Applications of HMMs in Computational Biology

BMI/CS 576

www.biostat.wisc.edu/bmi576.html

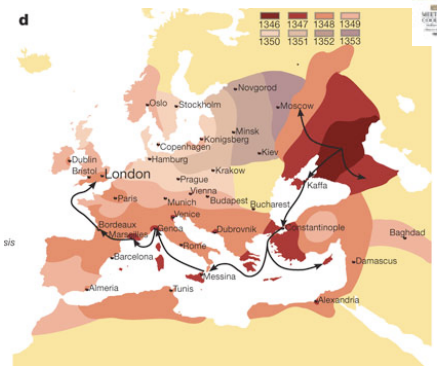
Mark Craven

craven@biostat.wisc.edu

Fall 2011

Sequencing news this month

Yersinia pestis

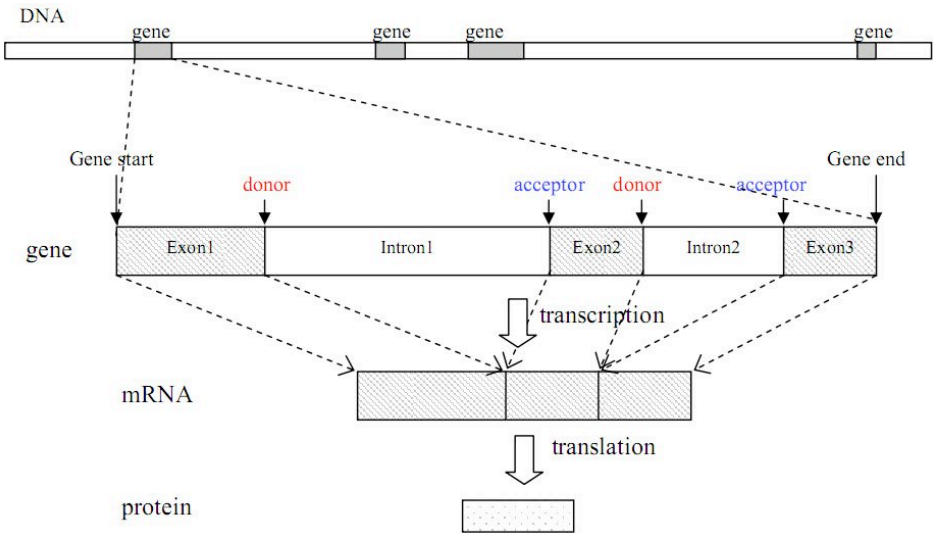


Cannabis sativa

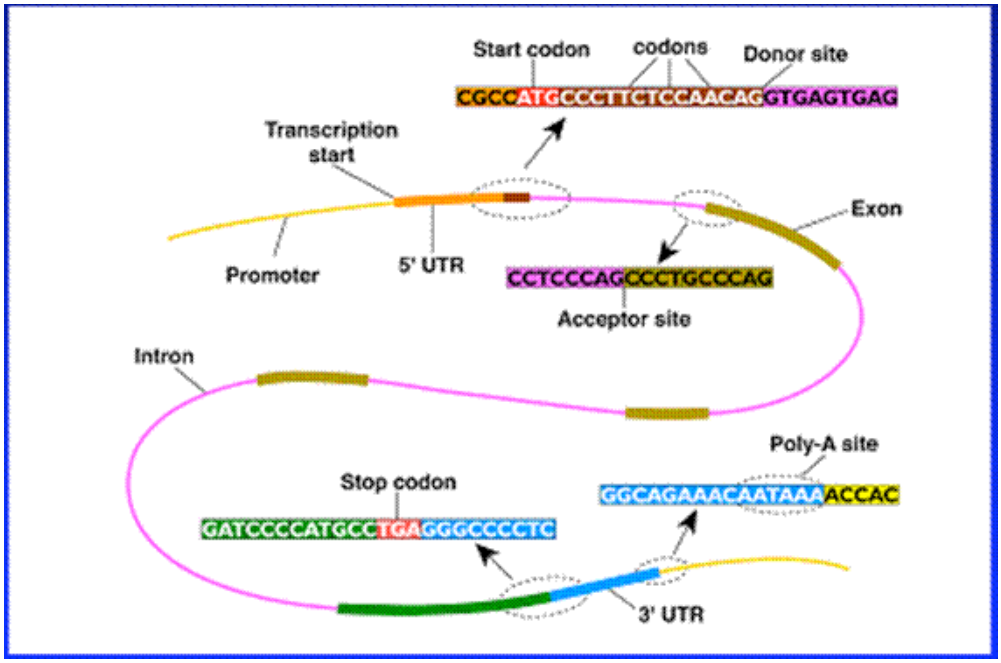
The gene finding task

Given: an uncharacterized DNA sequence

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



Eukaryotic gene structure



Sources of evidence for gene finding

- **signals**: the sequence *signals* (e.g. splice junctions) involved in gene expression
- **content**: statistical properties that distinguish protein-coding DNA from non-coding DNA
- **conservation**: signal and content properties that are conserved across related sequences (e.g. syntenic regions of the mouse and human genome)

