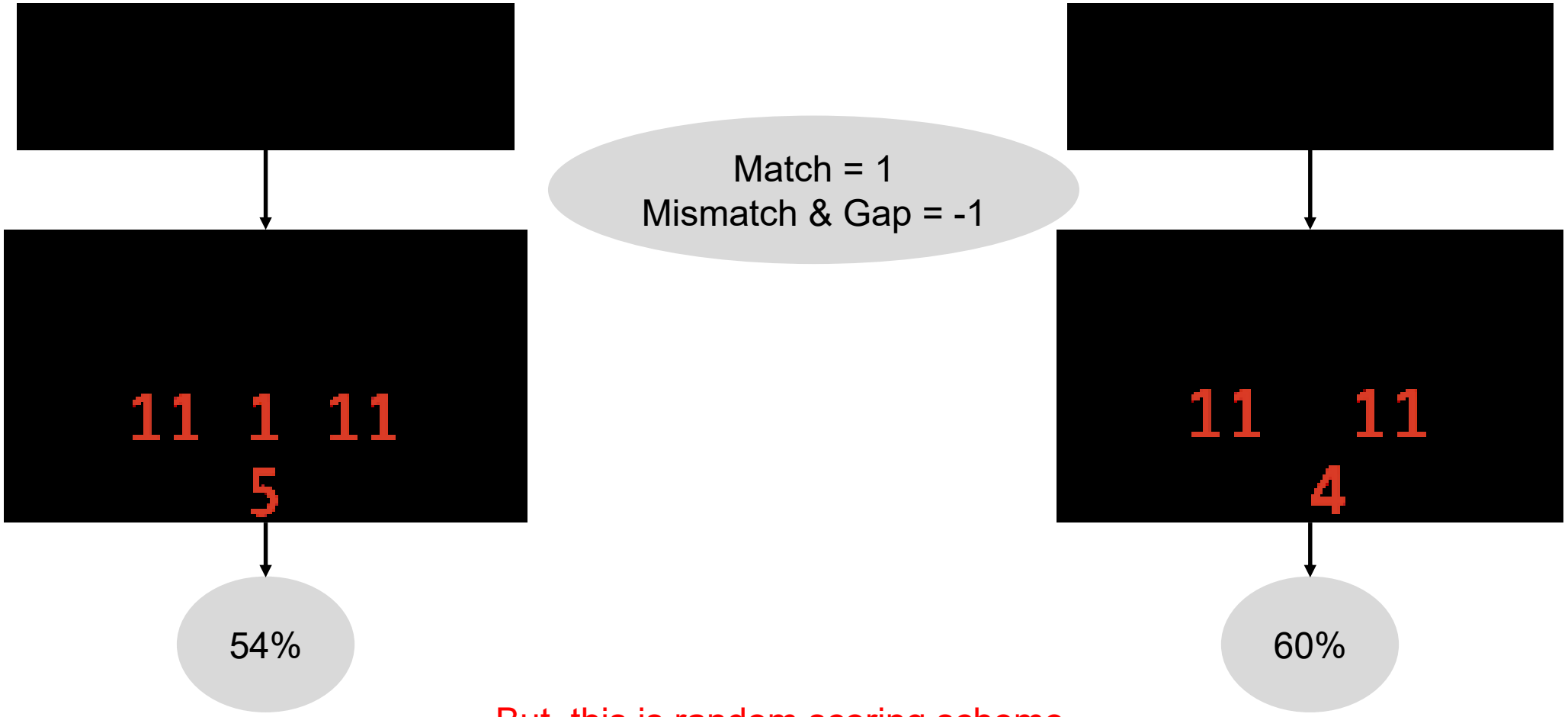
Scoring (or substitution) Matrices

Scoring schemes



But, this is random scoring scheme.

Scoring (or substitution) matrices

- It is possible to measure sequence similarity in many different ways such as:

- **Hamming Distance:** counting the number of differences between them.

- **Levenshtein Distance:** counting the number of insertions, deletions and substitutions required to make two sequences identical.

- **Percentage Similarity:** percent identity or just use an arbitrary scoring system for matches, mismatches, insertions and deletions.

- All these methods yield a measure of a ***relationship between the sequences, but none reflect any biological association between them***.

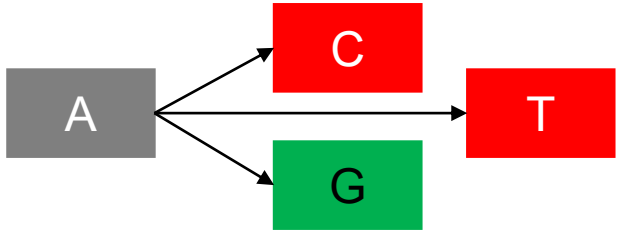
- Sequences may be more or less similar by sheer random chance, and consequently, we need a method to ***distinguish a random similarity from the similarity caused by evolutionary relationship***.

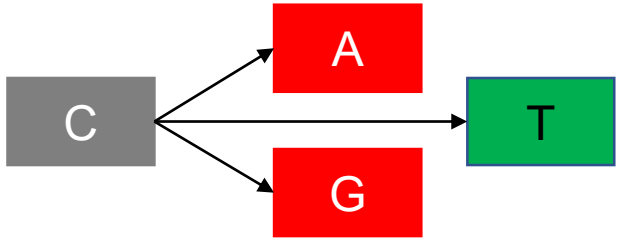
- Being able to determine if two sequences have the same function is useful in appraising the function of an unknown protein and gene by comparison to a known one.

