PAM and BLOSUM substitution matrices

Substitution scoring matrices

There are two main families of amino acids substitution scoring matrices:

- **PAM** substitution matrices

  based on the rate of divergence between sequences

- **BLOSUM** substitution matrices

  based on the conservation of domains in proteins

Another popular substitution matrix was proposed by Gonnet et al (1992):

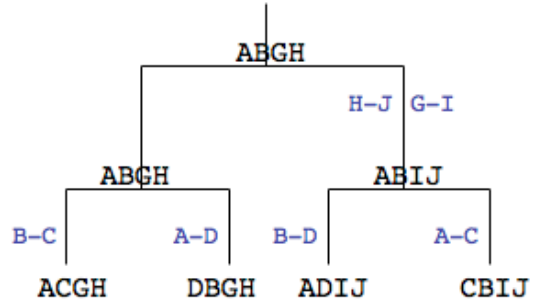
- **GONNET** substitution matrix

  based on an exhaustive sequence alignment analysis

PAM scoring matrices

The substitution score is expected to depend on the rate of divergence between sequences.

accepted mutations



The **PAM matrices** derived by Dayhoff (1978):

- are based on evolutionary distances.
- have been obtained from carefully aligned closely related protein sequences (71 gapless alignments of sequences having at least 85% similarity).

M. Dayhoff

Reference: Dayhoff *et al.* (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

PAM scoring matrices

PAM = Percent (or Point) Accepted Mutation

The PAM matrices are **series of scoring matrices**, each reflecting a certain level of divergence:

PAM = unit of evolution (1 PAM = 1 mutation/100 amino acid)

- PAM1 proteins with an evolutionary distance of 1% mutation/position
- PAM50 idem for 50% mutations/position
- PAM250 250% mutations/position (a position could mutate several times)

Reference: Dayhoff *et al.* (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

Derivation of the PAM matrices

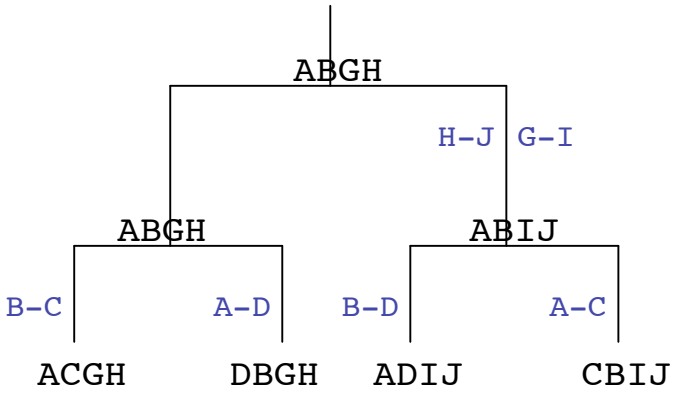
To illustrate how the PAM substitution matrices have been derived, we will consider the following artificial ungapped aligned sequences:

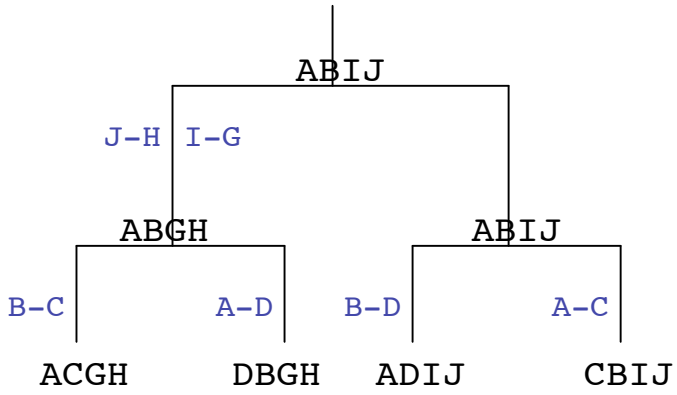
```
A C G H
D B G H
A D I J
C B I J
```

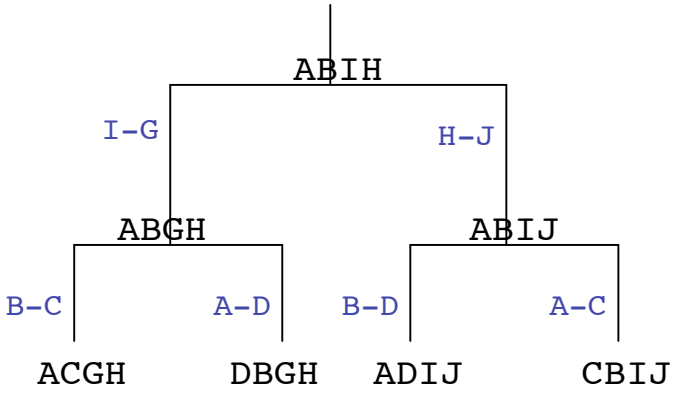
Reference: Borodovsky & Ekisheva (2007) Problems and Solutions in Biological sequence analysis. *Cambridge Univ Press*.

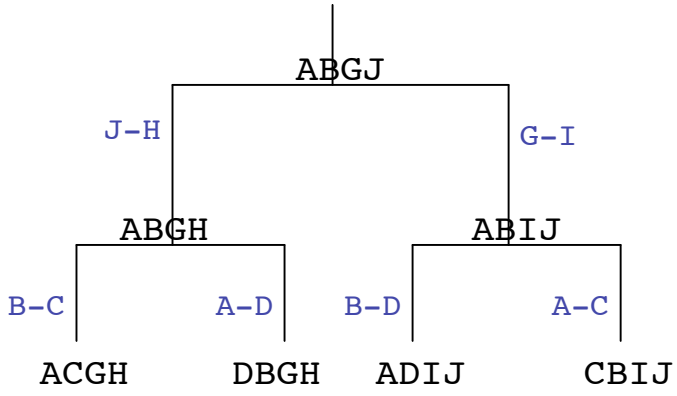
Derivation of the PAM matrices

Phylogenetic trees (maximum parsimony)









Here are represented the four more parsimonious (minimum of substitutions) phylogenetic trees for the alignment given above.

Derivation of the PAM matrices

Matrix of accepted point mutation counts (A)

	A	B	C	D	G	H	I	J
A		0	4	4	0	0	0	0
B	0		4	4	0	0	0	0
C	4	4		0	0	0	0	0
D	4	4	0		0	0	0	0
G	0	0	0	0		0	4	0
H	0	0	0	0	0		0	4
I	0	0	0	0	4	0		0
J	0	0	0	0	0	4	0	

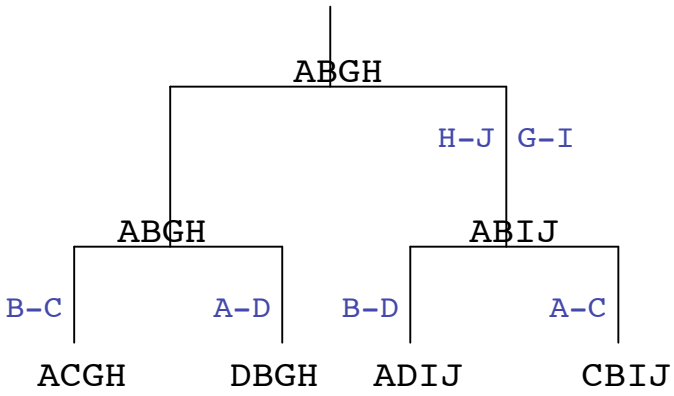
For each pair of different amino acids (i,j), the total number a_{ij} of substitutions from i to j along the edges of the phylogenetic tree is calculated.

(they are indicated in blue on the previous slide)

Derivation of the PAM matrices

Each edge of a given tree is associated with the ungapped alignment of the two sequences connected by this edge.

Thus, any tree shown above generates 6 alignments. For example the first phylogenetic tree generates the following alignments:



A B G H	A B G H	A B G H
A B G H	A B I J	A C G H
A B G H	A B I J	A B I J
D B G H	A D I J	C B I J

Those alignments can be used to assess the "relative mutability" of each amino acid.

