Notes on GenScan

The Two Main Concepts.

1. There is a low-level recognition module ("sensor") for each relevant part of a gene (e.g., promoter, $5'$ UTR, initial exon, etc.) and a high-level procedure that assembles the recognized pieces into predicted genes. The methods for building a low-level procedure can differ from that used for the top-level procedures, such as a "maximal dependence decomposition" for recognizing donor splice sites (i.e., those at the $5'$ ends of *introns*) and a hidden Markov model at the top level.

2. It is useful to think in terms of (low-level and high-level) procedures that *generate* DNA sequences according to different probabilities. Later, we will convert them to methods that *analyze* (i.e., parse) any given DNA sequence.

Some Points Made in "Finding the genes in genomic DNA" by Burge and Karlin.

Bacterial gene-finding. Genes in bacteria (prokaryotes) are easier to find because they have no introns and are packed closely together. However, a bacterium may have a few genes with substantially different properties from most of the genes. For instance, about 15 percent of *E. coli* genes have been "laterally transfered" into the genome from other organisms.

Sensitivity and selectivity. The *sensitivity* of a prediction mechanism (e.g., for coding regions) is defined as the proportion of true sites that are correctly predicted. The *selectivity* of a prediction mechanism is defined as the proportion of predicted sites that are correct. Prediction mechanisms are frequently "tuned" to make their sensitivity approximately equal to their selectivity.

Four kinds of exons: (1) initial (from a start codon to the first $5'$ (donor) splice junction, (2) internal (from a $3'$ splice site to the $5'$ splice site for the next intron), (3) terminal (from a $3'$ splice site to a stop codon) and (4) single-exon (intronless) gene.

Transcriptional signals. There are some weak signals for the start of transcription, including a "TATA box", which is a short T+A-rich region about 30 bp before the transcription start site in about 70% of human genes. About 60% of human genes (probably slightly less in the mouse) have their $5'$ end in or near a CpG island. (GenScan doesn't use CpG islands.) About 50% of human genes have AATAAA about 30 bp upstream of the end of their $3'$ UTR.

Translational signals. The consensus "Kozak signal" for the translation start site is `RNNATGG`, where `R` denotes `A` or `G` and `N` matches any nucleotide. (The first translated codon is the `ATG`.) Certain weaker variants are also found.