Applications of HMMs in Computational Biology

BMI/CS 576

www.biostat.wisc.edu/bmi576.html

Colin Dewey

cdewey@biostat.wisc.edu

Fall 2008

The Gene Finding Task

Given: an uncharacterized DNA sequence

Do: locate the genes in the sequence, including the coordinates of individual *exons* and *introns*

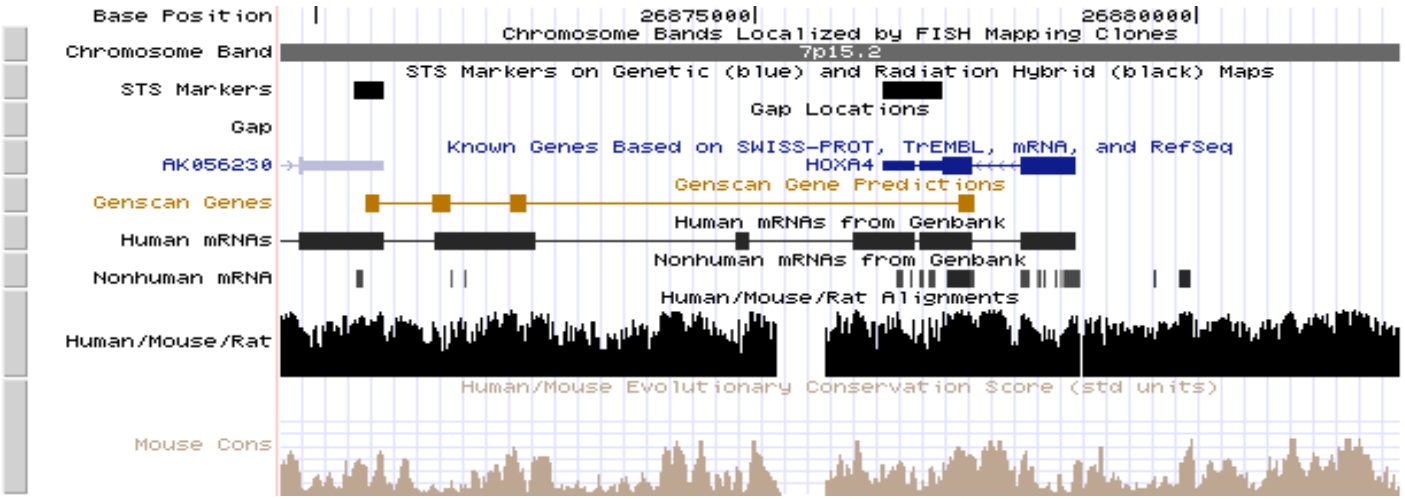
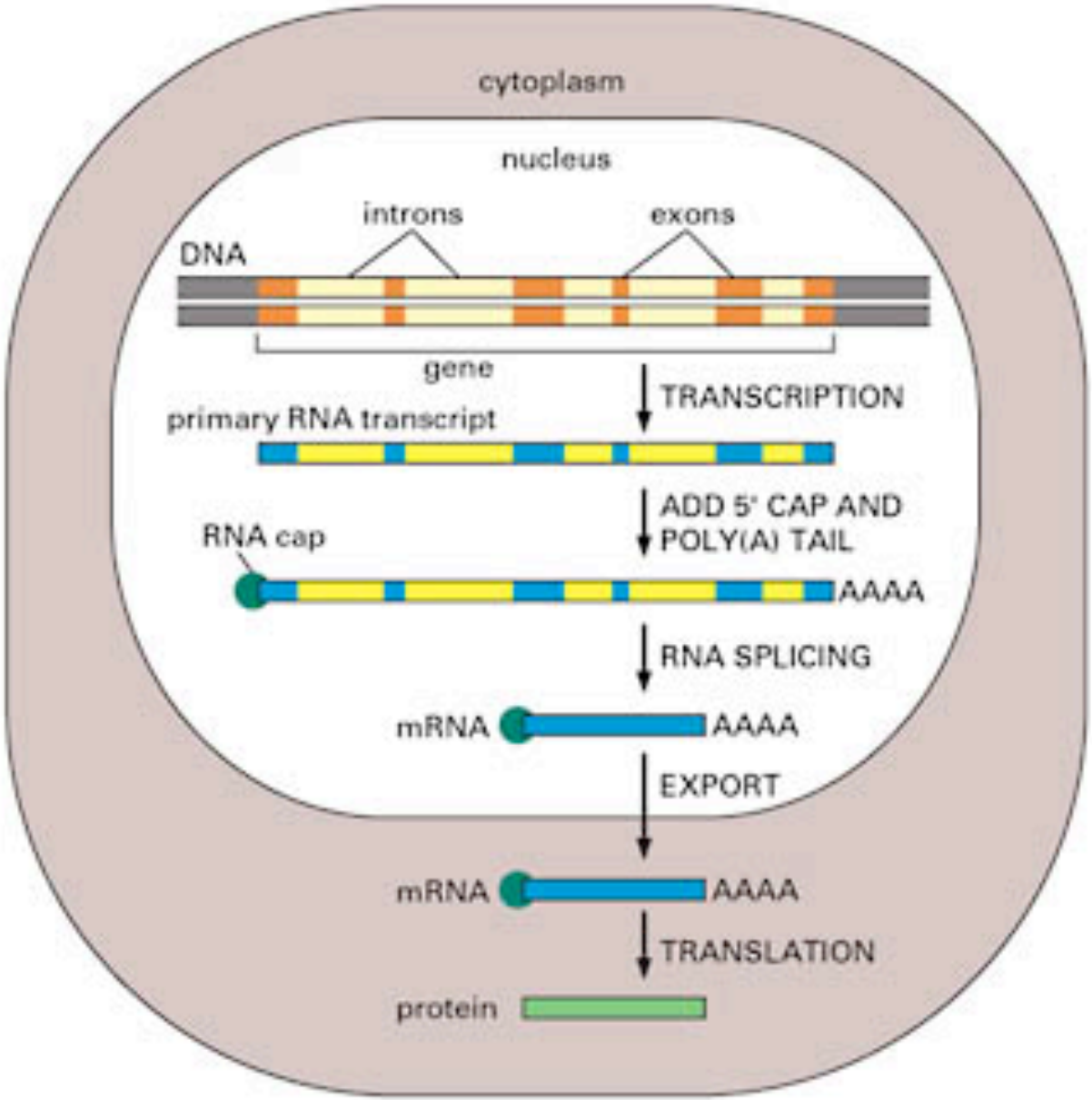


