Gene Finding

Lecturer: Michael Brudno
scribed by Hui Lan, Feb 17, 2006

Gene finding refers to identifying stretches of sequences (genes) in genomic DNA that are biologically functional. (Figure 1. shows the structure of a protein-coding gene in eukaryotes.) Gene finding is crucial in understanding the genome of a species.

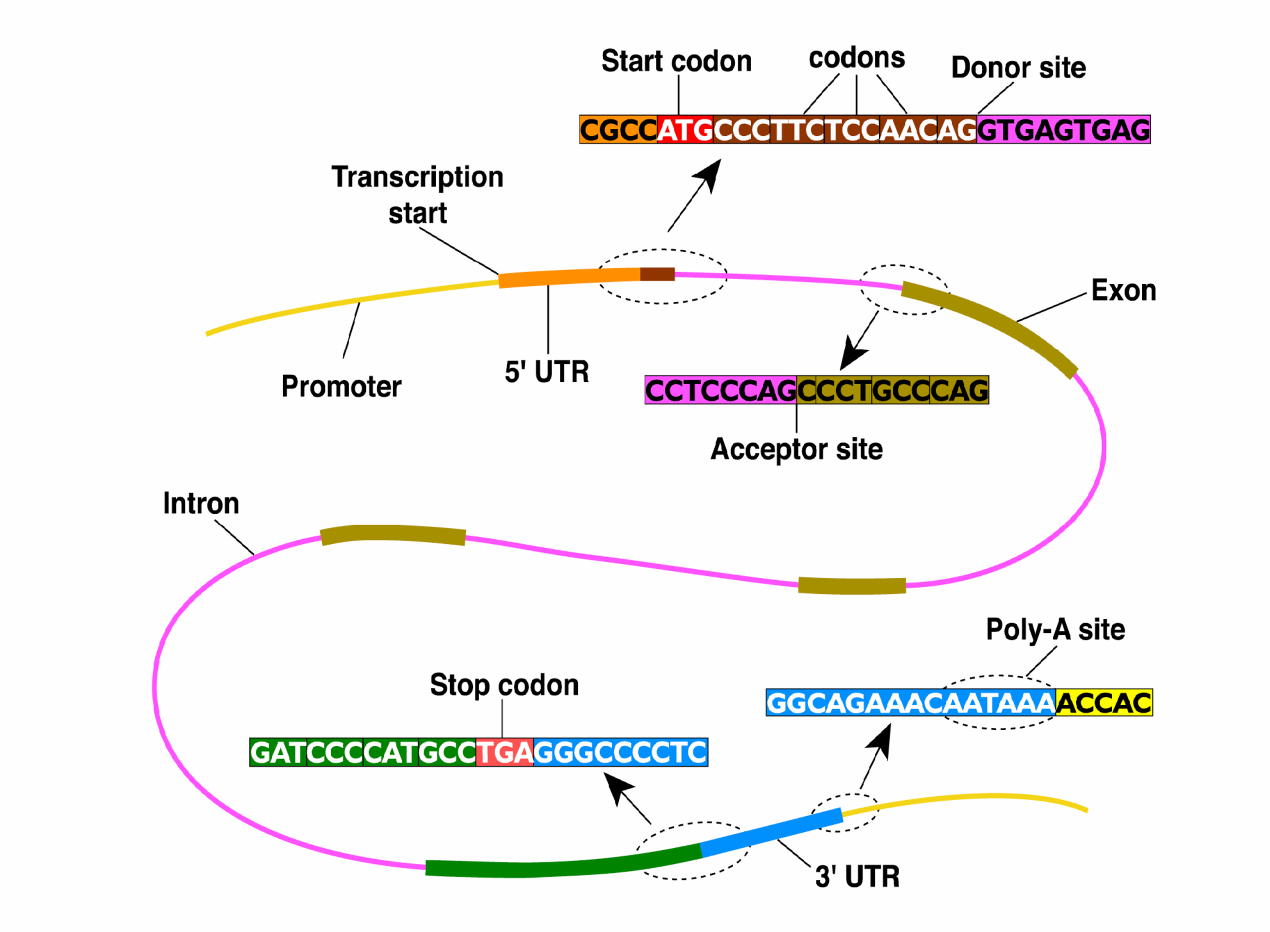


Figure 1: A protein-coding gene. (This picture is from Sanja Rogic's lecture slides, "Computational Gene Finding")

Approaches

Computational methodology for finding genes in a genome has evolved significantly over the last 20 years. Many approaches have been proposed to find genes in both prokaryotes and eukaryotes. These approaches mainly fall into three categories: homology-based approaches, *Ab Initio* approaches and comparative genomics approaches.