Derivation of the PAM matrices

Relative mutability (m_i)

The relative mutability is defined by the ratio of the total number of times that amino acid *j* has changed in all the pair-wise alignments (in our case 6x4=24 alignments) to the number of times that *j* has occurred in these alignments, i.e.

$$m_j = \frac{number\ of\ changes\ of\ j}{number\ of\ occurrences\ of\ j}$$

Relative amino acid mutability values m_j *for our example*

Amino acid	A	B	I	H	G	J	C	D
Changes (substitutions)	8	8	4	4	4	4	8	8
Frequency of occurrence	40	40	24	24	24	24	8	8
Relative mutability m_j	0.2	0.2	0.167	0.167	0.167	0.167	1	1

The relative mutability accounts for the fact that the different amino acids have different mutation rates. This is thus the probability to mutate.

Derivation of the PAM matrices

Relative mutability of the 20 amino acids

aa	m_i	aa	m_i
Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Values according Dayhoff (1978)

The value for Ala has been arbitrarily set at 100.

Trp and **Cys** are less mutable

Cys is known to have several unique, indispensable function (attachment site of heme group in cytochrome and of FeS clusters in ferredoxin). It also forms cross-links such as in chymotrypsin or ribonuclease.

Big groups like Trp or Phe are less mutable due to their particular chemistry. On the other extreme, the low mutability of Cys must be due to its unique smallness that is advatageous in many places.

Asn, **Ser**, **Asp** and **Glu** are most mutable

Although Ser sometimes functions in the active center, it more often performs a function of lesser importance, easily mimicked by several other amino acids of similar physical and chemical properties.

Derivation of the PAM matrices

Effective frequency (f_i)

The notion of effective frequency f_i takes into account the difference in variability of the primary structure conservation in proteins with different functional roles. Two alignment blocks corresponding to 2 different families may contribute differently to f_i even if the number of occurrence of amino acid j in these blocks is the same.

$$\begin{pmatrix} relative\ frequency\ of \\ exposure\ to\ mutation \end{pmatrix} = \begin{pmatrix} average\ composition \\ of\ each\ group \end{pmatrix} \times \begin{pmatrix} number\ of\ mutations\ in \\ the\ corresponding\ tree \end{pmatrix}$$

Derivation of the PAM matrices

Effective frequency (f_i)

The effective frequency is defined as

$$f_j = k \sum_b q_j^{(b)} N^{(b)}$$

where the sum is taken over all alignment blocks b
$q_j^{(b)}$ is the observed frequency of amino acid j in block b,
$N^{(b)}$ is the number of substitutions in a tree built for b
and the coefficient k is chosen the ensure that the sum of the frequences $f_j = 1$.

In our example, there is only one block, therefore the effective frequencies are equal to the compositional frequencies ($f_i = q_j$):

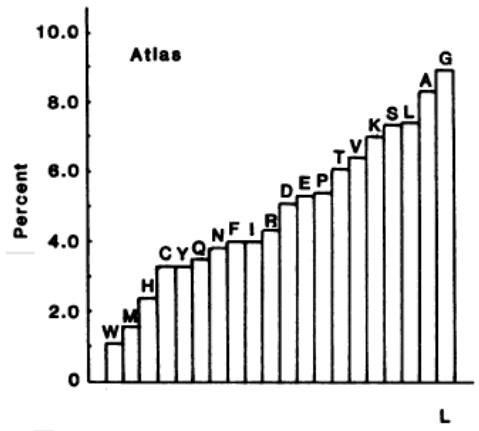
Amino acid	A	B	I	H	G	J	C	D
Frequency *f*	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

Derivation of the PAM matrices

Effective frequency of the 20 amino acids determined for the original alignment data (Dayhoff *et al.*, 1978)

Amino acid	**Gly**	**Ala**	**Leu**	**Lys**	**Ser**	**Val**	**Thr**
Frequency *f*	0.089	0.087	0.085	0.081	0.070	0.065	0.058
Amino acid	**Pro**	**Glu**	**Asp**	**Arg**	**Asn**	**Phe**	**Gln**
Frequency *f*	0.051	0.050	0.047	0.041	0.040	0.040	0.038
Amino acid	**Ile**	**His**	**Cys**	**Tyr**	**Met**	**Trp**	
Frequency *f*	0.037	0.034	0.033	0.030	0.015	0.010	

Source: Dayhoff, 1978



Distribution of amino acids found in 1081 peptides and proteins listed in the *Atlas of Protein Sequence and Structure* (1981).

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? *Science*. 214:149-59.

Derivation of the PAM matrices

Mutational probability matrix (*M*)

Let's define M_{ij} the probability of the amino acid in column *j* having been substituted by an amino acid in row *i* over a given evolutionary time unit.

Non-diagonal elements of M:

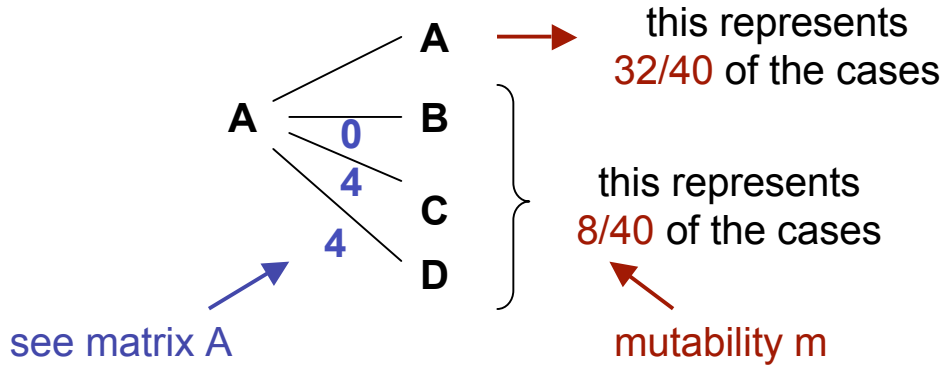
$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}}$$

Diagonal elements of M:

$$M_{ii} = 1 - \lambda m_i$$

In these equations, m is the relative mutability and A is the matrix of accepted point mutations. The constant λ represents a degree of freedom that could be used to connect the matrix M with an evolutionary time scale.

In our example:



If A is mutated, the probability that it is mutated into D is

$A_{DA}/(A_{BA}+A_{CA}+A_{DA}) = 4/8$

Thus the probability that A is mutated into D is:

$M_{DA} = 4/8 * 8/40 = 4/40$

and the probability that A is not mutated is:

$M_{AA} = 1 - 8/40 = 32/40$

Derivation of the PAM matrices

Mutational probability matrix (*M*)

Let's define M_{ij} the probability of the amino acid in column *j* having been substituted by an amino acid in row *i* over a given evolutionary time unit.

Non-diagonal elements of M:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}}$$

Diagonal elements of M:

$$M_{ii} = 1 - \lambda m_i$$

In these equations, m is the relative mutability and A is the matrix of accepted point mutations. The constant λ represents a degree of freedom that could be used to connect the matrix M with an evolutionary time scale.

The coefficient λ could be adjusted to ensure that a specific (small) number of substitutions would occur on average per hundred residues. This adjustement was done by Dayhoff *et al* in the following way. The expected number of amino acids that will remain inchanged in a protein sequence 100 amino acid long is given by:

$$100 \sum_j f_j M_{jj} = 100 \sum_j f_j (1 - \lambda m_j)$$

If only one substitution per residue is allowed, then λ is calculated from the equation:

$$100 \sum_j f_j (1 - \lambda m_j) = 99$$

Derivation of the PAM matrices

Mutational probability matrix

In our example, $\lambda = 0.0261$ and the mutation probability matrix (PAM1) is:

	A	B	C	D	G	H	I	J
A	0.9948	0	0.0131	0.0131	0	0	0	0
B	0	0.9948	0.0131	0.0131	0	0	0	0
C	0.0026	0.0026	0.9740	0	0	0	0	0
D	0.0026	0.0026	0	0.9740	0	0	0	0
G	0	0	0	0	0.9957	0	0.0043	0
H	0	0	0	0	0	0.9957	0	0.0043
I	0	0	0	0	0.0043	0	0.9957	0
J	0	0	0	0	0	0.0043	0	0.9957

Note that *M* is a non-symmetric matrix.

Derivation of the PAM matrices

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.**

Source: Dayhoff, 1978

Derivation of the PAM matrices

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

This matrix is the mutation probability matrix for an evolution time of **1 PAM**.