image from the UCSC Genome Browser
http://genome.ucsc.edu/

Eukaryotic Gene Structure

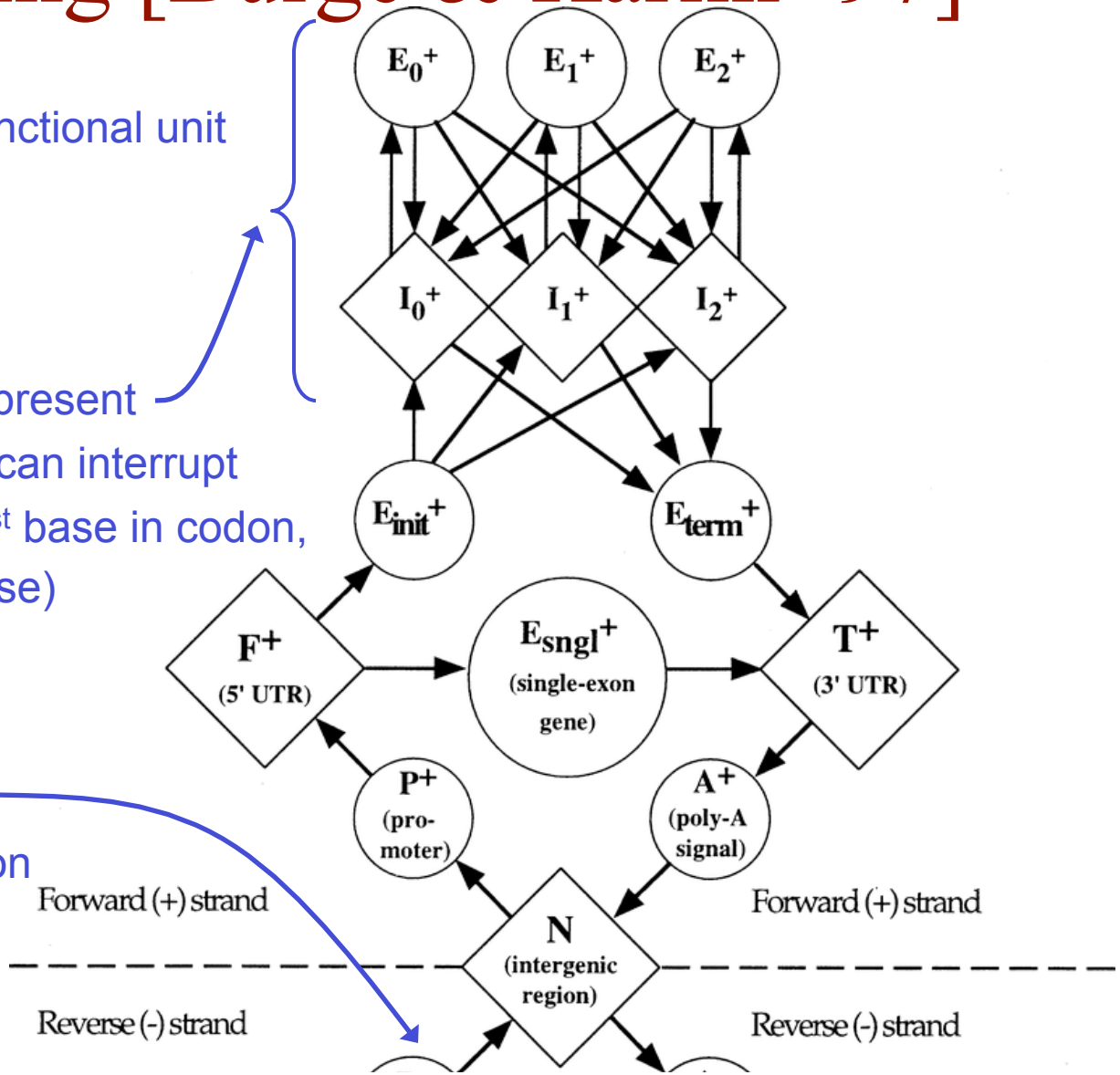


The GENSCAN HMM for Eukaryotic Gene Finding [Burge & Karlin '97]

Each shape represents a functional unit of a gene or genomic region

Pairs of intron/exon units represent the different ways an intron can interrupt a coding sequence (after 1st base in codon, after 2nd base or after 3rd base)

