Applications of HMMs in Computational Biology

BMI/CS 576

www.biostat.wisc.edu/bmi576.html

Mark Craven

craven@biostat.wisc.edu

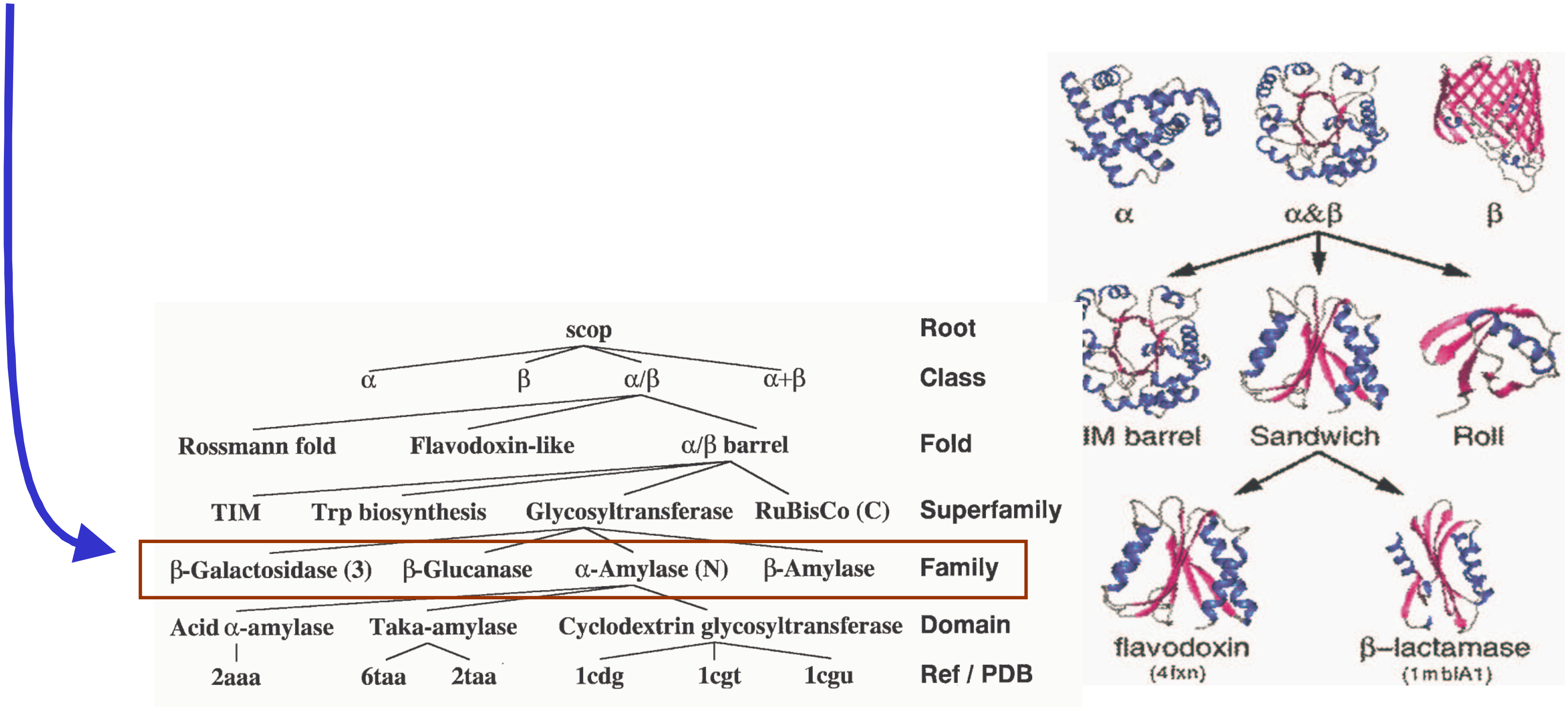
Fall 2011

The protein classification task

Given: amino-acid sequence of a protein

Do: predict the *family* to which it belongs

GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVCVLAHHFGKEFTPPVQAAYAKVVAGVANALAHKYH



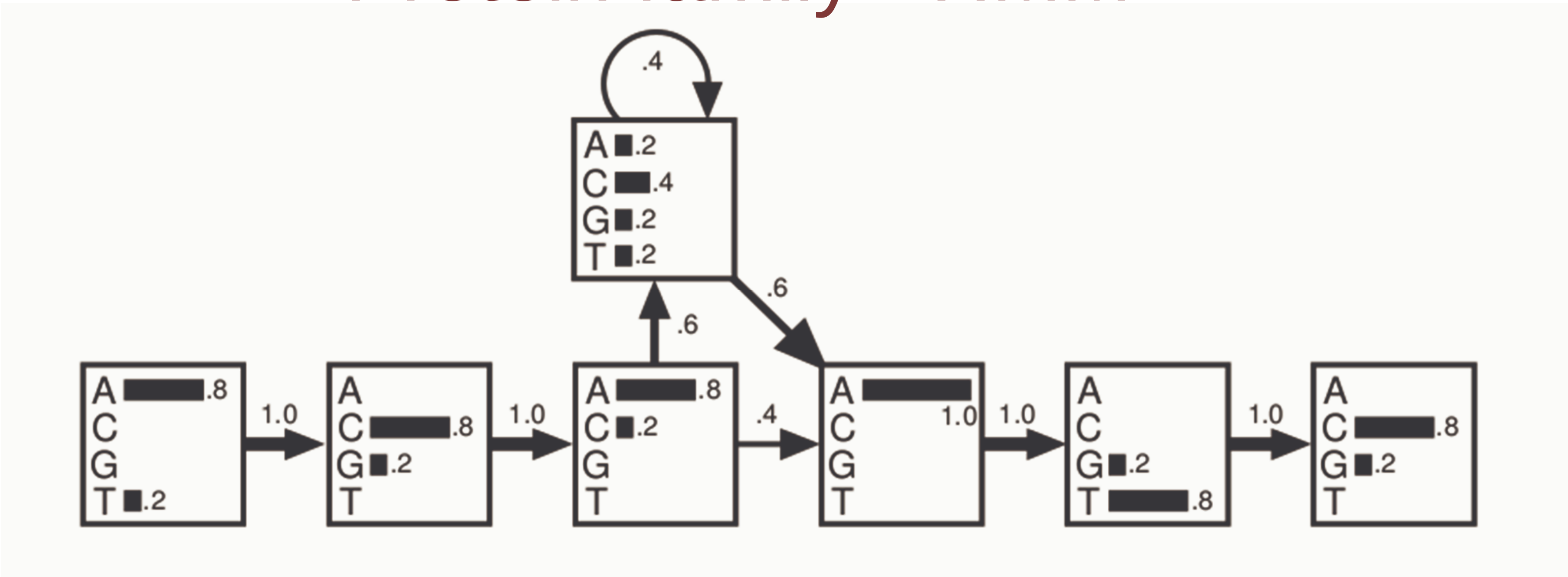
Protein family - a simplified view

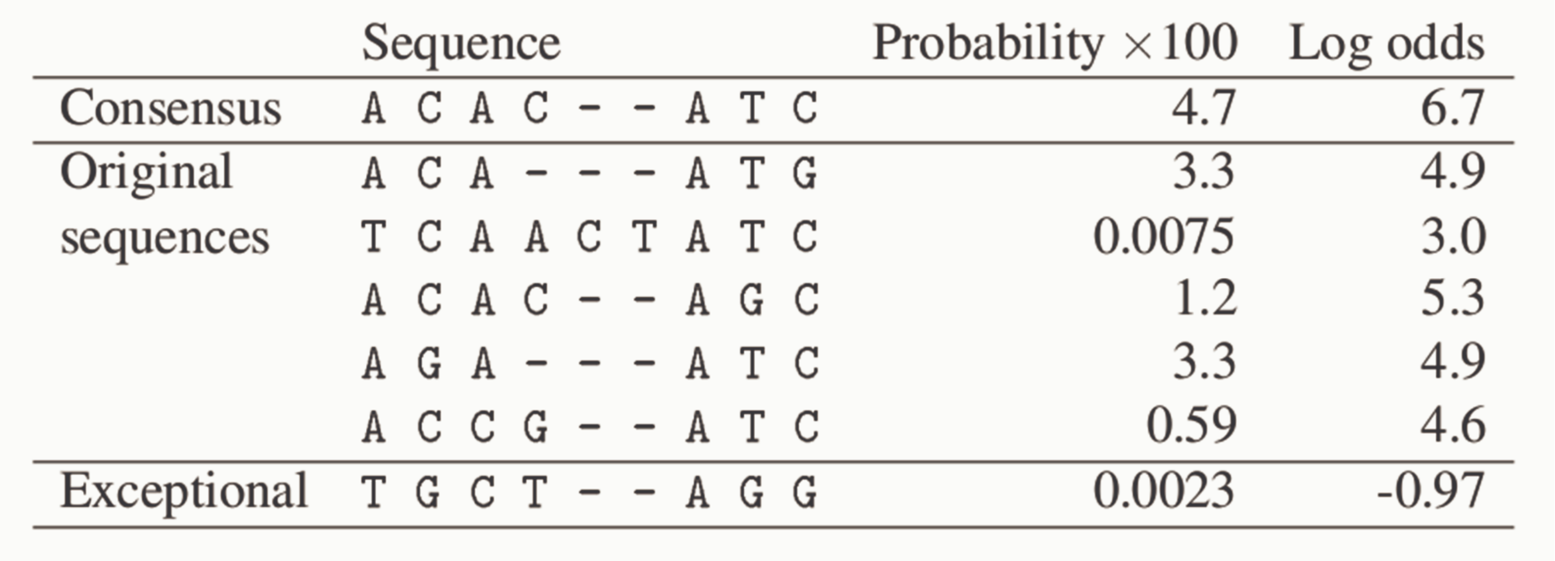
A	C	A	–	–	–	A	T	G	family
T	C	A	A	C	T	A	T	C	