Gene finding: search by content

- encoding a protein affects the statistical properties of a DNA sequence

```
UUU F 0.46   UCU S 0.19   UAU Y 0.44   UGU C 0.46
UUC F 0.54   UCC S 0.22   UAC Y 0.56   UGC C 0.54
UUA L 0.08   UCA S 0.15   UAA * 0.30   UGA * 0.47
UUG L 0.13   UCG S 0.05   UAG * 0.24   UGG W 1.00

CUU L 0.13   CCU P 0.29   CAU H 0.42   CGU R 0.08
CUC L 0.20   CCC P 0.32   CAC H 0.58   CGC R 0.18
CUA L 0.07   CCA P 0.28   CAA Q 0.27   CGA R 0.11
CUG L 0.40   CCG P 0.11   CAG Q 0.73   CGG R 0.20

AUU I 0.36   ACU T 0.25   AAU N 0.47   AGU S 0.15
AUC I 0.47   ACC T 0.36   AAC N 0.53   AGC S 0.24
AUA I 0.17   ACA T 0.28   AAA K 0.43   AGA R 0.21
AUG M 1.00   ACG T 0.11   AAG K 0.57   AGG R 0.21

GUU V 0.18   GCU A 0.27   GAU D 0.46   GGU G 0.16
GUC V 0.24   GCC A 0.40   GAC D 0.54   GGC G 0.34
GUA V 0.12   GCA A 0.23   GAA E 0.42   GGA G 0.25
GUG V 0.46   GCG A 0.11   GAG E 0.58   GGG G 0.25

[Codon/a.a./fraction per codon per a.a.]
Homo sapiens data from the Codon Usage Database
```

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

Each shape denotes a functional unit of a gene or genomic region and is represented by a submodel in the HMM