Complementary submodel (not shown) detects genes on opposite DNA strand



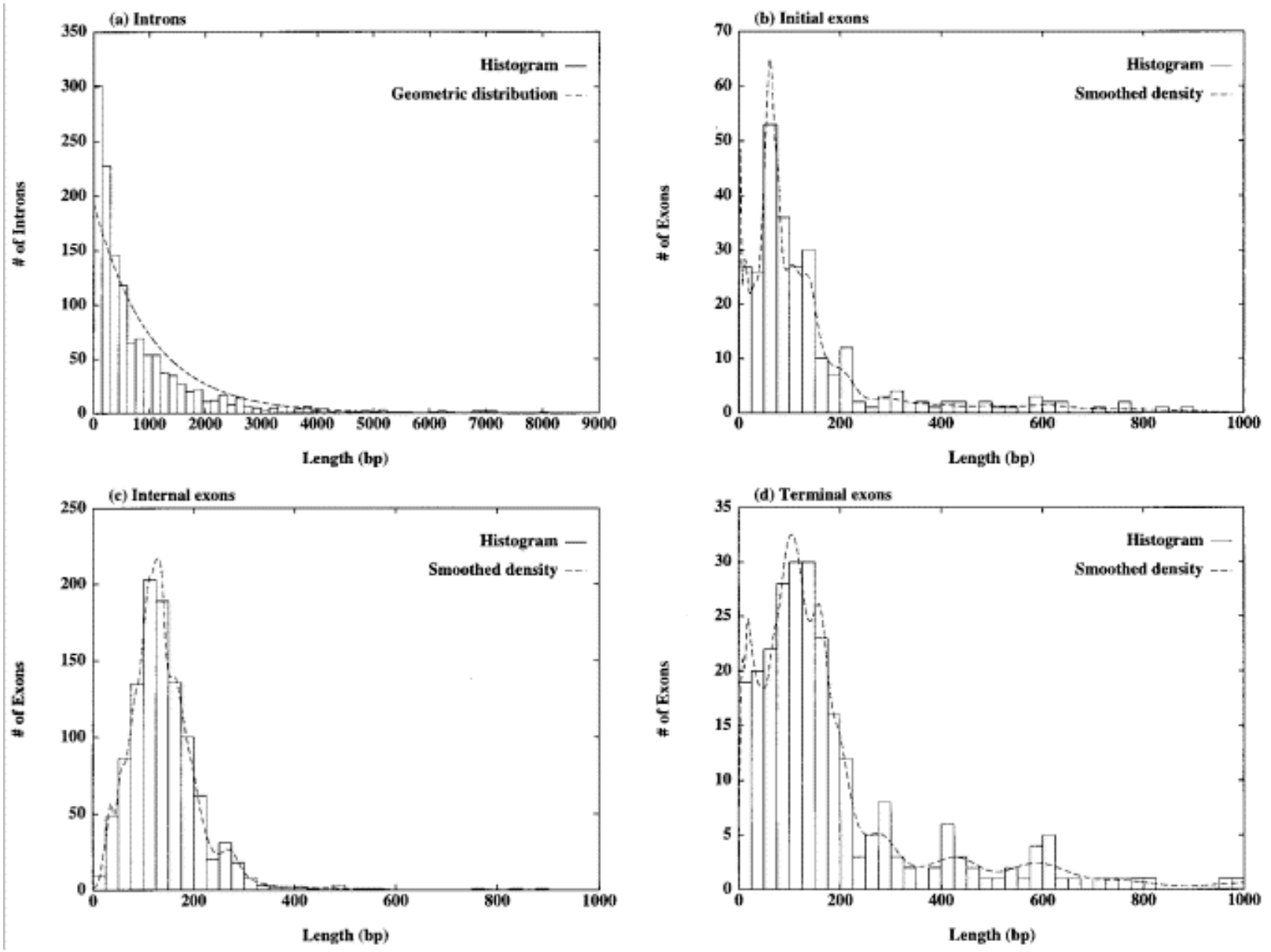
The GENSCAN HMM

- for each sequence type, GENSCAN models
  - the length distribution
  - the sequence composition

- length distribution models vary depending on sequence type
  - nonparametric (using histograms)
  - parametric (using geometric distributions)
  - fixed-length

- sequence composition models vary depending on type
  - 5^{th}-order, inhomogeneous
  - 5^{th} -order homogenous
  - 0^{th} and 1^{st}-order inhomogeneous
  - tree-structured variable memory

Human Intron & Exon Lengths



Representing Exons in GENSCAN

- for exons, GENSCAN uses
  - Histograms to represent exon lengths
  - 5^{th}-order, inhomogeneous Markov models to represent exon sequences

- 5^{th}-order, inhomogeneous models can represent statistics about pairs of neighboring codons

