Analysing Molecular Sequence Families

BISS 2014

Sergey Nurk

ABLAB SPbAU

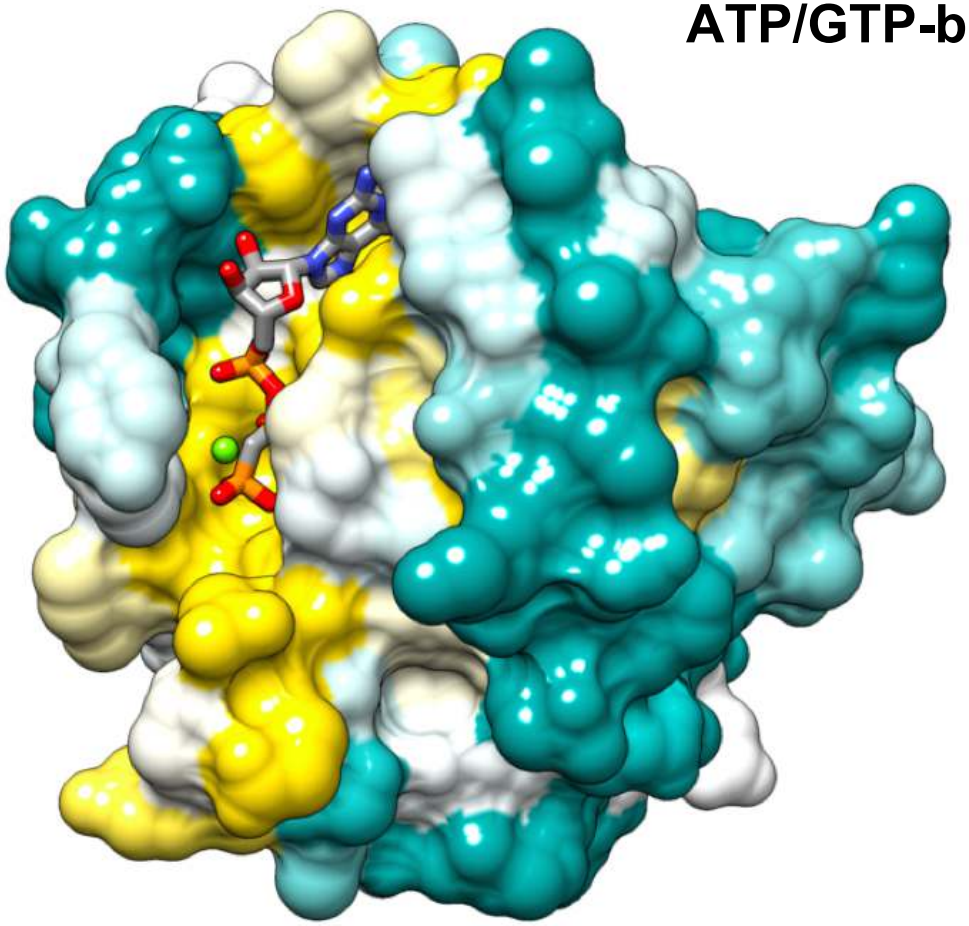
Motivation

Within a protein or nucleic acid sequence there may be a number of characteristic residues that occur consistently

These conserved “sequence fingerprints” (or **motifs**) usually contain functionally important elements

Analysis of such fingerprints and search for their occurrences has **A LOT** of applications

Protein functional domains



Conservation

ATP/GTP-binding proteins

```
    *       ***
FYGPPGLGKTSNIGG
LYGPPGLGKTANMGV
LFGPPGLGKTAHLGV
LIGPPGLGKTACLGV
LSGPPGLGKTAFMNA
ISGPIGTGKSAGIGI
LHGNPFTGKTASFSA
VCGLPGMGKTVETGF
VAGTPGVGKTVKLRF
IAGTPGVGKTVKMKF
IHGVPGTGKTMKKGY
G     GKT
```

Protein families

Part of multiple alignment of 7 globin sequences

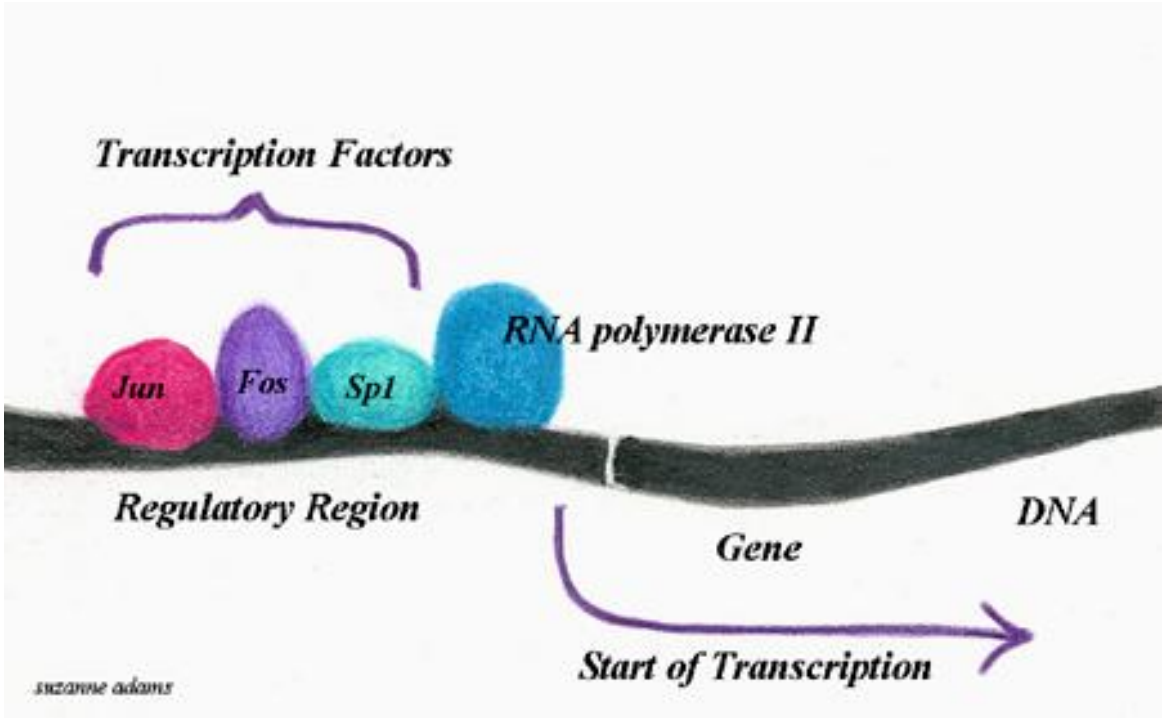
```
Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE                FFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    .. . v..Hg kv. a   a...l   d   . a l. l   H  .
```

Transcription factors

DNA binding proteins that modulate transcription of protein coding genes

Perhaps 5-10%, of all human proteins

Recognize specific DNA patches



Transcription factor binding sites

Binding motif of the yeast TF Pho4p

R06098 \TCA**CACGTG**GGA\
R06099 \GGC**CACGTG**CAG\
R06100 \TGA**CACGTG**GGT\
R06102 \CAG**CACGTG**GGG\
R06103 \TTC**CACGTG**CGA\
R06104 \ACG**CACGTT**GGT\
R06097 \CAG**CACGTT**TTC\
R06101 \TAC**CACGTT**TTC\

TFBS

- Complex “code”
- Short patches (4-8 bp)
- Often near each other (1 turn = 10 bp)
- Often reverse-complements (dimer symmetry)
- Not perfect matches

Goals

Compact representation allowing accurate member testing

Why is it useful?

- Protein classification
- Protein functional site annotation
- Determining TF regulating the gene
- DNA annotation (in particular gene search)
- ...

Representing sequence families

- **Pattern:** Describes a motif using a qualitative consensus sequence. Mismatches are not tolerated!
- **Profile:** Describes a motif using quantitative information captured in a position specific scoring matrix. Profiles quantify similarity and often span larger stretches of sequence.
- **Logos:** A useful visual representation of sequence motifs
- **Profile HMMs:** Generalization of profiles to support insertions deletions

PROSITE patterns

[LFI]-x-G-[PT]-P-G-x-G-K-[TS]-[AGSI]

- Each position in pattern is separated with a hyphen
- x can match any residue
- [ ] are used to indicate ambiguous positions in the pattern e.g., [LFI] means the pattern can match L, F, or I at this position
- { } are used to indicate residues that are not allowed at this position e.g., {S} means NOT S (not Serine)
- ( ) surround repeated residues, e.g., A(3) means AAA

PROSITE contains > 1700 patterns and profiles: http://prosite.expasy.org/

IUPAC codes

Symbol[2]	Description	Bases represented				
A	**A**denine	A				1
C	**C**ytosine		C			
G	**G**uanine			G		
T	**T**hymine				T	
U	**U**racil				U	
W	**W**eak	A			T	2
S	**S**trong		C	G		
M	a**M**ino	A	C			
K	**K**eto			G	T	
R	pu**R**ine	A		G		
Y	p**Y**rimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	a**N**y base (not a gap)	A	C	G	T	4

Patterns pros/cons

Advantages:

- Intuitive
- Straightforward to apply
- Databases with large numbers of proteins are available.

