"... each with its own beauty, and each with a story to tell."
-- Stephen Jay Gould

Homology:
$P [a_i, a_j \mid H] = q_{ij}$ = (frequency of a_i, a_j pairs in homologous protein alignments)

Random sequence:
$P [a_i, a_j \mid R] = p_i\, p_j$ = (frequency of a_i) * (frequency of a_j)

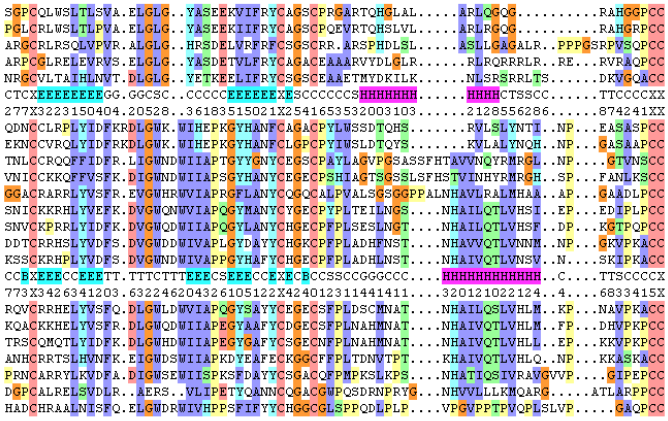
The maximum local alignment score (similarity) is the alignment that maximizes the log odds ratio of H vs R:

$$s(x,y) = \log (q_{xy} / p_x\, p_y)$$

The logarithm is necessary for an additive scoring scheme.
Each column of alignment is independent

Simple maximum likelihood estimate.

In principle, given a large set of confirmed alignments we can calculate: q_{xy} , p_x , and p_y



However,

- This simple approach for calculating a scoring matrix has two problems:

1. It is difficult to obtain a good random sample of protein sequences.
2. This approach does not take into account the effect of evolutionary distances.
   - Short evolutionary distance -> small q_{xy}
   - Large evolutionary distance -> q_{xy} is same p_x and p_y

Deriving scoring parameters

- Maximum likelihood estimate
  - From a set of known good alignments
- Appropriate homology model depends on evolutionary distance of protein sequences
  - (Many) different scoring matrices.

The optimal (local) alignment using the wrong scoring matrix might tell a very implausible evolutionary story for your sequences.

PAM (point accepted mutation)

- Dayhoff, Schwartz, & Orcutt (1978)
- Identify substitution matrix for closely related (easy to determine) alignments
- Extrapolate to longer evolutionary distances

Not estimating joint P_{ab} but rather $P(b|a, t)$!

Goal was to derive a matrix for which expectation:

$$\Sigma_{a,b}\, p_a p_b P(b|a, t=1) = 0.01$$

i.e. 1% expected number mutations which define as t = 1.

Expected score?

$$E(X) = \sum_{x=1}^{Z} x_i p_{x_i}$$

The ***expected value*** of a random variable:

- intuitively, is the long-run average value of repetitions of the experiment it represents.
- measure of the center of the distribution of the variable.
- is the probability-weighted average of all possible values.

$$E(s_{a,b}) = \Sigma_{a,b}\, p_a p_b s(x,y)$$

Not estimating joint P_{ab} but rather $P(b|a, t)$!

Goal was to derive a matrix for which expectation:

$$\Sigma_{a,b}\, p_a p_b P(b|a, t=1) = 0.01$$

i.e. 1% expected number mutations which define as t = 1.

