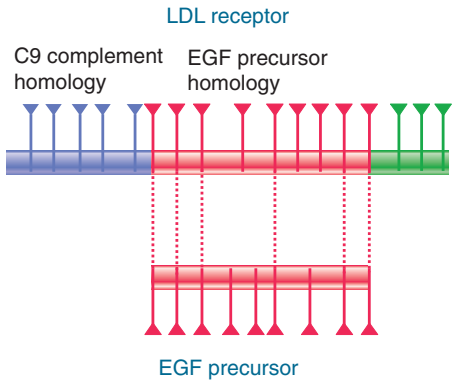


FIGURE 4.19 The LDL receptor gene consists of 18 exons, some of which are related to EGF precursor exons and some of which are related to the C9 blood complement gene. Triangles mark the positions of introns.

Exons tend to be fairly small (see Figure 6.11), around the size of the smallest polypeptide that can assume a stable folded structure (~20 to 40 residues). It may be that proteins were originally assembled from rather small modules. Each individual module need not correspond to a current function; several modules could have combined to generate a new functional unit. Larger genes tend to have more exons, which is consistent with the view that proteins acquire multiple functions by successively adding appropriate modules.

KEY CONCEPTS

- Many exons can be equated with encoding polypeptide sequences that have particular functions.
- Related exons are found in different genes.

CONCEPT AND REASONING CHECK

Would you expect two membrane-bound enzymes that catalyze different reactions to have (1) completely different exons, (2) some similar exons and some different exons, or (3) all similar exons? Why?

4.8 Members of a Gene Family Have a Common Organization

Many genes in a multicellular eukaryotic genome are related to others in the same genome. A **gene family** is defined as a group of genes that encode related or identical products as a result of gene duplication events. After the first duplication event, the two copies are identical, but then they diverge as different mutations accumulate in them. Further duplications and divergences extend the family. The globin genes are an example of a family that can be divided into two subfamilies (α globin and β globin), but all its members have the same basic structure and function. In some cases, we can find genes that are more distantly related, but still can be recognized as having common ancestry. Such a group of gene families is called a **superfamily**.

A fascinating case of evolutionary conservation is presented by the α and β globins and two other proteins related to them. Myoglobin is a monomeric oxygen-binding protein in animals. Its amino acid sequence suggests a common (though ancient)

gene family A set of genes within a genome that encode related or identical proteins or RNAs. The members originated from duplication of an ancestral gene followed by accumulation of changes in sequence between the copies. Most often the members are related but not identical.

superfamily A set of genes all related by presumed descent from a common ancestor but now showing considerable variation.

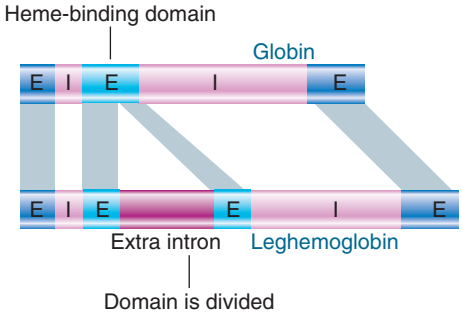


FIGURE 4.20 The exon structure of globin genes corresponds to protein function, but leghemoglobin has an extra intron in the central domain.

origin with globins. Leghemoglobins are oxygen-binding proteins found in legume plants; like myoglobin, they are monomeric and share a common origin with the other heme-binding proteins. Together, the globins, myoglobin, and leghemoglobins make up the globin superfamily—a set of gene families all descended from an ancient common ancestor.

Both α- and β-globin genes have three exons and two introns at conserved positions (see Figure 4.6). The central exon represents the heme-binding domain of the globin chain. There is a single myoglobin gene in the human genome and its structure is essentially the same as that of the globin genes. The three-exon structure therefore predates the common ancestor of the myoglobin and globin genes. Leghemoglobin genes contain three introns, the first and last of which are homologous to the two introns in the globin genes. This remarkable similarity suggests an exceedingly ancient origin for the interrupted structure of heme-binding proteins, as illustrated in **FIGURE 4.20**.

