Lecture 8
Gene Prediction

Saleet Jafri
BINF 730

Gene Prediction

- Analysis by sequence similarity can only reliably identify about 30% of the protein-coding genes in a genome
- 50-80% of new genes identified have a partial, marginal, or unidentified homolog
- Frequently expressed genes tend to be more easily identifiable by homology than rarely expressed genes

Gene Finding

- Process of identifying potential coding regions in an uncharacterized region of the genome
- Still a subject of active research
- There are many different gene finding software packages and no one program is capable of finding everything

Genes aren't the only thing we're looking for

•Biologically significant sites include:

•Splice sites

•Protein binding sites

•DNA 3D structure features

•etc.

In a lot of cases, we don't even know what constitutes one of these sites, so all we can do is look for repeating patterns

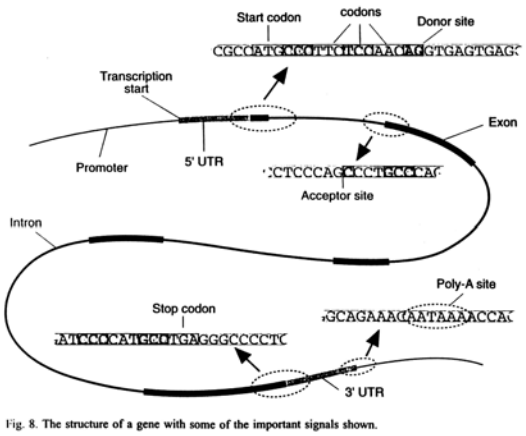


Fig. 8. The structure of a gene with some of the important signals shown.

Eukaryotes vs Prokaryotes

- Eukaryotic DNA wrapped around histones that might result in repeated patterns for histone binding. The promotor regions might be near these sites so that they remain hidden.
- Prokaryotes have no introns.
- Promotor regions and start sites more highly conserved in Prokaryotes
- Different codon use frequencies

Gene finding is species-specific

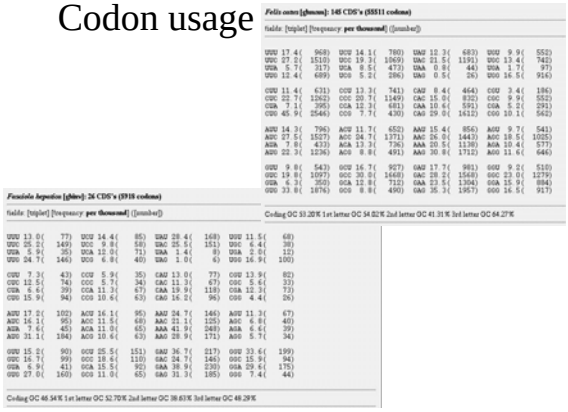
- Codon usage patterns vary by species
- Functional regions (promoters, splice sites, translation initiation sites, termination signals) vary by species
- Common repeat sequences are species-specific
- Gene finding programs rely on this information to identify coding regions

The genetic code

Table of Standard Genetic Code

	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

Codon usage

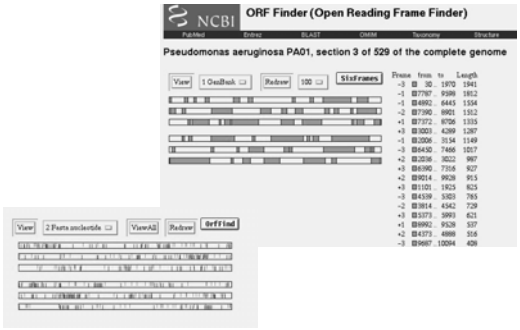


Identifying ORFs

- Simple first step in gene finding
- Translate genomic sequence in six frames. Identify stop codons in each frame
- Regions without stop codons are called "open reading frames" or ORFs
- Locate and tag all of the likely ORFs in a sequence
- The longest ORF from a Met codon is a good prediction of a protein encoding sequence.
- SOFTWARE: NCBI ORF Finder

ORF Finder input

ORF finder results



Tests of the Predicted ORF

- Check if the third base in the codons tends to be the same one more often than by chance alone.
- Are the codons used in the ORF the same as those used in other genes (need codon usage frequency).
- Compare the amino acid sequence for similarity with other know amino acid sequences.

Problems with ORF finding

- A single-character sequencing error can hide a stop codon or insert a false stop codon, preventing accurate identification of ORFs
- Short exons can be overlooked
- Multiple transcripts or ORFs on complementary strand can confuse results

Pattern-based gene finding

- ORF finding based on start and stop codon frequency is a pattern-based procedure
- Other pattern-based procedures recognize characteristic sequences associated with known features and genes, such as ribosome binding sites, promoter sites, histone binding sites, etc.
- Statistically based.

Content-based gene finding

- Content-based gene finding methods rely on statistical information derived from known sequences to predict unknown genes
- Some evaluative measures include: "coding potential" (based on codon bias), periodicity in the sequence, sequence homogeneity, etc.

A standard content-based alignment procedure

- Select a window of DNA sequence from the unknown. The window is usually around 100 base pairs long
- Evaluate the window's potential as a gene, based on a variety of factors
- Move the window over by one base
- Repeat procedure until end of sequence is reached; report continuous high-scoring regions as putative genes

Combining measures

- Programs rarely use one measure to predict genes
- Different values are combined (using probabilistic methods, discriminant analysis, neural net methods, etc.) to produce one "score" for the entire window

Drawbacks to window-based evaluation

- A sequence length of at least 100 b.p. is required before significant information can be gained from the analysis
- Results in a +/- 100 b.p. uncertainty in the start site of predicted coding regions, unless an unambiguous pattern can also be found to indicate the start.

