

Comparative Genomics & Annotation

The Foundation of Comparative Genomics

The main methodological tasks of CG Annotation:

Protein Gene Finding

RNA Structure Prediction

Signal Finding

Overlapping Annotations:

Protein Genes

Protein-RNA

Combining Grammars

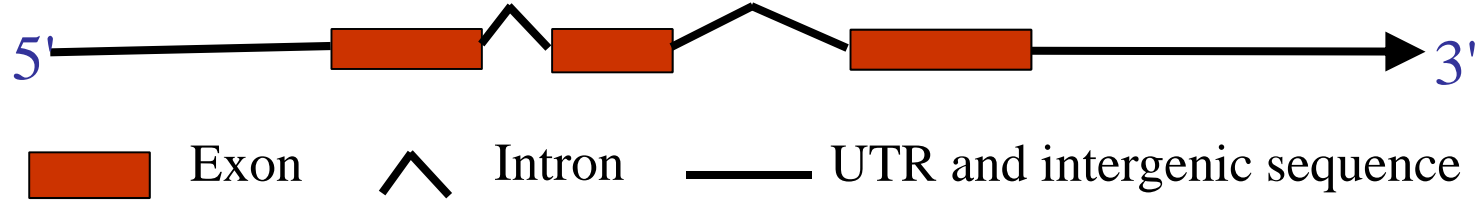
Ab Initio Gene prediction

Ab initio gene prediction: prediction of the location of genes (and the amino acid sequence it encodes) given a **raw DNA sequence**.

... tttttgcagtagctcccgggccctctgttggggcctccccttctctccaggggtggagtcgaggaggcgggggtgccccctccttatctctagagccggccctggctctctggcgccggggcccttagtccgggctttttgccatggggctctctgttccctctgtcgctgctgttttttttggcggccgcctaccgggagttgggagcgcgctgggacgccggactaagcggggcgaagccccaagggtagccctctcgcgccctccgggacctcagtgcccttctgggtgcgcatgagcccggagttcgtggctgtgcagccggggaagtcaagtgcagctcaattgcagcaacagctgtccccagccgcagaattccagcctccgcagcccgctgcccccaaggcaaacgctcagagtcgggggtgggtgtcttaccagctgctcgacgtgagggccctggagctccctcgcgactgctcgtgacctgcgaggaacacgctgggcatctcagatcactgctacagtgaggatggggtctcccggctgggggtgaggggagggggctggaagaggtgggggaaggtagttgacagtcgctctatagggagcctgggtgggctctcaggggtccccttggctggcagcctggagcgtatcttggagcctccggctctaaagggcaggaatacactttgcgctgccacgtgacgcaggtgttcccgggtgggctacttgggtgggtgacctgagggcatggaagccgggtcatctattccgaaagcctggagcgttaccggcctggatctggccaacgtgacctgacctacgagtttgctgctggaccccgcgacttctggcagcccgtgatctgccacgcgcgctcaatctcgacggcctgggtggctcgcaacagctcggcaccattacactgatgctcgggtgagggacccctgtaaccctggggactaggaggaagggggcagagagagttatgaccccagaggggcgcacagaccaagcgtgagctccacgcgggtcgacagacctccctgtgttccgcttctaattctcgccttctgctcccagcttggagccccgcgcccacagcttggcctccgggtccatcgctgcccttgtagggatcctcctcactgtggggcgtgctgacctatgcaagtgccctagctatgaagtcccaggcgtaaaggggatgttctatgcccggctgagcgagaaaaagaggaatatgaaacaatctggggaaatggccatacatggtg...

Input data

Output:

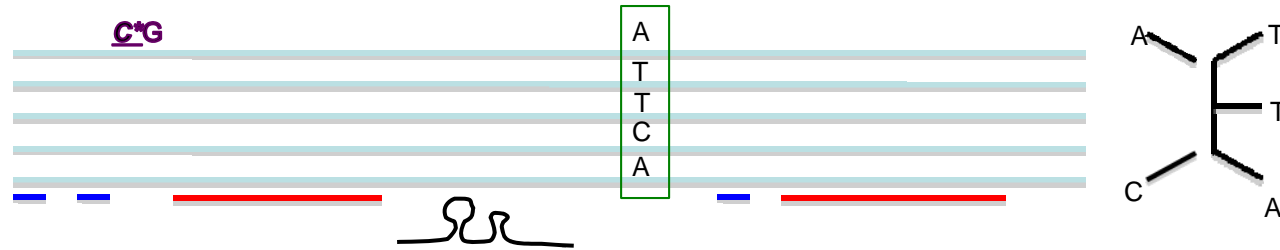


5'... tttttgcagtagctcccgggccctctgttggggcctccccttctctccaggggtggagtcgaggaggcgggggtgccccctccttatctctagagccggccctggctctctggcgccggggcccttagtccgggctttttgcccATGGGGTCTCTGTTCCCTCTGTCGCTGCTGTTTTTTTTTGGCGGCCGCCTACCCGGGAGTTGGGAGCGCGCTGGGACGCCGGACTAAGCGGGCGCAAAGCCCAAGGGTAGCCCTCTCGCGCCCTCCGGGACCTCAGTGCCCTTCTGGGTGCGCATGAGCCCAGGAGTTCGTGGCTGTGCAGCCGGGGAAGTCAGTGCAGCTCAATTGCAGCAACAGCTGTCCCCAGCCGCAAAATTCAGCCTCCGCACCCCGCTGCGGCAAGGCAAGACGCTCAGAGGGCCGGGTTGGGTGTCTTACCAGCTGCTCGACGTGAGGGCCTGGAGCTCCCTCGCACTGCCTCGTGACCTGCGCAGGAAAAACACGCTGGGCCACCTCCAGGATCACCGCTACAgtaggggacaggggctcgggtcccggctgggggtgaggggagggggctggaagaggtggggaaaggtagttgacagtcgctctatagggagcggcccgacactcactcagaggtccccttgccttagAACCGCCCCACAGCGTGATTTTGGAGCCTCCGGTCTTAAAGGGCAGGAAATACACTTTGCGCTGCCACGTGACGCAGGTGTTCCCGGTGGGCTACTTGGTGGTGACCCTGAGGCATGGAAGCCGGGTATCTATTCCGAAAGCCTGGAGCGCTTACCCGGCCTGGATCTGGCCAACGTGACCTTGACCTACGAGTTTGCTGCTGGACCCCGGACTTCTGGCAGCCCCGTGATCTGCCACGCgcgctcaatctcgacggcctgggtggctcggcaacagctcggcaccattacactgatgctcgggtgagggacccctgtaaccctggggactaggaggaagggggcagagagagttatgaccccagaggggcgcacagaccaagcgtgagctccacgcgggtcgacagacctccctgtgtccgcttctaattctcgccttctgctcccagcttggagccccgcgcccacagcttggcctccgggtccatcgctgcccttgtagggatcctcctcactgtggggcgtgctgacctatgcaagtgccctagctatgaagtcccaggcgtaaaggggatgttctatgcccggctgagcgagaaaaagaggaatatgaaacaatctggggaaatggccatacatggtg.... 3'

Levels of Annotation

“Annotation”: Tagging regions and nucleotides with information about function, structure, knowledge, additional data,....

Homologous Genomes



Annotation levels

Protein coding genes including alternative splicing

RNA structure

Regulatory signals – fast/slow, prediction of TF, binding constants,...

Selection Strength,...

Epigenomics – methylation, histone modification

Further complications

Integration of levels – RNA structure of mRNA, signals in coding regions,..