Disadvantages:

- Patterns are qualitative and deterministic
- Lose information about relative frequency of each residue at a position ([GAC] *vs* 0.6 G, 0.28 A, and 0.12 C)
- Cannot represent subtle sequence motifs

Position frequency matrix

c_{ja} - *occurrences of residue* ***a*** *at position* ***j***

f_{ja} - *relative frequency of residue i at position j*

w - *matrix width*

Binding motif of the yeast TF Pho4p (TRANSFAC matrix F$PHO4_01)

$$f_{ja} = c_{ja} / \sum_{a'} c_{ja'}$$

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.13	0.38	0.25	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.13	0.25
C	0.25	0.25	**0.38**	**1.00**	0.00	**1.00**	0.00	0.00	0.00	0.25	0.00	0.25
G	0.13	0.25	**0.38**	0.00	0.00	0.00	**1.00**	0.00	**0.63**	**0.50**	**0.63**	0.25
T	0.50	0.13	0.00	0.00	0.00	0.00	0.00	**1.00**	0.38	0.25	0.25	0.25
Sum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Pseudo-counts

Would like to avoid zeros

q_a - *prior residue probability for residue* ***a***

k - *pseudo weight*

$$f'_{ja} = \frac{c_{ja} + q_a k}{\sum_{a'} c_{ja'}}$$

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.15	0.37	0.26	0.04	**0.93**	0.04	0.04	0.04	0.04	0.04	0.15	0.26
C	0.24	0.24	**0.35**	**0.91**	0.02	**0.91**	0.02	0.02	0.02	0.24	0.02	0.24
G	0.13	0.24	**0.35**	0.02	0.02	0.02	**0.91**	0.02	**0.58**	**0.46**	**0.58**	0.24
T	0.48	0.15	0.04	0.04	0.04	0.04	0.04	**0.93**	0.37	0.26	0.26	0.26
Sum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Position-specific scoring matrix

PSSM (PWM, sequence profile) gives a *quantitative* description of a sequence motif

$$W_{ja} = ln\left(\frac{f'_{ja}}{q_a}\right)$$

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.175	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.175	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.325	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1.000	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

PSSM Score

Profiles assign a score to a query sequence of potential family member

S - *query string* $(s_1, \ldots, s_w)$

$$W_S = \sum_{j=1}^{w} W_{j s_j}$$

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	**-0.79**	0.13	-0.23	-2.20	1.05	-2.20	**-2.20**	**-2.20**	**-2.20**	-2.20	-0.79	-0.23
C	0.32	0.32	0.70	**1.65**	-2.20	1.65	-2.20	-2.20	-2.20	0.32	**-2.20**	0.32
G	-0.29	0.32	**0.70**	-2.20	**-2.20**	-2.20	1.65	-2.20	1.19	**0.97**	1.19	0.32
T	0.39	**-0.79**	-2.20	-2.20	-2.20	**-2.20**	-2.20	1.05	0.13	-0.23	-0.23	**-0.23**
residue r	A	T	G	C	G	T	A	A	A	G	C	T
W(r)	-0.79	-0.79	0.70	1.65	-2.20	-2.20	-2.20	-2.20	-2.20	0.97	-2.20	-0.23
Weight	**-11.67**											

What does it mean?

Adapted from http://rsat.ulb.ac.be/course/index.html

Probabilistic sequence models

Two simple generative sequence models:

Foreground (M): residue a is emitted at position j with probability f'_{ja}

Background (B): residue a is emitted with probability q_a at all positions

NB: Assumed independence between positions!

Likelihoods

Foreground:

$$P(S|M) = \prod_{j=1}^{w} f'_{js_j}$$

Background:

$$P(S|B) = \prod_{j=1}^{w} q_{s_j}$$

Reasonable score for discrimination -- **log odds**:

$$W_S = ln\left(\frac{P(S|M)}{P(S|B)}\right)$$

Score revisited

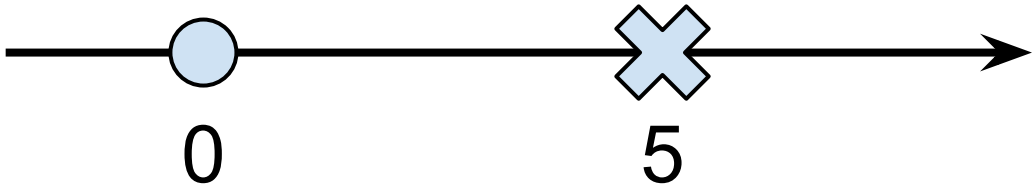
$$W_S = ln\left(\frac{P(S|M)}{P(S|B)}\right) = ln\left(\frac{\prod_{j=1}^{w} f'_{js_j}}{\prod_{j=1}^{w} q_{s_j}}\right)$$

$$= \sum_{j=1}^{w} ln\left(\frac{f'_{js_j}}{q_{s_j}}\right) = \sum_{j=1}^{w} W_{js_j}$$

Score interpretation

Got $\mathbf{W_s = 5}$

Is it a significant evidence that **S** is a family member?



Depends on the profile!

Simple heuristic threshold:

60% x (Max possible score for profile)

Score interpretation

Let’s think in terms of statistical testing

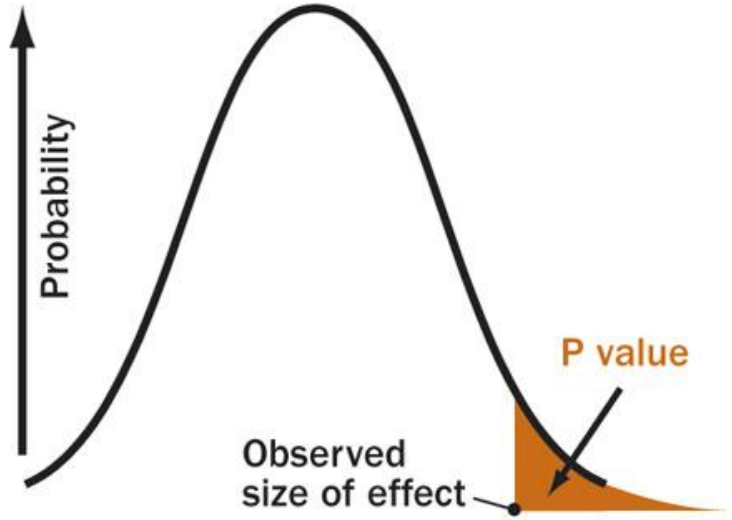
Null hypothesis: S is from the background

Alternative: S belongs to the family

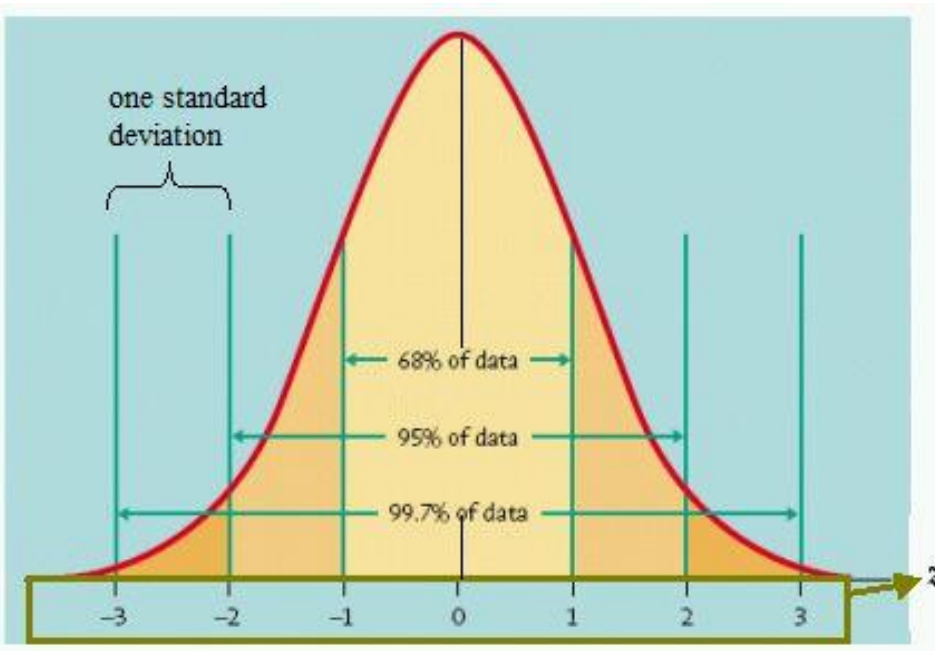
Consider distribution of scores under Null (simulated)

Score interpretation

p-value: probability of getting score as extreme under Null



Z-score (standard score): distance from mean in standard deviations



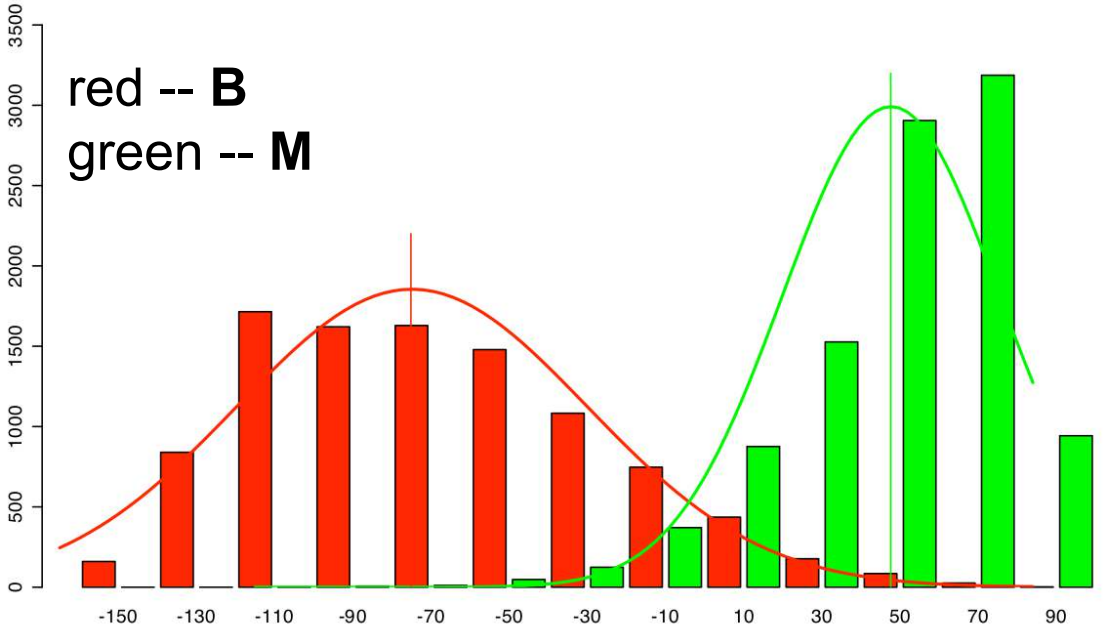
Score interpretation

What about **type II** errors?

