

## Bioinformatics: Life Sciences on Your Computer

### Eukaryotic Gene Prediction

*This is supplemental reading for the video lecture on Eukaryotic Gene Structure.*

#### Eukaryotic promoter structure

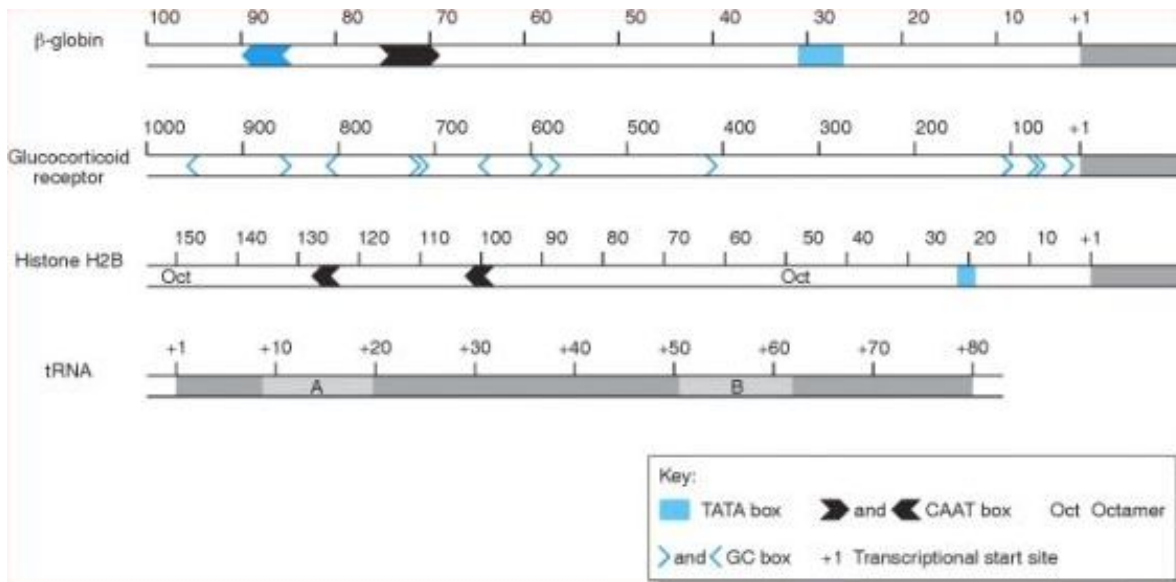
Several upstream elements are found in eukaryotic promoters--but eukaryotes are complex. Promoters vary a lot and many genes have some or none of these elements.

GC box at -90 (20 bp)

CAAT box at -75 (22 bp)

TATA promoter at -30 (15 bp consensus)

Again, -30 means 30 bases upstream from the transcription start site, **NOT 30 bases upstream from the start codon!**



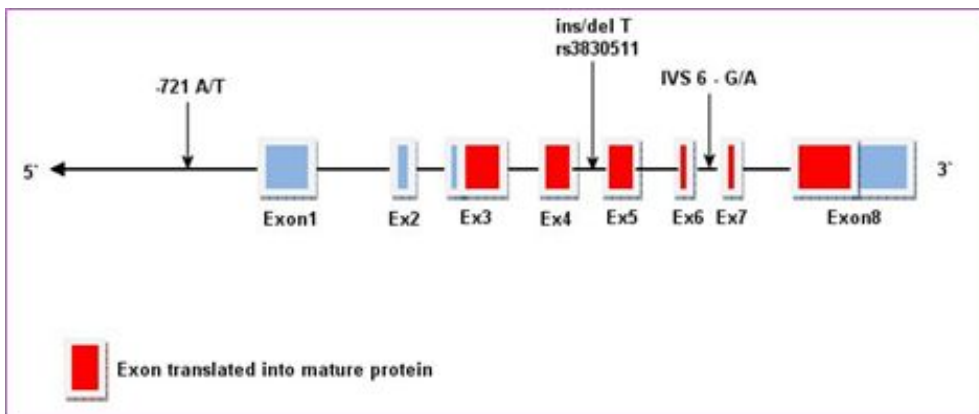
--image from Strachan & Read (1999). *Human Molecular Genetics*, 2nd edition, Wiley-Liss.

From the figure above, you can see how promoters are not one size fits all! There are many varieties.

#### Eukaryotic exon structure

Eukaryotic mRNA has a CAP site (8 bp) on the 5'-end and a polyA signal on the 3'-end. Exons and introns alternate in genomic DNA, but remember that alternative splicing means that more than one different transcript product may be present. The diagram below shows noncoding regions of mRNA in blue and coding regions of mRNA in red. Note that some exons are completely noncoding--that's normal and common.

The straight lines that connect the blue and red boxes represent introns.