- **orthologous genes (orthologs)** Related genes in different species.
- **homologous genes (homologs)** Related genes in the same species, such as alleles on homologous chromosomes or multiple genes in the same genome sharing common ancestry.

Orthologous genes, or **orthologs**, are genes that are **homologous** (**homologs**) due to speciation; in other words, they are related genes in different species. Comparison of orthologs that differ in structure may provide information about their evolution. An example is insulin. Mammals and birds have only one gene for insulin, except for rodents, which have two. **FIGURE 4.21** illustrates the structures of these genes. We use the principle of parsimony in comparing the organization of orthologous genes by assuming that a common feature predates the evolutionary separation of the two species. In chickens, the single insulin gene has two introns; one of the two homologous rat genes has the same structure. The common structure implies that the ancestral insulin gene had two introns. However, since the second rat gene has only one intron, it must have evolved by a gene duplication in rodents that was followed by the precise removal of one intron from one of the homologs.

The organizations of some orthologs show extensive discrepancies between species. In these cases, there must have been deletion or insertion of introns during evolution. A well-characterized example is that of the actin genes. The common features of actin genes are a nontranslated leader of <100 bases, a coding region of ~1200 bases, and a trailer of ~200 bases. Most actin genes have introns, and their positions can be aligned with regard to the coding sequence (except for a single intron sometimes found in the leader). **FIGURE 4.22** shows that almost every actin gene is different in its pattern of intron positions. Among all the genes being compared, introns occur at 19 different sites. However, the range of intron number per gene is zero to six. How did this situation arise? If we suppose that the primordial actin gene had introns, and that all current actin genes are related to it by loss of introns, different introns have been lost in each evolutionary branch. Probably some introns have been lost entirely, so the primordial gene could well have had 20 introns or more. The alternative is to suppose that a process of intron insertion continued independently in the different lines of evolution.

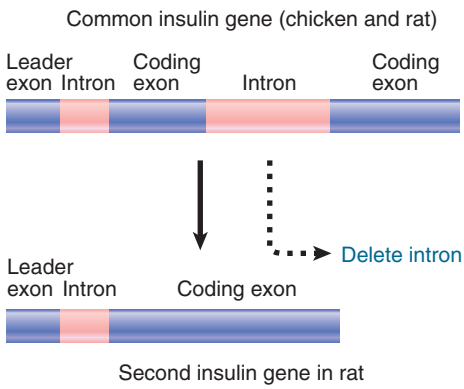


FIGURE 4.21 The rat insulin gene with one intron evolved by loss of an intron from an ancestor with two introns.

The relationship between individual exons and functional protein domains is somewhat erratic. In some cases there is a clear 1:1 relationship; in others no pattern can be discerned. One possibility is that the removal of introns has fused previously adjacent exons. This means that the intron must have been precisely removed, without changing the integrity of the coding region. An alternative is that some introns arose by

FIGURE 4.22 Actin genes vary widely in their organization. The sites of introns are indicated by dark boxes. The bar at the top summarizes all the intron positions among the different orthologs.

insertion into an exon encoding a single domain. Together with the variations that we see in exon placement in cases such as the actin genes, the conclusion is that intron positions can evolve.

The correspondence of at least some exons with protein domains and the presence of related exons in different proteins leave no doubt that the duplication and juxtaposition of exons have played important roles in evolution. It is possible that the number of ancestral exons—from which all proteins have been derived by duplication, variation, and recombination—could be relatively small, perhaps as little as a few thousand. The idea that exons are the building blocks of new genes is consistent with the "introns early" model for the origin of genes encoding proteins (see *Section 8.6, How Did Interrupted Genes Evolve?*).

KEY CONCEPTS

- A common feature in a set of genes is assumed to identify a property that preceded their separation in evolution.
- All globin genes have a common form of organization with three exons and two introns, suggesting that they are descended from a single ancestral gene.

CONCEPT AND REASONING CHECK

"Parsimony" is the least complex explanation for a given observation. For example, the explanation that modern actin genes evolved from an ancestral gene with many exons is more parsimonious than the explanation that modern actin genes gained introns independently. Evaluate the statement, "Selecting the most parsimonious explanation guarantees the choice of the correct explanation."