Very dependant on the family!

The more it differs from the background the better!

Can look at score distribution for **M** (simulated)



Information content

Want some measure of difference between foreground and background model

Usually use Kullback-Leibler divergence

$$D_{KL}(P||Q) = \sum P(s) log_2 \left(\frac{P(s)}{Q(s)} \right)$$

$$IC_j = \sum_a f_{ja} log_2 \left(\frac{f_{ja}}{q_a} \right)$$

Sequence logo

Useful visual representation of sequence motifs

- Height on each position equal to its IC
- Height of each residue is proportional to its frequency

Sequence logo (cont.)

For DNA often make assumption of uniform background. Then

$$IC_j = 2 - H_j + e(n)$$

e(n) -- small sample correction

$$H_j = -\sum_a f_{ja} log_2 f_{ja}$$

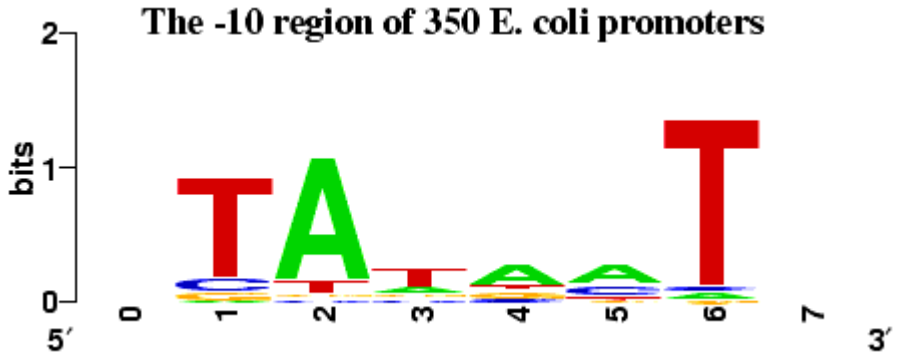
Looks familiar? :)

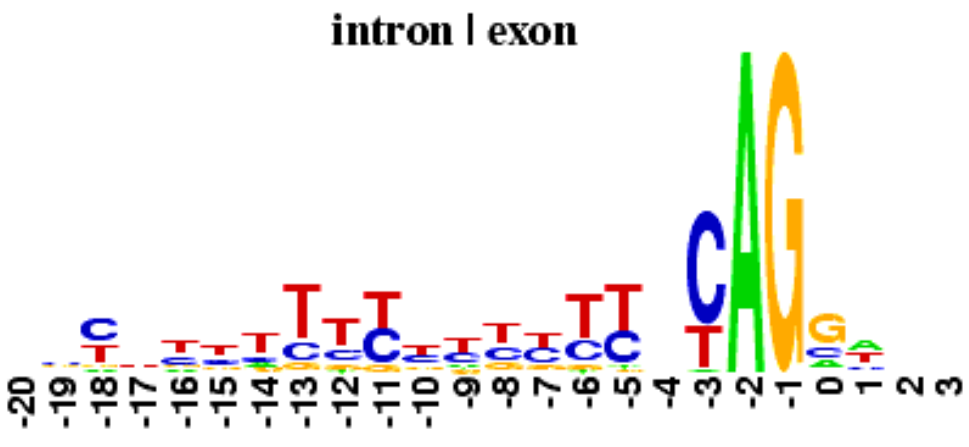
Height is interpreted as measure of conservation

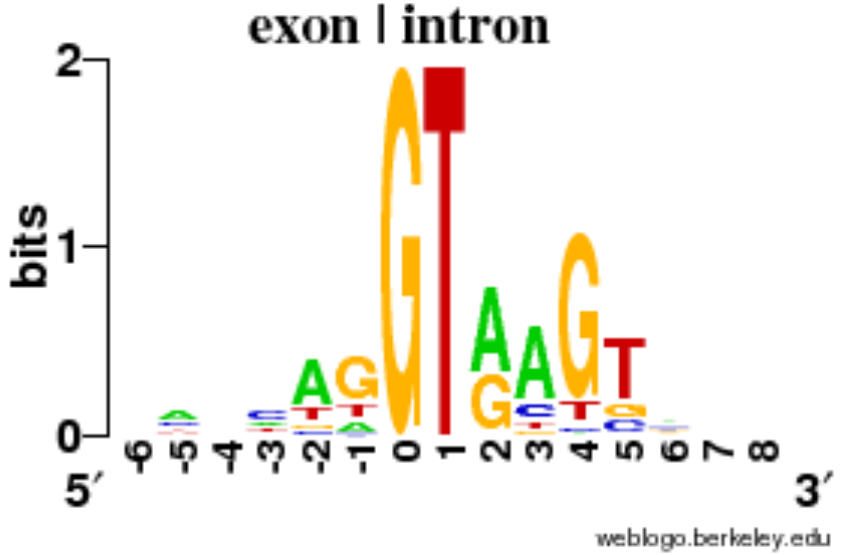
weblogo.berkley.edu

More examples

Pribnow box in *E.coli*

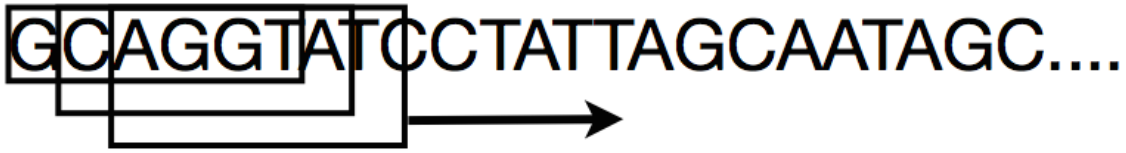


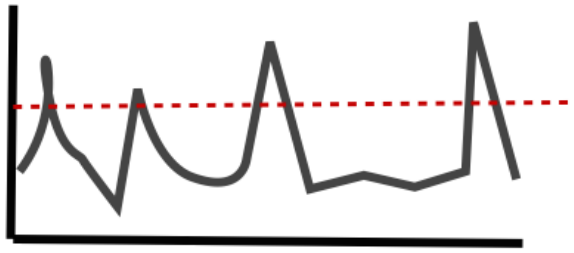




Searching for PSSM matches

If we do not allow gaps: apply PSSM at each "sliding window"





For gaps: dynamic programming

Multiple hypothesis testing! :)

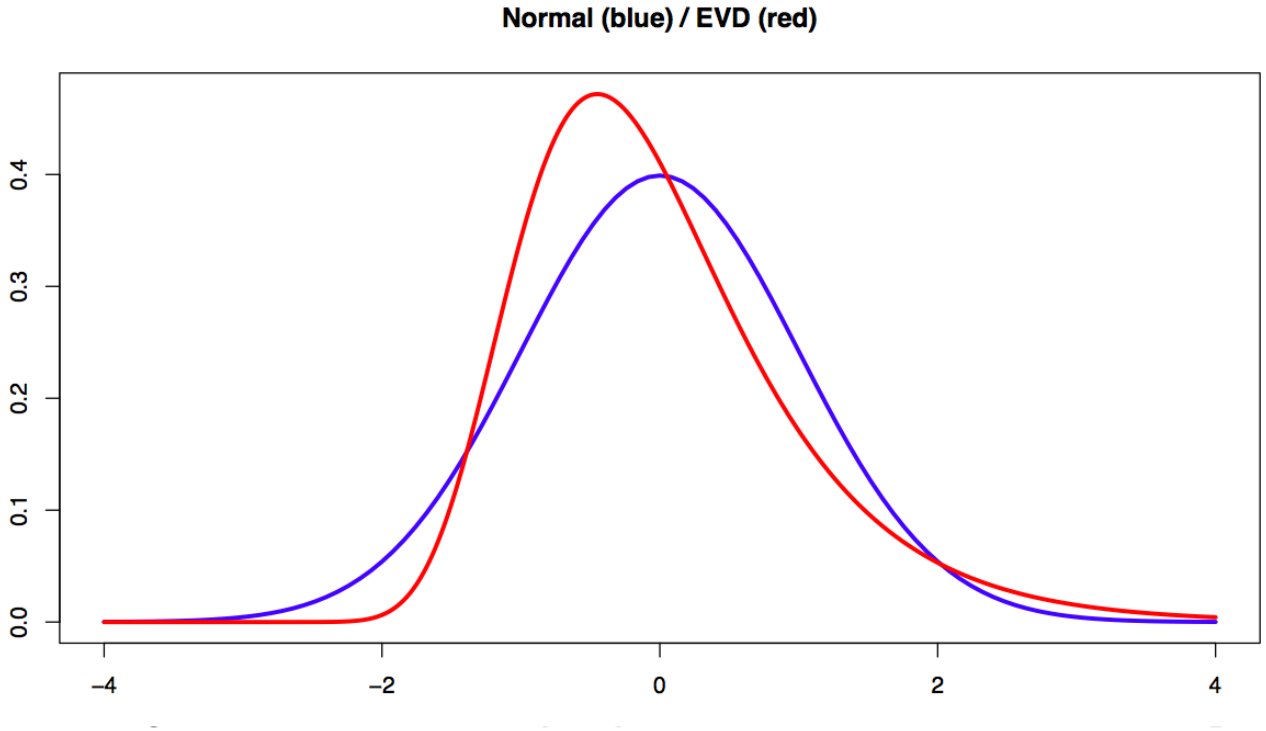
E-value - expected number of hits with scores at least as high in the random background sequence of the length equal to target

Best hit significance

What is the distribution of maximum score in the long random background sequence?

