THE INTERRUPTED GENE

# 1.2.1 Introduction

## Key Terms

An **exon** is any segment of an interrupted gene that is represented in the mature RNA product.

An **intron (Intervening sequence)** is a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it.

A **transcript** is the RNA product produced by copying one strand of DNA. It may require processing to generate a mature RNA.

**RNA splicing** is the process of excising the sequences in RNA that correspond to introns, so that the sequences corresponding to exons are connected into a continuous mRNA.

A **structural gene** codes for any RNA or protein product other than a regulator.

## Key Concepts

- Eukaryotic genomes contain interrupted genes in which exons (represented in the final RNA product) alternate with introns (removed from the initial transcript).
- The exon sequences occur 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.

Until eukaryotic genes were characterized by molecular mapping, we assumed that they would have the same organization as prokaryotic genes. We expected the gene to consist of a length of DNA that is colinear with the protein. But a comparison between the structure of DNA and the corresponding mRNA shows a discrepancy in many cases. The mRNA always includes a nucleotide sequence that corresponds exactly with the protein product according to the rules of the genetic code. *But the gene includes additional sequences that lie within the coding region, interrupting the sequence that represents the protein.* (For a description of the discovery see *Great Experiments 4.1 The discovery of RNA splicing* and *Great Experiments 4.2 The discovery of split genes and RNA splicing*.)

The sequences of DNA comprising an interrupted gene are divided into the two categories depicted in **Figure 2.1**:

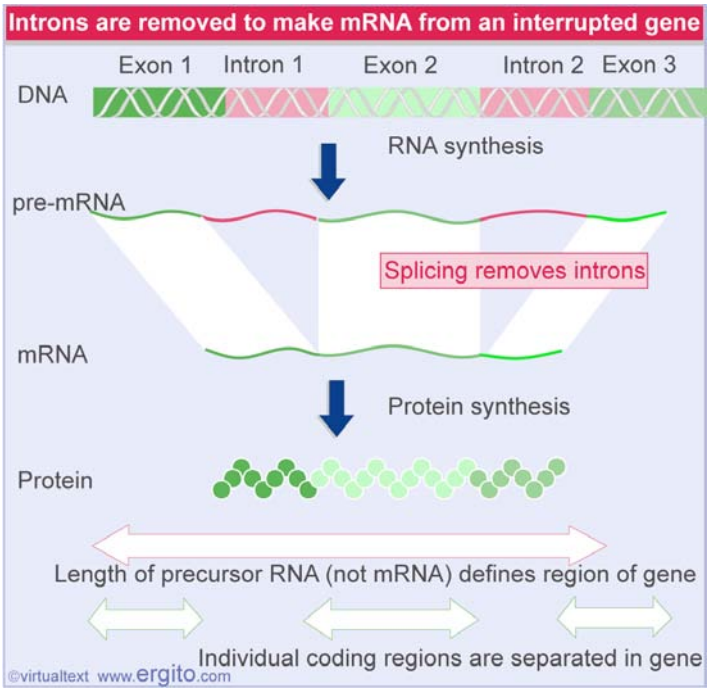


**Figure 2.1** Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.

- The **exons** are the sequences represented in the mature RNA. By definition, a gene starts and ends with exons, corresponding to the 5′ and 3′ ends of the RNA.
- The **introns** are the intervening sequences that are removed when the primary transcript is processed to give the mature RNA.

The expression of interrupted genes requires an additional step that does not occur for uninterrupted genes. The DNA gives rise to an RNA copy (a **transcript**) that exactly represents the genome sequence. But this RNA is only a precursor; it cannot be used for producing protein. 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**. It involves a precise deletion of an intron from the primary transcript; the ends of the RNA on either side are joined to form a covalently intact molecule (see *Molecular Biology 5.24 RNA splicing and processing*).

The **structural gene** comprises the region in the genome between points corresponding to the 5′ and 3′ terminal bases of mature mRNA. We know that transcription starts at the 5′ end of the mRNA, but usually it extends beyond the 3′ end, which is generated by cleavage of the RNA (see *Molecular Biology 5.24.19 The 3′ ends of mRNAs are generated by cleavage and polyadenylation*). The gene is considered to include the regulatory regions on both sides of the gene that are required for initiating and (sometimes) terminating gene expression.

*This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.2.1*

THE INTERRUPTED GENE

# 1.2.2 An interrupted gene consists of exons and introns

## Key Concepts

- Introns are removed by the process of RNA splicing, which occurs only in *cis* on an individual RNA molecule.
- Only mutations in exons can affect protein sequence, but mutations in introns can affect processing of the RNA and therefore prevent production of protein.

How does the existence of introns change our view of the gene? Following splicing, the exons are always joined together in the same order in which they lie in DNA. So the colinearity of gene and protein is maintained between the individual exons and the corresponding parts of the protein chain. **Figure 2.2** shows that the *order* of mutations in the gene remains the same as the order of amino acid replacements in the protein. But the *distances* in the gene do not correspond at all with the distances in the protein. Genetic distances, as seen on a recombination map, have no relationship to the distances between the corresponding points in the protein. The length of the gene is defined by the length of the initial (precursor) RNA instead of by the length of the messenger RNA.

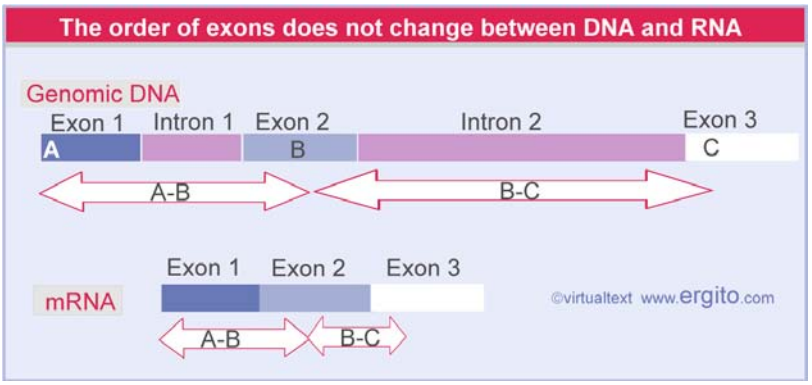


**Figure 2.2** Exons remain in the same order in mRNA as in DNA, but distances along the gene do not correspond to distances along the mRNA or protein products. The distance from A-B in the gene is smaller than the distance from B-C; but the distance from A-B in the mRNA (and protein) is greater than the distance from B-C.

All the exons are represented on the same molecule of RNA, and their splicing together occurs only as an *intra*molecular reaction. There is usually no joining of exons carried by *different* RNA molecules, so the mechanism excludes any splicing together of sequences representing different alleles. Mutations located in different exons of a gene cannot complement one another; thus they continue to be defined as members of the same complementation group.

Mutations that directly affect the sequence of a protein must lie in exons. What are

the effects of mutations in the introns? Since the introns are not part of the messenger RNA, mutations in them cannot directly affect protein structure. However, they can prevent the production of the messenger RNA – for example, by inhibiting the splicing together of exons. A mutation of this sort acts only on the allele that carries it. So it fails to complement any other mutation in that allele, and constitutes 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 an aberrant site. The most common result is to introduce a termination codon that results in truncation of the protein sequence. About 15% of the point mutations that cause human diseases are caused by disruption of splicing (for review see 3645).

Eukaryotic genes are not necessarily interrupted. Some correspond directly with the protein product in the same manner as prokaryotic genes. In yeast, most genes are uninterrupted. In higher eukaryotes, most genes are interrupted; and the introns are usually much longer than exons, creating genes that are very much larger than their coding regions (for review see 8).

*Last updated on 3-10-2003*

## Reviews

8. Breathnach, R. and Chambon, P. (1981). *Organization and expression of eukaryotic split genes coding for proteins.* Annu. Rev. Biochem. 50, 349-383.

3645. Faustino, N. A. and Cooper, T. A. (2003). *Pre-mRNA splicing and human disease.* Genes Dev. 17, 419-437.

THE INTERRUPTED GENE

# 1.2.3 Restriction endonucleases are a key tool in mapping DNA

## Key Terms

**Restriction endonucleases** recognize specific short sequences of DNA and cleave the duplex (sometimes at target site, sometimes elsewhere, depending on type).

A **restriction map** is a linear array of sites on DNA cleaved by various restriction enzymes.

A **kilobase (kb)** is a measure of length and may be used to refer to DNA (1000 base pairs) or to RNA (1000 bases).

A **megabase (Mb)** is 1 million base pairs of DNA.

## Key Concepts

- Restriction endonucleases can be used to cleave DNA into defined fragments.
- A map can be generated by using the overlaps between the fragments generated by different restriction enzymes.

The characterization of eukaryotic genes was made possible by the development of techniques for physically mapping DNA. The techniques can be extended to (single-stranded) RNA by making a (double-stranded) DNA copy of the RNA. A physical map of any DNA molecule can be obtained by breaking it at defined points whose distance apart can be accurately determined. Specific breaks are made possible by the ability of **restriction endonucleases** to recognize rather short sequences of double-stranded DNA as targets for cleavage.

Each restriction enzyme has a particular target in duplex DNA, usually a specific sequence of 4-6 base pairs. The enzyme cuts the DNA at every point at which its target sequence occurs. Different restriction enzymes have different target sequences, and a large range of these activities (obtained from a wide variety of bacteria) now is available.

A **restriction map** represents a linear sequence of the sites at which particular restriction enzymes find their targets. Distance along such maps is measured directly in base pairs (abbreviated *bp*) for short distances; longer distances are given in **kb**, corresponding to kilobase ($10^3$) pairs in DNA or to kilobases in RNA. At the level of the chromosome, a map is described in megabase pairs (1 **Mb** = $10^6$ bp).

When a DNA molecule is cut with a suitable restriction enzyme, it is cleaved into distinct fragments. These fragments can be separated on the basis of their size by gel electrophoresis, as shown in **Figure 2.3**. The cleaved DNA is placed on top of a gel made of agarose or polyacrylamide. When an electric current is passed through the gel, each fragment moves down at a rate that is inversely related to the log of its molecular weight. This movement produces a series of bands. Each band corresponds

to a fragment of particular size, decreasing down the gel.