A	C	A	C	–	–	A	G	C	
A	G	A	–	–	–	A	T	C	
A	C	C	G	–	–	A	T	C	

A	C	A	C	–	–	A	T	C	query 1
A	A	A	C	–	–	A	T	C	query 2
T	G	C	T	–	–	A	T	C	query 3

An example from Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.

Protein family - HMM



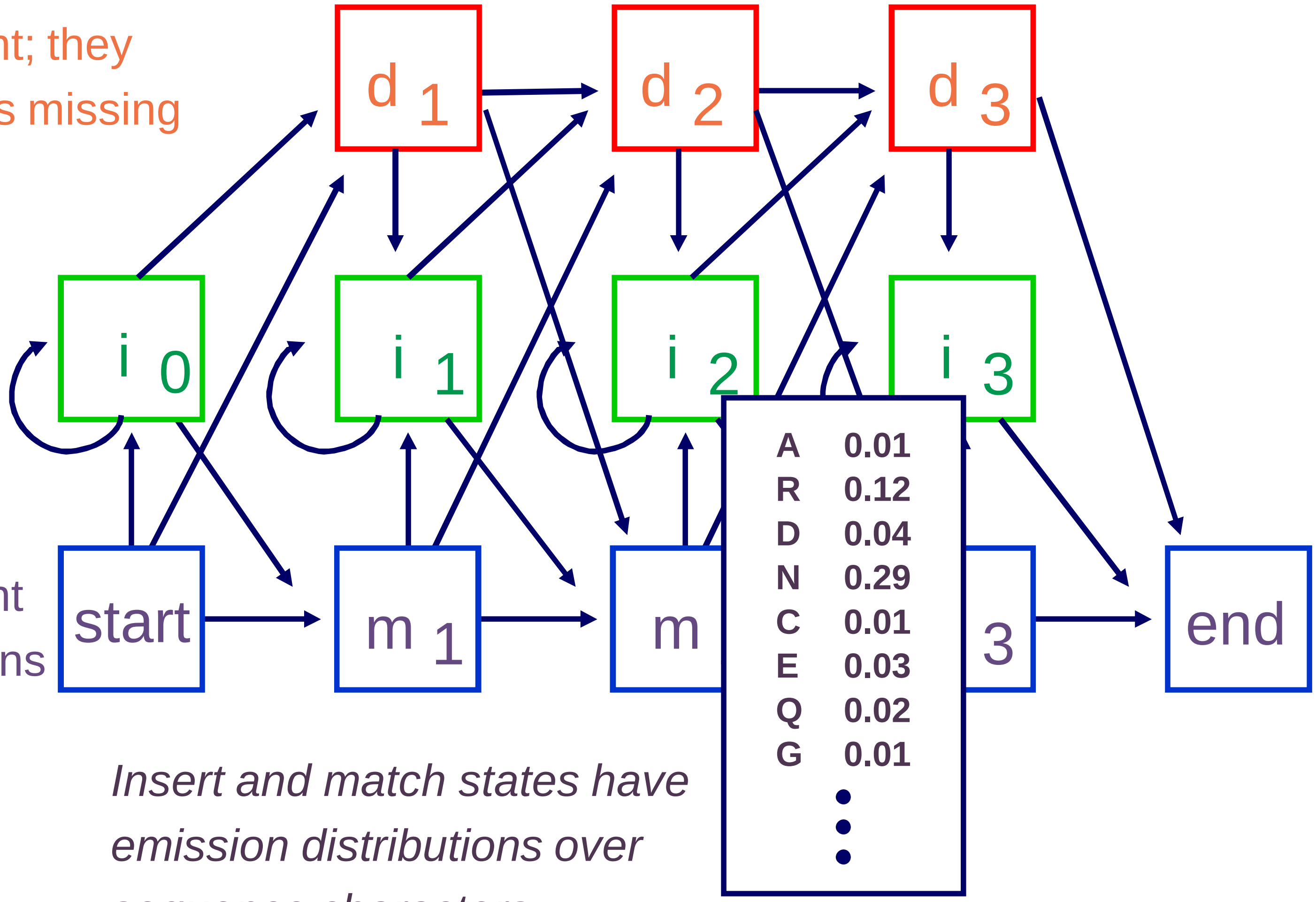


	Sequence	Probability × 100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original sequences	A C A - - - A T G	3.3	4.9
	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97

An example from Krogh: An Introduction to HMMs for Biological Sequences, CMMB 1998.