Inference with the Gene-Finding HMM

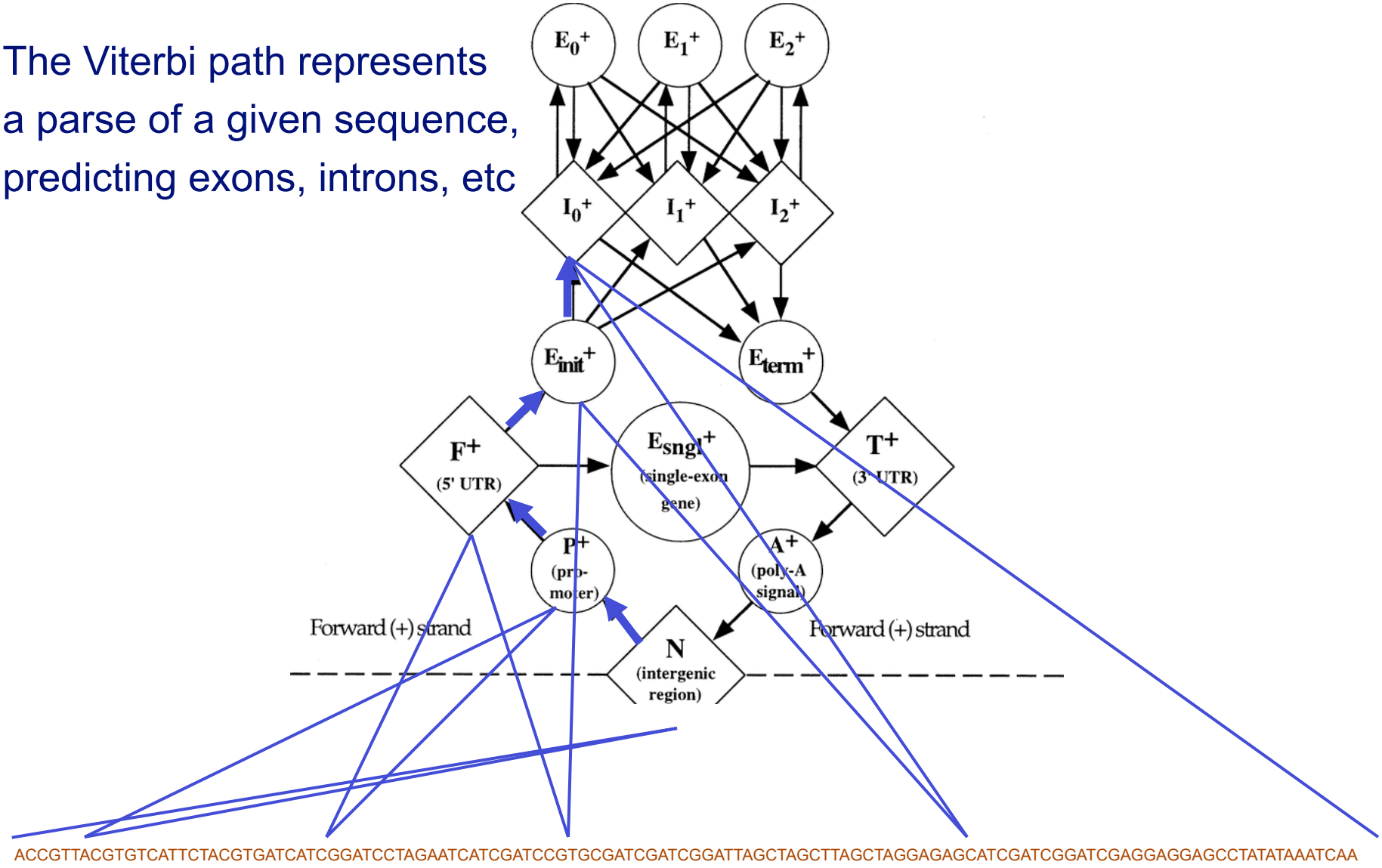
given: an uncharacterized DNA sequence

do: find the most probable path through the model for the sequence

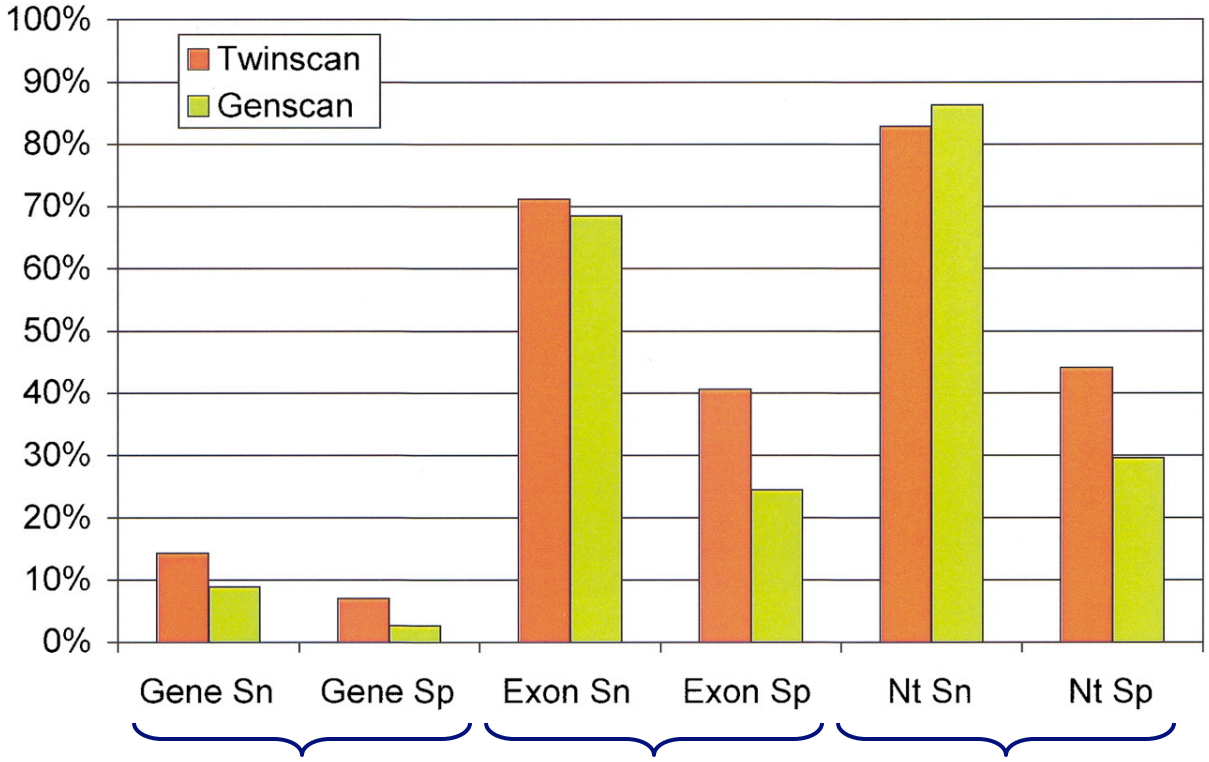
- This path will specify the coordinates of the predicted genes (including intron and exon boundaries)
- The Viterbi algorithm is used to compute this path

Parsing a DNA Sequence

The Viterbi path represents a parse of a given sequence, predicting exons, introns, etc



Accuracy of GENSCAN (and TWINSCAN)



$$\text{sensitivity (Sn)} = \frac{TP}{TP + FN}$$

$$\text{specificity (Sp)} = \frac{TP}{TP + FP}$$

genes exactly correct?

exons exactly correct?

nucleotides correct?

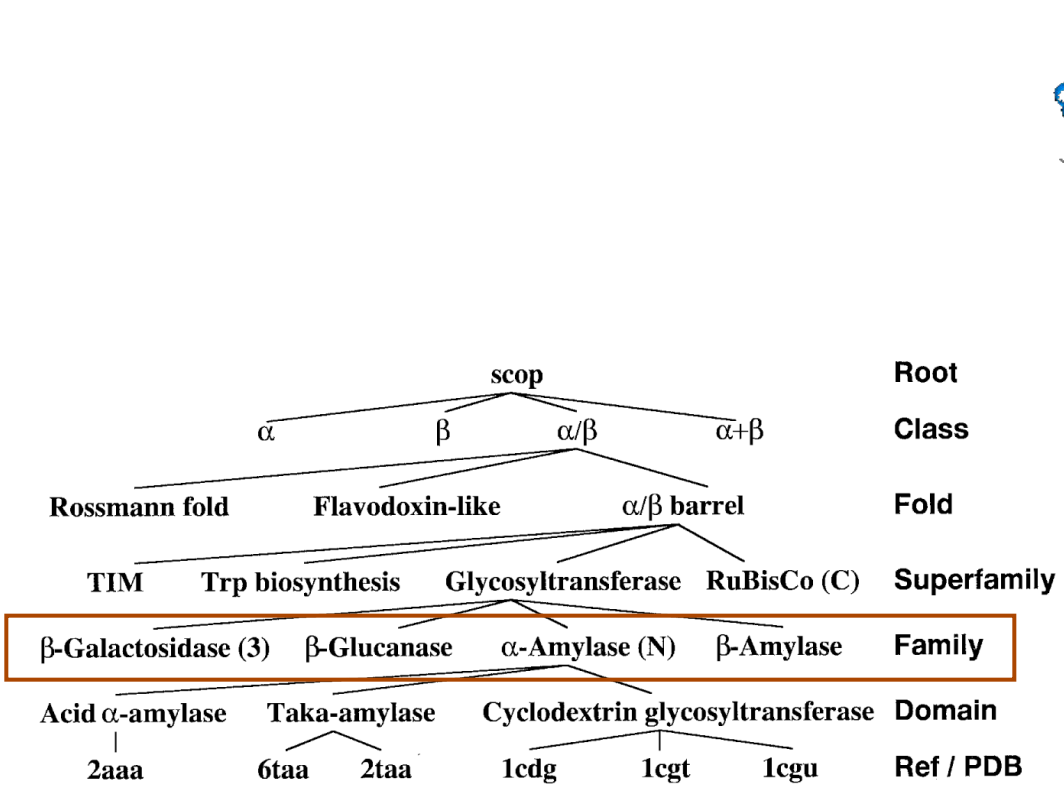
Figure from Flicek et al., *Genome Research*, 2003