1. Homology-based Approaches

These approaches are based on the similarity of sequences. Given a library of sequences of other organisms, we search target sequence in this library and identify library sequences (known genes) that resemble the target sequence. Also, we could compare the target sequence with expressed sequence tags (ETSs) of the same organism to identify regions corresponding to processed mRNA. If the identified sequences are genes, the target sequence is probably (putatively) a gene. These approaches are able to find biologically relevant genes. However, they could not identify genes that code for proteins not already in the library. BLAST (Basic Local Alignment Search Tool) is a well-known search tool in this category.

2. *Ab Initio* Approaches

Ab Initio gene finding searches for certain signals of protein coding genes. There are two types of organisms: Prokaryotes and Eukaryotes.

Prokaryotes have small genomes (0.5 ~ 10,000,000 bp) and high coding density (>90%). There are no introns in Prokaryotes (right part in Figure 2). Gene finding for Prokaryotes is relatively easy since Prokaryotes genes have specific signals such as transcription factor binding site and Pribnow box that are easy to identify. Moreover, the protein-coding sequence is a contiguous open reading frame (ORF), starting with a start codon (ATG) and ending with a stop codon (TAG/TGA/TAA).

However, for Eukaryotes genes, *Ab Initio* gene finding is more difficult for the following reasons. First, genes are separated by large intergenic regions. Second, a gene is not contiguous. The gene is divided into exons and introns by the splicing mechanisms in eukaryotic cells (left part in Figure 2). The split genes make it difficult to define ORFs. Second, the signals (e.g., promoters) are more difficult to identify than that in prokaryotes since these signals are more complex and unspecified. Two such signals are CpG islands and binding sites for a Poly-A tail.

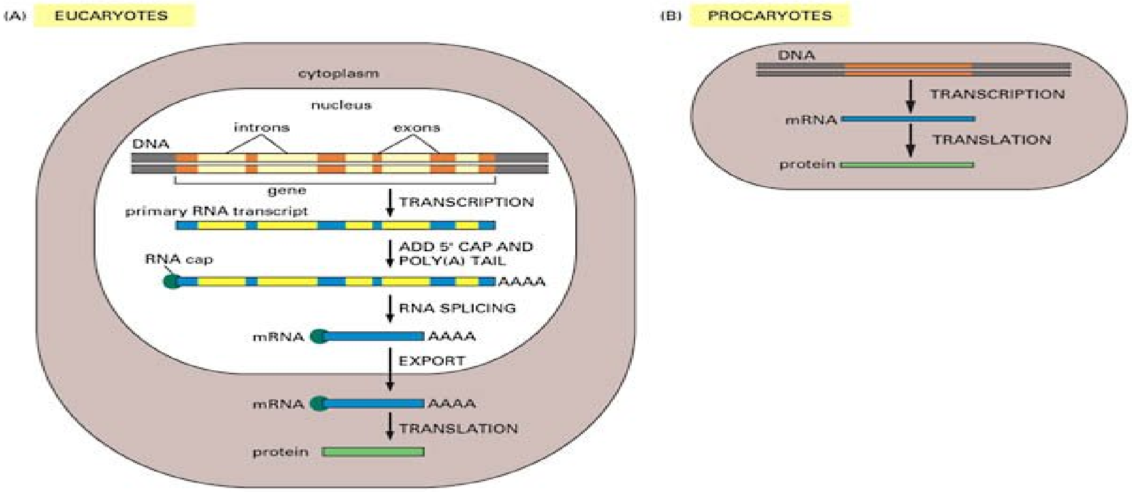


Figure 2: Gene structure for Eukaryotes and Prokaryotes. (This picture is from Sanja Rogic‘s lecture slides, “Computational Gene Finding”.)