Profile HMMs

- profile HMMs are used to model families of sequences



A	0.01
R	0.12
D	0.04
N	0.29
C	0.01
E	0.03
Q	0.02
G	0.01

⋮

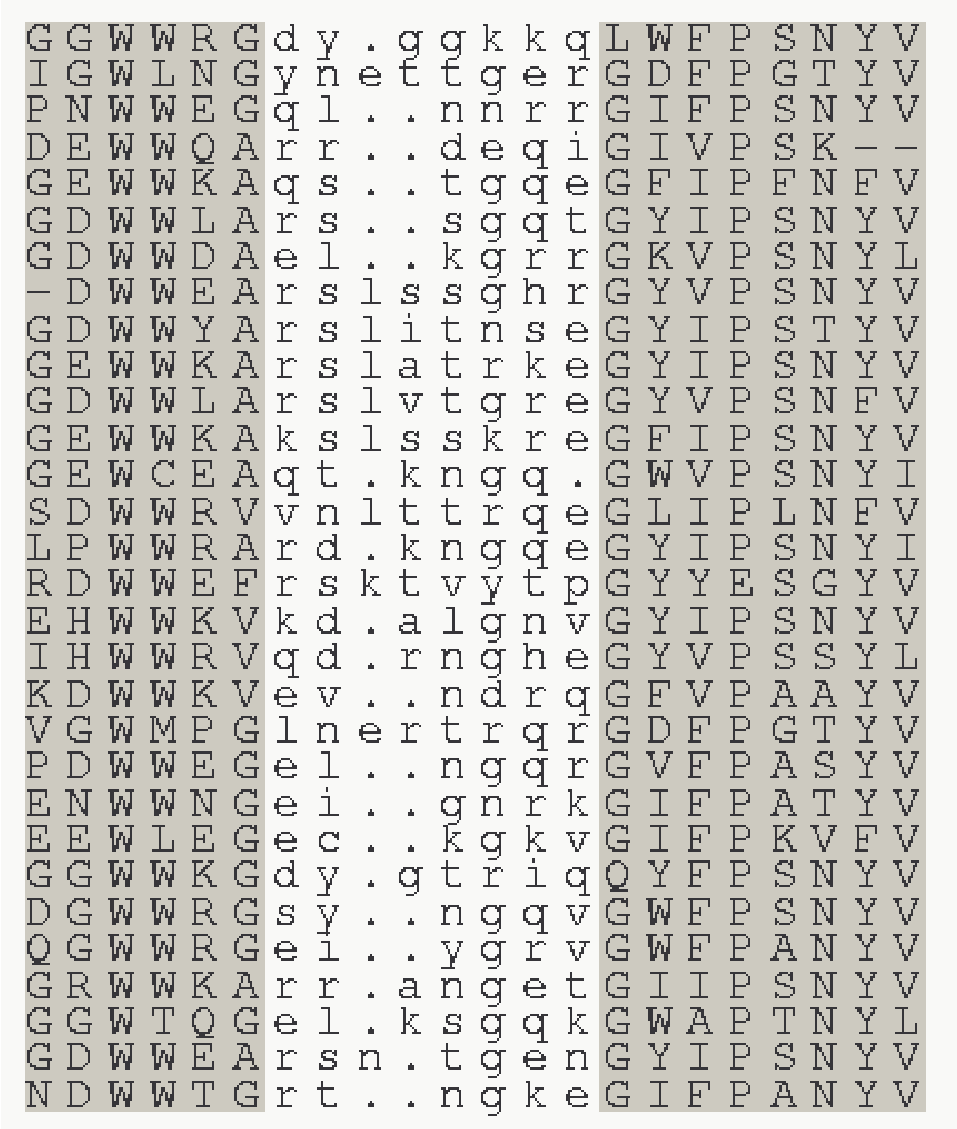
Delete states are silent; they Account for characters missing in some sequences