Knowledge and annotation transfer – experimental knowledge might be present in other species

Evolution of Feature – regulatory signals > RNA > protein

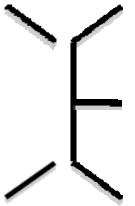
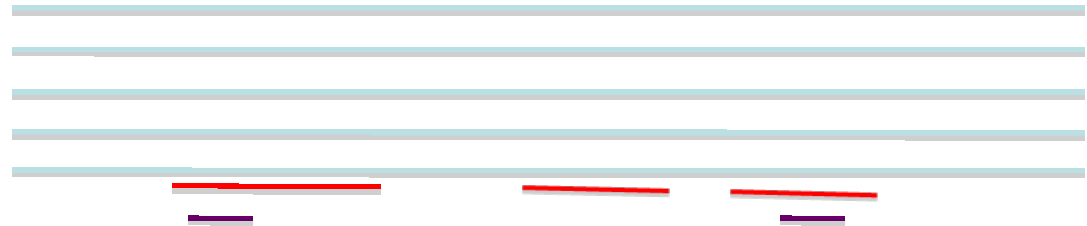
Combining with non-homologous analysis – tests for common regulation.

Combining specie and population perspective

Observables, Hidden Variables, Evolution & Knowledge

Observables

$$P(X) = \pi(X)$$

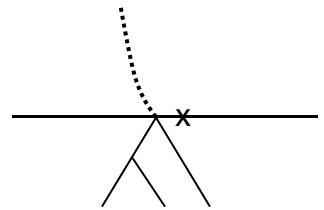


Hidden Variable

$$P(X) = \sum_H P(X|H)P(H) = \sum_H \pi(X|H)P(H)$$

Evolution

$$P(X) = \sum_H \pi(X|H)P(X_{dyna}|H)P(H)$$



Knowledge (Constraints)

$$P(X) = [PW]^{-1} \sum_H P(X|H)P(H)w(H) \stackrel{\text{If knowledge deterministic}}{\downarrow} [PW]^{-1} \sum_{H \cap w=1} P(X|H)P(H)$$

Genscan

Exons of phase 0, 1 or 2

■ State with length distribution

Initial exon

Introns of phase 0, 1 or 2

Terminal exon

Exon of single exon gene

5' UTR

3' UTR

Poly-A signal

Promoter

Forward (+) strand

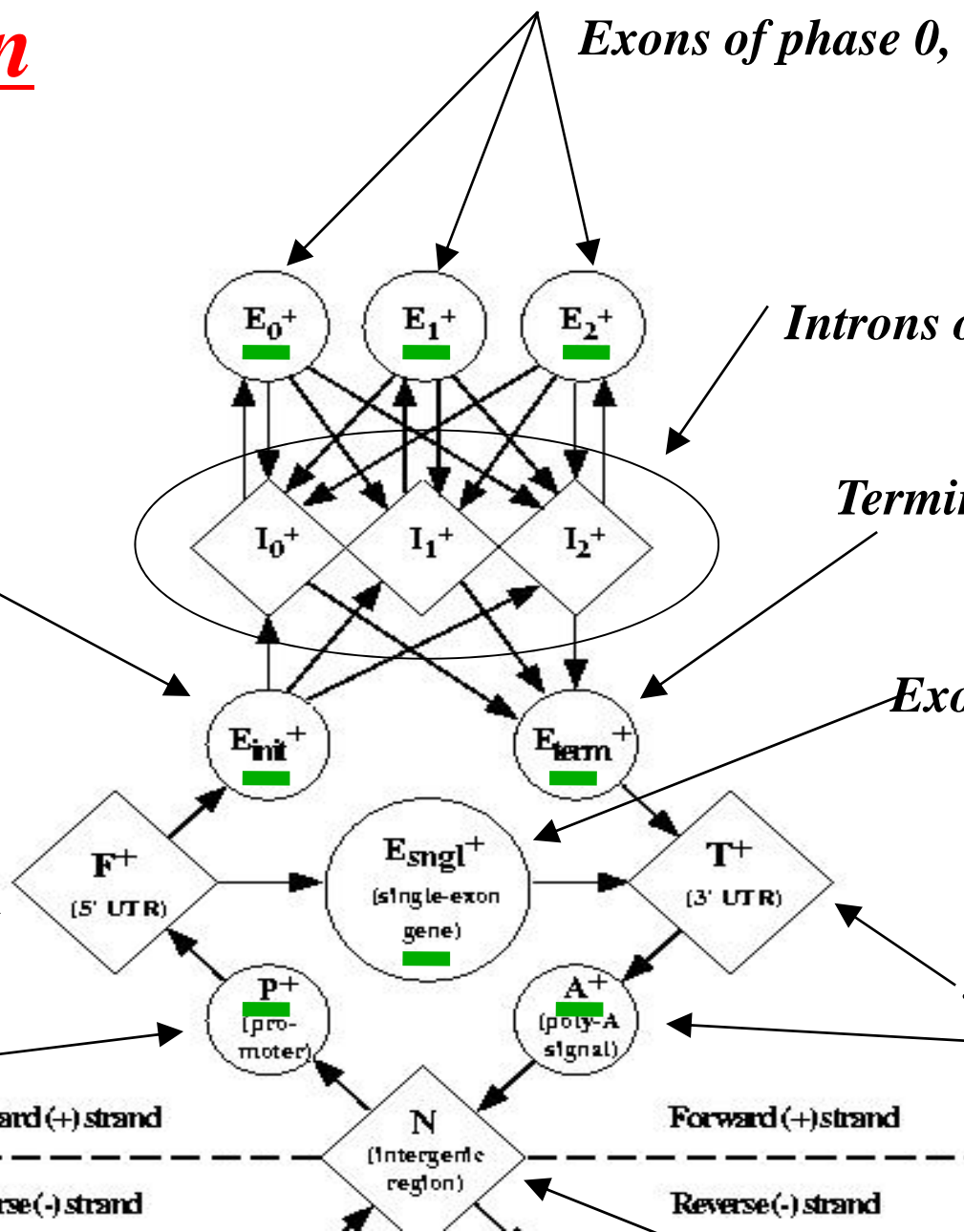
Forward (+) strand

Reverse (-) strand

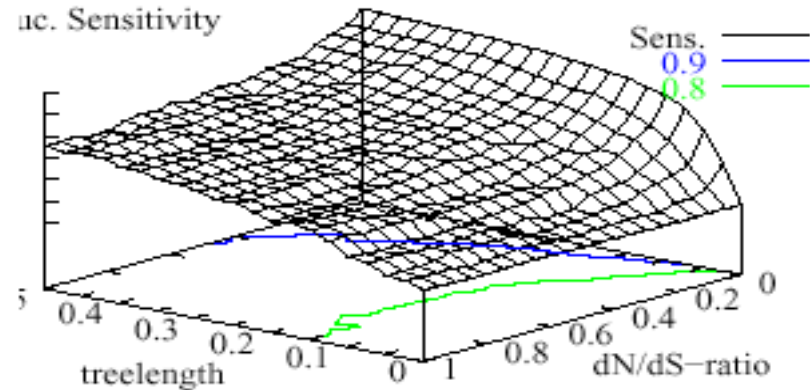
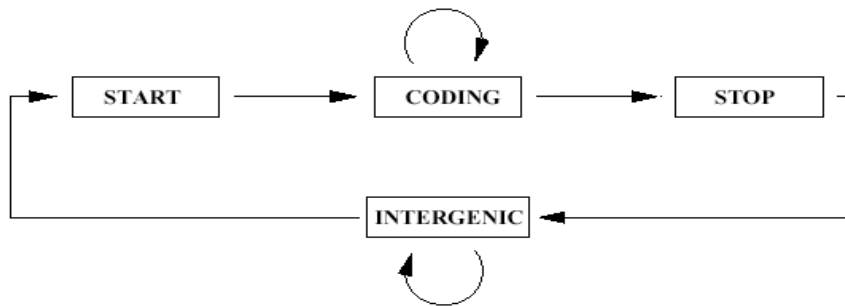
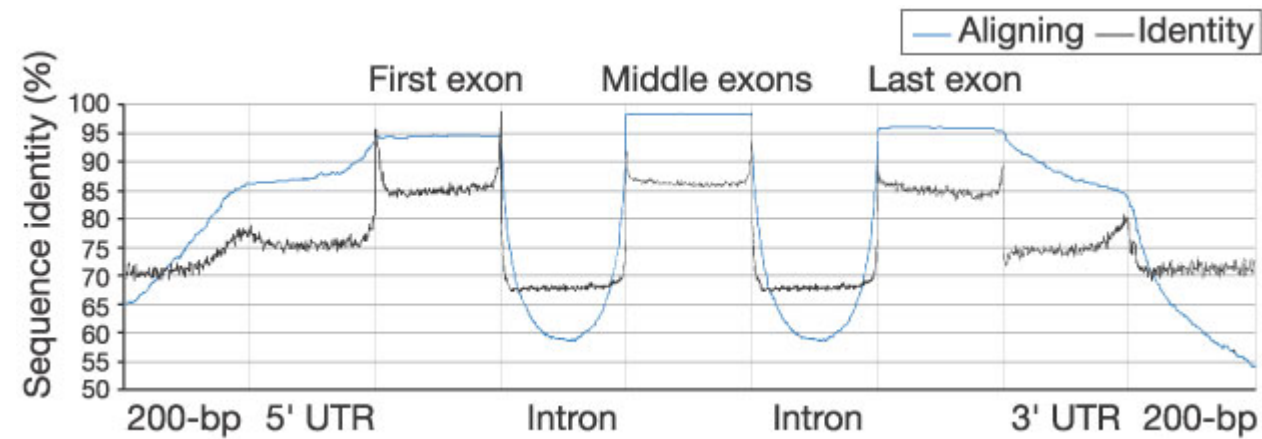
Reverse (-) strand

