Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

- PAM matrices, Dayhoff *et al*, 1978
- BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

Overall strategy for both PAM and BLOSUM

1. Trusted amino acid alignments
2. Obtain amino acid pair counts $(A_{xy}^N)$with corrections for
   - Evolutionary divergence
   - Sample biases
3. Estimate substitution frequencies, q_{xy}^N, from pair counts, A_{xy}^N
4. Log odds substitution matrix: $S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$

Log odds substitution matrices

Two sequences have *N* PAMs divergence , if, on average, *N* amino acid replacements per 100 residues occurred since their separation

$$S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$$

Frequency of x aligned with y in sequences with divergence *N*

Frequency of x aligned with y in "random" sequences

Scaling constant

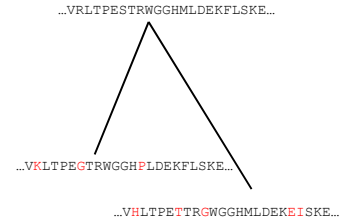
PAM: A unit of evolutionary divergence

- PAM: Percent Accepted Mutation
  - *Accepted Mutations* are mutations that are retained and passed on to future generations
- We say the divergence between two sequences is *N* PAMs, if, on average, *N* amino acid replacements per 100 residues (including multiple substitutions) occurred since their separation.

Measures of sequence divergence

It is useful to have a measure of divergence. However, we cannot observe PAMs directly.

What actually occurred:

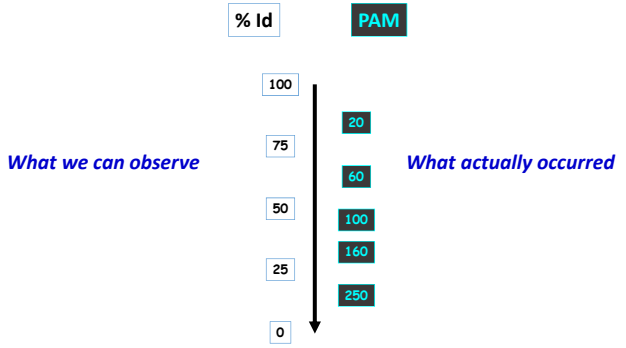


***N* replacements per 100 residues**

What we can observe:

…VKLTPEGTRWEGHPLDEKFLSKE…
…VHLTPETTRGGGHMLDEKEISKE…

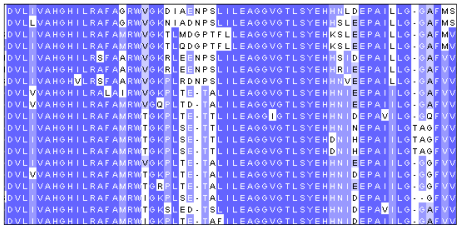
% identity



1. Trusted multiple sequence alignments

Examined 1572 changes in 71 groups of closely related proteins

At least 85% identical



2. Obtain amino acid pair counts (A_{xy}) with corrections for evolutionary divergence and sample biases

Counting amino acid pairs on a tree:

For each unrooted tree with *k* leaves

- ➢ Select the tree(s) that require the fewest substitutions to explain the data
- Count amino acid pairs on the branches of the tree

3. Estimate substitution frequencies, q_{xy}^N, from pair counts, A_{xy}

- Markov model with 20 states (A, C, D, E ... Y)
- Estimate 1 PAM transition matrix P^1 from A_{xy}
- N-PAM transition matrix: $P^1 = (P^1)^N$
- $q_{xy}^N = p_x P_{xy}^N$
- $S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$

Is P_{xy}^N a symmetric matrix? No. (Check this algebraically).

Is $S^N[x,y]$ a symmetric matrix? Yes (Check this algebraically).

Zvelebil, Baum, Understanding bioinformatics.

PAM 120

Positive score – frequency of substitutions is greater than would have occurred by random chance.

Zero score – frequency is equal to that expected by chance.

Negative score – frequency is less than would have occurred by random chance.

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

small, polar

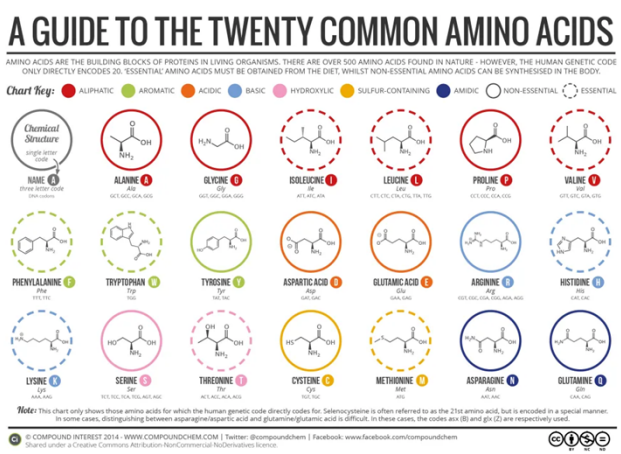
small, nonpolar

polar or acidic

basic

large, hydrophobic

aromatic



Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

- PAM matrices, Dayhoff *et al*, 1978

➢BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

BLOSUM Matrices

- Trusted data
  - 2000 blocks of conserved regions in ~500 groups of proteins
- Count amino acid pairs: A_{xy}^N
  - Parameterize by evolutionary distance, *N*
  - Correct for sample bias
- Calculate amino acid frequencies:
  - Related pairs: q_{xy}^N
  - Background pair frequencies calculated from blocks: E_{xy}
- Log likelihood scoring matrix
  - $S^N = 2\ log_2 \frac{q_{xy}^N}{E_{xy}}$