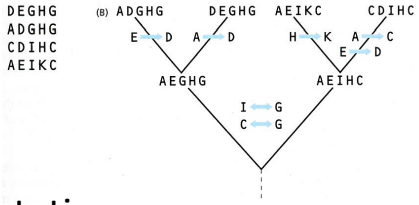
Scaled and rounded => PAM_1

$$PAM_n = (PAM_1)^n$$

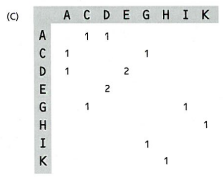
We will not formally derive the PAM matricies in this class, but it really isn't *that* hard …

The basic idea

- High confidence alignments are built relative to an evolutionary tree:



- Acceptable point mutations are tallied from the tree:



Mutational probability matrix derived by Dayhoff for the 20 amino acids

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

For clarity, the values have been multiplied by 10000

This matrix corresponds to an evolution time period giving 1 mutation/100 amino acids, and is refered to as the **PAM1 matrix.**

$P(b|a, t=1)$

Source: Dayhoff, 1978

Note that we convert this into our log-odds scoring scheme by:

$$s(x,y) = \log (q_{xy} / p_x p_y)$$

PAM matrix was: $P(B|A, t= 1)$

Recall: $p(A)p(B|A) = P(AB)$

Log_2 is often used to make the scores represent information content (bits).

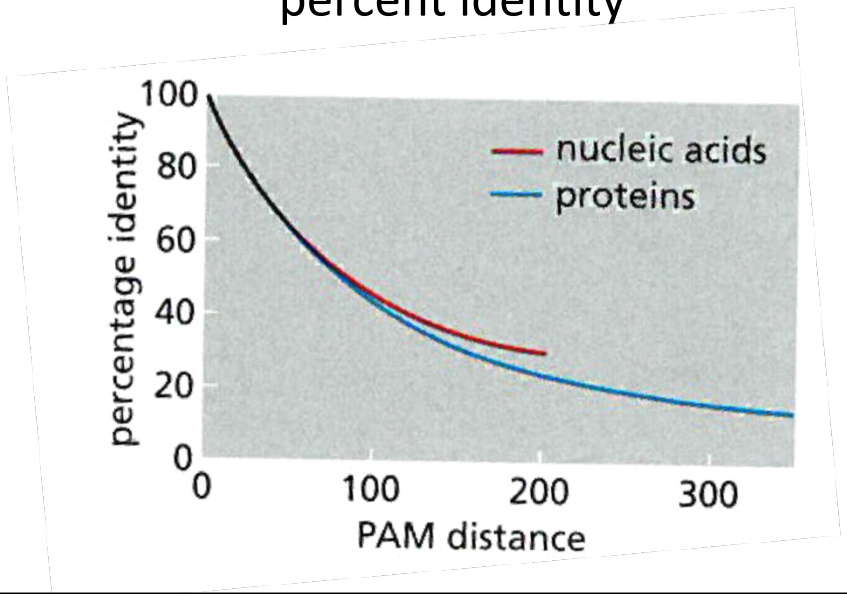
Therefore …

$$S(a,b) = c \log (p(b|a, t=1) / p_b)$$

Why a c? To account for the evolutionary distance – scaling!

The resulting matrix is, in fact, symmetrical.

Relationship between PAM and percent identity



PAM summary

- The scores derived through the PAM model are an accurate description of the information content (or the relative entropy) of an alignment (Altschul, 1991).
- PAM-1 corresponds to about 1 million years of evolution
- PAM-250 is the traditionally most popular matrix

PAM250 log odds scoring matrix

$$\frac{3}{\log 2} s^n(a,b); n = 250$$

values rounded to the nearest integer

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

PAM limitations

- Small dataset for derivation of PAM.
- Substitution data is calculated for *small evolutionary distance and extrapolated to longer times*.
- Raising $PAM_1(a,b)$ to a higher power, to give for instance a PAM250 matrix does not capture the true difference between short time substitutions and long term ones [Gonnet, 1992].

Deriving scoring parameters

- Maximum likelihood estimate
  - From a set of known good alignments
- Appropriate homology model depends on evolutionary distance of protein sequences
  - (Many) different scoring matrices.

The BLOSSUM approach leverages **both** strategies !!!