4.9 Summary

Virtually all eukaryotic genomes contain interrupted genes. The proportion of interrupted genes is low in some fungi, but few genes are uninterrupted in multicellular eukaryotes. Introns are found in all classes of eukaryotic genes. The structure of the interrupted gene is the same in all tissues: exons are spliced together in RNA in the same order as they are found in DNA, and the introns, which usually have no coding function, are removed from RNA by splicing.

Some genes are expressed by alternative splicing patterns, in which a particular sequence is removed as an intron in some situations, but retained as an exon in others. Often, when the organizations of orthologous genes are compared, the positions of

introns are conserved. Intron sequences vary—and may even be unrelated—although exon sequences are clearly related. The conservation of exon sequence and position can be used to isolate related genes in different species.

The size of a gene is determined primarily by the lengths of its introns. Large introns probably first appeared in the multicellular eukaryotes, and there is an evolutionary trend toward increased intron (and consequently, gene) size. The range of gene sizes in mammals is generally from 1 to 100 kb, but it is possible to have even larger genes.

Some genes share only some of their exons with other genes, suggesting that they have been assembled by addition of exons representing functional "modular units" of the protein. Such modular exons may have been incorporated into a variety of different proteins.

CHAPTER QUESTIONS

1. Mutations that affect splicing:
- **A.** are usually inconsequential.
- **B.** are usually deleterious.
- **C.** are always deleterious.
- **D.** increase gene expression.

2. In which group of organisms are most genes interrupted (have introns)?
- **A.** bacteria
- **B.** yeast
- **C.** animals
- **D.** more than one of the above

3. The length of genes in a gene family often varies; this variation is most often determined by:
- **A.** length of the 5′ untranslated region.
- **B.** number and size of exons.
- **C.** number and size of introns.
- **D.** length of the 3′ untranslated region.

4. In general, for two genes that are related, which shows the closest relationship?
- **A.** exons
- **B.** introns
- **C.** both introns and exons
- **D.** the promoter region and first exon

5. What percentage of genes is uninterrupted in animals?
- **A.** <5%
- **B.** <10%
- **C.** <20%
- **D.** ~33%

6. In general, as genome size increases in different organisms:
- **A.** exons also increase in size.
- **B.** introns increase in size.
- **C.** both exons and introns increase in size.
- **D.** 5′ and 3′ untranslated regions increase in size.

7. Which group of organisms has the longest average intron size?
- **A.** bacteria
- **B.** flies
- **C.** worms
- **D.** mammals

8. The hypothesis that new genes are constructed by combining exons from existing genes is supported by the observation that genes with different functions:
   A. have all related introns.
   B. have some related exons.
   C. have all related exons.
   D. have some related introns.
9. Genes that encode related or identical proteins in an organism are part of:
   A. a gene family.
   B. a superfamily.
   C. homologous genes.
   D. orthologous genes.
10. α globin and β globin are:
   A. unrelated.
   B. orthologs of each other.
   C. paralogs of myoglobin.
   D. paralogs of each other.

KEY TERMS

alternative splicing
cDNA
exon
gene family
homologous genes (homologs)
interrupted gene
intron
mature transcript
orthologous genes (orthologs)
overlapping gene
paralogous genes (paralogs)
primary transcript
RNA splicing
superfamily

FURTHER READING

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72,** 291–336. An in-depth review of specific examples of alternative splicing, including the *Drosophila* sex determination system.

Faustino, N. A., and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes Dev.* **17,** 419–437. A review of the mechanisms by which problems with alternative splicing of pre-mRNA can result in human diseases.

Ponting, C. P., and Russell, R. R. (2002). The natural history of protein domains. *Annu. Rev. Biophys. Biomol.* **31,** 45–71. A discussion of the origin and evolution of protein domains, including the suggestion that domains be classified in a hierarchical taxonomic system.

Reddy, A. S. N. (2007). Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.* **58,** 267–294. A review of unexpected recent discoveries of alternative splicing of plant genes.

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F. J. (2006). Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.* **40,** 47–76. A review of classic and newer hypotheses about the origins of introns.