Omitted: reverse strand part of the HMM

Intergenic sequence



Comparative Gene Annotation



AGGTATATA**ATGCG**..... $P_{\text{coding}}\{\text{ATG} \rightarrow \text{GTG}\}$ or
 AGCCATTTA**GTGCG**..... $P_{\text{non-coding}}\{\text{ATG} \rightarrow \text{GTG}\}$

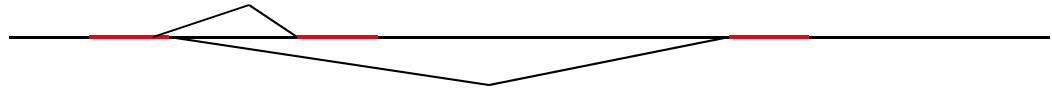
Gene Finding & Protein Homology

(Gelfand, Mironov & Pevzner, 1996)

Protein Database



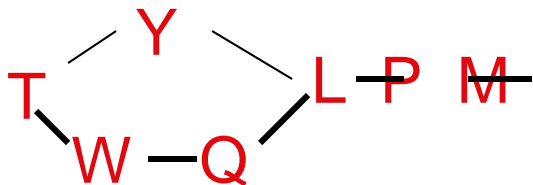
Exon Ordering Graph



Spliced Alignment:

1. Define set of potential exons in new genome.
2. Make exon ordering graph - EOG.
3. Align EOG to protein database.

TYGHL P



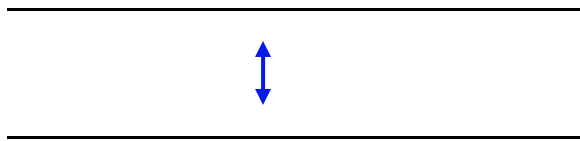
TYGHL P

TY - - L P M

Simultaneous Alignment & Gene Finding

Bafna & Huson, 2000, T.Scharling,2001 & Blayo,2002.

Align by minimizing Distance/
Maximizing Similarity:



Align genes with structure
Known/unknown:

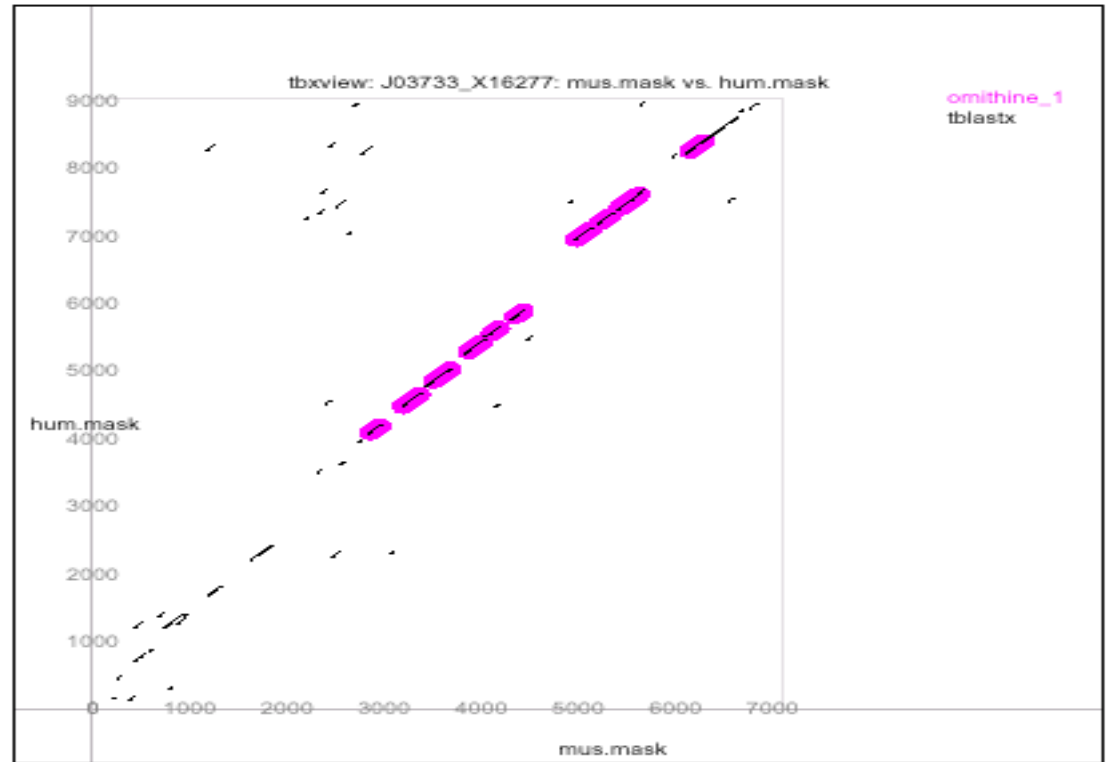
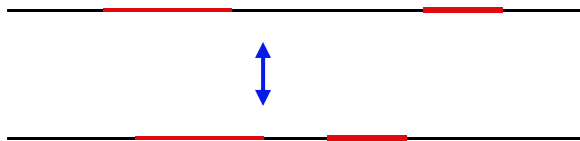


Figure 1: The horizontal axis represents 7100 bps of mouse DNA and the vertical axis represents 9043 bps of human DNA (Genbank accession numbers: J03733 and X16277). Thin black diagonal lines represent HSPs computed by TBLASTX (version 1.4.6) on the +/+ strands. Thick gray diagonal lines represent the exons of ornithine-1. Repeat-Masker (Smit & Green 1995) was used to mask repeats in both sequences and only HSPs with score 17 or more are shown. This plot was produced by the author's own software.

Secondary Structure Generators

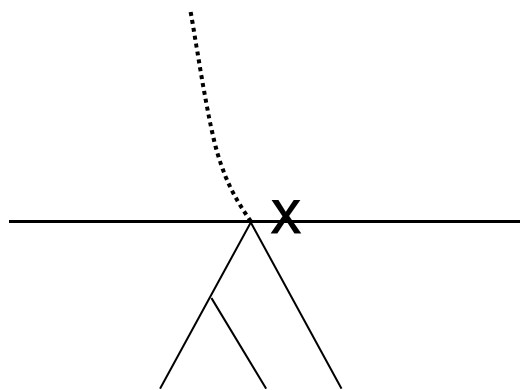
S	-->	LS L	.869	.131
F	-->	dFd LS	.788	.212
L	-->	s dFd	.895	.105

S → *LS* → *LLLLLLLS* → *LLLLLLLLL*
 → *ssLsssss* → *ssdFdsssss*
 → *ssdddFdddsssss*
 → *ssdddLSdddsssss*
 → *ssdddLLLdddsssss*
 → *ssdddssssdddsssss*