1. Trusted multiple sequence alignments

~2000 blocks representing 500+ groups of proteins

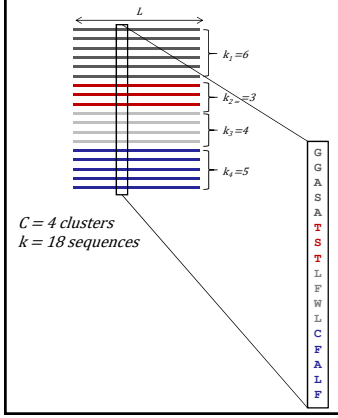
2. Count amino acid pairs: A_{xy}^N

Parameterize by evolutionary distance, *N*

Correct for sample bias

- Cluster sequences such that if *s1* and *s2* are in different clusters, then *identity(s1, s2) < N%*
- Count amino acid pairs in *s1* aligned with *s2* *only* if *s1* and *s2* are in different clusters
- Normalize for cluster size

Normalizing for cluster size



- Clusters C_i & C_j contribute $k_i \cdot k_j$ comparisons per column.
- Normalize counts so each cluster pair contributes equally.
- Sum over all cluster pairs and all columns.

Cluster counts in Column *l*:

$$A_{xy}^N(l) = \frac{B_l(C_i,x)B_l(C_j,y) + B_l(C_i,y)B_l(C_j,x)}{k_i k_j}$$

where $B_l(C_i, x)$ = the number *x's* in column *l* of cluster C_i

Example: LF pairs in Clusters 3 and 4

$B_l(C_3,F)=1$ $B_l(C_3,L)=2$ $B_l(C_4,F)=2$ $B_l(C_4,L)=1$

$$A_{FL}^N(l) = \frac{1 \cdot 1 + 2 \cdot 2}{4 \cdot 5} = \frac{1}{4}$$

An example...

BLOSUM clustering example

1: KKRK
2: KKKK
3: KNRN
4: NRNR
5: KNKN
6: KRNR

Percent Sequence Identity

	[2]	[3]	[4]	[5]	[6]
[1]	0.75	0.5	0	0.25	0.25
[2]		0.25	0	0.5	0.25
[3]			0	0.75	0.25
[4]				0	0.75
[5]					0.25

Unclustered sequences: Every sequence is at least 25% identical

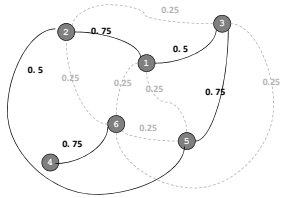
Percent Sequence Identity

	[2]	[3]	[4]	[5]	[6]
[1]	**0.75**	**0.5**	0	0.25	0.25
[2]		0.25	0	**0.5**	0.25
[3]			0	**0.75**	0.25
[4]				0	**0.75**
[5]					0.25

1: KKRK
2: KKKK
3: KNRN
5: KNKN

< 45% identical

4: NRNR
6: KRNR



< 65% id

1: KKRK
2: KKKK

< 65% id

3: KNRN
5: KNKN

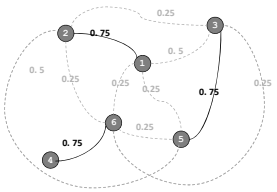
< 65% id

4: NRNR
6: KRNR

< 65% id

Percent Sequence Identity

	[2]	[3]	[4]	[5]	[6]
[1]	**0.75**	0.5	0	0.25	0.25
[2]		0.25	0	0.5	0.25
[3]			0	**0.75**	0.25
[4]				0	**0.75**
[5]					0.25



Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

- PAM matrices, Dayhoff *et al*, 1978
- BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

Similarities and differences between PAM and BLOSUM

	PAM	BLOSUM
Evolutionary model	Explicit evolutionary model	None
Data	Full length MSAs of closely related sequences.	Conserved blocks. i.e., ungapped local MSAs
Bias correction	Trees	Clustering
Multiple substitutions	Markov model: $P^n = (P^1)^n$	Implicitly represented in data (clustering)
Evolutionary distance	Markov model: $P^n = (P^1)^n$	Clustering
Matrices	Transition and log odds scoring matrices	Log odds scoring matrix only.
Parameter *n*	Distance increases with *n*	Distance decreases with *n*
Biophysical properties	Derived indirectly from data	Derived indirectly from data

Comparing PAM and BLOSUM matrices