EVD (Extreme Value Distribution)

Much heavier right tail than of normal distribution



$$P(y \leq z) \approx \exp(-KNe^{-\lambda(z-\mu)})$$

Emission probabilities revisited

Why did we choose **M** emission probabilities as $f_{j\,a}$?

Among models with independent emission probabilities, $f_{j\,a}$ deliver maximum likelihood of observing input family sequences. Same with **B** and q_a

Might give poor estimates if we have small sample size (insufficient data).

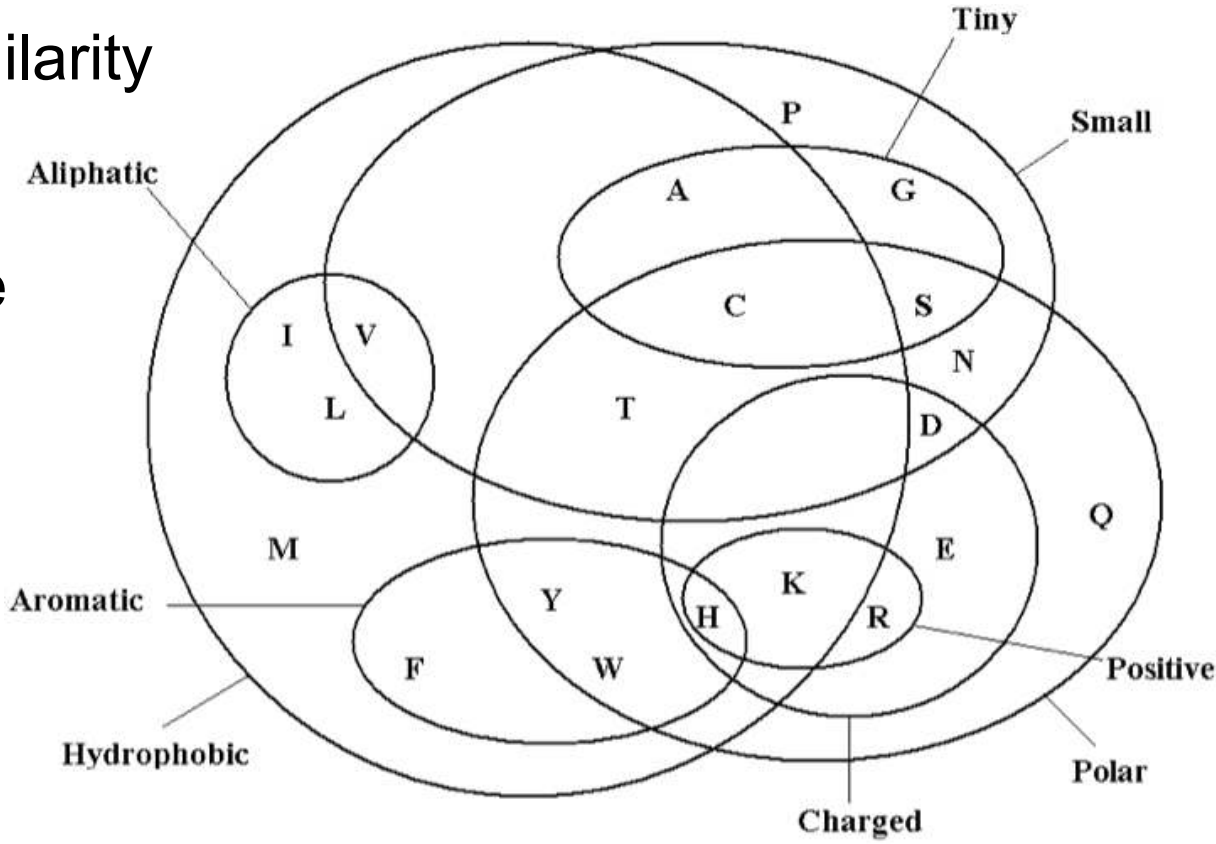
Pseudocounts solve the problem for DNA more or less…

But still poor performance for proteins!

Amino acid scoring matrices

Similarity scoring matrices can be constructed from based on different properties

- coding codon similarity
- hydrophilicity
- charge
- molecular volume
- …



Scoring matrices

Most commonly used scoring matrices were developed from experimental data

- PAM family (Dayhoff 1978)

Improvements:

  - Pet91
  - Gonnet matrices

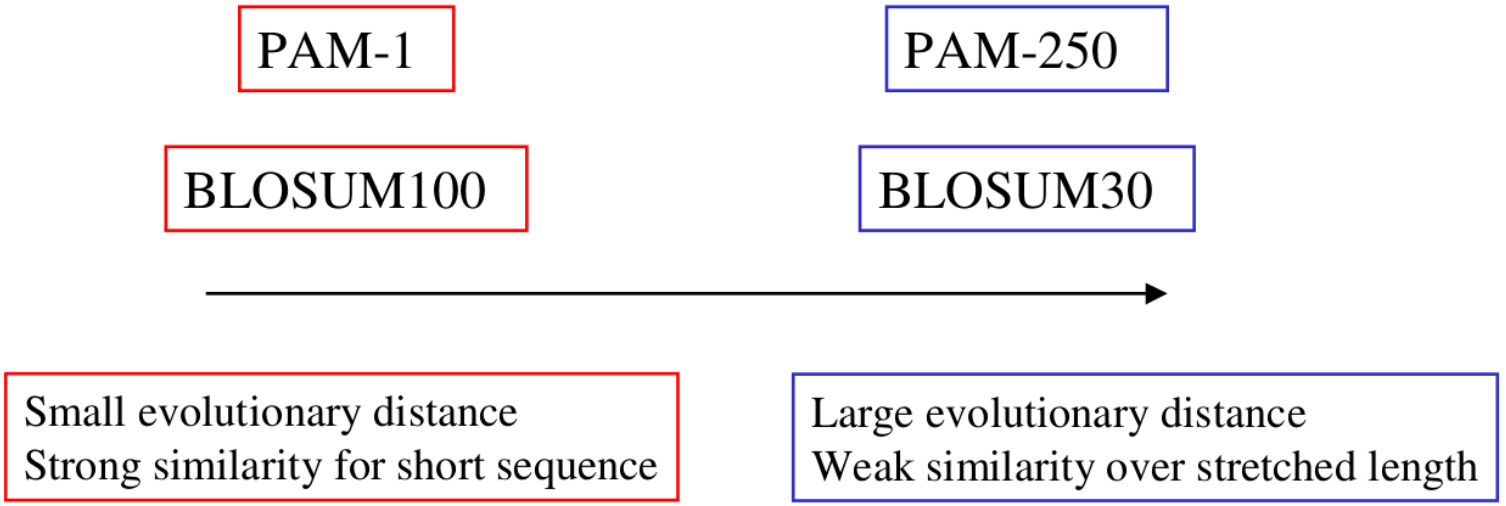
- BLOSUM family (Henikoff & Henikoff 1992)

All of them contain **log-odds scores**!

BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Bottom line



For general purposes: BLOSUM 62, PAM 120

Tons of recommendations… Most useful one:

In all cases it is recommended to use more than one matrix for any purpose :)

Average score method

$$W_{ja} = \sum_{b} f_{jb} * S_{ab}$$

Widely used!

No obvious interpretation of resulting score :(

Substitution matrix mixtures

$$\alpha_{ja} = A \sum_b f_{jb} P(a|b)$$

$$e_{M_j}(a) = \frac{c_{ja} + \alpha_{ja}}{\sum_{a'} c_{ja'} + \alpha_{ja'}}$$

A -- positive constant

Reasonable (not well theoretically justified) way to use substitution probabilities for obtaining pseudo-counts

Estimation based on ancestor

$$P(y_j = a|\text{alignment}) = \frac{q_a \prod_k P(x_j^k|a)}{\sum_{a'} q_{a'} \prod_k P(x_j^k|a')}$$