HMM (Hidden Markov Model) is widely use for finding both Prokaryotic and Eukaryotic genes. Given a genome ***G*** of length ***L***, HMM outputs the most probable hidden state path ***S*** that generates the observed genome ***G*** using Viterbi algorithm. The probability of the hidden state sequence ***S*** given ***G*** is computed using the following Bayes’ rule, ***P{S|G} = P{S,G}/Σ P{S’,G}***, where ***S’*** is in the set of all the possible hidden state path of length ***L***.

Figure 3 shows the HMM for a gene finder called GenScan (Chris Burge 1997). Each diamond or circle is a state of a genomic region that corresponds to a basic functional unit of a eukaryotic gene, e.g., intergenic region, exon, intron and so on.

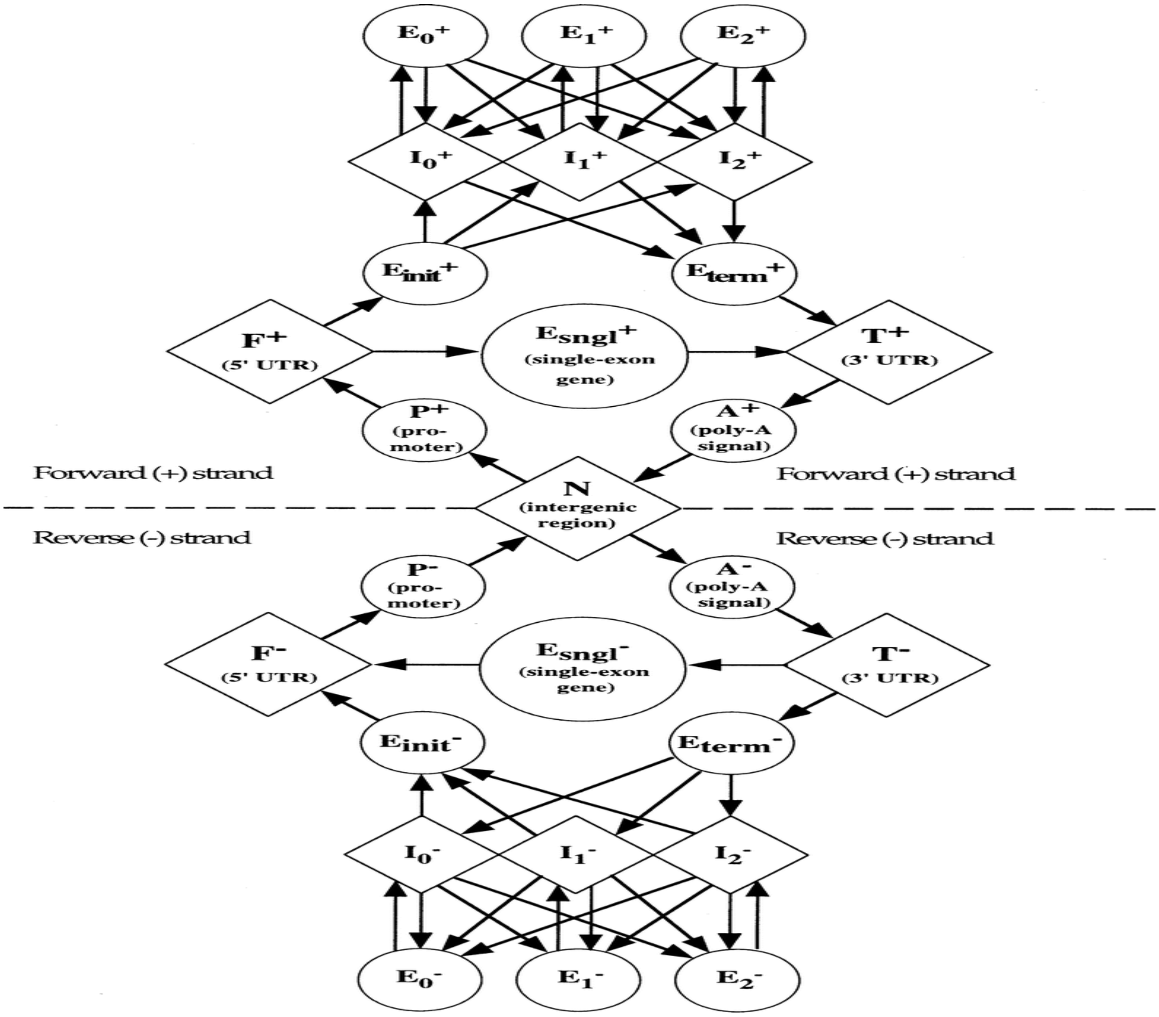


Figure 3: Hidden Markov Model in GeneScan.