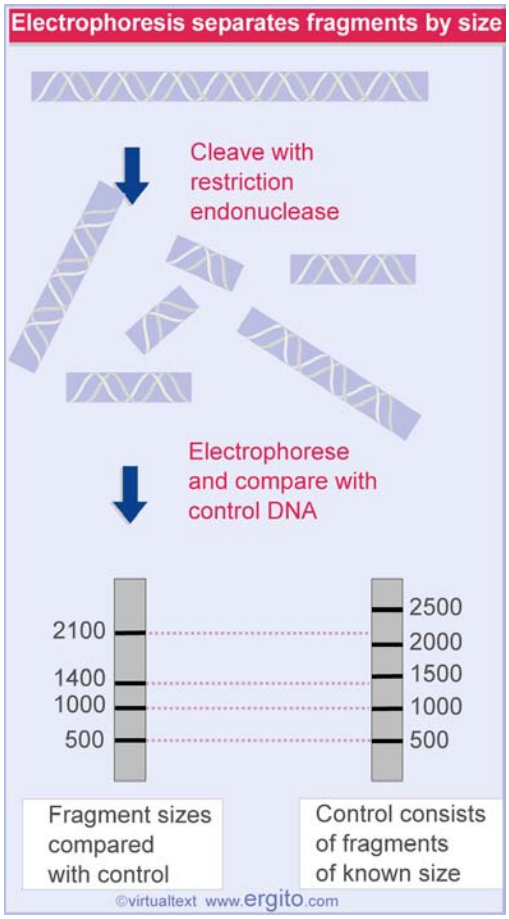


**Figure 2.3** Fragments generated by cleaving DNA with a restriction endonuclease can be separated according to their sizes.

By analyzing the restriction fragments of DNA, we can generate a map of the original molecule in the form shown in **Figure 2.4**. The method is explained in detail in *Molecular Biology Supplement 32.11 Restriction mapping*. The map shows the positions at which particular restriction enzymes cut DNA; the distances between the sites of cutting are measured in base pairs. *So the DNA is divided into a series of regions of defined lengths that lie between sites recognized by the restriction enzymes.* An important feature is that a restriction map can be obtained for any sequence of DNA, *irrespective of whether mutations have been identified in it,* or, indeed, whether we have any knowledge of its function (392) (for review see 6; 10).

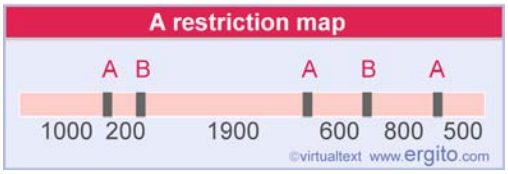


**Figure 2.4** A restriction map is a linear sequence of sites separated by defined distances on DNA. The map identifies the sites cleaved by enzymes A and B, as defined by the individual fragments produced by the single and double digests.

## Reviews

6. Nathans, D. and Smith, H. O. (1975). *Restriction endonucleases in the analysis and restructuring of DNA molecules.* Annu. Rev. Biochem. 44, 273-293.

10. Wu, R. (1978). *DNA sequence analysis.* Annu. Rev. Biochem. 47, 607-734.

## References

392. Danna, K. J., Sack, G. H., and Nathans, D. (1973). *Studies of SV40 DNA VII A cleavage map of the SV40 genome*. J. Mol. Biol. 78, 363-376.

THE INTERRUPTED GENE

# 1.2.4 Organization of interrupted genes may be conserved

## 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, but the ultimate definition is based on comparison of sequences.
- The positions of introns are usually conserved when homologous genes are compared between different organisms, but the lengths of the corresponding introns may vary greatly.
- Introns usually do not code for proteins.

When a gene is uninterrupted, the restriction map of its DNA corresponds exactly with the map of its mRNA.

When a gene possesses an intron, the map at each end of the gene corresponds with the map at each end of the message sequence. But within the gene, the maps diverge, because additional regions are found in the gene, but are not represented in the message. Each such region corresponds to an intron. The example of **Figure 2.5** compares the restriction maps of a β-globin gene and mRNA. There are two introns. Each intron contains a series of restriction sites that are absent from the cDNA. But the pattern of restriction sites in the exons is the same in both the cDNA and the gene (387; 388; 389; 390; 391).

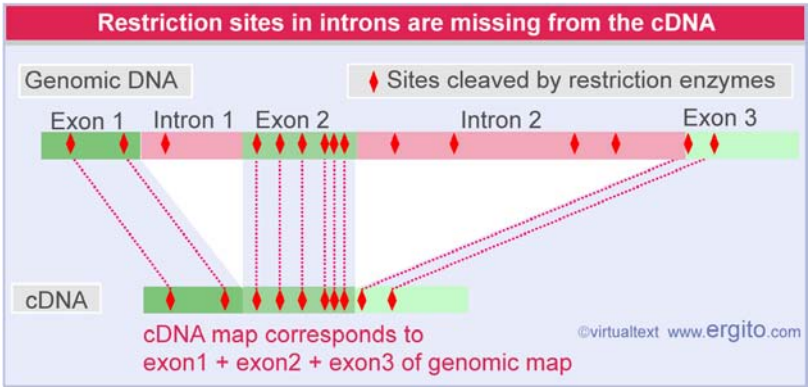


**Figure 2.5** Comparison of the restriction maps of cDNA and genomic DNA for mouse β -globin shows that the gene has two introns that are not present in the cDNA. The exons can be aligned exactly between cDNA and gene.

Ultimately a comparison of the nucleotide sequences of the genomic and mRNA sequences precisely defines the introns. As indicated in **Figure 2.6**, an intron usually has no open reading frame. An intact reading frame is created in the mRNA sequence by the removal of the introns.

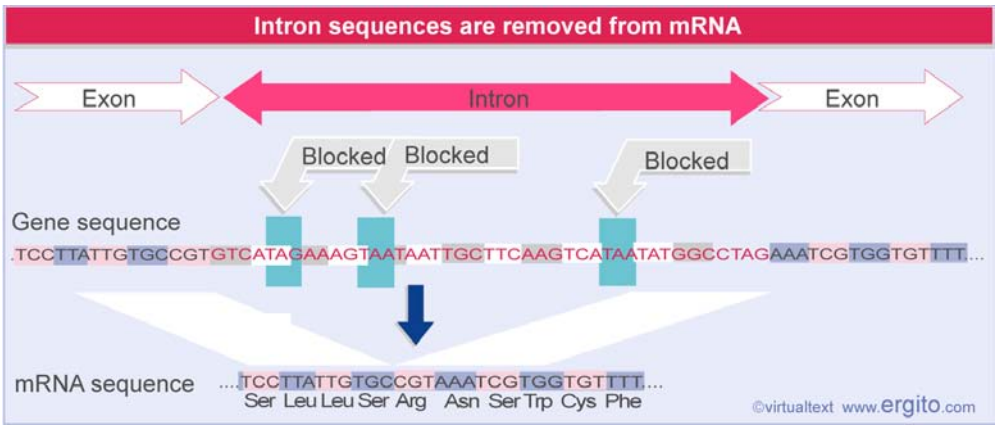


**Figure 2.6** An intron is a sequence present in the gene but absent from the mRNA (here shown in terms of the cDNA sequence). The reading frame is indicated by the alternating open and shaded blocks; note that all three possible reading frames are blocked by termination codons in the intron.

The structures of eukaryotic genes show extensive variation. Some genes are uninterrupted, so that the genomic sequence is colinear with that of the mRNA. Most higher eukaryotic genes are interrupted, but the introns vary enormously in both number and size.

All classes of genes may be interrupted: nuclear genes coding for proteins, nucleolar genes coding for rRNA, and genes coding for tRNA. Interruptions also are found in mitochondrial genes in lower eukaryotes, and in chloroplast genes. Interrupted genes do not appear to be excluded from any class of eukaryotes, and have been found in bacteria and bacteriophages, although they are extremely rare in prokaryotic genomes.

Some interrupted genes possess only one or a few introns. The globin genes provide an extensively studied example (see *Molecular Biology 1.2.11 The members of a gene family have a common organization*). The two general types of globin gene, α and β, share a common type of structure. The consistency of the organization of mammalian globin genes is evident from the structure of the "generic" globin gene summarized in **Figure 2.7**.

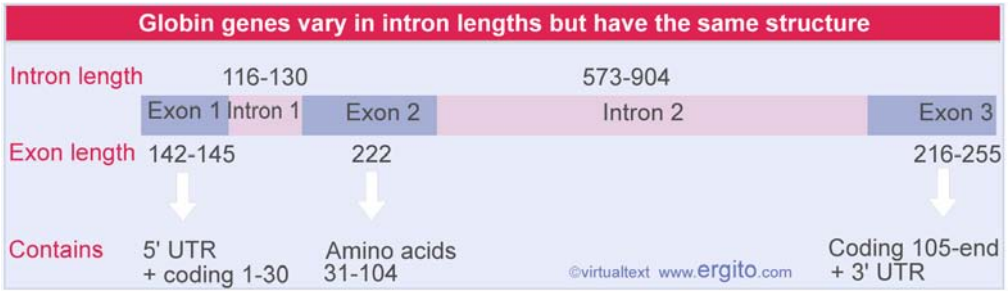


**Figure 2.7** All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian β -globin genes.

Interruptions occur at homologous positions (relative to the coding sequence) in all known active globin genes, including those of mammals, birds, and frogs. The first intron is always fairly short, and the second usually is longer, but the actual lengths

can vary. Most of the variation in overall lengths between different globin genes results from the variation in the second intron. In the mouse, the second intron in the α-globin gene is only 150 bp long, so the overall length of the gene is 850 bp, compared with the major β-globin gene where the intron length of 585 bp gives the gene a total length of 1382 bp. The variation in length of the genes is much greater than the range of lengths of the mRNAs ( α-globin mRNA = 585 bases, β-globin mRNA = 620 bases).

The example of DHFR, a somewhat larger gene, is shown in **Figure 2.8**. The mammalian DHFR (dihydrofolate reductase) gene is organized into 6 exons that correspond to the 2000 base mRNA. But they extend over a much greater length of DNA because the introns are very long. In three mammals the exons remain essentially the same, and the relative positions of the introns are unaltered, but the lengths of individual introns vary extensively, resulting in a variation in the length of the gene from 25-31 kb.

