3.9: Introns and Exons

Multiple, large introns can make some eukaryotic genes very large. Eukaryotic genes can be split into many (>60), sometimes very small exons (e.g. <60 bp, coding for <20 amino acids), separated by very large introns (as large as >100kb), resulting in some enormous genes (>500 kb). E.g. the DMD gene (which when mutated can cause Duchenne's muscular dystrophy) is almost 1 Mb, about 1/4 the size of the E. coli chromosome! The average size of genes from more complex organisms is considerably larger than those of simpler ones, but the avg. size of mRNA is about the same, reflecting the presence of more and larger introns in the more complex organisms. tRNA and rRNA genes also contain introns.

Finding exons in long genomic sequences using computer programs

Far more exons and introns have been discovered (or more accurately, predicted) through the analysis of genomic DNA sequences than could ever be discovered by direct experimentation. The different types of exons, the enormous length of introns, and other factors have complicated the task of finding reliable diagnostic signatures for exons in genomic sequences. However, considerable progress has been made and continues in current research. Some of the commonly used approaches are summarized in Figure 3.27.
Finding exons with computers

- *Ab initio* computation
  - E.g. Genescan:
    - http://genes.mit.edu/GENSCAN.html
  - Uses an explicit, sophisticated model of gene structure, splice site properties, etc to predict exons
- Compare with genomics and cDNA sequences
  - BLAST2 alignments between cDNA and genomic sequences

Find exons for *HBB*

- Sequence for human beta-globin gene (*HBB*):
  - accession number L43217
  - Thalassemia variant
- Sequence for *HBB* mRNA
  - NN_000518
- Retrieve those from GenBank at NCBI (or the course website)
  - Get the files in FASTA format
- Run Genescan and BLAST2 sequences

Genescan analysis of *HBB* gene

BLAST2: *HBB* gene vs. cDNA

![Gene and cDNA](https://bio.libretexts.org/Bioinformatics/Bookshelf/Genetics/Book%3A_Working_with_Molecular_Genetics_(Hardison)/Unit_1%3A_Genes...)

**Figure 3.27.** Introns in the b-globin gene can be reliably identified computationally
Introns are removed by splicing RNA precursors

![Diagram]

**Figure 3.28.** Introns are removed from pre-mRNA to generate mRNA.

**Alternative splicing** generates more than one polypeptide from the same gene

![Diagram]

**Figure 3.29.**

Some segments of RNA may be included in the mature mRNA (exons) but not included on other spliced products. The alternative products may be made in different tissues or at different developmental stages - i.e. alternative splicing can be regulated.
Split genes may enhance the rate of evolution

Many exons encode a unit very close to a protein domain, e.g. the exons of leghemoglobin, or the variable and constant regions of immunoglobulins, or domains (e.g. "kringle") in EGF precursor that are also found in part of the LDL receptor. The exon organization tends to be well conserved in highly divergent species. Introns tend to occur between those portions of genes that encode structural domains of proteins.

Duplication of the exons encoding structural domains and subsequent recombination can lead to more rapid evolution of a new protein, essentially using the parts from earlier evolved genes. Analogous to building a house from prefabricated parts, as opposed to one nail and one board at a time - start with preassembled walls, roof joists etc.

However, the relationship between exons and structural domains of proteins is not exact, and some exon-intron boundaries vary (a little) in genes for different species. A different model holds that the introns are transposable elements (some certainly are - see later). They can insert anywhere in a gene, but they are least disruptive at domain boundaries, and these latter insertions are more likely to be fixed in a population than insertions into the middle of a region encoding a domain. So the results after long years of evolution is that the introns tend to be between region coding domains, but the gene was originally intact, not assembled from discrete exons.

Multigene families and gene clusters

Many eukaryotic genes are found in multiple copies. Some of them are developmentally regulated, such as HOX gene clusters and globin gene clusters.

Figure 3.30.

A multigene family contains multiple genes of similar sequence encoding similar proteins; e.g. globin genes (Figure 3.30). Globin genes are expressed at different times of development. The order of developmental expression is the same as their order along the chromosome, e.g. the e-globin gene is expressed in early embryonic red cells, the g-
globin gene is expressed at a high level in fetal red cells, and the b-globin gene is expressed in red cells after birth. As we will see later, this correlates with their distance from a dominant control element at the 5’ end of the cluster, the Locus Control Region.