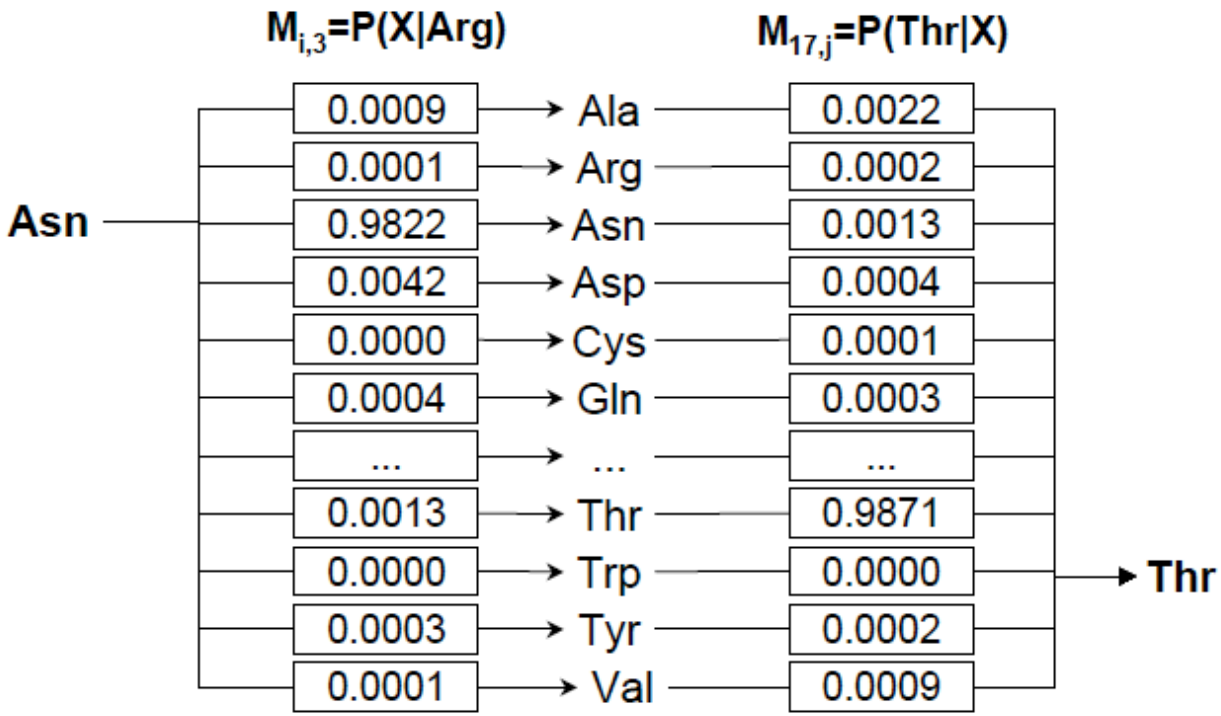
The **diagonal** represents the probability to still observe the same residue after 1 PAM. Therefore the diagonal represents the **99% of the case of non-mutation**.

Note that this does not mean that there was no mutation during this time interval. Indeed, the conservation of a residue could reflect either a conservation during the whole period, or a succession of two or more mutations ending at the initial residue

Source: J. van Helden

Derivation of the PAM matrices

From PAM1 to PAM2



P(Asn -> Thr)= P(Asn -> Ala -> Thr) + P(Asn -> Arg -> Thr) + ... + P(Asn -> Val -> Thr)
= (0.0009)(0.0001) + (0.0001)(0.0002) + ... + (0.0001)(0.009)

line 3 of PAM1 column 17 of PAM1

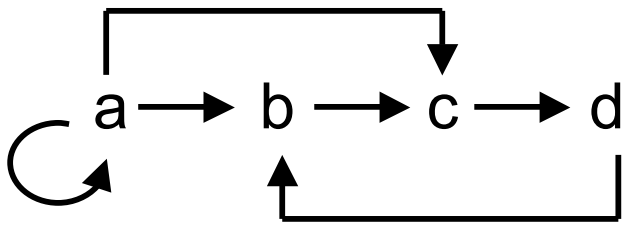
=> Matrix product: PAM2 = PAM1 x PAM1

Source: J. van Helden

Derivation of the PAM matrices

From PAM1 to PAM2, PAM100, PAM250, etc...

Remark (from graph theory)



	a	b	c	d
a	1	1	1	0
b	0	0	1	0
c	0	0	0	1
d	0	1	0	0

Matrix **Q** indicates the number of paths going from one node to another in 1 step

	a	b	c	d
a	1	1	2	1
b	0	0	0	1
c	0	1	0	1
d	0	1	1	1

Matrix $\mathbf{Q}^2$ indicates the number of paths going from one node to another in 2 steps

	a	b	c	d
a	...	...	...	...
b	...	...	...	...
c	...	...	...	...
d	...	...	...	...

Matrix $\mathbf{Q}^n$ indicates the number of paths going from one node to another in *n* steps

Source: J. van Helden

Derivation of the PAM matrices

From PAM1 to PAM2, PAM100, PAM250, etc...

Similarly:

PAM1	gives the probability to observe the changes $i \rightarrow j$ per 100 mutations
PAM2 = $PAM1^2$	gives the probability to observe the changes $i \rightarrow j$ per 200 mutations
PAM100 = $PAM1^{100}$	gives the probability to observe the changes $i \rightarrow j$ per 10 000 mutations
PAM250 = $PAM1^{250}$	gives the probability to observe the changes $i \rightarrow j$ per 25 000 mutations
PAMn = $PAM1^n$	gives the probability to observe the changes $i \rightarrow j$ per $100 \times n$ mutations.

Convergence: it can be verified that

PAM∞ = $PAM1^{\infty}$ converges to the observed frequencies:

$$\lim_{n \to \infty} M^n = \begin{pmatrix} f_A & f_A & \cdots & f_A \\ f_R & f_R & \cdots & f_R \\ \cdots & \cdots & & \cdots \\ f_V & f_V & \cdots & f_V \end{pmatrix}$$

Dayhoff *et al.* (1978) checked this convergence by computing M^{2034}.

Derivation of the PAM matrices

PAM250 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	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

For clarity, the values have been multiplied by 100

This matrix corresponds to an evolution time period giving 250 mutation/100 amino acids (i.e. an evolutionary distance of 250 PAM), and is refered to as the **PAM250 matrix.**

Source: Dayhoff, 1978

Derivation of the PAM matrices

Interpretation of the PAM250 matrix

	A	R	N	D	...
A	13	6	9	9	...
R	3	17	4	3	...
N	4	4	6	7	...
D	5	4	8	11	...
C	2	1	1	1	...
Q	3	5	5	6	...
E	5	4	7	11	...
G	12	5	10	10	...
H	2	5	5	4	...
I	3	2	2	2	...
L	6	4	4	3	...
K	6	18	10	8	...
M	1	1	1	1	...
F	2	1	2	1	...
P	7	5	5	4	...
S	9	6	8	7	...
T	8	5	6	6	...
W	0	2	0	0	...
Y	1	1	2	1	...
V	7	4	4	4	...

In comparing 2 sequences at this evolutionary distance (250 PAM), there is:

```
* * * * A * * * * *
```

↓ **250 PAM**

```
* * * * A * * * * *   probability of 13%
* * * * R * * * * *   probability of 3%
* * * * N * * * * *   probability of 4%
* * * * W * * * * *   probability of 0%
...
```

Source: Dayhoff, 1978

Derivation of the PAM matrices

From probabilities to scores

So far, we have obtained a **probability matrix**, but we would like a **scoring matrix**.

A **score** should reflect the significance of an alignment occurring as a result of an evolutionary process with respect to what we could expect by chance.

A score should involve the ratio between the probability derived from non-random (evolutionary) to random models:

$$r_n(i,j) = \frac{M^n_{ji}}{f_j} = \frac{P_{ji,n}}{f_i f_j}$$

← probability to see a pair (i,j) due to evolution

← probability to see a pair (i,j) by chance

The matrix $\boldsymbol{M_{ji}^n}$ is the mutational probability matrices at PAM distance *n*. Matrices M^1 and M^{250} have been shown before.

$\boldsymbol{P_{ji,n} = f_i M_{ji}^n}$ is the probability that two aligned amino acids have diverged from a common ancestor *n*/2 PAM unit ago, assuming that the substitutions follow a Markov process (for details, see Borodovsky & Ekisheva, 2007).

Note that ***R*** (the odd-score or relatedness matrix) is a symmetric matrix.