Probability that ancestor has amino-acid **a** at position **j**

$$e_{M_j}(a) = \sum_{a'} P(a|a') P(y_j = a'|\text{alignment}).$$

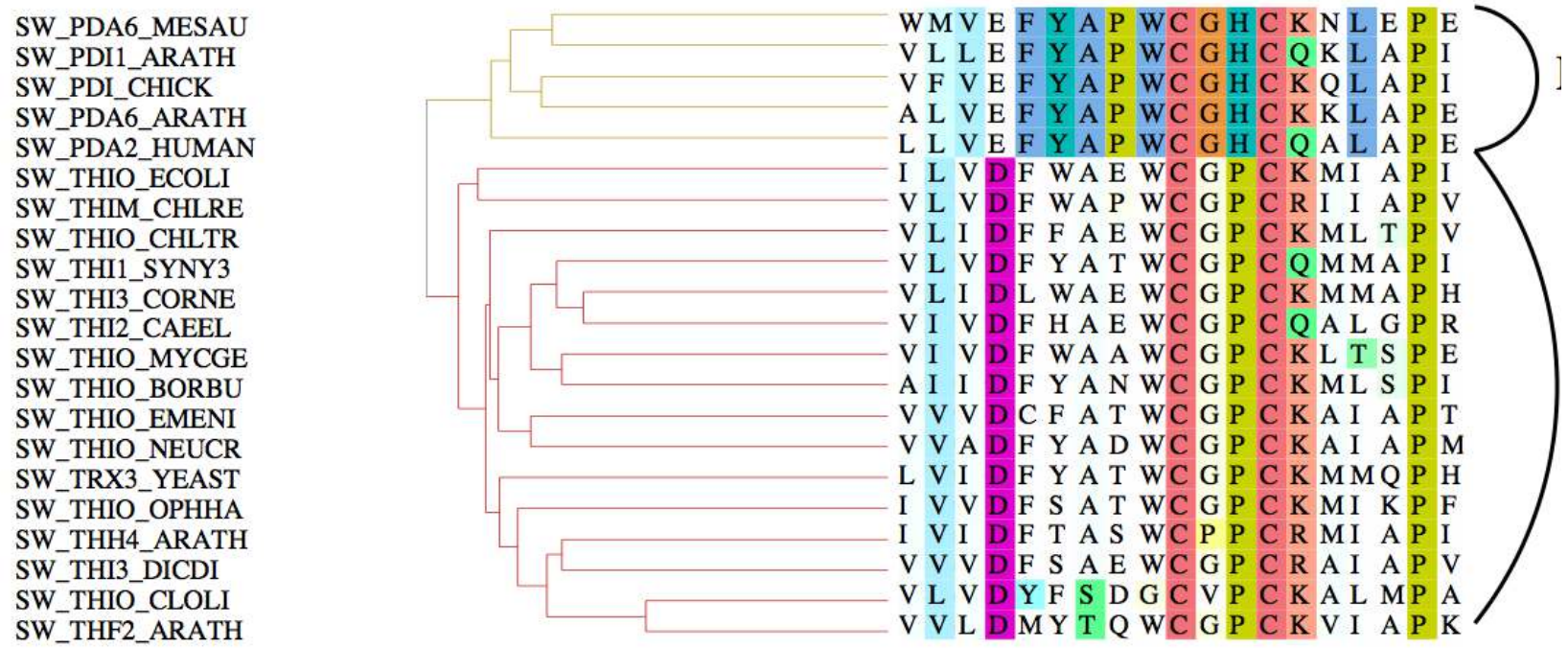
Probability to get AA **a** at position j wrt ancestor mixture

Sequence weighting

An MSA is often made of a few distinct sets of related sequences, or sub-families

Sub-families can be very differently populated, influencing observed residue frequencies

Solution: sequence weighting



Some PSSM problems

- Did not consider dependencies between positions.
  - Generalizations possible, but usually don’t have enough data

- Hard to work with insertions/deletions
  - Solution: **Profile HMMs**

Markov chains

Stochastic processes of transitions between a finite series of states **1,2,...,n**

Markov property:
state at the next step depends **only** on the current state
Moreover the transition probabilities do not depend on **t.**

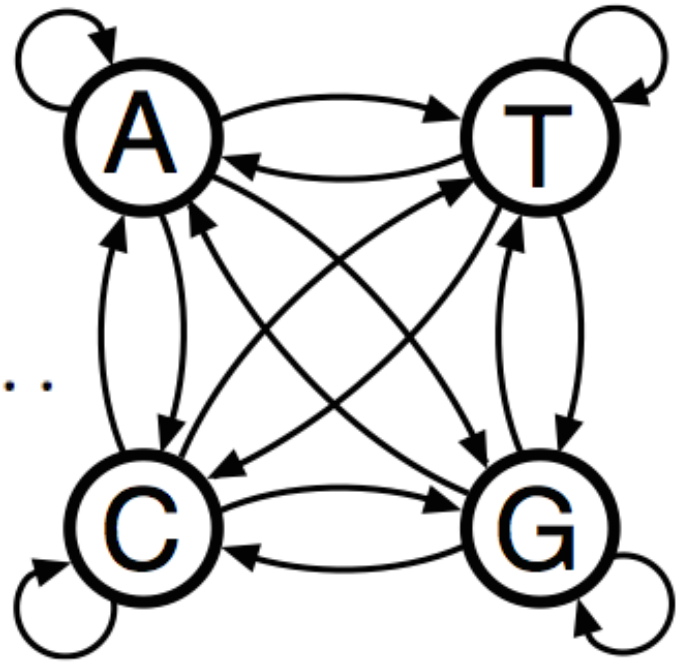
$$P(\pi_1, \pi_2, \pi_3, \ldots) = P(\pi_1)P(\pi_2|\pi_1)P(\pi_3|\pi_2)\ldots$$

Markov chains

Markov chains can be represented by state diagrams, which consist of states and connecting transitions

- States: **1,2,3,...,*n***
- Initial probabilities: a_{0k}
- Transition probabilities: a_{kl}

$$P(\pi_1, \pi_2, \pi_3, \ldots) = a_{0\pi_1} a_{\pi_1\pi_2} a_{\pi_2\pi_3} \ldots$$



Hidden Markov Models

States: $1, 2, 3, \ldots$

Paths: sequences of states $\pi = (\pi_1, \pi_2, \ldots)$

Transitions: $a_{k,l} = P(\pi_i = l \mid \pi_{i-1} = k)$

Emissions: $e_k(b) = P(x_i = b \mid \pi_i = k)$

Observed data: emission sequence

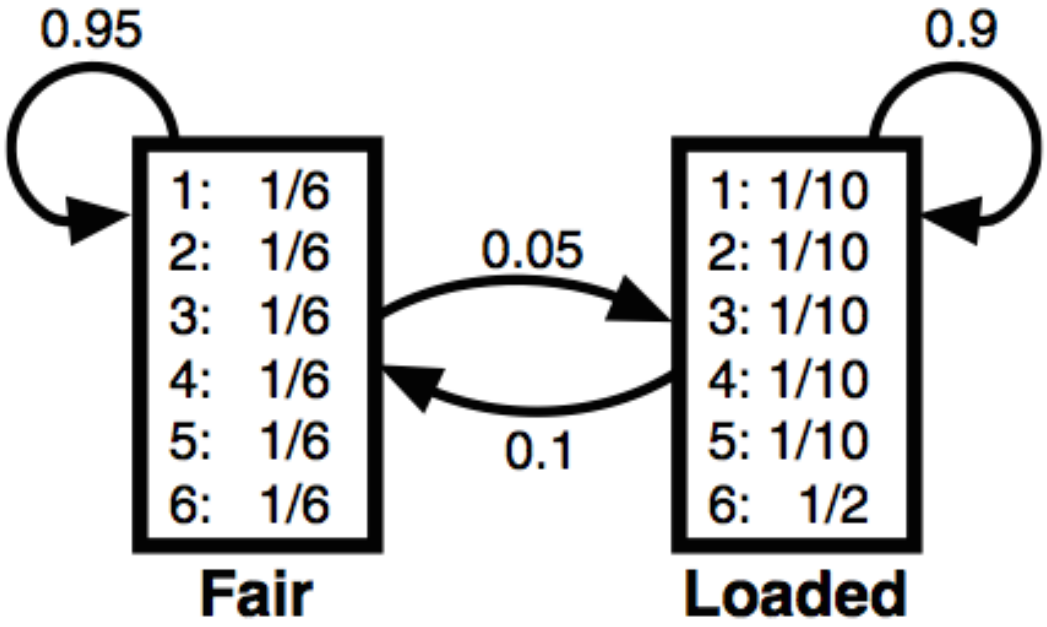
Hidden data: state/transition sequence

Probability of observed and hidden paths

$$P(x, \pi) = a_{0,\pi_1} \prod_{i=1}^{n} e_{\pi_i}(x_i) \cdot a_{\pi_i,\pi_{i+1}}$$

Occasionally dishonest casino

1 fair die, 1 “loaded” die, occasionally swapped



HMM-related problems

- Find probability of observed sequence

$$P(x)$$

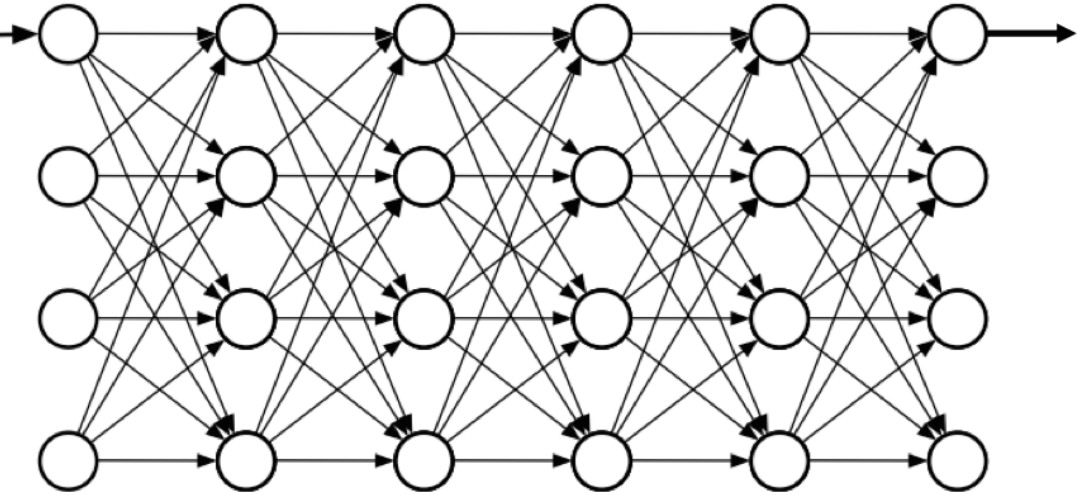
- Find most likely sequence of hidden states

$$\pi^* = \arg\max_{\pi} P(x, \pi)$$

- Find most probable state at step ***i***
  - Sequence of most probable states -- posterior decoding