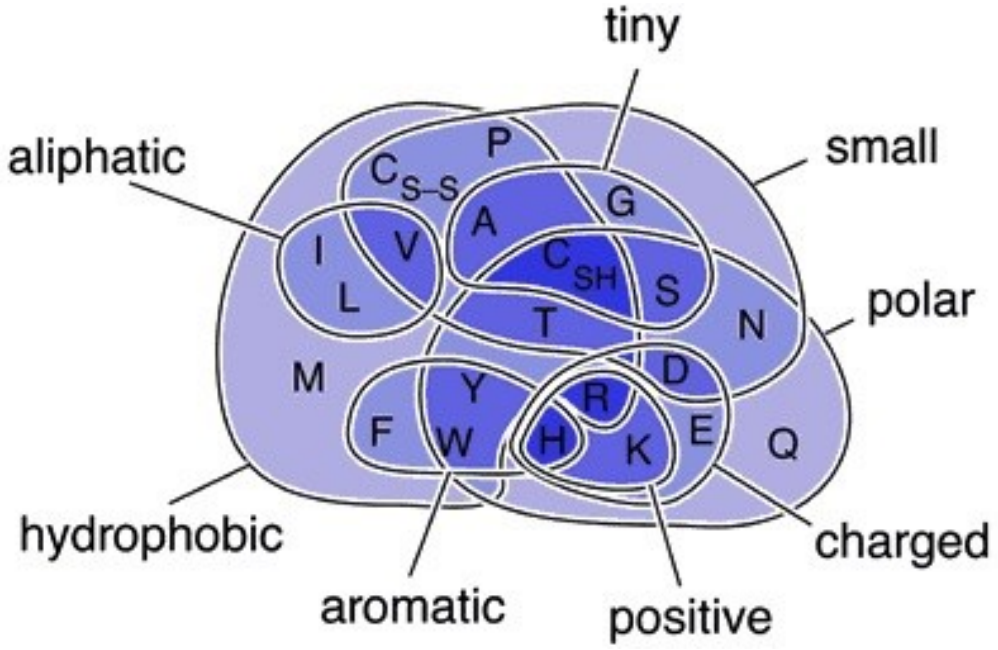
Scoring schemes

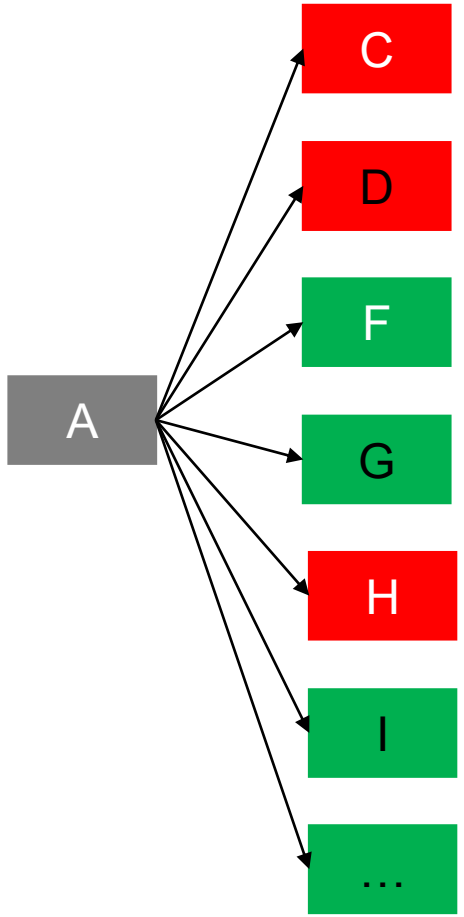
- The construction of DNA and protein sequence alignments is the same, the difference lies in how we score substitutions (mismatches). In DNA sequence alignments we only have identities.

- But protein sequence alignments contain amino acid pairings that are similar.









Construction of substitution matrices

- Consequently, by observing mutations among orthologous protein sequences, we can determine which amino acid changes are possible without altering the function of a protein.

- Further, by enumerating the frequencies of these changes, we can construct scoring systems such as:

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	P_{AA}																			
C	P_{CA}	P_{CC}																		
D	P_{DA}	P_{DC}	P_{DD}																	
E	P_{EA}	P_{EC}	P_{ED}	P_{EE}																
F	P_{FA}	P_{FC}	P_{FD}	P_{FE}	P_{FF}															
G	P_{GA}	P_{GC}	P_{GD}	P_{GE}	P_{GF}	P_{GG}														
H	P_{HA}	P_{HC}	P_{HD}	P_{HE}	P_{HF}	P_{HG}	P_{HH}													
I	P_{IA}	P_{IC}	P_{ID}	P_{IE}	P_{IF}	P_{IG}	P_{IH}	P_{II}												
K	P_{KA}	P_{KC}	P_{KD}	P_{KE}	P_{KF}	P_{KG}	P_{KH}	P_{KI}	P_{KK}											
L	P_{LA}	P_{LC}	P_{LD}	P_{LE}	P_{LF}	P_{LG}	P_{LH}	P_{LI}	P_{LK}	P_{LL}										
M	P_{MA}	P_{MC}	P_{MD}	P_{ME}	P_{MF}	P_{MG}	P_{MH}	P_{MI}	P_{MK}	P_{ML}	P_{MM}									
N	P_{NA}	P_{NC}	P_{ND}	P_{NE}	P_{NF}	P_{NG}	P_{NH}	P_{NI}	P_{NK}	P_{NL}	P_{NM}	P_{NN}								
P	P_{PA}	P_{PC}	P_{PD}	P_{PE}	P_{PF}	P_{PG}	P_{PH}	P_{PI}	P_{PK}	P_{PL}	P_{PM}	P_{PN}	P_{PP}							
Q	P_{QA}	P_{QC}	P_{QD}	P_{QE}	P_{QF}	P_{QG}	P_{QH}	P_{QI}	P_{QK}	P_{QL}	P_{QM}	P_{QN}	P_{QP}	P_{QQ}						
R	P_{RA}	P_{RC}	P_{RD}	P_{RE}	P_{RF}	P_{RG}	P_{RH}	P_{RI}	P_{RK}	P_{RL}	P_{RM}	P_{RN}	P_{RP}	P_{RQ}	P_{RR}					
S	P_{SA}	P_{SC}	P_{SD}	P_{SE}	P_{SF}	P_{SG}	P_{SH}	P_{SI}	P_{SK}	P_{SL}	P_{SM}	P_{SN}	P_{SP}	P_{SQ}	P_{SR}	P_{SS}				
T	P_{TA}	P_{TC}	P_{TD}	P_{TE}	P_{TF}	P_{TG}	P_{TH}	P_{TI}	P_{TK}	P_{TL}	P_{TM}	P_{TN}	P_{TP}	P_{TQ}	P_{TR}	P_{TS}	P_{TT}			
V	P_{VA}	P_{VC}	P_{VD}	P_{VE}	P_{VF}	P_{VG}	P_{VH}	P_{VI}	P_{VK}	P_{VL}	P_{VM}	P_{VN}	P_{VP}	P_{VQ}	P_{VR}	P_{VS}	P_{VT}	P_{VV}		
W	P_{WA}	P_{WC}	P_{WD}	P_{WE}	P_{WF}	P_{WG}	P_{WH}	P_{WI}	P_{WK}	P_{WL}	P_{WM}	P_{WN}	P_{WP}	P_{WQ}	P_{WR}	P_{WS}	P_{WT}	P_{WV}	P_{WW}	
Y	P_{YA}	P_{YC}	P_{YD}	P_{YE}	P_{YF}	P_{YG}	P_{YH}	P_{YI}	P_{YK}	P_{YL}	P_{YM}	P_{YN}	P_{YP}	P_{YQ}	P_{YR}	P_{YS}	P_{YT}	P_{YV}	P_{YW}	P_{YY}

Point Accepted Mutation (PAM) Matrices

Point Accepted Mutation (PAM) Matrix

- Margaret Dayhoff and colleagues (in 1978) compared closely related homologous sequences and counted the frequency of each type of substitution.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	50	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Protein sequences from 34 families grouped into 71 evolutionary trees resulting in a total count of 1,572 accepted point mutations. The displayed counts are original counts times 10.