Derivation of the PAM matrices

Log-odd scores

In practice, we often use the log-odd scores defined by

$$s_n(i,j) = \log \frac{M_{ji}^n}{f_j} = \log \frac{P_{ji,n}}{f_i f_j}$$

This definition has convenient practical consequences:

A **positive score** ($s_n > 0$) characterizes the accepted mutations
A **negative score** ($s_n < 0$) characterizes the unfavourable mutations

Another property of the log-odd scores is that they can be added to produce the score of an alignment:

```
T  A  H  G  K
Y  S  D  G  D
```

$$S_{alignment} = s(T,Y) + s(A,S) + s(H,D) + s(G,G) + s(K,D)$$

Derivation of the PAM matrices

PAM250 matrix: log-odds scores

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

For clarity, the values have been multiplied by 10

Source: Dayhoff, 1978

Derivation of the PAM matrices

PAM250 matrix (log-odds)

		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Cys	C	12																			
Ser	S	0	2																		
Thr	T	-2	1	3																	
Pro	P	-1	1	0	6																
Ala	A	-2	1	1	1	2															
Gly	G	-3	1	0	-1	1	5														
Asn	N	-4	1	0	-1	0	0	2													
Asp	D	-5	0	0	-1	0	1	2	4												
Glu	E	-5	0	0	-1	0	0	1	3	4											
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4										
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp
Hydrophobic	C			P	A	G								M	I	L	V			
Aromatic											H							F	Y	W
Polar		S	T				N			Q									Y	
Basic											H	R	K							
Acidic								D	E											

Source: J. van Helden

PAM matrices: exercise

The original PAM250 substitution matrix scores a substitution of *Gly* by *Arg* by a negative score -3 (decimal logarithm and scaling factor 10 are used, with rounding to the nearest neighbour). The average frequency of *Arg* in the protein sequence database is 0.041. Use this information as well as the method described above to estimate the probability that *Gly* will be substituted by *Arg* after a PAM250 time period.

Source: Borodovsky & Ekisheva (2007)

PAM matrices: exercise

The original PAM250 substitution matrix scores a substitution of *Gly* by *Arg* by a negative score -3 (decimal logarithm and scaling factor 10 are used, with rounding to the nearest neighbour). The average frequency of *Arg* in the protein sequence database is 0.041. Use this information as well as the method described above to estimate the probability that *Gly* will be substituted by *Arg* after a PAM250 time period.

The element s_{ij} of the PAM250 substitution matrix and the frequency of amino acid j (f_j) in a protein sequence database are connected by the following formula:

$$s_{ij} = \left(10\log\frac{P(i \rightarrow j \text{ in } 250\ PAM)}{f_j}\right)$$

Therefore, the probability of substitution of *Gly* by *Arg* is:

$$P(Gly \rightarrow Arg \text{ in } 250\ PAM) = 0.041 \times 10^{-0.3} = 0.0205$$

Source: Borodovsky & Ekisheva (2007)

Derivation of the PAM matrices

Scoring an alignment

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

A scoring matrix like PAM250 can be used to score an alignment

```
T  A  H  G  K
Y  S  D  G  D
```

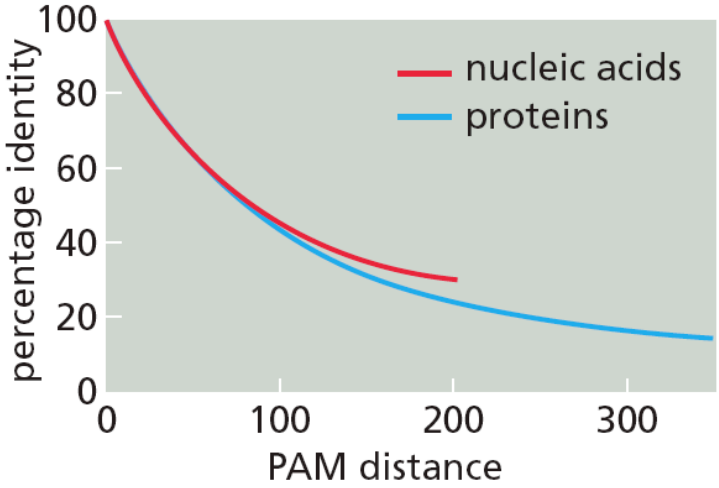
$$S_{alignment} = s(T,Y) + s(A,S) + s(H,D) + s(G,G) + s(K,D)$$
$$= -3 + 1 + 1 + 5 + 0$$
$$= 4$$

Choosing the appropriate PAM matrix

How to choose the appropriate PAM matrix?

Correspondance between the observed percent of amino acid difference *d* between the evolutionary distance *n* (in PAM) between them:

$$100 \sum_j f_j M_{jj}^n = 100 - d$$



identity (%)	*difference d (%)*	*PAM index n*
99	1	1
95	5	5
90	10	11
85	15	17
80	20	23
75	25	30
70	30	38
60	40	56
50	50	80
40	60	112
30	70	159
20	80	246
14	**86**	**350**

twilight zone
(detection limit)

Choosing the appropriate PAM matrix

How to choose the appropriate PAM matrix?

Altschul SF(1991) Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 219:555-65.

- PAM120 matrix is the most appropriate for database searches
- PAM200 matrix is the most appropriate for comparing two specific proteins with suspected homology

Remark:

In the PAM matrices, the **index** indicates the percentage of substitution per position.

Higher indexes are more appropriate for **more distant** proteins (PAM250 better than PAM100 for distant proteins).

Improved PAM matrices

Update of PAM matrices

With the rapid growth of protein data, updates and variants of the PAM matrices series have been proposed:

- Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8: 275-82.
- Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18 : 691-9.

BLOSUM scoring matrices

BLOSUM matrices were designed to find conserved regions in proteins (Henikoff & Henikoff, 1992).

Contrarily to the PAM matrices, the **BLOSUM** matrices:

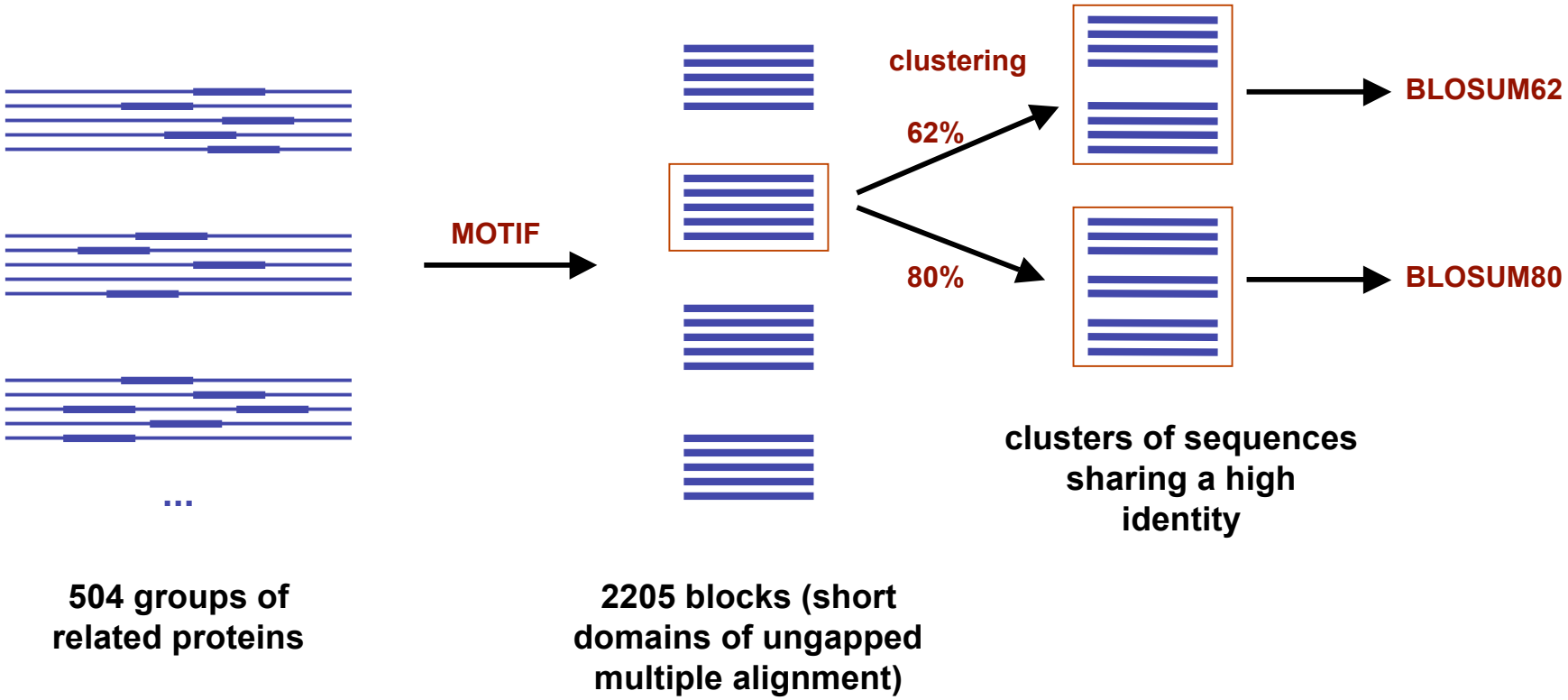
- are not based on evolutionary distances
- are based on ungapped aligned regions from proteins families called the block database (Henikoff & Henikoff, 1991)

BLOSUM = BLOcks SUbstitution Matrix

Reference: Henikoff S and Henikoff JG (1992). Amino acid substitution matrices from protein blocks. *PNAS* 89:10915-10919.