BLOSUM matrices

- Henikoff & Henikoff (1992)
- **Blo**cks **Su**bstitution **M**atrix. Scores for each position are obtained frequencies of substitutions in blocks of local alignments of protein sequences.
- Motivation: At increasingly longer times, mutational process observed wasn't well represented in scaled PAM matricies.

18

Conserved blocks in alignments

```
AKLGGREAVEAAVDKFYNKIVADPTVSTYFSNTDMKVQRSKQFAFLAYALG   GAHFQAVARHLSDTLTELGV
AKLGGREAVEAAVDKFYNKVVADPTVSVFFSKTDMKVQRSKQFAFLAYALG   GAHFQAVVRHLSDTLAELGV
DKIGGHEAIEVVVEDFYVRVLADDQLSAFFSGTNMSRLKGKQVEFFAAALG   GPHFSLVAGHLADALTAAGV
DNIGGQPAIEQVVDELHKRIATDSLLAPVFAGTDMVKQRNHLVAFLAQIFE   GPHFDAIAKHLGERMAVRGV
DNIGGQPAIEQVVDELHKRIATDSLLAPIFAGTDMAKQRNHLVAFLGQIFE   GPHFDAIAKHLGEAMAVRGV
EKLGGTTAVDLAVDKFYERVLQDDRIKHFFADVDMAKQRAHQKAFLTYAFG   GTHFDAVAEDLLATLKEMGV
EQLGGQAAVQAVTAQFYANIQADATVATFFNGIDMPNQTNKTAAFLCAALG   GPQFTTVIGHLRSALTGAGV
EKLGGENAMKAAVPLFYKKVLADERVKHFFKNTDMDHQTKQQTDFLTMLLG   GPHFDAIIENLAATLKELGV
EKLGGQAAMHAAVPLFYKKVLADDRVKHYFKNTNMEHQAKQQEDFLTMLLG   GPHFDAIIENLAATLKELGV
YEAIGEELLSQLVDTFYERVASHPLLKPIFPSDLTETARKQKQFLTQYLGG   PPRADAWLSCMKDAMDHVGL
EQLGGEAAVHAVTTQFYANIAADATVANFFNGINMPTQTDKTAAFLCAALG   GPQFTTVIGHLRSALTGAGV
EQLGGEAAVTAVTTQFYANIQADATVANFFNGINMADQTNKTASFLCAALG   GPQFTTVIGHLRSALTSAGV
```

Yes, there is a little circularity here − calculating alignment scores from alignments!

19

Collecting substitution statistics

1. Count amino acids pairs in each column; e.g.,
   - 6 AA pairs, 4 AB pairs, 4 AC, 1 BC, 0 BB, 0 CC.
   - Total = 6+4+4+1=15
2. Normalize results to obtain probabilities (p_X's and q_{xy}'s)
3. Compute log-odds score matrix from probabilities:
   $s(x,y) = \log (q_{xy} / (p_X p_y))$

```
A
A
B
A
C
A
```

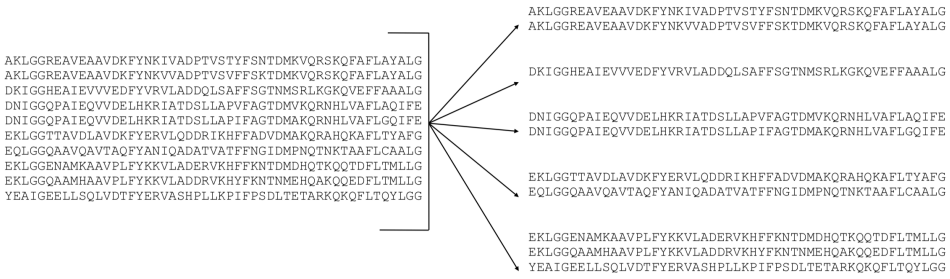
20

Constructing BLOSUM *r*

- To avoid bias in favor of a certain protein, first eliminate sequences that are more than *r*% identical
- The elimination is done by either
  - removing sequences from the block, or
  - finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster.
- BLOSUM *r* is the matrix built from blocks with no more the *r*% of similarity
  - E.g., BLOSUM62 is the matrix built using sequences with no more than 62% similarity.
  - Note: BLOSUM 62 is the default matrix for protein BLAST