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

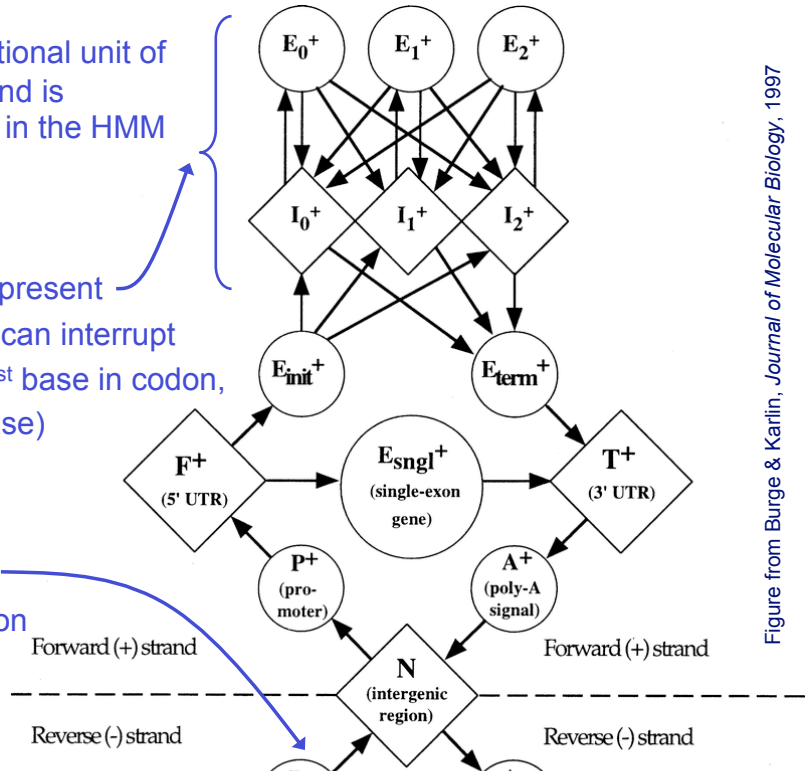


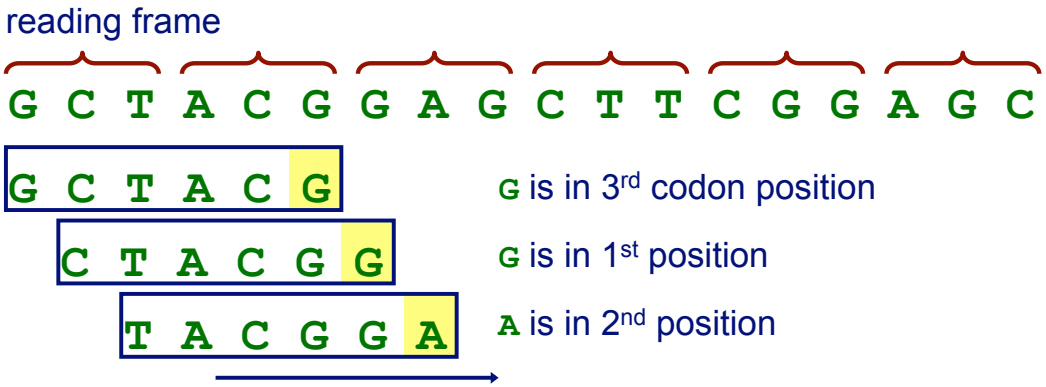
Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

GENSCAN uses a variety of submodel types

sequence feature	model
exons	5th order inhomogenous
introns, intergenic regions	5th order homogenous
poly-A, translation initiation, promoter	0th order, fixed-length
splice junctions	tree-structured variable memory

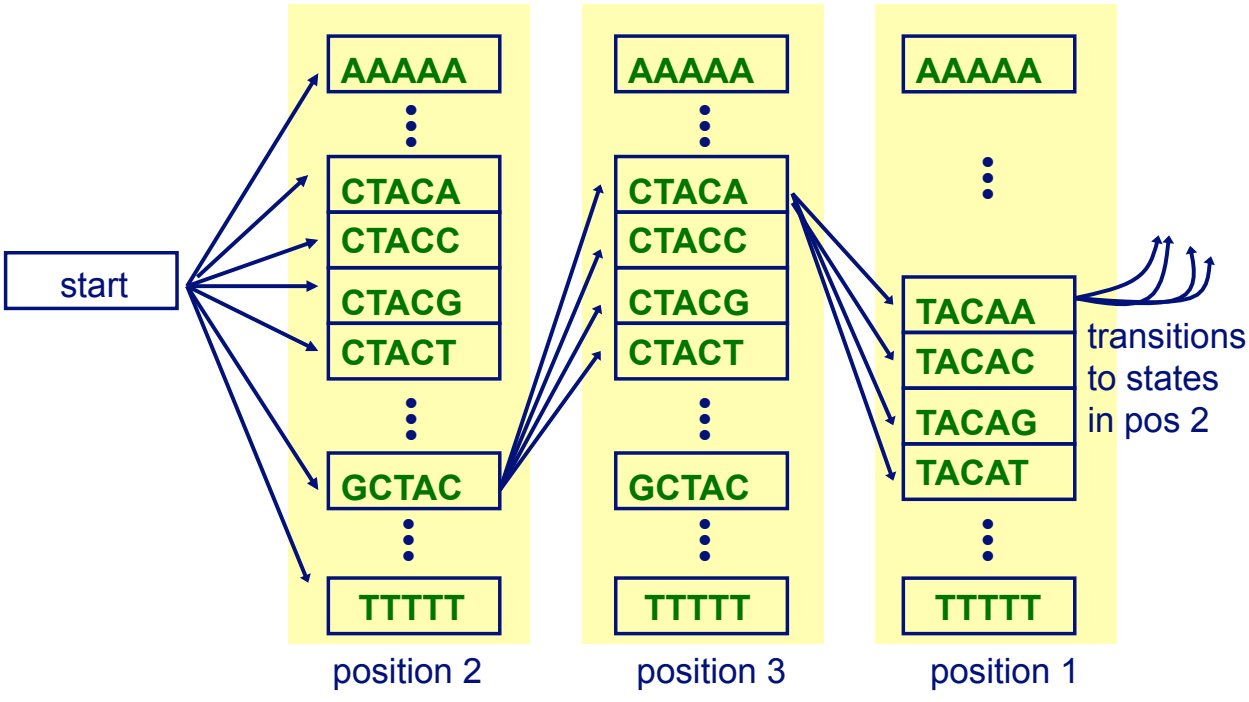
Markov models & exons

- consider modeling a given coding sequence
- for each "word" we evaluate, we'll want to consider its position with respect to the reading frame we're assuming