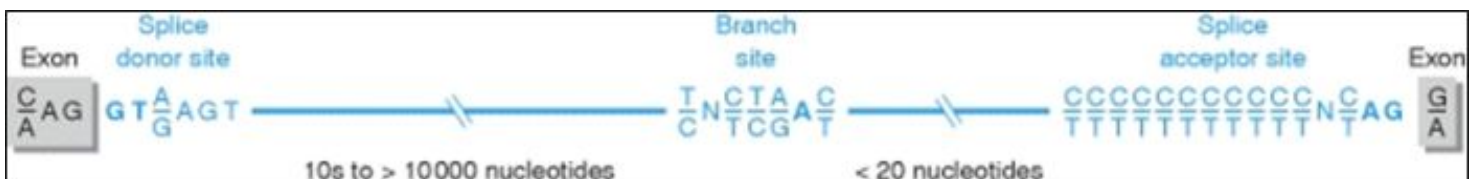


What is VERY important to note is that while the actual gene structure includes the 5' UTR and the 3' UTR, that since those are noncoding regions, gene predictors cannot detect the UTRs well. So gene prediction programs **only predict the CDS regions** of the transcript (the red regions in the diagram above. Gene predictors would ignore exon 1 in that diagram.

### Intron structure

The predicted coding regions are interrupted in genomic DNA by introns. Introns have a very specific structure, particularly exons spliced by the major spliceosome. The majority of transcripts are spliced with catalysis from the major spliceosome.

1. Introns start with a donor site consensus--almost always beginning with GT (GU in RNA).
2. There is a branch site near the 3' end of the intron, with a sequence of TACTAAC, although there is some variation in that sequence.
3. The intron ends in AG



--image from Strachan & Read (1999). *Human Molecular Genetics*, 2nd edition, Wiley-Liss.

The GT and AG are almost invariant, which is a useful clue.

### Exon types

Here are several types of exons found in eukaryotic genes:

**Noncoding exon:** Completely in 3' or 5' UTR, these exons are ignored by *de novo* gene prediction algorithms.

**Initial exon:** Begins with the start codon and ends with the 5' splice site (splice donor site)

**Internal exon(s):** Begin with 3' splice site (splice acceptor site) and end with 5' splice site (splice donor site)

**Terminal exon:** Begins with 3' splice site (splice acceptor site) and ends with stop codon

**Single exon:** In a gene with no introns, the only coding exon is the CDS.

Each exon type is a little different, so assembling a gene structure SHOULD begin with an initial exon and end with a terminal exon.

**Discussion topic: Which exon type would you think would be the easiest to find? Briefly explain.**

### Prominent de novo programs

Here are some prominent *de novo* gene prediction programs:

**GENSCAN:** <http://genes.mit.edu/GENSCAN.html>

**FGENESH:** <http://linux1.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind>

**GeneMark:** <http://opal.biology.gatech.edu/GeneMark/>

**Augustus:** <http://bioinf.uni-greifswald.de/augustus/submission>

**JIGSAW:** <http://www.cbcb.umd.edu/software/jigsaw/>

### Genscan output

There is one thing about Genscan that can be very confusing. Some of the output is marked "Intr," which does NOT stand for "intron." The "Intr" stands for "internal exon!" Here are the exon notations:

**Init** = Initial exon (ATG to 5' splice site)

**Intr** = Internal exon (3' splice site to 5' splice site)

**Term** = Terminal exon (3' splice site to stop codon)

**Sngl** = Single-exon gene (ATG to stop)

**Prom** = Promoter (TATA box / initiation site)

**PlyA** = poly-A signal (consensus: AATAAA)

So, the output below has a four-exon gene on the + strand of DNA:

| Gn.Ex | Type | S | .Begin | ...End | .Len | Fr | Ph  | I/Ac | Do/T | CodRg | P.... | Tscr.. |
|-------|------|---|--------|--------|------|----|-----|------|------|-------|-------|--------|
| 1.01  | Init | + | 1990   | 3577   | 1588 | 0  | 1   | 88   | 109  | 1292  | 0.997 | 122.22 |
| 1.02  | Intr | + | 6449   | 6734   | 286  | 0  | 1   | 108  | 78   | 234   | 0.994 | 21.34  |
| 1.03  | Intr | + | 6964   | 7093   | 130  | 1  | 1   | 65   | 73   | 149   | 0.999 | 12.30  |
| 1.04  | Term | + | 7445   | 7711   | 267  | 1  | 0   | 105  | 37   | 224   | 0.971 | 14.83  |
| 1.05  | PlyA | + | 8051   | 8056   | 6    | 1  | .05 |      |      |       |       |        |

The polyA signal is not an exon or part of the CDS, but does represent the end of the transcript.

### Comparative genomics

Genomic alignment can work well to find genes. For example, compare the mouse and human genomes though alignment. Coding regions (exons) will tend to be more conserved than non-coding regions (introns and other noncoding DNA). Here are some goals of comparative genomics:

**Goal:** Small-scale: To use similar genomes to find (or confirm) the **exon** locations of a gene.

**Goal:** Large-scale: To use similar genomes to identify potentially undiscovered genes.

**Goal:** General: To improve on *de novo* prediction by combining it with alignment to a similar genome.

There is a very good web-based [tutorial](#) on how comparative genomics enhances *de novo* gene prediction. Follow the web address and then follow the directions. It gives a good overview of the process and how **comparative genomics** enhances that process. *In fact, the tutorial in itself is a pretty good lecture, so it's highly recommended.*