BLOSUM scoring matrices

Collection of protein blocks



Reference: Henikoff S and Henikoff JG (1991). Automated assembly of protein blocks for database searching. *Nucl Acids Res* 23:6565-6572.
Web: http://blocks.fhcrc.org/blocks/

Derivation of the BLOSUM matrices

As done for the PAM matrices, we will illustrate how the BLOSUM matrices can be derived using a simple example.

Let's start from the following sample sequences:

```
A D A D
A D C D
A C C D

D C A A
D C A A

A A C C
D A C C
```

The sequences are first clustered into "blocks" according the % of identity.

(here: 75% - the BLOSUM matrix obtained will thus have the index 75)

Reference: Borodovsky & Ekisheva (2007) Problems and Solutions in Biological sequence analysis. *Cambridge Univ Press*.

Derivation of the BLOSUM matrices

Matrix of weighted counts (*F*)

```
A D A D
A D C D
A C C D
```

each element in the first block
has a weight of 1/3

```
D C A A
D C A A
```

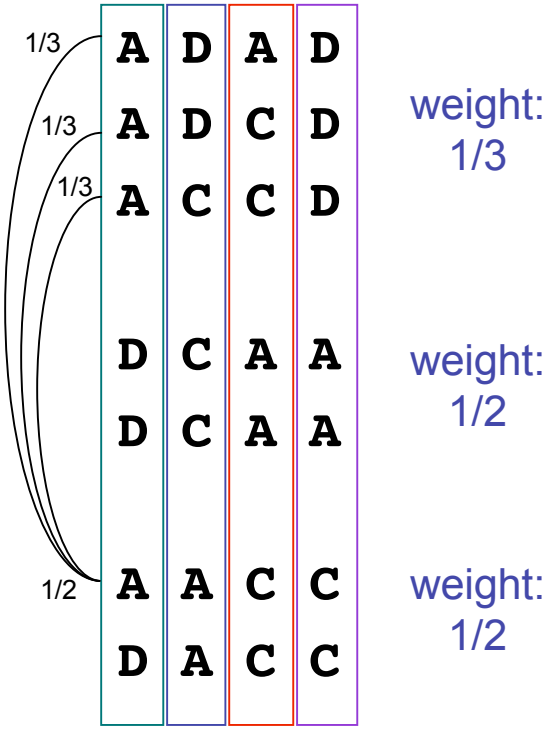
each element in the second block
has a weight of 1/2

```
A A C C
D A C C
```

each element in the third block
has a weight of 1/2

Derivation of the BLOSUM matrices

Matrix of weighted counts (*F*)



	A	C	D
A	5/6	13/3	11/3
C	13/3	1	5/3
D	11/3	5/3	1/2

Each element f_{ij} of matrix ***F*** is the weighted count of substitution of element *i* (in a cluster) by *j* (in another cluster).

For example, here

$$f_{AA} = \underbrace{1/3*1/2 + 1/3*1/2 + 1/3*1/2}_{\text{1st column}} + \underbrace{0}_{\text{2nd column}} + \underbrace{1/3*1/2 + 1/3*1/2}_{\text{3rd column}} + \underbrace{0}_{\text{4th column}} = 5/6$$

Derivation of the BLOSUM matrices

Observed frequencies of occurrence (*Q*)

The observed frequencies of occurrence of a pair (*i,j*) is defined by:

$$q_{ij} = \frac{f_{ij}}{\sum_{i}\sum_{j=1}^{i} f_{ij}}$$

In our example, we get the following observed frequencies of occurrence:

f_{ij}	A	C	D
A	5/6	13/3	11/3
C		1	5/3
D			1/2

q_{ij}	A	C	D
A	5/72	13/36	11/36
C		1/12	5/36
D			1/24

$$\sum_{i}\sum_{j=1}^{i} f_{ij} = f_{AA}+f_{AC}+f_{AD}+f_{CC}+f_{CD}+f_{DD} = 12$$

Examples: $q_{AA} = f_{AA}/12 = 5/72$
$q_{CA} = f_{CA}/12 = f_{AC}/12 = 5/72$

Derivation of the BLOSUM matrices

Expected frequency (*E*)

By definition, the expected frequencies are:

$$e_{ii} = p_i^2 \qquad e_{ij} = 2p_i p_j \quad \text{(for } i \neq j\text{)}$$

where p_i is the probability of occurrence of amino acid *i*:

$$p_i = q_{ii} + \frac{1}{2}\sum_{j \neq i} q_{ij}$$

we assume that there are as much C → A than A → C ($q_{ij} = p_{i \to j} + p_{j \to i}$)

In our example, we get the following expected frequencies:

$p_A = q_{AA} + 1/2\ (q_{CA}) + 1/2\ (q_{DA}) = 29/72$
$p_C = q_{CC} + 1/2\ (q_{AC}) + 1/2\ (q_{DC}) = 19/72$
$p_D = q_{DD} + 1/2\ (q_{AD}) + 1/2\ (q_{CD}) = 1/3$

e_{ij}	A	C	D
A	0.1622	0.2683	0.2125
C		0.1108	0.1757
D			0.0696

Derivation of the BLOSUM matrices

Log-odd ratio (*S*)

Finally, we calculate the log-odd ratio as

$$s_{ij} = 2\log_2 \frac{q_{ij}}{e_{ij}}$$

q_{ij} → Observed frequencies of occurrence of a pair (i,j)

e_{ij} → Expected frequencies of occurrence of a pair (i,j)

Calculating the log-odd ratio in our example, we obtained the following substitution score matrix:

s_{ij}	A	C	D
A	-2	1	1
C	1	-1	-1
D	1	-1	-1

This matrix corresponds to the BLOSUM matrix

Because the original clustering was done for a threshold identity of 75%, this is a BLOSUM75 matrix

BLOSUM62 matrix

Here is the BLOSUM matrix obtained by Henikoff & Henikoff on the basis of 2205 block of proteins. In each group, the sequences have been clustered together if they were 62% identical. This matrix is referred to as **BLOSUM62**.

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	

Source: J. van Helden

BLOSUM62 matrix

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Substitutions between acidic residues

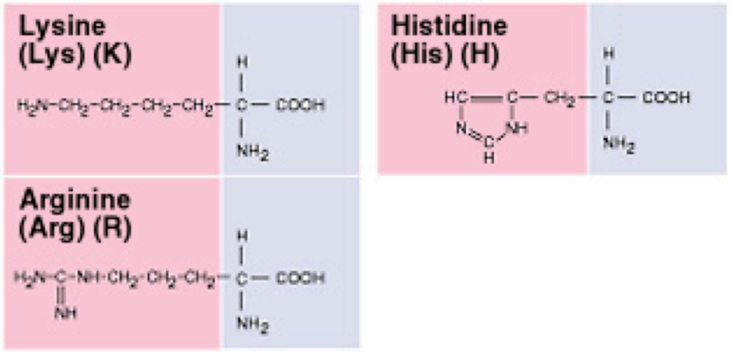
	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P									
Aromatic									H					F				W						
Polar			N			Q										S	T							
Basic		R							H			K												
Acidic				D			E																	

Ambiguous Amino Acids	3-Letter	1-Letter
Asparagine or aspartic acid	Asx	B
Glutamine or glutamic acid	Glx	Z
Leucine or Isoleucine	Xle	J
Unspecified or unknown amino acid	Xaa	X