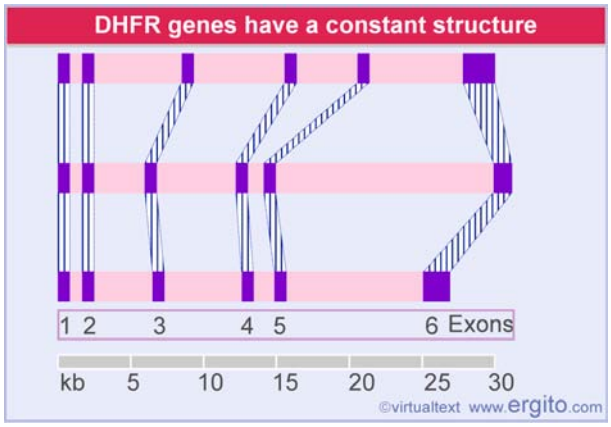


**Figure 2.8** Mammalian genes for DHFR have the same relative organization of rather short exons and very long introns, but vary extensively in the lengths of corresponding introns.

The globin and DHFR genes present examples of a general phenomenon: *genes that are related by evolution have related organizations, with conservation of the positions of (at least some) of the introns. Variations in the lengths of the genes are primarily determined by the lengths of the introns.*

## References


387. Wenskink, P. et al. (1974). *A system for mapping DNA sequences in the chromosomes ofD. melanogaster.* Cell 3, 315-325.

388. 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.

389. 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.

390. 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.

391. 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.

THE INTERRUPTED GENE

# 1.2.5 Exon sequences 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 well related.
- Introns evolve much more rapidly than exons because of the lack of selective pressure to produce a protein with a useful sequence.

Is a structural gene unique in its genome? The answer can be ambiguous. The entire length of the gene is unique as such, but its exons often are 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 the introns. In an extreme case, the exons of two genes may code for the same protein sequence, but the introns may be different. This implies that the two genes originated by a duplication of some common ancestral gene. Then differences accumulated between the copies, but they were restricted in the exons by the need to code for protein functions.

As we see later when we consider the evolution of the gene, exons can be considered as basic building blocks that are assembled in various combinations. A gene may have some exons that are related to exons of another gene, but the other exons may be unrelated. Usually the introns are not related at all in such cases. Such genes may arise by duplication and translocation of individual exons.

The relationship between two genes can be plotted in the form of the dot matrix comparison of **Figure 2.9**. A dot is placed to indicate each position at which the same sequence is found in each gene. The dots form a line at an angle of 45° if two sequences are identical. The line is broken by regions that lack similarity, and it is displaced laterally or vertically by deletions or insertions in one sequence relative to the other.

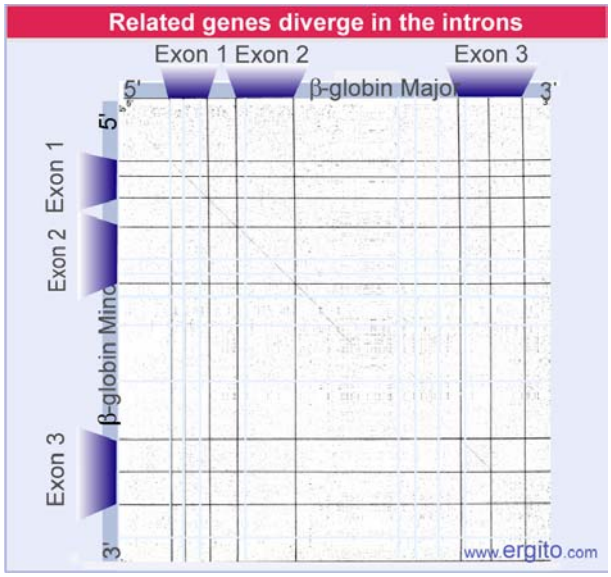


**Figure 2.9** The sequences of the mouse $\alpha^{maj}$ and $\alpha^{min}$ globin genes are closely related in coding regions, but differ in the flanking regions and large intron. Data kindly provided by Philip Leder.

When the two β-globin genes of the mouse are compared, such a line extends through the three exons and through the small intron. The line peters out in the flanking regions and in the large intron. This is a typical pattern, in which coding sequences are well related, the relationship can extend beyond the boundaries of the exons, but it is lost in longer introns and the regions on either side of the gene.

The overall degree of divergence between two exons is related to the differences between the proteins. It is caused mostly by base substitutions. In the translated regions, the exons are under the constraint of needing to code for amino acid sequences, so they are limited in their potential to change sequence. Many of the changes do not affect codon meanings, because they change one codon into another that represents the same amino acid. Changes occur more freely in nontranslated regions (corresponding to the 5′ leader and 3′ trailer of the mRNA).

In corresponding introns, the pattern of divergence involves both changes in size (due to deletions and insertions) and base substitutions. Introns evolve much more rapidly than exons. When a gene is compared in different species, sometimes the exons are homologous, while the introns have diverged so much that corresponding sequences cannot be recognized.

Mutations occur at the same rate in both exons and introns, but are removed more effectively from the exons by adverse selection. However, in the absence of the constraints imposed by a coding function, an intron is able quite freely to accumulate point substitutions and other changes. These changes imply that the intron does not have a sequence-specific function. Whether its presence is at all necessary for gene function is not clear.

*This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.2.5*

THE INTERRUPTED GENE

# 1.2.6 Genes can be isolated by the conservation of exons

## Key Terms

A **zoo blot** describes the use of Southern blotting to test the ability of a DNA probe from one species to hybridize with the DNA from the genomes of a variety of other species.

**Exon trapping** inserts a genomic fragment into a vector whose function depends on the provision of splicing junctions by the fragment.

## Key Concepts

- Conservation of exons can be used as the basis for identifying coding regions by identifying fragments whose sequences are present in multiple organisms.

Some major approaches to identifying genes are based on the contrast between the conservation of exons and the variation of introns. In a region containing a gene whose function has been conserved among a range of species, the sequence representing the protein should have two distinctive properties:

- it must have an open reading frame;
- and it is likely to have a related sequence in other species.

These features can be used to isolate genes.

Suppose we know by genetic data that a particular genetic trait is located in a given chromosomal region. If we lack knowledge about the nature of the gene product, how are we to identify the gene in a region that may be (for example) >1 Mb?

A heroic approach that has proved successful with some genes of medical importance is to screen relatively short fragments from the region for the two properties expected of a conserved gene. First we seek to identify fragments that cross-hybridize with the genomes of other species. Then we examine these fragments for open reading frames.

The first criterion is applied by performing a **zoo blot**. We use short fragments from the region as (radioactive) probes to test for related DNA from a variety of species by Southern blotting. If we find hybridizing fragments in several species related to that of the probe – the probe is usually human – the probe becomes a candidate for an exon of the gene.

The candidates are sequenced, and if they contain open reading frames, are used to isolate surrounding genomic regions. If these appear to be part of an exon, we may

then use them to identify the entire gene, to isolate the corresponding cDNA or mRNA, and ultimately to identify the protein.

This approach is especially important when the target gene is spread out because it has many large introns. This proved to be the case with Duchenne muscular dystrophy (DMD), a degenerative disorder of muscle, which is X-linked and affects 1 in 3500 of human male births. The steps in identifying the gene are summarized in **Figure 2.10**.

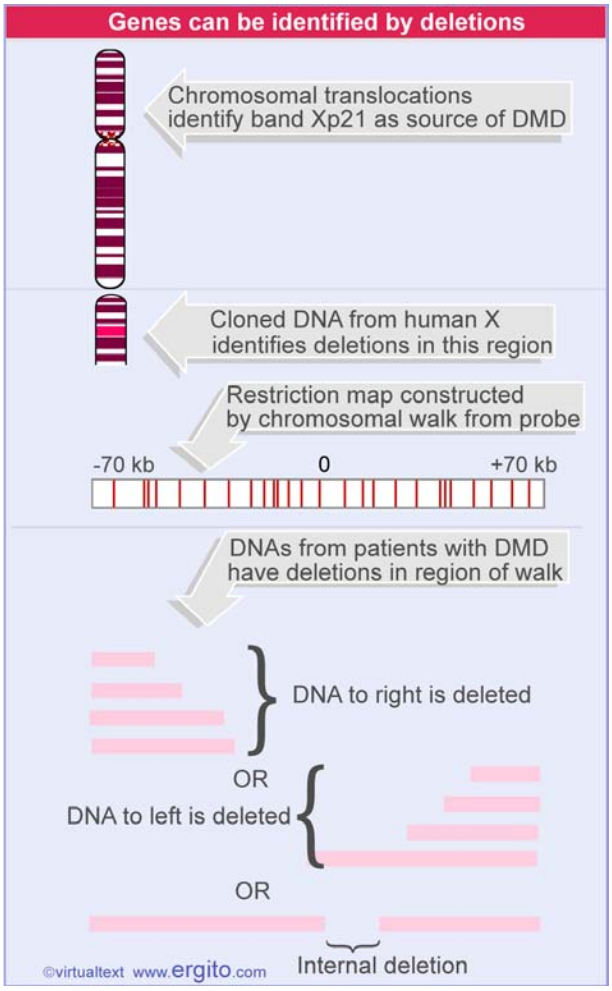


**Figure 2.10** The gene involved in Duchenne muscular dystrophy was tracked down by chromosome mapping and walking to a region in which deletions can be identified with the occurrence of the disease.

Linkage analysis localized the DMD locus to chromosomal band Xp21. Patients with the disease often have chromosomal rearrangements involving this band. By comparing the ability of X-linked DNA probes to hybridize with DNA from patients and with normal DNA, cloned fragments were obtained that correspond to the region that was rearranged or deleted in patients' DNA.

Once some DNA in the general vicinity of the target gene has been obtained, it is

possible to "walk" along the chromosome until the gene is reached (see *Molecular Biology Supplement 32.12 Genome mapping*). A chromosomal walk was used to construct a restriction map of the region on either side of the probe, covering a region of >100 kb. Analysis of the DNA from a series of patients identified large deletions in this region, extending in either direction. The most telling deletion is one contained entirely within the region, since this delineates a segment that must be important in gene function and indicates that the gene, or at least part of it, lies in this region (3032; 3033).

Having now come into the region of the gene, we need to identify its exons and introns. A zoo blot identified fragments that cross-hybridize with the mouse X chromosome and with other mammalian DNAs. As summarized in **Figure 2.11**, these were scrutinized for open reading frames and the sequences typical of exon-intron junctions. Fragments that met these criteria were used as probes to identify homologous sequences in a cDNA library prepared from muscle mRNA.