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



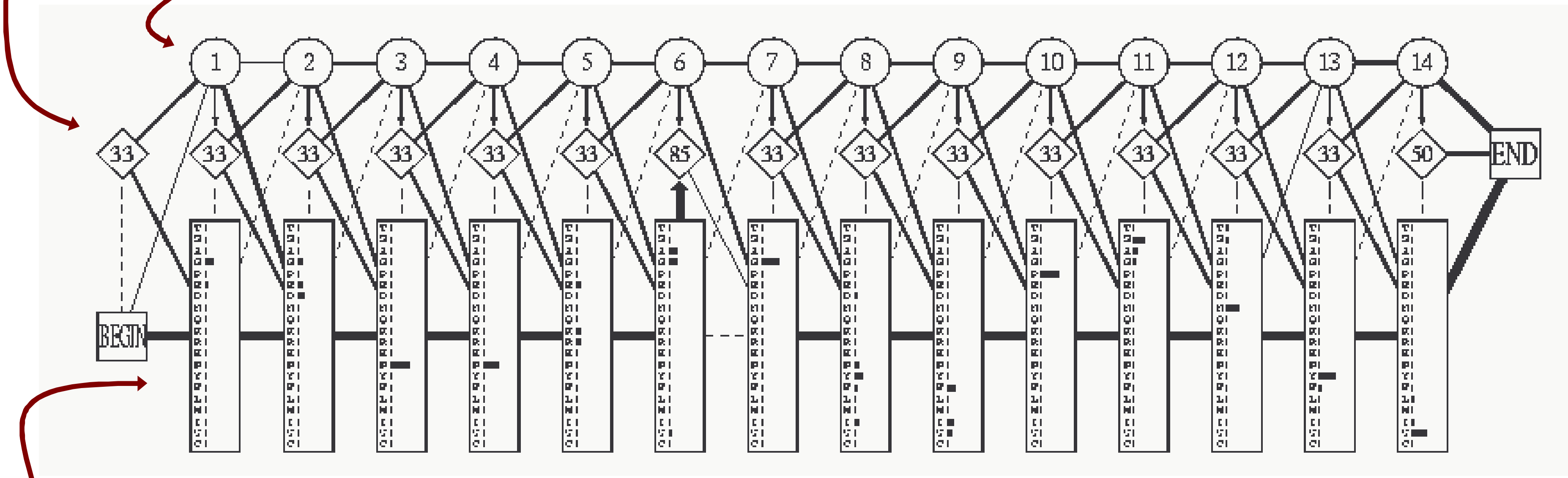
```
GGWWRGdy.ggkkqLWFPSNYV
IGWLNGynettgerGDFPGTYV
PNWWEGql..nnrrGIFPSNYV
DEWWQArr..deqiGIVPSK--
GEWWKAqs..tgqeGFIPFNFV
GDWWLArs..sgqtGYIPSNYV
GDWWDAel..kgrrGKVPSNYL
-DWWEArslssghrGYVPSNYV
GDWWYArslitnseGYIPSTYV
GEWWKArslatrkeGYIPSNYV
GDWWLArslvtgreGYVPSNFV
GEWWKAkslsskreGFIPSNYV
GEWCEAqt.kngq.GWVPSNYI
SDWWRVvnlttrqeGLIPLNFV
LPWWRArd.kngqeGYIPSNYI
RDWWEFrsktvytpGYYESGYV
EHWWKVkd.algnvGYIPSNYV
IHWWRVqd.rngheGYVPSSYL
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

insert states

delete states
(silent)



match states

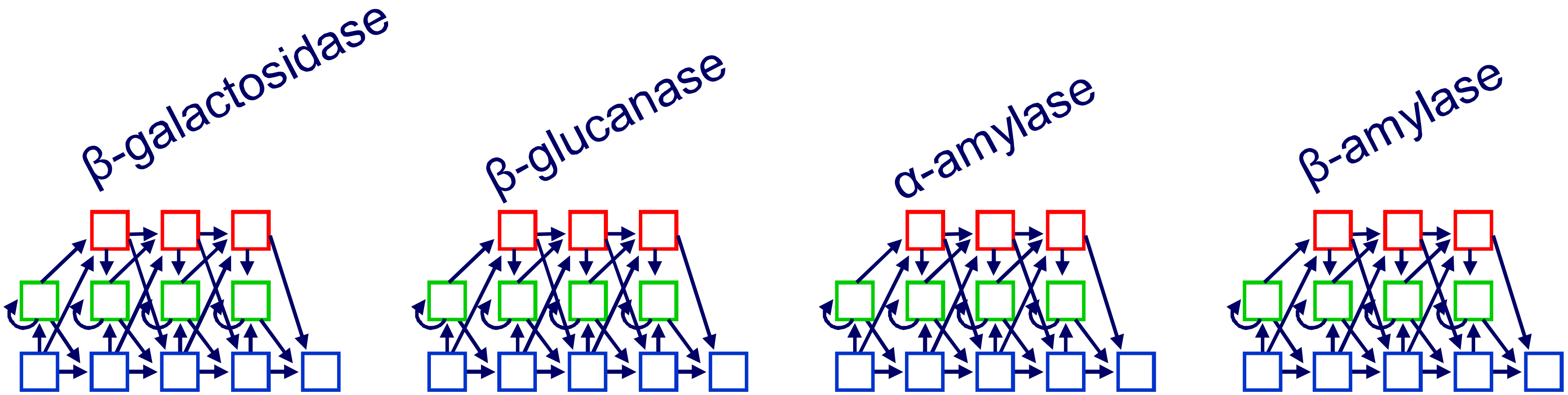
Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

Profile HMMs

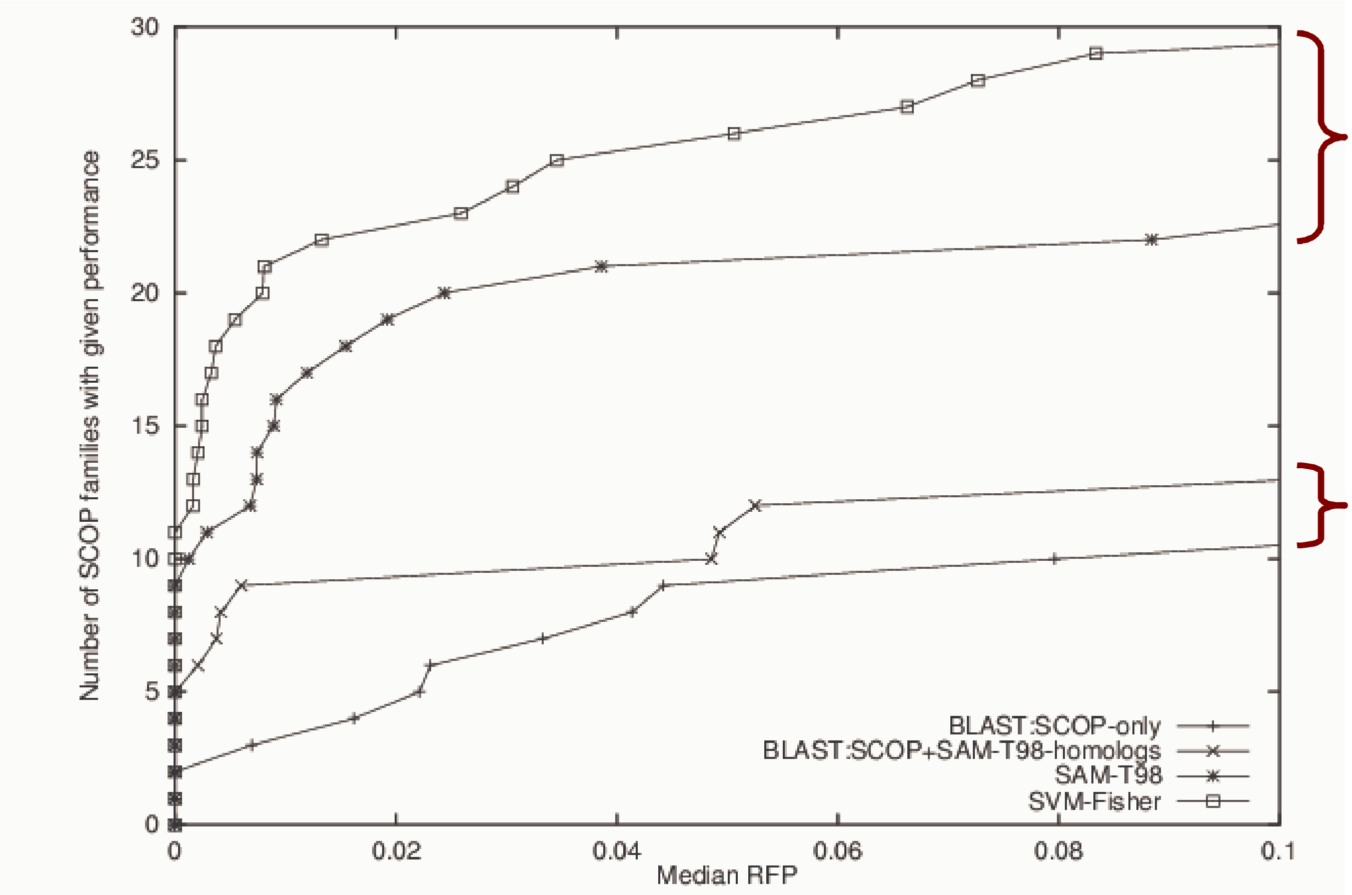
- to classify sequences according to family, we can train a profile HMM to model the proteins of each family of interest