The assumption is that the likelihood of substitution XY is the same as YX; thus, the method does not allow measurements over evolutionary distances, but the distances are in PAMs.

PAM matrices

PAM matrices are amino acid substitution matrices that **encode the expected evolutionary change at the amino acid level**. **Each PAM matrix is designed to compare two sequences which are a specific number of PAM units apart**.

For example - the PAM120 score matrix is designed to compare between sequences that are 120 PAM units apart: The score it gives a pair of sequences is the (log of the) probabilities of such sequences evolving during 120 PAM units of evolution.

For any specific pair (A_i, A_j) of amino acids, the (i,j) entry in the PAMn matrix reflects the frequency at which A_i is expected to replace with A_j in two sequences that are ***n* PAM units diverged**. These frequencies should be estimated by gathering statistics on replaced amino acids.

Collecting statistics about amino acids substitution in order to compute the PAM matrices is relatively difficult for sequences that are distantly diverged. But for sequences that are highly similar, i.e., the PAM divergence distance between them is small, finding the position correspondence is relatively easy since only few insertions and deletions took place.

Point Accepted Mutation (PAM) Matrix

- The assumption in this evolutionary model is that the amino acid substitutions observed over short periods of evolutionary history can be extrapolated to longer distances.

- In deriving the PAM matrices, each change in the current amino acid at a particular site is assumed to be independent of previous mutational events at that site. Thus, the probability of change of any amino acid **a** to amino acid **b** is the same, regardless of the previous changes at that site and also regardless of the position of amino acid **a** in a protein sequence.

- ***Amino acid substitutions in a protein sequence are thus viewed as a Markov model, characterized by a series of changes of state in a system such that a change from one state to another does not depend on the previous history of the state.***

PAM units

PAM units are used to measure the amount of evolutionary distance between two amino acid sequences. Two strings S_1 and S_2 are said to be one PAM unit diverged if a series of accepted point mutations (and no insertions of deletions) has converted S_1 to S_2 with an average of one accepted point-mutation event per 100 amino acids.

The term ”**accepted**" here means a mutation that was incorporated into the protein and passed to its progeny. Therefore, either the mutation did not change the function of the protein or the change in the protein was beneficial to the organism.

Note that **two strings which are one PAM unit diverged do not necessarily differ in one percent**, as often mistakenly thought, because a single position may undergo more than a single mutation. The difference between the two notions grows as the number of units does.

PAM units

There are two main problems with the notion of the PAM units:

First, practically all the sequences we can obtain today are extracted from extant organisms. One almost does not know any protein sequences where one is actually derived from the other. The lack of ancestral protein sequences is handled by assuming that amino acid mutations are reversible and equally likely in either direction. This assumption, together with the additivity property of the PAM units derived from its definition, imply that given two amino acid sequences S_i and S_j whose mutual ancestor is S_{ij}:

$$d(S_i,S_j) = d(S_i,S_{ij}) + d(S_{ij},S_j)$$

when $d(S_i,S_j)$ is the PAM distance between amino acid sequences S_i and S_j.

The second problem, which is more difficult to overcome, is that insertions and deletions are disregarded which may occur during evolution, hence one can not be sure of the correct correspondence between sequence positions. In order to know the exact correspondence one has to be able to identify the true historical gaps, or, at least to identify large intervals along the two sequences where the correspondence is correct. This can not always be done with certainty, especially when the two sequences are distantly diverged.

Derivation of PAM matrices

Dayhoff developed this model using the four step approach. Specifically:

Step 1:
As training data, Dayhoff and her colleagues used a set of ungapped, global multiple sequence alignments of 71 groups of closely related sequences. Within each group, the sequence identity was 85% or greater. The rationale is that sequences with at least 85% identity will contain no site that has sustained more than one mutation.

Step 2:
Observed amino acid pair frequencies were tabulated from the 71 multiple alignments. Sample bias was corrected by counting the minimum number of changes required to fit the data to a tree. This requires inferring the unrooted tree that describes the evolutionary relationships between the sequences in each aligned family and then estimating the number of amino acid replacements that occurred on each branch of that tree.

Derivation of PAM matrices

Let us consider the following alignment of four amino acid sequences of length four:

```
1:   AEIR
2:   DEIR
3:   QKLH
4:   AHLH
```

Derivation of PAM matrices

For an alignment with four sequences, there are three unrooted trees with four leaves

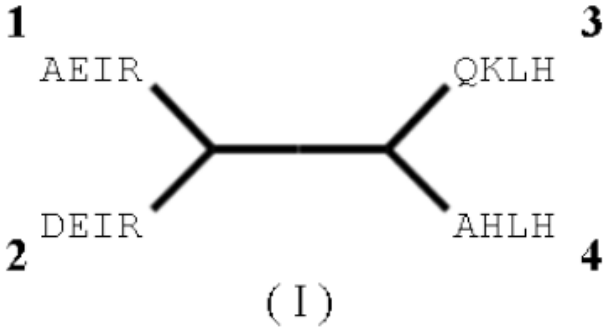
If there are *n* leaf nodes or taxa, how many different trees are possible?

N_R = Number of Possible Rooted Trees

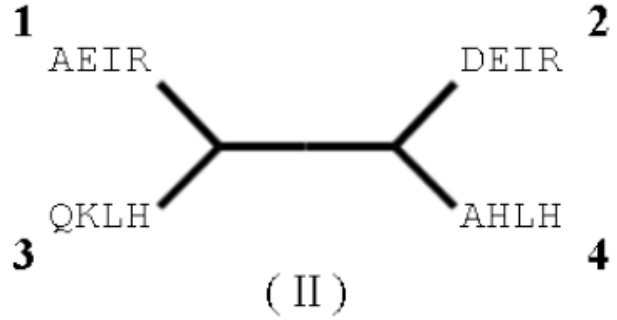
$$= \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

N_U = Number of Possible Unrooted Trees

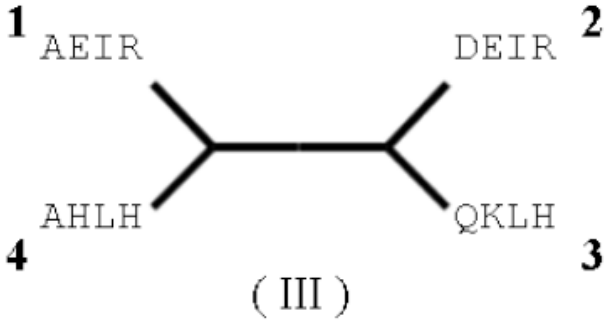
$$= \frac{(2n-5)!}{2^{n-3}(n-3)!}$$



(I)



(II)



(III)