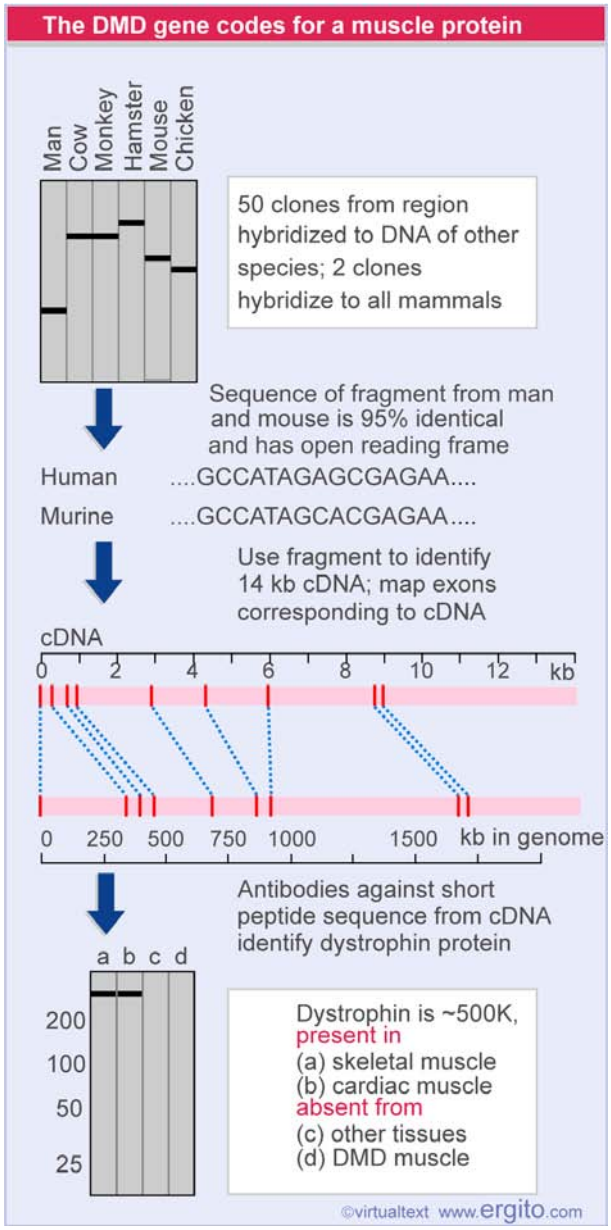


**Figure 2.11** The Duchenne muscular dystrophy gene was characterized by zoo blotting, cDNA hybridization, genomic hybridization, and identification of the protein.

The cDNA corresponding to the gene identifies an unusually large mRNA, ~14 kb. Hybridization back to the genome shows that the mRNA is represented in >60 exons, which are spread over ~2000 kb of DNA. This makes DMD the longest gene identified; in fact, it is 10× longer than any other known gene (3034; 3035).

The gene codes for a protein of ~500 kD, called dystrophin, which is a component of muscle, present in rather low amounts. All patients with the disease have deletions at this locus, and lack (or have defective) dystrophin.

Muscle also has the distinction of having the largest known protein, titin, with almost 27,000 amino acids. Its gene has the largest number of exons (178) and the longest single exon in the human genome (17,000 bp).

Another technique that allows genomic fragments to be scanned rapidly for the presence of exons is called **exon trapping** (3030). **Figure 2.12** shows that it starts with a vector that contains a strong promoter, and has a single intron between two exons. When this vector is transfected into cells, its transcription generates large amounts of an RNA containing the sequences of the two exons. A restriction cloning site lies within the intron, and is used to insert genomic fragments from a region of interest. If a fragment does not contain an exon, there is no change in the splicing pattern, and the RNA contains only the same sequences as the parental vector. But if the genomic fragment contains an exon flanked by two partial intron sequences, the splicing sites on either side of this exon are recognized, and the sequence of the exon is inserted into the RNA between the two exons of the vector. This can be detected readily by reverse transcribing the cytoplasmic RNA into cDNA, and using PCR to amplify the sequences between the two exons of the vector. So the appearance in the amplified population of sequences from the genomic fragment indicates that an exon has been trapped. Because introns are usually large and exons are small in animal cells, there is a high probability that a random piece of genomic DNA will contain the required structure of an exon surrounded by partial introns. In fact, exon trapping may mimic the events that have occurred naturally during evolution of genes (see *Molecular Biology 1.2.9 How did interrupted genes evolve?*).

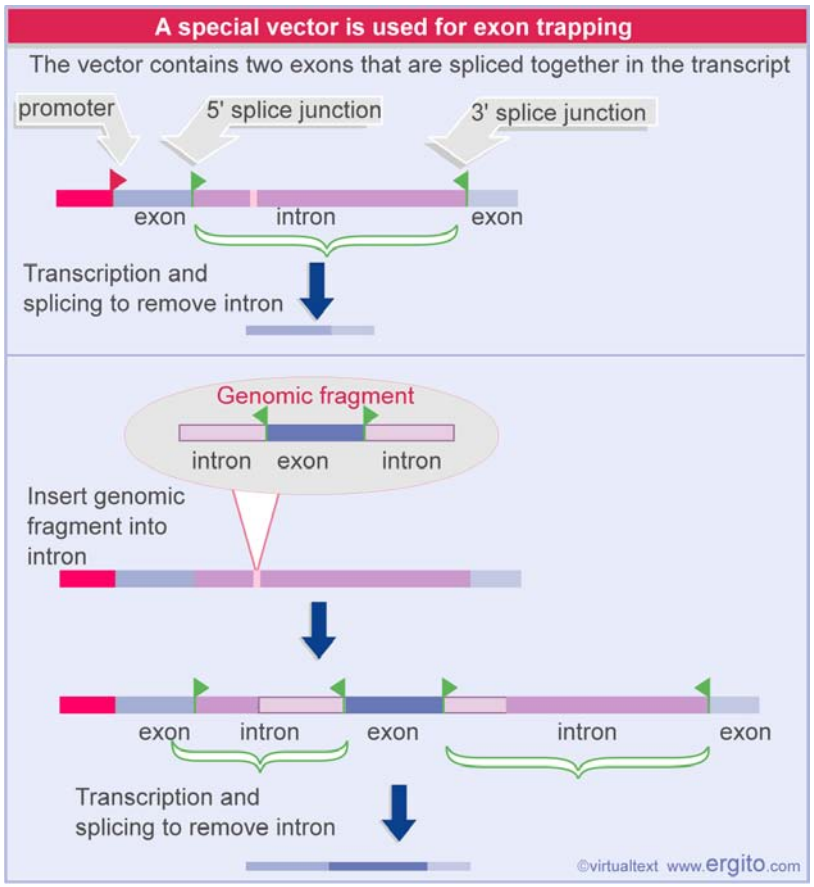


**Figure 2.12** A special splicing vector is used for exon trapping. If an exon is present in the genomic fragment, its sequence will be recovered in the cytoplasmic RNA, but if the genomic fragment consists solely of sequences from within intron, splicing does not occur, and the mRNA is not exported to the cytoplasm.

*Last updated on 2-16-2001*

## References

3030. Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A., and Housman, D. E. (1991). *Exon amplification: a strategy to isolate mammalian genes based on RNA splicing.* Proc. Natl. Acad. Sci. USA 88, 4005-4009.

3032. Kunkel, L. M., Monaco, A. P., Middlesworth, W., Ochs, H. D., and Latt, S. A. (1985). *Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion.* Proc. Natl. Acad. Sci. USA 82, 4778-4782.

3033. Monaco, A.P., Bertelson, C. J., Middlesworth, W., Colletti, C. A., Aldridge, J., Fischbeck, K. H., Bartlett, R., Pericak-Vance, M. A., Roses, A. D., and Kunkel, L. M. (1985). *Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA segment.* Nature 316, 842-845.

3034. van Ommen, G. J., Verkerk, J. M., Hofker, M. H., Monaco, A. P., Kunkel, L. M., Ray, P., Worton, R., Wieringa, B., Bakker, E., and Pearson, P. L. (1986). *A physical map of 4 million bp around the Duchenne muscular dystrophy gene on the human X-chromosome.* Cell 47, 499-504.

3035. Koenig, M., Hoffman, E. P., Bertelson, C. J., Monaco, A. P., Feener, C., and Kunkel, L. M. (1987). *Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals.* Cell 50, 509-517.

THE INTERRUPTED GENE

# 1.2.7 Genes show a wide distribution of sizes

## Key Concepts

- Most genes are uninterrupted in yeasts, but are interrupted in higher eukaryotes.
- Exons are usually short, typically coding for <100 amino acids.
- Introns are short in lower eukaryotes, but range up to several 10s of kb in length in higher eukaryotes.
- The overall length of a gene is determined largely by its introns.

**Figure 2.13** shows the overall organization of genes in yeasts, insects, and mammals. In *S. cerevisiae*, the great majority of genes (>96%) are not interrupted, and those that have exons usually remain reasonably compact. There are virtually no *S. cerevisiae* genes with more than 4 exons.

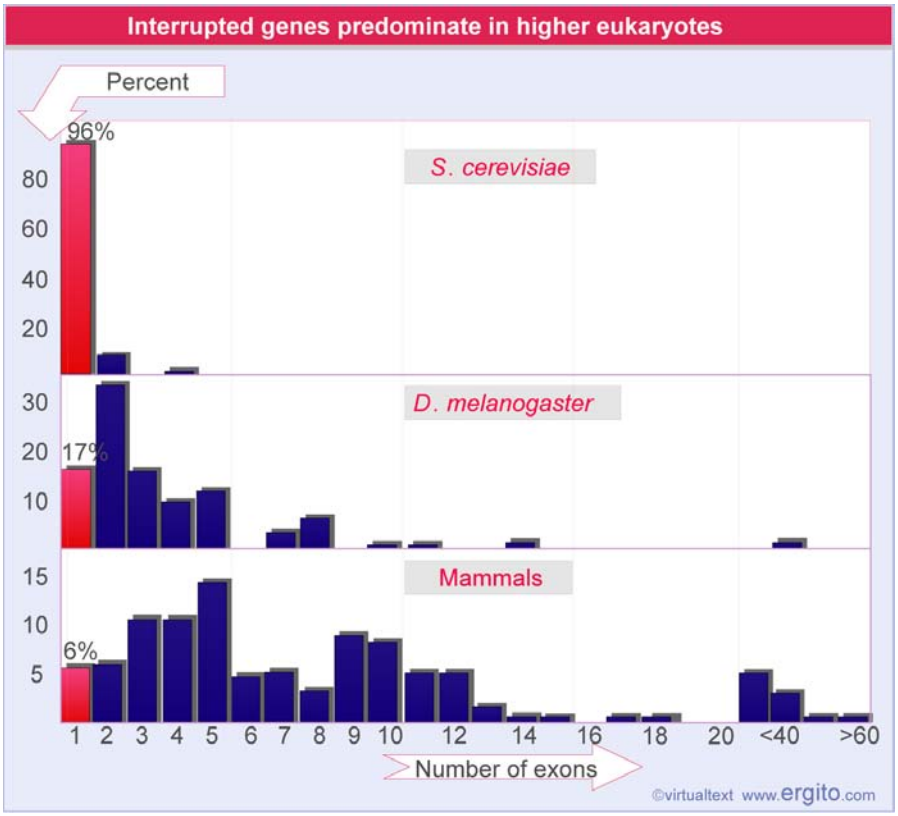


**Figure 2.13** Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only 1 exon, and are totaled in the leftmost column.)

