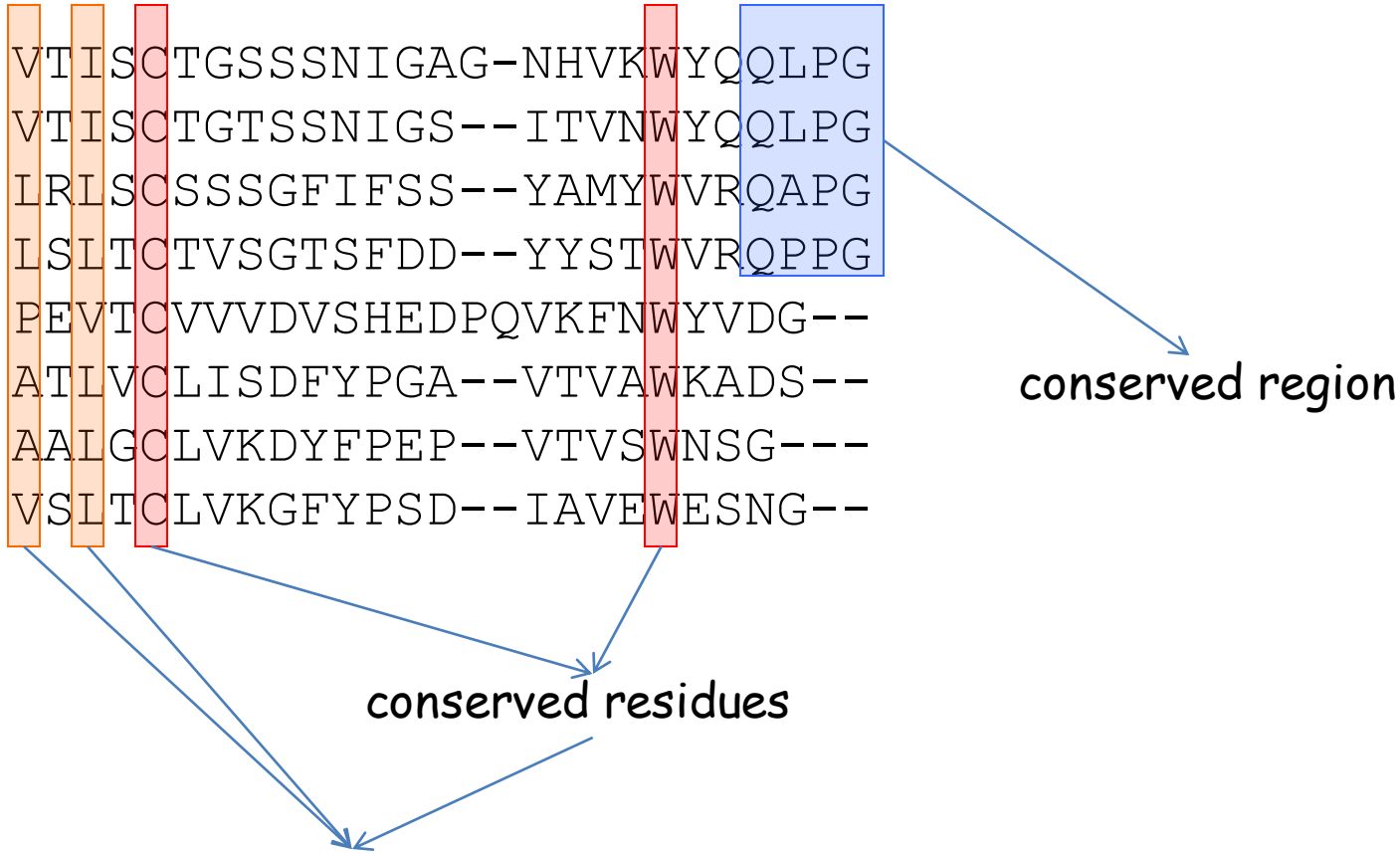
Sequence Profiles

by Ahmet Sacan

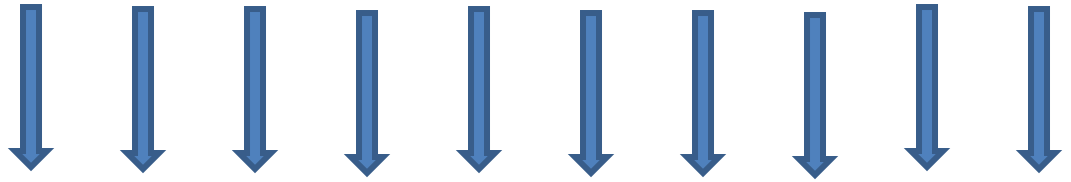
Images are from "Understanding Bioinformatics" and Wikipedia.

