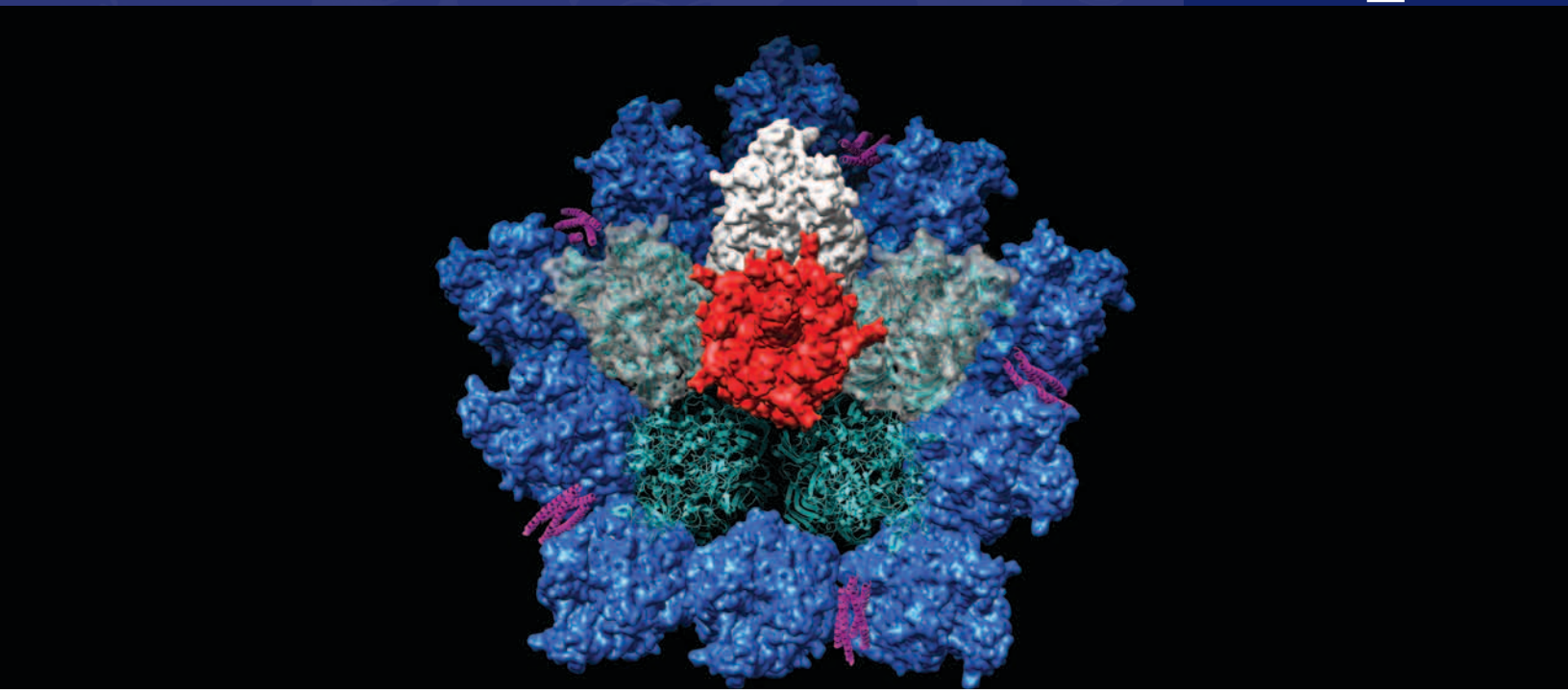
4



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

The Interrupted Gene

Edited by Donald Forsdyke

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 sequence of $3N$ bases encodes a polypeptide of N amino acids. However, in eukaryotes ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and messenger RNAs (mRNAs) are first synthesized as long precursor transcripts that are subsequently shortened (see the chapter titled *RNA Splicing and Processing*). Thus eukaryotic genes are much longer than the functional transcripts they produce. It is reasonable to assume that the shortening involved a trimming of additional, perhaps regulatory, sequences at the 5′ and/or 3′ ends of transcripts, leaving the rRNA or protein-encoding sequence of the precursor intact.

However, as it happens a eukaryotic gene can include additional sequences that lie both *within* and outside the region that is operational with respect to phenotype. Protein-encoding sequences can be interrupted, as can the 5′ and 3′ sequences (UTRs) that flank the protein-encoding sequences within mRNA. The interrupting sequences are removed from the **primary (RNA) transcript** (or **pre-mRNA**) during gene expression, generating an mRNA that includes a continuous base sequence corresponding to the polypeptide product as determined by the genetic code. The sequences of DNA comprising an interrupted protein-encoding gene are divided into the two categories depicted in FIGURE 4.1:

- **Exons** are the sequences retained in the mature RNA product. A **mature transcript** starts 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 RNA 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 mature RNA product because of the presence of the introns.

The processing of interrupted genes requires an additional step that is not necessary in uninterrupted genes. The DNA of an interrupted gene is transcribed to an RNA copy (a *transcript*) that is exactly complementary to the original DNA 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 a series of exons. This process is called **RNA splicing** (see the chapter titled *Genes Encode RNAs and Polypeptides*) and involves precisely deleting the introns from the primary transcript and then joining the ends of the RNA on either side of each intron to form a covalently intact molecule (see the chapter titled *RNA Splicing and Processing*).

The original eukaryotic gene comprises the region in the genome between points corresponding to the 5′ and 3′ terminal bases of mature RNA. 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 RNA, which is generated by cleavage of the 3′ extension. The gene is also considered to include the regulatory regions on both sides of the gene that are required for the initiation and (sometimes) termination of transcription.

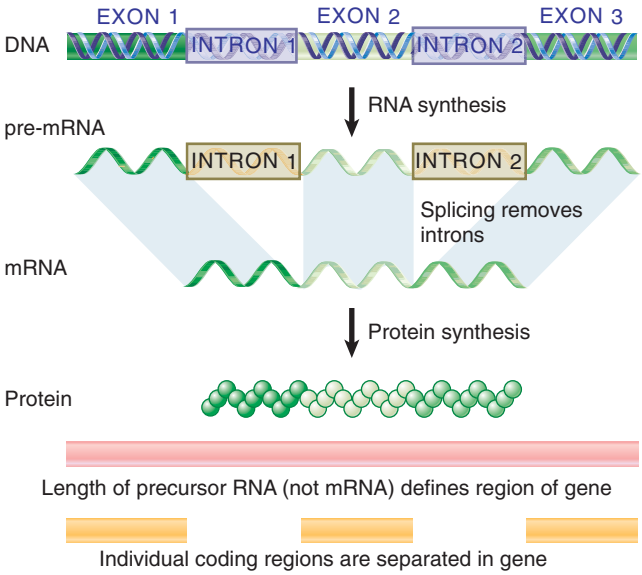


FIGURE 4.1 Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.

