Eukaryotic Gene Structure
Terminology

- **Genome** – entire genetic material of an individual
- **Transcriptome** – set of transcribed sequences
- **Proteome** – set of proteins encoded by the genome
Gene

• Basic physical and functional units of heredity.
• Specific sequences of DNA bases that encode instructions on how to make **proteins**.
• In eukaryotes, the genes have their coding sequences (**exons**) interrupted by non-coding sequences (**introns**).
Eukaryotic Gene Structure

- Exons
- Introns
- Promoter sequences
- Terminator sequences
- Enhancers or silencers regulatory sequences which may be up stream or down stream, near or far from the gene

Signals:
- Up stream sequence signal for addition of Cap
- Down stream sequence signal for addition of poly A tail
Eukaryotic Gene Structure
• The gene is the sequence of chromosomal DNA that is required for production of functional product; mRNA, polypeptide, tRNA, rRNA.

• The sequence of gene consists of:
  • **Exons**: coding sequence, transcribed and translated. Coding for the amino acids in the polypeptide chain. Vary in number, sequence and length from one gene to another.
• **Introns**; non-coding sequence, transcribed but not translated. Also vary in sequence, number and length from one gene to another but they are sharing the base sequence **GT** in the 5` end and **AG** in the 3` end of all introns.

**Promoters**; sequence of DNA nucleotides up-stream of the initial base of transcription (at which RNA polymerase binds and initiates transcription), control individual gene expression. Two distinct sequences:

- **TATA or hogens box**; between 30-80 nucleotides upstream from the transcription site
- **CAT**; upstream the TATA box.

These together with general transcription factors are responsible for binding of the enzyme RNA polymerase II which is responsible for transcription.
• **Up-stream (5`)**

• Signal for addition of cap (7 methyl guanisine) to the 5` end of mRNA. The cap facilitates the initiation of translation through recognition by ribosomes and helps stabilization of mRNA.

• Addition of cap and tail to primary RNA (hn—RNA) occur after transcription.
- **Down stream (3` end)**
- The translation termination codon TAA.
- AATAAA sequence which is a signal for addition of a poly (A) tail. The addition of poly (A) tail makes mRNA more stable and resistant to digestion by endogenous cellular nucleases.
- Note, poly A tail is not transcribed from DNA but its added after transcription and gradually shorten in the cytoplasm.
- **Terminator**
  A DNA sequence just downstream of the coding segment of a gene, which is recognized by RNA polymerase as a signal to stop transcription.

- **Enhancer**
  A regulatory DNA sequence that greatly enhances the transcription of a gene.

- **Silencer**
  A DNA sequence that helps to reduce or shut off the expression of a nearby gene.
Eukaryote Promoter

- Goldberg-Hogness or TATA located at –30
- Additional regions at –100 and at –200
- Possible distant regions acting as enhancers or silencers (even more than 50 kb).
Gene → RNA

- Exons and introns are numbered in the 5’ to 3’ direction of the coding strand.
- Both exons and introns are transcribed into a precursor RNA (primary transcript).
- The first and the last exons usually contain sequences that are not translated. These are called the 5’ untranslated region (5’ UTR) of exon 1 and the 3’ UTR at the 3’ end of the last exon.
- The noncoding segments (introns) are removed from the primary transcript and the exons on either side are connected by a process called splicing.
Gene → RNA cont...

- Splicing must be very precise to avoid an undesirable change of the correct reading frame.
- Introns almost always start with the nucleotides GT in the 5’ to 3’ strand (GU in RNA) and end with AG.
- The sequences at the 5’ end of the intron beginning with GT are called **splice donor site** and at the 3’ end, ending with AG, are called the **splice acceptor site**.
- Mature mRNA is modified at the 5’ End by adding a stabilizing structure (cap) and by adding many adenines at the 3’ end (polyadenylation)
RNA Modification

- **Trimming**: removing bases from the 5’ and 3’ ends

- **Capping**: adding a methylated G to the 5’ end
  - Necessary for RNA localization to the ribosome

- **Tailing**: addition of A’s to the 3’ end of the mRNA
  - More A’s = more stable mRNA

- **Splicing**: removing introns prior to mRNA transport to the nucleus
Genetic Code
The Genetic Code

• How DNA code for amino acid sequence in polypeptide chain?
• How many nucleotides are necessary for amino acid specification?
• 20 amino acids
• One?
  • 4 bases (AGCU) = 4 possible codons
• Two?
  • 4 bases at two positions = $4^2 = 16$ codons
• Three?
  • 4 bases at three positions = $4^3 = 64$ codons
• A triplet code (codon) is the most efficient way to code for all 20 amino acids
• Shown by Crick et al in 1961
The Genetic Code

<table>
<thead>
<tr>
<th>First base</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>UUC</td>
<td>UUA</td>
<td>UUG</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Leu</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td>CUC</td>
<td>CUA</td>
<td>CUG</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>AUU</td>
<td>AUC</td>
<td>AUA</td>
<td>AUG</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>GUC</td>
<td>GUA</td>
<td>GUG</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Stop Codons**
- Also called termination codons or nonsense codons
Initiation Codons and Reading Frame

• **Initiation codon:** AUG
  – Bacteria: specifies N-formylmethionine
  – Eukaryotes: specifies methionine

• Genetic Code is **non-overlapping**
  – Except in some viruses

• 3 possible **reading frames**
Nucleotide sequence: A U A C G A G U C
Nonoverlapping code: A U A, C G A, G U C, ...
Overlapping code: A U A, C G A, A G U C

3 different reading frames
- tRNAs carry the same amino acid but have different anticodons.
- Sense codons specify an amino acid
- 61 sense codons
- Only 20 amino acids.
- The genetic code is a **degenerate code**
**Wobble**: "Flexibility" in the pairing of the 5′ base of the anticodon with the 3′ base of the codon.
The Genetic Code

- There are $4^3$, (64 codon) total different triplets that can be created but only 20 different amino acids.
- The DNA has a **triplet code** using only the 4 nucleotides, A, C, G, and T. Only 3 nucleotides form a triplet which, when in a gene, codes for a part of a protein.
- The code is **degenerate** i.e. the amino acid may be specified by more than one codon.
- Methionine and tryptophan each is specified by single codon.
- There are **3 stop codons** which are signal for translation termination.
- The genetic code is **universal** i.e. same in all organisms, with exception in mitochondria.
Deviations from the universal genetic code

<table>
<thead>
<tr>
<th>Codon</th>
<th>Universal</th>
<th>Deviation</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>UGA</td>
<td>Stop</td>
<td>Tryptophan</td>
<td><em>Mycoplasma, mitochondria of some species</em></td>
</tr>
<tr>
<td>CUG</td>
<td>Leucine</td>
<td>Threonine</td>
<td><em>Mitochondria in yeast</em></td>
</tr>
<tr>
<td>UAA, UAG</td>
<td>Stop</td>
<td>Glycine</td>
<td><em>Acetabularia, Tetrahymena, Paramecium</em></td>
</tr>
<tr>
<td>UGA</td>
<td>Stop</td>
<td>Cysteine</td>
<td><em>Euplothes</em></td>
</tr>
</tbody>
</table>
(a) gene in DNA (template strand)

(b) mRNA (codons)

(c) tRNA (anticodons)

(d) protein (amino acids)

methionine glycine valine etc.