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 several 10s of exons. ~50% of mammalian genes have >10 introns.

Examining the consequences of this type of organization for the overall size of the gene, we see in **Figure 2.14** that there is a striking difference between yeast and the higher eukaryotes. The average yeast gene is 1.4 kb long, and very few are longer than 5 kb. The predominance of interrupted genes in high eukaryotes, however, means that the gene can be much larger than the unit that codes for protein. Relatively few genes in flies or mammals are shorter than 2 kb, and many have lengths between 5 kb and 100 kb. The average human gene is 27 kb long (see **Figure 3.22**).

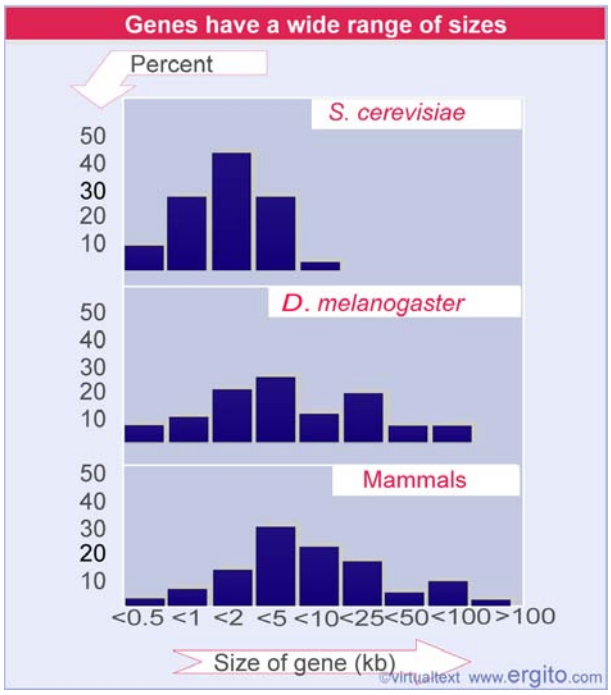


**Figure 2.14** Yeast genes are small, but genes in flies and mammals have a dispersed distribution extending to very large sizes.

The switch from largely uninterrupted to largely interrupted genes occurs in the lower eukaryotes. In fungi (excepting the yeasts), the majority of genes are interrupted, but they have a relatively small number of exons (<6) and are fairly short (<5 kb). The switch to long genes occurs within the higher eukaryotes, and genes become significantly larger in the insects. With this increase in the length of the gene, the relationship between genome complexity and organism complexity is lost (see **Figure 3.5**).

As genome size increases, the tendency is for introns to become rather large, while exons remain quite small.

**Figure 2.15** shows that the exons coding for stretches of protein tend to be fairly small. In higher eukaryotes, the average exon codes for ~50 amino acids, and the general distribution fits well with the idea that genes have evolved by the slow addition of units that code for small, individual domains of proteins (see *Molecular Biology 1.2.9 How did interrupted genes evolve?*). There is no very significant difference in the sizes of exons in different types of higher eukaryotes, although the distribution is more compact in vertebrates where there are few exons longer than 200 bp. In yeast, there are some longer exons that represent uninterrupted genes where the coding sequence is intact. There is a tendency for exons coding for untranslated 5′ and 3′ regions to be longer than those that code for proteins.

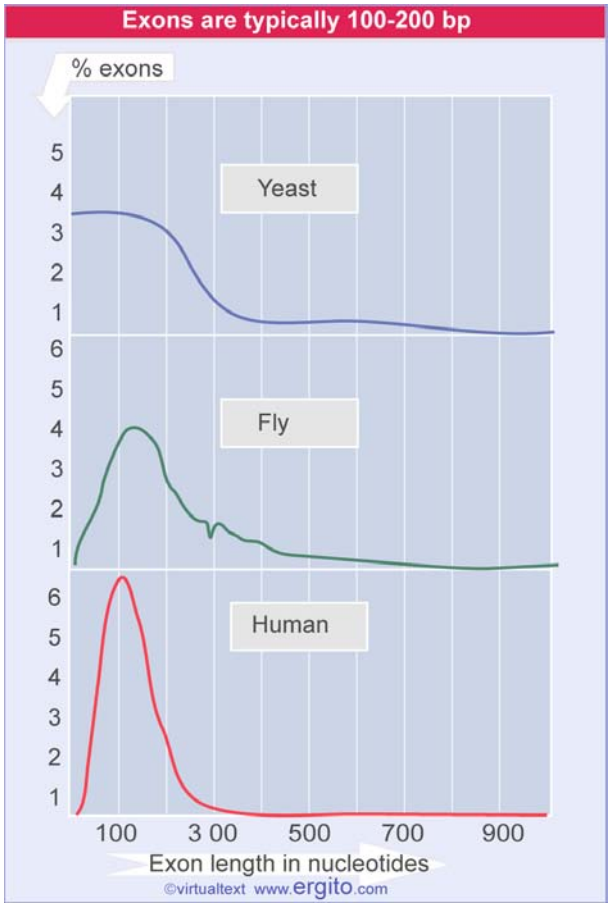


**Figure 2.15** Exons coding for proteins are usually short.

**Figure 2.16** shows that introns vary widely in size. 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 a significant proportion. In vertebrates, the size distribution is much wider, extending from approximately the same length as the exons (<200 bp) to lengths measured in 10s of kbs, and extending up to 50-60 kb in extreme cases.

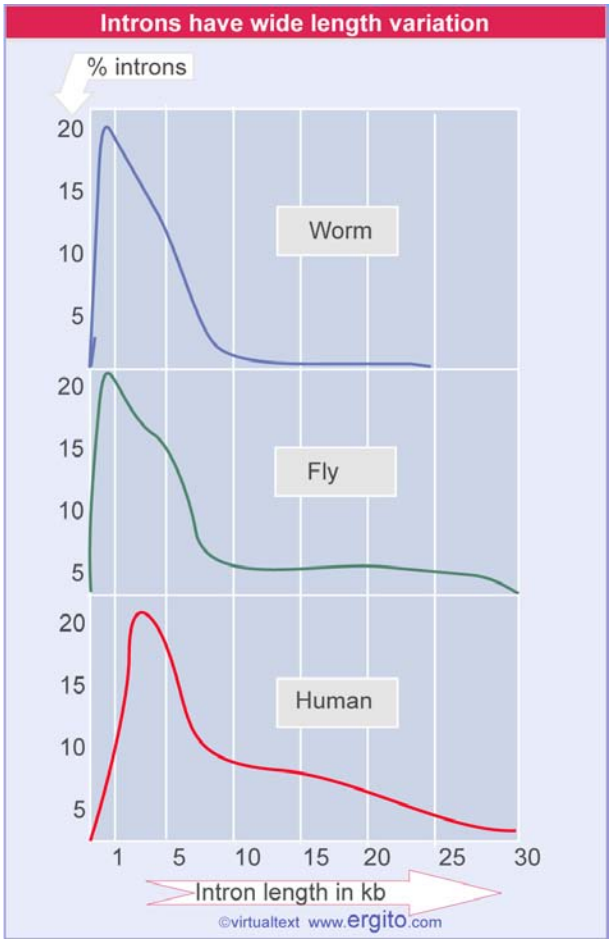


**Figure 2.16** Introns range from very short to very long.

Very long genes are the result of very long introns, not the result of coding for longer products. There is no correlation between gene size and mRNA size in higher eukaryotes; nor is there a good correlation between gene size and the number of exons. The size of a gene therefore depends primarily on the lengths of its individual introns. In mammals, insects, and birds, the "average" gene is approximately 5× the length of its mRNA.

*Last updated on 2-16-2001*

*This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.2.7*

THE INTERRUPTED GENE

# 1.2.8 Some DNA sequences code for more than one protein

## Key Concepts

- The use of alternative initiation or termination 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.

Most genes consist of a sequence of DNA that is devoted solely to the purpose of coding for one protein (although the gene may include noncoding regions at either end and introns within the coding region). However, there are some cases in which a single sequence of DNA codes for more than one protein.

*Overlapping genes* occur in the relatively simple situation in which one gene is part of the other. The first half (or second half) of a gene is used independently to specify a protein that represents the first (or second) half of the protein specified by the full gene. This relationship is illustrated in **Figure 2.17**. The end result is much the same as though a partial cleavage took place in the protein product to generate part-length as well as full-length forms.

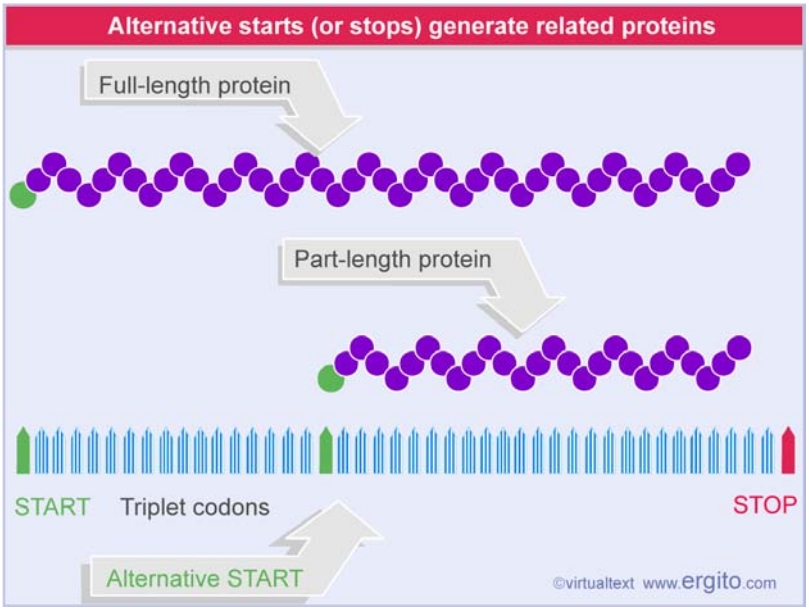


**Figure 2.17** Two proteins can be generated from a single gene by starting (or terminating) expression at different points.

