Sequence profiles, Hidden Markov models and homology modeling

Morten Nielsen,
CBS, Department of Healt Technology, DTU
and
Instituto de Investigaciones Biotecnológicas,
Universidad de San Martín, Argentina

Identification of essential residues in protein sequences


```
TVNGQ--FPGPRLAGVAREGDQVLVKVVNHVAENITIHWHGVQLGTGWADPPAYVTQCPI

TKAVVLTFNTSVEICLVMQGTSIV----AAESHPLHLHGFNFPSNFNLVDGMERNTAGVP
```

Summary

- Sequence logo is a power tool to visualize (binding) motifs
  - Information content identifies essential residues for function and/or structural stability
- Weight matrices can be derived from very limited number of data using the techniques of
  - Sequence weighting
  - Pseudo counts

Sequence Profiles and Weight matrices

- Alignments based on conventional scoring matrices (BLOSUM62) scores all positions in a sequence in an equal manner
- Some positions are highly conserved, some are highly variable (more than what is described in the BLOSUM matrix)
- Sequence profile are ideal suited to describe such position specific variations

Sequence alignment

- Conventional sequence alignment uses a (Blosum) scoring matrix to identify amino acids matches in the two protein sequences

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

Alignment scoring matrices

- Blosum62 score matrix. Fg=1. Ng=0?

	L	A	G	D	S	D
F						
I						
G						
D						
S						
L						

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

Alignment scoring matrices

- Blosum62 score matrix. Fg=1. Ng=0?

	L	A	G	D	S	D
F	0	-2	-3	-3	-2	-3
I	2	-1	-4	-3	-2	-3
G	-4	0	6	-1	0	-1
D	-4	-2	-1	6	0	6
S	-2	1	0	0	4	0
L	4	-1	-4	-4	-2	-4

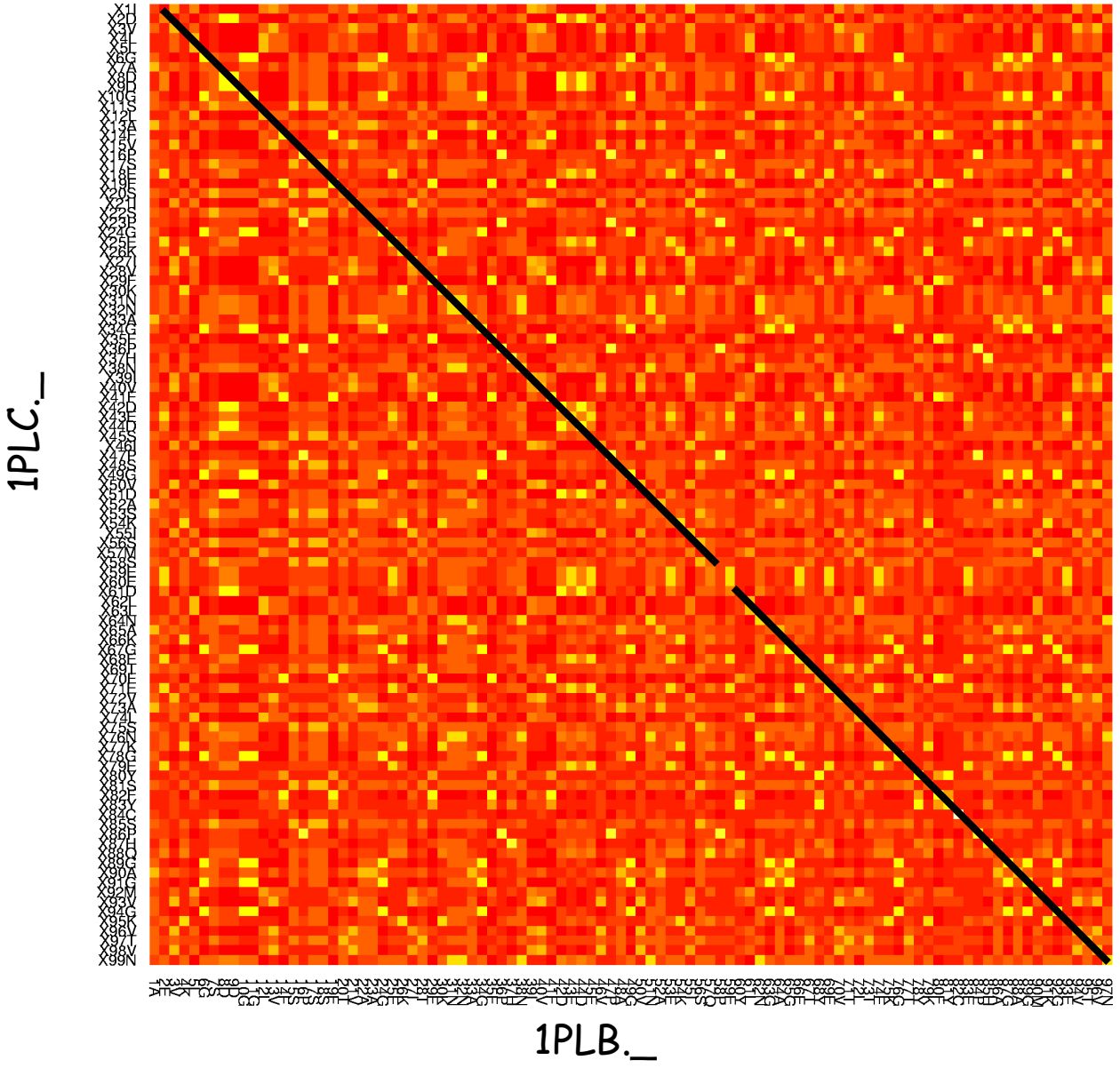
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2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

- Score =2-1+6+6+4=17

LAGDS

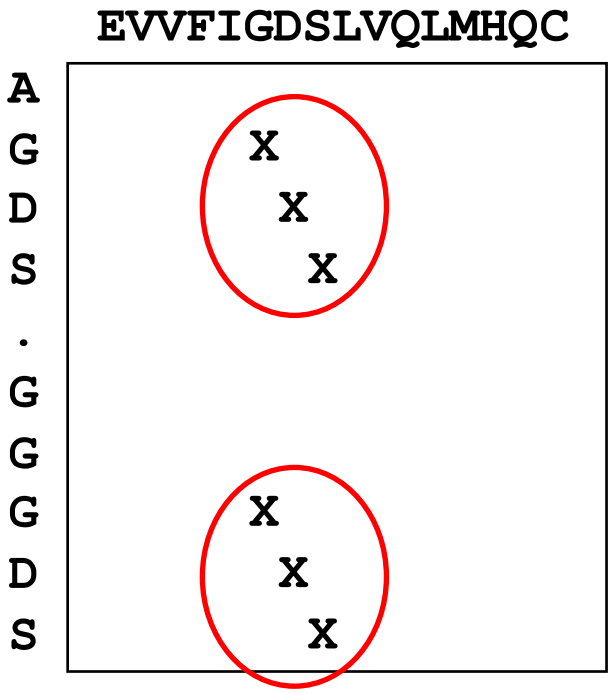
I-GDS

When Blast works!

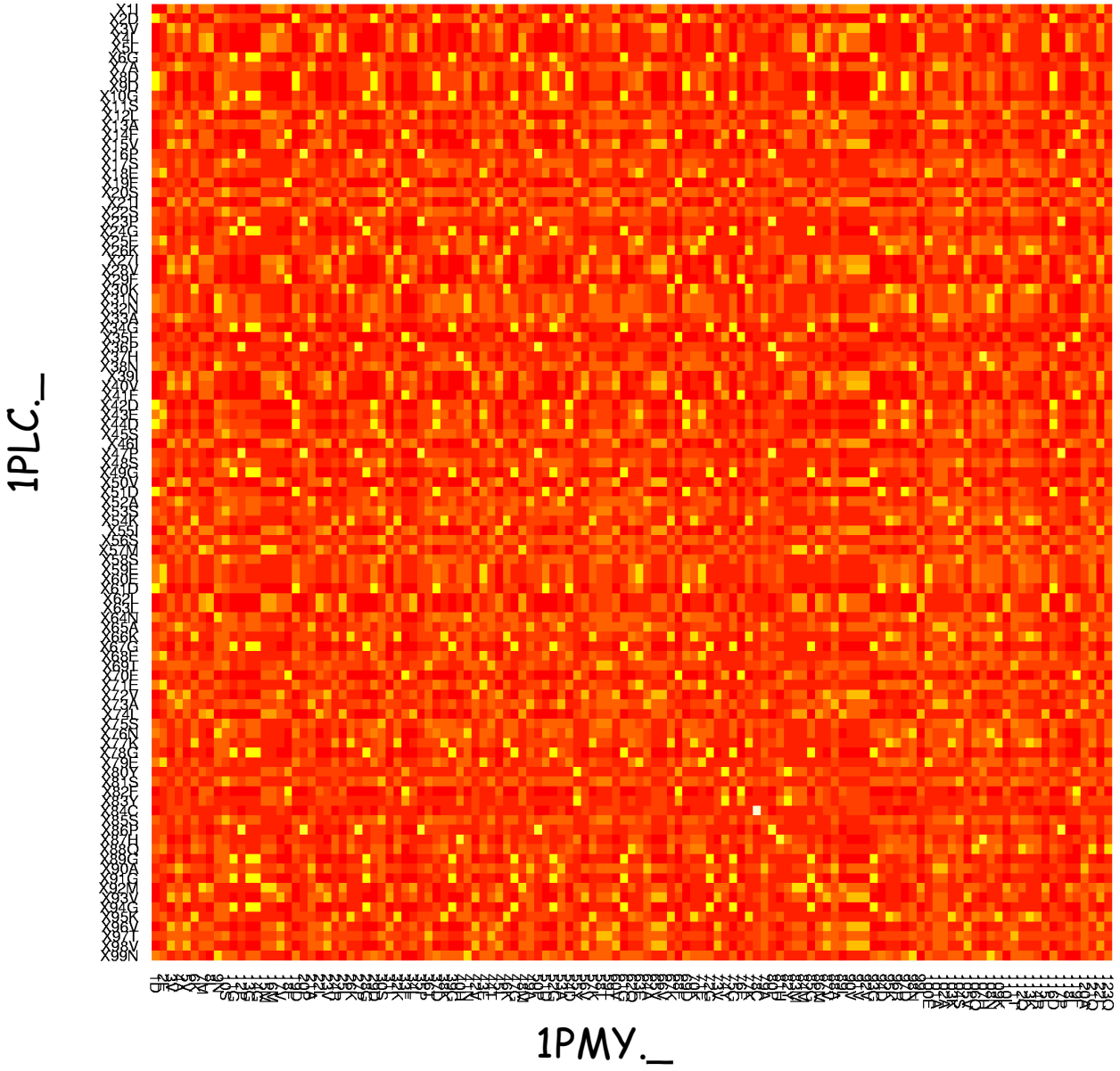


What goes wrong when Blast fails?

- Conventional sequence alignment uses a (Blosum) scoring matrix to identify amino acids matches in the two protein sequences
- This scoring matrix is identical at all positions in the protein sequence!

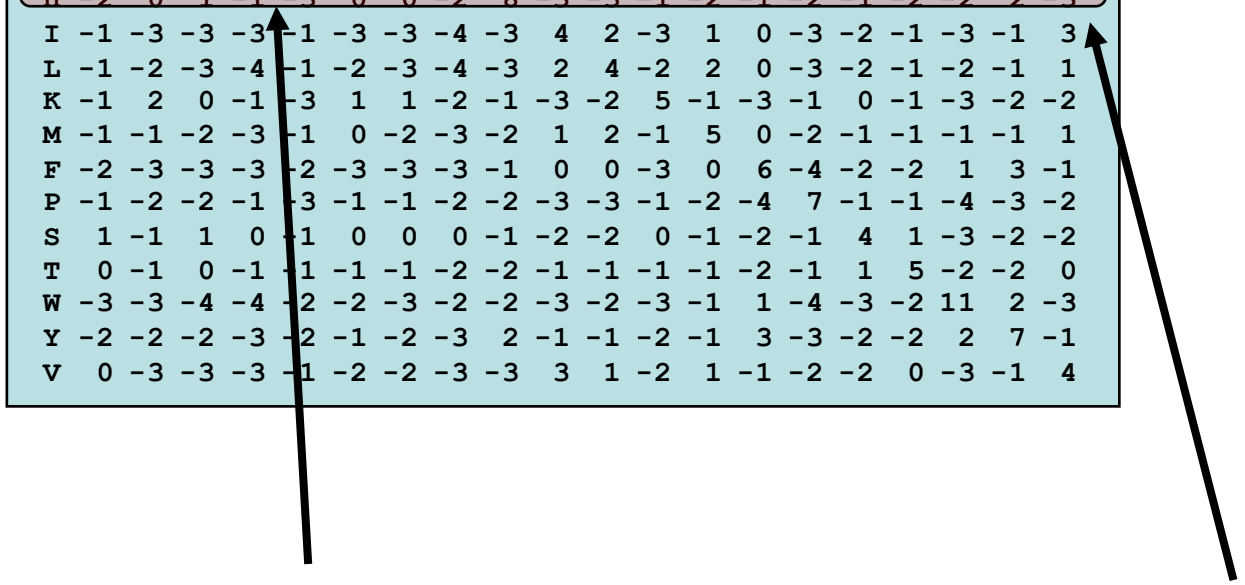


When Blast fails!



Alignment match scores

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

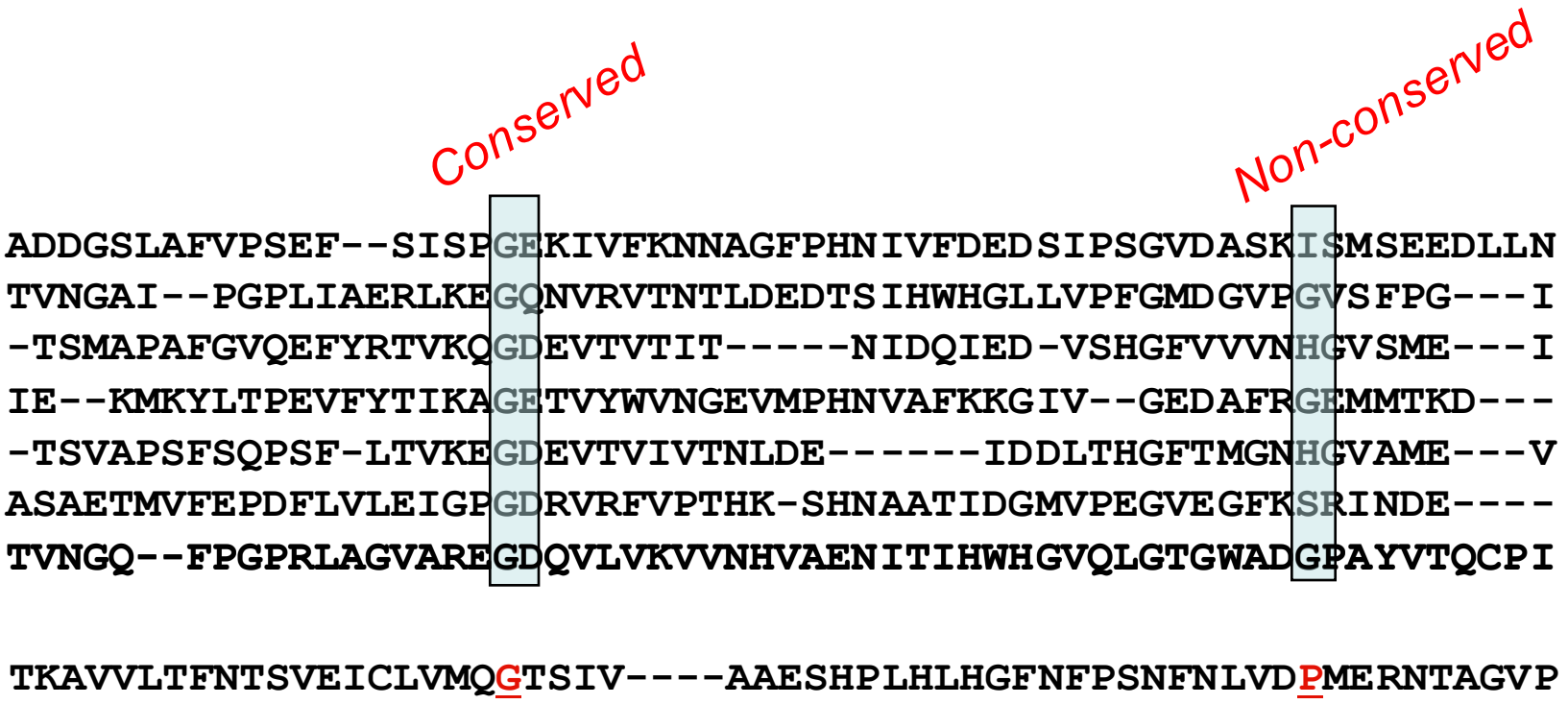


TVNGQ--FPGPRLAGVAREGDQVLVKVVNHVAENITIHWHGVQLGTGWADGPAYVTQCPI

Sequence profiles

- In reality not all positions in a protein are equally likely to mutate
  - Some amino acids (active cites) are highly conserved, and the score for mismatch must be very high
  - Other amino acids can mutate almost for free, and the score for mismatch should be lower than the BLOSUM score
- Sequence profiles can capture these differences

Sequence profiles



Matching any thing
but G => large
negative score

Any thing can match

How to make sequence profiles

1. Align (BLAST) sequence against large sequence database (Swiss-Prot)
2. Select significant alignments and make sequence profile
3. Use profile to align against sequence database to find new significant hits
4. Repeat 2 and 3 (normally 3 times!)