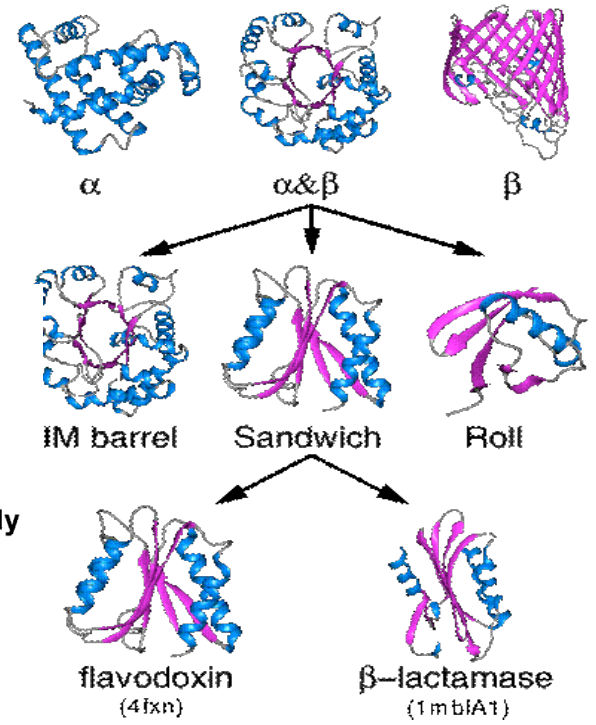
The Protein Classification Task

Given: amino-acid sequence of a protein

Do: predict the *family* to which it belongs

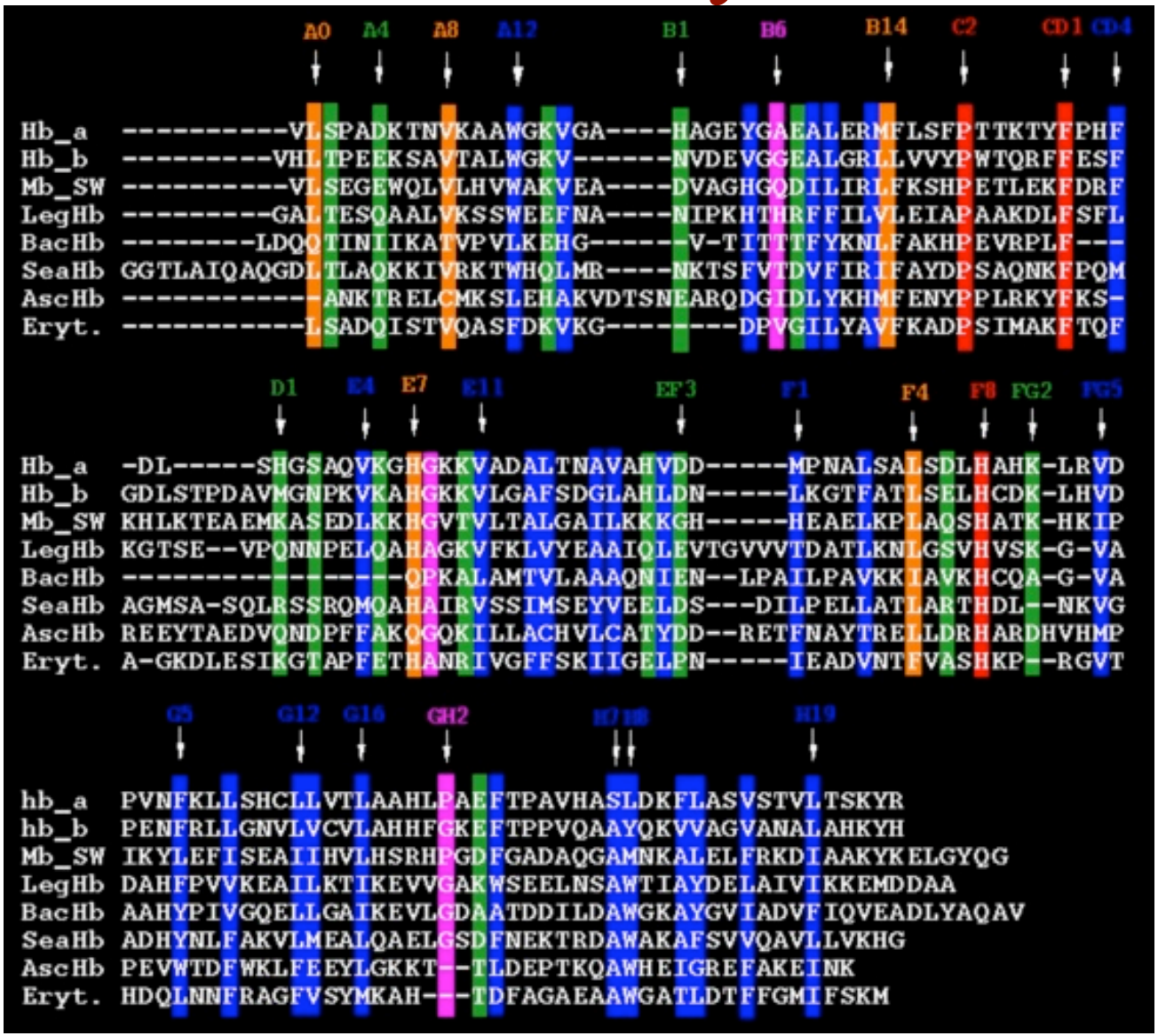
GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVCVLAHHFGKEFTPPVQAAYAKVVAGVANALAHKYH