Two genes overlap in a more subtle manner when the same sequence of DNA is shared between two *nonhomologous* proteins. This situation arises when the same sequence of DNA is translated in more than one reading frame. In cellular genes, a DNA sequence usually is read in only one of the three potential reading frames, but in some viral and mitochondrial genes, there is an overlap between two adjacent genes that are read in different reading frames. This situation is illustrated in **Figure 2.18**. The distance of overlap is usually relatively short, so that most of the sequence representing the protein retains a unique coding function.

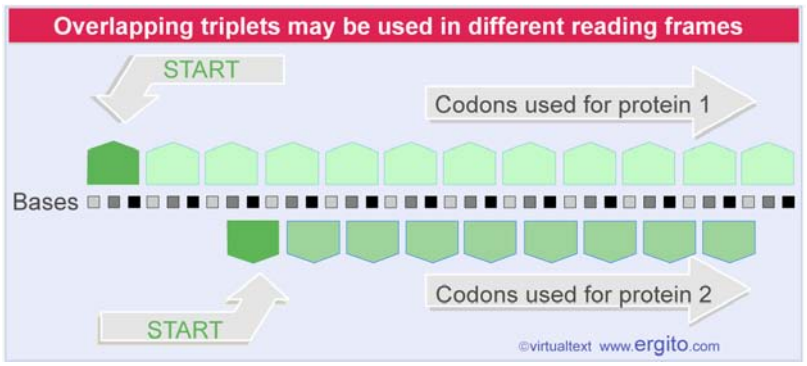


**Figure 2.18** Two genes may share the same sequence by reading the DNA in different frames.

In some genes, *alternative* patterns of gene expression create switches in the pathway for connecting the exons. A single gene may generate a variety of mRNA products that differ in their content of exons. The difference may be that certain exons are optional – they may be included or spliced out. Or there may be exons that are treated as mutually exclusive – one or the other is included, but not both. The alternative forms produce proteins in which one part is common while the other part

is different.

In some cases, the alternative means of expression do not affect the sequence of the protein; for example, changes that affect the 5′ nontranslated leader or the 3′ nontranslated trailer may have regulatory consequences, but the same protein is made. In other cases, one exon is substituted for another, as indicated in **Figure 2.19**.

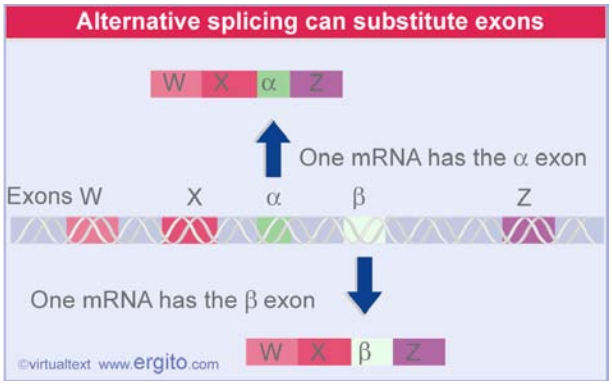


**Figure 2.19** Alternative splicing generates the α and β variants of troponin T.

In this example, the proteins produced by the two mRNAs contain sequences that overlap extensively, but that are different within the alternatively spliced region. The 3′ half of the troponin T gene of rat muscle contains 5 exons, but only 4 are used to construct an individual mRNA. Three exons, *WXZ*, are the same in both expression patterns. However, in one pattern the α exon is spliced between *X* and *Z;* in the other pattern, the β exon is used. The α and β forms of troponin T therefore differ in the sequence of the amino acids present between sequences 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 to form an individual mRNA, but both cannot be used in the same mRNA.

**Figure 2.20** illustrates an example in which alternative splicing leads to the inclusion of an exon in some mRNAs, while it is left out of others. A single type of transcript is made from the gene, but it can be spliced in either of two ways. In the first pathway, two introns are spliced out, and the three exons are joined together. In the second pathway, the second exon is not recognized. As a result, 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 part of the single intron. The pathways produce two proteins that are the same at their ends, but one of which has an additional sequence in the middle. So the region of DNA codes for more than one protein. (Other types of combinations that are produced by alternative splicing are discussed in *Molecular Biology 5.24.12 Alternative splicing involves differential use of splice junctions*).

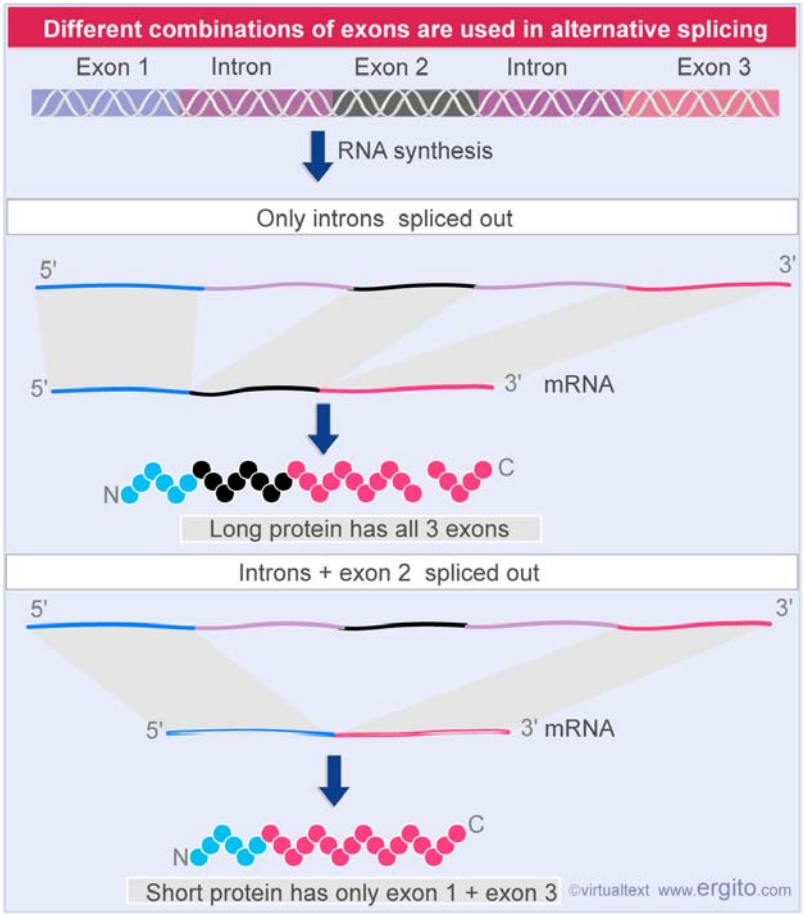


**Figure 2.20** Alternative splicing uses the same pre-mRNA to generate mRNAs that have different combinations of exons.

Sometimes two pathways operate simultaneously, a certain proportion of the RNA being spliced in each way; sometimes the pathways are alternatives that are expressed under different conditions, one in one cell type and one in another cell type.

So alternative (or differential) splicing can generate proteins with overlapping sequences from a single stretch of DNA. It is curious that the higher eukaryotic genome is extremely spacious in having large genes that are often quite dispersed, but at the same time it may make multiple products from an individual locus. Alternative splicing expands the number of proteins relative to the number of genes by ~15% in flies and worms, but has much bigger effects in man, where ~60% of genes may have alternative modes of expression (see *Molecular Biology 1.3.11 The human genome has fewer genes than expected*). About 80% of the alternative splicing events result in a change in the protein sequence.

*Last updated on 3-10-2003*

*This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.2.8*

THE INTERRUPTED GENE

# 1.2.9 How did interrupted genes evolve?

## Key Concepts

- The major evolutionary question is whether genes originated as sequences interrupted by exons or whether they were originally uninterrupted.
- Most protein-coding genes probably originated in an interrupted form, but interrupted genes that code for RNA may have originally been uninterrupted.
- A special class of introns is mobile and can insert itself into genes.

The highly interrupted structure of eukaryotic genes suggests a picture of the eukaryotic genome as a sea of introns (mostly but not exclusively unique in sequence), in which islands of exons (sometimes very short) are strung out in individual archipelagoes that constitute genes.

What was the original form of genes that today are interrupted?

- The "introns early" model supposes that introns have always been an integral part of the gene. Genes originated as interrupted structures, and those without introns have lost them in the course of evolution.
- The "introns late" model supposes that the ancestral protein-coding units consisted of uninterrupted sequences of DNA. Introns were subsequently inserted into them.

A test of the models is to ask whether the difference between eukaryotic and prokaryotic genes can be accounted for by the acquisition of introns in the eukaryotes or by the loss of introns from the prokaryotes.

The introns early model suggests that the mosaic structure of genes is a remnant of an ancient approach to the reconstruction of genes to make novel proteins. Suppose that an early cell had a number of separate protein-coding sequences. One aspect of its evolution is likely to have been the reorganization and juxtaposition of different polypeptide units to build up new proteins.

If the protein-coding unit must be a continuous series of codons, every such reconstruction would require a precise recombination of DNA to place the two protein-coding units in register, end to end in the same reading frame. Furthermore, if this combination is not successful, the cell has been damaged, because it has lost the original protein-coding units.

But if an approximate recombination of DNA could place the two protein-coding units within the same transcription unit, splicing patterns could be tried out at the level of RNA to combine the two proteins into a single polypeptide chain. And if

these combinations are not successful, the original protein-coding units remain available for further trials. Such an approach essentially allows the cell to try out controlled deletions in RNA without suffering the damaging instability that could occur from applying this procedure to DNA. This argument is supported by the fact that we can find related exons in different genes, as though the gene had been assembled by mixing and matching exons (see *Molecular Biology 1.2.10 Some exons can be equated with protein functions*).