Multiple alignment



PSSM: Position-Specific-Scoring-Matrix

gene1:	**A**	**A**	**G**	**A**	**G**	**T**	**—**	**—**	**A**	**A**
gene2:	**A**	**A**	**G**	**A**	**C**	**—**	**—**	**—**	**T**	**A**
gene3:	**G**	**A**	**G**	**A**	**C**	**T**	**G**	**C**	**T**	**A**
gene4:	**G**	**A**	**—**	**A**	**C**	**C**	**G**	**C**	**A**	**A**
gene5:	**T**	**A**	**G**	**T**	**G**	**C**	**G**	**C**	**T**	**A**



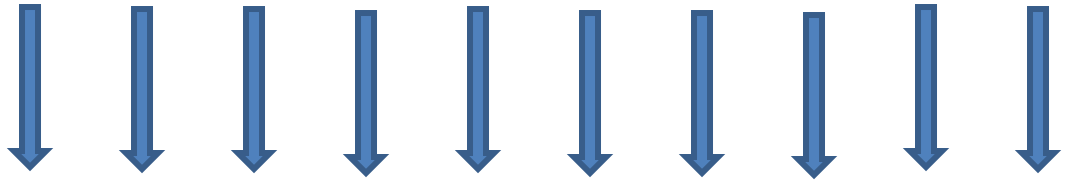
%A:	**40**	**100**	0	**80**	0	0	0	0	**40**	**100**
%C:	0	0	0	0	**60**	**40**	0	**60**	0	0
%G:	**40**	0	**80**	0	**40**	0	**60**	0	0	0
%T:	**20**	0	0	**20**	0	**40**	0	0	**60**	0

$$f_{u,a} = \frac{n_{u,a}}{N_{seq}}$$

- $n_{u,a}$: number of residues of type a at column u
- N_{seq}: number of sequences

PSSM: Position-Specific-Scoring-Matrix

protein1:	**A**	**R**	**S**	**N**	**C**	**P**	**—**	**—**	**A**	**A**
protein2:	**A**	**R**	**S**	**N**	**C**	**—**	**—**	**—**	**T**	**A**
protein3:	**L**	**R**	**C**	**N**	**C**	**P**	**G**	**C**	**T**	**A**
protein4:	**L**	**R**	**—**	**D**	**C**	**C**	**G**	**C**	**A**	**A**
protein5:	**I**	**R**	**C**	**D**	**G**	**C**	**G**	**C**	**T**	**A**



%A:	**40**	0	0	0	0	0	0	0	**40**	**100**
%R:	0	**100**	0	0	0	0	0	0	0	0
%N:	0	0	0	**60**	0	0	0	0	0	0
%D:	0	0	0	**40**	0	0	0	0	0	0
%C:	0	0	**40**	0	**80**	**40**	0	**60**	0	0
…	..	…	…	…	…	…	…	…	…	…

Pseudo-counts

protein1:	A	R	S	N	C	P	–	–	A	A
protein2:	A	R	S	N	C	–	–	–	T	A
protein3:	L	R	C	N	C	P	G	C	T	A
protein4:	L	R	–	D	C	C	G	C	A	A
protein5:	I	R	C	D	G	C	G	C	T	A
fakeprotein A:	A	A	A	A	A	A	A	A	A	A
fakeprotein R:	R	R	R	R	R	R	R	R	R	R
fakeprotein N:	N	N	N	N	N	N	N	N	N	N
…	…	…	…	…	…	…	…	…	…	…

Correct for lack of or bias in data

- Use pseudocounts:

$$f'_{u,a} = \frac{n_{u,a} + 1}{N_{seq} + 20}$$

- Use pseudocounts by background frequencies

$$f'_{u,a} = \frac{n_{u,a} + \beta p_a}{N_{seq} + \beta}$$

  - Lower the contribution of pseudocounts or substitution matrix if there is enough data.

- Use a substitution matrix

$$f'_{u,a} = \sum_b f_{u,b} s_{a,b}$$

- Weight the sequence contributions
  - Lower the weights of highly similar sequences

Representing profiles using logos

- Entropy (uncertainty) in a column c:

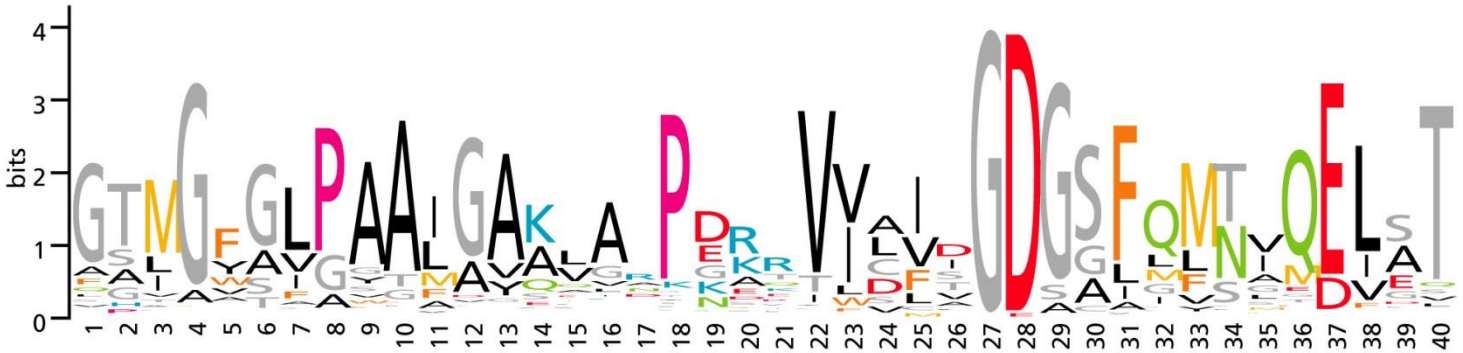
$$H_c = -\sum_a f_{c,a} \log_2(f_{c,a})$$

- Information

$$I_c = \log_2 20 - H_c$$

- Contribution of a residue a

$$I_{c,a} = f_{c,a} I_u$$



weblogo.berkeley.edu

Sequence Logo Example

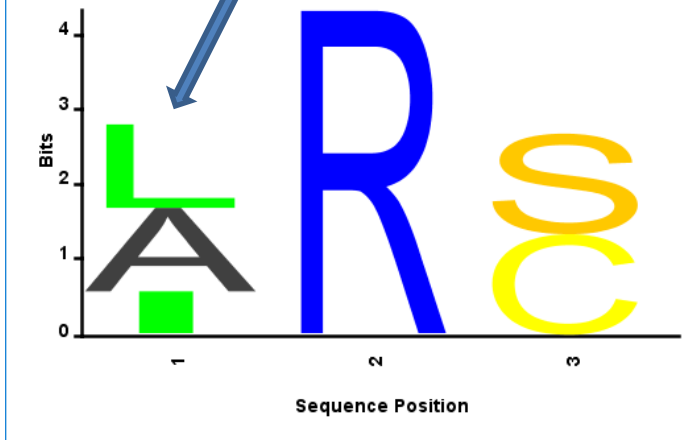
protein1:	A	R	S	N	C	P	–	–	A	A
protein2:	A	R	S	N	C	–	–	–	T	A
protein3:	L	R	C	N	C	P	G	C	T	A
protein4:	L	R	–	D	C	C	G	C	A	A
protein5:	I	R	C	D	G	C	G	C	T	A

20

%A:	40	0	0	0	0	0	0	0	40	100
%R:	0	100	0	0	0	0	0	0	0	0
%N:	0	0	0	60	0	0	0	0	0	0
%D:	0	0	0	40	0	0	0	0	0	0
%C:	0	0	40	0	80	40	0	60	0	0
...	..	...	...	...	...	...	...	...	...	...

$$I_{1,A} = 0.4 * 2.8 = 1.12$$
$$I_{1,L} = 0.4 * 2.8 = 1.12$$
$$I_{1,I} = 0.2 * 2.8 = 0.56$$

$$H_1 = -(\ f_A * \log_2 f_A + f_L * \log_2 f_L + f_I * \log_2 f_I\ )$$
$$= -(\ 0.4 * \log_2 0.4 + 0.4 * \log_2 0.4 + 0.2 * \log_2 0.2\ )$$
$$= 1.52$$
$$I_1 = \log_2 20 - H_1 = \mathbf{2.8}$$