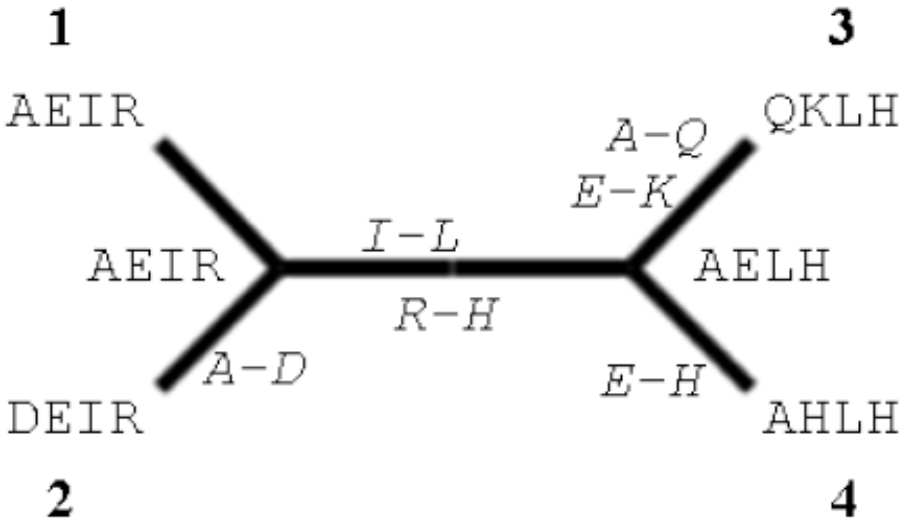
For each tree, the leaves are annotated with the corresponding present-day sequences. The sequences on internal nodes are unknown, since they correspond to ancestral sequences.

Derivation of PAM matrices

First, we will illustrate how to estimate the number of substitutions, given the evolutionary tree. Then, we will return to the question of how to infer the tree that best explains a given alignment.

Dayhoff inferred the sequences on the internal nodes according to the parsimony criterion, which states that the best hypothesis is the one that requires the fewest amino acid replacements to explain the data. Consistent with this criterion, sequences were assigned to the internal nodes of each tree in such a way that the total number of changes along branches of the tree is minimized.

For example, suppose that we have determined that Tree I is the best hypothesis for the evolutionary history of the four sequences in the alignment. Ancestral sequences that satisfy the parsimony criterion for Tree I are



Derivation of PAM matrices

With these ancestral sequences, six substitutions (shown on their respective branches) are required to explain the evolution of the four present day sequences.

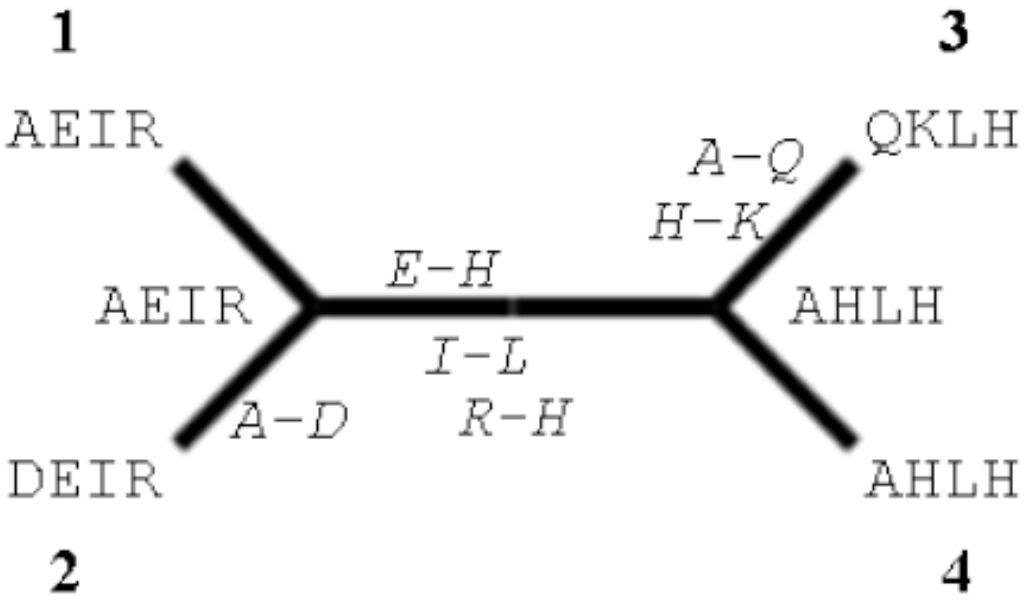
Once ancestral sequences have been inferred, the counts for each amino acid pair are tabulated.

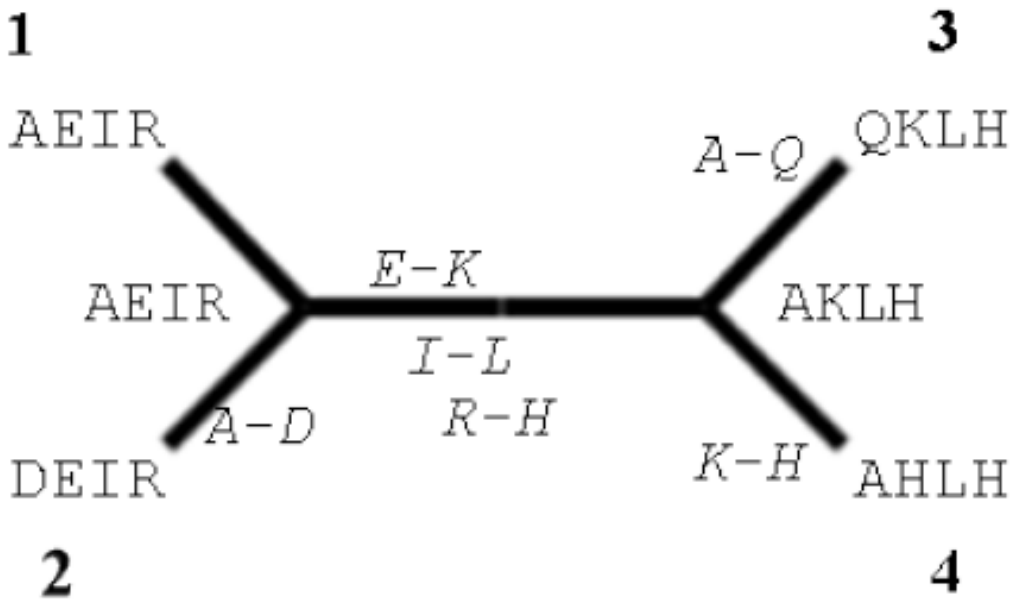
Axy, the number of x,y pairs observed, is determined by counting the number of edges connecting x and y, for x ≠ y. Note that Axy = Ayx, since every edge connecting x with y also connects y with x. Axx is defined to be twice the number of edges connecting x and x. This is because the edges connecting two dissimilar residues are also counted twice, once in the xy direction and once in the yx direction. The tabulated counts for all amino acid pairs are given