- can do this using an inhomogeneous model

A fifth-order inhomogenous Markov chain



Inference with the gene-finding HMM

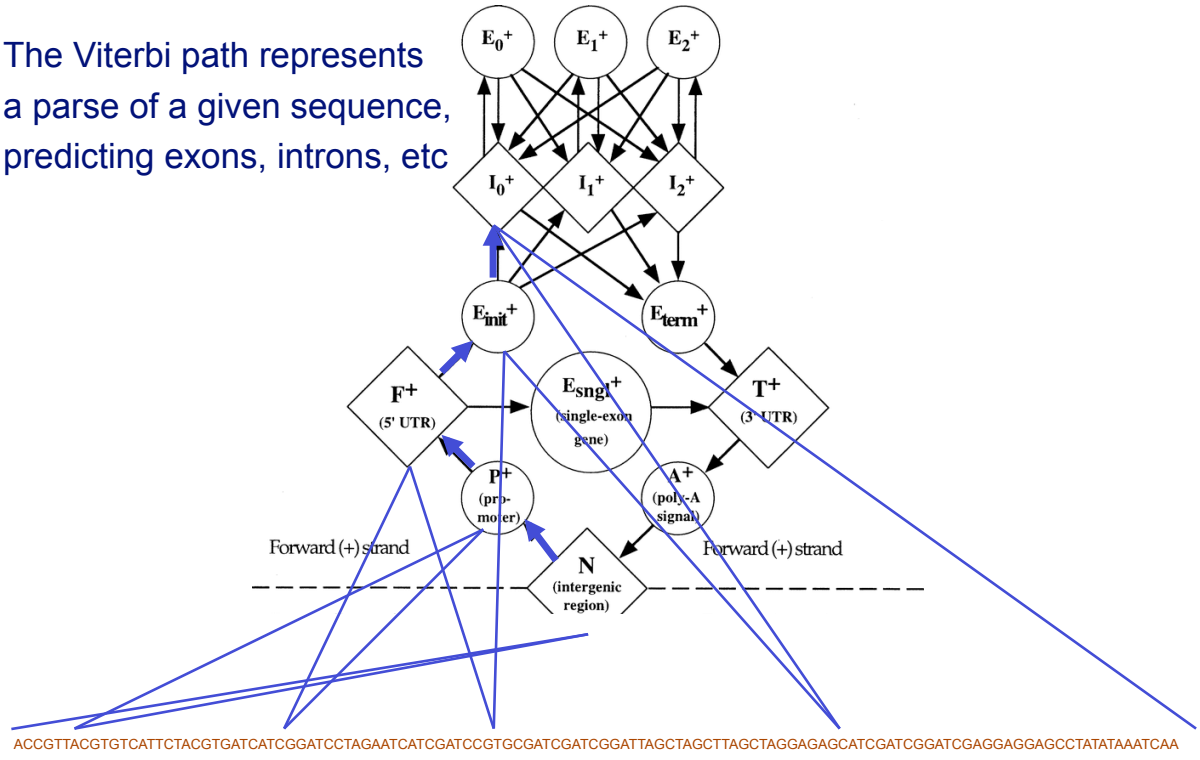
given: an uncharacterized DNA sequence

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



Assessing the accuracy of a trained model

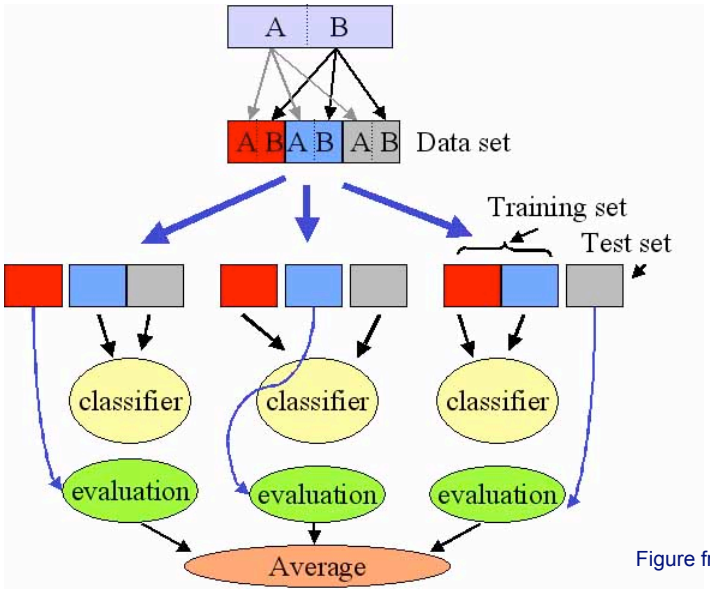
- two issues
  - What data should we use?
  - Which metrics should we use?