BLOSUM62 matrix

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	

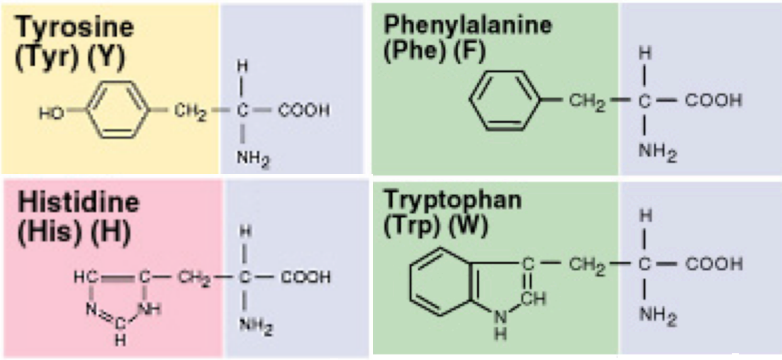


Substitutions between basic residues

BLOSUM62 matrix

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	

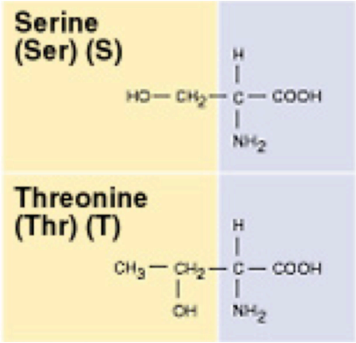


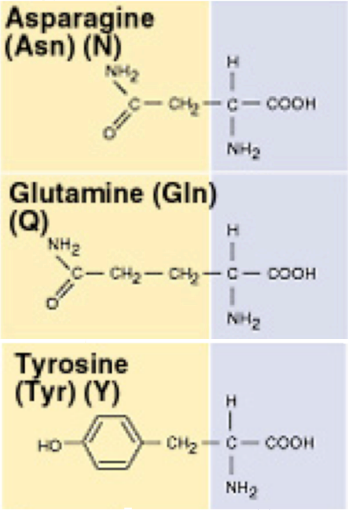
Substitutions between aromatic residues

BLOSUM62 matrix

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	





Substitutions between polar residues

BLOSUM62 matrix

Substitutions between hydrophobic residues

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	

RBLOSUM62

BLOSUM62 miscalculations improve search performance

MP Styczynski, KL Jensen, I Rigoutsos, G Stephanopoulos
Nat. Biotech. 26: 274–275, 2008

The BLOSUM family of substitution matrices, and particularly BLOSUM62, is the *de facto* standard in protein database searches and sequence alignments. In the course of analyzing the evolution of the Blocks database, we noticed errors in the software source code used to create the initial BLOSUM family of matrices (available online at ftp://ftp.ncbi.nih.gov/repository/blocks/ unix/blosum/blosum.tar.Z). The result of these errors is that the BLOSUM matrices—BLOSUM62, BLOSUM50, etc.—are quite different from the matrices that should have been calculated using the algorithm described by Henikoff and Henikoff. Obviously, minor errors in research, and particularly in software source code, are quite common. This case is noteworthy for three reasons: first, the BLOSUM matrices are ubiquitous in computational biology; second, these errors have gone unnoticed for 15 years; and third, the ‘incorrect’ matrices perform better than the ‘intended’ matrices.

RBLOSUM62

BLOSUM62 miscalculations improve search performance

MP Styczynski, KL Jensen, I Rigoutsos, G Stephanopoulos
Nat. Biotech. 26: 274–275, 2008

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

Supplementary Figure 5. The revised BLOSUM matrix, RBLOSUM62.

Values in red are one greater than they would be in BLOSUM62, while values in green are one less than they would be in BLOSUM62.

The entropy of this matrix (based on raw matrix values) is 0.6626 bits.

PAM vs BLOSUM matrices

BLOSUM62 scores - PAM160 scores

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

BLOSUM62 score matrix

Both matrices have identical relative entropies (0.70)

Source: Henikoff & Henikoff, 1992

PAM vs BLOSUM matrices

A comparison of the matrices can be done on the basis of their "information content" (see precise definition later)

more conserved proteins

PAM100 ≡ BLOSUM90

PAM120 ≡ BLOSUM80

PAM160 ≡ BLOSUM60

PAM200 ≡ BLOSUM52

PAM250 ≡ BLOSUM45

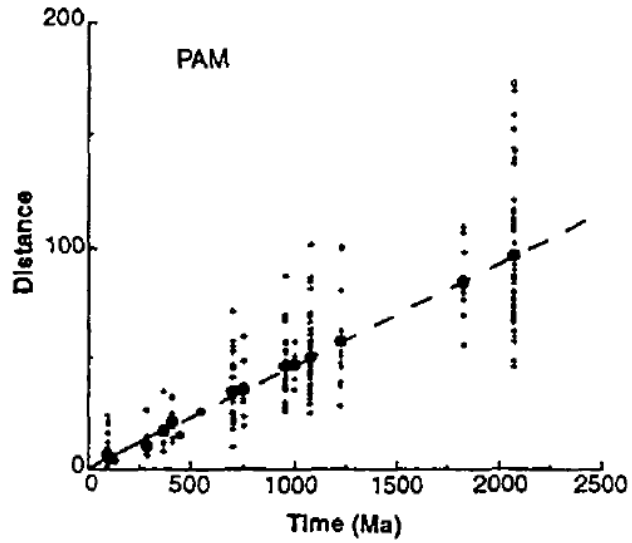
more distant proteins

PAM/BLOSUM matrices: link with real time?

It is very difficult to relate the substitution matrices with the real evolution time because the rate of mutations depends
- on the time
- on the species
- on the protein type

Nevertheless, using 57 enzymes from various organisms (animals, plants, fungi, bacteria) and under simplifying assumptions, Doolittle *et al* (1996) could relate the scores obtained with PAM250/BLOSUM62 matrices (converted into a distance) with the evolution time.

Using a linear fitting, they estimated that eukaryotes and eubacteria last shared a common ancestor about 2 billion (2.10^9) years ago.



Doolittle et al (1996) Determining divergence time of the major kingdoms of living organisms with a protein clock. *Science* 271: 470-477.