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			6	1		1			
H			1	4					1
I					4		1		
K			1						
L					1		4		
Q	1								
R				1					4

Derivation of PAM matrices

In general, there can be more than one way to assign sequences to internal nodes such that the total change is minimized. Each most parsimonious set of internal node labels will result in different amino acid pair counts. For example, there are two additional assignments of ancestral sequences for which six substitutions are sufficient to explain the present-day sequences





Derivation of PAM matrices

The pair counts resulting from these two alternate sets of labels are

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			4	1					
H			1	6		1			1
I					4		1		
K				1					
L					1		4		
Q	1								
R				1					4

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			4			1			
H				4		1			1
I					4		1		
K			1	1		2			
L					1		4		
Q	1								
R				1					4

Derivation of PAM matrices

Comparison of the original multiple alignment with the pair counts derived from the trees demonstrates how this approach compensates for sample bias and leads to different amino acid pair statistics. If we derived amino acid pairs directly from the alignment, each sequence would be compared to three other sequences, effectively counting the replacement of the same amino acid more than once. In contrast, when counting amino acid pairs on a tree, each sequence is compared to one other sequence, i.e., the inferred ancestral sequence.

For example, since D and Q both appear in the first column of the alignment, obtaining amino acid pair counts directly from the alignment would result in a non-zero value of A_{DQ}. However, no D-Q replacement appears on the branches of the labeled trees and $A_{DQ} = 0$.

Derivation of PAM matrices

Having demonstrated how to infer ancestral sequences for a given evolutionary tree, we return to the question of how to infer the tree that is the best hypothesis for the aligned sequences. Dayhoff also used the parsimony principle to select the tree. For a given tree, the minimum number of changes required to explain the present day sequences, over all possible internal labelings, is called the parsimony score of that tree. Tree I has a parsimony score of 6, for example.

Given an alignment of a family of k sequences, all unrooted trees with k sequences were considered and the parsimony score was estimated for each tree. In general, there can be more than one most parsimonious tree for a given set of present-day sequences (there is only one in this example).

Having found the set of most parsimonious trees, Dayhoff estimated amino acid pair frequencies by averaging the counts over all most parsimonious labelings of all most parsimonious trees, yielding

$$A_{xy} = \frac{1}{n_T} \sum_T A_{xy}^T,$$

where nT is the number of labeled trees with an optimal parsimony score and T is an indicator variable that enumerates such trees.

Derivation of PAM matrices

Since there is no way of knowing which set of inferred ancestral sequences is the best estimate, all possibilities must be considered. Dayhoff does this by averaging the counts over all most parsimonious labelings.

	A	D	E	H	I	K	L	Q	R
A	6	1						1	
D	1								
E			4.7	1		0.7			
H			1	4.7		0.7			1
I					4		1		
K			0.7	0.7		0.7			
L					1		4		
Q	1								
R				1					4

Derivation of PAM matrices

Step 3:
To estimate substitution frequencies from amino acid pair counts, Dayhoff constructed a family of Markov models representing evolution at a single site, i, in an amino acid sequence (model assumes site independence.) All models in the family have twenty states, one for each amino acid.

If the model visits state x at time t, we say that the amino acid at site i was an x at time t. The models differ in their transmission probability matrices, which reflect the propensity for amino acid replacement at various evolutionary divergences.

Dayhoff derived $P^{(1)}xy$, transition matrix for the 1 PAM model, from closely related alignments that may be assumed to contain no multiple substitutions. $P^{(1)}xy$ is the probability that amino acid x will be replaced by amino acid y in sequences separated by 1 PAM of evolutionary distance.

Next, Dayhoff derived the PAM-N transition matrix, $P^{(N)}xy$, by extrapolating from the PAM-1 transition probability.

Derivation of PAM matrices

The transition matrix $P^{(1)}xy$ is derived from the counts, Axy as follows:

$$P_{xy}^{(1)} = m_x \frac{A_{xy}}{\sum_{h \neq x} A_{xh}}, \quad x \neq y$$

$$P_{xx}^{(1)} = 1 - m_x$$

Here, m_x is the “mutability" of amino acid x and is defined to be

$$m_x = \frac{1}{L p_x z} \sum_{l \neq x} A_{xl},$$

where p_x is the background frequency of x, L is the length of the alignment, and z is a scaling that guarantees that the transition matrix will correspond to exactly 1 PAM.

Derivation of PAM matrices

We select the scaling factor, z, so that

$$\sum_{x=1}^{20}(p_x m_x) = \frac{1}{100}.$$

This scaling factor is required because although the training alignments are sufficiently conserved to contain no multiple substitutions, but the frequency of replacements in each alignment may not be exactly one in a hundred.

$$z = \frac{100}{L}\sum_{x=1}^{20}\sum_{l\neq x} A_{xl}.$$

$$m_x = \frac{0.01}{p_x}\ \frac{\sum_{l\neq x} A_{xl}}{\sum_h \sum_{l\neq h} A_{hl}}.$$

$$P_{xy}^{(1)} = \frac{0.01}{p_x}\ \frac{A_{xy}}{\sum_h \sum_{l\neq h} A_{hl}}.$$

Derivation of PAM matrices

Note that $P^{(1)}xy$ is consistent with the definition of a Markov chain: the rows of the transition matrix sum to 1 and it is history independent. This Markov chain is definite, aperiodic and irreducible (“connected"). Therefore, it has a stationary distribution.

We now derive the PAM-2 transition matrix. Note that the residue at site i can change from x to y in two time steps via several state paths: x → x → y; x → y → y, or x → l → y, where l is a third amino acid, not equal to x or y.

Recall that the probability of changing from x to y in two time steps is

$$P_{xy}^{(2)} = \sum_{l} P_{xl}^{(1)} P_{ly}^{(1)}$$

$P^{(2)}$ can be derived by squaring the matrix P(1) by matrix multiplication. This is the transition probability of a second order Markov chain that models amino acid replacements that occur in two time steps.

Similarly, we can use matrix multiplication to derive the PAM-N transition matrix for any N ≥ 2 as follows:

$$P^{(N)} = \left(P^{(1)}\right)^{N}.$$

Derivation of PAM matrices

Step 4:
We obtain a log likelihood scoring matrix from the transition probability matrix as follows. Let $q^{(N)}_{xy} = p_x P^{(N)}_{xy}$ be the probability that we see amino acid x aligned with amino acid y at a given position in an alignment of sequences with N PAMs of divergence; i.e., that amino acid x has been replaced by amino acid y after N PAMs of mutational change.