**Figure 2.21** illustrates the outcome when a random sequence that includes an exon is translocated to a new position in the genome. Exons are very small relative to introns, so it is likely that the exon will find itself within an intron. Because only the sequences at the exon-intron junctions are required for splicing, the exon is likely to be flanked by functional 3′ and 5′ splice junctions, respectively. Because splicing junctions are recognized in pairs, the 5′ splicing junction of the original intron is likely to interact with the 3′ splicing junction introduced by the new exon, instead of with its original partner. Similarly, the 5′ splicing junction of the new exon will interact with the 3′ splicing junction of the original intron. The result is to insert the new exon into the RNA product between the original two exons. So long as the new exon is in the same coding frame as the original exons, a new protein sequence will be produced. This type of event could have been responsible for generating new combinations of exons during evolution. Note that the principle of this type of event is mimicked by the technique of exon trapping that is used to screen for functional exons (see **Figure 2.12**).

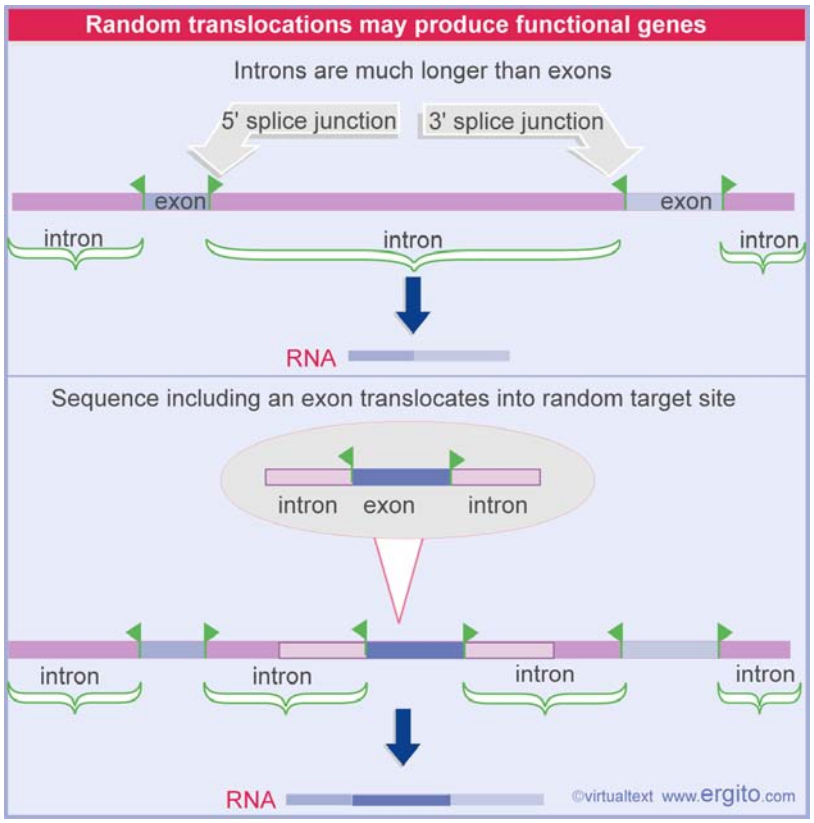


**Figure 2.21** An exon surrounded by flanking sequences that is translocated into an intron may be spliced into the RNA product.

Alternative forms of genes for rRNA and tRNA are sometimes found, with and without introns. In the case of the tRNAs, where all the molecules conform to the same general structure, it seems unlikely that evolution brought together the two regions of the gene. After all, the different regions are involved in the base pairing that gives significance to the structure. So here it must be that the introns were inserted into continuous genes.

Organelle genomes provide some striking connections between the prokaryotic and eukaryotic worlds. Because of many general similarities between mitochondria or chloroplasts and bacteria, it seems likely that the organelles originated by an *endosymbiosis* in which an early bacterial prototype was inserted into eukaryotic cytoplasm. Yet in contrast with the resemblances with bacteria – for example, as seen in protein or RNA synthesis – some organelle genes possess introns, and therefore resemble eukaryotic nuclear genes.

Introns are found in several chloroplast genes, including some that have homologies with genes of *E. coli*. This suggests that the endosymbiotic event occurred before introns were lost from the prokaryotic line. If a suitable gene can be found, it may therefore be possible to trace gene lineage back to the period when endosymbiosis occurred.

The mitochondrial genome presents a particularly striking case. The genes of yeast and mammalian mitochondria code for virtually identical mitochondrial proteins, in spite of a considerable difference in gene organization. Vertebrate mitochondrial genomes are very small, with an extremely compact organization of continuous genes, whereas yeast mitochondrial genomes are larger and have some complex interrupted genes. Which is the ancestral form? The yeast mitochondrial introns (and certain other introns) can have the property of mobility – they are self-contained sequences that can splice out of the RNA and insert DNA copies elsewhere – which suggests that they may have arisen by insertions into the genome (see *Molecular Biology 5.26.5 Some group I introns code for endonucleases that sponsor mobility* and *Molecular Biology 5.26.6 Some group II introns code for reverse transcriptases*).

THE INTERRUPTED GENE

# 1.2.10 Some exons can be equated with protein functions

## Key Concepts

- Facts suggesting that exons were the building blocks of evolution and the first genes were interrupted are:
  - Gene structure is conserved between genes in very distant species.
  - Many exons can be equated with coding for protein sequences that have particular functions.
  - Related exons are found in different genes.

If current proteins evolved by combining ancestral proteins that were originally separate, the accretion of units is likely to have occurred sequentially over some period of time, with one exon added at a time (for review see 9). Can the different functions from which these genes were pieced together be seen in their present structures? In other words, can we equate particular functions of current proteins with individual exons ?

In some cases, there is a clear relationship between the structures of the gene and protein. The example *par excellence* is provided by the immunoglobulin proteins, which are coded by genes in which every exon corresponds exactly with a known functional domain of the protein. **Figure 2.22** compares the structure of an immunoglobulin with its gene.

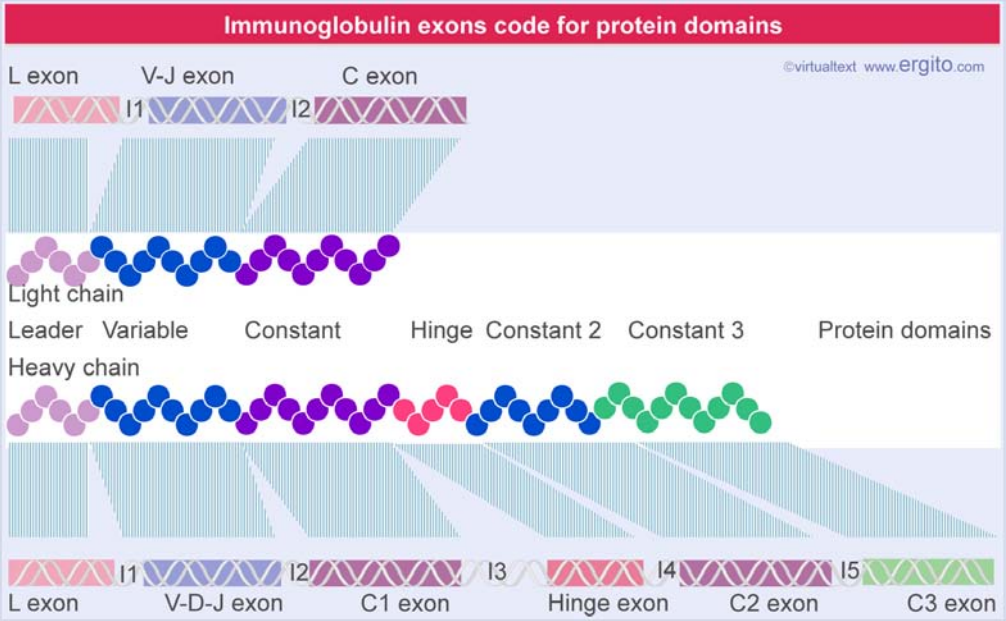


**Figure 2.22** Immunoglobulin light chains and heavy chains are coded by genes whose structures (in their expressed forms) correspond with the distinct domains in the protein. Each protein domain corresponds to an exon; introns are numbered 1-5.

An immunoglobulin is a tetramer of two light chains and two heavy chains, which aggregate to generate a protein with several distinct domains. Light chains and heavy chains differ in structure, and there are several types of heavy chain. Each type of chain is expressed from a gene that has a series of exons corresponding with 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, the first exon, coding for the N-terminal region of the polypeptide, often specifies the signal sequence involved in membrane secretion. An example is insulin.

The view that exons are the functional building blocks of genes is supported by cases in which two genes may have some exons that are related to one another, while other exons are found only in one of the genes. **Figure 2.23** summarizes the relationship between the receptor for human LDL (plasma low density lipoprotein) and other proteins. In the center of the LDL receptor gene is a series of exons related to the exons of the gene for the precursor for EGF (epidermal growth factor). In the N-terminal part of the protein, a series of exons codes for a sequence related to the blood protein complement factor C9. So the LDL receptor gene was created by assembling *modules* for its various functions. These modules are also used in different combinations in other proteins.

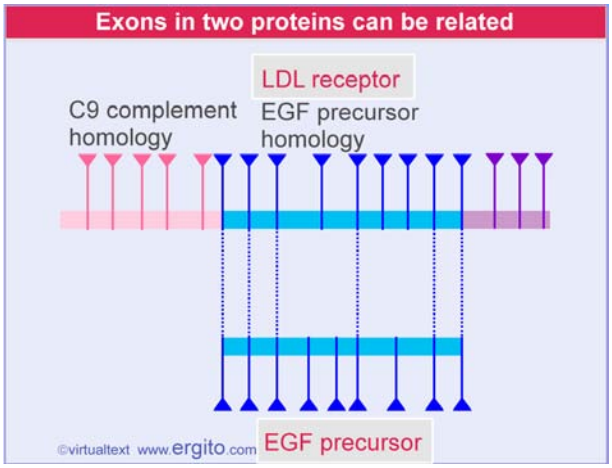


**Figure 2.23** The LDL receptor gene consists of 18 exons, some of which are related to EGF precursor and some to the C9 blood complement gene. Triangles mark the positions of introns. Only some of the introns in the region related to EGF precursor are identical in position to those in the EGF gene.

Exons tend to be fairly small (see **Figure 2.12**), around the size of the smallest polypeptide that can assume a stable folded structure, ~20-40 residues. Perhaps proteins were originally assembled from rather small modules. Each module need not necessarily correspond to a current function; several modules could have combined to generate a function. The number of exons in a gene tends to increase with the length of its protein, which is consistent with the view that proteins acquire multiple functions by successively adding appropriate modules.

This idea might explain another feature of protein structure: it seems 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.

