

## 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

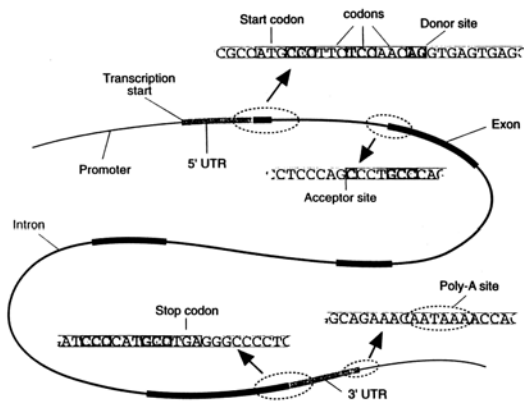


Fig. 8. The structure of a gene with some of the important signals shown.

## Eukaryotes vs Prokaryotes

- Eukaryotic DNA wrapped around histones that might result in repeated patterns for histone binding. The promoter regions might be near these sites so that they remain hidden.
- Prokaryotes have no introns.
- Promoter 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)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)
	TTC "	TCC "	TAC "	TGC "
	TTA Leu (L)	TCA "	TAA <b>Ter</b>	TGA <b>Ter</b>
	TTG "	TCG "	TAG <b>Ter</b>	TGG Trp (W)
C	CTT Leu (L)	CCT Pro (P)	CAT His (H)	CGT Arg (R)
	CTC "	CCC "	CAC "	CGC "
	CTA "	CCA "	CAA Gln (Q)	CGA "
	CTG "	CCG "	CAG "	CGG "
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)
	ATC "	ACC "	AAC "	AGC "
	ATA "	ACA "	AAA Lys (K)	AGA Arg (R)
	<b>ATG Met (M)</b>	<b>AGG "</b>	<b>AAG "</b>	<b>AGG "</b>
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)
	GTC "	GCC "	GAC "	GGC "
	GTA "	GCA "	GAA Gln (E)	GGA "
	GTG "	GGG "	GAG "	GGG "

## Codon usage

Full table (40000): 146 CDS's (88111 codons)  
 Table [table] [frequency per thousand] [download]

```

UUU 17.4( 969) UUU 14.1( 780) UAU 12.2( 683) UUG 9.9( 552)
UUC 17.5( 970) UUA 15.3( 849) UAU 21.5( 1193) UUG 18.4( 1025)
UUA 5.7( 317) UUA 8.5( 473) UAA 0.8( 44) UUA 1.7( 97)
UUG 12.4( 689) UUG 5.2( 286) UAG 0.5( 28) UUG 16.9( 949)

CUU 11.4( 631) CUU 13.3( 743) CUA 8.4( 464) CUG 9.9( 552)
CUC 12.7( 1262) CUC 25.7( 1409) CAC 15.0( 832) CCG 9.9( 552)
CUA 7.1( 395) CUA 12.3( 681) CAA 10.6( 593) CAA 5.2( 292)
UUG 45.9( 2545) CUG 7.7( 430) UAG 29.0( 1613) CUG 10.1( 562)

AUU 14.2( 796) AUU 11.7( 652) AAU 15.4( 856) AUA 9.7( 543)
AUC 27.2( 1527) AUC 28.7( 1717) AAC 26.2( 1443) AAC 18.5( 1035)
AAA 7.0( 437) AAA 11.3( 736) AAA 10.5( 1139) AAA 10.4( 577)
AUG 22.3( 1236) AUG 9.8( 493) AAG 30.6( 1712) AUG 11.6( 646)

GUU 9.8( 547) GUU 14.7( 927) GAU 17.7( 983) GUA 9.2( 510)
GUC 19.8( 1097) GUC 39.0( 1608) GAC 28.2( 1568) GAC 23.0( 1279)
GUA 6.3( 360) GUA 12.8( 712) GAA 22.6( 1306) GAA 15.9( 884)
GUG 33.8( 1876) GUG 9.8( 490) GAG 35.3( 1957) GUG 14.5( 917)
  
```

Partial sequence (40000): 26 CDS's (1918 codons)  
 Table [table] [frequency per thousand] [download]