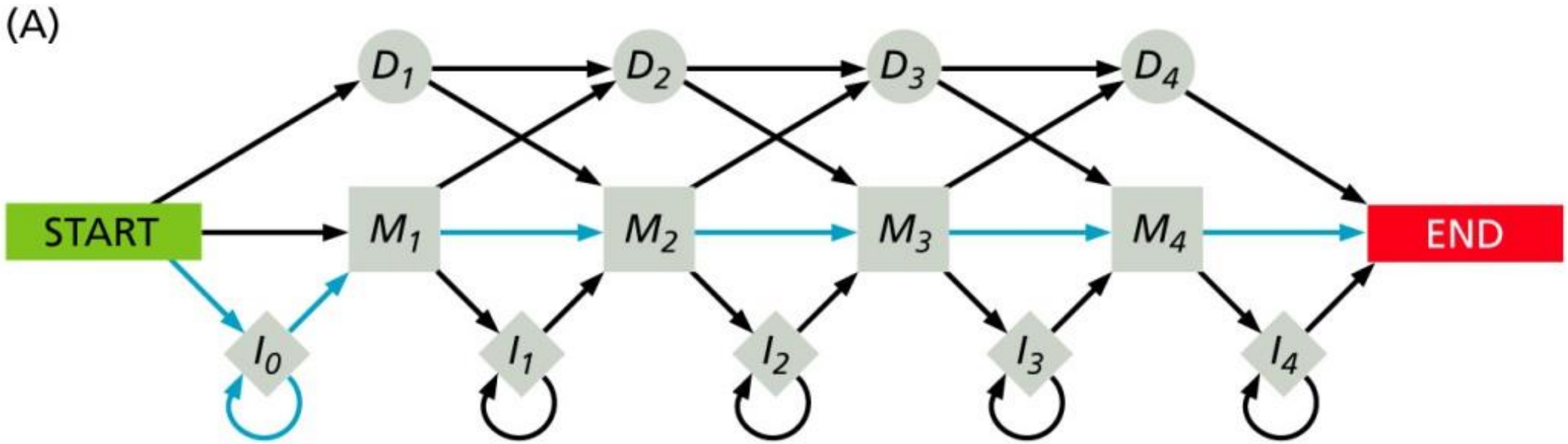


Prosite patterns

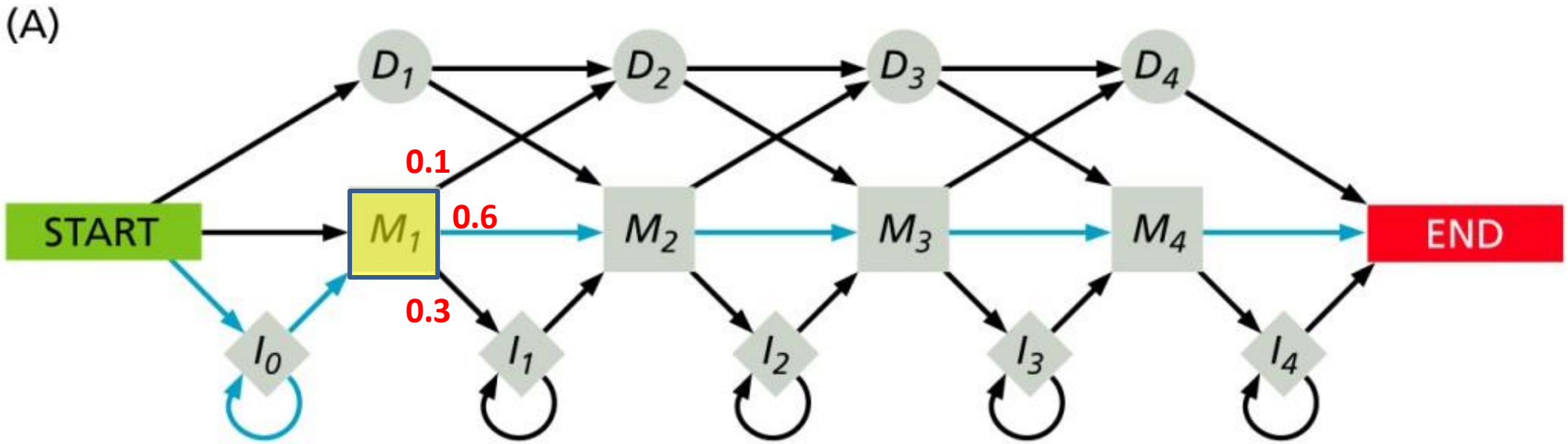
- [AC]-x-V-x(4)-{ED}
  - [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
- <A-x-[ST](2)-x(0,1)-V
  - Nterminal Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

Prosite	Regular Expression	
x	**.**	any character
[ALT]	**[ALT]**	any of A, L, or T
{AM}	**[^AM]**	anything but A or M
A(3)	**A{3}**	AAA
A(2,4)	**A{2,4}**	AA, or AAA, or AAAA
<A	**^A**	A at the N-terminus
A>	**A$**	A at the C-terminus