Most are web-based, but...

- Submit sequence; input sequence length may be limited
- Select parameters, if any
- Interpret results
- Most software is first or second generation; results come in non-graphical formats.

GRAIL

- Gene finder for human, mouse, arabidopsis, drosophila, E. coli
- Based on neural networks
- Masks human and mouse repetetive elements
- Incorporates pattern-based searches for several types of promoters and simple repeats
- Accuracy in 75-95% range

Glimmer

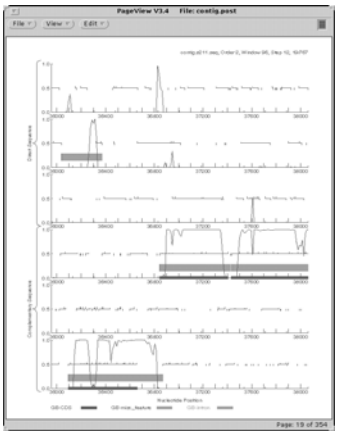
- Genefinder for bacterial and archaebacterial genomes
- Uses an "interpolated Markov model" approach (a Markov model is a model for computing probabilities in the context of sequential events)
- Predicts genes with around 98% accuracy when compared with published annotations
- No web server

GENSCAN

- Genefinder for human and vertebrate sequences
- Probabilistic method based on known genome structure and composition: number of exons per gene, exon size distributions, hexamer composition, etc.
- Only protein coding genes predicted
- Maize and arabidopsis-optimized versions now available
- Accuracy in 50-95% range

GeneMark

- Gene finder for bacterial and archaebacterial sequences
- Markov model-based
- GeneMark and GeneMarkHMM available as web servers
- Accuracy in 90-99% range



CRITICA

- Gene finder for bacterial and archaebacterial genomes
- Combines sequence homology-based prediction with content-based statistical (dicodon probability) analysis
- Accuracy in 90-99% range
- No web server

GeneParser

- Predicts the most likely combination of exons and introns using dynamic programming.
- The intron an exon positions are aligned subject to the constraint that they alternate.
- A neural network is used to adjust the weights given to the sequence indicators of know exon and intron regions such as codon usage, information content, length distribution, hexamer frequencies, and scoring matrices.

Other software

- Generation
- GeneID
- Genie
- GenView
- EcoParse
- etc...

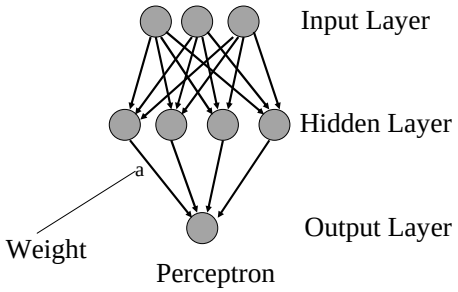
tRNAscan

- Locating tRNA genes is less difficult than other types of gene identification
- pol III promoter is simple; RNA secondary structure is conserved
- SOFTWARE: tRNAscan-SE

Gene finding strategy for beginners

- Choose the appropriate type of gene finder! Make sure that you're using gene finders for microbial (intronless) sequences only to analyze bacteria and archaea!
- If there is no organism-specific gene finder for your system, at least use one that makes sense (i.e. use an arabidopsis gene finder for other plants)

Neural Network Topology



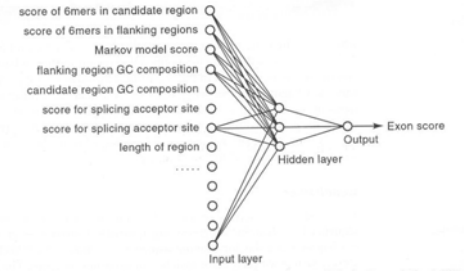
Making Neural Networks

- Take known data and divide into two sets: the *training* set and *test* set.
- Use the optimize the weights so that the neural net gives the best outputs for the training set.
- Test the neural net with the test set to see if it works
- If data is limited, you can permute the data so that you have multiple training and test sets

Caveats with Neural Nets

- The net only performs as well as the training set.
- In other words, it can only find things it is trained to do.
- As more diverse data becomes available, the neural net gets better

Grail II Neural Net



- Finds exons in eukaryotic genes, that is, takes inputs and predicts if a gene is present.

Markov Model

- A process is Markov if it has no memory, that is, if the next state it assumes, depends only on its present state and not on any previous states.
- The states can be observed and the transition probabilities between states is known
- Example – rolling a die has 6 possible states each with a probability of 1/6

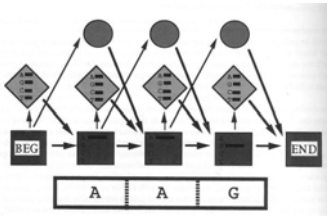
Hidden Markov Model

- Also has the Markov property.
- Some of the state or transition probabilities information is missing.
- The process emits sequences of results.
- The emission probabilities is the probability of each outcome in a given state.
- The model is trained so that the training set is the most likely outcome for the model

Training and Testing the HMM

- The parameters of the model are fit on a training set, ie., the parameters are chosen so that the training set is the most likely outcome for the model.
- A test set is used to make sure the model is well-trained.
- If so, the model can be used on new data.

HMM of *E. Coli* Gene



- HMM for finding the most probable set of genes in *E. coli* gene sequences of unknown gene composition.
- A similar model exists for each of the 61 codons

HMM of *E. Coli* Genes

- Assumes that there is no relationship each codon and codons used later in the sequence.
- This assumption works, however, analysis of sequential codons in a gene have shown that some pairs are found at greater/lesser frequencies than would occur at random.
- GeneMark.HMM uses sequence information from the previous 5 bases instead of the previous 2 bases.