- given a sequence x, use Bayes' rule to make classification

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

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



Profile HMM accuracy



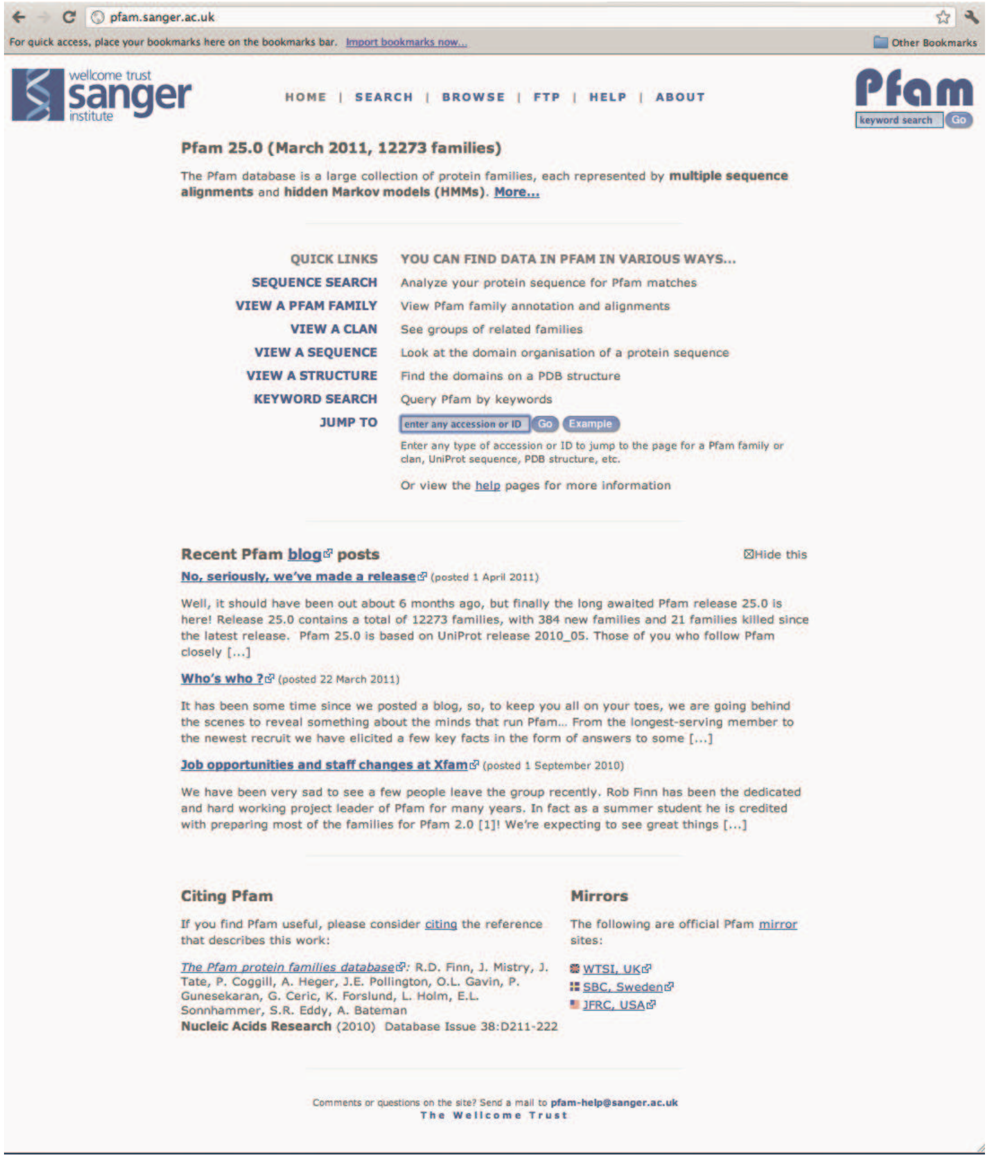
profile HMM-based methods

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

See Pfam database for a large collection profile HMMs


pfam.sanger.ac.uk
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Other Bookmarks
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go

Pfam 25.0 (March 2011, 12273 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc. Or view the help pages for more information

Recent Pfam blog posts ☒Hide this

No, seriously, we've made a release (posted 1 April 2011)