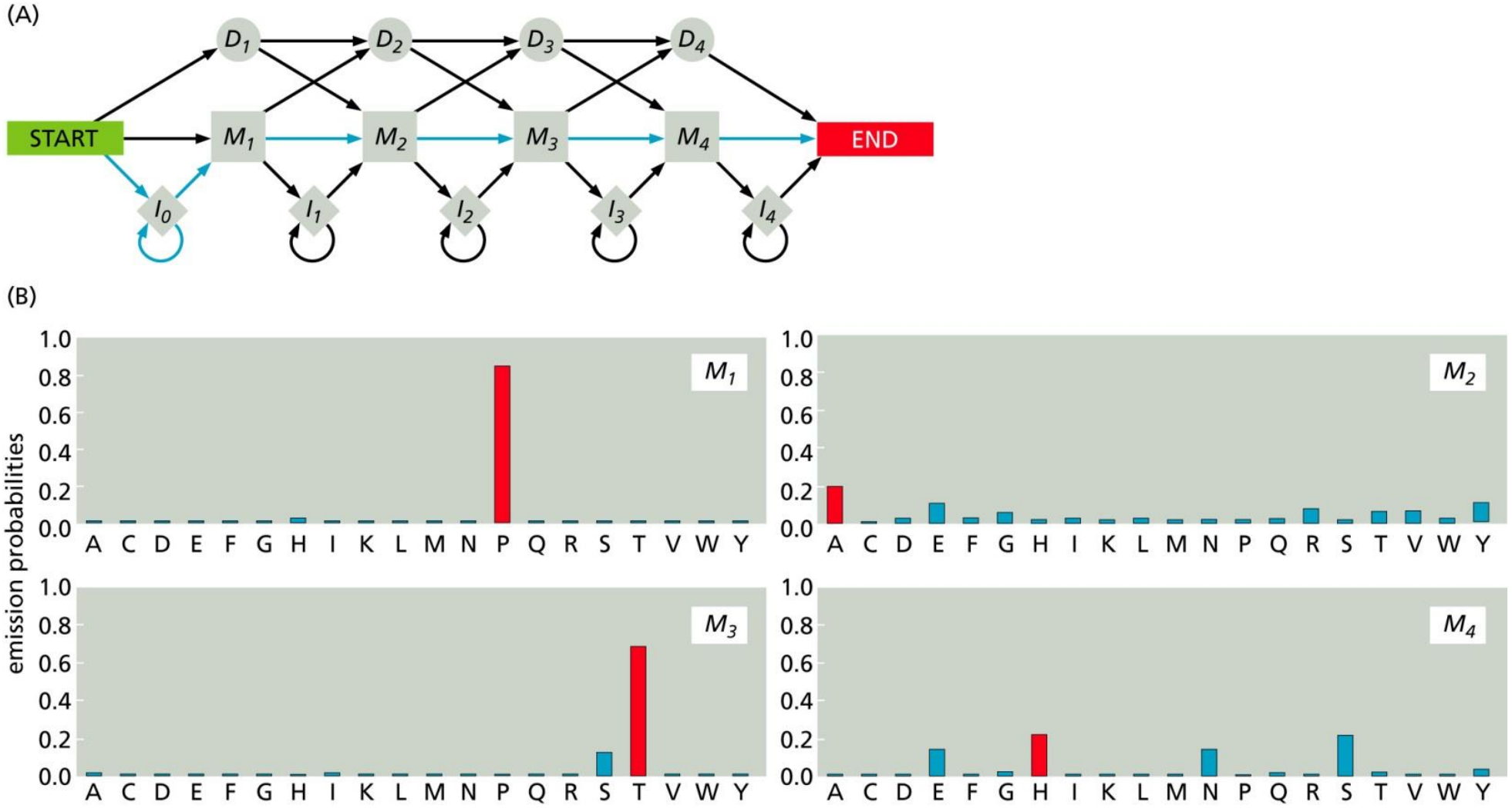
Hidden Markov Models (HMM)