Sequence logos.
Visualization of sequence profiles

$$I = \sum_a p_a \log(\frac{p_a}{q_a})$$

$P_A = 6/10 = 0.6$
$P_G = 2/10 = 0.2$
$P_T = P_K = 1/10 = 0.1$
$P_C = P_D = \ldots P_V = 0.0$

$q_A = 0.07$
$q_G = 0.07$
$q_T = 0.05$
$q_K = 0.06$

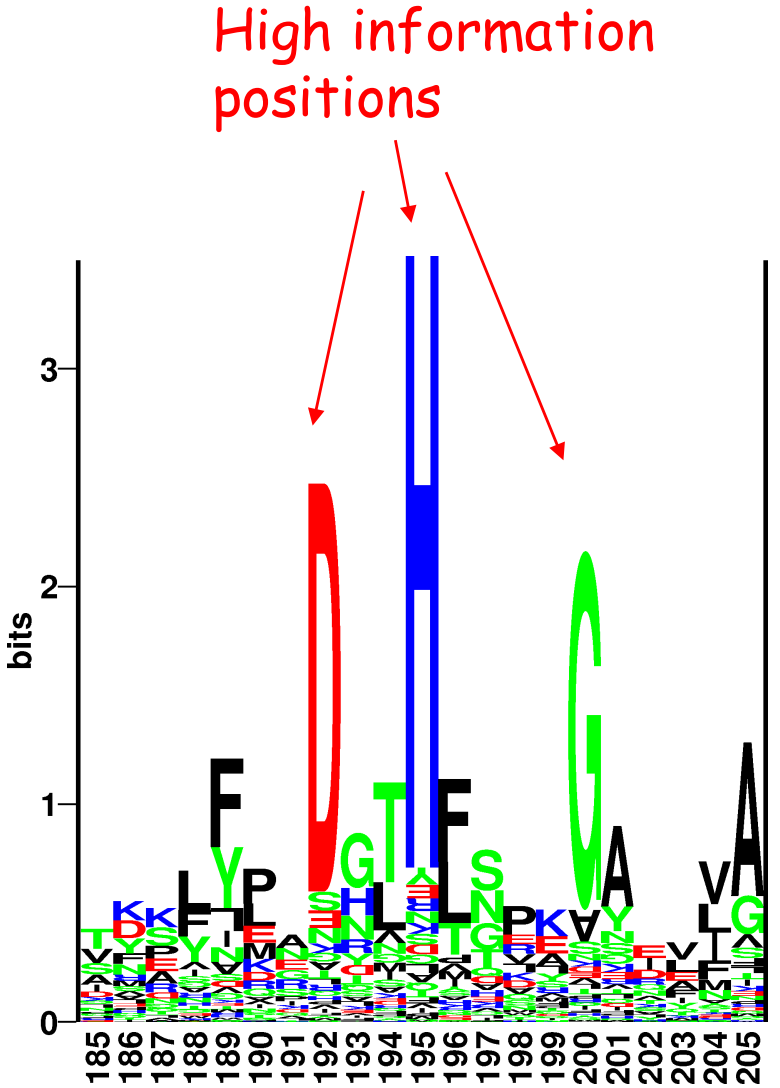
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Sequence logos

$$I = \sum_a p_a \log(\frac{p_a}{q_a})$$

- Height of a column equal to I
- Relative height of a letter is p

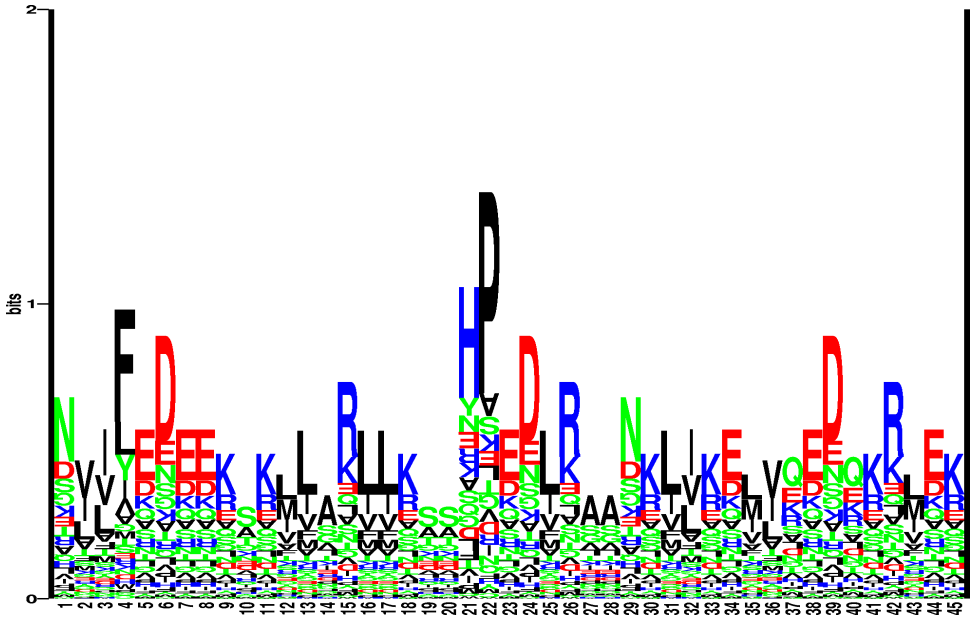
(letters are upside down if q>p)



Sequence profiles (1J2J.B)

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

IFEDEEKSKMLARLLKSSHPEDLRAANKLIKELVQEDQKRLEK



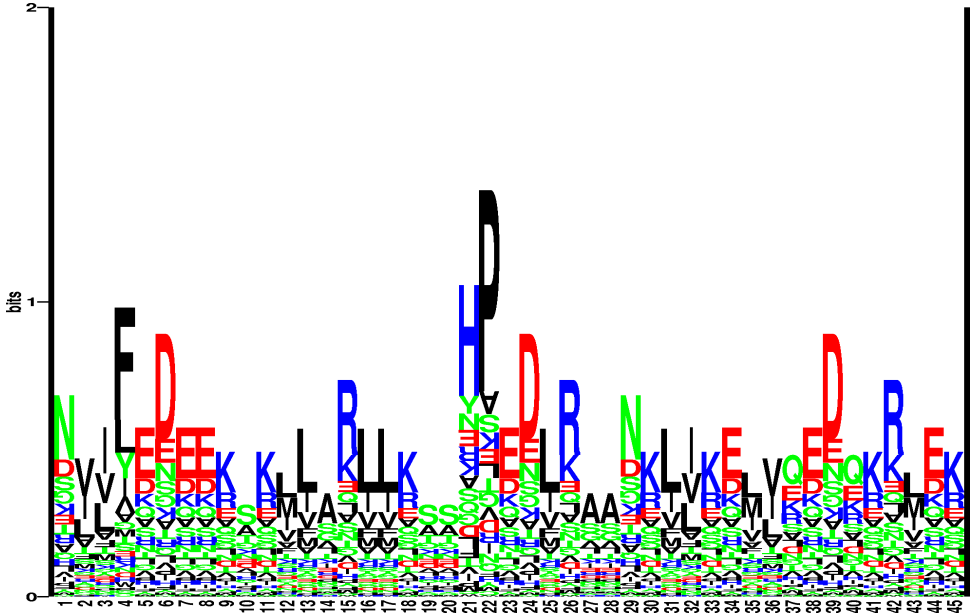
Blosum62

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

$$W_{ij} = \log(p_{ij}/q_j)$$

Sequence profiles (1J2J.B)

IFEDEEKSKMLARLLKSSHPEDLRAANKLIKELVQEDQKRLEK

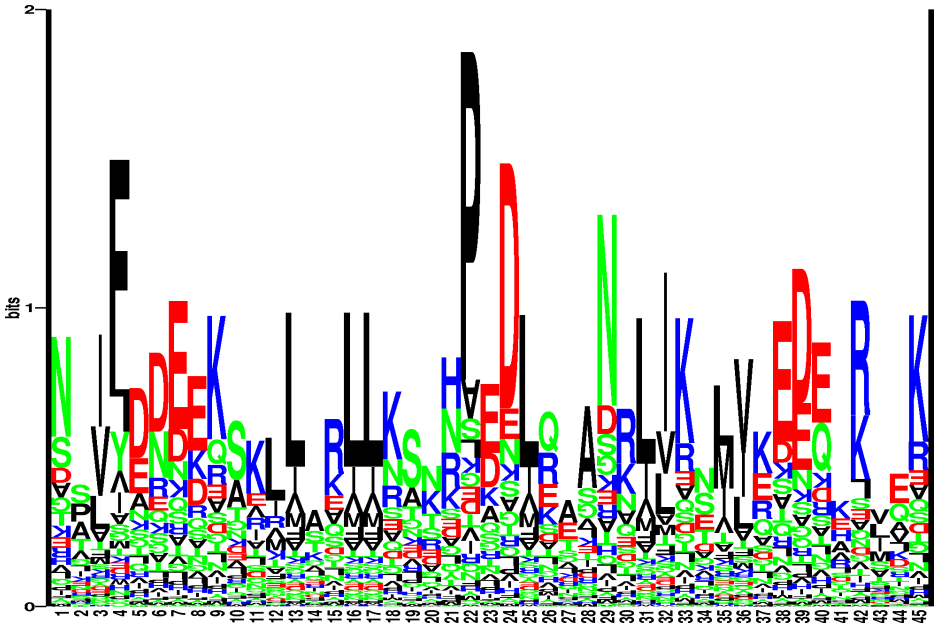


		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
1	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
2	E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
3	D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
4	E	-1	0	0	2	-4	2	5	-2	0	-3	-3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

Sequence profiles (1J2J.B)

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

IFEDEEKSKMLARLLKSSHPEDLRAANKLIKELVQEDQKRLEK



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

Blast2logo

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Blast2logo 1.0 Server

Instructions | Output format

SUBMISSION

Paste a single sequence in FASTA format into the field below:

```
>Ex
VALAELYIPEVARRLGQGWHEDECTFAEVTIGTARLQAILRDIATSWSADEGGMRDGPAVLVLLPPG
EQHTLGAMVAVAKLRRLGVSVCLRMSTGPAELRELFGKRRFDAIMISLAHAEMLEVGRKLVKTLKD
MTGGRIPVAMGGALFLDGTEAASIPEADIVTNDIEAALQ
```

Submit a file in FASTA format directly from your local disk:

Choose File no file selected

Upload a file in BLAST PROFILE format:

Choose File no file selected

Blast Database SP

Number of Blast iterations 1

Blast E-value cutoff 0.00001

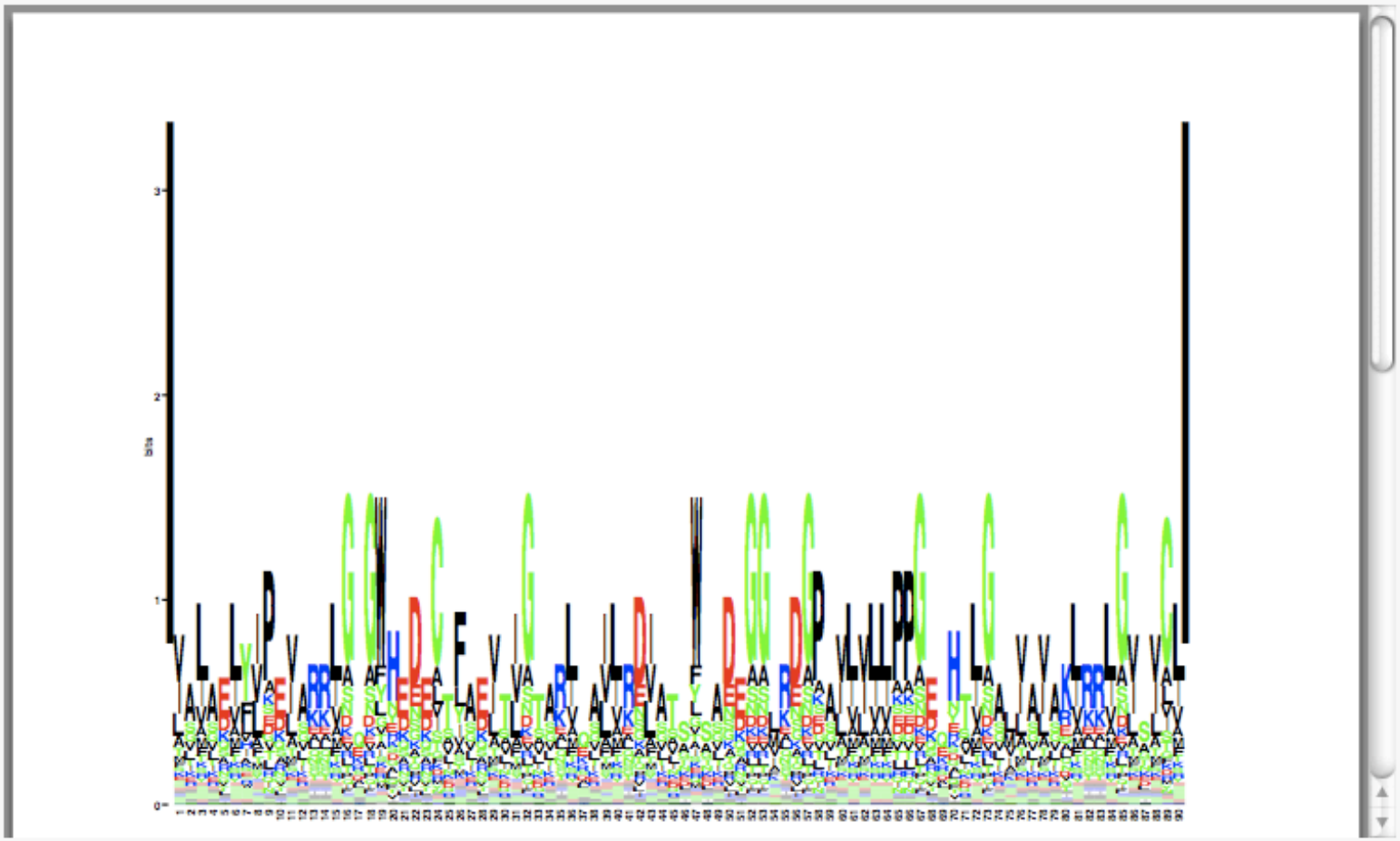
Stack Linesize 90

Plot Kullback-Leibler logo ☐

File format for logo file PDF

Submit Clear fields

Blast2logo



Download logo file Logo

Link to Blastprofile output file Blast.prof

Blast2logo

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Last position-specific scoring matrix computed

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

.
.

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

Blast2logo

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Blast2logo 1.0 Server

Instructions | Output format

SUBMISSION

Paste a single sequence in FASTA format into the field below:

```
>Ex
VALAELYIPEVARRLGQGWHEDECTFAEVTIGTARLQAILRDIATSWSADEGGMRDGPAVLVLLPPG
EQHTLGAMVAVAKLRRLGVSVCLRMSTGPAELRELFGKRRFDAIMISLAHAEMLEVGRKLVKTLKD
MTGGRIPVAMGGALFLDGTEAASIPEADIVTNDIEAALQ
```

Submit a file in FASTA format directly from your local disk:

Choose File no file selected

Upload a file in BLAST PROFILE format:

Choose File no file selected

Blast Database NR70

Number of Blast iterations 3

Blast E-value cutoff 0.00001

Stack Linesize 90

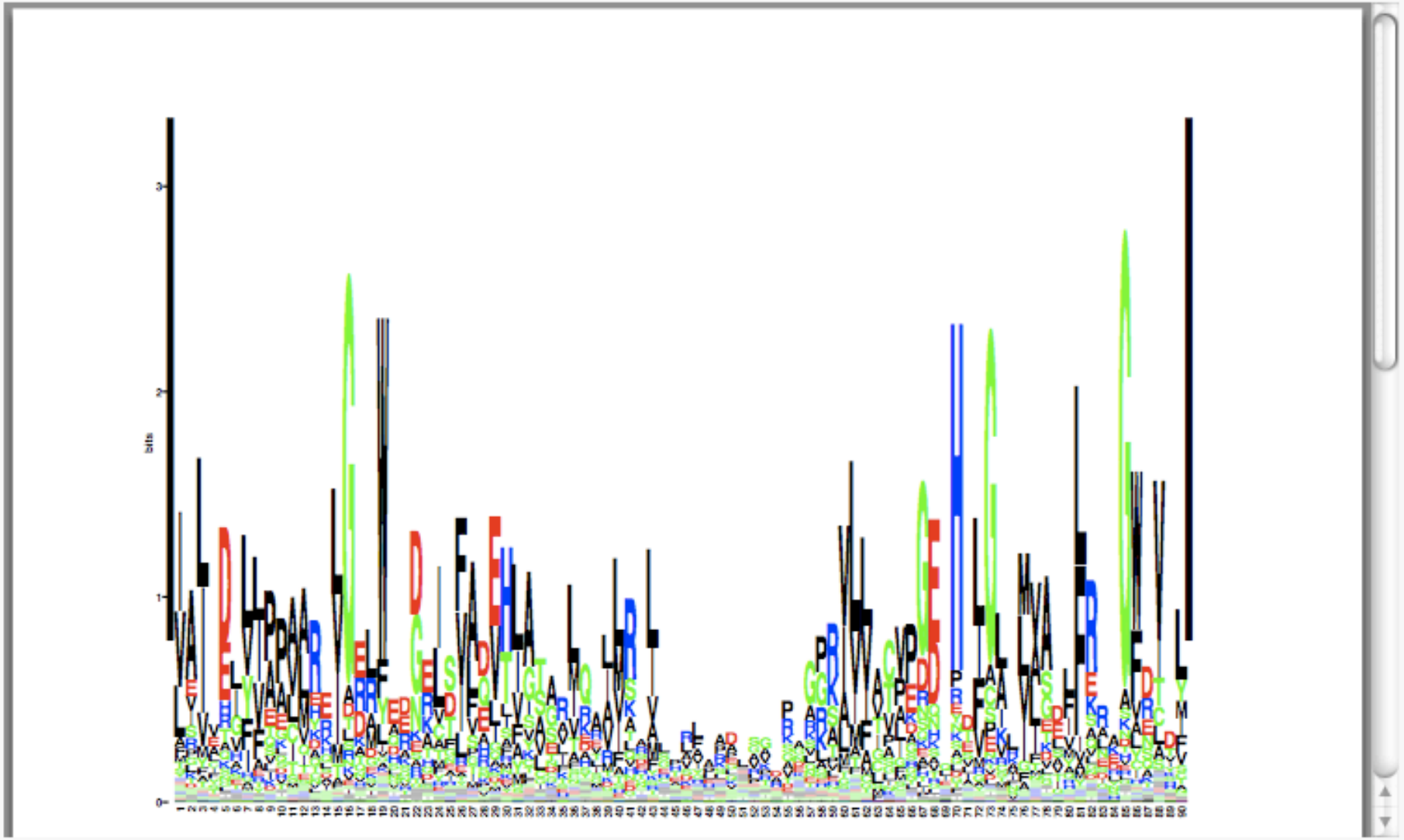
Plot Kullback-Leibler logo ☐

File format for logo file PDF

Submit Clear fields

Restrictions:
At most 1 sequences per submission; each sequence not more than 20,000 amino acids.

Blast2logo



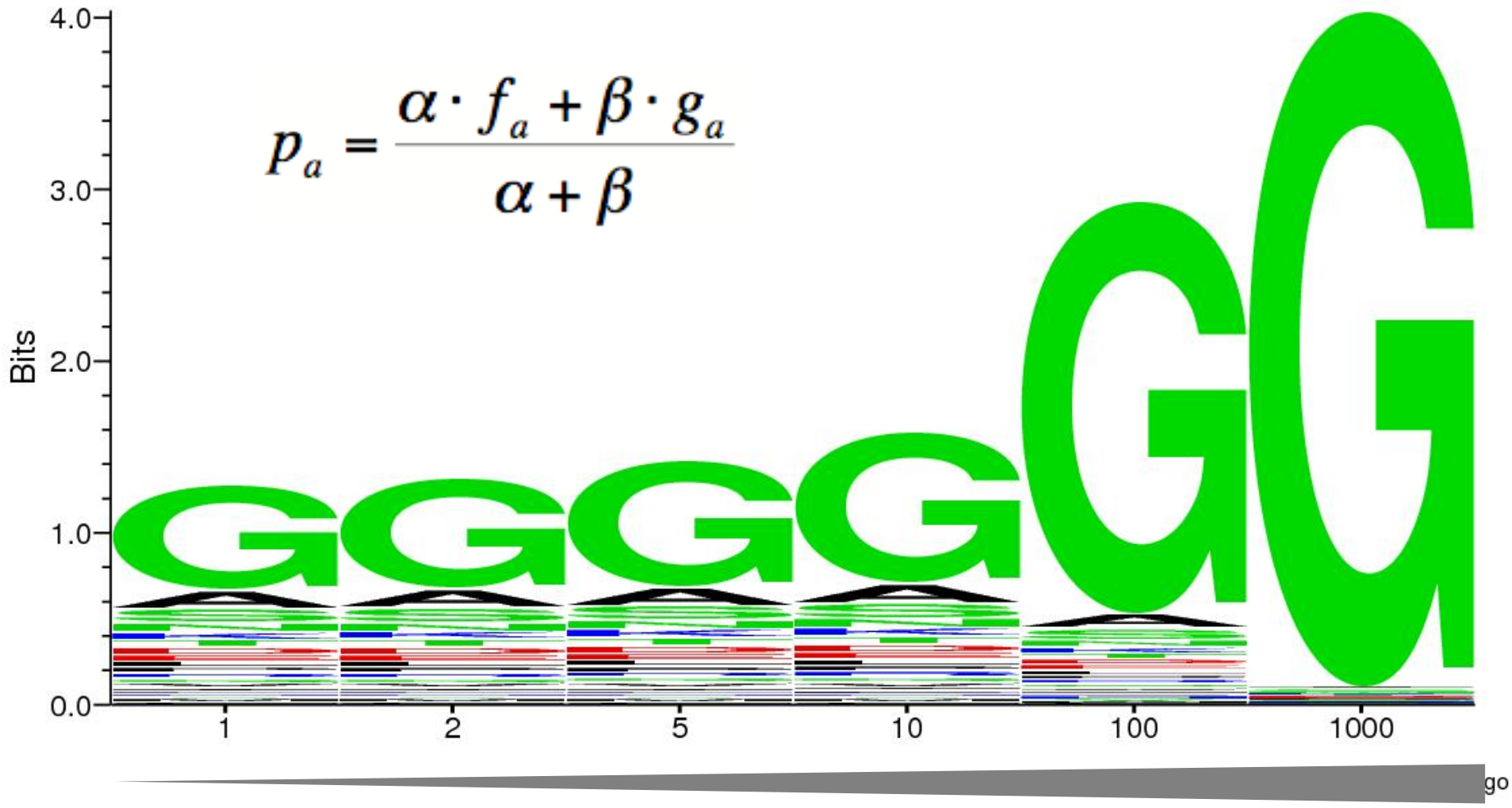
Blast2logo

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Last position-specific scoring matrix computed,

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

Sequence profiles or Gaining confidence



Example.