Well, it should have been out about 6 months ago, but finally the long awaited Pfam release 25.0 is here! Release 25.0 contains a total of 12273 families, with 384 new families and 21 families killed since the latest release. Pfam 25.0 is based on UniProt release 2010_05. Those of you who follow Pfam closely [...]

Who's who ? (posted 22 March 2011)

It has been some time since we posted a blog, so, to keep you all on your toes, we are going behind the scenes to reveal something about the minds that run Pfam... From the longest-serving member to the newest recruit we have elicited a few key facts in the form of answers to some [...]

Job opportunities and staff changes at Xfam (posted 1 September 2010)

We have been very sad to see a few people leave the group recently. Rob Finn has been the dedicated and hard working project leader of Pfam for many years. In fact as a summer student he is credited with preparing most of the families for Pfam 2.0 [1]! We're expecting to see great things [...]

Citing Pfam

If you find Pfam useful, please consider citing the reference that describes this work:

The Pfam protein families database: R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman
Nucleic Acids Research (2010) Database Issue 38:D211-222

Mirrors

The following are official Pfam mirror sites:

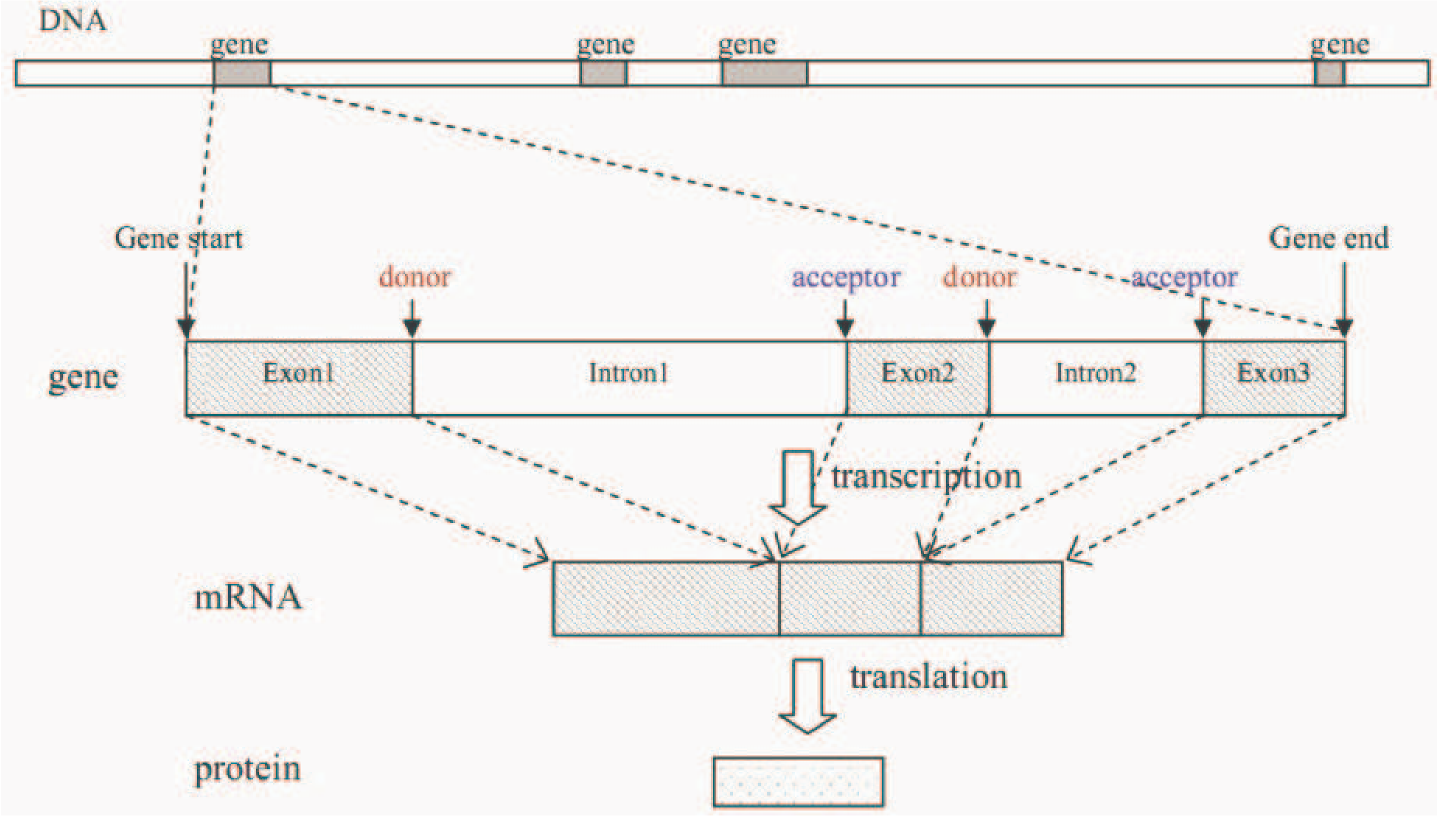
- WTSI, UK
- SBC, Sweden
- JFRC, USA

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**
The Wellcome Trust

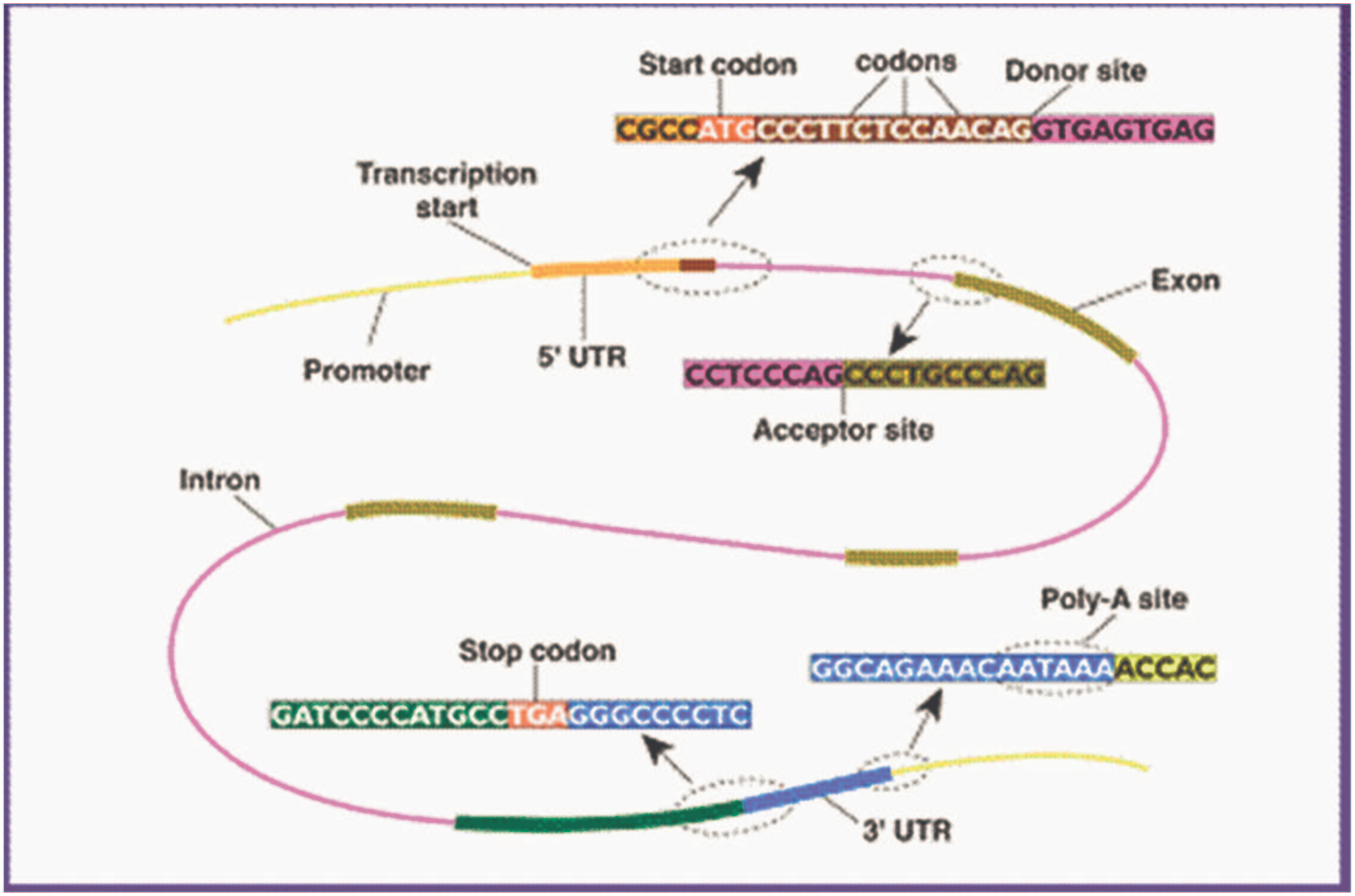
The gene finding task

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

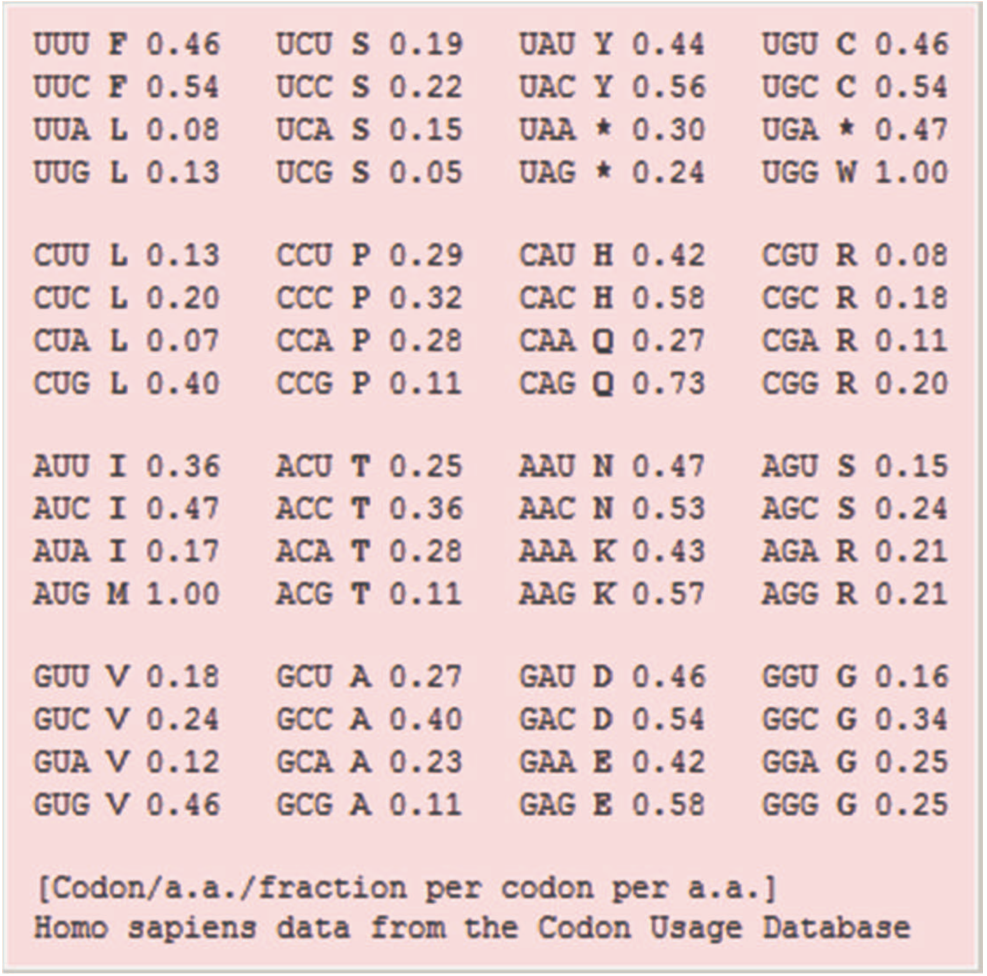
- **signals**: the sequence *signals* (e.g. splice junctions) involved in gene expression

- **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