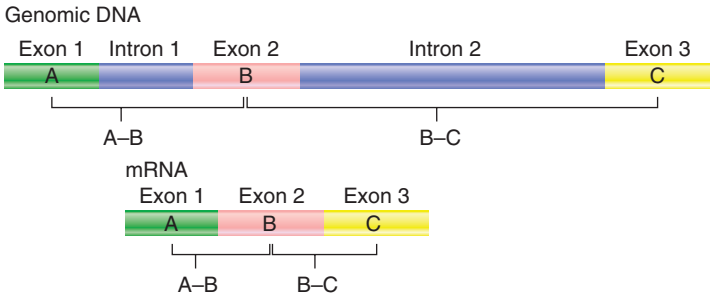


FIGURE 4.2 Exons remain in the same order in mRNA as in DNA, but distances along the gene do not correspond to distances along the mRNA or polypeptide products. The distance from A–B in the gene is smaller than the distance from B–C, but the distance from A–B in the mRNA (and polypeptide) is greater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

4.2 An Interrupted Gene Consists of Exons and Introns

Key concepts

- Introns are removed by RNA splicing, which occurs in *cis* in individual RNA molecules.
- Mutations in exons can affect polypeptide sequence; mutations in introns can affect RNA processing and hence may influence the sequence and/or production of a polypeptide.

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 a 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 a gene is defined by the length of the primary mRNA transcript instead of the length of the mature mRNA. All exons of a gene are on one RNA molecule, and their splicing together is an *intra*molecular reaction. There is usually no joining of exons carried by different RNA molecules, so there is rarely cross-splicing of sequences. (However, in a process known as *trans*-splicing, sequences from different mRNAs are ligated together into a single molecule for translation.)

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 mature mRNA, so mutations in them cannot directly affect the polypeptide sequence. However, they may

affect the processing of the mRNA production by inhibiting the splicing of exons. A mutation of this sort acts only on the allele that carries it.

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 shortens the polypeptide sequence. Thus, intron mutations may affect not only the production of a polypeptide, but also its 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 the yeast *Saccharomyces cerevisiae*, most genes are uninterrupted. In multicellular eukaryotes most genes are interrupted, and the introns are usually much longer than exons, so that genes are considerably larger than their coding regions.

4.3 Exon and Intron Base Compositions Differ

Key concepts

- The four "rules" for DNA base composition are the first and second parity rules, the cluster rule, and the GC rule. Exons and introns can be distinguished on the basis of all rules except the first.
- The second parity rule suggests an extrusion of structured stem-loop segments from duplex DNA, which would be greater in introns.
- The rules relate to genomic characteristics, or "pressures," that constitute the genome phenotype.

In the 1940s Erwin Chargaff initiated studies of DNA base composition that led to four "rules," beginning with the *first parity rule* for duplex DNA (see the chapter titled *Genes Are DNA*). This rule applies to most regions of DNA, including both exons and introns. Base A in one strand of the duplex is matched by a complementary base (T) in the other strand, and base G in one strand of the duplex is matched by a complementary base (C) in the other strand. By extension, the rule applies not only to single bases, but also to dinucleotides, trinucleotides, and oligonucleotides. Thus, GT pairs with its reverse complement AC, and ATG pairs with its reverse complement CAT. In addition to the well known first parity rule, later work by Chargaff led him to propose a *second parity rule*. The little-known second parity rule is that, to a close approximation, there are equal amounts of A and T, and equal amounts of C and G, in each single strand of the duplex. Like the first parity rule, this extends to oligonucleotide sequences: For example, in a very long strand there are approximately equal numbers of AC and TG dinucleotides. The reasons for the existence of this rule are not clear, but sequencing of many genomes has shown it to be nearly universally true. The second parity rule applies more closely to introns than to exons, partly due to a further rule—that purines tend to cluster on one DNA strand and pyrimidines tend to cluster on the other. This *cluster rule* as applied to exons is that the purines, A and G, tended to be clustered in one DNA strand of the DNA duplex (usually the nontemplate strand) and these are complemented by clusters of the pyrimidines, T and C, in the template strand.

The fact that in single-stranded DNA an oligonucleotide is accompanied in series by equal quantities of its reverse complementary oligonucleotide suggests that duplex DNA has the potential to extrude folded stem-loop structures, the stems of which can display base parity and the loops of which can display some degree of base clustering. Indeed, the potential for such secondary structure is found to be greater in introns than exons, especially in exons under positive selection pressure (see the section titled *Exon Sequences Under Positive Selection Vary but Introns Are Conserved* later in this chapter).

Finally, there is the *GC rule*, which is that the overall proportion of G+C in a genome (*GC content*) tends to be a species-specific character (although individual genes within that genome tend to have distinctive values). The GC content tends to be greater in exons than in introns. Chargaff's four rules are seen to relate to characters or "pressures" that are *intrinsic* to the genome, contributing to what was termed the *genome phenotype* (see the section titled *There Are Many Forms of Information in DNA* later in this chapter).

4.4 Organization of Interrupted Genes May Be Conserved

Key concepts

- Introns can be detected when genes are compared with their RNA transcription products by either restriction mapping, electron microscopy, or sequencing.
- The positions of introns are usually conserved when homologous genes are compared between different organisms. The lengths of the corresponding introns may vary greatly.
- Introns usually do not encode proteins.

When a gene is uninterrupted, the restriction map of its DNA corresponds with the map of its mRNA. When a gene possesses an intron, the map at each end of the gene corresponds to the map at each end of the message sequence. Within the gene, however, the maps diverge because additional regions that are found in the gene are not represented in the mature mRNA. Each such region corresponds to an intron. The example in **FIGURE 4.3** compares the restriction maps of a β-globin gene and its mRNA. There are 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. The finer comparison of the base sequences of a gene and its mRNA permits precise identification of introns. An intron usually has no open reading frame. An intact reading frame is created in an mRNA sequence by the removal of the introns from the primary transcript.

The structures of eukaryotic genes show extensive variation. Some genes are uninterrupted and their sequences are colinear with those of the corresponding mRNAs. Most multicellular eukaryotic genes are interrupted, but the introns vary enormously in both number and size.

Genes encoding polypeptides, rRNA, or tRNA may all have introns. Introns also are 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.

Some interrupted genes have only one or a few introns. The globin genes provide a much studied example (see the section titled *Members of a Gene Family Have a Common Organization* later in this chapter). 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 4.4**.

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 α-globin gene is only 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 (α-globin mRNA = 585 bases; β-globin mRNA = 620 bases).

The example of dihydrofolate reductase (DHFR), a somewhat larger gene, is shown in **FIGURE 4.5**. The mammalian DHFR gene is organized into six exons that correspond to a 2000-base mRNA. The gene itself is long because the introns are very long. In three mammal species the exons are essentially the same and the relative positions of

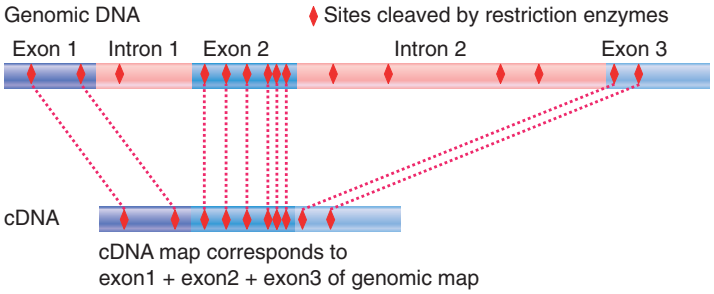


FIGURE 4.3 Comparison of the restriction maps of cDNA and genomic DNA for mouse β-globin shows that the gene has two introns that are not present in the cDNA. The exons can be aligned exactly between cDNA and the gene.