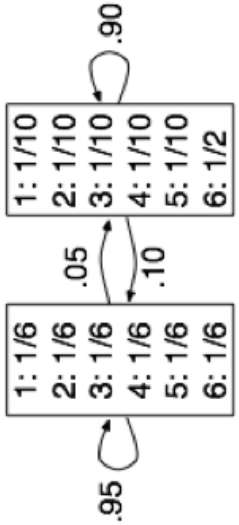
$$\hat{\pi}_i = \arg\max_{k} P(\pi_i = k \mid x)$$

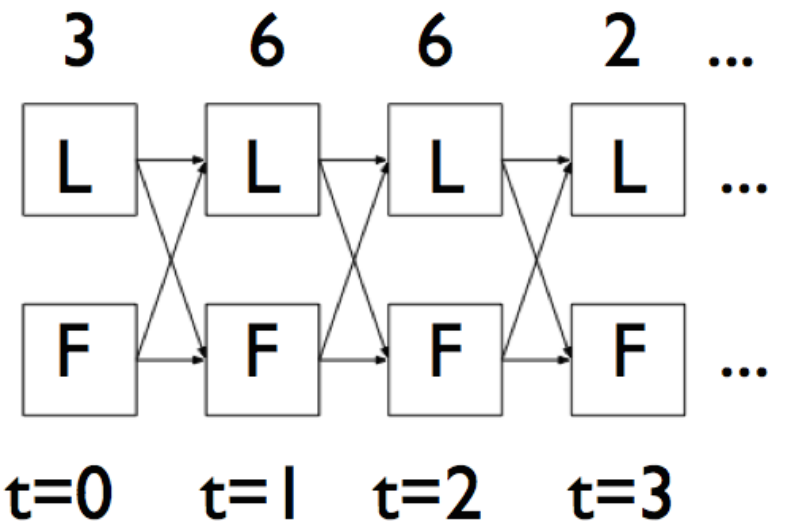
“Unrolling” HMM



x_1 x_2 x_3 x_4

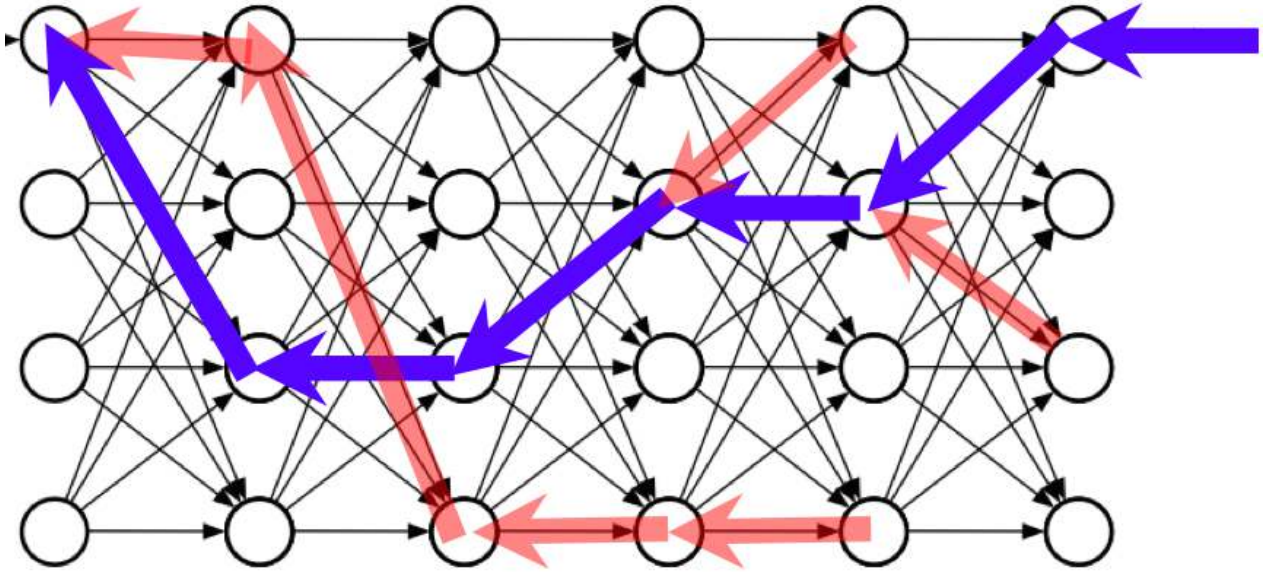
Emissions/sequence positions





Viterbi algorithm

$v_l(i) =$ probability of the most probable path emitting $x_1, x_2, \ldots, x_i$ and ending in state l



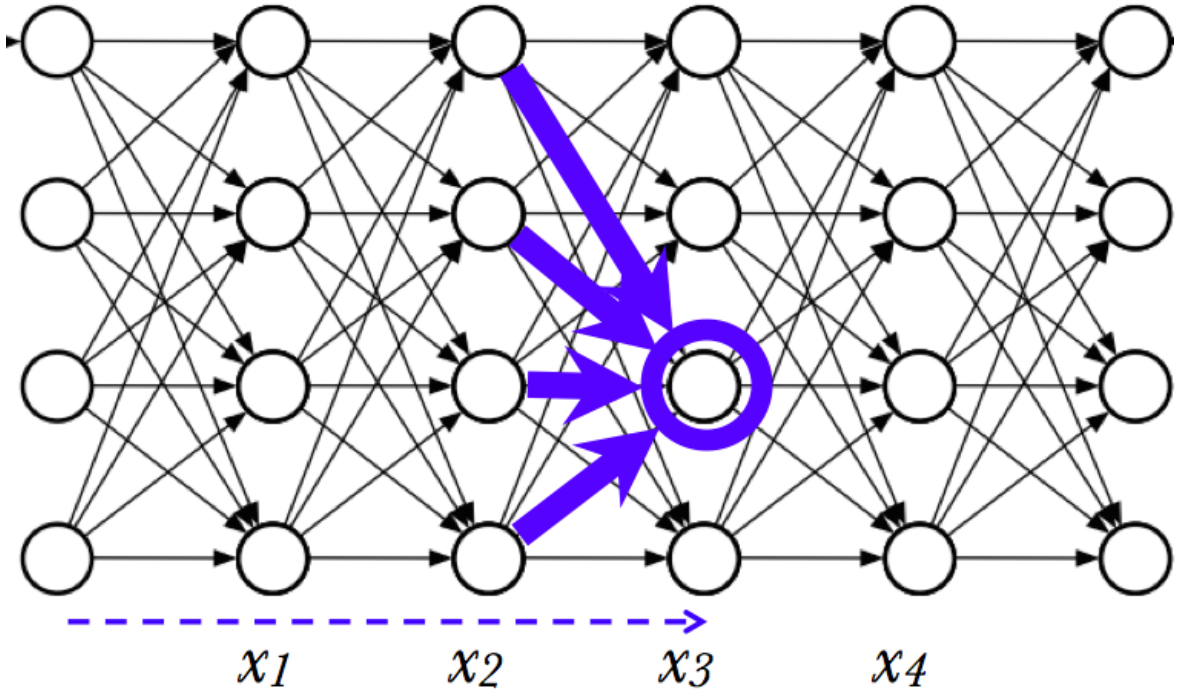
$$v_l(i+1) = e_l(x_{i+1}) \cdot \max_k (v_k(i)\, a_{k,l})$$

$$back_l(i+1) = \arg\max_k (v_k(i)\, a_{k,l})$$

Occasionally dishonest casino

```
Rolls   315116246446644245311321631164152133625144543631656626566666
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls   651166453132651245636664631636663162326455236266666625151631
Die     LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLLFFFFFFFFF
Viterbi LLLLLLFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls   222555441666566563564324364131513465146353411126414626253356
Die     FFFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls   366163666466232534413661661163252562462255265252266435353336
Die     LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls   233121625364414432335163243633665562466662632666612355245242
Die     FFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
```