Then, we define the PAM-N scoring matrix to be

$$\begin{aligned} S^N[x,y] &= \lambda \log \frac{q^{(N)}_{xy}}{p_x p_y} \\ &= \lambda \log \frac{P^{(N)}_{xy}}{p_y}, \end{aligned}$$

where λ is a constant chosen to scale the matrix to a convenient range. Typically λ = 10 and the entries of S^N are rounded to the nearest integer. It is a log likelihood ratio, where $q^{(N)}_{xy}$ is the probability of seeing x and y aligned under the alternate hypothesis that x and y share common ancestry with divergence N and $p_x p_y$ is the probability that x and y are aligned by chance.

Derivation of PAM matrices

It is easy to verify that the PAM-N transition matrix is not symmetric; that is, $P^{(N)}_{xy} \neq P^{(N)}_{yx}$. This makes sense since replacing amino acid x with amino acid y may have different consequences than replacing y with x.

In contrast, the substitution matrix is symmetric; that is, $S^N[x; y] = S^N[y; x]$. This makes sense because in an alignment, we cannot determine direction of evolution, so we assign the same score when x is aligned with y, and when y is aligned with x.

Making of the mutation probability matrix for the evolutionary distance of one PAM (PAM1)

- The PAM1 matrix consists of sequences that are 1% different, i.e., a single mutation per 100 amino acids and is the evolutionary distance of one PAM.

Original amino acid

Replacement amino acid	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	Sum
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18	1.007
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1	0.998
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1	0.996
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1	0.999
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2	0.999
Q	3	9	4	5	0	9876	27	1	23	1	2	6	4	0	6	2	2	0	0	1	0.997
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2	1.000
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5	1.003
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1	0.998
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33	0.996
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15	1.008
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1	1.007
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4	0.990
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0	1.000
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2	0.999
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2	1.004
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9	1.000
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0	0.998
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1	0.999
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901	1.004
Sum	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

Extrapolation of PAM1 mutation matrix

PAMn=PAM1xPAM(n-1)

A	B	C		J	K	L	=	[AJ + BM + CP]	[AK + BN + CQ]	[AL + BO + CR]
D	E	F	x	M	N	O	=	[DJ + EM + FP]	[DK + EN + FQ]	[DL + EO + FR]
G	H	I		P	Q	R	=	[GJ + HM + IP]	[GK + HN + IQ]	[GL + HO + IR]

Original amino acid

PAM 2 Probability Matrix

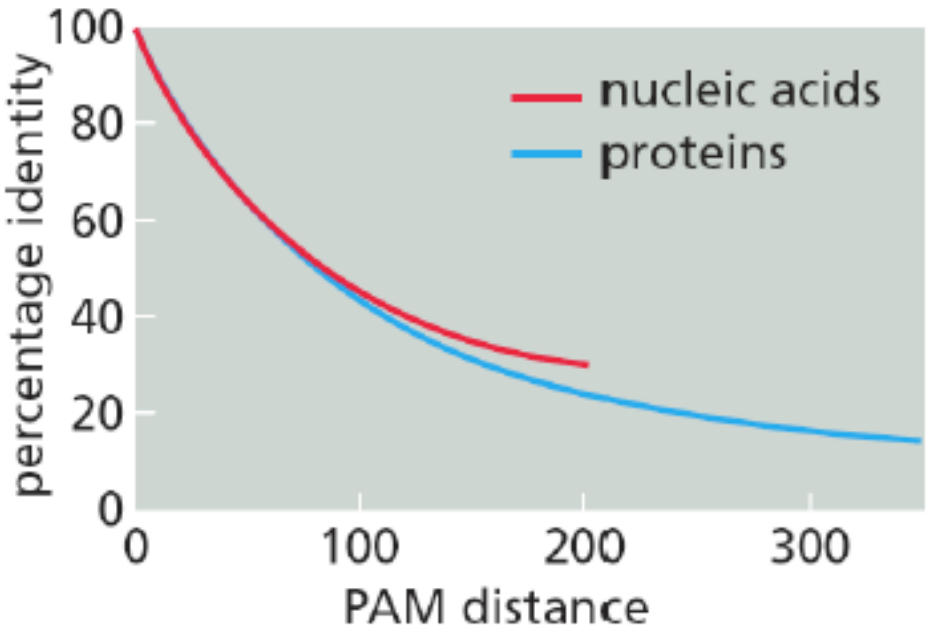
Replacement amino acid

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9736	2	8	11	2	7	19	42	2	5	7	4	2	1	25	56	43	0	1	26
R	5	9828	3	0	2	19	0	2	16	5	3	74	3	1	10	21	3	4	0	3
N	17	3	9647	82	0	8	15	24	35	6	6	50	0	1	4	67	26	0	6	2
D	21	0	71	9720	0	10	111	22	6	2	0	12	0	0	1	13	8	0	0	2
C	6	2	0	0	9947	0	0	2	2	3	0	0	0	0	2	22	2	0	6	6
Q	15	20	8	13	0	9754	69	5	40	1	12	24	3	0	15	8	6	0	0	4
E	34	0	12	105	0	54	9731	14	3	4	2	13	1	0	5	11	4	0	1	5
G	41	1	11	12	1	2	8	9870	1	0	1	4	1	1	4	32	4	0	0	7
H	4	19	42	8	2	46	4	2	9825	1	8	4	0	4	9	5	3	1	8	6
I	11	5	6	2	3	1	6	0	1	9746	43	7	10	15	1	3	22	0	2	113
L	7	1	3	0	0	6	1	1	3	19	9894	3	15	12	3	2	4	1	2	22
K	5	37	25	7	0	12	8	5	2	3	3	9852	7	0	3	13	16	0	1	1
M	12	7	0	0	0	9	3	3	0	24	89	38	9751	7	2	9	12	0	0	33
F	3	1	1	0	0	0	0	3	3	14	27	0	3	9891	1	6	2	2	41	2
P	43	8	3	1	1	12	5	6	6	1	5	5	1	1	9852	33	9	0	0	6
S	70	12	39	9	11	4	8	41	2	2	3	15	2	4	24	9684	63	2	2	4
T	63	2	18	6	1	4	3	6	2	14	6	22	3	1	8	75	9744	0	2	20
W	0	16	2	0	0	0	0	0	2	0	8	0	0	6	0	10	0	9952	4	0
Y	4	1	8	0	6	0	2	0	8	3	5	2	0	55	0	5	5	1	9891	4
V	36	2	1	2	3	3	4	10	3	64	30	2	8	1	5	4	18	0	2	9804

How to choose the appropriate PAM matrix?

Correspondence between the observed percent of amino acid difference d and the evolutionary distance n (in PAM):

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



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

Limitations: Based on only one original dataset, examines proteins with few differences (85% identity) and based mainly on small globular proteins so the matrix is biased.