- Can we measure accuracy on the data set that was used to train the model?

  NO! This will result in accuracy estimates that are biased (too high).

Assessing the accuracy of a trained model

- need to have a *test set* that is disjoint from the training set
- more generally, can use *cross validation*



3-fold CV illustrated

Figure from http://gepas.bioinfo.cipf.es/

Accuracy (for 2-class problems)

predicted class \ actual class	positive	negative
positive	true positives (TP)	false positives (FP)
negative	false negatives (FN)	true negatives (TN)

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Accuracy metrics

predicted class \ actual class	positive	negative
positive	true positives (TP)	false positives (FP)
negative	false negatives (FN)	true negatives (TN)

$$\text{sensitivity (recall)} = \frac{TP}{\text{actual pos}} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{\text{actual neg}} = \frac{TN}{TN + FP}$$

$$\text{precision} = \frac{TP}{\text{predicted pos}} = \frac{TP}{TP + FP}$$

sometimes specificity is defined this way

Accuracy of GENSCAN

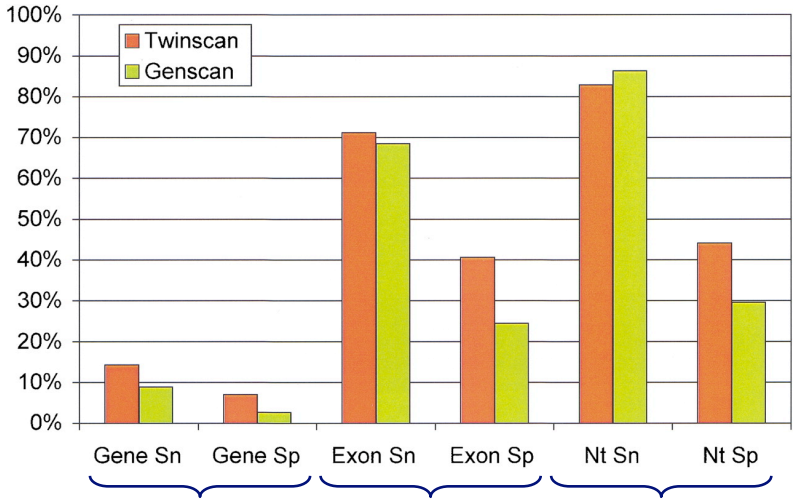
Table 1. Performance comparison for Burset/Guigó set of 570 vertebrate genes
A *Comparison of GENSCAN with other gene prediction programs*

		Accuracy per nucleotide				Accuracy per exon				
Program	Sequences	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
GENSCAN	570 (8)	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
FGENEH	569 (22)	0.77	0.88	0.78	0.80	0.61	0.64	0.64	0.15	0.12
GeneID	570 (2)	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
Genie	570 (0)	0.76	0.77	0.72	n/a	0.55	0.48	0.51	0.17	0.33
GenLang	570 (30)	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
GeneParser2	562 (0)	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
GRAIL2	570 (23)	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
SORFIND	561 (0)	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	570 (28)	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13
GeneID+	478 (1)	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	478 (1)	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09

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

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

Accuracy of GENSCAN on a different test set



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

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

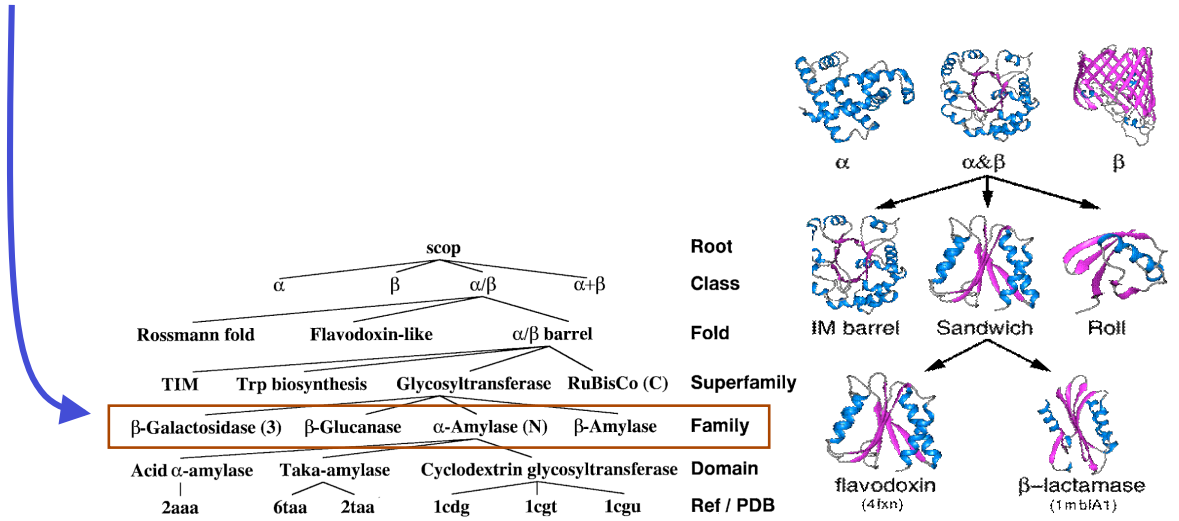
Figure from Flicek et al., *Genome Research*, 2003