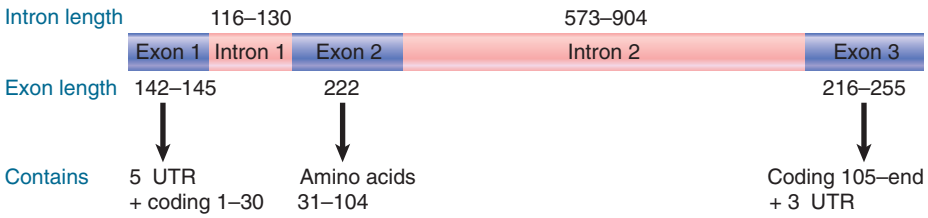


FIGURE 4.4 All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian β-globin genes.

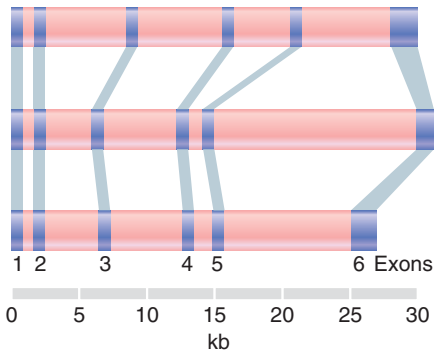


FIGURE 4.5 Mammalian genes for DHFR have the same relative organization of rather short exons and very long introns, but vary extensively in the lengths of introns.

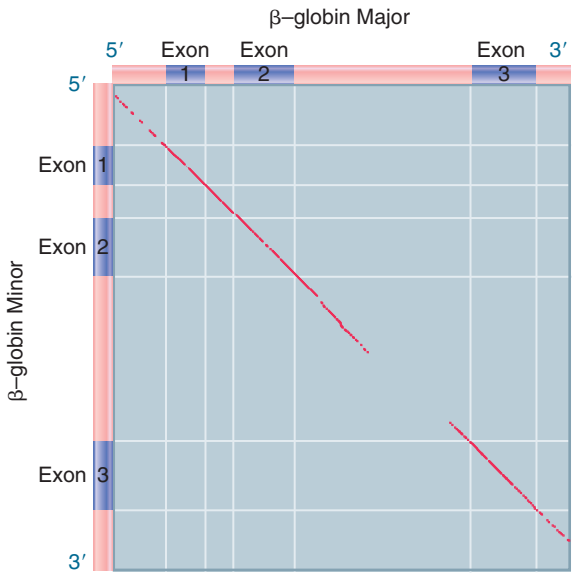


FIGURE 4.6 The sequences of the mouse β^{maj}- and β^{min}-globin genes are closely related in coding regions but differ in the flanking UTRs and the long intron. Data provided by Philip Leder, Harvard Medical School.

the introns are unaltered, but the lengths of individual introns vary extensively, resulting in a variation in the length of the gene from 25 to 31 kb.

The globin and DHFR 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.*

4.5 Exon Sequences Under Negative Selection Are Conserved but Introns Vary

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.

Is a single-copy structural gene completely unique among other genes in its 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 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 the chapter titled *Genome Evolution*, where 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 4.6**. 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 in 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. In other words, the exon sequences are conserved by the *negative selection* of individuals in which the sequences have changed (have not been conserved) to result in a phenotype that is less able to survive and produce fertile progeny. Many of the preserved changes do not affect codon meanings because they change a codon into another for the same amino acid (i.e., they are synonymous substitutions). In this case, the polypeptide will not change and negative selection will not operate on the phenotype conferred by the polypeptide. Similarly, there are higher rates of change in untranslated 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 the exons are under negative selection pressure. When a 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. Indeed, it is sometimes possible to locate exons in uncharted sequences by virtue of their conservation relative to introns (see the chapter *The Content of the Genome*). From this description it is all too easy to conclude that introns do not have a sequence-specific function. Genes under positive selection, however, cast a different light on the problem.

4.6 Exon Sequences Under Positive Selection Vary but Introns Are Conserved

Key concepts

- Under positive selection an individual with an advantageous mutation survives (i.e., is able to produce more fertile progeny) relative to others without the mutation.
- Due to intrinsic genomic pressures, such as that which conserves the potential to extrude stem-loops from duplex DNA, introns evolve more slowly than exons that are under positive selection pressure.

A mutation that confers a more advantageous phenotype to an organism, relative to individuals in the same population without the mutation, may result in the preferential survival (*positive selection*) of that organism. Pathogenic bacteria are killed by an antibiotic, but a bacterium with a mutation that confers antibiotic resistance survives (i.e., is positively selected). Mutations conferring venom-resistance to prey of venomous snakes can result in the positive selection of that prey relative to its fellows that succumb to the poison (i.e., are negatively selected). Likewise, a snake that, when confronted by a venom-resistant prey population, has a mutation that enhances the power of its venom, will be positively selected. This can trigger an attack–defense cycle—an "arm's race" between two protagonist species.

In such situations the pattern of exon conservation and intron variation seen in genes under negative selection can be reversed because exons evolve faster than introns. Thus, a plot similar to Figure 4.6 will have lines in introns and gaps in exons. Another way of showing this is to plot base substitutions along the length of a gene. **FIGURE 4.7** shows a plot of the substitutions observed when two snake venom alkaline phosphatase genes are compared. The protein-encoding parts of exons (2, 3, and the first half of exon 4) have many base substitutions (i.e., they are varying), whereas the three introns have relatively few (i.e., they are conserved).

What is being conserved in introns? First, intron sequences needed for RNA splicing—the 5′ and 3′ splice sites and the branch site—are conserved (see the chapter titled *RNA Splicing and Processing*). In addition to these, base order has been adapted to promote the potential of the duplex DNA in the region to extrude stem–loop structures (*fold potential*). Thus, a plot of base order–dependent fold potential along the

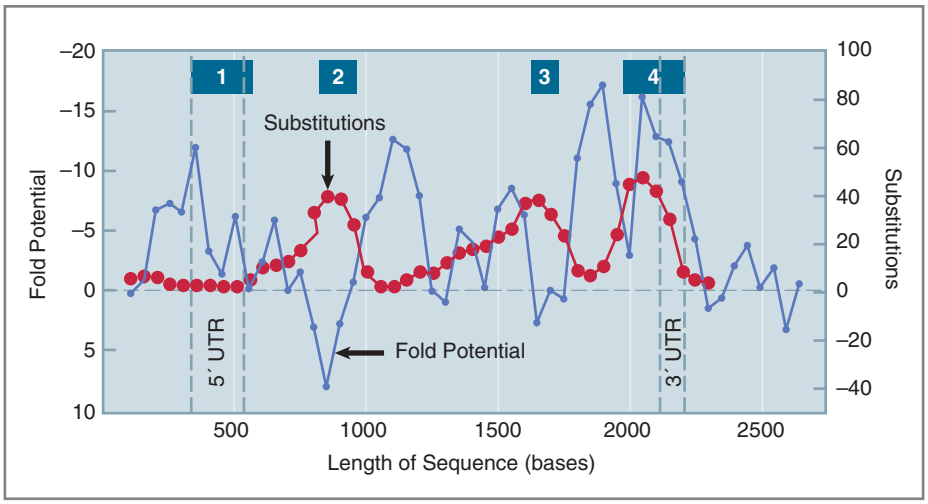


FIGURE 4.7 The sequences of snake venom phospholipase genes differ in coding regions, but are closely related in introns and flanking regions. Fold potential (here the contribution of base order to the potential to extrude stem–loop structures) is low (more positive) in the protein-encoding exons and high (more negative) in introns. The positions of the four exons are shown as numbered boxes. Modified from D. R. Forsdyke, Conservation of Stem-Loop Potential in Introns of Snake Venom Phospholipase A2 Genes: An Application of FORS-D Analysis, *Mol. Biol. Evol.*, vol. 12(6), pp. 1157–1165, by permission of Oxford University Press.

length of the gene shows that fold potential (measured in negative units) is high (more negative) in introns, and low (more positive) in exons (Figure 4.7). This reciprocal relationship between substitution frequency and the contribution of base order to fold potential is a characteristic of DNA sequences under positive selection. Indeed, the low (more positive) value of fold potential in an exon provides evaluation of the extent to which it has been under positive selection, without the need to compare two sequences (the classical way of determining if selection is positive or negative).