GONNET matrix

GONNET scoring matrix

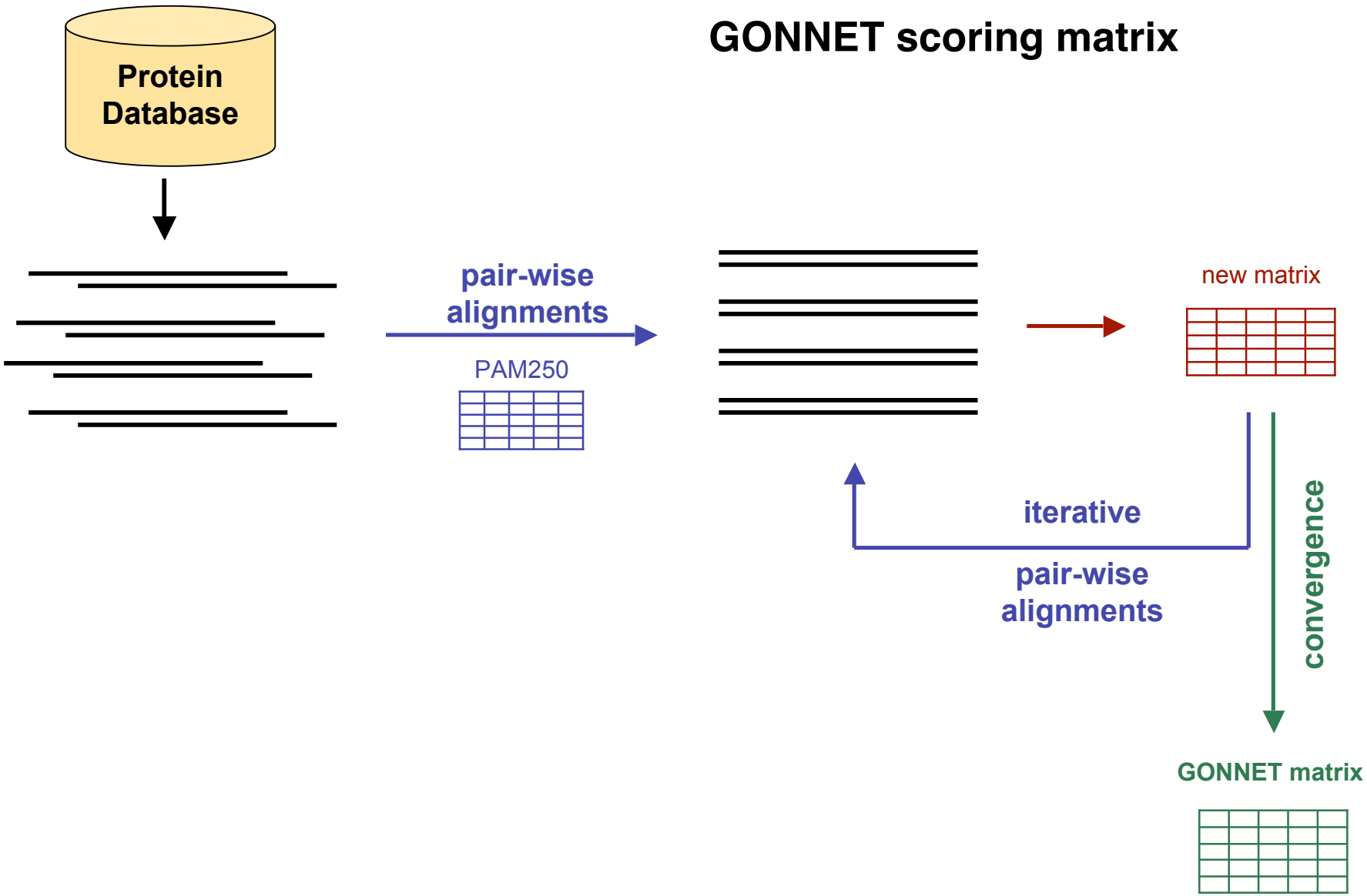
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	11.5																			
S	0.1	2.2																		
T	-0.5	1.5	2.5																	
P	-3.1	0.4	0.1	7.6																
A	0.5	1.1	0.6	0.3	2.4															
G	-2.0	0.4	-1.1	-1.6	0.5	6.6														
N	-1.8	0.9	0.5	-0.9	-0.3	0.4	3.8													
D	-3.2	0.5	0.0	-0.7	-0.3	0.1	2.2	4.7												
E	-3.0	0.2	-0.1	-0.5	0.0	-0.8	0.9	2.7	3.6											
Q	-2.4	0.2	0.0	-0.2	-0.2	-1.0	0.7	0.9	1.7	2.7										
H	-1.3	-0.2	-0.3	-1.1	-0.8	-1.4	1.2	0.4	0.4	1.2	6.0									
R	-2.2	-0.2	-0.2	-0.9	-0.6	-1.0	0.3	-0.3	0.4	1.5	0.6	4.7								
K	-2.8	0.1	0.1	-0.6	-0.4	-1.1	0.8	0.5	1.2	1.5	0.6	2.7	3.2							
M	-0.9	-1.4	-0.6	-2.4	-0.7	-3.5	-2.2	-3.0	-2.0	-1.0	-1.3	-1.7	-1.4	4.3						
I	-1.1	-1.8	-0.6	-2.6	-0.8	-4.5	-2.8	-3.8	-2.7	-1.9	-2.2	-2.4	-2.1	2.5	4.0					
L	-1.5	-2.1	-1.3	-2.3	-1.2	-4.4	-3.0	-4.0	-2.8	-1.6	-1.9	-2.2	-2.1	2.8	2.8	4.0				
V	0.0	-1.0	0.0	-1.8	0.1	-3.3	-2.2	-2.9	-1.9	-1.5	-2.0	-2.0	-1.7	1.6	3.1	1.8	3.4			
F	-0.8	-2.8	-2.2	-3.8	-2.3	-5.2	-3.1	-4.5	-3.9	-2.6	-0.1	-3.2	-3.3	1.6	1.0	2.0	0.1	7.0		
Y	-0.5	-1.9	-1.9	-3.1	-2.2	-4.0	-1.4	-2.8	-2.7	-1.7	2.2	-1.8	-2.1	-0.2	-0.7	0.0	-1.1	5.1	7.8	
W	-1.0	-3.3	-3.5	-5.0	-3.6	-4.0	-3.6	-5.2	-4.3	-2.7	-0.8	-1.6	-3.5	-1.0	-1.8	-0.7	-2.6	3.6	4.1	14.2

A different method to measure differences among amino acids was developed by Gonnet, Cohen and Benner (1992) using exhaustive pairwise alignments of the protein databases as they existed at that time. They used classical distance measures to estimate an alignment of the proteins. They then used this data to estimate a new distance matrix. This was used to refine the alignment, estimate a new distance matrix and so on iteratively. They noted that the distance matrices (all first normalised to 250 PAMs) differed depending on whether they were derived from distantly or closely homologous proteins. They suggest that for initial comparisons their resulting matrix should be used in preference to a PAM250 matrix, and that subsequent refinements should be done using a PAM matrix appropriate to the distance between proteins.

Gonnet, Cohen, Benner (1992). Exhaustive matching of the entire protein sequence database. *Science*. *256*:1443-1445.

Source: http://www.ebi.ac.uk/help/matrix.html

GONNET matrix



Other substitution matrices for amino acids

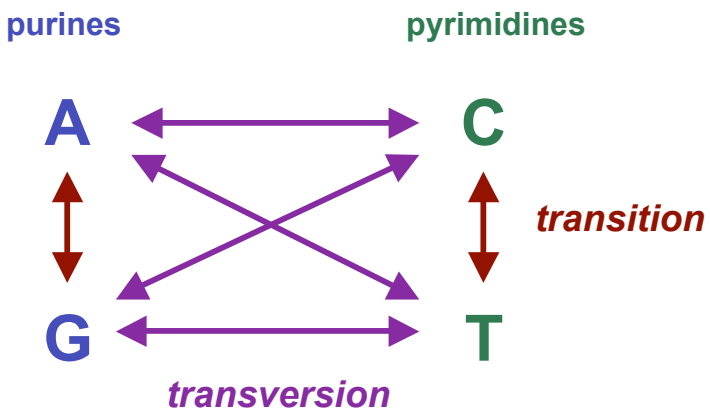
Many other families of substitution matrices for amino acids have been proposed:

- Simple identity matrices
- Matrices based on the genetic code changes (score the minimum of nucleotide changes to change a codon for one amino acid into a codon for another) (Fitch, 1966)
- Matrices based on chemical similarities of amino acid side chains (molecular volume, polarity, hydrophobicity) (Vogt et al, 1995)
- Matrices based on structurally aligned 3D structures (Risler et al, 1998; Henikoff & Henikoff, 1993)
- Dipeptide substitution matrices (Gonnet et al, 1994)
- Specific substitution matrices for transmembrane proteins (Jones et al, 1994)

Source: Mount, 2004

Substitution matrices for nucleotides

Substitution matrices for nucleotides have also been obtained in a similar way than the PAM matrices for proteins



PAM10	A	C	G	T
A	90.7	-3.70	-2.19	-3.70
C	-3.70	90.7	-3.70	-2.19
G	-2.19	-3.70	90.7	-3.70
T	-3.70	-2.19	-3.70	90.7

States et al (1991) Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* 3: 66-70.

Substitution matrices for nucleotides

Probability matrix for nucleotides

	A	G	T	C
A	**0.99**			
G	0.00333	**0.99**		
T	0.00333	0.00333	**0.99**	
C	0.00333	0.00333	0.00333	**0.99**

Probability matrix based on a model of uniform mutation rates among nucleotides

	A	G	T	C
A	**0.99**			
G	0.006	**0.99**		
T	0.002	0.002	**0.99**	
C	0.002	0.002	0.006	**0.99**

Probability matrix based on a model of 3-fold higher transition (substitution between the purines A and G or pyrimidine C and T) **that transversion** (substitution form a purine to a pyrimidine or a from pyrimidine to a purine) (Li & Graur, 1991)

Source: Mount, 2004

Substitution matrices for nucleotides

Probability matrix for nucleotides

	A	G	T	C
A	**0.99**			
G	0.00333	**0.99**		
T	0.00333	0.00333	**0.99**	
C	0.00333	0.00333	0.00333	**0.99**

	A	G	T	C
A	**0.99**			
G	0.006	**0.99**		
T	0.002	0.002	**0.99**	
C	0.002	0.002	0.006	**0.99**

Exercise

Here are given the probability matrices.

What would be the corresponding scoring matrices (PAM1-like) if we assume a equal probability of each nucleotide ($f_i = 0.25$)? Give the log-odd score and use $\log_2$.