>1K7C.A
TTVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADV
VTAGDYVIVEFGHNDGGSLSTDNGRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKL
FTAKGAKVILSSQTPNNPWETGTFVNSPTRFVEYAELAAEVAGVEYVDHWSYVDSIYETL
GNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSLKSVLTTTSFEGTCL

- What is the function
- Where is the active site?

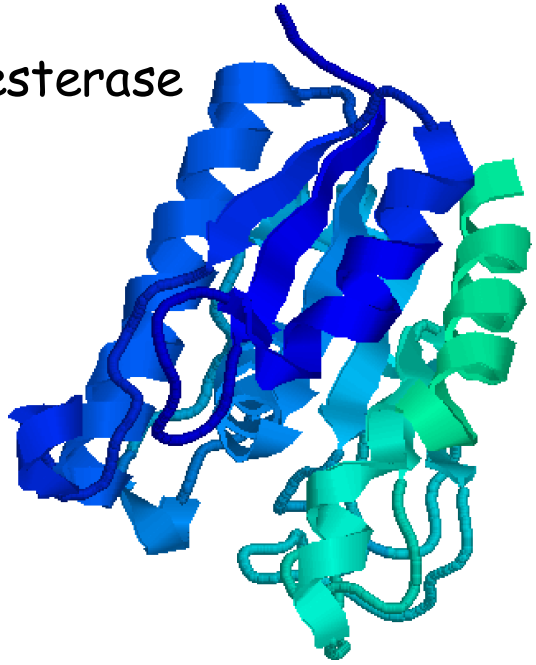
What would you do?

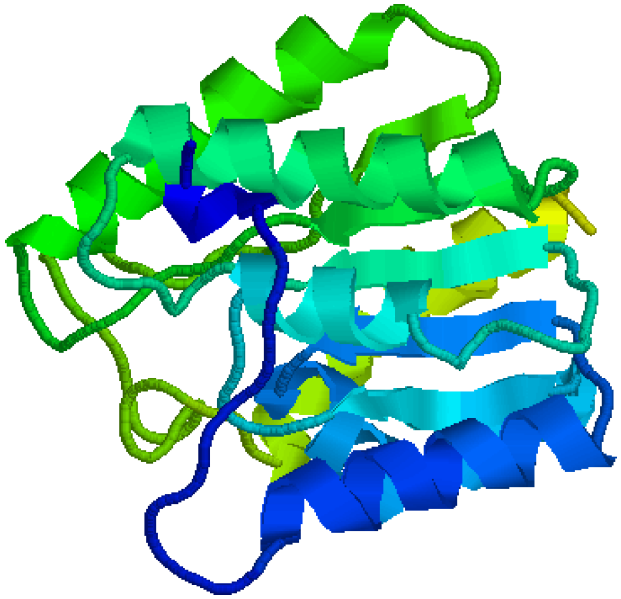
- Function
  - Run Blast against PDB
    - No significant hits
  - Run Blast against NR (Sequence database)
    - Function is Acetylesterase?
- Where is the active site?

Example. Where is the active site?

1USW Hydrolase

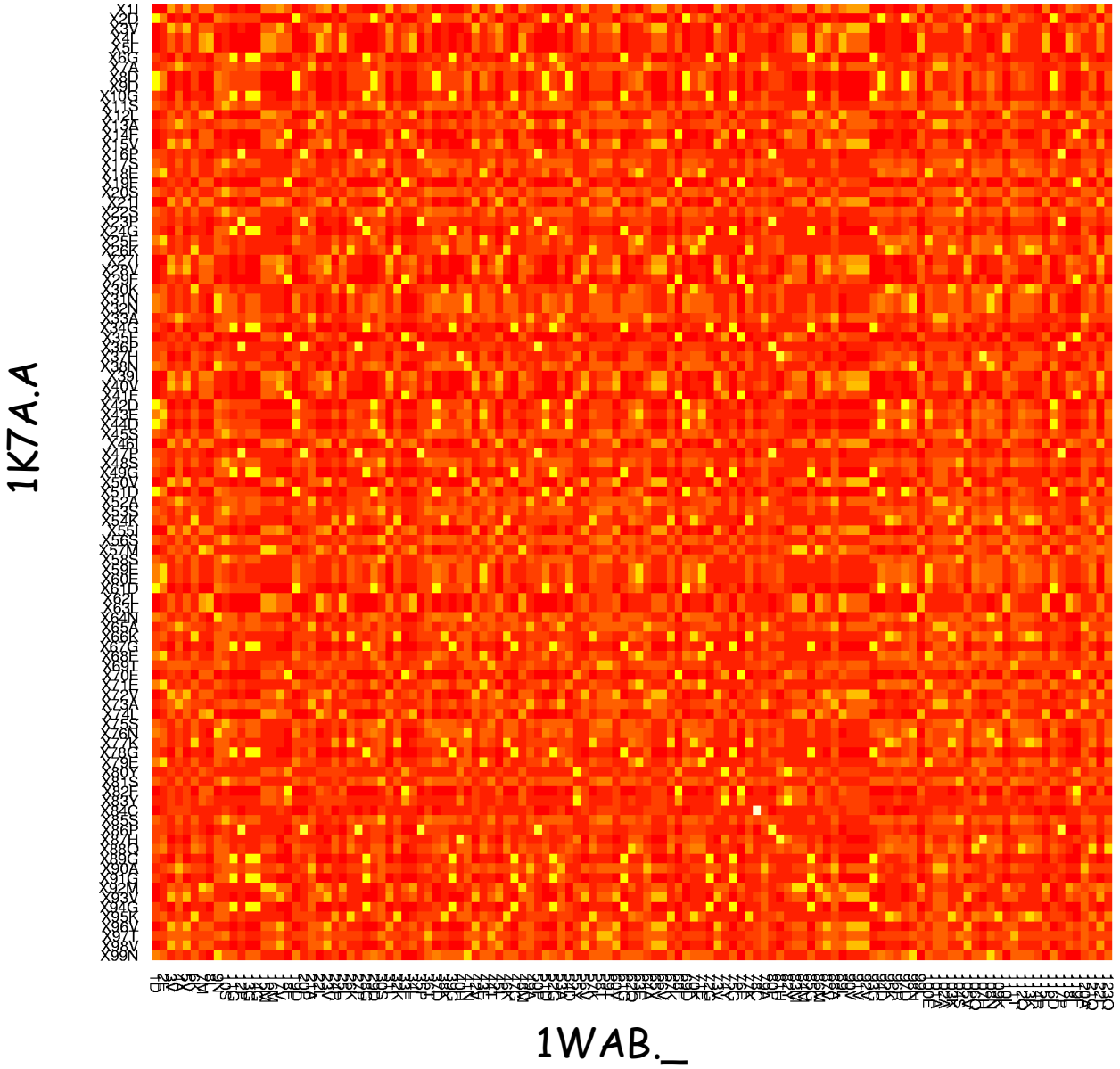
1G66 Acetylxylan esterase





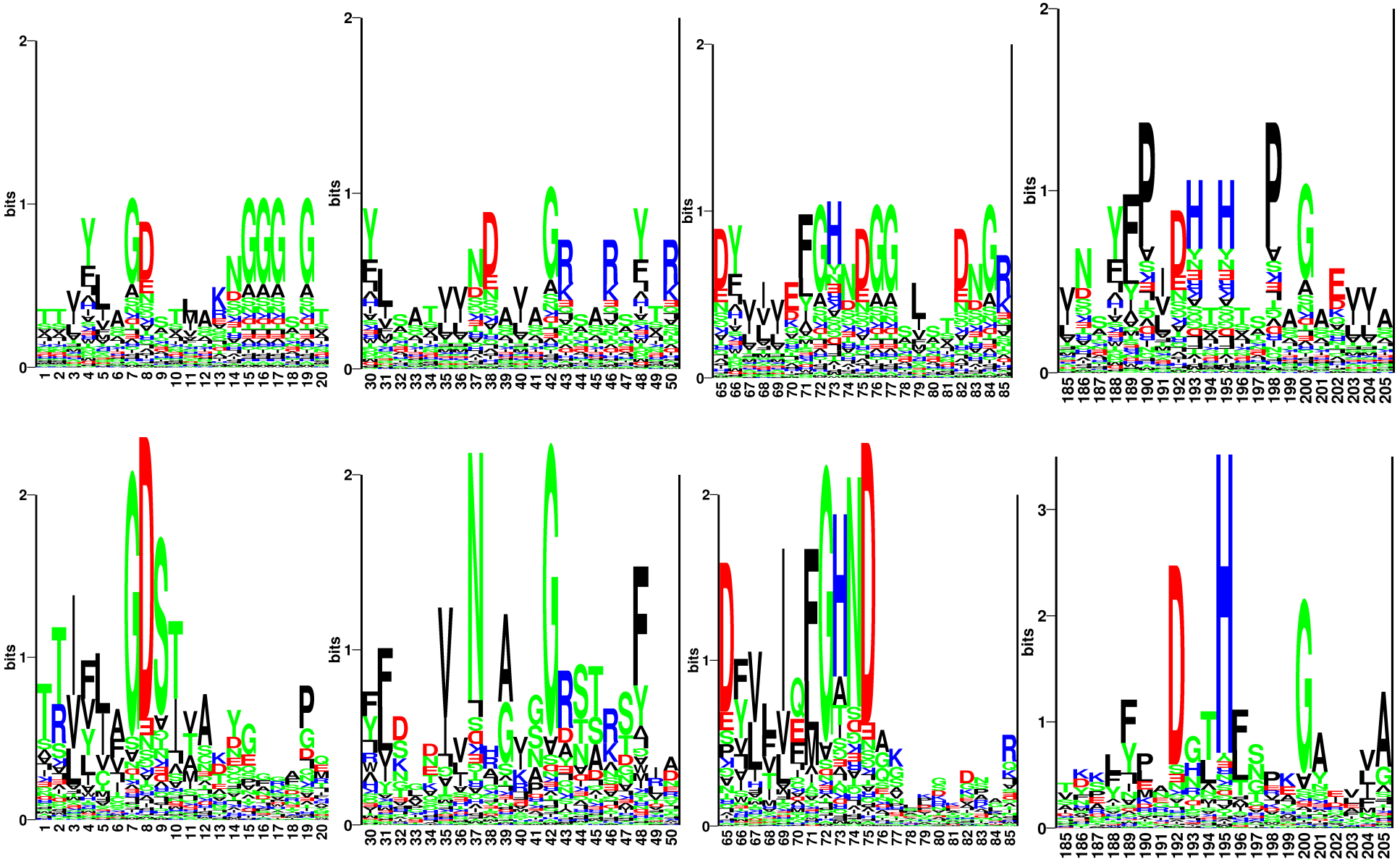
1WAB Acetylhydrolase

When Blast fails!



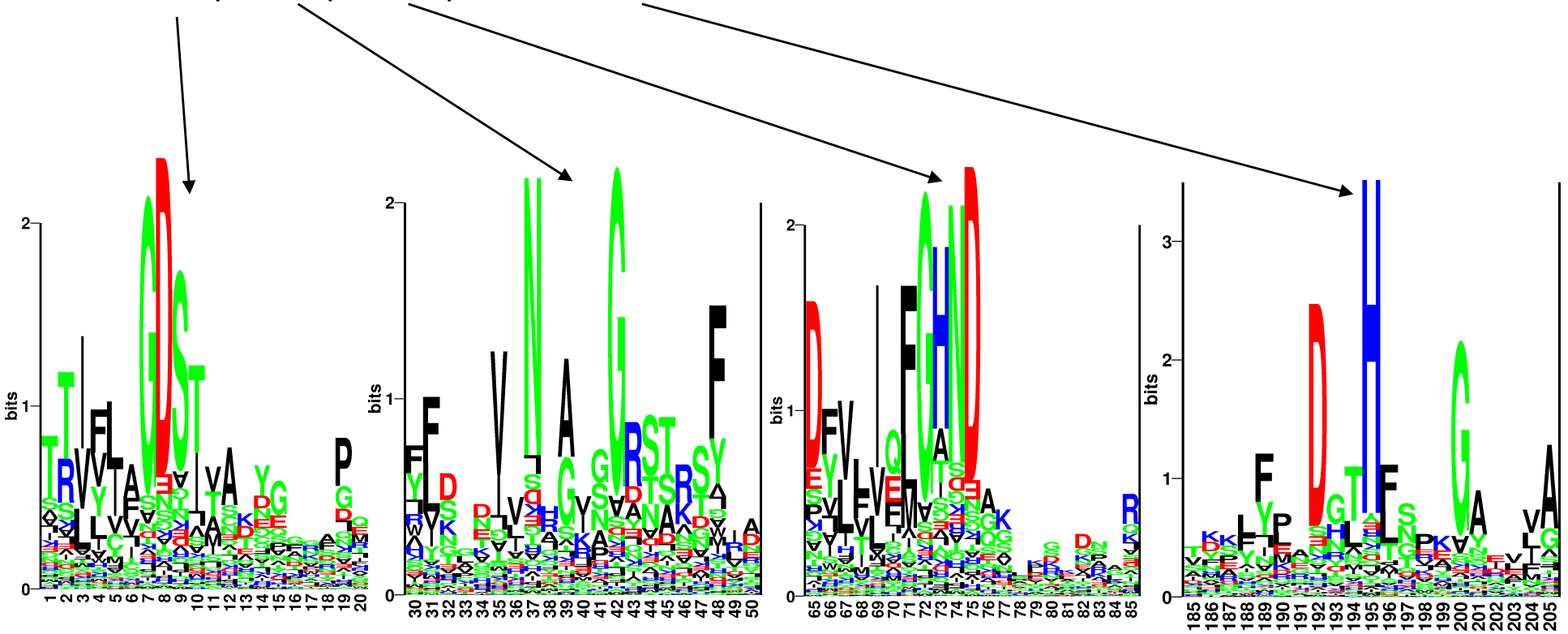
Example. (SGNH active site)

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

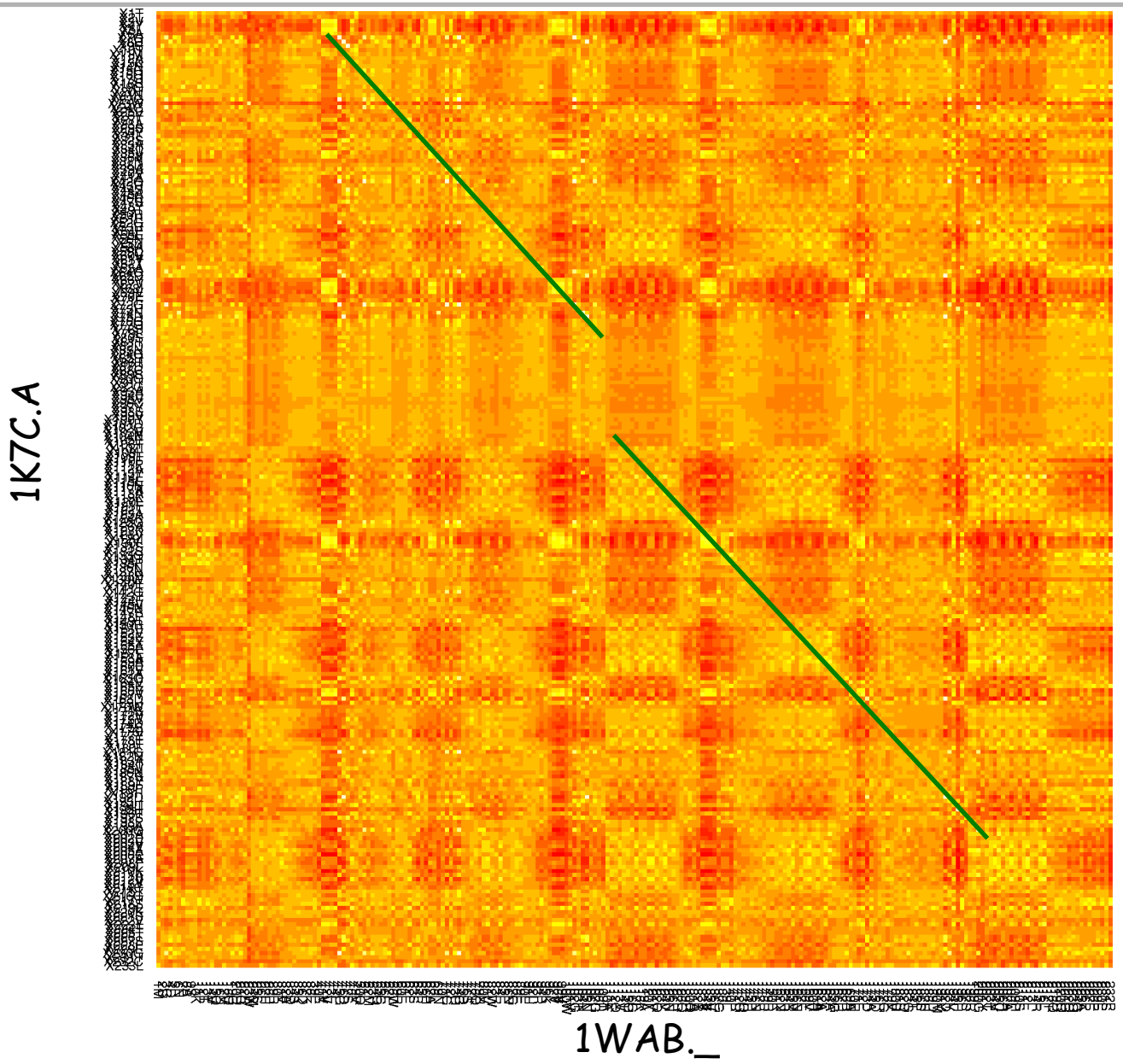


Example. Where is the active site?