Alignment of Globin Family Proteins

- The sequences in a family may vary in length
- Some positions are more conserved than others



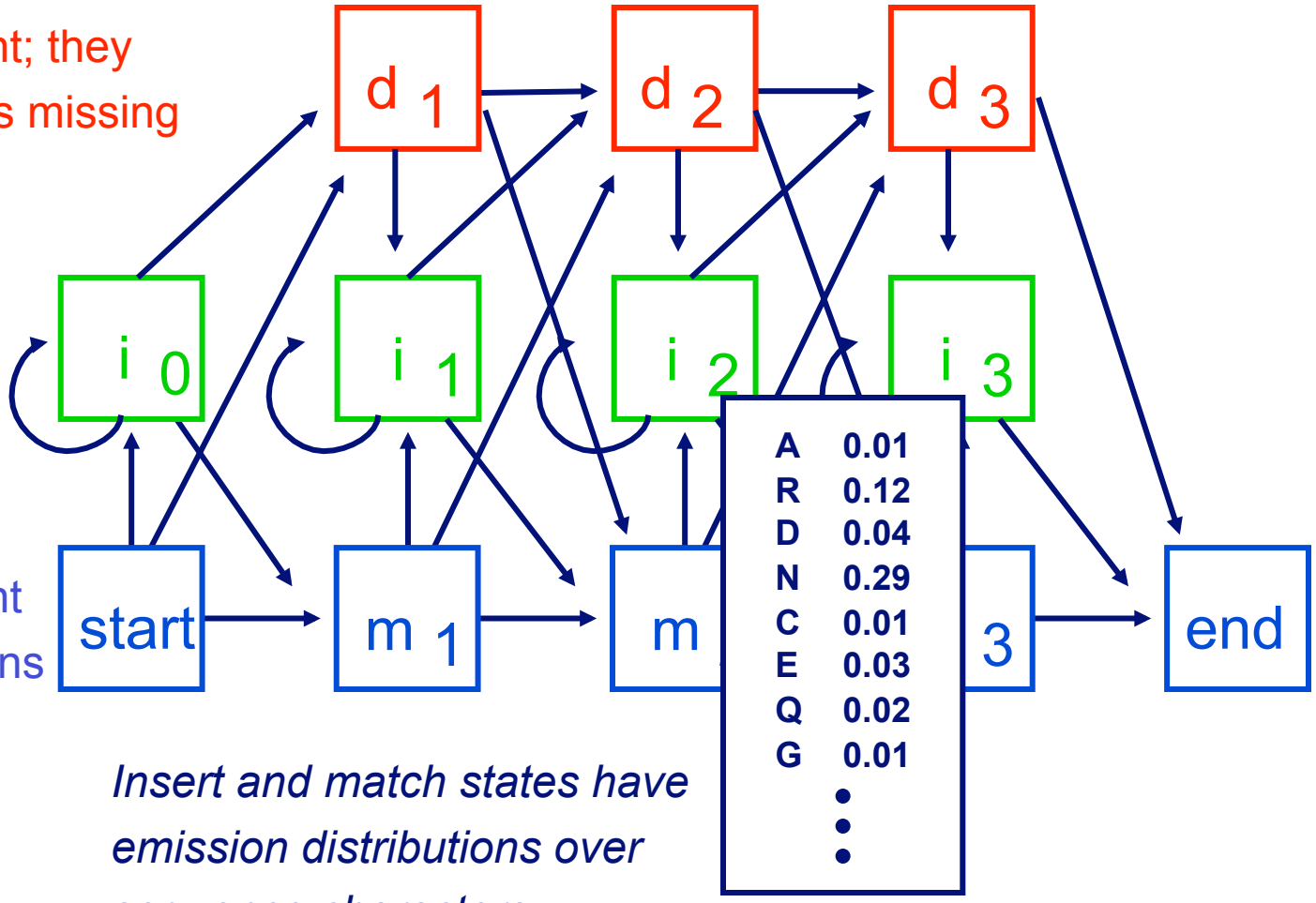
Profile HMMs

- profile HMMs are commonly used to model families of sequences

Delete states are silent; they Account for characters missing in some sequences

Insert states account for extra characters in some sequences