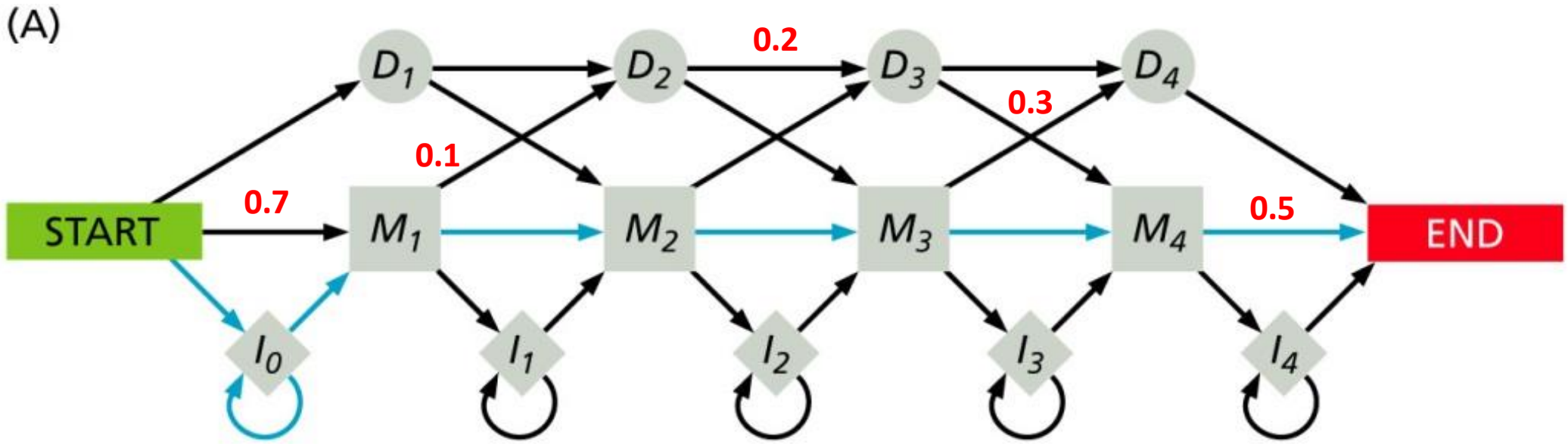
Hidden Markov Models (HMM)



Hidden Markov Models (HMM)

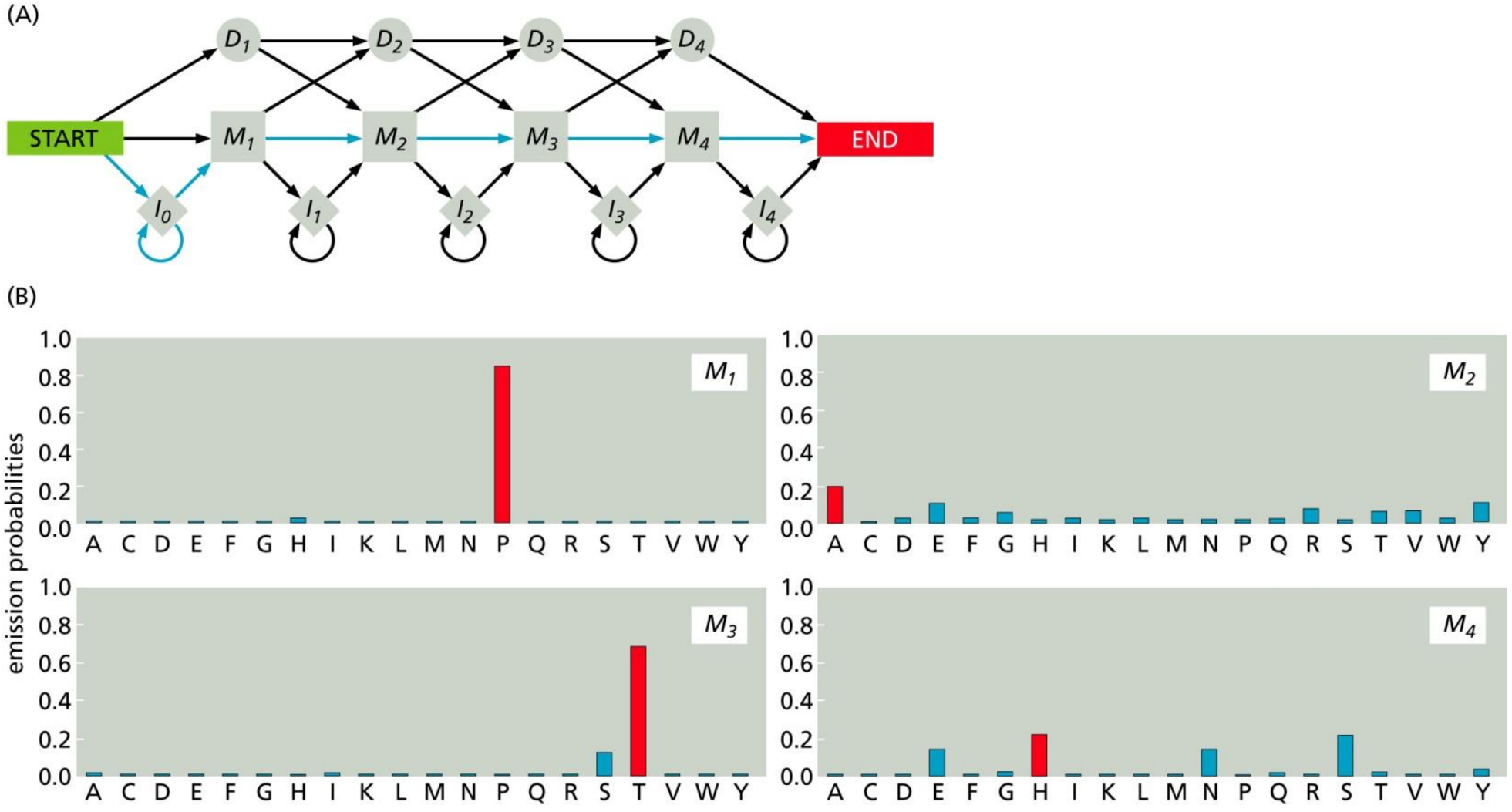


Hidden Markov Models (HMM)