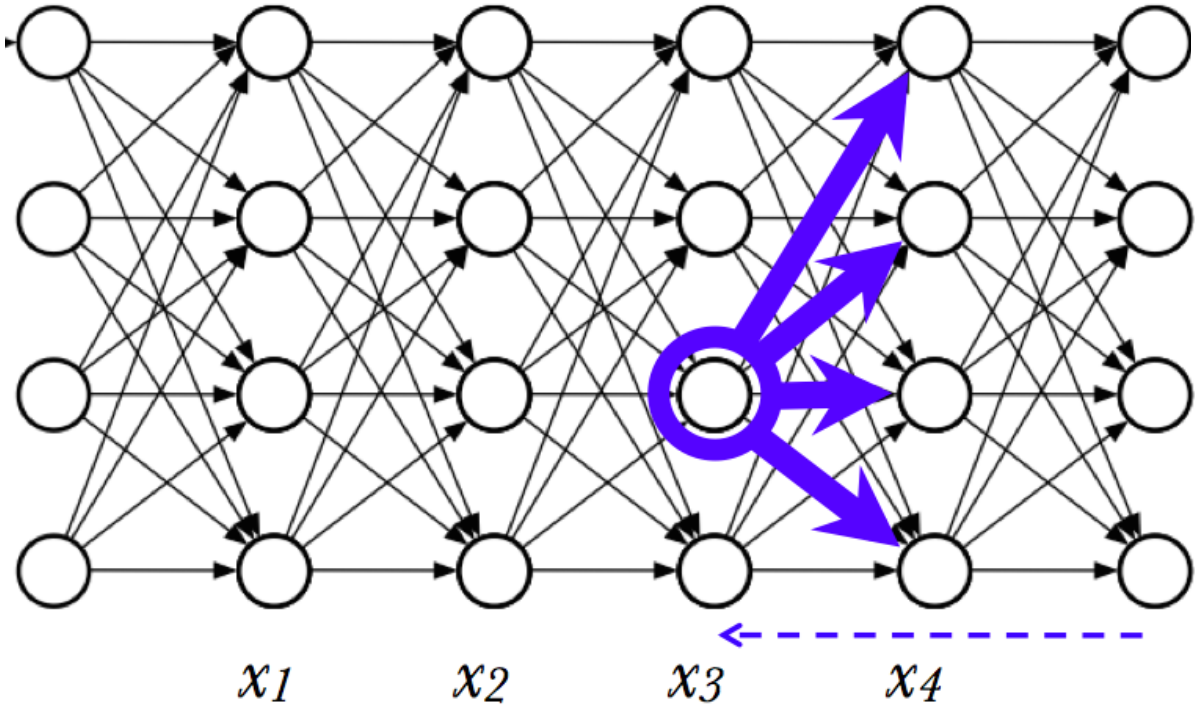
Forward algorithm



$$f_k(i) \triangleq P(x_1 \ldots x_i, \pi_i = k)$$

$$f_l(i+1) = e_l(x_{i+1}) \sum_k f_k(i) a_{k,l}$$

Backward algorithm



$$b_k(i) \triangleq P(x_{i+1} \cdots x_n \mid \pi_i = k)$$

$$b_k(i) = \sum_l a_{k,l}\, e_l(x_{i+1})\, b_l(i+1)$$

Posterior decoding

$$
\begin{aligned}
P(x, \pi_i = k) \\
&= P(x_1, \ldots, x_i, \pi_i = k) \cdot P(x_{i+1}, \ldots, x_n \mid x_1, \ldots, x_i, \pi_i = k) \\
&= P(x_1, \ldots, x_i, \pi_i = k) \cdot P(x_{i+1}, \ldots, x_n \mid \pi_i = k) \\
&= f_k(i) \cdot b_k(i)
\end{aligned}
$$

$$
P(\pi_i = k \mid x) = \frac{P(x, \pi_i = k)}{P(x)} = \frac{f_k(i) \cdot b_k(i)}{P(x)}
$$

Occasionally dishonest casino (posterior probabilities)

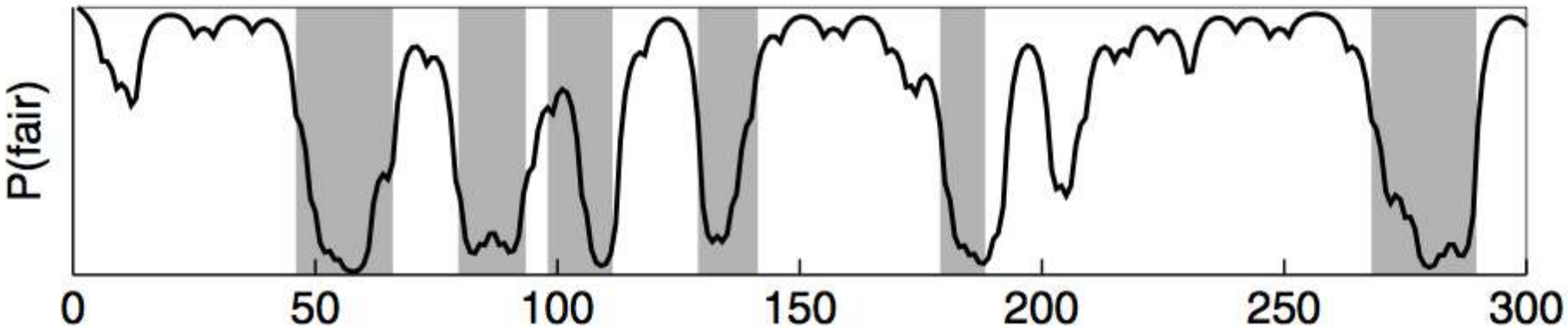


Figure 3.6 *The posterior probability of being in the state corresponding to the fair die in the casino example. The x axis shows the number of the roll. The shaded areas show when the roll was generated by the loaded die.*

HMM Training

Are hidden state paths for training sequences known?

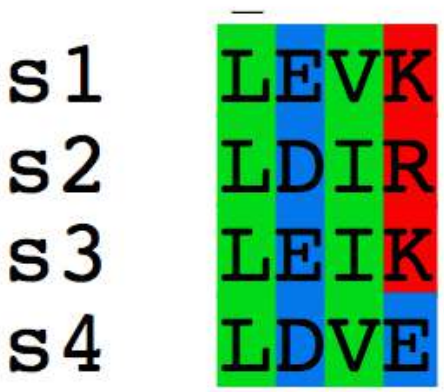
Yes: easy!

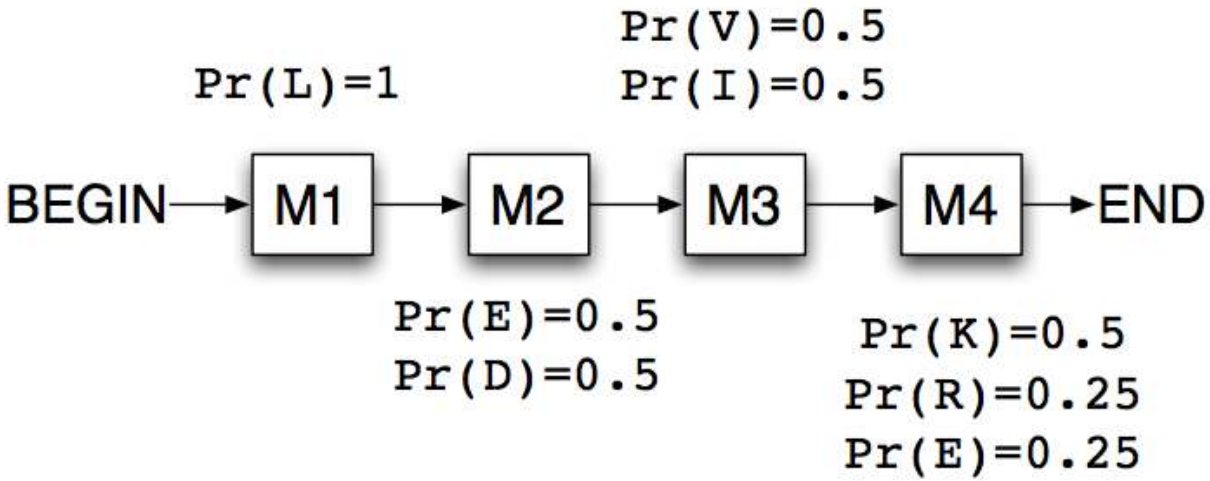
- count how often each transition and emission occurs
- normalize to get probabilities

No: iterative optimization...

Baum-Welch algorithm or Viterbi training

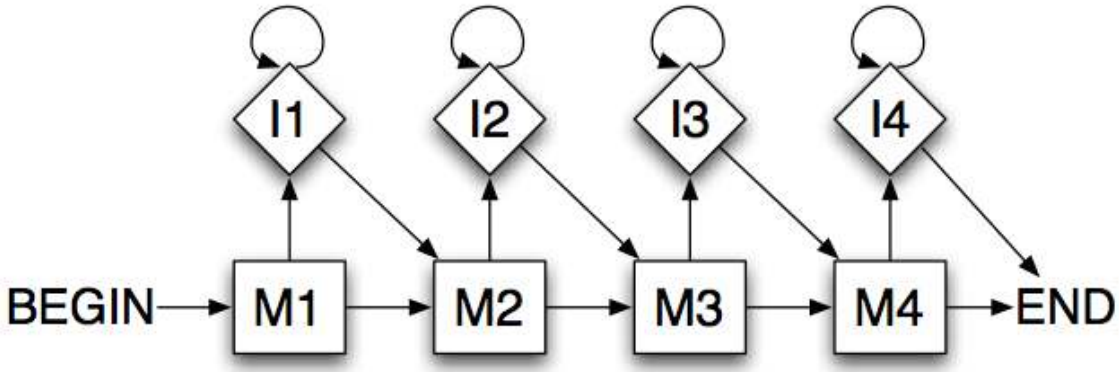
Profile HMM: Match states





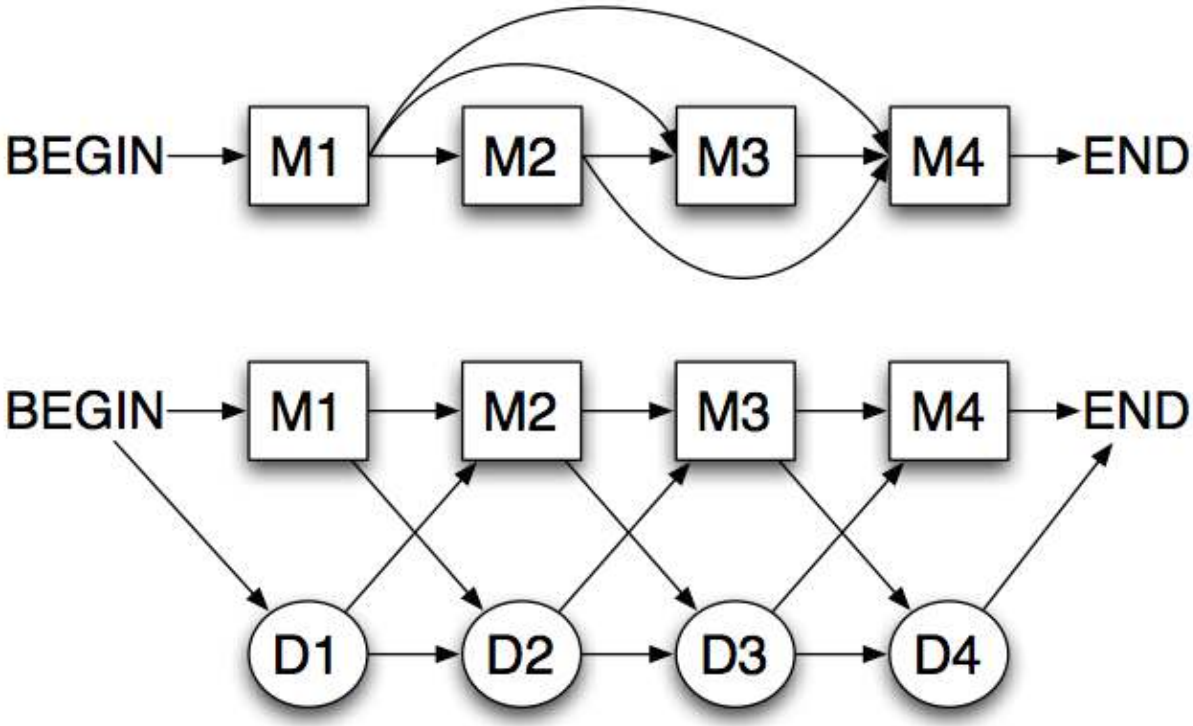
Profile HMM: Insertion states

```
s1      LE---VK
s2      LD---IR
s3      LE---IK
s4      LD---VE
query1  LDA--VK
query2  LDAAAVK
```

