3.8: Eukaryotic Gene Structure

Much can be learned about any gene after it has been isolated by recombinant DNA techniques. The structure of coding and noncoding regions, the DNA sequence, and more can be deduced. This is true for bacterial and viral genes, as well as eukaryotic cellular genes. The next sections of this chapter will focus on analysis of eukaryotic genes, showing the power of examining purified copies of genes.

Split genes and introns

Precursors to mRNA longer than mRNA

Initial indications of a complex structure to eukaryotic genes came from analysis of nuclear RNAs during the 1970’s. The precursors to messenger RNA, or pre-mRNAs, were found to be surprisingly long, considerably larger than the average mRNA size (Figure \
\cite{PagelIndex{1}}).
Denaturing sucrose gradients (with high concentration of formamide, e.g. >50%) separate RNAs on the basis of size. Analysis of nuclear RNA showed that the average size was much larger than the average size of cytoplasmic RNA. Labeled RNA could be "chased" from the nucleus to the cytoplasm - i.e. nuclear RNA was a precursor to mRNA and other cytoplasmic RNAs. Was the extra RNA at the ends? or in the middle of the pre-mRNA? More precisely, one could examine specific RNAs by hybridizing fractions from the denaturing sucrose gradients to labeled copies of, e.g. globin mRNA. The hybridizing RNA from the nucleus was about 11S (as well as mature 8S message), whereas cytoplasmic RNA of about 8S hybridized. Thus the nuclear RNA encoding globin is larger than the cytoplasmic mRNA.

**Visualization of mRNA-DNA heteroduplexes revealed extra sequences internal to the mRNA-coding segments**

R-loops are hybrids between RNA and DNA that can be visualized in the EM, under conditions where DNA-RNA duplexes are favored over DNA-DNA duplexes (Figure \(\PageIndex{2}\)). For a simple gene structure, one sees a continuous RNA-DNA duplex (smooth, slowly curving) and a displaced single strand of DNA (thinner, many more turns and curves – single stranded DNA is not a rigid as double stranded nucleic acid, either duplex DNA or RNA-DNA).
EM pictures of duplexes between purified adenovirus mRNAs and the genomic DNA showed extensions at both the 3' (poly A) and 5' ends, which are encoded elsewhere on the genome. All late mRNAs have the same sequence at the 5' end; this is derived from the tripartite leader. R-loops between late mRNAs and adenovirus DNA fragments including the major late promoter showed duplexes with the leader segments, separated by loops of duplex DNA (Figure 3.23, bottom panel). The RNA-DNA hybrids identify regions of DNA that encode RNA. The surprising result is that RNA-coding portions of a gene are separated by loops of duplex DNA in the R-loop analysis. Examples of R-loops in genes with introns are shown in Figure \( \PageIndex{3} \).

These data showed that the adenovirus RNAs are encoded in different segments of the viral genome; i.e. the genes are split. The portion of a gene that encodes mRNA was termed an exon. The part of gene does not code for sequences in the mature mRNA is called an intron. These observations led to the Nobel Prize for Phil Sharp and Rich Roberts. Louise Chow and Sue Berget were also key players in the discovery of introns.
Figure 3: R-loops between clones of rabbit beta-like globin genes (now called HBE and HBG) and mRNA from rabbit embryonic erythroid cells. A photograph from the electron microscope is shown at the top of each panel, and an interpretive drawing is included below it. The displaced nontemplate strand of DNA forms partial or complete duplexes with the template strand in the large intron. A small intron is also visible in panel C. Panel G shows the two genes together on one large clone.

**Interruptions in cellular genes** were discovered subsequently, in the late 1970's, in globin genes, immunoglobulin genes and others. We now realize that most genes in complex eukaryotes are split by multiple introns.

Exons are more conserved than introns (in most cases), since alterations in protein-coding regions that alter or decrease function are selected against, whereas many sequences in introns can be altered without affecting the function of the gene product. Important sequences in introns (such as splice junctions, the branch point, and occasionally enhancers) are covered in some detail in Part Three.

**Differences in restriction maps between cDNA and genomic clones reveal introns**

Restriction maps based on copies of the mRNA (cDNA) were different from those in genomic DNA - the genes were cleaved by some restriction endonucleases that the cDNAs were not, and some restriction sites were further apart in the genomic DNA. These observations were explained by the presence of intervening sequences or introns (Figure 4).
The experimental procedures to do this involve making a **restriction map** of the clones of genomic DNA, and then identifying the regions that encode mRNA by hybridization of labeled cDNA probes to the restriction digests. Cloned genomic DNA digested with appropriate restriction endonucleases, separated by size on an agarose gel, and then transferred onto a nylon or nitrocellulose solid support. This **Southern blot** is then hybridized with a labeled probe specific to the cDNA (composed only of exons). The pattern of labeled fragments on the resulting autoradiogram shows the fragments that contain exons. Alignment of these with the restriction map of the gene gives an approximation of the position of the exons.

The blot-hybridization approach can be combined with a PCR (polymerase chain reaction) analysis for higher resolution. Primers are synthesized that will anneal to adjacent exons. The difference in size of the PCR amplification product between genomic DNA and cDNA is the size of the intron. The PCR product can be cloned and sequenced for more detailed information, e.g. to precisely define the exon/intron junctions.

Subsequently, the nucleotide sequence of exonic regions and preferably the entire gene is determined. The presence of introns were confirmed and their locations defined precisely in DNA sequences of isolated clones of the genes.

### Types of Exons

Eukaryotic genes are a combination of introns and exons. However, not all exons do the same thing (Figure 4). In particular, the protein-coding regions or genes are a subset of the sequences in exons. Exons include both the untranslated regions and the protein-coding, translated regions. Introns are the segments of genes that are present in the primary transcript (or precursor RNA) but are removed by splicing in the production of mature RNA. Methods used to detect coding regions will not find all exons.
Figure \(\PageIndex{5}\): Types of exons