What would be the PAM-like scoring matrices corresponding to a distance of 2 PAM?

What would be the PAM-like scoring matrices corresponding to a distance of 10 PAM? of 100 PAM?

Substitution matrices for nucleotides

Scoring matrix for nucleotides

	A	G	T	C
A	**2**			
G	-6	**2**		
T	-6	-6	**2**	
C	-6	-6	-6	**2**

Scoring matrix based on a model of uniform mutation rates among nucleotides (PAM1-like matrix)

	A	G	T	C
A	**2**			
G	-5	**2**		
T	-7	-7	**2**	
C	-7	-7	-5	**2**

Scoring matrix based on a model of 3-fold higher transition that transversion (PAM1-like matrix)

Source: Mount, 2004

Substitution matrices for nucleotides

Scoring matrix for nucleotides

	A	G	T	C
A	**2**			
G	-5	**2**		
T	-5	-5	**2**	
C	-5	-5	-5	**2**

Scoring matrix based on a model of uniform mutation rates among nucleotides (PAM2-like matrix)

	A	G	T	C
A	**2**			
G	-4	**2**		
T	-6	-6	**2**	
C	-6	-6	-4	**2**

Scoring matrix based on a model of 3-fold higher transition that transversion (PAM2-like matrix)

Source: Mount, 2004

Substitution matrices for nucleotides

Scoring matrix for nucleotides

	A	G	T	C
A	**1.86**			
G	-3.01	**1.86**		
T	-3.01	-3.01	**1.86**	
C	-3.01	-3.01	-3.01	**1.86**

Scoring matrix based on a model of uniform mutation rates among nucleotides (PAM10-like matrix)

	A	G	T	C
A	**1.86**			
G	-2.18	**1.86**		
T	-3.70	-3.70	**1.86**	
C	-3.70	-3.70	-2.18	**1.86**

Scoring matrix based on a model of 3-fold higher transition that transversion (PAM10-like matrix)

Source: Mount, 2004

Substitution matrices for nucleotides

Scoring matrix for nucleotides

	A	G	T	C
A	**0.83**			
G	-0.46	**0.83**		
T	-0.46	-0.46	**0.83**	
C	-0.46	-0.46	-0.46	**0.83**

Scoring matrix based on a model of uniform mutation rates among nucleotides (PAM100-like matrix)

	A	G	T	C
A	**0.88**			
G	**0.069**	**0.88**		
T	-0.86	-0.86	**0.88**	
C	-0.86	-0.86	**0.069**	**0.88**

Scoring matrix based on a model of 3-fold higher transition that transversion (PAM100-like matrix)

Note the positive values obtained for the transitions A-G and C-T!

Source: Mount, 2004

Substitution matrices for nucleotides

Scoring matrix for nucleotides

	A	G	T	C
A	**2**			
G	-6	**2**		
T	-6	-6	**2**	
C	-6	-6	-6	**2**

	A	G	T	C
A	**2**			
G	-5	**2**		
T	-7	-7	**2**	
C	-7	-7	-5	**2**

Exercise

Here are scoring PAM-1 matrices. By multiplying the probability matrix with itself, one can easily get PAM-n matrices.

Calculate the correspondance between those PAM-n matrices and the percent of identity.

Substitution matrices for nucleotides

Correspondance between the PAM distance and the identity level

identity %	difference %	PAM index (n)
99	1	1
90.6	9.4	10
78.7	21.3	25
63.5	36.5	50
44.8	55.2	100

model of uniform mutation rates among nucleotides

Note the mismatch scores in this model tend to 0 as PAM distance increases.
Thus, this matrix is not very informative at high PAM distance and should be used only for comparing sequences that are quite similar (using a low index PAM matrix).

identity %	difference %	PAM index (n)
99	1	1
90.7	9.3	10
79.0	21.0	25
64.2	35.8	50
46.3	53.7	100

model of 3-fold higher transition that transversion

Note that as PAM distance increases, the mismatch scores in this model become positive and appear as conservative substitutions!
Thus, this model can provide much more information than the uniform mutation model and should be used for distantly related sequences.

Source: Mount, 2004

Substitution matrices for nucleotides

Other substitution matrices for nucleotides have been derived...

Chiaromonte et al (2002) have obtained matrices by analysis of alignments of distinct regions of the human genome with different G+C content.

(A)

	A	C	G	T
A	67	-96	-20	-117
C	-96	100	-79	-20
G	-20	-79	100	-96
T	-117	-20	-96	67

CFTR region (37% G+C)

(B)

	A	C	G	T
A	91	-114	-31	-123
C	-114	100	-125	-31
G	-31	-125	100	-114
T	-123	-31	-114	91

HOXD region (47% G+C)

(C)

	A	C	G	T
A	100	-123	-28	-109
C	-123	91	-140	-28
G	-28	-140	91	-123
T	-109	-28	-123	100

hum16pter region (53% G+C)

Source: Zvelebil & Baum, 2007

Substitution matrices for nucleotides

Nucleotide scoring matrices used in FASTA and BLAST

FASTA and BLAST uses arbitrary nucleotide matrices

FASTA and WU-BLAST

	A	G	T	C
A	**5**			
G	-4	**5**		
T	-4	-4	**5**	
C	-4	-4	-4	**5**

should be used to detect homologous alignments that are 65% identical (edge of the twilight zone).

NCBI-BLAST

	A	G	T	C
A	**1**			
G	-2	**1**		
T	-2	-2	**1**	
C	-2	-2	-2	**1**

should be used to detect homologous alignments that are 95% identical (almost perfect match)

Source: Eddy, 2004

Substitution matrices: summary

- **Substitution matrices allow to detect similarities between more distant proteins than what would be detected with the simple identity of residues.**

- Different substitution scoring matrices have been established
  - PAM (Dayhoff, 1979)
  - BLOSUM (Henikoff & Henikoff, 1992)
  - Residue categories (Phylip)
  - ...

- Limitations of the substitution scoring matrices
  - They assumed independance between successive residues
  - They have been derived from manually aligned sequences
  - They have been built from a limited data set

Substitution matrices: summary

The matrix must be chosen carefully, depending on the expected rate of conservation between the sequences to be aligned.

Beware

- With **PAM** matrices

  The score indicates the percentage of substitution per position
  => **higher index** are appropriate for **more distant** proteins

- With **BLOSUM** matrices

  The score indicates the percentage of conservation
  => **higher index** are appropriate for **more conserved** proteins

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent ←		→ *More divergent*

Matrices used in BLAST Source: NCBI

Substitution matrices: further reading

PAM

Dayhoff *et al.* (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

BLOSUM

Henikoff & Henikoff (1992). Amino acid substitution matrices from protein blocks. *PNAS* 89:10915-10919.

Others substitution matrices for amino acids

Fitch (1966) An improved method of testing for evolutionary homology, *J Mol Biol* 21: 112-125.

Vogt et al (1995) An assessment of amino acid exchange matrices: the twilight zone re-visited. *J Mol Biol* 249: 816-831.

Risler et al (1988) Amino acid substitution in structurally related proteins: a pattern recognition approach, *J Mol Biol* 204: 1019-1029.

Henikoff & Henikoff (1993).Performance evaluation of amino acid substitution matrices, *Prot Struct Funct Genet* 17: 49-61.

Gonnet (1994) Analysis of amino acid substitution during divergent evolution: the 400 x 400 dipetide substitution matrix, *Biochem Biophys Res Commun* 199: 489-496.

Jones (1994) A mutational data matrix for transmembrane proteins, *FEBS Lett.* 339: 269-275.

Substitution matrices: further reading

Substitution matrices for nucleotides

States et al (1991) Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* 3: 66-70.

Chiaromonte et al (2002) Scoring pair-wise genomic sequence alignments. *Pac Symp Biocomput* 7: 115-126.

Comparison and discussion of substitution matrices

Doolittle (1981) Similar amino acid sequences: chance or common ancestry? *Science* 214: 149-159

Henikoff & Henikoff (1993). Performance evaluation of amino acid substitution matrices, *Prot Struct Funct Genet* 17: 49-61.

Pearson (1995) Comparison of methods for searching proteins sequence databases. Protein Sci 4: 1150-1160.

Pearson (1996) Effective protein sequence comparison. Meth. Enzymol 266: 227-258.

Altschul (1991) Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 219:555-65.

Reviews

Eddy (2004) Where did the BLOSUM62 alignment score matrix come from? *Nature Biotech* 22: 1035-1036.