The protein classification task

Given: amino-acid sequence of a protein

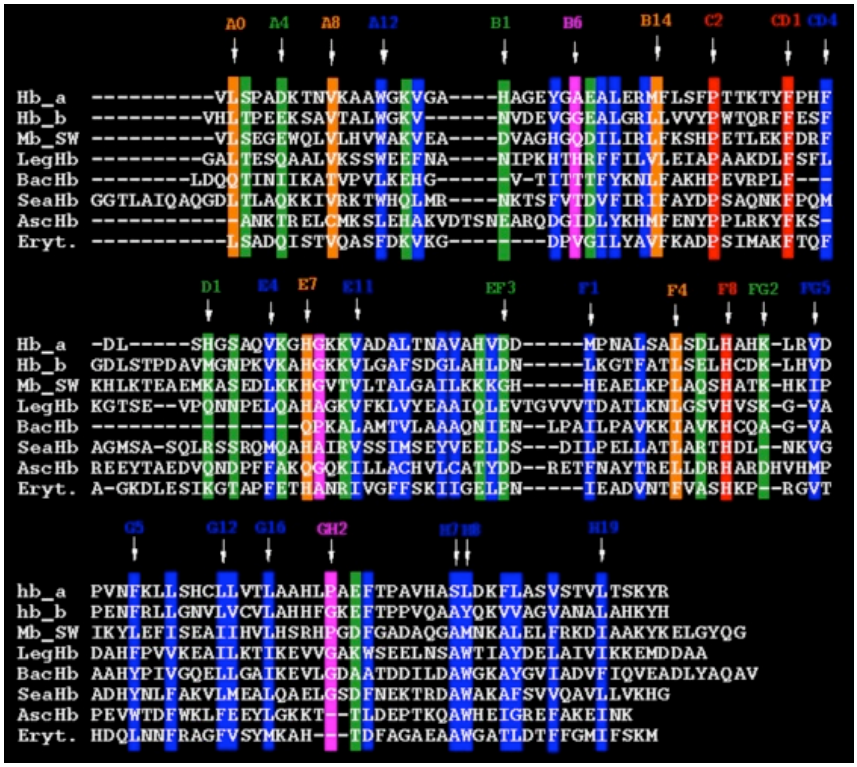
Do: predict the *family* to which it belongs

GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVCVLAHHFGKEFTPPVQAAYAKVVAGVANALAHKYH



Alignment of globin family proteins

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



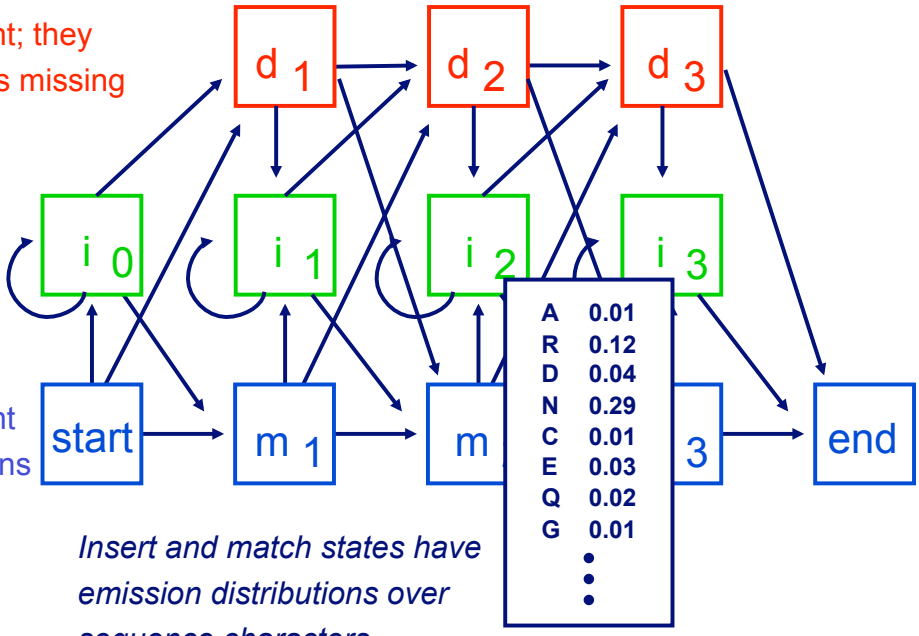
Profile HMMs

- profile HMMs are 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

```
GGWWRGdy.ggkkqLWFPSNYV
IGWLNGynetfgerGDFPGTYV
PNWWEGql..nnrrGIFPSNYV
DEWWQArr..deqiGIVPSK--
GEWWKAqs..tgqeGFIPFNFV
GDWWLArs..sgqtGYIPSNYV
GDWWDAel..kgrrGKVPSNYL
-DWWEArslssghrGYVPSNYV
GDWWYArslitnseGYIPSTYV
GEWWKArslatrkeGYIPSNYV
GDWWLArslvtgreGYVPSNFV
GEWWKAkslsskreGFIPSNYV
GEWCEAqt.kngq.GWVPSNYI
SDWWRVvnlttrqeGLIPLNFV
LPWWRArd.kngqeGYIPSNYI
RDWWEFrsktvytpGYYESGYV
EHWWKVkd.algnvGYIPSNYV
IHWWRVqd.rnghePGYVPSSYL
KDWWKVev..ndrqGFVPAAYV
VGWMPGlnertrqrGDFPGTYV
PDWWEGel..ngqrGVFPASYV
ENWWNGei..gnrkGIFPATYV
EEWLEGec..kgkvGIFPKVFV
GGWWKGdy.gtriqQYFPSNYV
DGWWRGsy..ngqvGWFPSNYV
QGWWRGei..ygrvGWFPANYV
GRWWKArr.angetGIIPSNYV
GGWTQGel.ksgqkGWAPTNYL
GDWWEArsn.tgenGYIPSNYV
NDWWTGrt..ngkeGIFPANYV
```