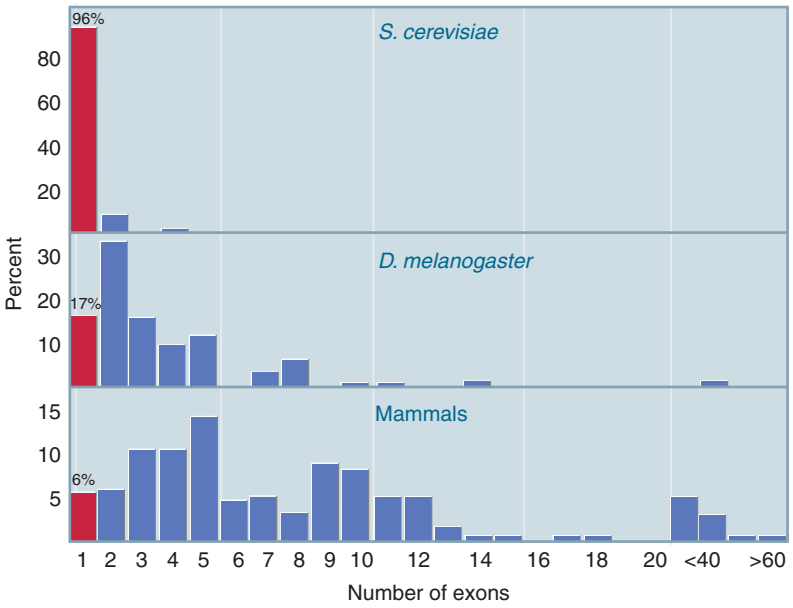


FIGURE 4.8 Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only one exon and are totaled in the leftmost column in red.)

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

Key concepts

- Most genes are uninterrupted in *Saccharomyces cerevisiae* but are interrupted in multicellular eukaryotes.
- Exons are usually short, typically encoding fewer than 100 amino acids.
- Introns are short in unicellular/oligocellular eukaryotes but can be many kb in multicellular eukaryotes.
- The overall length of a gene is determined largely by its introns.

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 uninterrupted, and those that have exons 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 small number of exons, typically fewer than 10. Mammalian genes are split into more pieces and some have more than 60 exons. Approximately 50% of mammalian genes have more than 10 introns. If we examine the effect of intron number variation on the total size of genes, we see in **FIGURE 4.9** 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. 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.

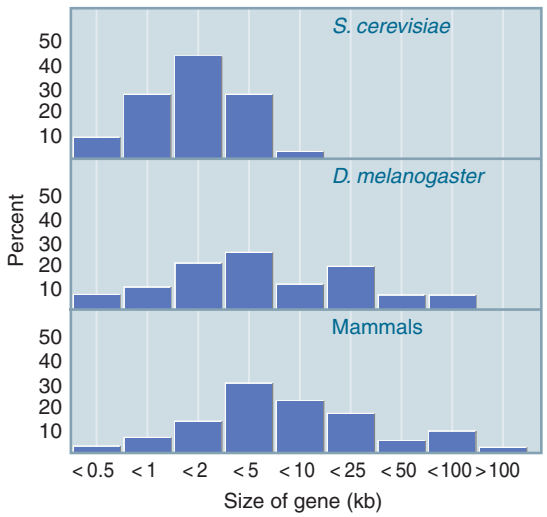


FIGURE 4.9 Yeast genes are short, but genes in flies and mammals have a dispersed bimodal distribution extending to very long sizes.

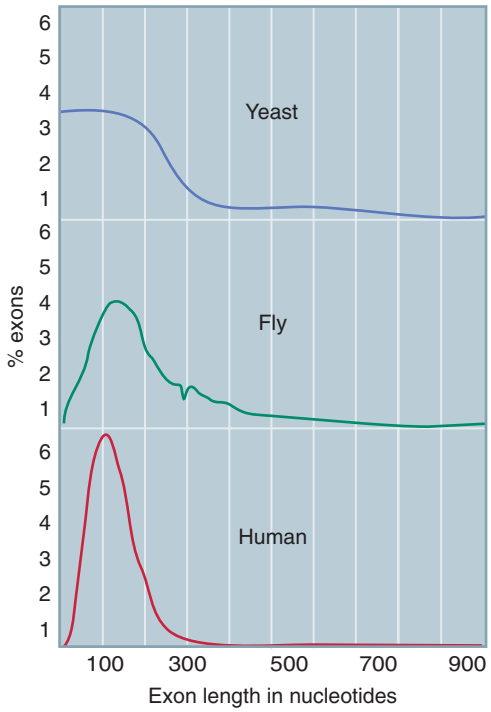


FIGURE 4.10 Exons encoding polypeptides are usually short.

FIGURE 4.10 shows that exons encoding stretches of protein tend to be fairly small. 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 encode short, functionally independent protein domains (see the *Genome Evolution* chapter). 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. There is a tendency for exons containing untranslated 5′ and 3′ regions to be longer than those that encode proteins.

FIGURE 4.11 shows that introns vary widely in size among multicellular eukaryotes. (Note that the scale of the x-axis differs from that of Figure 4.10.) In worms and flies, the average intron is not 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 regions than mammals have.)

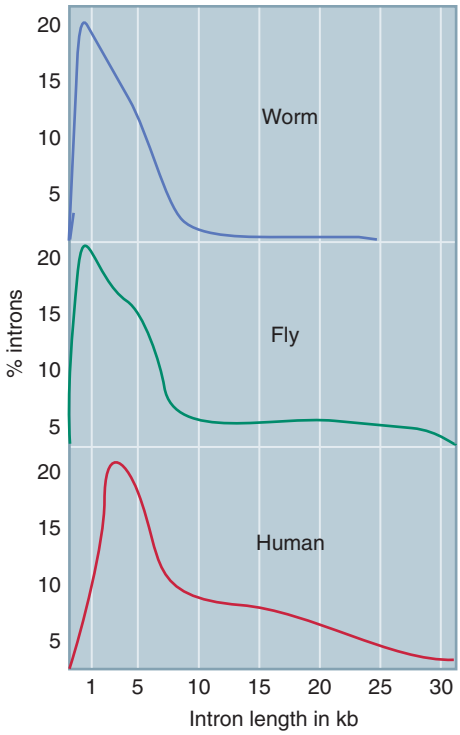


FIGURE 4.11 Introns range from very short to very long.