The order of HOX genes is also aligned with their spatial expression in the embryo. This is another example of alignment between chromosomal position and regulation of expression.

Other multi-gene families include those encoding histones, immunoglobulins, actins, cyclins, cyclin-dependent protein kinases, and rRNAs. Some of these families are linked in gene clusters, but others are dispersed around the genome. Having multiple copies of genes may be more the rule than the exception in eukaryotic genomes.

Experimental techniques that reveal multigene families include the following.

**Purification and analysis of a particular kind of protein**, e.g. hemoglobins, immunoglobulins, and many enzymes, *may reveal heterogeneity*. Further purification (via chromatography and electrophoresis) and sequencing can show that the observed heterogeneity is a result of related but not identical proteins, and one deduces that these similar proteins are encoded by multiple genes with similar sequences, i.e. a multigene family.

**Analysis of the clones obtained by screening a library of cloned genomic DNA may reveal multiple related sequences**, each with a distinctive restriction map. In many cases these are clones of different, related genes that comprise a multigene family (Figure 3.31).

Southern blot-hybridization of restriction-cleaved genomic DNA can reveal multiple copies of genes, simply as multiple bands on the hybridized blot. Although the number of fragments generated from total genomic DNA is too many to resolve on a gel, after transfer to a membrane, particular fragments can be visualized by hybridization with a specific probe. The number of hybridizing fragments is roughly correlated with the number of copies of related genes. Some genes are cleaved by the restriction enzyme, producing multiple bands, but some fragments can have multiple genes. A true measure of the number of related genes comes from more detailed restriction mapping or sequencing.

**Figure 3.31.** Blot-hybridization analysis of clones of genomic DNA and genomic DNA showing that multiple copies of...
genes are present. A set of overlapping clones containing rabbit genomic DNA were digested and run on an agarose gel (panel A), blotted onto a membrane and hybridized with a radiolabeled probe that detected embryonic hemoglobin genes, and exposed to X-ray film. The resulting autoradiogram is shown in panel B. Panel C shows the results of a blot-hybridization analysis of rabbit total genomic DNA, using the same probe. Many of the same bands are seen as in the cloned DNA, confirming the existence of multiple hybridizing fragments. Mapping the fragments showed that they represented separate genes.

Keeping multigene families homogeneous

Sometimes multiple copies of genes are maintained as virtually identical over the course of evolution: e.g. rRNA genes, histone genes, a-globin genes (in primates). In these cases, the multiple copies are coevolving (concerted evolution).

sequence differences

Human: $\text{A|A|A}$ among human genes: 1%
between human & chimp: 5%
Chimp: $\text{A|A|A}$ among chimp genes: 1%
between chimp & monkey: 10%
Monkey: $\text{A|A|A}$ among monkey genes: 1%

Since all three primates have 3 A genes, we infer that the common ancestor had 3 genes (the duplications preceded the speciation events). If in the time since human and chimp diverged, the A genes have diverged 5%, why haven't the A genes in human (e.g.) also diverged 5% from each other? They have been apart even longer than the human and chimp chromosomes carrying them! The A genes within a species are "talking to each other", or co-evolving or evolving in concert.

Sequence homogeneity in a multigene family can arise because of recent gene amplification (Figure 3.32 part1). In this case the genes have not been separate from each other long enough to accumulate variation in their sequences. Other multigene families have existed for a long time, but maintain sequence homogeneity despite ample opportunity for divergence. Two mechanisms have been seen that maintain similarity. The first is multiple rounds of unequal crossing over. As illustrated in Figure 3.32, part 2, the expansions and contractions of repeated genes can result in a new variant predominanting in the gene cluster. The other method for maintaining homogeneity is gene conversion between homologs. When a new mutation arises, it can be removed by conversion with the unmutated allele, or the mutation can be passed on the the other allele. Either way, the sequences of the two alleles becomes the same.

Sometimes the products of the gene duplications, or duplicative transpositions, accumulate mutations so they are no longer functional. These remnants of once-active genes are called pseudogenes.
Mechanisms for maintaining homogeneity in a tandem multigene family

1. Gene amplification

2. Multiple rounds of unequal crossover during sister chromatid exchanges

3. Gene conversion

Figure 3.32.