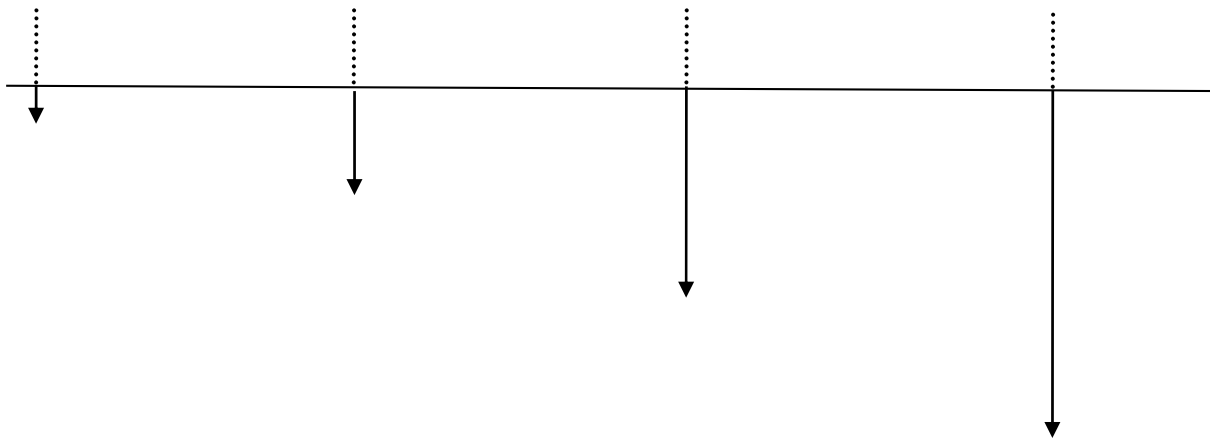
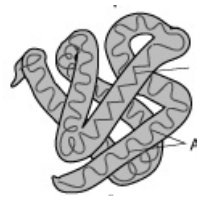
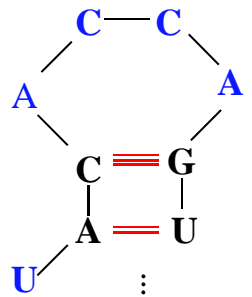
s *ss* *s*
d-d
d-d
ss *d-d* *sssss*

Observing Evolution has 2 parts

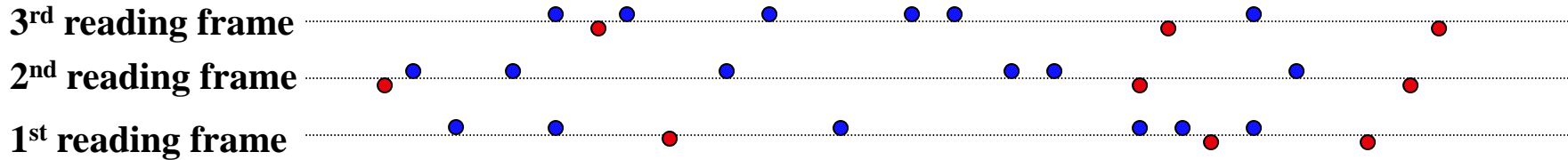
$P(x)$:



$P(\text{Further history of } x)$:



Hidden Markov Model for Overlapping Genes

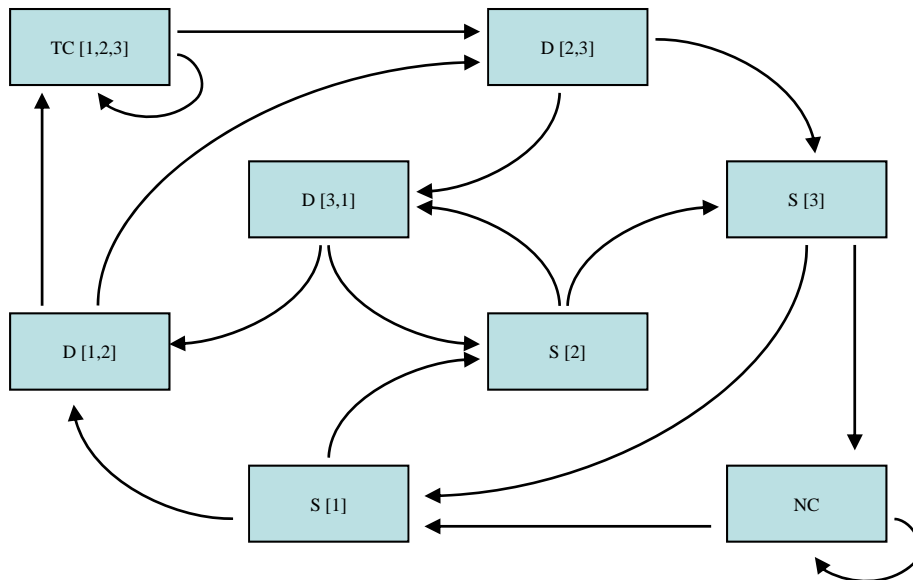


Virus genome

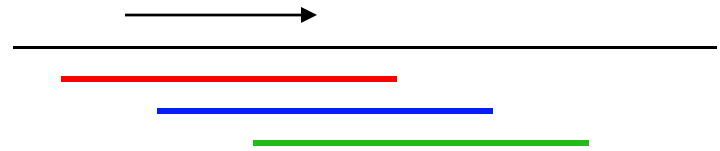


Hidden States

Amplification



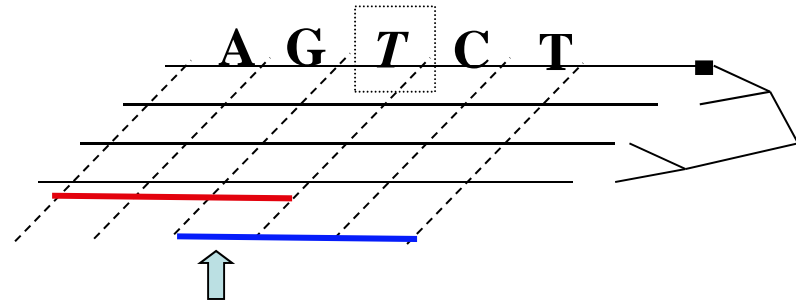
Scanning



- Only starts in AUG (0.06) ●
- Will Stop in "STOP" (1.0) ●

Molecular Evolution: Known Reading Frames

Known fixed context throughout phylogeny



Assume multiplicativity of selection factors

$$f_{i,j}^{A,B} = f_{i,j}^A f_{i,j}^B$$

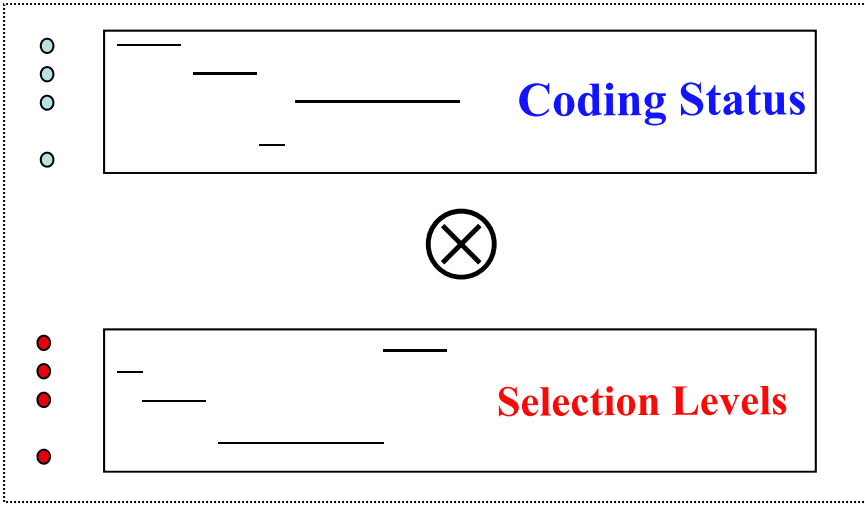
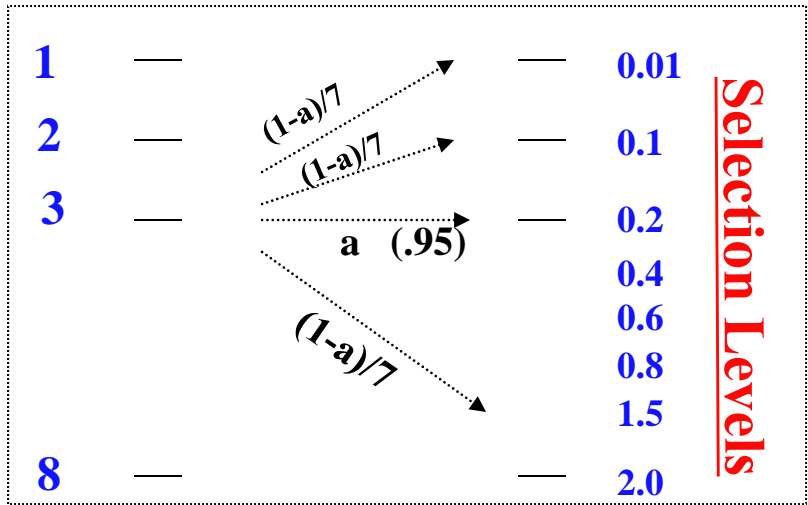
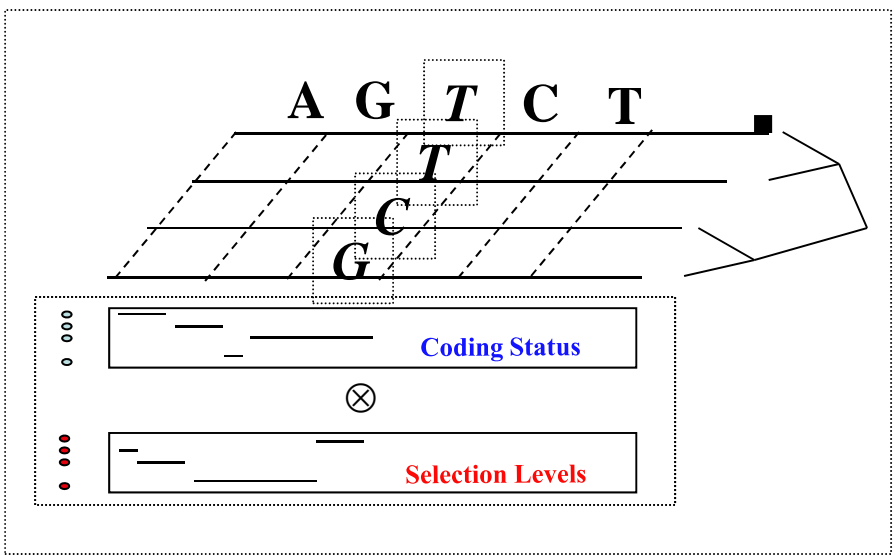
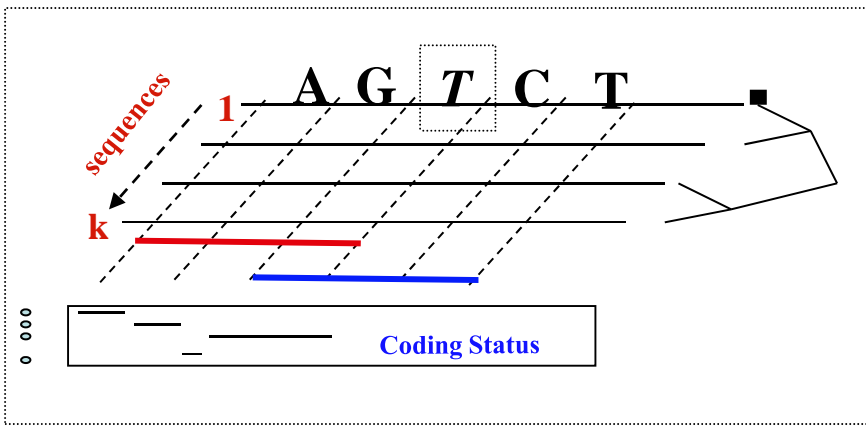
Selection rates on rates

$$q_{i,j} f_{i,j} : \begin{pmatrix} & q_{AC} f_{A,C} & q_{AG} f_{A,G} & q_{AT} f_{A,T} \\ q_{CA} f_{CA} & & q_{CG} f_{CG} & q_{CT} f_{CT} \\ q_{GA} f_{GA} & q_{GC} f_{GC} & & q_{GT} f_{GT} \\ q_{TA} f_{TA} & q_{TC} f_{TC} & q_{TG} f_{TG} & \end{pmatrix}$$

Simplify Genetic Code:

	2 nd	1 st			
4-fold			1-1-1-1	2-2	4
2-fold	1-1-1-1 sites		(f ₁ f ₂ a, f ₁ f ₂ b)	(f ₂ a, f ₁ f ₂ b)	(f ₂ a, f ₂ b)
	2-2		(f ₁ a, f ₁ f ₂ b)	(f ₂ a, f ₁ f ₂ b)	(a, f ₂ b)
(1-1-1-1)	4		(f ₁ a, f ₁ b)	(a, f ₁ b)	(a, b)

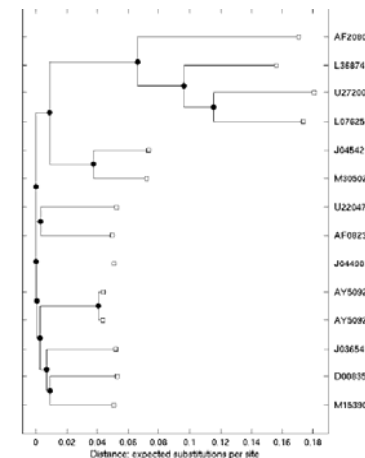
Un-known Reading Frames and varying selection.



HIV2 of 14 genomes: Evolution/Selection

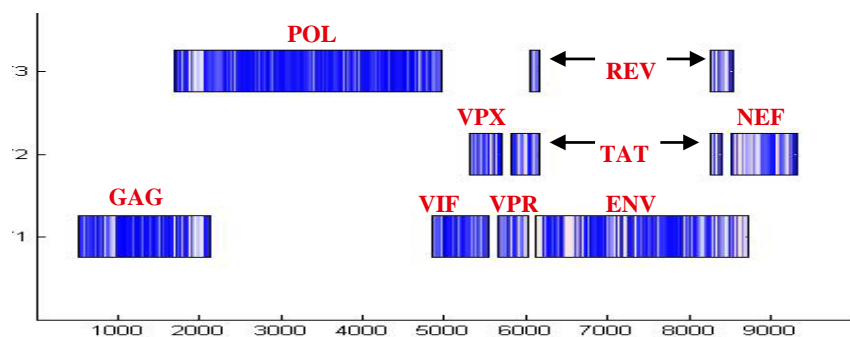
A. Phylogeny and Evolutionary Parameters.

Parameter	Estimate	+/- 1.96 Error
Transition	5.79	0.19
Transversion	1.03	0.05
Base SF	0.73	0.06
SF STOP	0.44	0.18
α	0.95	0.02



B. Selection Strengths for Genes and Positions

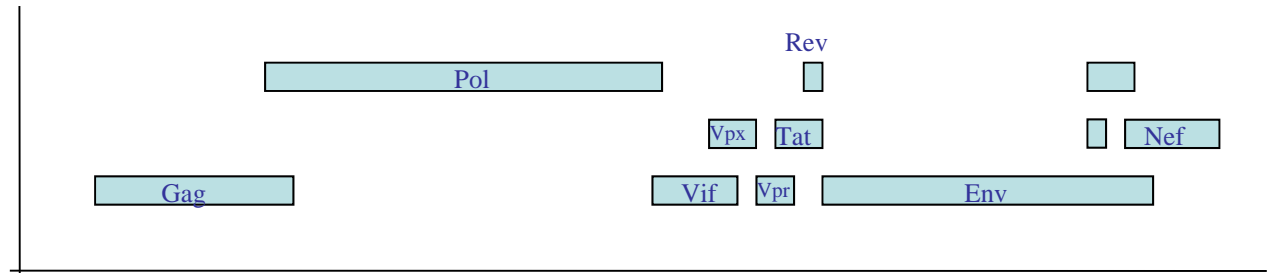
Rate Class	Single Coding	Double Coding	Triple Coding
0.0066	19.06%	5.71%	2.89%
0.066	21.06%	7.98%	4.13%
0.132	14.98%	8.40%	6.33%
0.264	10.53%	9.33%	10.77%
0.396	8.53%	10.98%	14.39%
0.528	8.20%	17.77%	18.00%
0.99	6.79%	22.01%	21.62%
1.32	10.86%	17.90%	22.91%