- $M_{1,P} = 0.9$, $M_{4,N} = 0.2$
- Path probability = 0.7*0.9*0.1*0.2*0.3*0.2*0.5 = 3.8e-04
- Sequence produced: **PN**

Hidden Markov Models (HMM)



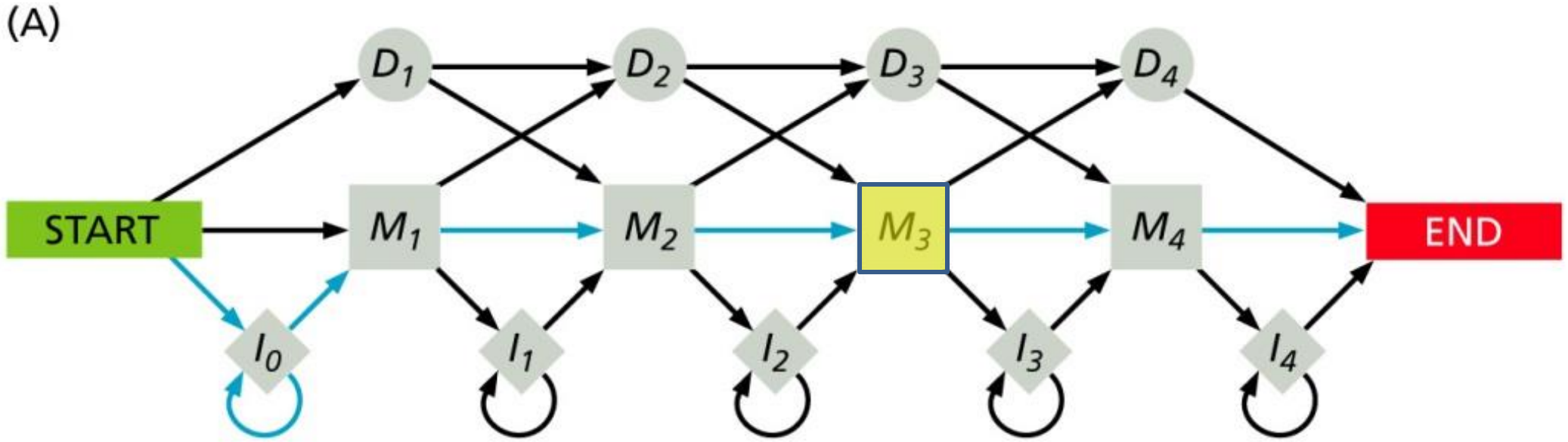
HMM questions

- What is the most likely path?
- What is the probability of a sequenced being produced?
- How do we construct the HMM and identify its parameters?

HMM answers

- Dynamic programming: Probability of a node can be decomposed into probabilities of transitioning into it from previous states.
- The most likely path
  - Viterbi algorithm
  - max() of each previous path

Viterbi Algorithm



H_{M_3}, most likely path into M3:

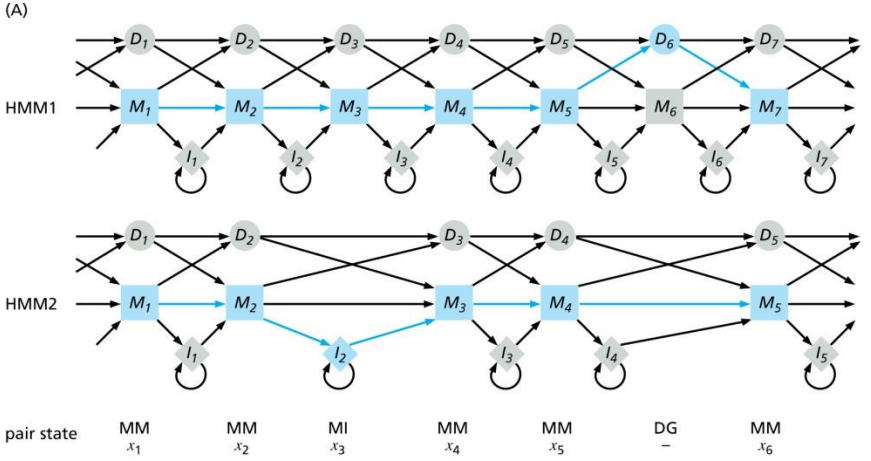
$$max \begin{cases} H_{D_2} * T_{D_2 \to M_3} \\ H_{M_2} * T_{M_2 \to M_3} \\ H_{I_2} * T_{I_2 \to M_3} \end{cases}$$