The upper half in Figure 3 shows the states of a gene on the forward strand (with a superscript '+' symbol in each state).; the lower half in Figure 3 shows the staes of a gene on the reverse (complementary) strand (with a superscript '-' symbol in each state). This symmetric structure allows GenScan to deal with forward strand states and reverse strand states simultaneously.

We explain the meaning of each state in Figure 3 (Chris Burge 1997). N is intergenic region. P is promoter. F is 5' untranslated region from the start of transcription to the translation initiation signal. T is 3' untranslated region from just after the stop codon to the poly-A signal. GenScan also explicitly models single, intial, internal and terminal exons. E_{sngl} is single-exon (prokaryotic) genes (translation start to stop codon). E_{init} is initial exon (acceptor splice site to donor splice site). E_k (0<=k<=2) is phase k internal exon. I_k (0<=k<=2) is phase k intron. E_{term} is terminal exon (acceptor splice site to stop codon). For convenience, GenScan treats donor and acceptor sites, translation initiation/termination signals as subcomponents of the associated exons. GenScan also considers intron states extending from just after a donor splice site to just before the branch acceptor/point site.

GenScan has explicit structure to account for the frame position of "interrupting" intron (Chris Burge 1997). It tracks "phase" (closely related to reading frame) of exons or introns. Internal exons and introns are divided according to "phase". An intron that is between codons is in phase 0 (I_0). An intron that is after the first base of a codon is in phase 1 (I_1). An intron that is after the second base of a codon is in phase 2 (I_2). Similarily, internal exons are divided according to the phase of the previous intron, giving states $E_{0,}$ E_1, and E_2. For example, initial exons start at codon position p_1 and end at codon position p_i such that *i mod 3* is the phase of the following intron state. Internal exons E_i start at codon position p_{i+1} and end at codon position p_j , such that *j mod 3* is the phase of following intron. The terminal exons begin at codon position p_{i+1} and end at codon position p_3, where i is the phase of the previous intron.

The advantages of GenScan include (1) It models single and multi-exon genes; (2) It models explicitly the length of introns and exons; (3) It models promoters, poly-A signals and intergenic sequences; (4) It allows forward and backward genes (i.e., genes from both one strand of a DNA and the complementary strand of a DNA). (5) It has advanced splice site modeling; (6) It is able to find multiple genes and partial or whole genes. The limitations of GenScan include (1) It could not handle overlapping transcription units; (2) It dose not address explicitly alternative splicing.

Beside GenScan, other HMM-based gene finders are FGENESH (Solovyev, 1997), HMMgene (Krogh, 1997), GENIE (Kulp 1996), GENMARKER (Borodovsky & Mclninch 1993) and VEIL (Henderson, Salzberg & Fasman 1997). Other tools include TwinScan, N-Scan and SLAM.

3. Comparative Genomics Approaches

Comparative genomics approaches to find genes are based on the obervation that the force of natural selection makes genes and other funcitonal elements mutates at a slower rate than the other parts of a genome. With more genomes sequenced in related species, genes can be identified by comparing these genomes to detect this conservation (Figure 5).