Rate Class	GAG	POL
0.0066	21.42%	21.38%
0.066	22.52%	25.21%
0.132	15.27%	17.85%
0.264	10.13%	10.18%
0.396	5.96%	7.42%
0.528	6.47%	6.86%
0.99	11.99%	7.10%
1.32	6.26%	4.00%

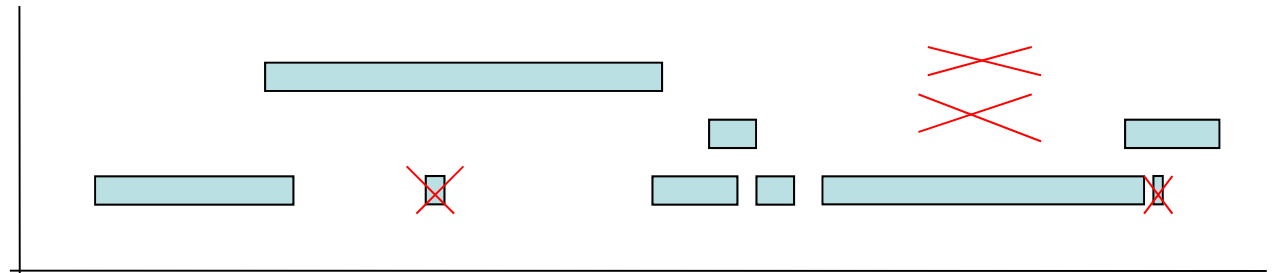
HIV2 of 14 genomes: Annotation

GenBank



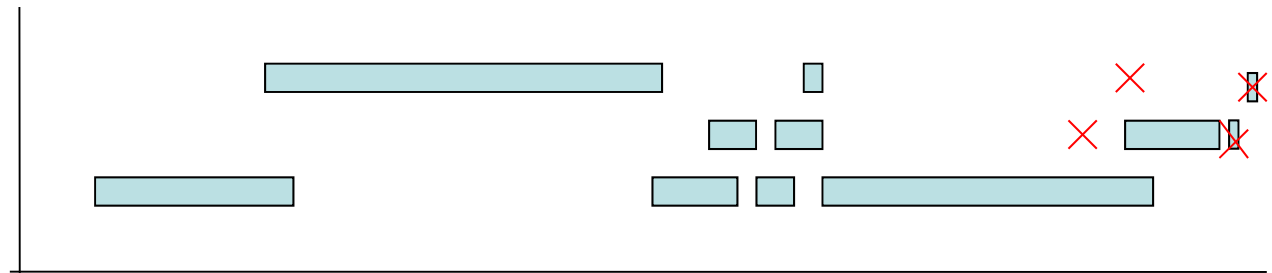
Single Sequence

Sensitivity: 0.9308
Specificity: 0.9939
LogLikelihood: -34939.32
ViterbiCont.: -34949.41

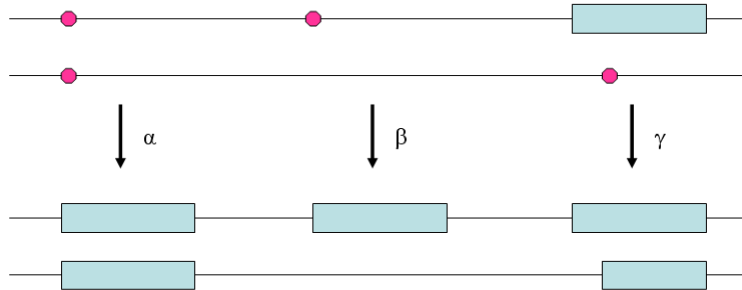
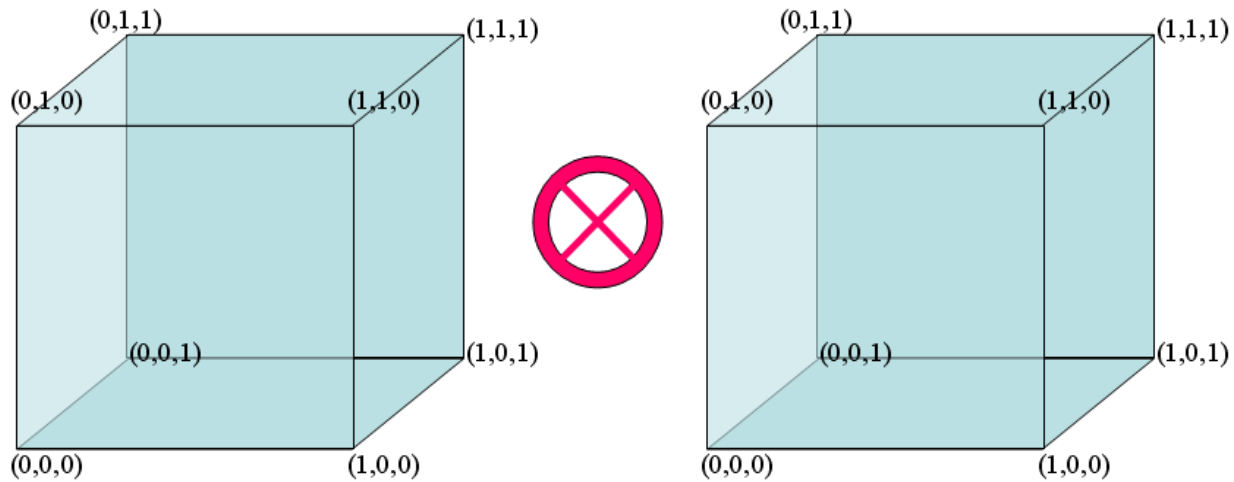


Phylo-HMM

Sensitivity: 0.9542
Specificity: 0.9965
LogLikelihood: -75939.18
ViterbiCont.: -75945.77



HMM extension: Stop/Start Skidding

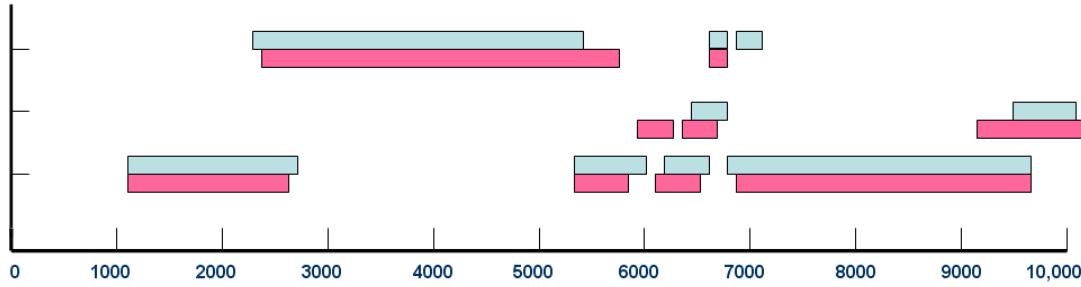


● = ATG

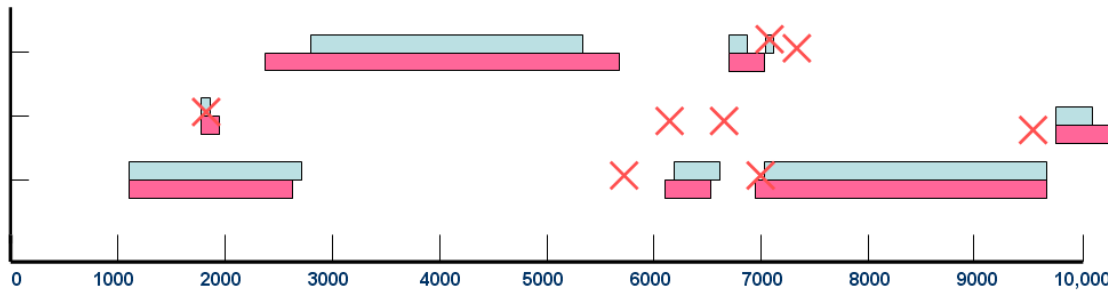
- Same evolutionary model as before, but different HMM topology
- 64 states
- 3 different types of transitions