Match states represent key conserved positions



Insert and match states have emission distributions over sequence characters

Multiple Alignment of SH3 Domain

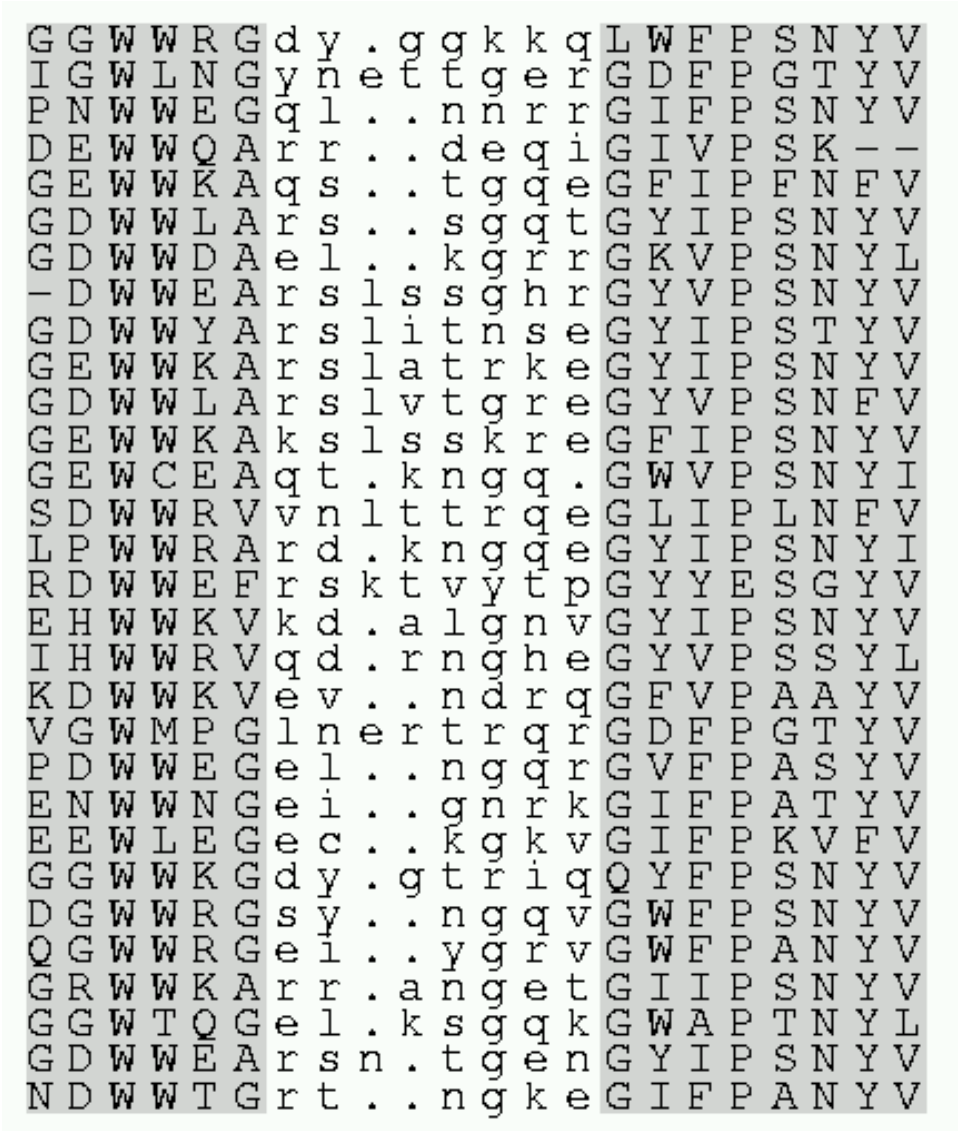


Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

A Profile HMM Trained for the SH3 Domain

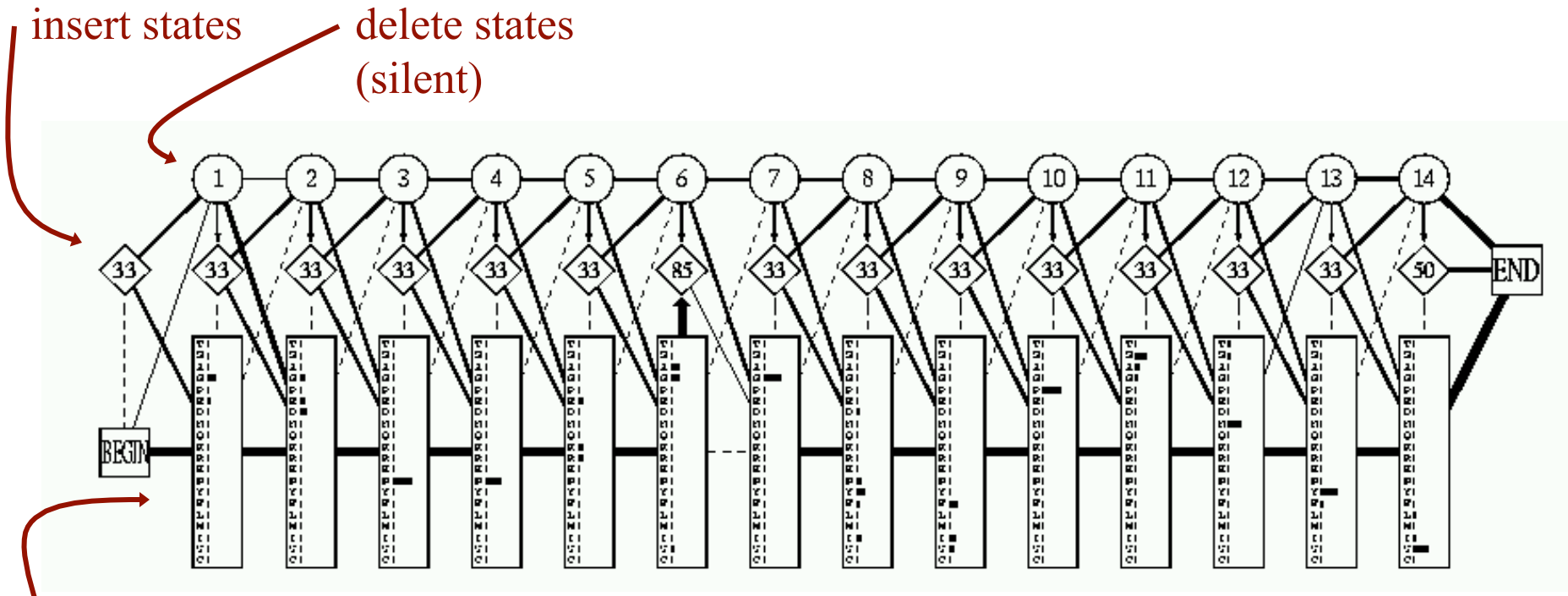
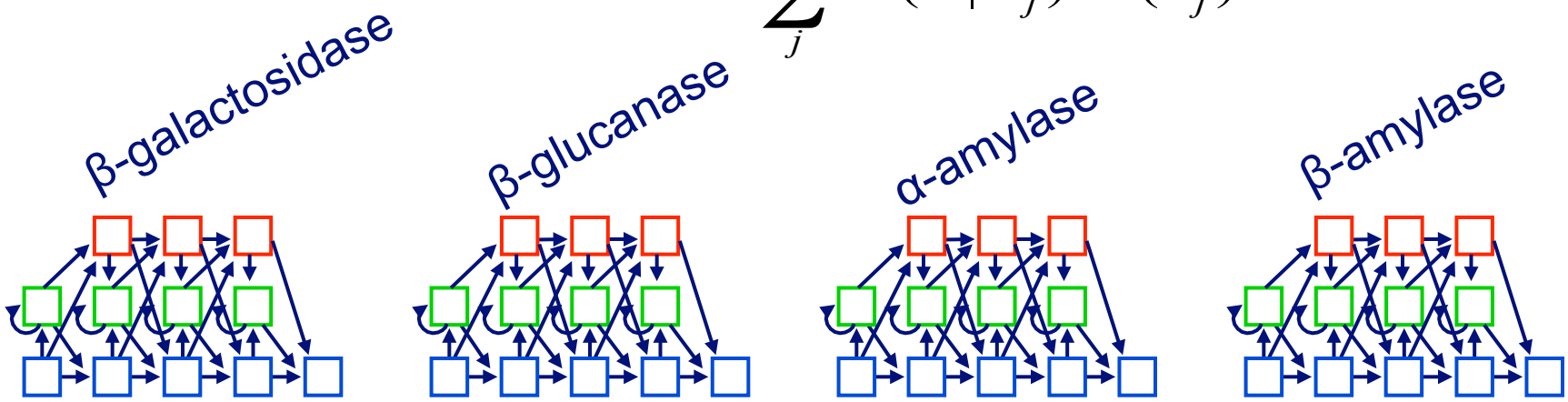


Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

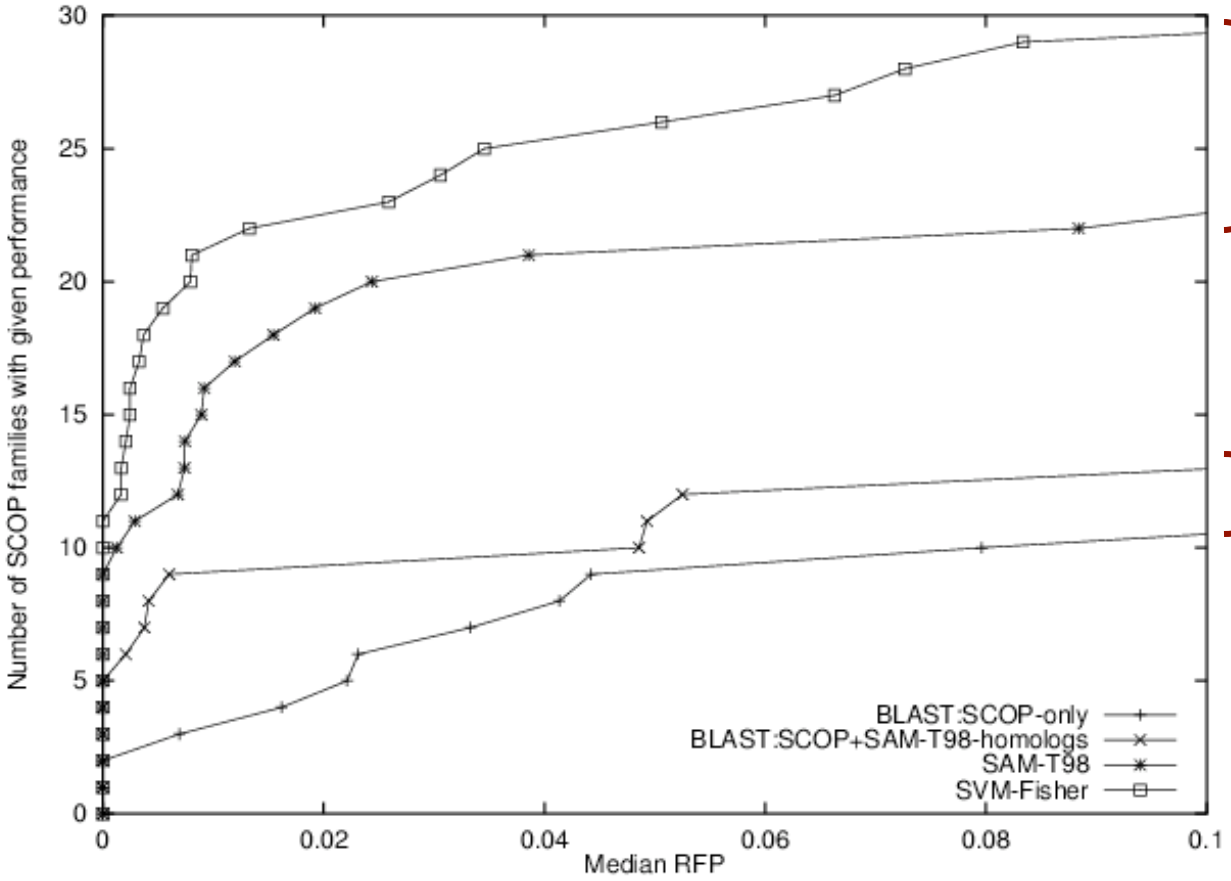
Profile HMMs

- To classify sequences according to family, we can train a profile HMM to model the proteins of each family of interest
- Given a sequence *x,* use Bayes' rule to make classification

$$\Pr(c_i \mid x) = \frac{\Pr(x \mid c_i)\Pr(c_i)}{\sum_j \Pr(x \mid c_j)\Pr(c_j)}$$



Profile HMM Accuracy



profile HMM-based methods

BLAST-based methods

Figure from Jaakola et al., ISMB 1999

- classifying 2447proteins into 33 families
- *x*-axis represents the median # of negative sequences that score as high as a positive sequence for a given family's model

Other Issues in Markov Models

- there are many interesting variants and extensions of the models/algorithms we considered here (some of these are covered in BMI/CS 776)
  - separating length/composition distributions with *semi-Markov models*
  - modeling multiple sequences with *pair HMMs*
  - learning the *structure* of HMMs
  - going up the Chomsky hierarchy: *stochastic context free grammars*
  - discriminative learning algorithms (e.g. as in *conditional random fields*)
  - etc.