Coding GC 51.28% 1st letter GC 54.82% 2nd letter GC 41.31% 3rd letter GC 54.23%

```

UUU 13.0( 77) UUU 14.4( 88) UAU 28.4( 168) UUG 11.5( 68)
UUC 15.3( 149) UUA 9.4( 59) UUA 25.4( 151) UUA 6.4( 39)
UUA 8.9( 26) UUA 12.0( 71) UAA 1.4( 8)
UUG 24.7( 146) UUG 6.8( 40) UAG 1.0( 6) UUG 14.9( 100)

CUU 7.3( 43) CUC 5.9( 36) CUA 13.0( 77) CUG 13.9( 82)
CUC 18.5( 74) CUC 5.7( 34) CUA 11.3( 67) CUG 15.6( 73)
CUA 6.4( 39) CUA 11.3( 67) CAA 19.9( 119) CAA 12.3( 73)
CUG 15.9( 94) CUG 10.4( 63) CUG 15.2( 90) CUG 4.4( 26)

AUU 17.0( 107) AUU 14.1( 85) AAU 24.7( 146) AUA 11.3( 67)
AUA 14.1( 85) AAC 11.5( 69) AAC 21.1( 125) AAC 6.8( 40)
AAA 7.4( 45) AAA 12.0( 65) AAA 42.9( 249) AAA 6.4( 39)
AUG 11.1( 104) AUG 10.4( 63) AUG 29.9( 171) AUG 5.7( 34)

GUU 15.2( 90) GUU 25.5( 153) GAU 26.7( 177) GUA 9.2( 199)
GUC 16.7( 99) GUC 19.4( 119) GAC 14.7( 146) GAC 12.9( 84)
GUA 6.9( 41) GUA 15.5( 92) GAA 39.9( 239) GAA 29.4( 175)
GUG 27.9( 169) GUG 11.0( 65) GAG 31.3( 195) GUG 7.4( 44)
  
```

Coding GC 46.54% 1st letter GC 52.70% 2nd letter GC 38.61% 3rd letter GC 48.29%

## 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

NCBI ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Seqapp sequence submission software.

Enter GI or ACCESSION:  OffFind Clear

or sequence in FASTA format

FROM:  TO:

Genetic codes:  Standard

## ORF finder results

NCBI ORF Finder (Open Reading Frame Finder)

Pseudomonas aeruginosa PA01, section 3 of 529 of the complete genome

View 1 OpenBack Prev 100 518frames

Frame	From	To	Length
-3	8	30	1970 1941
-1	87787	9398	1852
-1	84892	6443	1554
-2	87790	8901	1352
-1	87772	8796	1332
-3	85003	4289	1287
-1	82056	5136	1149
-2	86430	7486	1027
-2	82206	2622	987
-3	86390	7338	927
-2	89214	9928	915
-1	81101	1823	825
-3	84339	3383	785
-2	82814	6542	739
-1	85373	3993	621
-1	88992	9528	537
-2	84373	4886	516
-3	89987	10004	428

View 2 Parts on table ViewAll Prev OffFind

## 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 known 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 repetitive elements
- Incorporates pattern-based searches for several types of promoters and simple repeats
- Accuracy in 75-95% range

### Glimmer

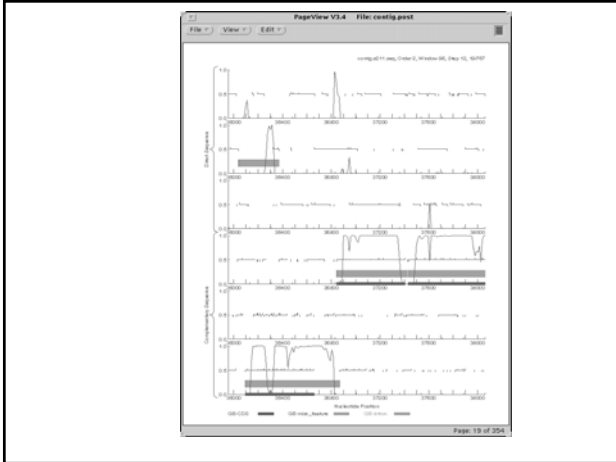
- Genefinder for bacterial and archaeobacterial genomes
- Uses an "interpolated Markov model" approach (a Markov model is a model for computing probabilities in the context of sequential events)
- Predicts genes with around 98% accuracy when compared with published annotations
- No web server

### 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 archaeobacterial sequences
- Markov model-based
- GeneMark and GeneMarkHMM available as web servers
- Accuracy in 90-99% range



## CRITICA

- Gene finder for bacterial and archaeobacterial 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 and 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 known 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...

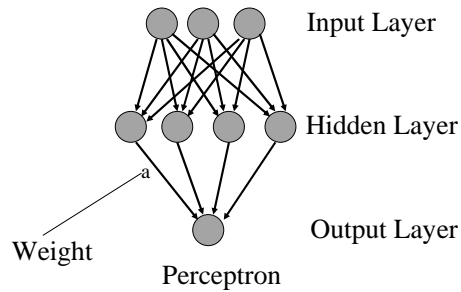
## tRNAscan

- 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



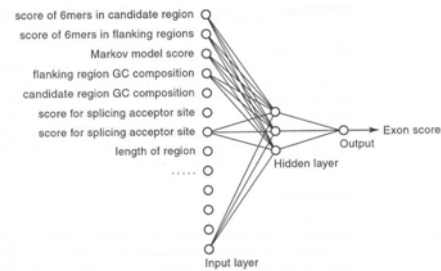
## Making Neural Networks

- 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

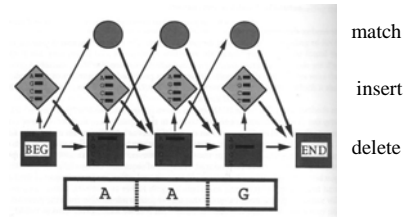
## Hidden Markov Model

- 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, i.e., 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.