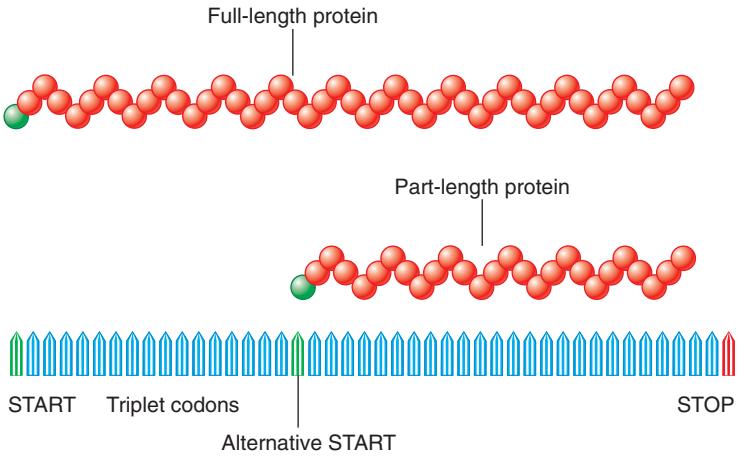


FIGURE 4.12 Two proteins can be generated from a single gene by starting (or terminating) expression at different points.

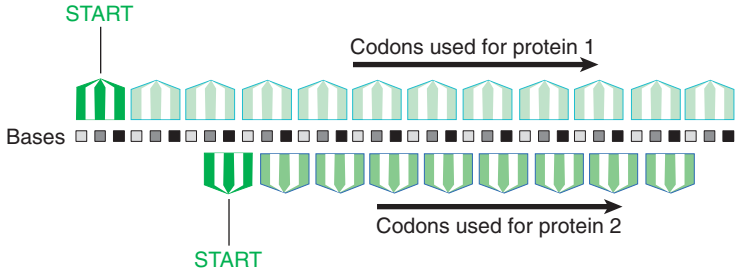


FIGURE 4.13 Two genes may overlap by reading the same DNA sequence in different frames.

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 is approximately 5× that of the total length of its exons.

4.8 Some DNA Sequences Encode More Than One Polypeptide

Key concepts

- The use of alternative initiation or termination codons allows multiple variants of a polypeptide chain.
- Different polypeptides can be produced from the same sequence of DNA when the mRNA is read in different reading frames (as two overlapping genes).
- Otherwise identical polypeptides, differing 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.

Many structural genes consist of a sequence 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.

In one simple example, a single DNA sequence may have two alternative start codons in the same reading frame (see **FIGURE 4.12**). Thus, under different conditions one or the other of the start codons may be used, allowing the production of either a short form of the polypeptide or a full-length form, where the short form is the last portion of the full-length form.

An actual **overlapping gene** occurs when the same sequence of DNA encodes two nonhomologous proteins because it uses 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 4.13**. The length of overlap is

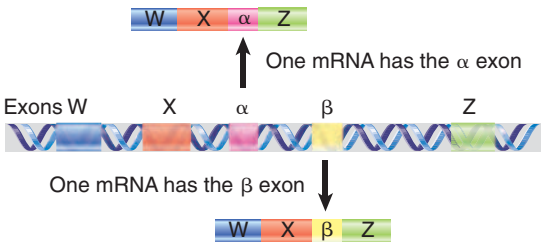


FIGURE 4.14 Alternative splicing generates the α and β variants of troponin T.

usually short, so that most of the DNA sequence encodes a unique polypeptide sequence.

In some cases, genes can be *nested*. This occurs when a complete gene is found within the intron of a larger "host" gene. Nested genes often lie on the strand opposite to that of the host gene.

In some genes there are switches in the pathway for splicing the exons that result in *alternative* patterns of gene expression. A single gene may generate a variety of mRNA products that differ in their exon content. Certain exons may be optional; in other words, they may be included or spliced out. There also may be a pair of exons 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.

In some cases, the alternative means of expression do not affect the sequence of the polypeptide. For example, changes that affect the 5′ UTR or the 3′ UTR may have regulatory consequences, but the same polypeptide is made. In other cases, one exon is substituted for another, as in **FIGURE 4.14**. In this example, the polypeptides produced by the two mRNAs contain sequences that overlap extensively, but are different within the alternatively spliced region. The 3′ half of the troponin T gene of rat muscle 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 differ in the sequence of the amino acids between *W* and *Z*, depending on which of the alternative exons (α or β) is used. Either one of the α and β exons can be used in an individual mRNA, but both cannot be used in the same mRNA.

FIGURE 4.15 shows that alternative splicing can lead to the inclusion of an exon in some mRNAs while leaving it out of others. A single

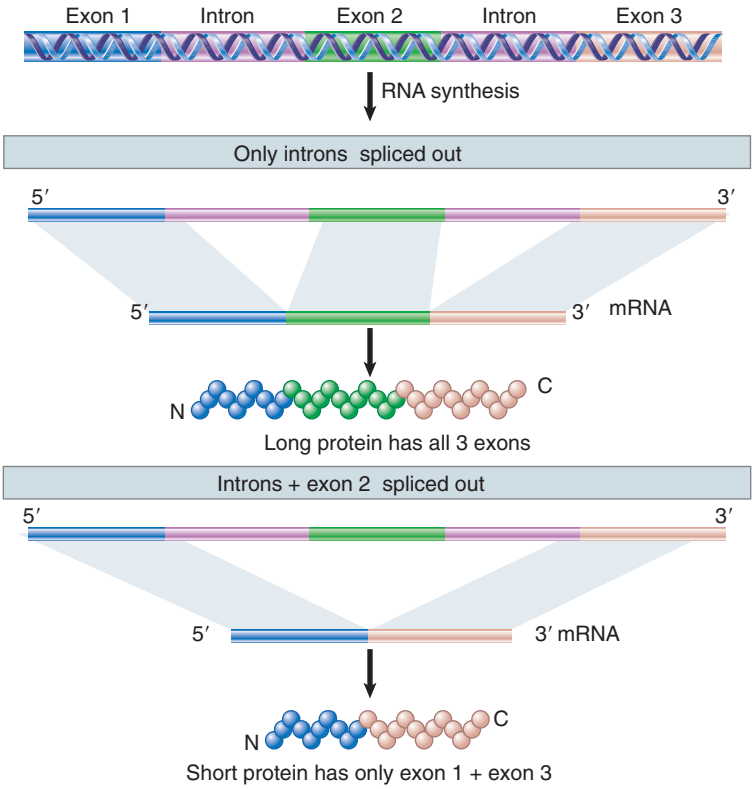


FIGURE 4.15 Alternative splicing uses the same pre-mRNA to generate mRNAs that have different combinations of exons.

primary transcript can be spliced in either of two ways. In the first (more standard) pathway, two introns are spliced out and the three exons are joined together. In the second pathway, the second exon is excluded as if a single large intron is spliced out. This intron consists of intron 1 + exon 2 + intron 2. In effect, exon 2 has been treated in this pathway as if it were part of a single intron. The pathways produce two polypeptides that 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 the *RNA Splicing and Processing* chapter).