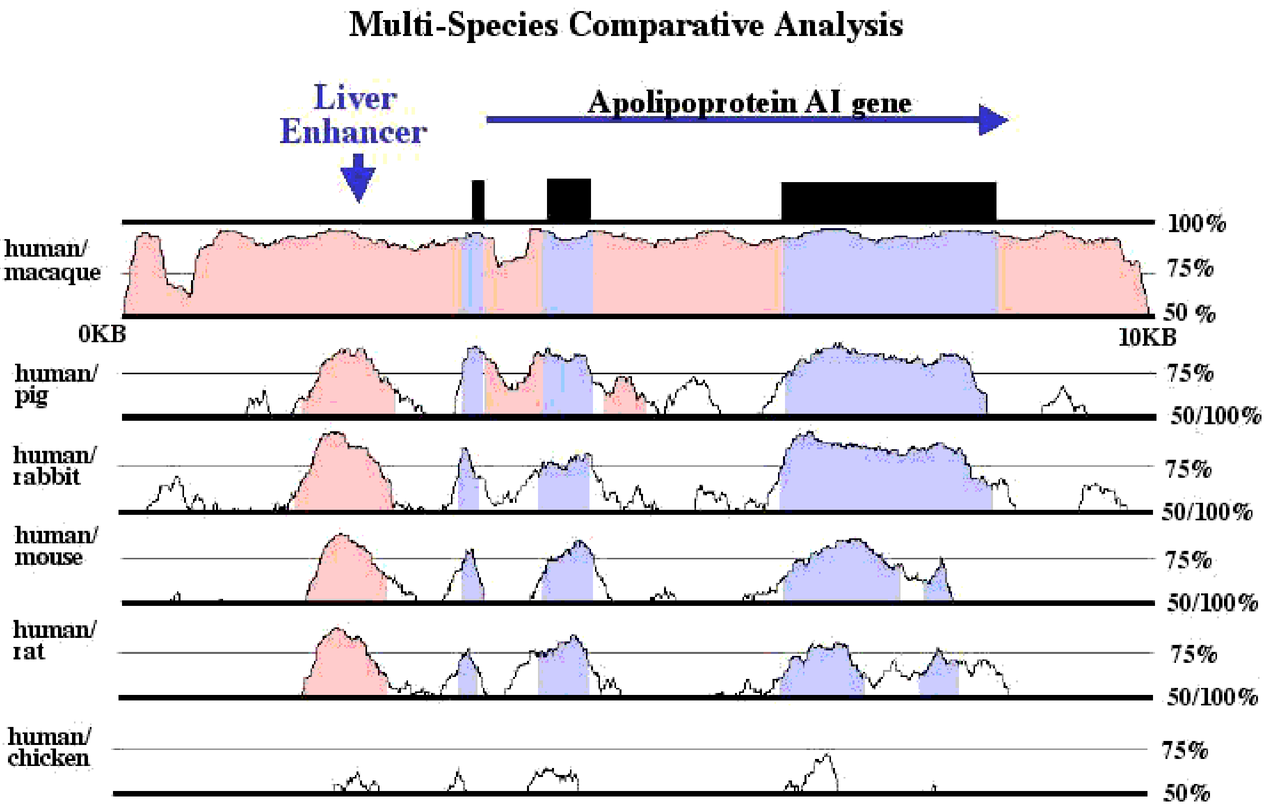
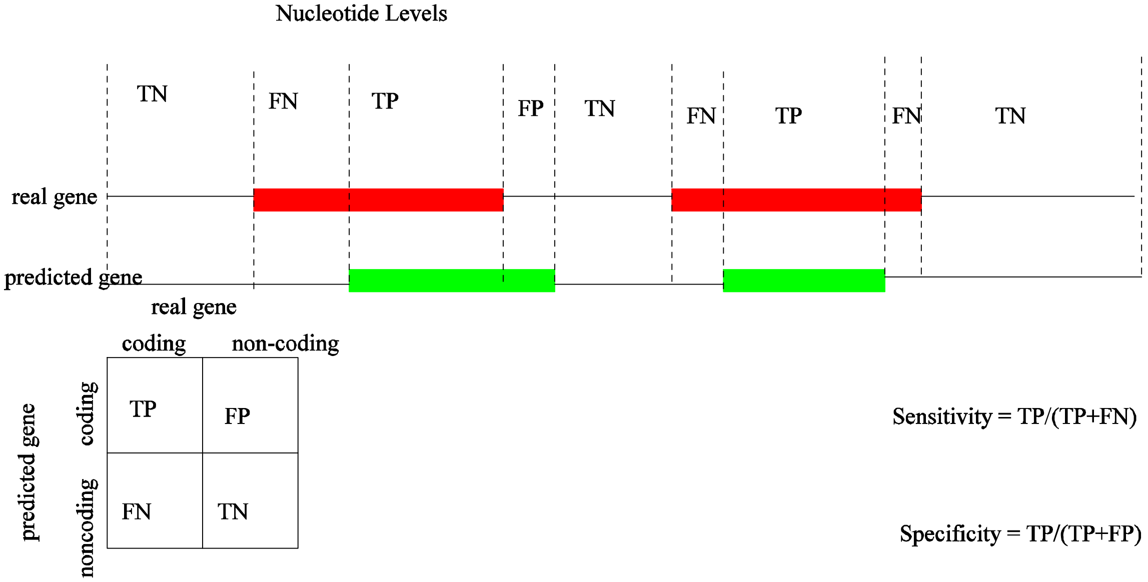