## Reviews

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

THE INTERRUPTED GENE

# 1.2.11 The members of a gene family have a common organization

## Key Terms

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

## 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 3 exons and 2 introns, suggesting that they are descended from a single ancestral gene.

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 of animals, whose amino acid sequence suggests a common (though ancient) origin with the globin subunits. Leghemoglobins are oxygen-binding proteins present in the legume class of plants; like myoglobin, they are monomeric. They too share a common origin with the other heme-binding proteins. Together, the globins, myoglobin, and leghemoglobin constitute the globin **superfamily**, a set of gene families all descended from some (distant) common ancestor.

Both α- and β-globin genes have three exons (see **Figure 2.7**). The two introns are located at constant positions relative to the coding sequence. The central exon represents the heme-binding domain of the globin chain.

Myoglobin is represented by a single gene in the human genome, whose structure is essentially the same as that of the globin genes. The three-exon structure therefore predates the evolution of separate myoglobin and globin functions.

Leghemoglobin genes contain three introns, the first and last of which occur at points in the coding sequence that are homologous to the locations of the two introns in the globin genes. This remarkable similarity suggests an exceedingly ancient origin for the heme-binding proteins in the form of a split gene, as illustrated in **Figure 2.24**.

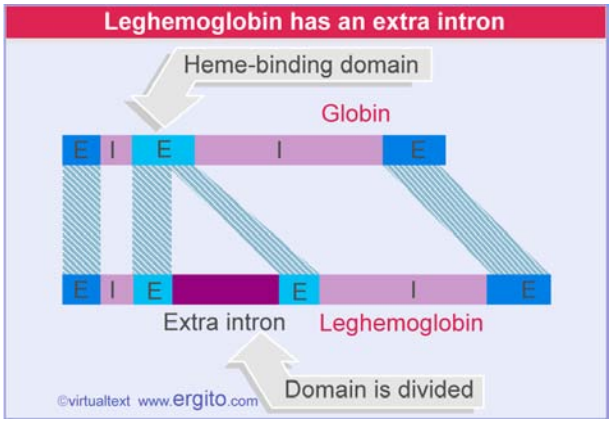


**Figure 2.24** The exon structure of globin genes corresponds with protein function, but leghemoglobin has an extra intron in the central domain.

The central intron of leghemoglobin separates two exons that together code for the sequence corresponding to the single central exon in globin. 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 at the start of plant evolution?

Cases in which homologous genes differ in structure may provide information about their evolution. An example is insulin. Mammals and birds have only one gene for insulin, except for the rodents, which have two genes. **Figure 2.25** illustrates the structures of these genes.

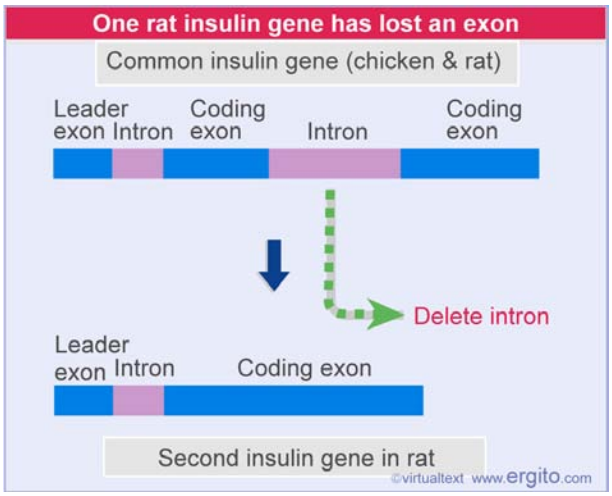


**Figure 2.25** The rat insulin gene with one intron evolved by losing an intron from an ancestor with two interruptions.

The principle we use in comparing the organization of related genes in different species is that *a common feature identifies a structure that predated the evolutionary separation of the two species.* In chicken, the single insulin gene has two introns; one of the two rat genes has the same structure. The common structure implies that the ancestral insulin gene had two introns. However, 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 copies.

The organization of some genes shows extensive discrepancies between species. In these cases, there must have been extensive removal or insertion of introns during evolution.

A well characterized case is represented by the actin genes. The typical actin gene has a nontranslated leader of <100 bases, a coding region of ~1200 bases, and a trailer of ~200 bases. Most actin genes are interrupted; the positions of the introns can be aligned with regard to the coding sequence (except for a single intron sometimes found in the leader).

**Figure 2.26** shows that almost every actin gene is different in its pattern of interruptions. Taking all the genes together, introns occur at 12 different sites. However, no individual gene has more than 6 introns; some genes have only one intron, and one is uninterrupted altogether. How did this situation arise? If we suppose that the primordial actin gene was interrupted, and 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 or more. The alternative is to suppose that a process of intron insertion continued independently in the different lines of evolution. The relationships between the intron locations found in different species may be used ultimately to construct a tree for the evolution of the gene.

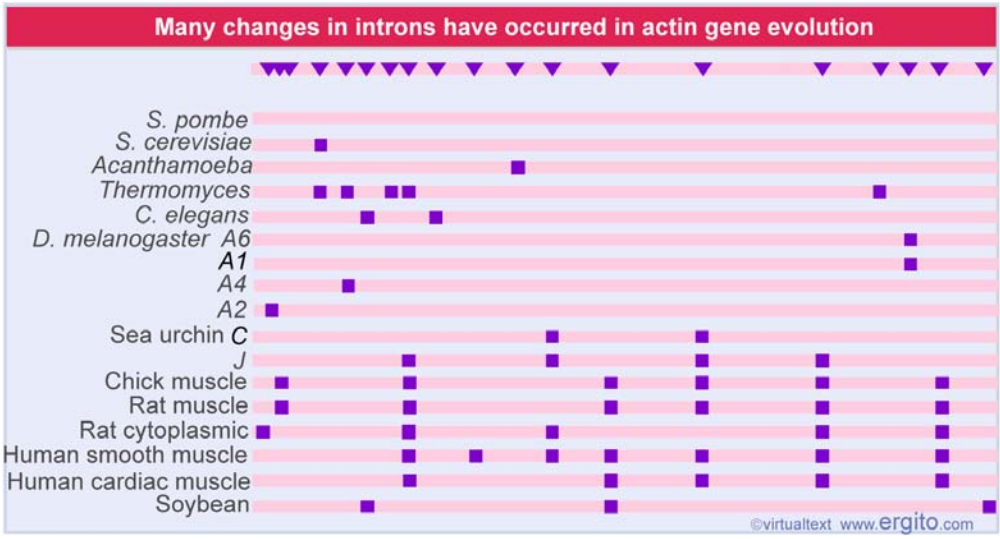


**Figure 2.26** Actin genes vary widely in their organization. The sites of introns are indicated in purple.

The relationship between exons and protein domains is somewhat erratic. In some cases there is a clear 1:1 relationship; in others no pattern is to be discerned. One possibility is that removal of introns has fused the 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 a coherent domain. Together with the variations that we see in exon placement in cases such as the actin genes, this argues that intron positions can be adjusted in the course of

evolution.

The equation of at least some exons with protein domains, and the appearance of related exons in different proteins, leaves no doubt that the duplication and juxtaposition of exons has played an important role 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 (a few thousands or tens of thousands). By taking exons as the building blocks of evolution, this view implicitly accepts the introns early model for the origin of genes coding for proteins.

THE INTERRUPTED GENE

# 1.2.12 Is all genetic information contained in DNA?

## Key Terms

**Positional information** describes the localization of macromolecules at particular places in an embryo. The localization may itself be a form of information that is inherited.

## Key Concepts

- The definition of the gene has reversed from "one gene : one protein" to "one protein : one gene".
- Positional information is also important in development.

The concept of the gene has evolved significantly in the past few years. The question of what's in a name is especially appropriate for the gene. We can no longer say that a gene is a sequence of DNA that continuously and uniquely codes for a particular protein. In situations in which a stretch of DNA is responsible for production of one particular protein, 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 the sequences representing proteins 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 actually responsible for production of the polypeptide (including introns as well as exons) as constituting the gene, while recognizing that from the perspective of another protein, part of this same sequence also belongs to *its* gene. This allows the use of descriptions such as "overlapping" or "alternative" genes.

We can now see how far we have come from the original one gene : one enzyme hypothesis. Up to that time, the driving question was the nature of the gene. Once it was discovered that genes represent proteins, the paradigm became fixed in the form of the concept that every genetic unit functions through the synthesis of a particular protein.

This view remains the central paradigm of molecular biology: a sequence of DNA functions either by directly coding for a particular protein or by being necessary for the use of an adjacent segment that actually codes for the protein. How far does this paradigm take us beyond explaining the basic relationship between genes and proteins?

The development of multicellular organisms rests on 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 functioning. The molecular nature of this regulatory network is largely unknown, but we assume that it consists of genes that code for products (probably protein, perhaps sometimes RNA) that act on other genes.

While 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 know how these particular regions are formed. But we may speculate that the existence of positional information in the egg leads to the differential expression of genes in the cells subsequently formed in these regions, which leads to the development of the adult organism, which leads to the development of an egg with the appropriate positional information#

This possibility prompts us to ask whether 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: when we read out 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?

## THE INTERRUPTED GENE

# 1.2.13 Summary

All types of eukaryotic genomes contain interrupted genes. The proportion of interrupted genes is low in yeasts and increases in the lower eukaryotes; few genes are uninterrupted in higher eukaryotes.

Introns are found in all classes of eukaryotic genes. The structure of the interrupted gene is the same in all tissues, exons are joined together in RNA in the same order as their organization in DNA, and the introns usually have no coding function. Introns 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.

Positions of introns are often conserved when the organization of homologous genes is compared between species. Intron sequences vary, and may even be unrelated, although exon sequences remain well related. The conservation of exons can be used to isolate related genes in different species.

The size of a gene is determined primarily by the lengths of its introns. Introns become larger early in the higher eukaryotes, when gene sizes therefore increase significantly. The range of gene sizes in mammals is generally from 1-100 kb, but it is possible to have even larger genes; the longest known case is dystrophin at 2000 kb.

Some genes share only some of their exons with other genes, suggesting that they have been assembled by addition of exons representing individual modules of the protein. Such modules may have been incorporated into a variety of different proteins. The idea that genes have been assembled by accretion of exons implies that introns were present in genes of primitive organisms. Some of the relationships between homologous genes can be explained by loss of introns from the primordial genes, with different introns being lost in different lines of descent.

*This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.2.13*