- Sequence profiles might show you where to look!
- The active site could be around
  - S9, G42, N74, and H195



Profile-profile scoring matrix



Example. Where is the active site?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Align using sequence profiles

ALN 1K7C.A 1WAB._ RMSD = 5.29522.14% ID

```
1K7C.A TVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADVVTAGDYVIVEFGHNDGGSLSTDN
            S                                   G                                 N
1WAB._ EVVFIGDSLVQLMHQCE---IWRELFS---PLHALNFGIGGDSTQHVLW--RLENGELEHIRPKIVVVWVGTNNHG------

1K7C.A GRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKLFTAK--GAKVILSSQTPNNPWETGTFVNSPTRFVEYAEL-AAEVA
1WAB._ --------------------HTAEQVTGGIKAIVQLVNERQPQARVVVLGLLPRGQ-HPNPLREKNRRVNELVRAALAGHP

1K7C.A GVEYVDHWSYVDSIYETLGNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSL
                                        H
1WAB._ RAHFLDADPG---FVHSDG--TISHHDMYDYLHLSRLGYTPVCRALHSLLLRL---L
```

Handout exercise

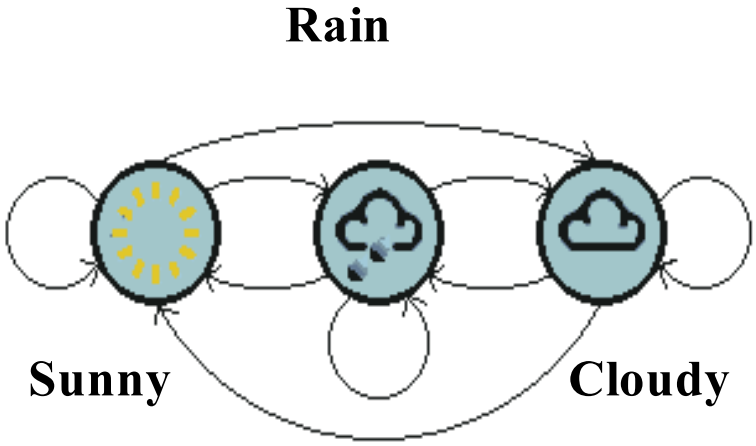
Using Psi-Blast Profiles

Sequence profiles take home message

- Blast will often fail to recognize sequence relationships for low homology sequence pairs
- Sequence profiles contain information on conserved/variable residues in a protein sequence
- Sequence profiles are calculated from (multiple) sequence alignments
- Iterative Blast enables homology recognition also for low sequence similarity
- Sequence profiles give information on residues essential for protein function and protein structure
- Can be used to predict impact of SNP's on protein function
  - This is often done using the Blosum matrix, but profiles are much more precise

Hidden Markov Models, HMM's

Markov Chains



States : Three states - sunny, cloudy, rainy.

weather today

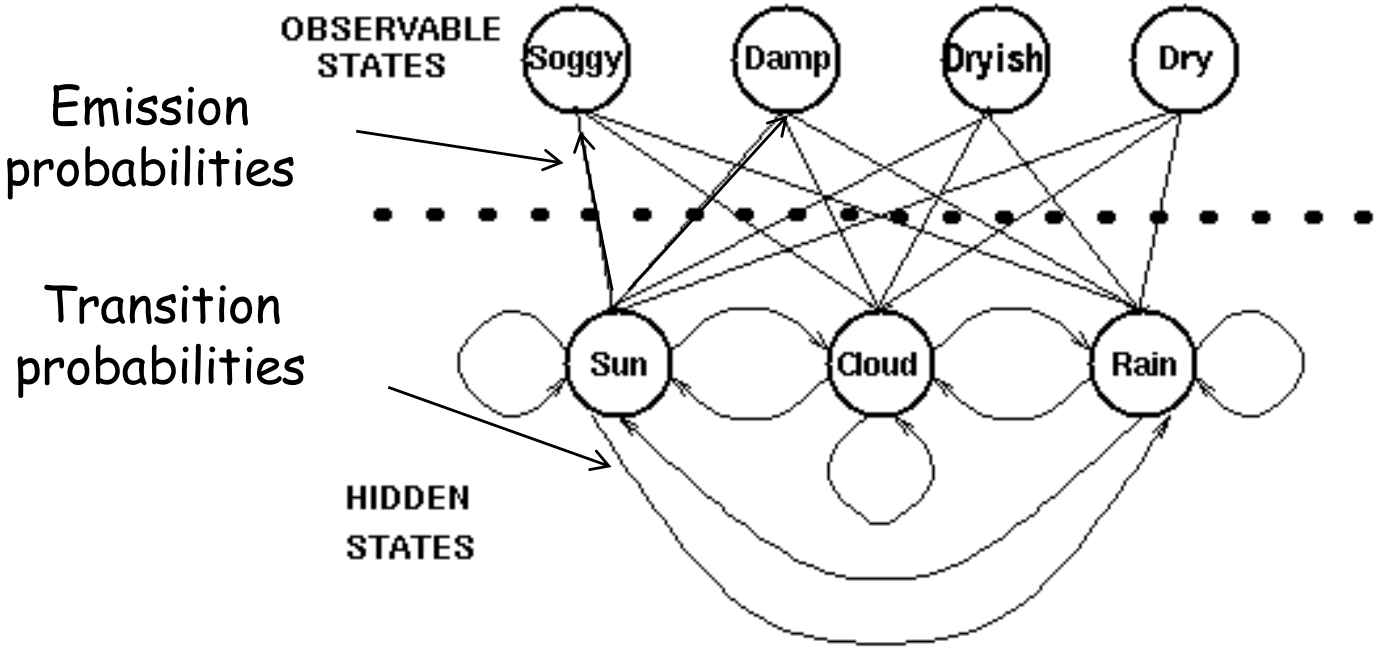
weather yesterday	Sun	Cloud	Rain
Sun	0.5	0.25	0.25
Cloud	0.375	0.125	0.375
Rain	0.125	0.625	0.375

State transition matrix : The probability of the weather given the previous day's weather.

Sun	Cloud	Rain
1.0	0.0	0.0

Initial Distribution : Defining the probability of the system being in each of the states at time 0.

Hidden Markov Models



Hidden states : the (TRUE) states of a system that may be described by a Markov process (e.g., the weather).

Observable states : the states of the process that are `visible' (e.g., seaweed dampness).

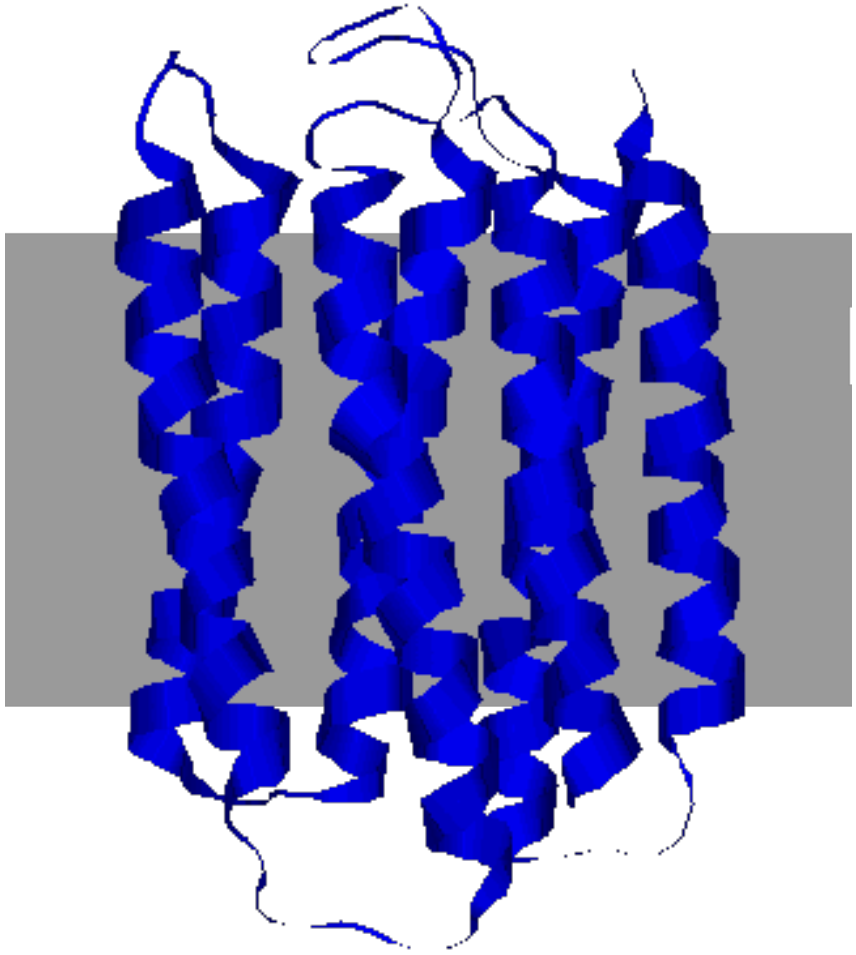
TMHMM (trans-membrane HMM) (Sonnhammer, von Heijne, and Krogh)

>3UTX:A|PDBID|CHAIN|SEQUENCE
QAQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAIATLVPAIAFTMYLSMLLGYGLTMVP
FGGEQNPIYWARYADWLFTTPLLLDLALLVDADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAIST
AAMLYILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIETLLFMVLDVS
AKVGFGLILLRSRAIFGEAEAPEPSAGDGAAATSD

Extra cellular

Trans membrane

Intra cellular



Weight matrix (PSSM) construction

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPFLVSV KLLEPVLLL LLDVPTAAV LLDVPTAAV LLDVPTAAV
LLDVPTAAV VLFRGGPRG MVDGTLLLL YMNGTMSQV MLLSVPLLL SLLGLLVEV ALLPPINIL TLIKIQHTL
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGFVFTL STNRQSGRQ GLDVLTAKV RILGAVAKV QVCERIPTI
ILFGHENRV ILMEHIHKL ILDQKINEV SLAGGIIGV LLIENVASL FLLWATAEA SLPDFGISY KKREEAPSL
LERPGGNEI ALSNLEVKL ALNELLQHV DLERKVESL FLGENISNF ALSDHHIYL GLSEFTEYL STAPPAHGV
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTAFARV RLDSYVRSL YMNGTMSQV GILGFVFTL ILKEPVHGV
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSLLVPFV FLPSDFFPS CLGGLLTMV FIAGNSAYE KLGEFYNQM
KLVALGINA DLMGYIPLV RLVTLKDIV MLLAVLYCL AAGIGILTV YLEPGPVTA LLDGTATLR ITDQVPFSV
KTWGQYWQV TITDQVPFS AFHHVAREL YLNKIQNSL MMRKLAILS AIMDKNIIL IMDKNIILK SMVGNWAKV
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTPLCVTL VLYRYGSFS YIGEVLVSV CINGVCWTV VMNILLQYV
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLV GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLAA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLTA AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRAFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLI LLGRNSFEV PLTFGWCYK VLEWRFDSR TLNAWVKVV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGSPY LLWTLVVLL SVRDRLARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFFVWL FLYGALLLA VLFSSDFRI LMWAKIGPV SLLLELEEV SLSRFSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMIIPLINV TLFIGSHVV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGTHTMEV MLGTHTMEV SLADTNSLA LLWAARPRL GVALQTMKQ
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPPGTRV KVAELVHFL IMIGVLVGV ALCRWGLLL LLFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDRLARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYI
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

PSSM construction

- Calculate amino acid frequencies at each position using
  - Sequence weighting
  - Pseudo counts
- Define background model
  - Use background amino acid frequencies
- PSSM is

$$S(a_i) = \log \frac{p(a_i)}{q(a)}$$

More on scoring

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

$$S = \sum_i S(a_i)$$

$$S = \sum_i \log \frac{p(a_i)}{q(a_i)}$$

$$S = \log \left(\frac{\prod_i p(a_i)}{\prod_i q(a_i)} \right)$$

Probability of observation given Model

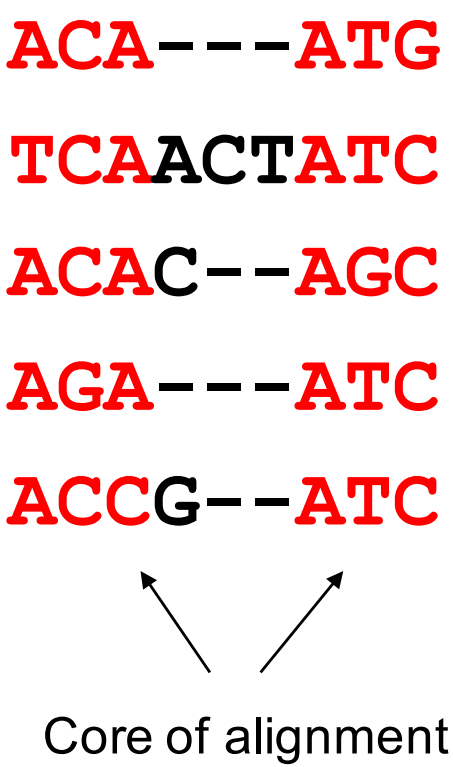
Probability of observation given Prior (background)

$$S = \log \left(\frac{P(a \mid M)}{P(a \mid B)} \right)$$

Hidden Markov Models

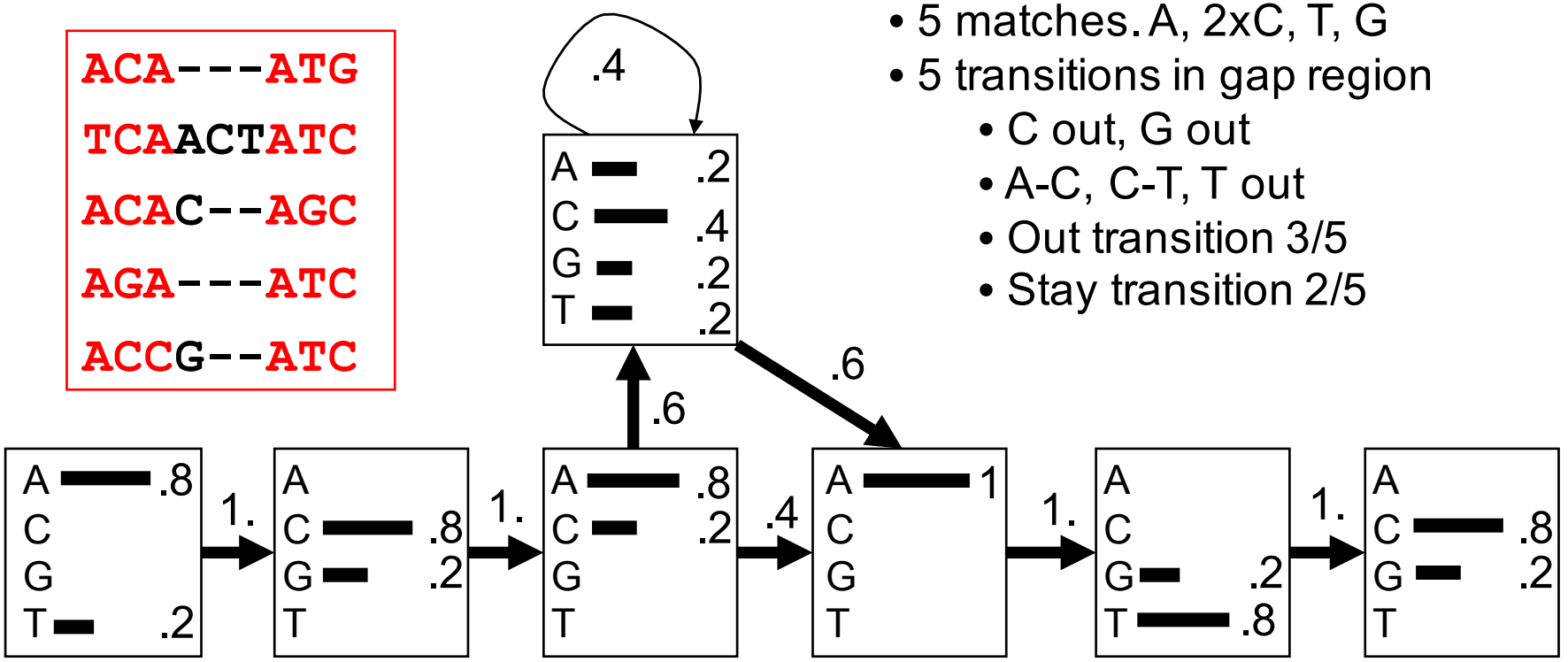
- Weight matrices do not deal with insertions and deletions
- In alignments, this is done in an ad-hoc manner by optimization of the two gap penalties for first gap and gap extension
- HMM is a natural framework where insertions/deletions are dealt with explicitly

HMM (a simple example)



- Example from A. Krogh
- Core region defines the number of states in the HMM (red)
- Insertion and deletion statistics are derived from the non-core part of the alignment (black)

HMM construction (supervised learning)



- 5 matches. A, 2xC, T, G
- 5 transitions in gap region
  - C out, G out
  - A-C, C-T, T out
  - Out transition 3/5
  - Stay transition 2/5

ACA---ATG 0.8x1x0.8x1x0.8x0.4x1x1x0.8x1x0.2 = 3.3×10^{-2}

Scoring a sequence to an HMM

ACA---ATG 0.8x1x0.8x1x0.8x0.4x1x0.8x1x0.2 = 3.3×10^{-2}

TCAACT**ATC** 0.2x1x0.8x1x0.8x0.6x0.2x0.4x0.4x0.4x0.2x0.6x1x1x0.8x1x0.8 = 0.0075×10^{-2}

ACAC--**AGC** = 1.2×10^{-2}

Consensus:

ACAC--**ATC** = 4.7×10^{-2}, **ACA---ATC** = 13.1×10^{-2}

Exceptional:

TGCT--**AGG** = 0.0023×10^{-2}

Align sequence to HMM - Null model

- Score depends strongly on length
- Null model is a random model. For length L the score is 0.25^L
- Log-odds score for sequence S
- Log(P(S)/0.25^L)
- Positive score means more likely than Null model

- **This is just like we did for PSSM log(p/q)!**

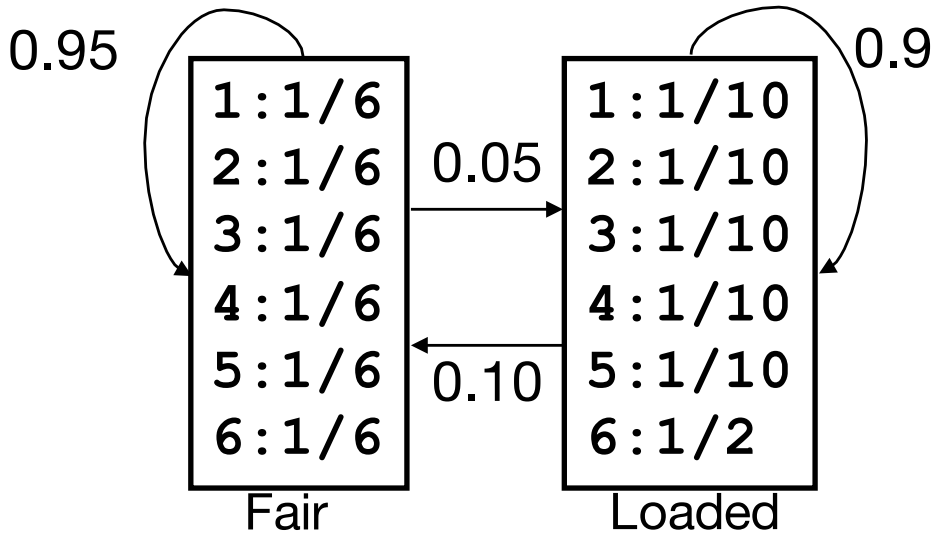
```
ACA---ATG = 4.9
TCAACTATC = 3.0
ACAC--AGC = 5.3
AGA---ATC = 4.9
ACCG--ATC = 4.6
Consensus:
ACAC--ATC = 6.7
ACA---ATC = 6.3
Exceptional:
TGCT--AGG = -0.97
```

Note!

Why hidden?