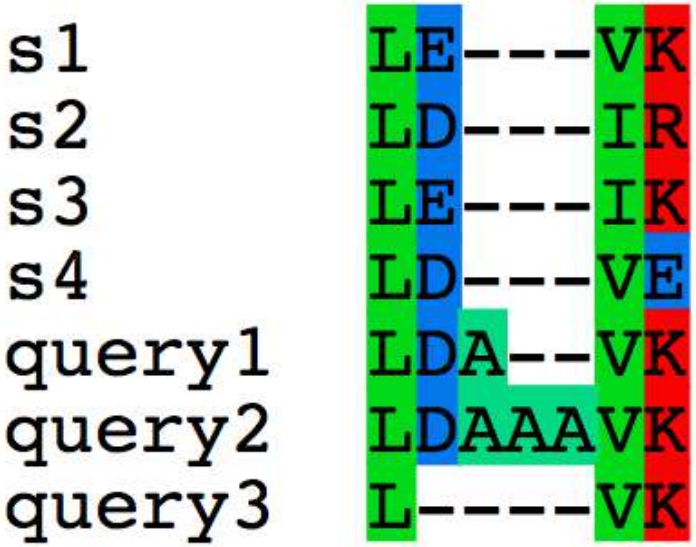


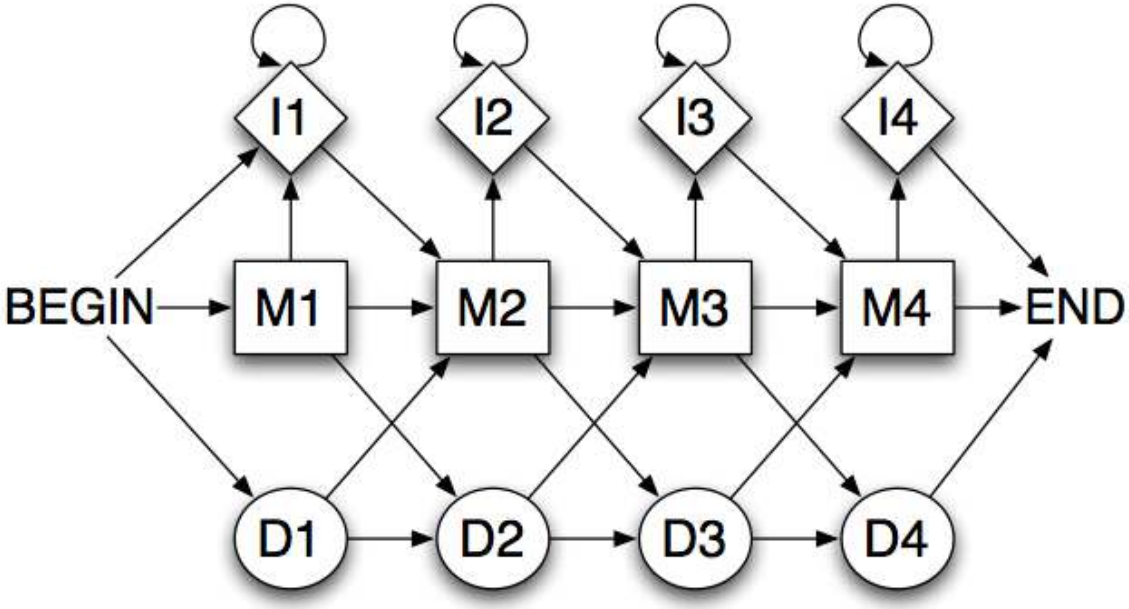
Profile HMM: Deletion states

```
s1       LEVK
s2       LDIR
s3       LEIK
s4       LDVE
query3   L-VK
```

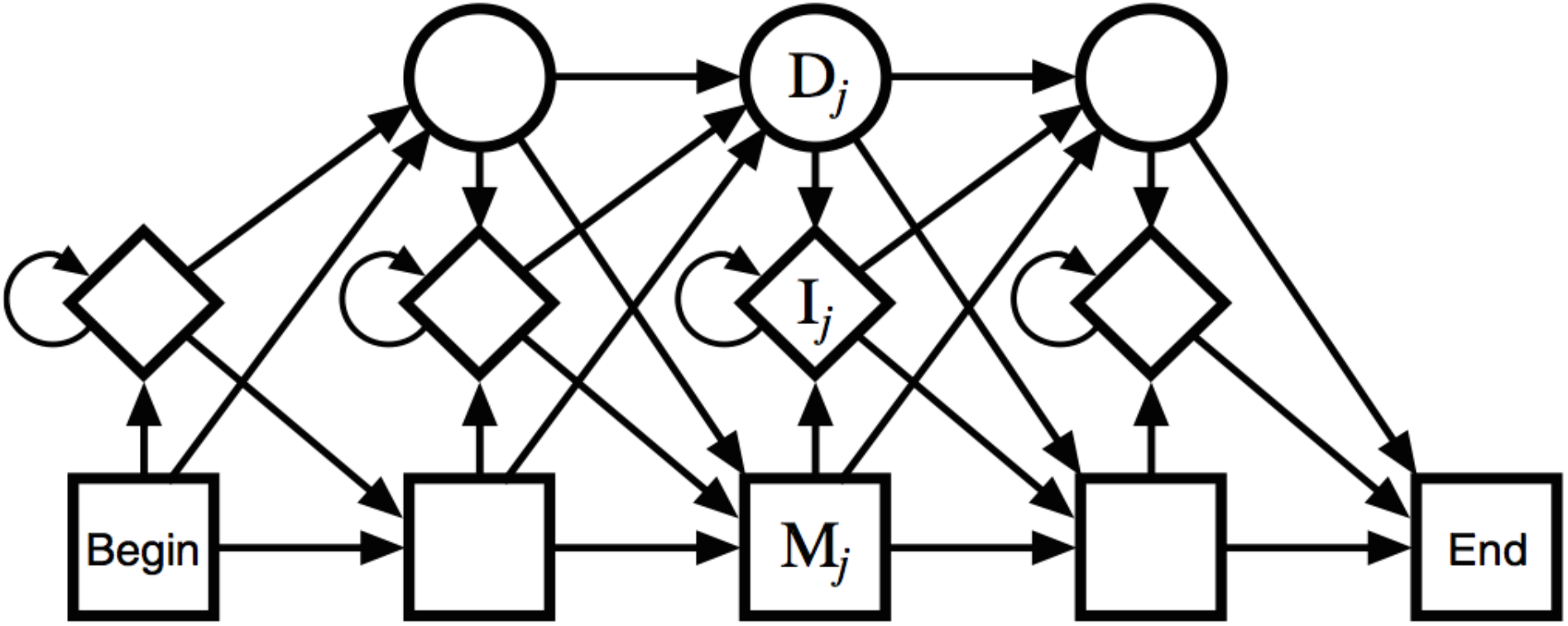


Profile HMM: Full model





Profile HMM: Full model



Profile HMM: Construction example

Usually start from multiple alignment

Need to chose columns that will correspond to match states

For each input sequence know its path through the model

```
s1  VGA--NAGRPY
s2  VG---NVDKPV
s3  VGA--NVAHPH
s4  VAA------PH
s5  VGS--TYEKPS
s6  FGA--NFEKPH
s7  IGAADNGARPY
```

Profile HMM notes

- For discrimination again consider
  log P(S|M)/P(S|B)
  - can modify forward algorithm to count it on the fly
- Numerical stability issues
- Can use Viterbi algorithm to find query alignment
- Probability estimation and sequence weighting strategies still apply + estimation of transition
- Easily adapted to local alignment (search)

HMMMER software

Building a model

hmmbuild From a multiple sequence alignment

Using a model

hmmalign Align sequences to an existing model (outputs a multiple alignment)

hmmconvert Convert a model into different formats

hmmcalibrate Empirically determines parameters for more sensitive searches

hmmemit Emit sequences probabilistically from a profile HMM

hmmsearch Search a sequence database for matches to an HMM

HMMs Databases

hmmfetch Get a single model from an HMM database

hmmindex: Index an HMM database (not available on the WEB server)

hmmpfam Search an HMM database for matches to a query sequence

Other programs

alistat: Show some simple statistics about a sequence alignment file

seqstat: Show some simple statistics about a sequence file

getseq: Retrieve a (sub-)sequence from a sequence file

sreformat: Reformat a sequence(s) or alignment file into a different format