BLOcks SUbstitution Matrix (BLOSUM) Matrices

BLOcks SUbstitution Matrix (BLOSUM)

- To measure the amino acid frequencies, Henikoff and Henikoff (in 1990) analyzed conserved regions of related protein sequences taken from BLOCKS database.

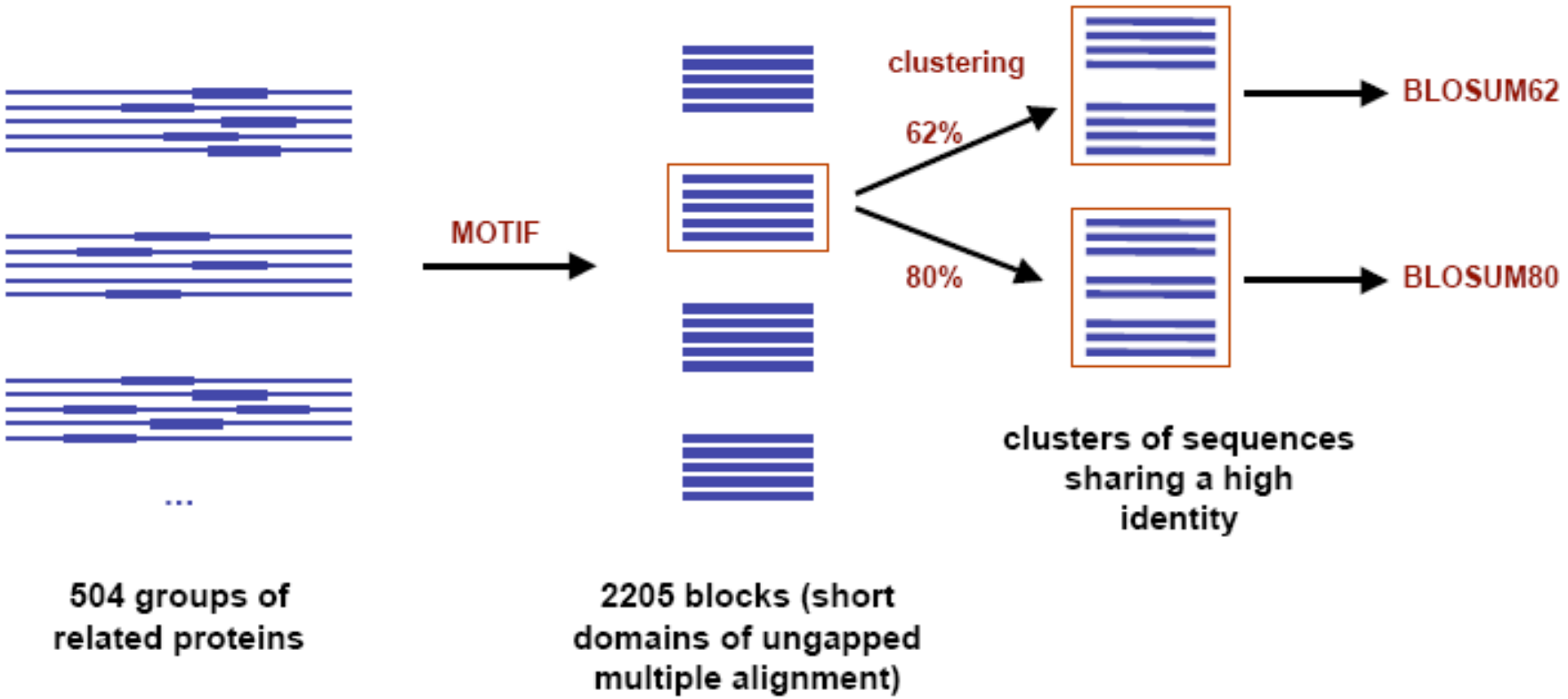
- In total, they examined 2,000 blocks *without gaps* and ~500 groups of related proteins by counting the number of matches and mismatches of each type of the 20 different amino acids.

- From the counts of each type, they created a ***frequency table*** and using these frequencies they further computed the ***probability*** of each type of match and mismatch and then ***converted the probabilities into logarithm of odds ratios***.

- To get the final scores in the matrix, they further converted the log-odds ratios into ***bit units*** and multiplied each bit score by a ***scaling factor of two*** and rounded to the nearest integer, producing the final scores in BLOSUM matrix.

- This way, the alignment score becomes ***zero*** if the observed frequencies are as ***expected***, ***negative*** score if frequencies are less than ***expected*** and ***positive*** score when the frequencies are over the ***expected*** frequencies.

Construction of BLOSUM matrices




```
   12345678
1  NLKIVSN…
2  DLKIVSN…
3  DLKIVSN…
4  DLKIVSN…
5  DLKIVSN…
6  DLKIVSN…
7  DLKIVSN…
8  DLKIVSN…
9  DLKIVSN…
10 DLKIVSN…
```

Construction of BLOSUM matrices

- Calculate the observed frequency q_{ij} of a pair (i, j) in the same column in the alignment block (A mutation can go in both directions, therefore the tally of A-V pair enters both q_{AV} and q_{VA} entries, while tally of A-A pair enters q_{AA} entry twice).

- Calculate the probability p_{ij} that a mutation occur between amino acid i and amino acid j.

$$p_{ij} = \frac{q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}}$$

- Calculate the marginal probability i.e. the expected probability of occurrence of amino acid i

$$p_i = \frac{\sum_{j=1,20} q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}}$$

- The final log-likelihood score is calculated as

$$S_{ij} = 2\log_2 \frac{p_{ij}}{p_i p_j}$$

An example of BLOSUM matrix construction

VVAPV
AAAPA
PVAPV
PAAAV

$N_{AA} = 0 + 1 + (4*3/2) + 0 + 0 = 7$

$N_{VV} = 0 + 1 + 0 + 0 + (3*2)/2 = 4$

$N_{PP} = 1 + 0 + 0 + (3*2)/2 + 0 = 4$

$N_{AV} = N_{VA} = 1 + 2*2 + 0 + 0 + 3 = 8$

$N_{AP} = N_{PA} = 2 + 0 + 0 + 3 + 0 = 5$

$N_{PV} = N_{VP} = 2 + 0 + 0 + 0 + 0 = 2$

N_{VP} is the number of V-P pairs

An example of BLOSUM matrix construction

q_{ij}=	A	V	P
A	14	8	5
V	8	8	2
P	5	2	8

q_{ij}: number of times amino acid j mutates to amino acid i.
A mutation could go in both directions, therefore the tally of A-V pair enters both q_{AV} and q_{VA} entries, while the tally of A-A pair enters q_{AA} entry twice.