The unfair casino: Loaded die p(6) = 0.5; switch fair to load:0.05; switch load to fair: 0.1

- Model generates numbers
  - 312453666641
- Does not tell which die was used
- Alignment (decoding) can give the most probable solution/path (Viterbi)
  - FFFFFFLLLLLL
- Or most probable set of states
  - FFFFFFLLLLLL



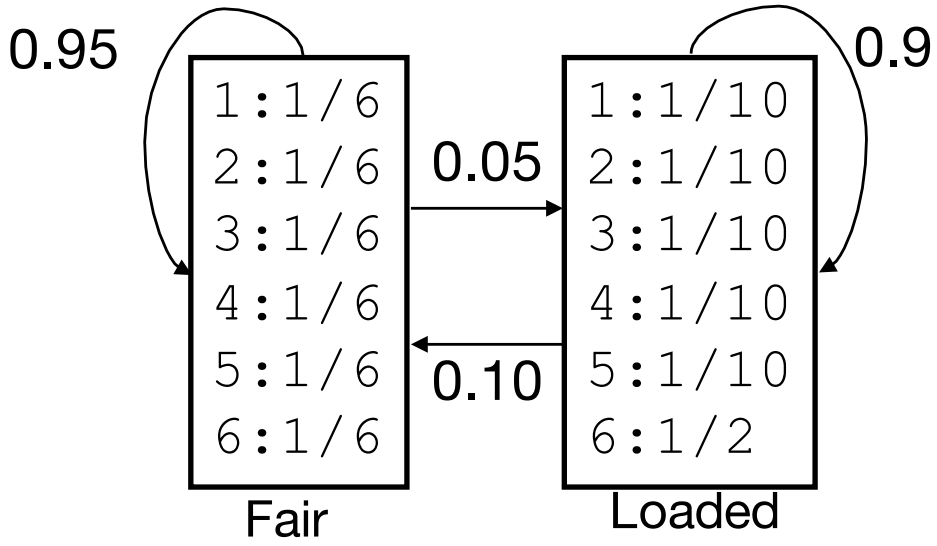
Aligning a sequence to an HMM

- Find the path through the HMM states that has the highest probability
  - For alignment, we found the path through the scoring matrix that had the highest sum of scores
- The number of possible paths rapidly gets very large making brute force search infeasible
  - Just like checking all path for alignment did not work
- Use dynamic programming
  - The Viterbi algorithm does the job

The Viterbi algorithm

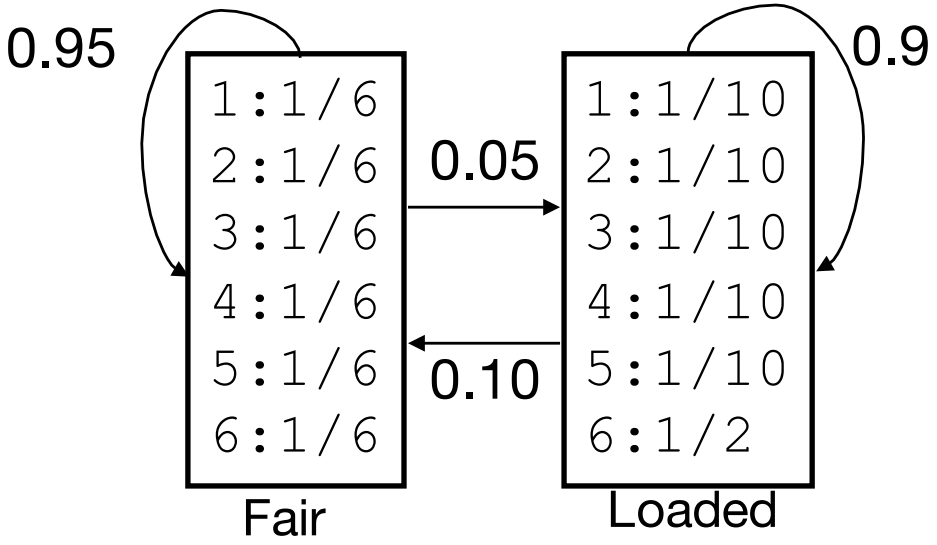
The unfair casino: Loaded dice p(6) = 0.5; switch fair to load:0.05; switch load to fair: 0.1

- Model generates numbers
  - 312453666641



Model decoding (Viterbi)

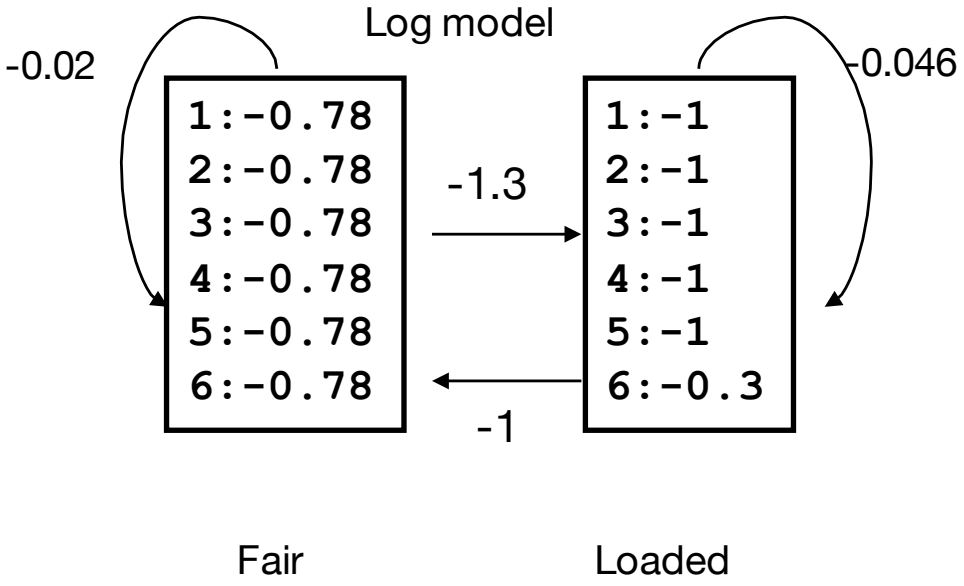
- Example: 566. What was the most likely series of dice used to generate this output?
- Use Brute force




```
FFF = 0.5*0.167*0.95*0.167*0.95*0.167 = 0.0021
FFL = 0.5*0.167*0.95*0.167*0.05*0.5 = 0.00333
FLF = 0.5*0.167*0.05*0.5*0.1*0.167 = 0.000035
FLL = 0.5*0.167*0.05*0.5*0.9*0.5 = 0.00094
LFF = 0.5*0.1*0.1*0.167*0.95*0.167 = 0.00013
LFL = 0.5*0.1*0.1*0.167*0.05*0.5 = 0.000021
LLF = 0.5*0.1*0.9*0.5*0.1*0.167 = 0.00038
LLL = 0.5*0.1*0.9*0.5*0.9*0.5 = 0.0101
```

Or in log space

- Example: 566. What was the most likely series of dice used to generate this output?

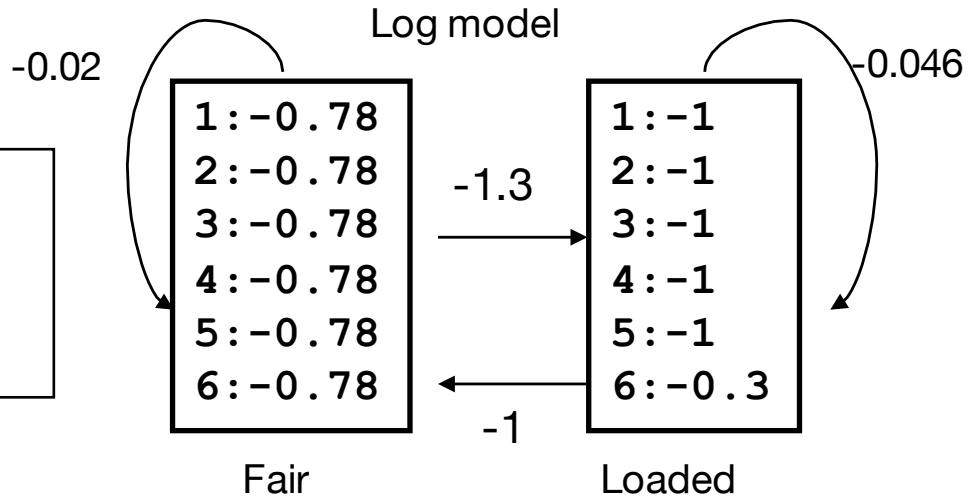



```
Log(P(LLL|M)) = log(0.5*0.1*0.9*0.5*0.9*0.5) = log(0.0101)
                                    or
Log(P(LLL|M)) = log(0.5)+log(0.1)+log(0.9)+log(0.5)+log(0.9)+log(0.5)
              = -0.3 -1 -0.046 -0.3 -0.046 -0.3 = -1.99
```

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?

```
F = 0.5*0.167
log(F) = log(0.5) + log(0.167) = -1.08
L = 0.5*0.1
log(L) = log(0.5) + log(0.1) = -1.30
```

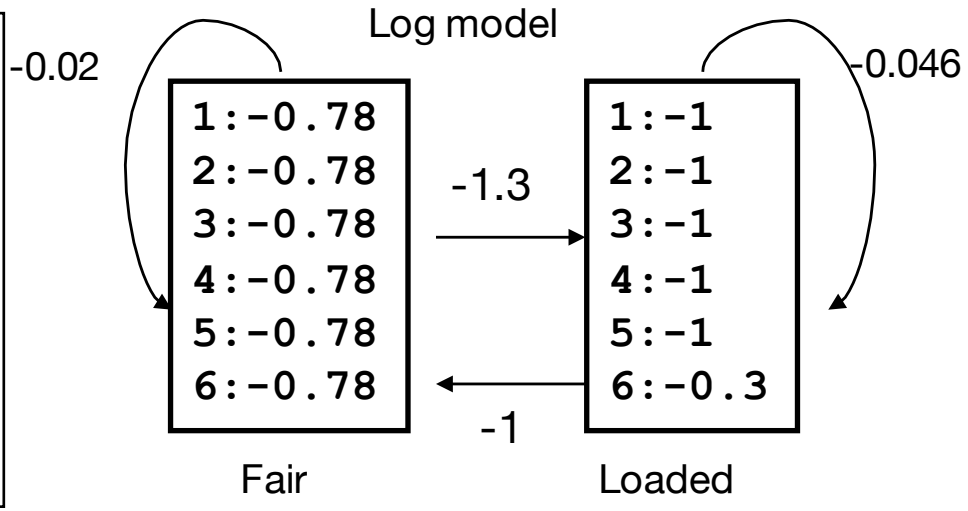


	5	6	6	6	1	1	2	3	4
F	-1.08								
L	-1.30								

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?

```
FF = 0.5*0.167*0.95*0.167
Log(FF) = -0.30 -0.78 - 0.02 -0.78 = -1.88
LF = 0.5*0.1*0.1*0.167
Log(LF) = -0.30 -1 -1 -0.78 = -3.08
FL = 0.5*0.167*0.05*0.5
Log(FL) = -0.30 -0.78 - 1.30 -0.30 = -2.68
LL = 0.5*0.1*0.9*0.5
Log(LL) = -0.30 -1 -0.046 -0.3 = -1.65
```

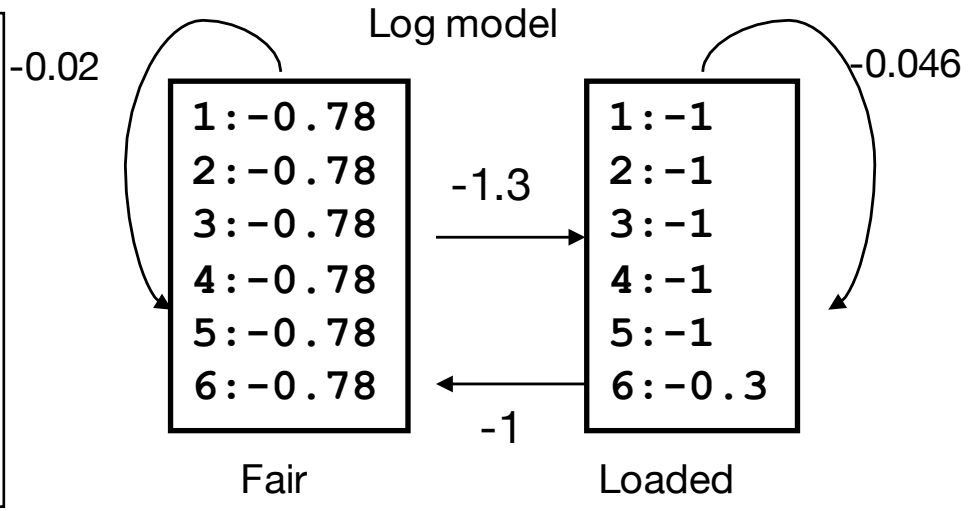


	5	6	6	6	1	1	2	3	4
F	-1.08								
L	-1.30								

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?

```
FF = 0.5*0.167*0.95*0.167
Log(FF) = -0.30 -0.78 - 0.02 -0.78 = -1.88
LF = 0.5*0.1*0.1*0.167
Log(LF) = -0.30 -1 -1 -0.78 = -3.08
FL = 0.5*0.167*0.05*0.5
Log(FL) = -0.30 -0.78 - 1.30 -0.30 = -2.68
LL = 0.5*0.1*0.9*0.5
Log(LL) = -0.30 -1 -0.046 -0.3 = -1.65
```

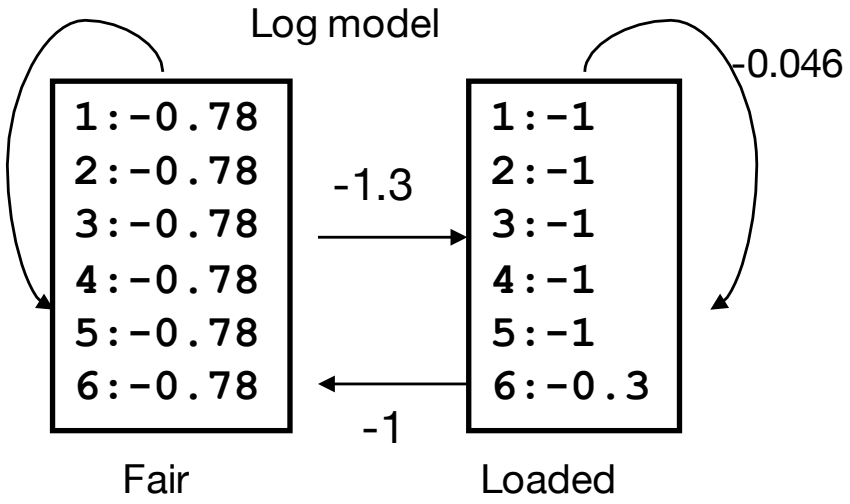


	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88							
L	-1.30	-1.65							

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?

```
FFF = 0.5*0.167*0.95*0.167*0.95*0.167 = 0.0021
FLF = 0.5*0.167*0.05*0.5*0.1*0.167 = 0.000035
LFF = 0.5*0.1*0.1*0.167*0.95*0.167 = 0.00013
LLF = 0.5*0.1*0.9*0.5*0.1*0.167 = 0.00038
FLL = 0.5*0.167*0.05*0.5*0.9*0.5 = 0.00094
FFL = 0.5*0.167*0.95*0.167*0.05*0.5 = 0.00333
LFL = 0.5*0.1*0.1*0.167*0.05*0.5 = 0.000021
LLL = 0.5*0.1*0.9*0.5*0.9*0.5 = 0.0101
```

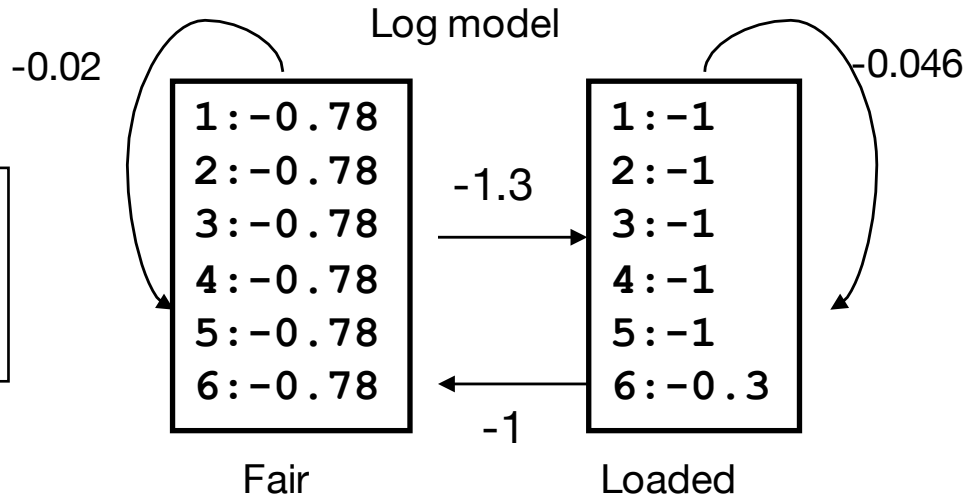


	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88							
L	-1.30	-1.65							

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?

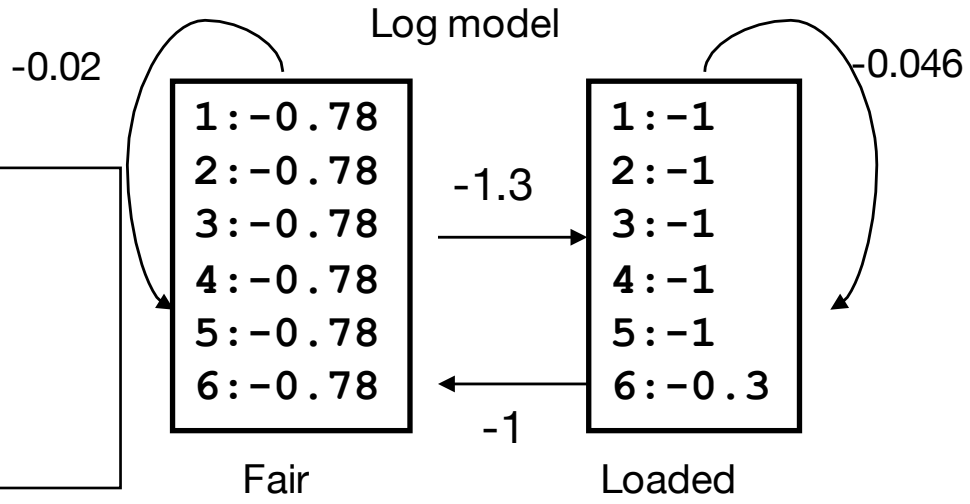
```
FFF = 0.5*0.167*0.95*0.167*0.95*0.167 = 0.0021
Log(P(FFF))=-2.68
LLL = 0.5*0.1*0.9*0.5*0.9*0.5 = 0.0101
Log(P(LLL))=-1.99
```



	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68						
L	-1.30	-1.65	-1.99						

Model decoding (Viterbi) – Or doing it recursively

- Example: 566611234. What was the most likely series of dice used to generate this output?

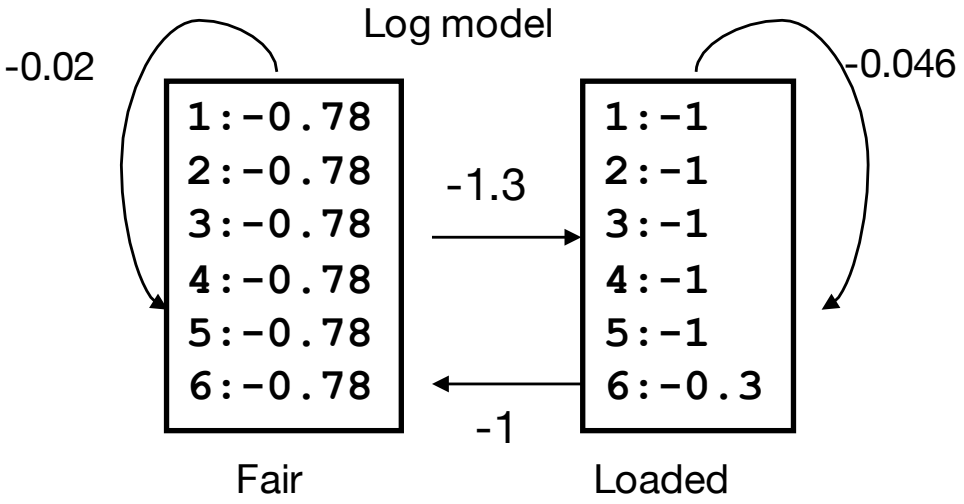



```
FFF = -0.78 + max(-1.88-0.02, -1.99-1)
    = -0.78 -1.88 -0.02
Log(P(FFF))=-2.68
LLL = -0.3 + max(-1.88-1.3, -1.65-0.046)
    = -0.3 - 1.65 - 0.046
Log(P(LLL))=-1.99
```

	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68						
L	-1.30	-1.65	-1.99						

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?