21

Cluster sequences by L% identity



BLOSUM 62

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

Blosum62 scoring matrix

$$\frac{2}{\log 2} s(a,b)$$

values rounded to the nearest integer

So what do the scores mean?

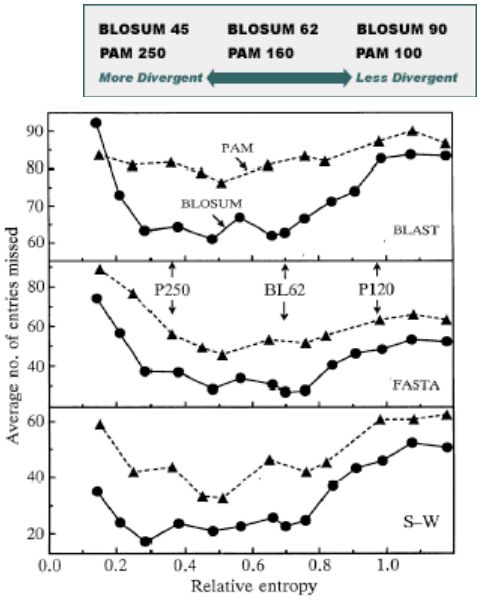
- Positive scores: The given amino acid pair is *more likely to occur* in an alignment than by chance.
- Negative scores: amino acid pair is *less likely to occur* than by chance.

```
GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLP
G  F+ G CP      +FD+  + G W+EI K+P
GQNFHLGKC                        P

LENENQGKC                        E
   E +G C                        E
ASFE-KGNC                        E

YMEGDL  A                        Q
 +  D  E++PD          KQ       K
VL--DKELSPDGTMNQVKGEAKQSNVSEPAKLEV

RVVNLVP----WVLATDYKNYAINYNCD-----Y
+   L+P    W+LATDY+NYA+ Y+C       +
QFFPLMPPAPYWILATDYENYALVYSCTTFFWLF

HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT
H D        WIL ++  L   T   + ++L
HVD------FFWILGRNPYLPPETITYLKDILT-
```

Y mutates to V receives -1
M mutates to L receives 2
E gets deleted receives -10
G gets deleted receives -10
D matches D receives 6
Total score = -13

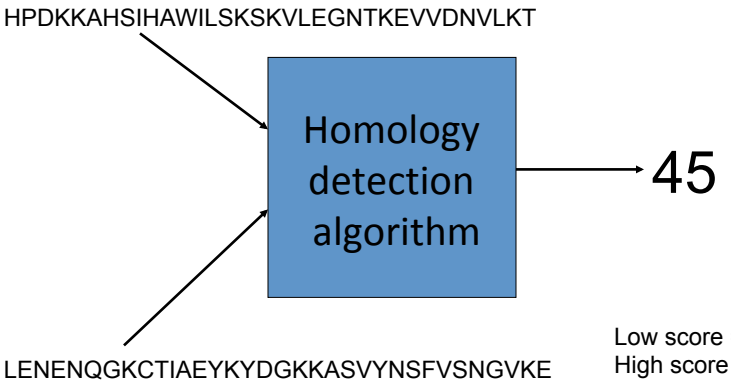
Comparison

- PAM is based on an evolutionary model using phylogenetic trees
- BLOSUM assumes no evolutionary model, but rather empirical from conserved "blocks" of proteins



Henikoff and Henikoff (1992)

Significance of scores: Karlin Altshul Statistics



Low score = unrelated
High score = homologs

How high is high enough?