Annotation Results: HIV1 vs. HIV2

GenBank annotation of ClustalW alignment of strains HIV1 and HIV2



Pairwise HMM annotation of ClustalW alignment of strains HIV1 and HIV2



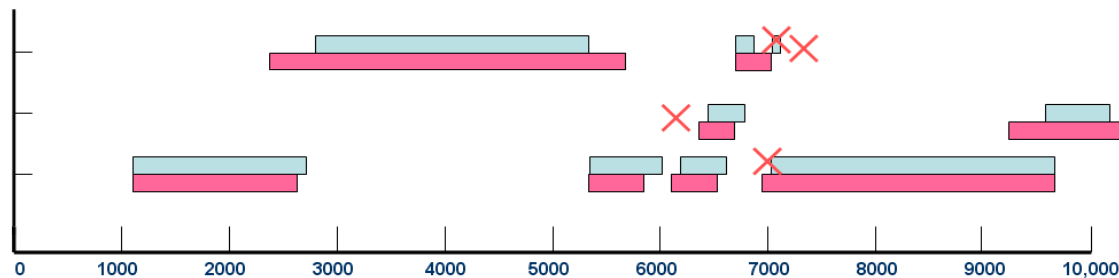
de novo annotation:

81.5% sensitivity (without non-homologous genes)

98.5% specificity

$\alpha = 0.23$ $\beta = 0.06$ $\gamma = 0.71$

Pairwise HMM annotation of ClustalW alignment of HIV2 given HIV1



Knowing HIV1 (fixing the Viterbi path for one cube):

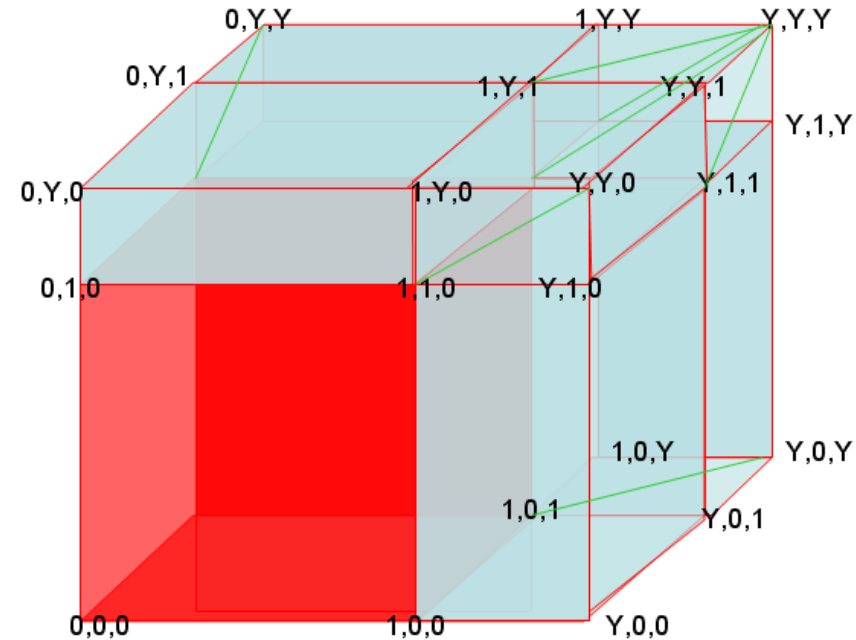
97.6% sensitivity (without non-homologous genes)

99.9% specificity

HMM Extension II: Introns

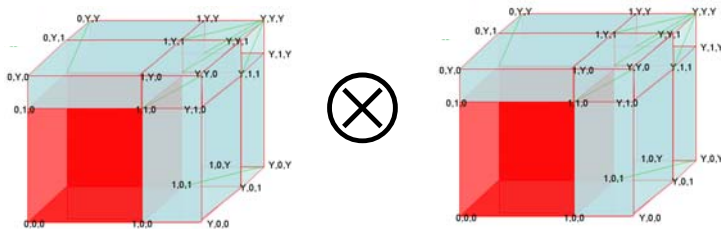
Single Sequence HMM

- Introns will almost always be 3k long
- 27 states



Pair HMM

- 729 states



Conserved RNA Structure in Protein Coding Genes

Problem: Gene Structure Known, RNA Structure Unknown.

RNA Structure:



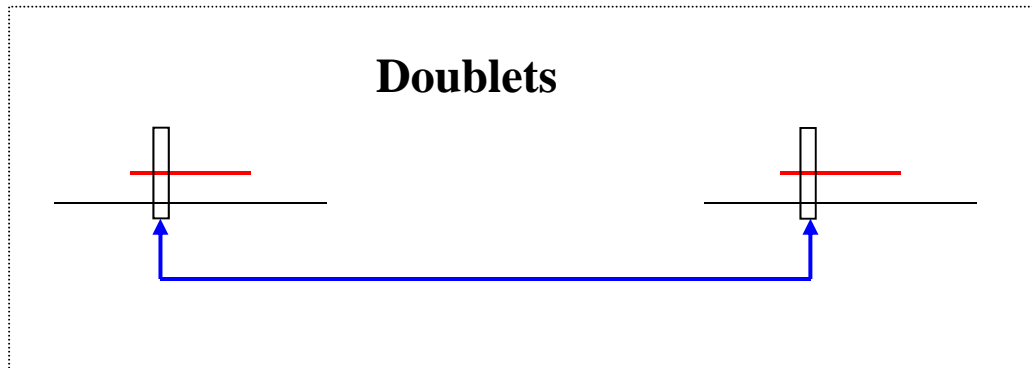
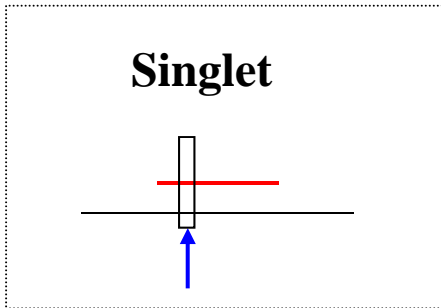
Exons:



Genome:



Protein-RNA Evolution:



Contagious Dependence

RNA + Protein Evolution

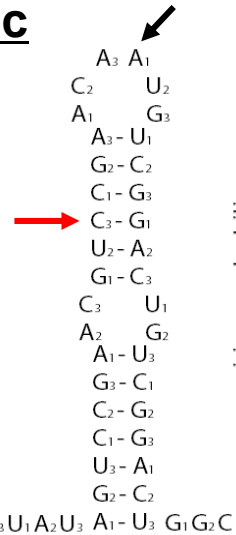
Codon Nucleotide Independence Heuristic

Singlet

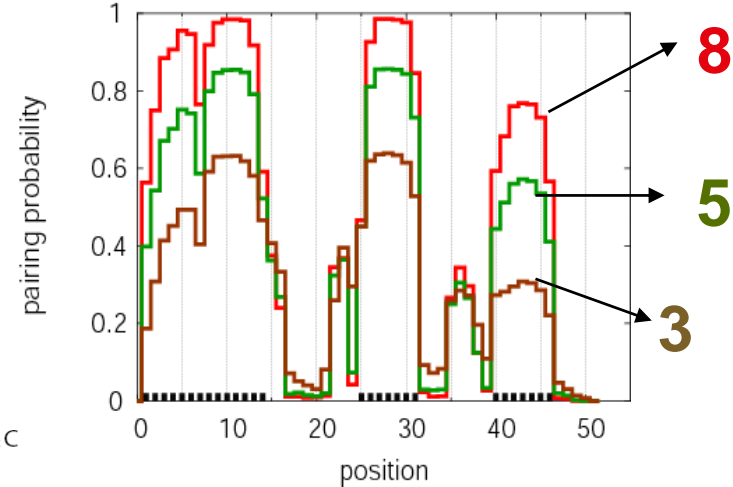
$$R_{i,j} = f^* q_{i,j}$$

Doublet

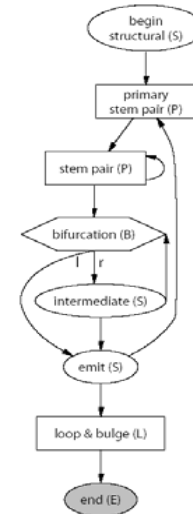
$$R_{(i1,i2),(j1,j2)} = f_1 * f_2 * q_{(i1,i2),(j1,j2)}$$