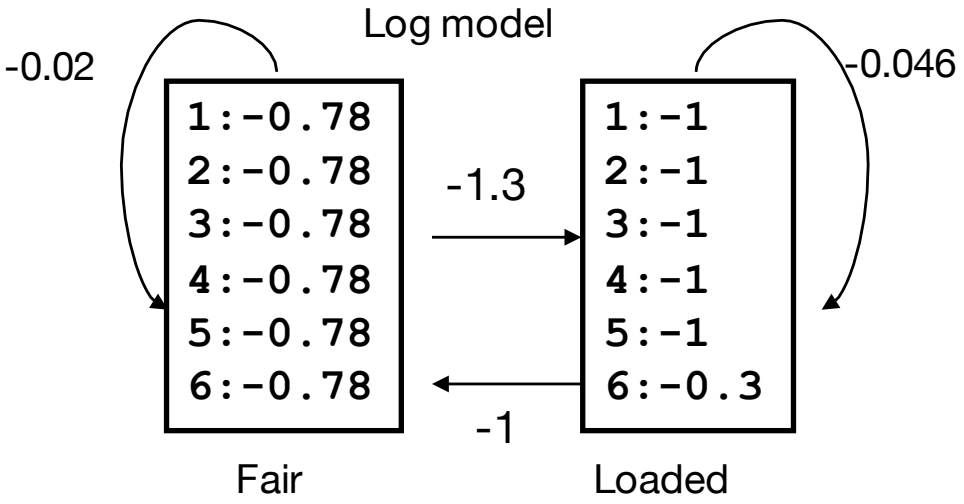
	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68						
L	-1.30	-1.65	-1.99						

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?

$$-0.78 - 0.02 - 2.68 = -3.48$$

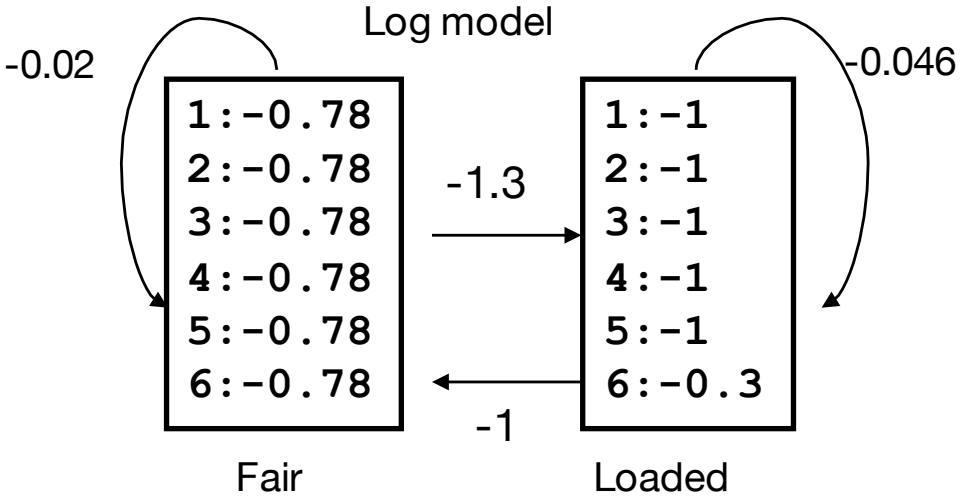
$$-0.78 - 1 - 1.99 = -3.77$$



	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68						
L	-1.30	-1.65	-1.99						

Model decoding (Viterbi)

- Example: 566611234. What was the most likely series of dice used to generate this output?



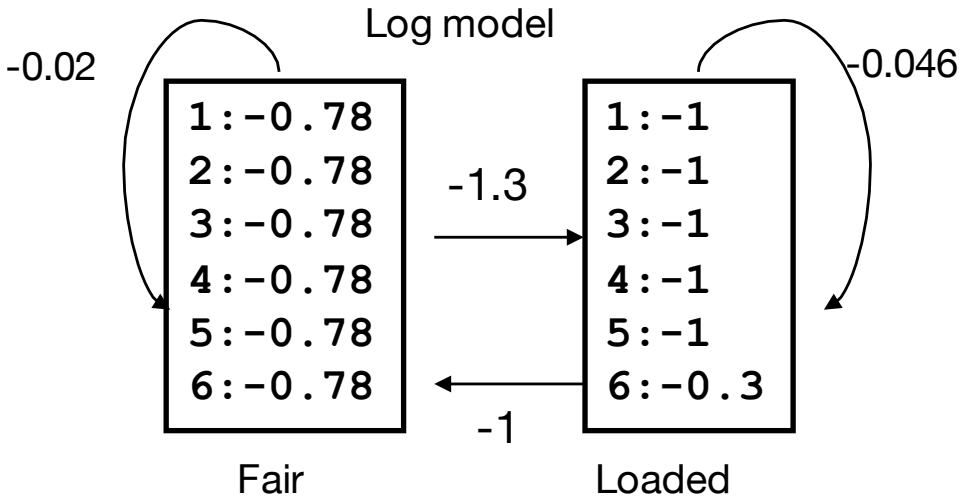
$$-0.78 - 0.02 - 2.68 = -3.48$$

$$-0.78 - 1 - 1.99 = -3.77$$

	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68	-3.48					
L	-1.30	-1.65	-1.99						

Model decoding (Viterbi)

- Now we can formalize the algorithm!



$$P_l(i+1) = p_l(i+1) \bullet \max_k(P_k(i) \bullet a_{kl}) \quad or$$

$$\log(P_l(i+1) = \log(p_l(i+1)) + \max_k(\log(P_k(i) + \log(a_{kl}))$$

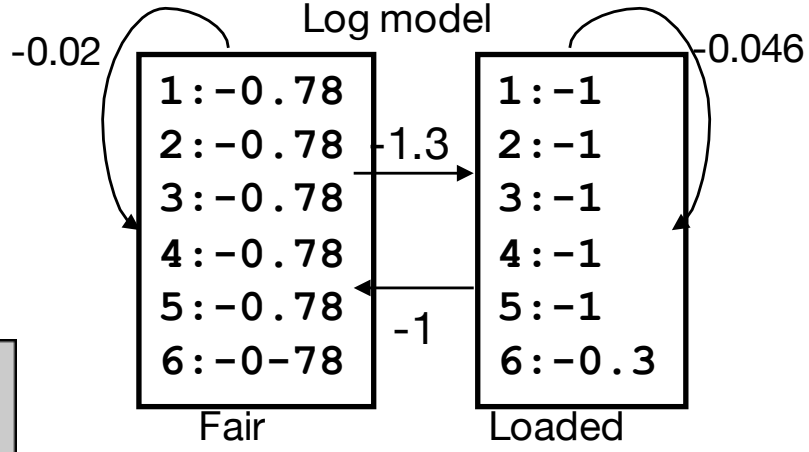
New match Old max score Transition

Model decoding (Viterbi). Can you do it?

- Example: 566611234. What was the mots likely series of dice used to generate this output?
- Fill out the table using the Viterbi recursive algorithm
  - Add the arrows for backtracking
- Find the optimal path

$$P_l(i+1) = p_l(i+1) \bullet \max_k(P_k(i) \bullet a_{kl}) \quad or$$

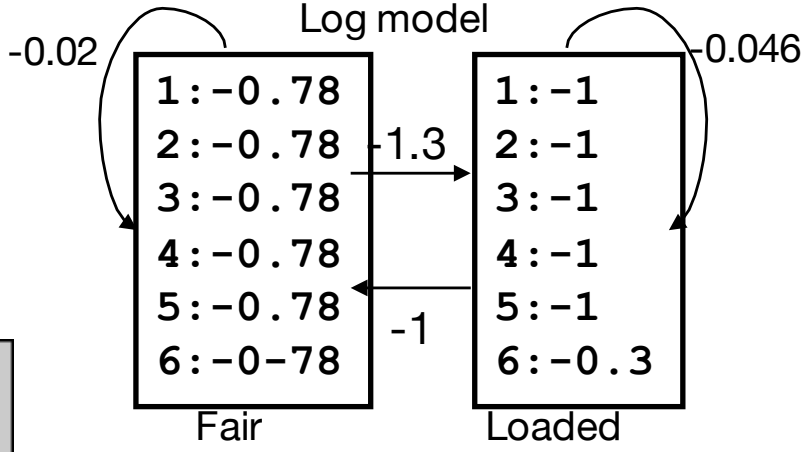
$$\log(P_l(i+1) = \log(p_l(i+1)) + \max_k(\log(P_k(i) + \log(a_{kl}))$$



	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68	-3.48					
L	-1.30	-1.65	-1.99						

Model decoding (Viterbi). Can you do it?

- Example: 566611234. What was the most likely series of dice used to generate this output?
- Fill out the table using the Viterbi recursive algorithm
  - Add the arrows for backtracking
- Find the optimal path



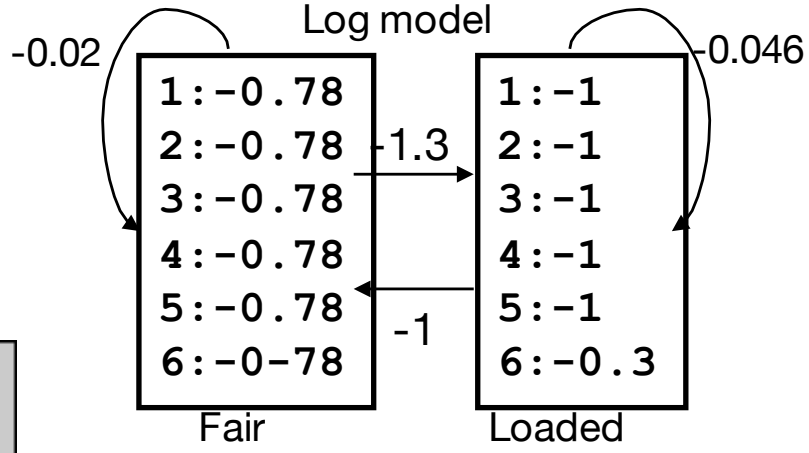
$$P_l(i+1) = p_l(i+1) \bullet \max_k(P_k(i) \bullet a_{kl}) \quad or$$

$$\log(P_l(i+1) = \log(p_l(i+1)) + \max_k(\log(P_k(i) + \log(a_{kl}))$$

	5	6	6	6	1	1	2	3	4
F	-1.08 ←	-1.88 ←	-2.68 ←	-3.48		-4.92 ←	-5.73 ←	-6.53	
L	-1.30 ←	-1.65 ←	-1.99		-3.39 ←	-4.44		-6.52 ←	-7.57

Model decoding (Viterbi). Can you do it?

- Example: 566611234. What was the most lilely series of dice used to generate this output?
- Fill out the table using the Viterbi recursive algorithm
  - Add the arrows for backtracking
- Find the optimal path



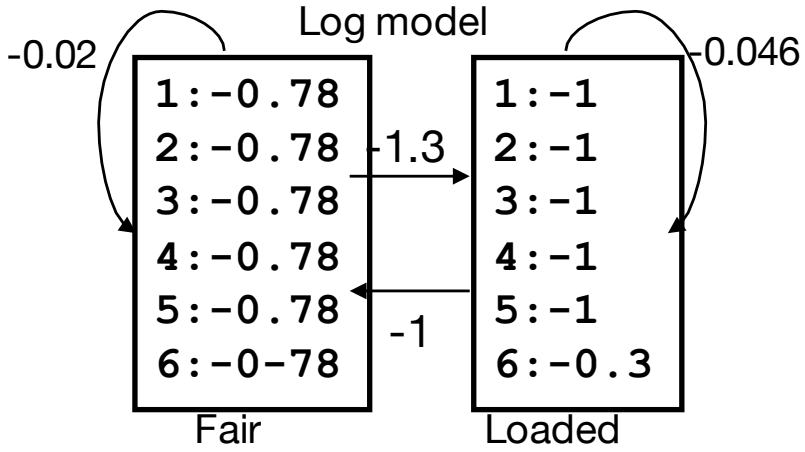
$$P_l(i+1) = p_l(i+1) \bullet \max_k(P_k(i) \bullet a_{kl}) \quad or$$

$$\log(P_l(i+1) = \log(p_l(i+1)) + \max_k(\log(P_k(i) + \log(a_{kl}))$$

	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68	-3.48	-4.12	-4.92	-5.73	-6.53	-7.33
L	-1.30	-1.65	-1.99	-2.34	-3.39	-4.44	-5.48	-6.52	-7.57

Model decoding (Viterbi). Can you do it?

- Example: 566611234. What was the most likely series of dice used to generate this output?
- The most likely path is
  - LLLLFFFFF



	5	6	6	6	1	1	2	3	4
F	-1.08	-1.88	-2.68	-3.48	-4.12	-4.92	-5.73	-6.53	-7.33
L	-1.30	-1.65	-1.99	-2.34	-3.39	-4.44	-5.48	-6.52	-7.57

Model decoding (Viterbi).

- What happens if you have three dice?

	5	6	6	6	1	1	2	3	4
F	-1.0								
L1	-1.2								
L2	-1.3								

$$P_l(i+1) = p_l(i+1) \bullet \max_k(P_k(i) \bullet a_{kl}) \quad or$$

$$\log(P_l(i+1) = \log(p_l(i+1)) + \max_k(\log(P_k(i) + \log(a_{kl}))$$

And if you have a trans-membrane model

- What is the most likely path (alignment) of a protein sequence to the model

	D	G	V	L	I	M	A	D	Q
iC	-1.0								
M	-1.2								
xC	-1.3								

$$P_l(i+1) = p_l(i+1) \bullet \max_k(P_k(i) \bullet a_{kl}) \quad or$$

$$\log(P_l(i+1) = \log(p_l(i+1)) + \max_k(\log(P_k(i) + \log(a_{kl}))$$

HMM's and weight matrices

- In the case of un-gapped alignments HMM's become simple weight matrices
- To achieve high performance, the emission frequencies are estimated using the techniques of
  - Sequence weighting
  - Pseudo counts

Profiles and profile HMM's

- Alignments based on conventional scoring matrices (BLOSUM62) scores all positions in a sequence in an equal manner
- Some positions are highly conserved, some are highly variable (more than what is described in the BLOSUM matrix)
- Profile HMM's are ideal suited to describe such position specific variations

Sequence profiles

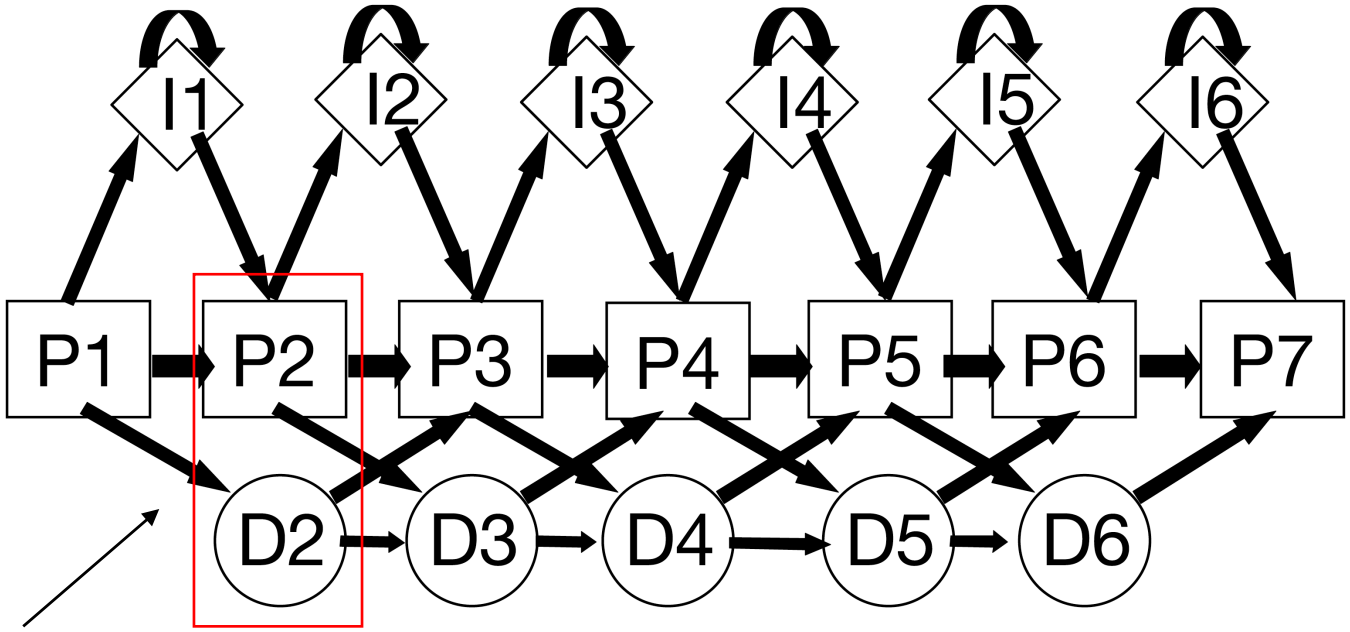
Matching any thing
but G => large
negative score

Any thing can match

HMM vs. alignment

- Detailed description of core
  - Conserved/variable positions
- Price for insertions/deletions varies at different locations in sequence
- These features cannot be captured in conventional alignments