Sometimes two alternative splicing pathways operate simultaneously, with a certain proportion of the primary RNA transcripts being spliced in each way. However, sometimes the pathways are alternatives that are expressed under different conditions, for example, one in one cell type and one in another cell type.

So, alternative (or differential) splicing can generate different polypeptides 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 polypeptides than genes in flies and worms, but it is estimated that the majority of human genes are alternatively spliced (see the chapter titled *Genome Sequences and Gene Numbers*).

4.9 Some Exons Can Be Equated with Protein Functional Domains

Key concepts

- Proteins can consist of independent functional modules, the boundaries of which, in some cases, can be equated with those of exons.
- The exons of some genes appear homologous to the exons of others, suggesting a common exon ancestry.

The issue of the evolution of interrupted genes is more fully considered in the *Genome Evolution* chapter. 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. Each addition would have to improve upon the advantages of prior additions in a sequence of positive selection events. Are the different function-encoding segments from which these genes may have originally been pieced together reflected in their present structures? If a protein sequence were randomly interrupted, sometimes the interruption would intersect a domain and sometimes it would lie between domains. If we can equate the functional domains of current proteins with the individual exons of the corresponding genes, then this would suggest selective interdomain interruptions rather than random ones.

In some cases there is a clear relationship between the structures of a gene and its protein product, but these may be special cases. The example *par excellence* is provided by the immunoglobulin (antibody) proteins—an extracellular system for self-/nonself-discrimination that aids in the elimination of foreign pathogens. Immunoglobulins are encoded by genes in which every exon corresponds exactly to a known functional protein domain. Banks of alternate sequence domains are tapped so that each cell acquires the ability to secrete a cell-specific immunoglobulin with distinctive binding capacity for a foreign antigen that the organism may one day encounter again (see the chapter titled *Somatic Recombination and Hypermutation in the Immune System*). **FIGURE 4.16** 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.

In many instances, some of the exons of a gene can be identified with particular functions. In secretory proteins, such as insulin,

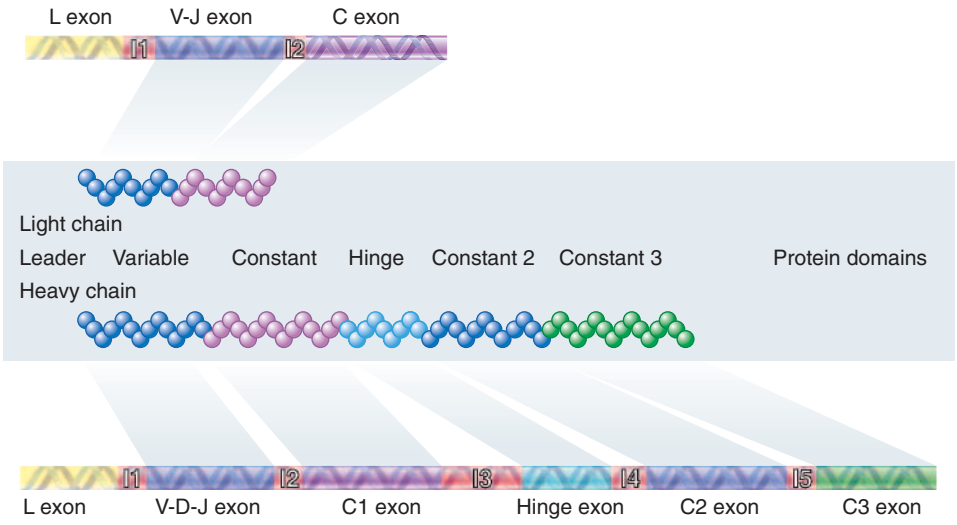


FIGURE 4.16 Immunoglobulin light chains and heavy chains are encoded by genes whose structures (in their expressed forms) correspond to the distinct domains in the protein. Each protein domain corresponds to an exon; introns are numbered I1 to I5.

the first exon that encodes for the N-terminal region of the polypeptide often specifies a signal sequence needed for transfer across a membrane.

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 4.17** summarizes the relationship between the receptor for human LDL (plasma low density lipoprotein) 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* for its various functions. These modules are also used in different combinations in other proteins.

Exons tend to be fairly small and are 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.

This suggestion might explain another aspect of protein structure. It appears that the sites represented at exon–intron boundaries often are located at the surface of a protein. As modules are added to a protein, the connections—at least of the most recently added modules—could tend to lie at the surface.

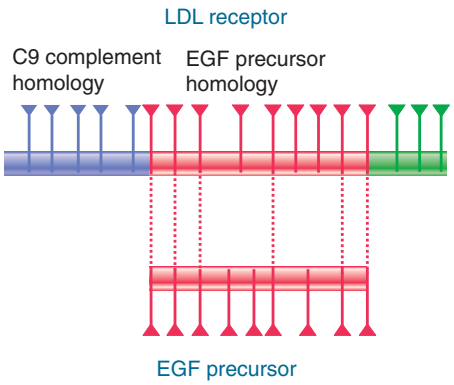


FIGURE 4.17 The LDL receptor gene consists of 18 exons, some of which are related to EGF precursor exons and some of which are related to the C9 blood complement gene. Triangles mark the positions of introns.

4.10 Members of a Gene Family Have a Common Organization

Key concepts

- A set of homologous genes should share common features that preceded their evolutionary separation.
- 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.
- Intron positions in the actin gene family are highly variable, which suggests that introns do not separate functional domains.

Many genes in a multicellular eukaryotic genome are related to others in the same genome, either *in series* (nonallelic) or *in parallel* (allelic). 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 divergence 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 (see the *Genome Evolution* chapter). In some cases, we can find genes that are more distantly related but that 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) origin with α and β globins. Leghemoglobins are oxygen-binding proteins present in legume plants; like myoglobin, they are monomeric and share a common origin with the other heme-binding proteins. Together, the globins, myoglobins, 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 in conserved positions (see Figure 4.4). 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 conserved 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 4.18**. The central intron of leghemoglobin separates two exons that together encode the sequence corresponding to the single central exon in globin; the functional heme-binding domain is split into two by an intron. Could the central exon of the globin gene have been derived by a fusion of two central exons in the ancestral gene? Or, is the single central exon the ancestral form? In this case, an intron must have been inserted into it early in plant evolution.

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 4.19** 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, because 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 extensive deletion or insertion of introns during evolution. A well characterized case is that of the actin genes. The common features of actin genes are an untranslated 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).

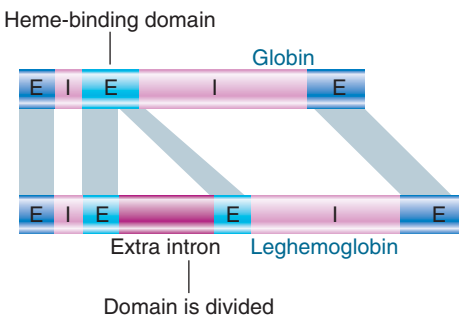


FIGURE 4.18 The exon structure of globin genes corresponds to protein function, but leghemoglobin has an extra intron in the central domain.