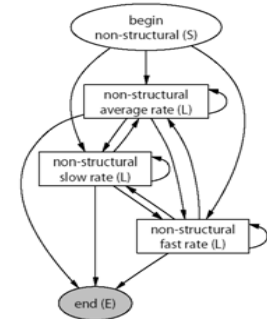
Prediction of stem-pairing regions for different number



Structure/non-Structure Grammars



Structural



Non-structural

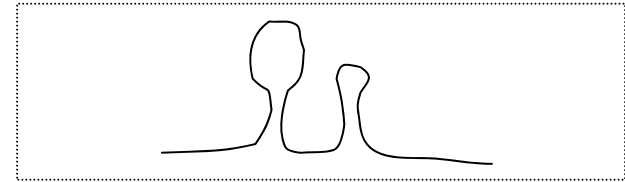
Combining Grammars: Multiple Hidden Layers

Present Approach:

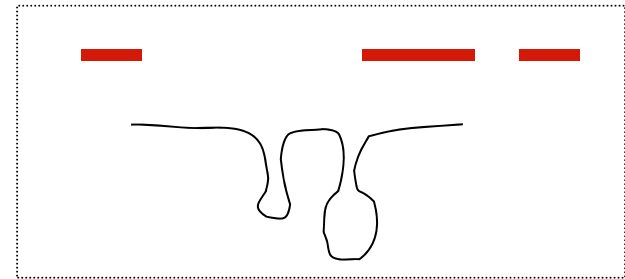
Two “independent” annotations

SCFG: RNA Structure

HMM: Protein Structure



Combine SCFG & HMM:
RNA, Gene Structure

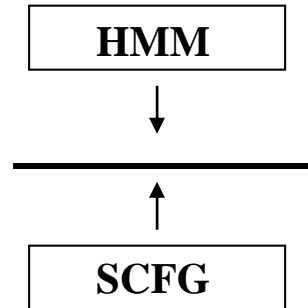


Ideal Approach:

Combined Annotation

Combining Grammars: Solution Attempts

Independence is non-trivial to define as they in principle are competing alternative models.



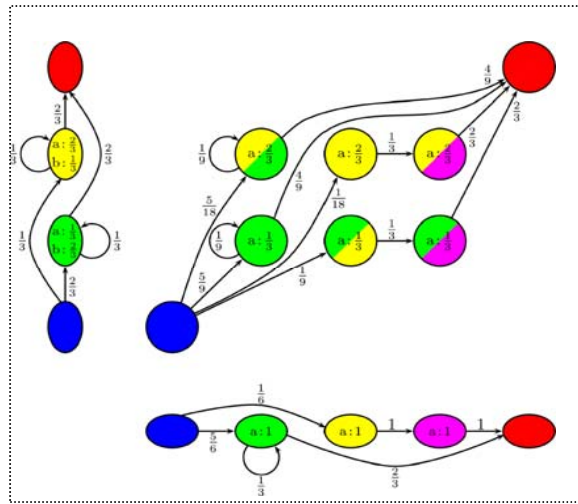
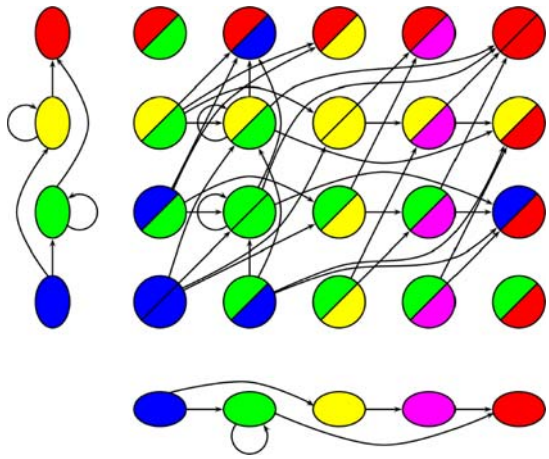
Let X be the stochastic variable giving the HMM annotation.

Let Y be the stochastic variable giving the SCFG annotation.

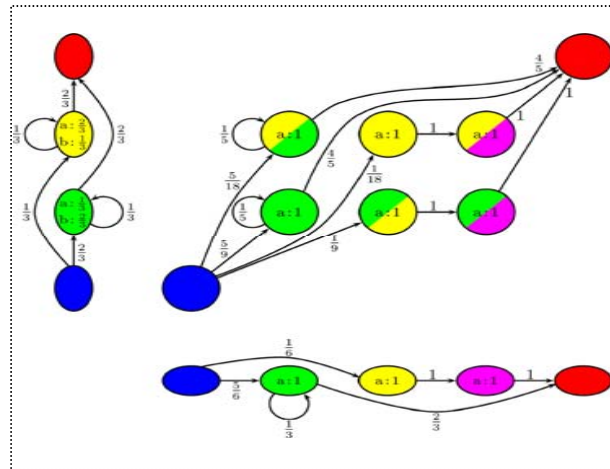
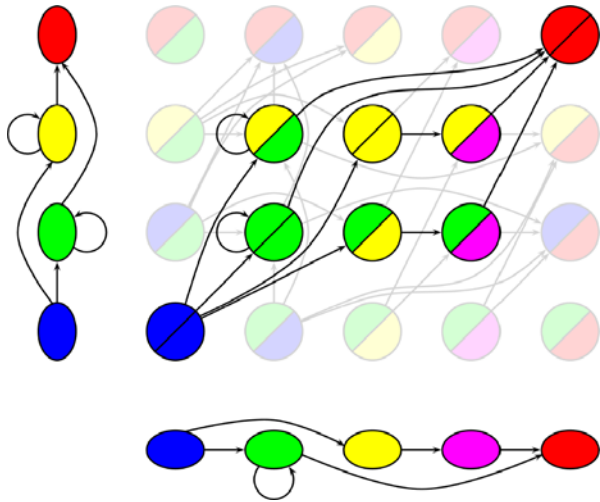
Is $P(X, Y | Data) = P(X | Data)P(Y | Data)$? **No.**

• Combined Grammars (HMM, SCGF) --> SCFG have been devised, but does not work well, have arbitrary designs and are very large.

Combinations of Viterbi and Posterior Decoding arises.



$$\begin{array}{c} \text{green green} \\ \text{yellow magenta} \\ \text{a a} \\ \hline 2 \\ 729 \end{array} < \begin{array}{c} \text{green green} \\ \text{green green} \\ \text{a a} \\ \hline 20 \\ 6561 \end{array} < \begin{array}{c} \text{yellow yellow} \\ \text{yellow magenta} \\ \text{a a} \\ \hline 4 \\ 729 \end{array} < \begin{array}{c} \text{yellow yellow} \\ \text{green green} \\ \text{a a} \\ \hline 40 \\ 6561 \end{array}$$



$$\begin{array}{c} \text{green green} \\ \text{yellow magenta} \\ \text{a a} \\ \hline 1 \\ 9 \end{array} > \begin{array}{c} \text{green green} \\ \text{green green} \\ \text{a a} \\ \hline 4 \\ 45 \end{array} > \begin{array}{c} \text{yellow yellow} \\ \text{yellow magenta} \\ \text{a a} \\ \hline 1 \\ 18 \end{array} > \begin{array}{c} \text{yellow yellow} \\ \text{green green} \\ \text{a a} \\ \hline 2 \\ 45 \end{array}$$