Profile HMM's



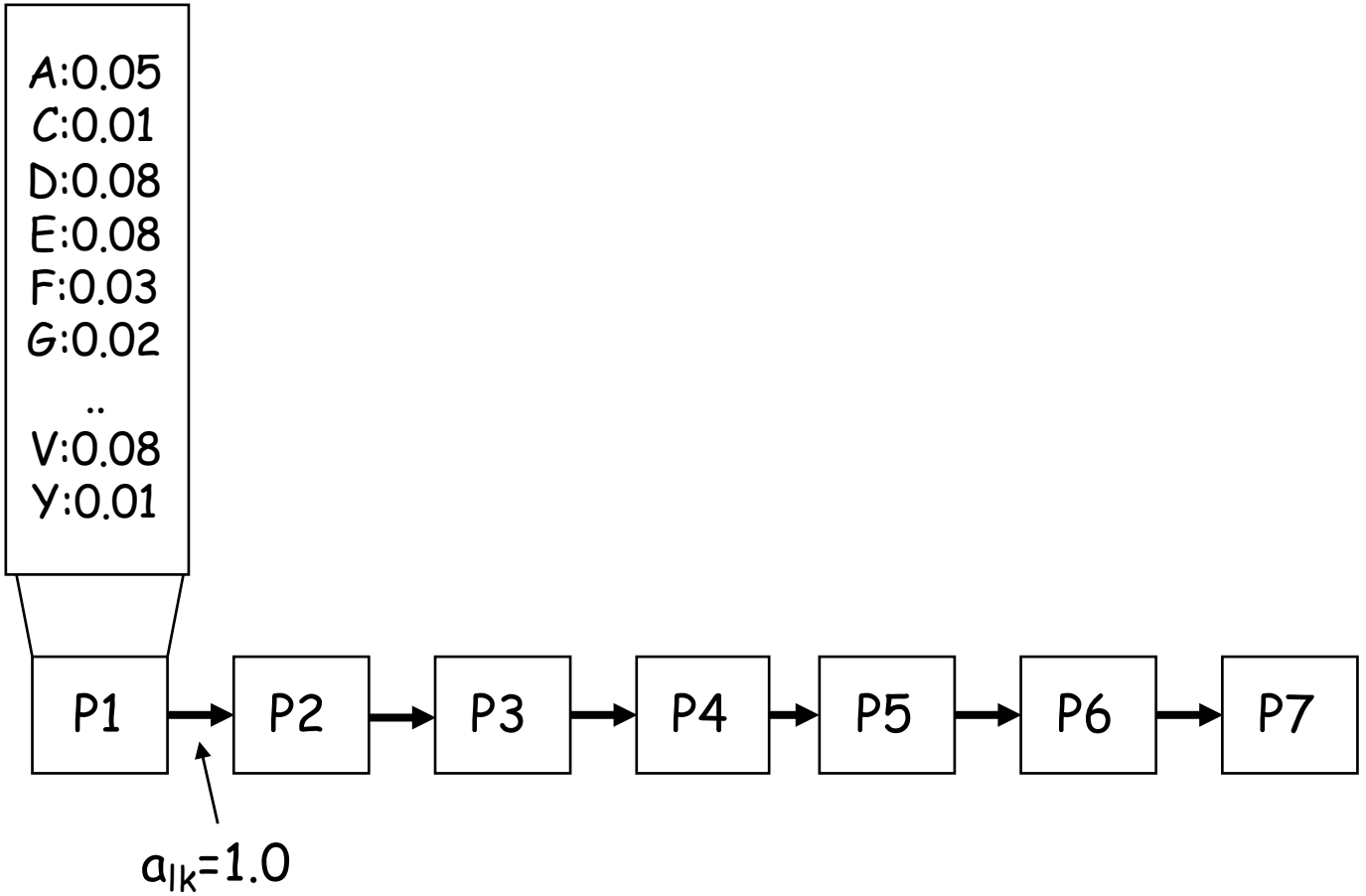
All P/D pairs must be visited once

$L_1 - Y_2A_3V_4R_5 - I_6$

$P_1D_2P_3P_4I_4P_5D_6P_7$

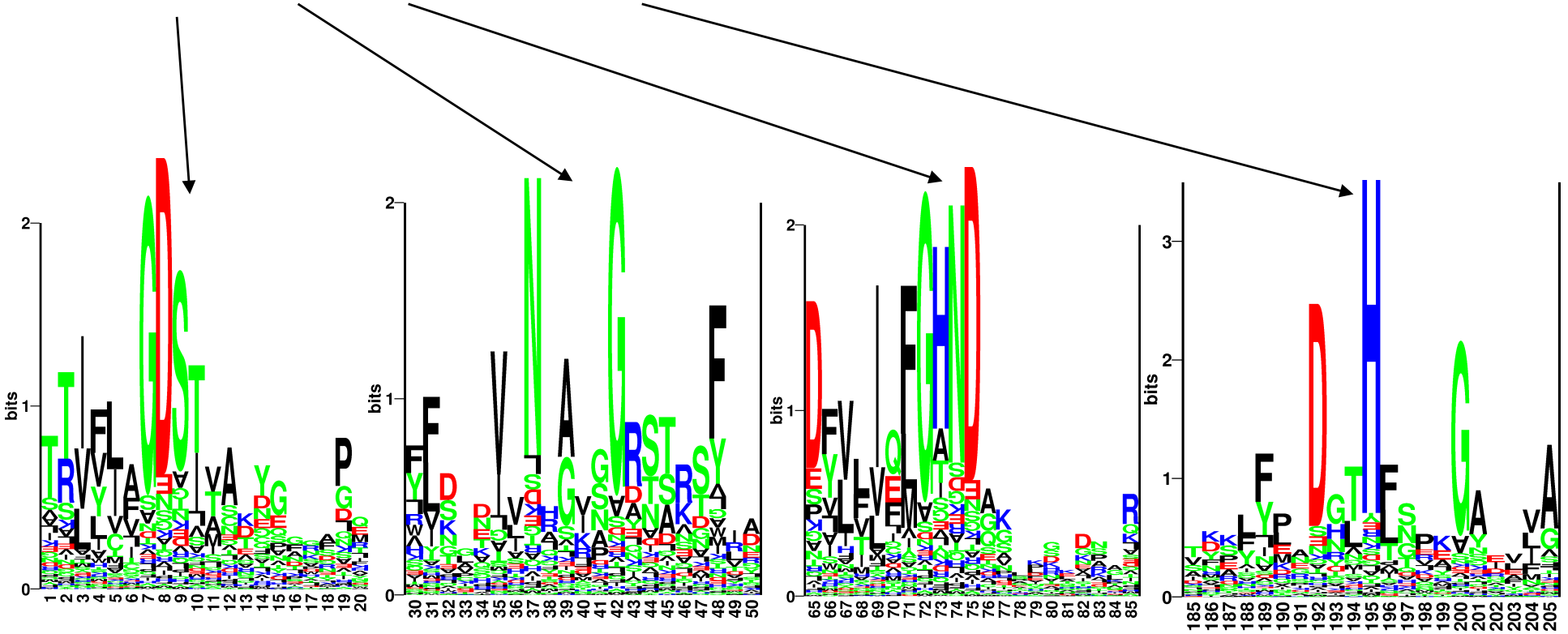
Profile HMM

- Un-gapped profile HMM is just a sequence profile



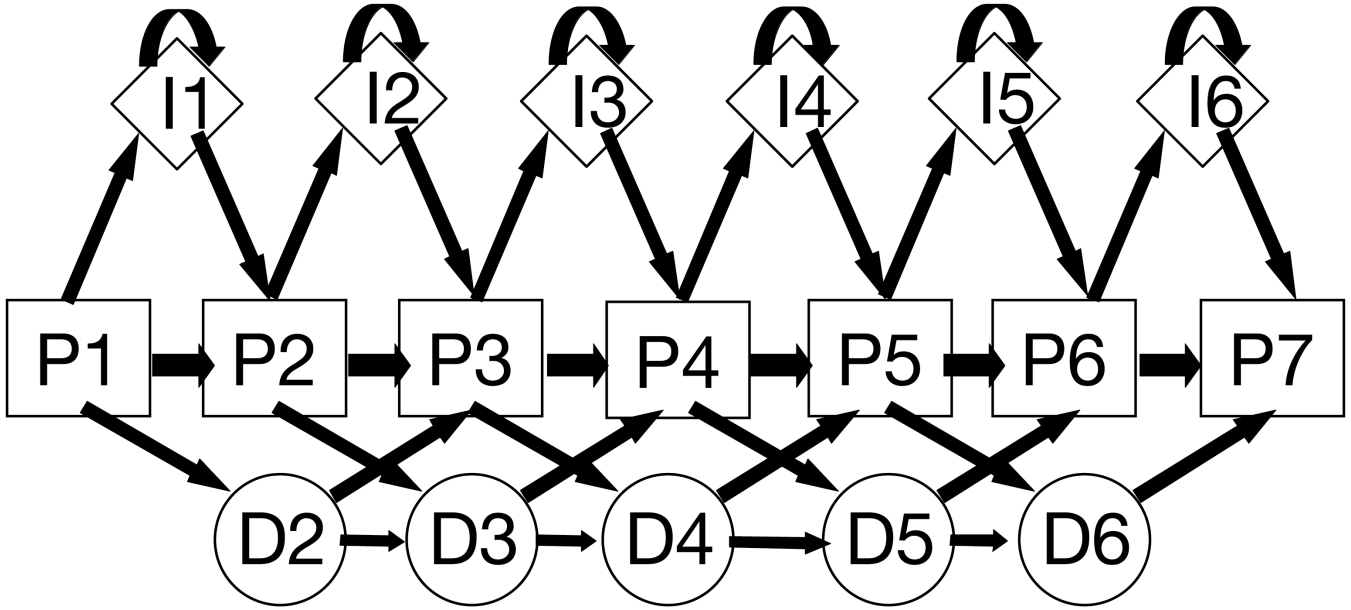
Example. Where is the active site?

- Sequence profiles might show you where to look!
- The active site could be around
  - S9, G42, N74, and H195



Profile HMM

- Profile HMM (deletions and insertions)



Profile HMM (deletions and insertions)

The HMMer program

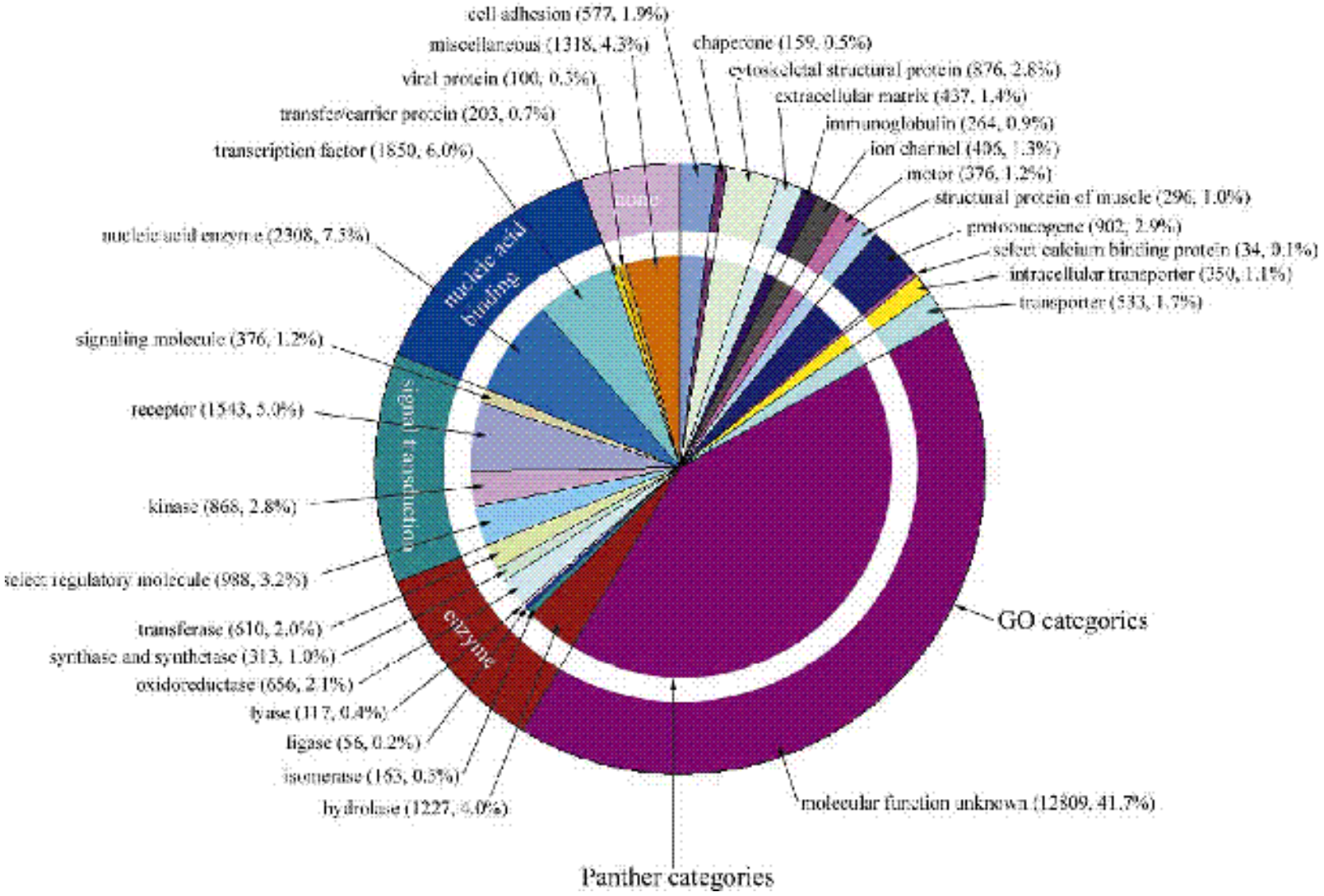
- HMMer is a open source program suite for profile HMM for biological sequence analysis
- Used to make the Pfam database of protein families
  - http://pfam.sanger.ac.uk/

Protein homology modeling and sequence profiles

Background. Why protein modeling?

- Because it works!
  - Close to 50% of all new sequences can be homology modeled
- Experimental effort to determine protein structure is very large and costly
- The gap between the size of the protein sequence data and protein structure data is large and increasing

Homology modeling and the human genome



Is it really impossible?

Protein homology modeling is only possible if %id greater than 30-50%

WRONG!!!!!!!

Why %id is so bad!!

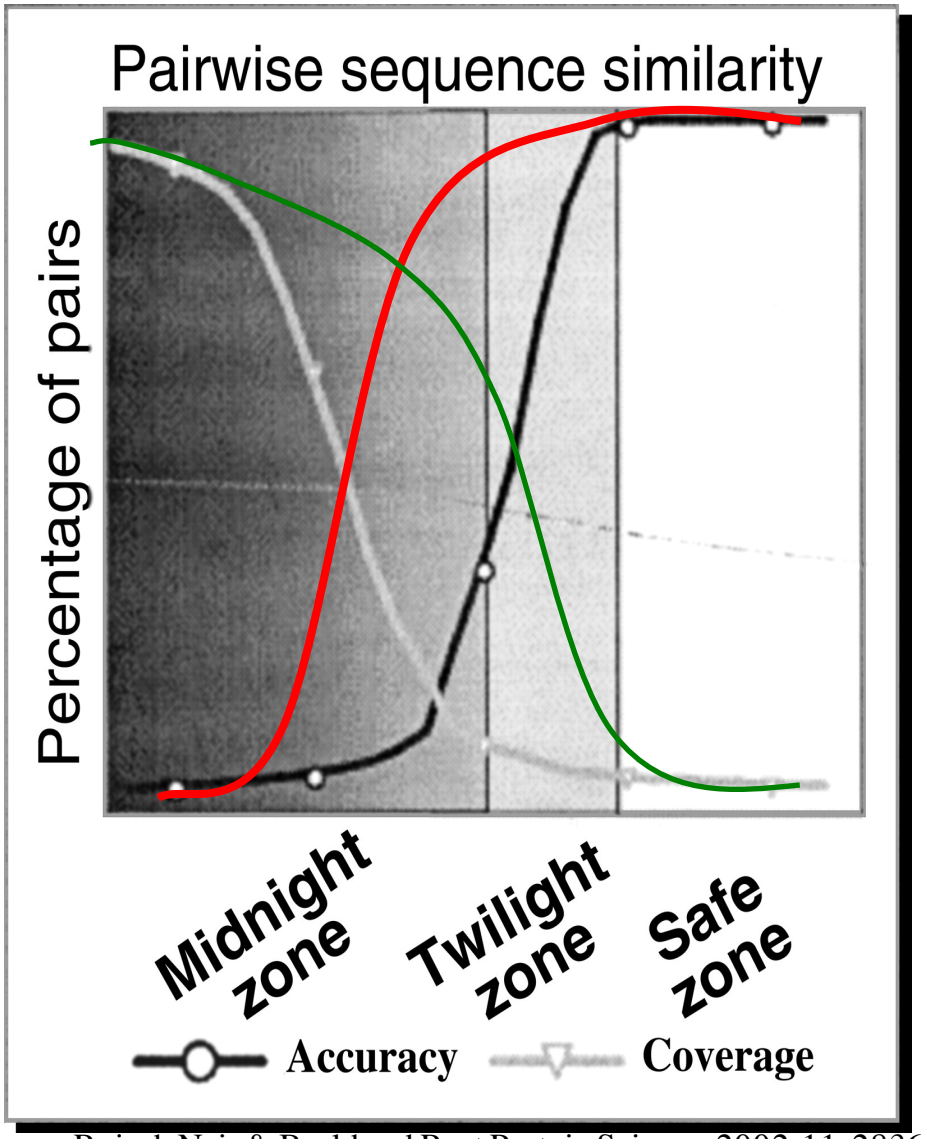
1200 models sharing 25-95% sequence identity with the submitted sequences (www.expasy.ch/swissmod)

Probabilities of SWISS-MODEL accuracy for target-template identity classes.

Percent sequence identity[a]	Total number of models[b]	Percent[c] models with rmsd lower than 1 Å	Percent models with rmsd lower than 2 Å	Percent models with rmsd lower than 3 Å	Percent models with rmsd lower than 4 Å	Percent models with rmsd lower than 5 Å	Percent models with rmsd higher than 5 Å
25-29	125	0	10	30	46	67	33
30-39	222	0	18	45	66	77	23
40-49	156	9	44	63	78	91	9
50-59	155	18	55	79	86	91	9
60-69	145	38	72	85	91	92	8
70-79	137	42	71	82	85	88	12
80-89	173	45	79	86	94	95	5
90-95	88	59	78	83	86	91	9

a: Range of sequence identity between target and template sequence.
b: Total number of models in any given class of sequence identity. The table summarises 1201 model – control structure pairs.
c: Probability in percent that a model, sharing X% sequence identity with its template, deviates by 1 Å or less from the corresponding experimental control structure. The following columns provide these probabilities for other rms deviations.

Identification of fold



- If sequence similarity is high, proteins share structure (Safe zone)
- If sequence similarity is low, proteins may share structure (Twilight zone)
- Most proteins do not have a high sequence homologous partner

Rajesh Nair & Burkhard Rost Protein Science, 2002, 11, 2836-47

How can we do it?

- Identify template(s) – initial alignment
  - Can give you the protein function
- Improve alignment
  - Can give you the active site
- Backbone generation
- Loop modeling
  - Most difficult part
- Side chains
- Refinement
- Validation

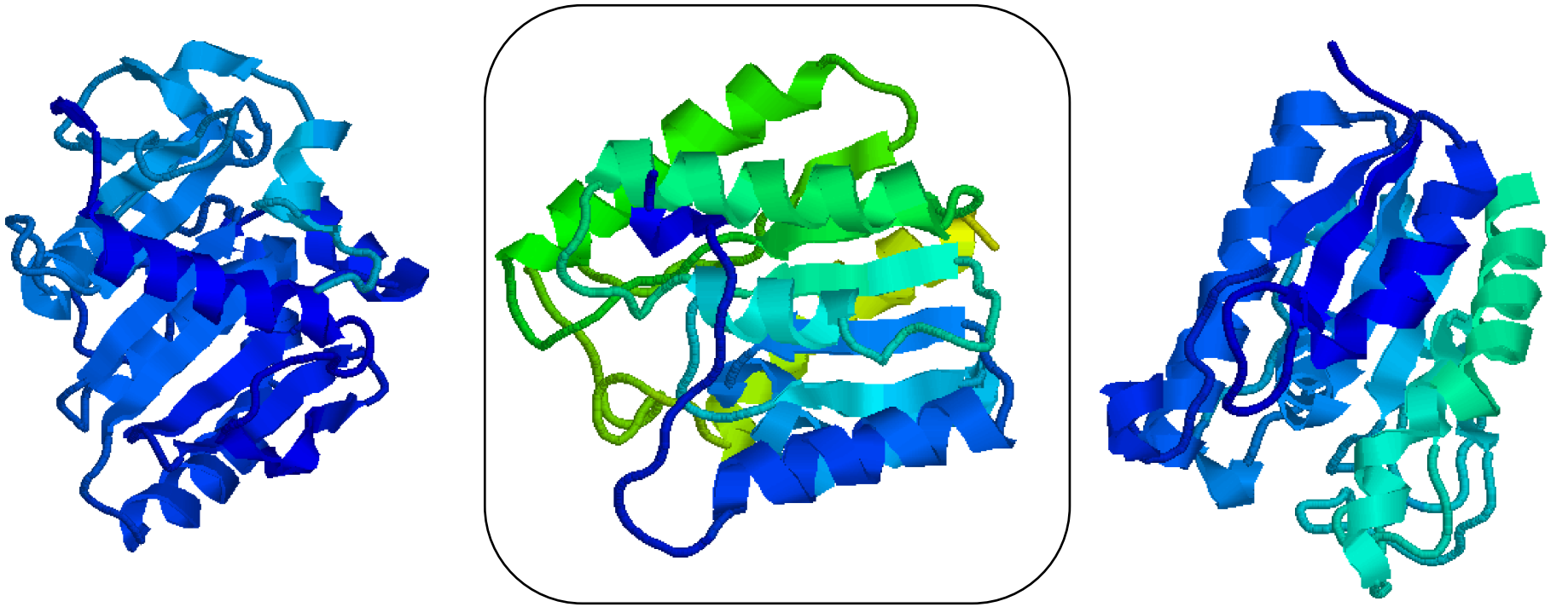
How to do it?