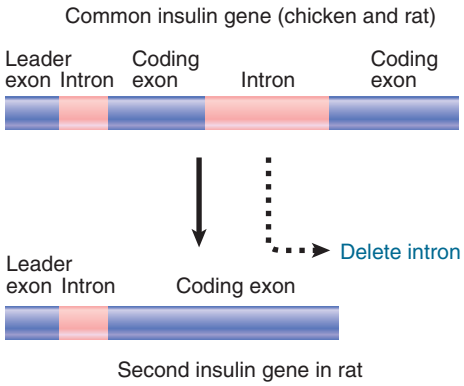


FIGURE 4.19 The rat insulin gene with one intron evolved by loss of an intron from an ancestor with two introns.

FIGURE 4.20 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 ancestral 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 ancestral 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 lineages.

Whether introns were present in actin genes early or late, there appears to have been no consistent influence from actin protein domains or subdomains as to where introns should locate. On the other hand, when exons are under negative selection (resulting in homology conservation), in-series recombination between members of an expanding gene family (that could cause a contraction in family size) would be decreased by intron diversification (resulting in loss of some homology), and introns would come to reside where this could best be achieved.

FIGURE 4.20 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.

Alleles would have similar exons and introns, so in-parallel interallelic recombination (as in meiosis) would be unimpaired until speciation occurred—a process that could be accompanied by intron relocations. The relationships between the intron locations among different species could then be used to construct a phylogenetic tree illustrating the evolution of the actin gene.

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 the 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 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 the *Genome Evolution* chapter).

4.11 There Are Many Forms of Information in DNA

Key concepts

- Genetic information includes not only that related to characters corresponding to the conventional phenotype, but also that related to characters (pressures) corresponding to the genome "phenotype."
- In certain contexts, the definition of the gene can be seen as reversed from "one gene–one protein" to "one protein–one gene."
- Positional information may be important in development.
- Sequences transferred "horizontally" from other species to the germline could land in introns or intergenic DNA and then transfer "vertically" through the generations. Some of these sequences may be involved in intracellular nonself-recognition.

The term *genetic information* can include all information that passes "vertically" through the germline, not just genic information. The word "gene" and its adjective "genic" have different meanings in different contexts, but in most circumstances there is little confusion when context is considered. In situations in which a sequence of DNA is responsible for production of one particular polypeptide, current usage regards the entire sequence of DNA—from the first point represented in the messenger RNA to the last point corresponding to its end—as comprising the "gene": exons, introns, and all.

When sequences encoding polypeptides overlap or have alternative forms of expression, we may reverse the usual description of the gene. Instead of saying "one gene–one polypeptide," we may describe the relationship as "one polypeptide–one gene." So we regard the

sequence involved in production of the polypeptide (including introns and exons) as constituting the gene, while recognizing that part of this same sequence also belongs to the gene of *another* polypeptide. This allows the use of descriptions such as "overlapping" or "alternative" genes.

We can now see how far we have come from the one gene–one enzyme hypothesis of the early part of the 20th century. The driving question at that time was the nature of the gene. It was thought that genes represented "ferments" (enzymes), but what was the fundamental nature of ferments? Once it was discovered that most genes encode proteins, the paradigm became fixed as the concept that every genetic unit functions through the synthesis of a particular protein. Either directly or indirectly, protein-encoding pressure was responsible for what we can now refer to as the *conventional phenotype*. We now recognize that genetic units encoding polypeptides may also include information corresponding to the *genome phenotype*, manifestations of which include *fold pressure*, *purine-loading (AG) pressure,* and *GC pressure*. There may be conflict between different pressures, such as competition for space in the gamete that will transfer genomic information to the next generation. For example, a protein might function most efficiently with the basic amino acid lysine (codon AAA) in a certain position, but GC pressure might require the substitution of another basic amino acid, such as arginine (codon CGG). Alternatively, fold pressure might require the corresponding nucleic acid to fold into a stem-loop structure where CCG would pair with the antiparallel arginine codon. A lysine codon in this position would disrupt the structure, so again a less efficient polypeptide would have to suffice.

The conventional phenotype, however, remains the central paradigm of molecular biology: A genic DNA sequence either directly encodes a particular polypeptide or is adjacent to the segment that actually encodes that polypeptide. How far does this paradigm take us beyond explaining the basic relationship between genes and proteins?

The development of multicellular organisms requires the use of different genes to generate the different cell phenotypes of each tissue. The expression of genes is determined by a regulatory network that takes the form of a cascade. Expression of the first set of genes at the start of embryonic development leads to expression of the genes involved in the next stage of development, which in turn leads to a further stage, and so on, until all the tissues of the adult are formed and functioning. The molecular nature of this regulatory network is still somewhat unknown, but we assume that it consists of genes that encode products (often protein, but sometimes RNA) that can influence the expression of other genes.

Although such a series of interactions is almost certainly the means by which the developmental program is executed, we can ask whether it is entirely sufficient. One specific question concerns the nature and role of *positional information*. We know that all parts of a fertilized egg are not equal; one of the features responsible for development of different tissue parts from different regions of the egg is location of information (presumably specific macromolecules) within the cell.

We do not fully understand how these particular regions are formed, though particular examples have been well studied (see the *mRNA Stability and Localization* chapter). We assume, however, that the existence of positional information in the egg leads to the differential expression of genes in the cells making up the tissues formed from these regions. This leads to the development of the adult organism, which in the next generation leads to the development of an egg with the appropriate positional information.

This possibility of positional information suggests that some information needed for development of the organism is contained in a form that we cannot directly attribute to a sequence of DNA (although the expression of particular sequences may be needed to perpetuate the positional information). Put in a more general way, we might ask the following: If we have the entire sequence of DNA comprising the genome of some organism and interpret it in terms of proteins and regulatory regions, could we in principle construct an organism (or even a single living cell) by controlled expression of the proper genes?

Once tissues and organs have developed, they not only have to be maintained, but also protected against potential pathogens. Groups of variable genes have diversified in the germline, and continue to diversify somatically, to allow multicellular organisms to (1) respond extracellularly by the synthesis of immunoglobulin antibodies directed against pathogens, and (2) "remember" past pathogens so that future responses will be faster and stronger (immunological memory; see the chapter titled

Somatic Recombination in the Immune System). Should it escape such *extracellular* defenses, though, the nucleic acid of a pathogenic virus could gain entry to cells and *intracellular* defenses would be needed.

We know that in bacteria infected by bacteriophages (see the chapter titled *Phage Strategies*), host defenses include rapid local or genome-wide transcription of DNA (which has been documented in eukaryotes in response to environmental insult or infection) to produce "antisense" transcripts that are capable of base-pairing with pathogen "sense" transcripts to form double-stranded RNAs. These RNAs then act as an alarm signal to trigger secondary defenses (see the example of bacterial CRISPRs discussed in the *Regulatory RNA* chapter). The host could store a "memory" of previous intracellular invaders by converting some pathogen transcripts into DNA through reverse transcription and inserting them into its genome in an inactive form for future rapid transcription of antisense RNAs in times of active infection by that pathogen. Thus, some pathogen nucleic acid might enter the germline "horizontally" (within a generation) and the parental memory of the pathogen could subsequently be transferred "vertically" to offspring. The diversity of some elements found within introns and extragenic DNA (see the chapter titled *Transposable Elements and Retroviruses*) could in part reflect such past pathogen attacks. There is recent evidence of such inherited antiviral immunity in several animal and plant species.