Figure 4: Multi-Species Comparative Analysis. (This picture is from Sanja Rogic‘s lecture slides, “Computational Gene Finding”)

4. Evaluation of methods

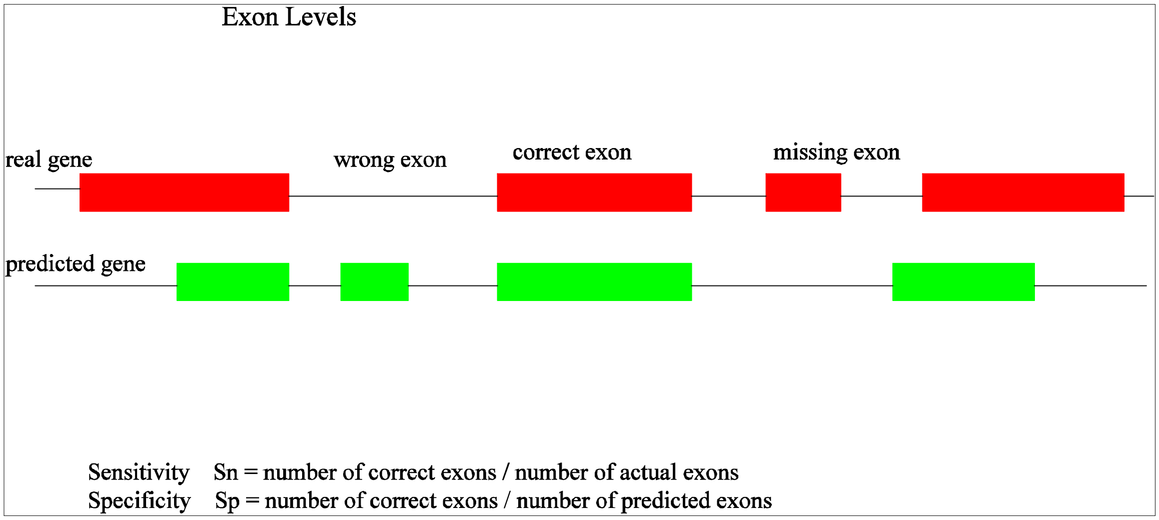
Sensitivity (Sn) and specificity (Sp) are two common measures for a gene finding tool. The performance of gene finding tool can be evaluated at nucleotide level and at exon level. Accuracy at nucleotide level reflects how close the predicted sequence and real coding sequence in an alignment (Figure 5). Accuracy at exon level reflects how well signals (e.g., start codons, stop codons, and spice sites) are identified (Figure 6).



Sensitivity = TP/(TP+FN)

Specificity = TP/(TP+FP)

Figure 5: Accuracy at nucleotide level. (This picture is adapted from Iosif Vaisman's lecture slides, "Bioinformatics and Gene Discovery".)



Sensitivity Sn = number of correct exons / number of actual exons
Specificity Sp = number of correct exons / number of predicted exons

Figure 6: Accuracy at exon level. (This picture is adapted from Iosif Vaisman's lecture slides, "Bioinformatics and Gene Discovery".)

References:

1. Allen, J.E., et al. (2004) Computational gene prediction using multiple sources of evidence. Genome Research. 14: 142-148 .
2. Burge C., Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology. 268(1):78-94.