- Identify fold (template) for modeling
  - Find the structure in the PDB database that resembles your query protein the most
- Can be used to predict function
  - and maybe active sites
- Align protein sequence to template
  - Simple alignment methods
  - Sequence profiles
- Model side chains and loops

How to do it?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

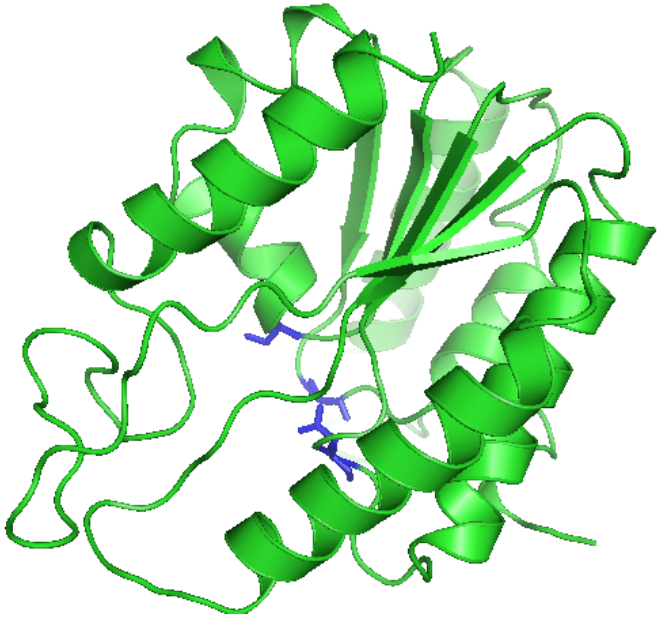
>1K7C.A
TTVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADV
VTAGDYVIVEFGHNDGGSLSTDNGRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKL
FTAKGAKVILSSQTPNNPWETGTFVNSPTRFVEYAELAAEVAGVEYVDHWSYVDSIYETL
GNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSLKSVLTTTSFEGTCL



How to do it?

>1K7C.A
TTVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADV
VTAGDYVIVEFGHNDGGSLSTDNGRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKL
FTAKGAKVILSSQTPNNPWETGTFVNSPTRFVEYAELAAEVAGVEYVDHWSYVDSIYETL
GNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSLKSVLTTTSFEGTCL

```
1K7C.A TVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADVVTAGDYVIVEFGHN
               S                                      G                          N
1WAB._ EVVFIGDSLVQLMHQCE---IWRELFS---PLHALNFGIGGDSTQHVLW--RLENGELEHIRPKIVVVWVGTN
```

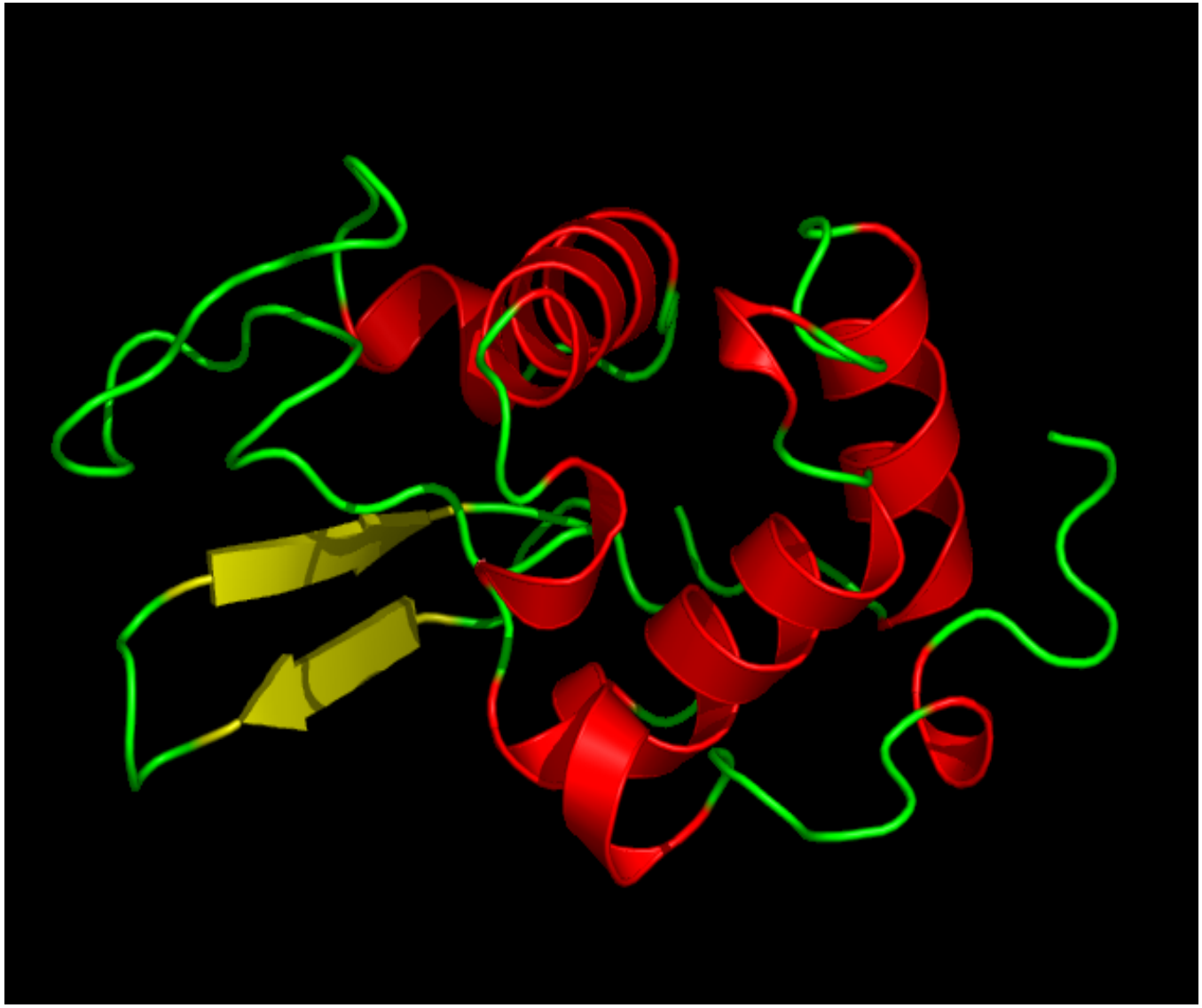


Including structure

- Sequence within in a protein superfamily share remote sequence similarity
- , but they share high structural similarity
- Structure is known for template
- Predict structural properties for query
  - Secondary structure
  - Surface exposure
- Position specific gap penalties derived from secondary structure and surface exposure

Predicting local Protein Structure

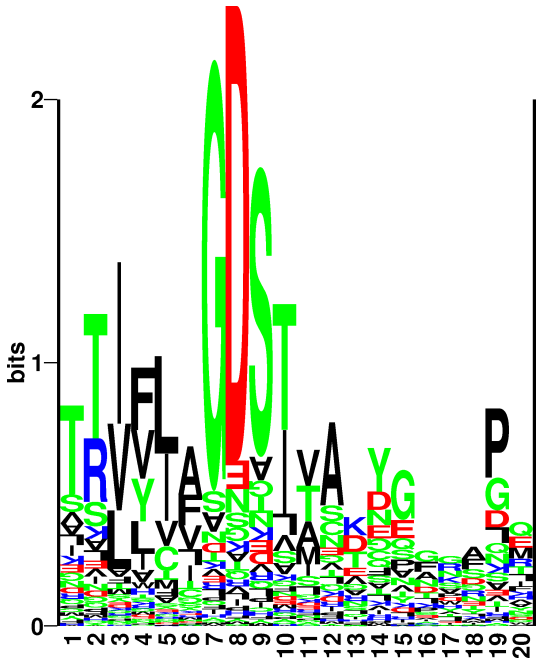
α+β: Lysozyme (1jsf)

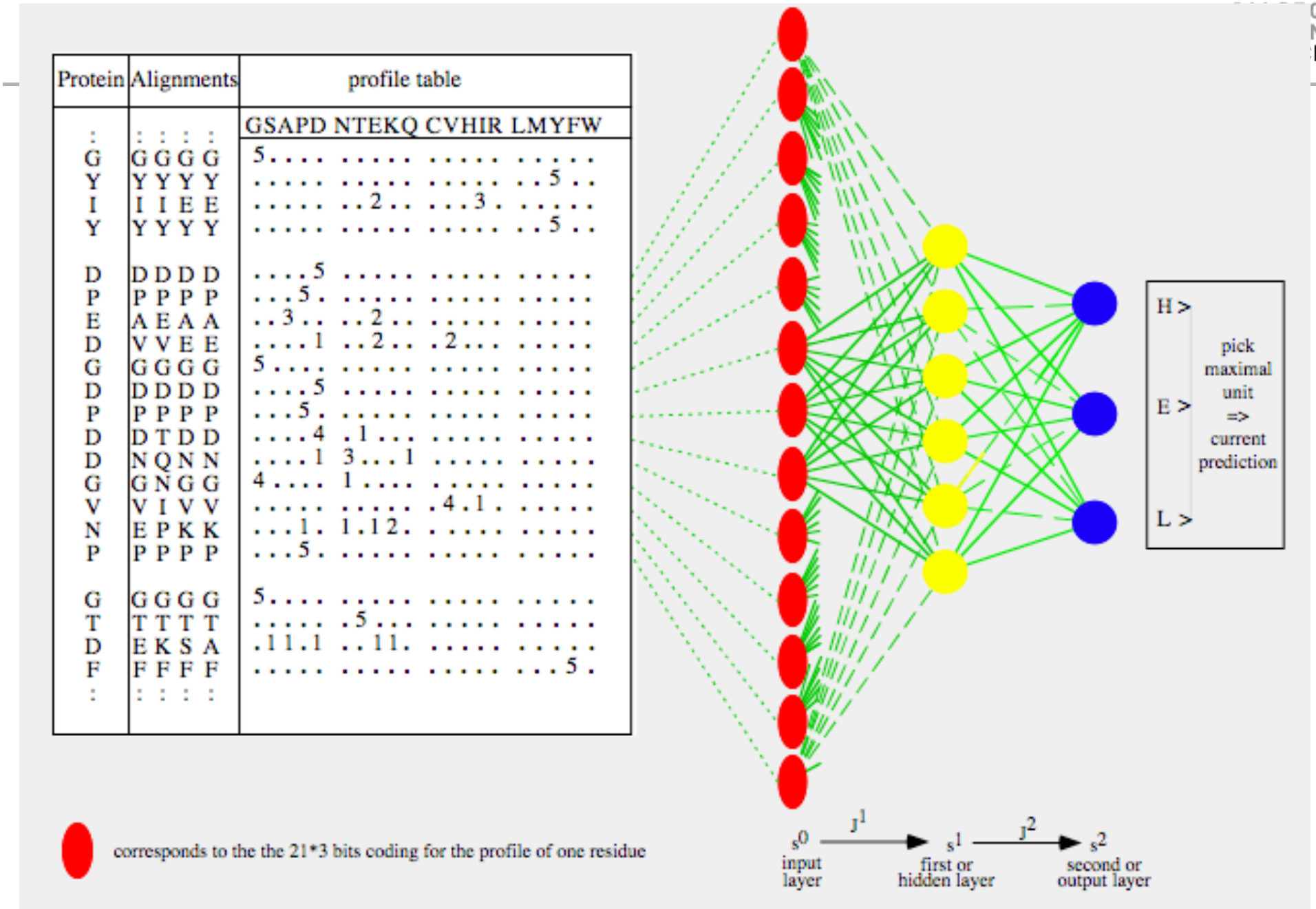


Improvement of accuracy

1974 Chou & Fasman	~50-53%
1978 Garnier	63%
1987 Zvelebil	66%
1988 Quian & Sejnowski	64.3%
1993 Rost & Sander	70.8-72.0%
1997 Frishman & Argos	<75%
1999 Cuff & Barton	72.9%
1999 Jones	76.5%
2000 Petersen et al.	77.9%

Sequence profiles

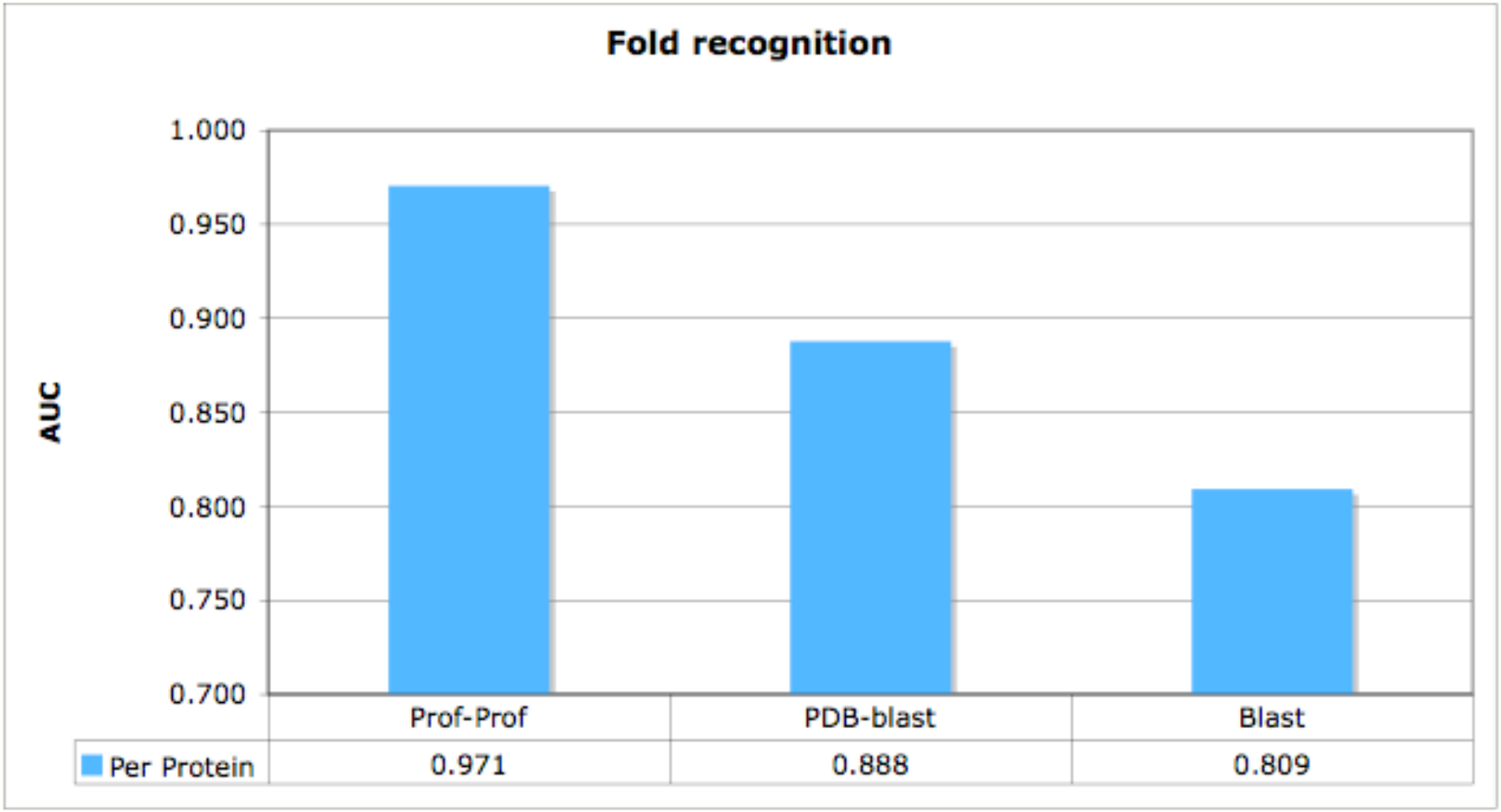




Slide courtesy by B. Rost 2004

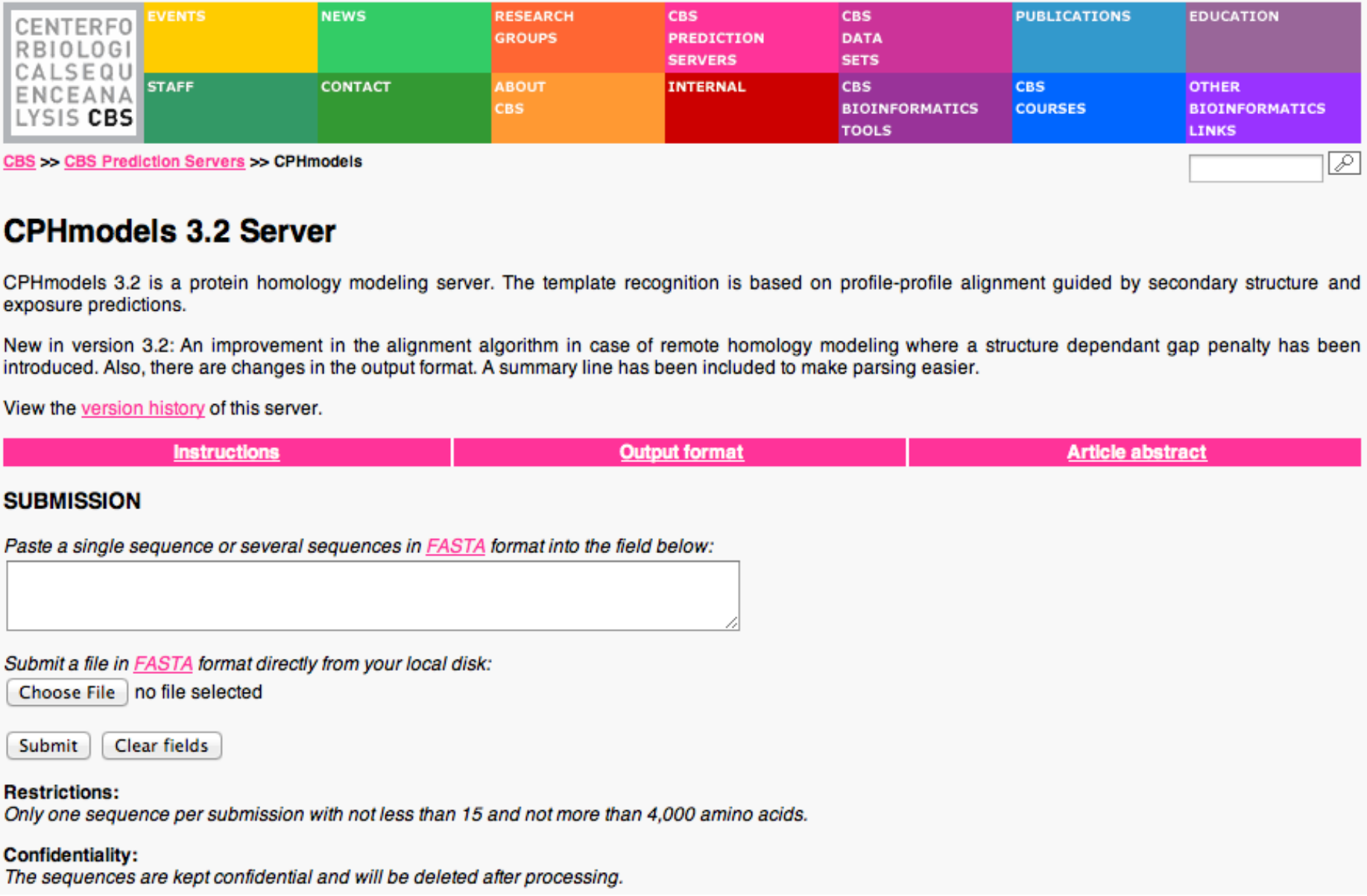
CpHModel - Fold recognition performance

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



CpHModels 3.2

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



Take home message

- Identifying the correct fold is only a small step towards successful homology modeling
- You **can** do reliable fold recognition **AND** homology modeling even for low sequence homology
- Use sequence profiles and local protein structure (predictions) to align sequences