An example of BLOSUM matrix construction

p_{ij} is the probability that a mutation occurs between amino acid i and amino acid j

$$p_{ij} = \frac{q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}}$$

p_{ij}=	A	V	P
A	14/60	8/60	5/60
V	8/60	8/60	2/60
P	5/60	2/60	8/60

An example of BLOSUM matrix construction

p_i is the marginal probability, meaning the expected probability of occurrence of amino acid i

$$p_i = \frac{\sum_{j=1,20} q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}}$$

p_i =	A	V	P
	9/20	6/20	5/20

An example of BLOSUM matrix construction

The BLOSUM log-likelihood matrix:

$$S_{ij} = 2\log_2 \frac{p_{ij}}{p_i p_j}$$

S_{ij}=	A	V	P
A	0.409		
V	-0.036	1.134	
P	-0.866	-2.34	2.19

An alignment that scores 6 means that the alignment by common ancestry is $2^{(6/2)}$=8 times as likely as expected than by chance.

An actual BLOSUM Matrix

BLOSUM 62 scoring matrix

(positive values are shaded)

	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	-3	**1**	**1**	-2	-1	-3	-2	**5**								
M	-1	-1	-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**

Note that these substitution matrices do not contain information for gap penalties and thus we need to assign them separately.

RBLOSUM62

- The BLOSUM family of substitution matrices, and particularly BLOSUM62, is the de facto standard in protein database searches and sequence alignments.

- The result of error in implementing the algorithm in coding 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.

- **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.**

BLOSUM62 miscalculations improve search performance

To the editor:

The BLOSUM[1] family of substitution matrices, and particularly BLOSUM62, is the *de facto* standard in protein database

The error that had the most impact was an incorrect normalization during a weighting procedure; this procedure, the error and its impact are discussed in greater

Styczynski et al., 2008, *Nat. Biotech*. 26, 274–275

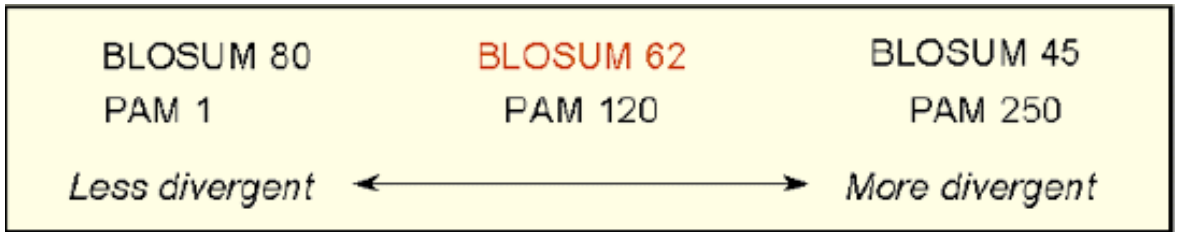
PAM vs BLOSUM Matrices

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**.

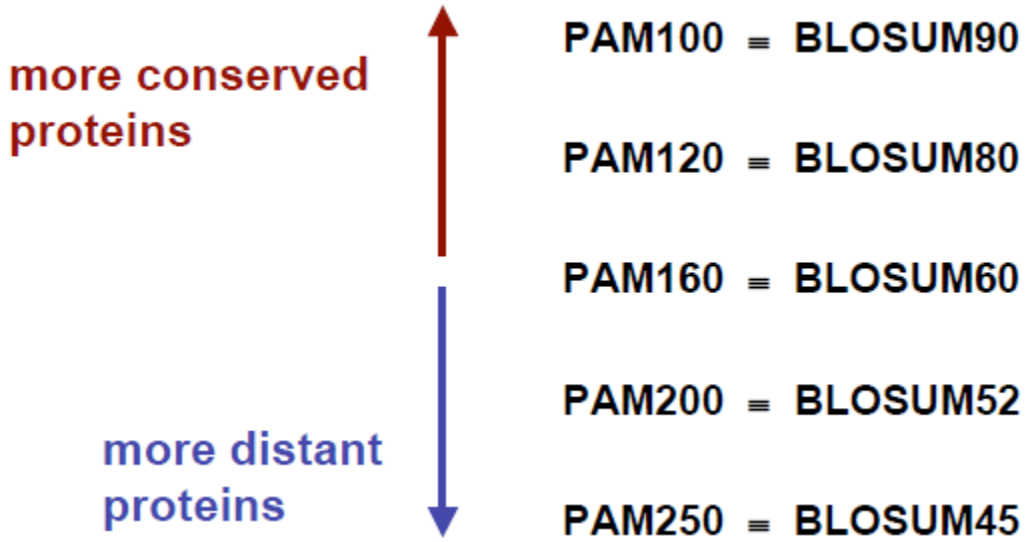


PAM vs BLOSUM Matrices

- The typical general purpose scoring matrices for protein sequence alignment are BLOSUM, PAM, variable time (VT), variable time maximum likelihood (VTML), and more recently published PFASUM matrices.

- In contrast to PAM matrices, in which the number refers to percentage differences, in BLOSUM matrices the number refers to similarity.

- A comparison of the matrices can be done on the basis on their "information content"



Gap costs

BLOSUM 62 scoring matrix

(positive values are shaded)

	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	-3	**1**	**1**	-2	-1	-3	-2	**5**								
M	-1	-1	-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**

Note that these substitution matrices do not contain information for gap penalties and thus we need to assign them separately.

Gap Penalties

The inclusion of gaps and gap penalties is necessary in order to obtain the best possible alignment between two sequences. A gap opening penalty for any gap (g) and a gap extension penalty for each element in the gap (r) is most often used, to give a total gap score w_x , according to the equation

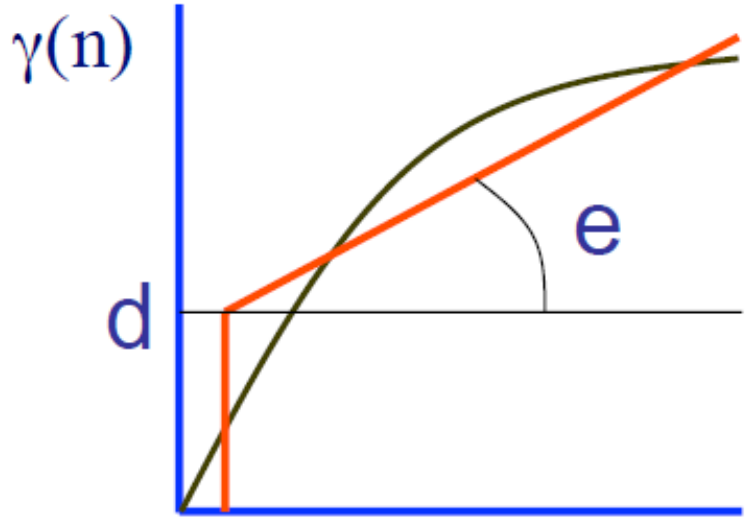
$$w_x = g + rx$$

where x is the length of the gap.

$$\gamma(n) = d + (n - 1) \times e$$

gap open

gap extend



Thank You