4.12 Summary

Most eukaryotic genomes contain genes that are interrupted by intron sequences. The proportion of interrupted genes is low in some fungi, but few genes are uninterrupted in multicellular eukaryotes. The size of a gene is determined primarily by the lengths of its introns. The range of gene sizes in mammals is generally from 1 to 100 kb, but there are some that are even larger.

Introns are found in all classes of eukaryotic genes, both those encoding protein products and those encoding independently functioning RNAs. The structure of an 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. In genes under negative selection pressure, intron sequences vary—and may even appear unrelated—although exon sequences remain closely related. This conservation of exons, which allows the conservation of important phenotypic characters, can be used to identify related genes in different species. In genes under positive selection pressure, however, exon sequences vary, although intron sequences can remain more similar. This conservation of introns relates to characters corresponding to the genome phenotype, such as fold pressure, which may relate to error correction in DNA.

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 and sometimes correspond to functional domains of those proteins. The idea that genes have been assembled by sequential addition of exons is consistent with the hypothesis that introns were present in the genes of ancestral organisms, thus facilitating the assembly process. Some of the relationships between homologous genes can be explained by loss of introns from the ancestral genes, with different introns being lost in different lines of descent.

References

4.1 Introduction

Reviews

Crick, F. (1979) Split genes and RNA splicing. *Science* 204, 264–271.

Harris, H. (1994) An RNA heresy in the fifties. *Trends Biochem. Sci.* 19, 303–305.

Hong, X., Schofield, D. G., and Lynch, M. (2006). Intron size, abundance, and distribution within untranslated regions of genes. *Mol. Biol. Evol.* 23, 2392–2404.

Research

Glover, D. M., and Hogness, D. S. (1977). A novel arrangement of the 8S and 28S sequences in a repeating unit of *D. melanogaster* rDNA. *Cell* 10, 167–176.

Scherrer, K., et al. (1970). Nuclear and cytoplasmic messenger-like RNAs and their relation to the active messenger RNA in polyribosomes of HeLa cells. *Cold Spring Harb. Symp. Quant. Biol.* 35, 539–554.

4.2 An Interrupted Gene Consists of Exons and Introns

Review

Forsdyke, D. R. (2011). Exons and introns. In: *Evolutionary Bioinformatics,* 2nd ed. Springer, New York, pp. 249–266. (*See also* http://post.queensu.ca/~forsdyke/introns.htm.)

4.3 Exon and Intron Base Compositions Differ

Reviews

Forsdyke, D. R., and Mortimer, J. R. (2000). Chargaff's legacy. *Gene* 261, 127–137. (*See* http://post.queensu.ca/~forsdyke/bioinfo2.htm.)

Forsdyke, D. R., and Bell, S. J. (2004). Purine-loading, stem-loops, and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics* 3, 3–8. (*See* http://post.queensu.ca/~forsdyke/bioinfo5.htm.)

Research

Babak, T., Blencowe, B. J., and Hughes, T. R. (2007). Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinf.* 8, article number 33.

Bechtel, J. M., et al. (2008). Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structure. *BMC Genomics* 9, article number 284.

Bultrini, E., et al. (2003). Pentamer vocabularies characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*. *Gene* 304, 183–192.

Ko, C. H., et al. (1998). U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize. *Plant Mol. Biol.* 36, 573–583.

Zhang, C., Li, W.-H., Krainer, A. R., and Zhang, M. Q. (2008). RNA landscape of evolution for optimal exon and intron discrimination. *Proc. Natl. Acad. Sci. USA* 105, 5797–5802.

4.4 Organization of Interrupted Genes May Be Conserved

Review

Fedoroff, N. V. (1979). On spacers. *Cell* 16, 697–710.

Research

Berget, S. M., Moore, C., and Sharp, P. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* 74, 3171–3175.

Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 mRNA. *Cell* 12, 1–8.

Jeffreys, A. J., and Flavell, R. A. (1977). The rabbit β-globin gene contains a large insert in the coding sequence. *Cell* 12, 1097–1108.

4.6 Exon Sequences Under Positive Selection Vary but Introns Are Conserved

Forsdyke, D. R. (1995). Conservation of stem-loop potential in introns of snake venom phospholipase A_2 genes: an application of FORS-D analysis. *Mol. Biol. Evol.* 12, 1157–1165.

Forsdyke, D. R. (1995). Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. *J. Mol. Evol.* 41, 1022–1037. (*See* http://post.queensu.ca/~forsdyke/hiv01.htm.)

Forsdyke, D. R. (1996). Stem-loop potential in MHC genes: a new way of evaluating positive Darwinian selection. *Immunogenetics* 43, 182–189.

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

Hawkins, J. D. (1988). A survey of intron and exon lengths. *Nucleic Acids Res.* 16, 9893–9905.

Naora, H., and Deacon, N. J. (1982). Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. *Proc. Natl. Acad. Sci. USA* 79, 6196–6200.

4.8 Some DNA Sequences Encode More Than One Polypeptide

Research

Pan, Q., et al. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40, 1413–1415.

Sultan, M., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960.

4.9 Some Exons Can Be Equated with Protein Functional Domains

Reviews

Blake, C. C. (1985). Exons and the evolution of proteins. *Int. Rev. Cytol.* 93, 149–185.

Doolittle, R. F. (1985). The genealogy of some recently evolved vertebrate proteins. *Trends Biochem. Sci.* 10, 233–237.

4.10 Members of a Gene Family Have a Common Organization

Review

Dixon, B., and Pohajdek, B. (1992). Did the ancestral globin gene of plants and animals contain only two introns? *Trends Biochem. Sci.* 17, 486–488.

Research

Matsuo, K., et al. (1994). Short introns interrupting the Oct-2 POU domain may prevent recombination between POU family

members without interfering with potential POU domain "shuffling" in evolution. *Biol. Chem. Hopp-Seyler* 375, 675–683.
Weber, K., and Kabsch, W. (1994). Intron positions in actin genes seem unrelated to the secondary structure of the protein. *EMBO. J.* 13, 1280–1286.

4.11 There Are Many Forms of Information in DNA

Reviews

Barrangou, R., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712.
Bernardi, G., and Bernardi, G. (1986). Compositional constraints and genome evolution. *J. Mol. Evol.* 24, 1–11.
Forsdyke, D. R. (2011). *Evolutionary Bioinformatics*, 2nd ed. Springer, New York, 509 pp.
Forsdyke, D. R., Madill, C. A., and Smith, S. D. (2002). Immunity as a function of the unicellular state: implications of emerging genomic data. *Trends Immunol.* 23, 575–579. (*See* http://post.queensu.ca/~forsdyke/theorimm.htm.)
Jeffares, D. C., Penkett, C. J., and Bähler, J. (2008). Rapidly regulated genes are intron poor. *Trends in Genetics* 24, 375–378.

Research

Bertsch, C., Beuve, M., Dolja, V. V., Wirth, M., Pelsy, F., Herrbach, E., and O. Lemaire. (2009). Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biology Direct* 4, 21.
Flegel, T. W. (2009). Hypothesis for heritable, antiviral immunity in crustaceans and insects. *Biology Direct* 4, 32.
Saleh, M.-C., Tassetto, M., van Rij, R. P., Goic. B., Gausson, V. Berry, B., Jacquier, C., Antoniewski, C., and R. Andino. (2009). Antiviral immunity in *Drosophila* requires systemic RNA interference spread. *Nature* 458, 346-350.