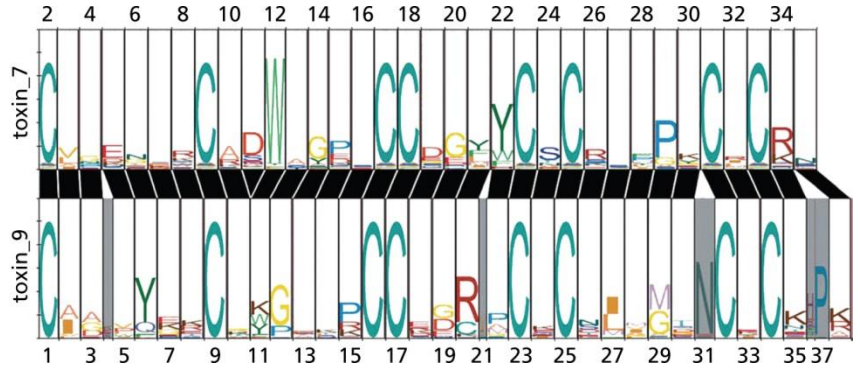
HMM answers

- The probability of a sequence to be emitted
  - Forward (or Backward) algorithm
  - sum() of previous paths
- HMM parameters can be estimated from unaligned sequences
  - Baum-Welch expectation maximization algorithm

Aligning families

- Two HMMs can be aligned
  - COACH, HHSEARCH programs



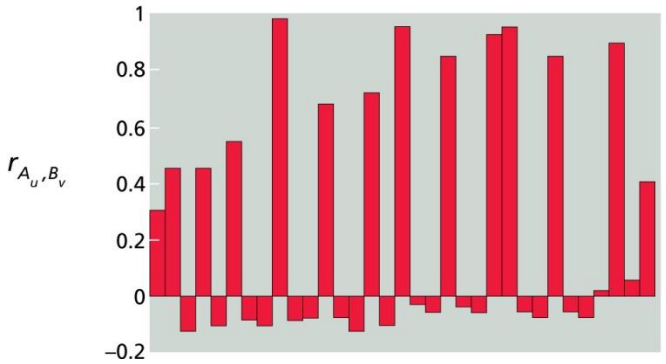


Aligning families

- Two PSSMs can be compared using Pearson correlation coefficient
  - LAMA program

```
OXDA_FUSSO   319   LDDETWIVHNYGHSGWGYQGSYGCAENVVQLVD   351
OXDD_BOVIN   294   DSRRLPVVHHYGHGSGGIAMHWGTALEATRLVN   326
OXDA_HUMAN   299   GPSNTEVIHNYGHGGYGLTIHWGCALEAAKLFG   331
OXDA_MOUSE   297   GSSSAEVIHNYGHGGYGLTIHWGCAMEAANLFG   329
OXDA_PIG     299   GSSNTEVIHNYGHGGYGLTIHWGCALEVAKLFG   331
OXDA_RABIT   299   GPSKTEVIHNYGHGGYGLTIHWGCALEAAKLFG   331
```

```
DHSA_BACSU   229   GEFIQIHPTAIPGDDKLRLMSESARGEGGRVWT   261
DHSA_ECOLI   234   QDMEMWQFHPTGIAGAGVLVTEGCRGEGGYLLN   266
FRDA_WOLSU   249   GNMEAVQFHPTPLFPSGILLTEGCRGDGGILRR   281
DHSA_BOVIN   289   QDLEFVQFHPTGIYGAGCLITEGCRGEGGILIN   321
DHSA_RICPR   238   QDMEFVQFHPTGIYGAGCLITEGARGEGGYLVN   270
DHSA_YEAST   279   QDLEFVQFHPSGIYGSGCLITEGARGEGGFLVN   311
FRDA_ECOLI   224   RDMEFVQYHPTGLPGSGILMTEGCRGEGGILVN   256
FRDA_PROVU   225   RDMEFVQYHPTGLPGSGILMTEGCRGEGGILVN   257
```



Summary

- Sequence profiles obtained from multiple sequence alignments
- Sequence logos and Prosite patterns are easy to interpret
- HMM more accurate for evaluating family membership