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

A profile HMM trained for the SH3 domain

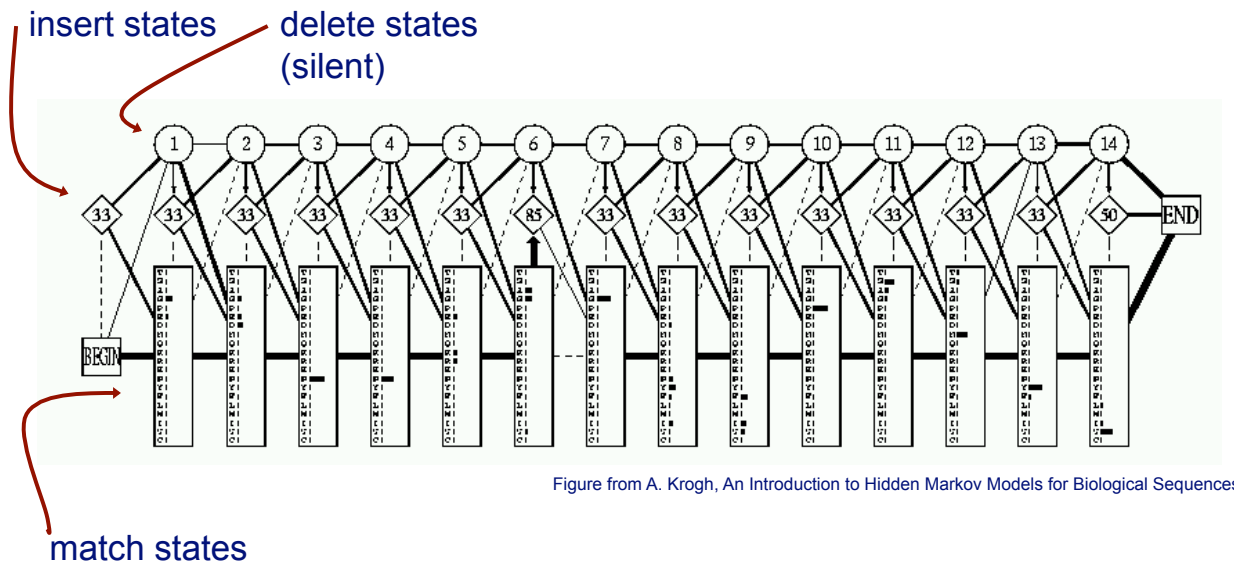


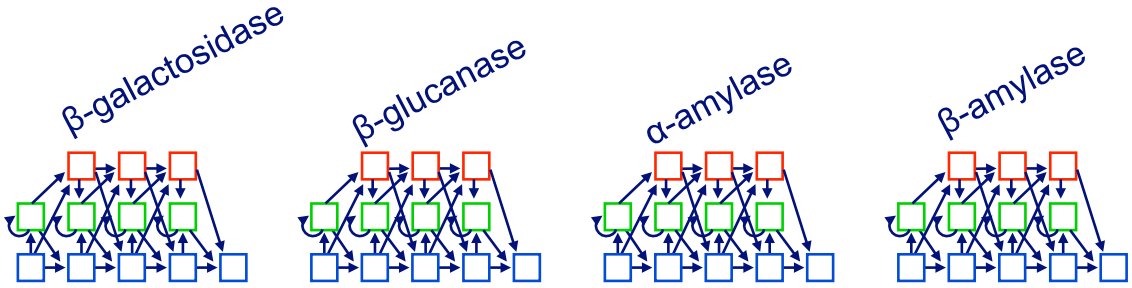
Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

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

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

- use Forward algorithm to compute $P(x \mid c_i)$ for each family c_i



Profile HMM accuracy

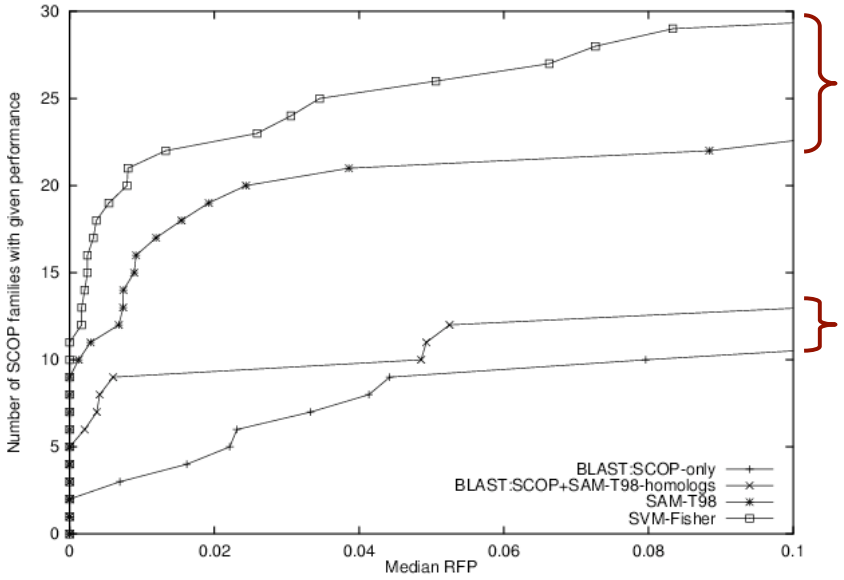
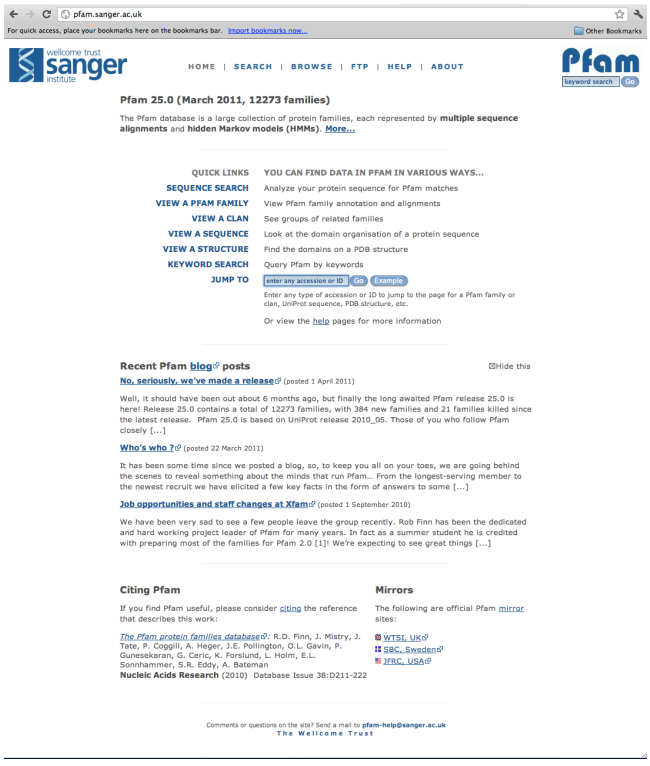


Figure from Jaakola et al., ISMB 1999

- 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

See Pfam database for a large collection profile HMMs



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.