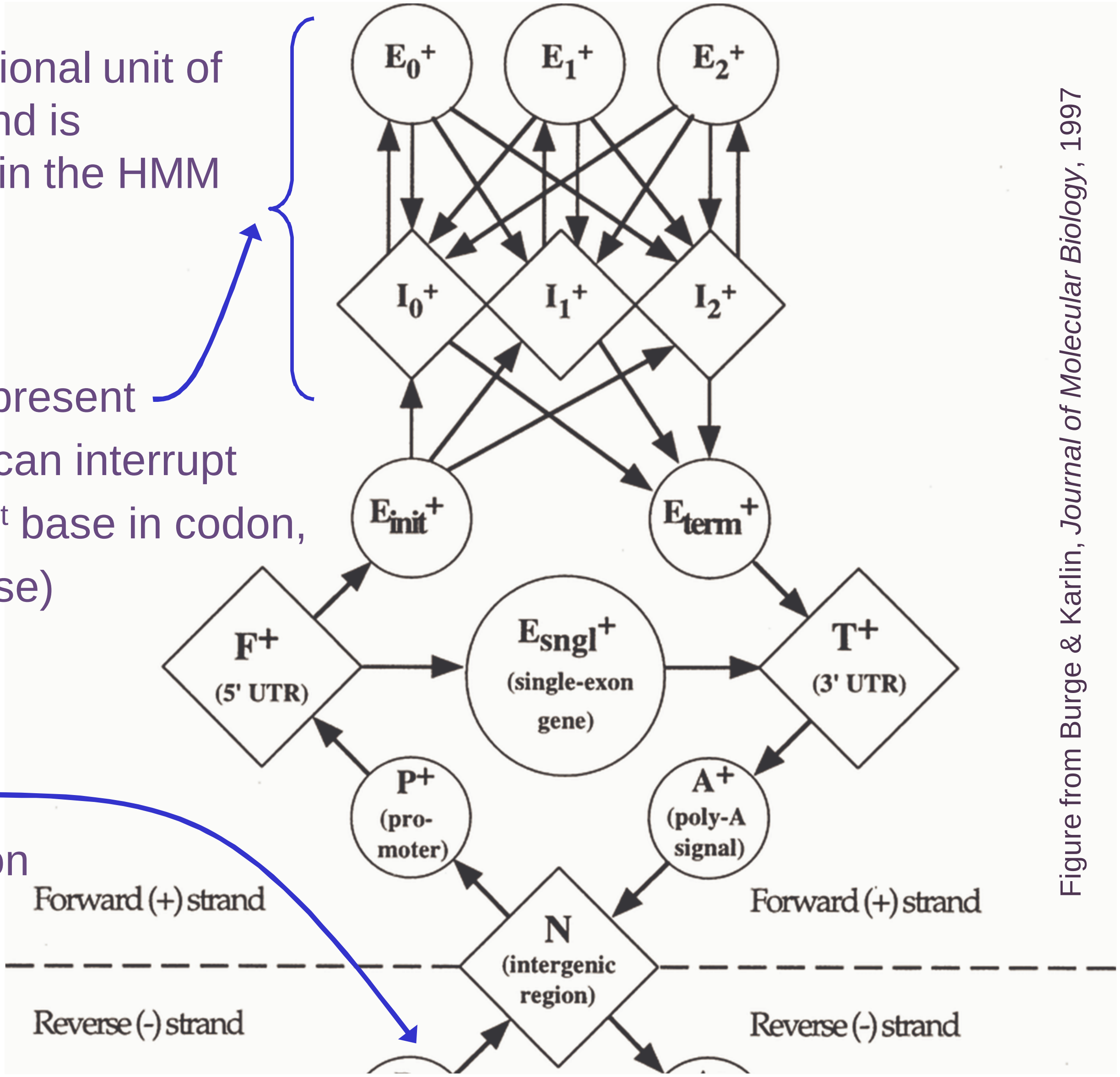


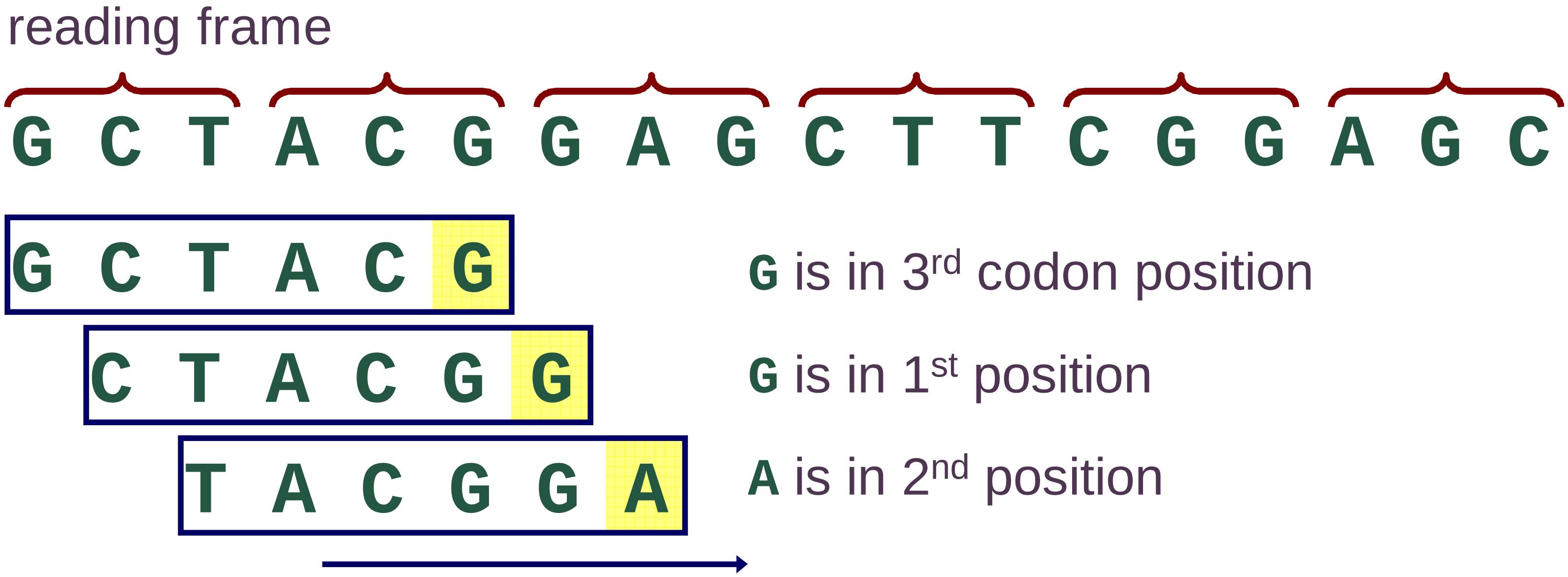
Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

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



reading frame

G C T A C G G A G C T T C G G A G C

G C T A C G

C T A C G G

T A C G G A

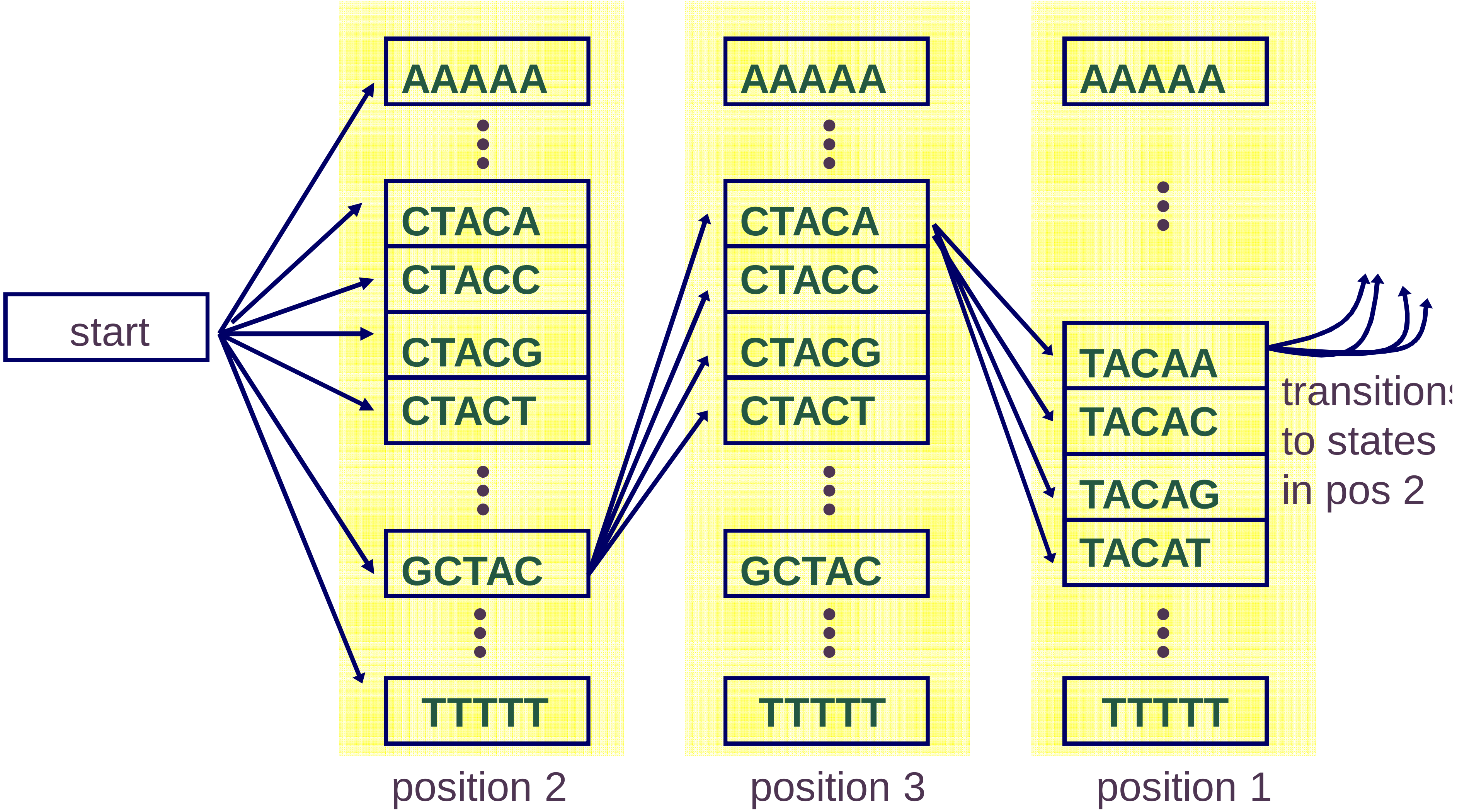
G is in 3rd codon position

G is in 1st position

A is in 2nd position

- can do this using an inhomogeneous model

A fifth-order inhomogenous Markov chain



Inference with the gene-finding HMM

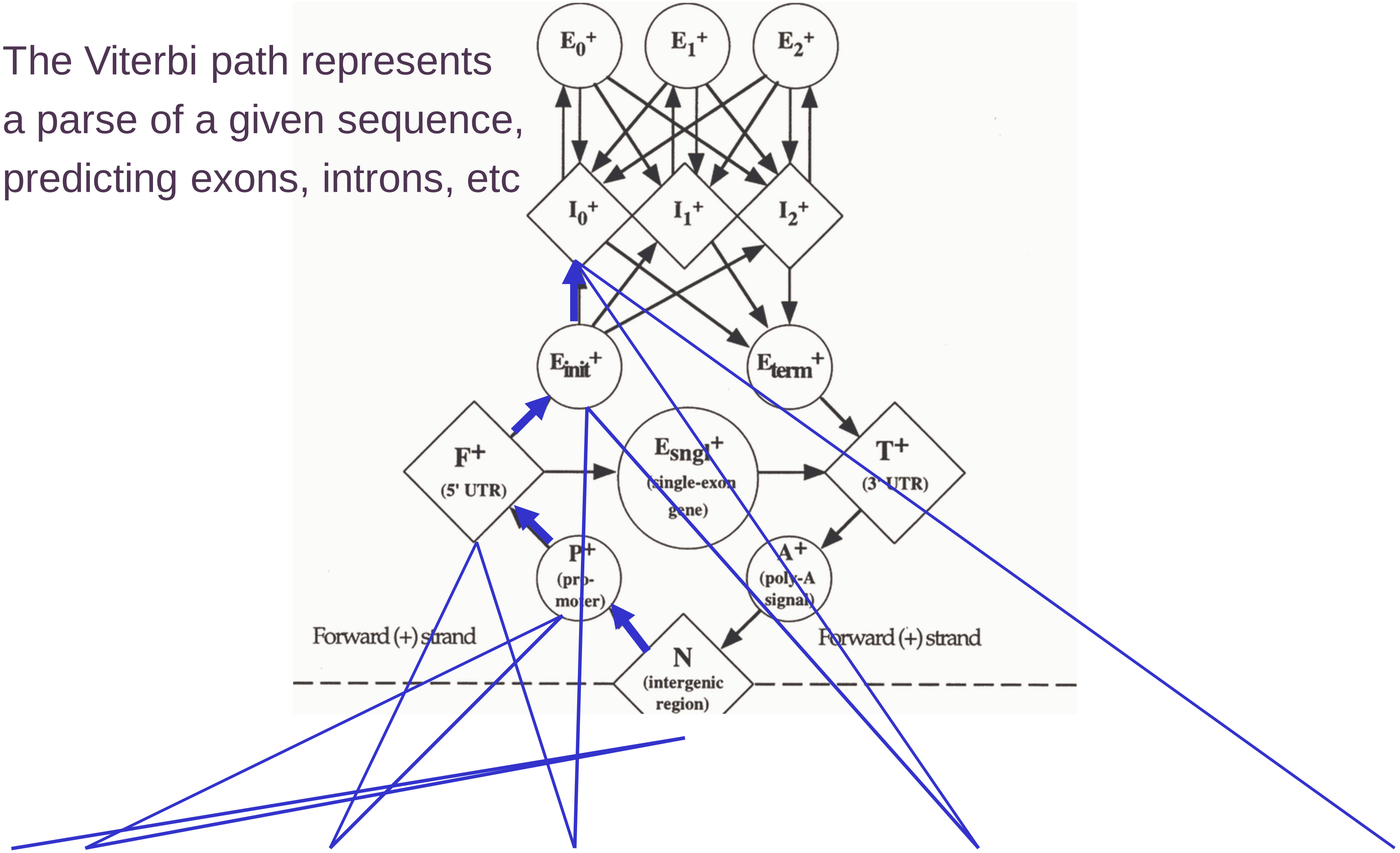
given: an uncharacterized DNA sequence

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



ACCGTTACGTGTCATTCTACGTGATCATCGGATCCTAGAATCATCGATCCGTGCGATCGATCGGATTAGCTAGCTTAGCTAGGAGAGCATCGATCGGATCGAGGAGGAGCCTATATAAATCAA

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.