Pairwise and Multiple Sequence Alignment

Substitution Matrix

A substitution matrix contains values proportional to the probability that amino acid *i* mutates into amino acid *j* for all pairs of amino acids.

Substitution matrices are constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.

Substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution.

The two major types of substitution matrices are PAM and BLOSUM.

PAM(Point Accepted Mutation)

- PAM matrices were introduced by Margaret Dayhoff in 1978.
- The calculation of these matrices were based on 1572 observed mutations in the phylogenetic trees of 71 families of closely related proteins.
- The proteins to be studied were selected on the basis of having high similarity with their predecessors.

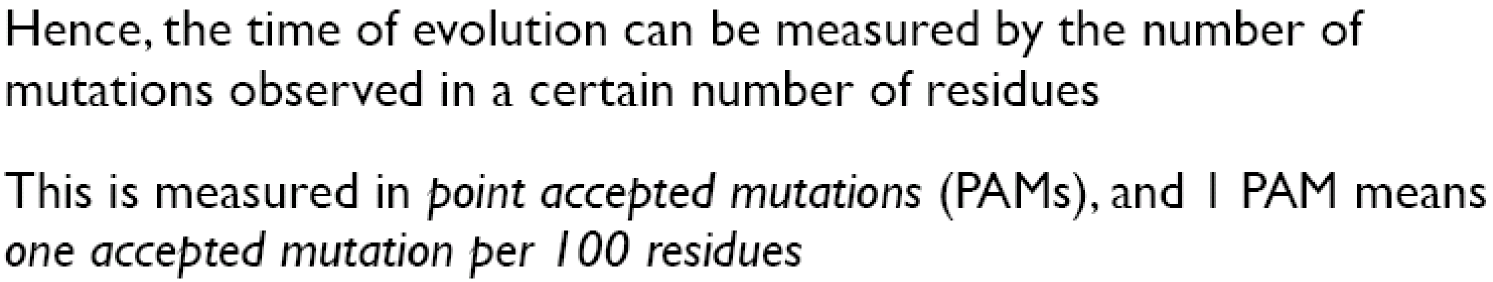
Point-accepted mutations

PAM matrices are based on global alignments of closely related proteins.

The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

Other PAM matrices are extrapolated from PAM1.

All the PAM data come from closely related proteins (>85% amino acid identity)



Hence, the time of evolution can be measured by the number of mutations observed in a certain number of residues

This is measured in *point accepted mutations* (PAMs), and 1 PAM means *one accepted mutation per 100 residues*

Dayhoff’s PAM1 mutation probability matrix

l amino acid

Replaced amino acid

	A Ala	R Ar	[illegible]	[illegible]	[illegible]	Q Gln	E Glu	G Gly	H His	I Ile
A	9867	2	[illegible]	[illegible]	[illegible]	8	17	21	2	6
R	1	9913	1	0	1	10	0	0	10	3
N	4	1	9822	36	0	4	6	6	21	3
D	6	0	42	[illegible]	[illegible]	[illegible]	[illegible]	6	4	1
C	1	1	0	[illegible]	[illegible]	[illegible]	[illegible]	0	1	1
Q	3	9	4	[illegible]	[illegible]	[illegible]	[illegible]	1	23	1
E	10	0	7	[illegible]	0	35	9865	4	2	3
G	21	1	12	11	1	3	7	9935	1	0
H	1	8	18	3	1	20	1	0	9912	0
I	2	2	3	1	2	1	2	0	0	9872

There is 98.67%chance that A will be replaced by A over an evolutionary distance of 1 PAM

Each element shows the probability that an original amino acid j (columns)will be replaced byanother amino acid i (rows) for 1% sequence divergence

Dayhoff's PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

PAM0	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A	100%	0%	0%	0%	0%	0%	0%	0%
R	0%	100%	0%	0%	0%	0%	0%	0%
N	0%	0%	100%	0%	0%	0%	0%	0%
D	0%	0%	0%	100%	0%	0%	0%	0%
C	0%	0%	0%	0%	100%	0%	0%	0%
Q	0%	0%	0%	0%	0%	100%	0%	0%
E	0%	0%	0%	0%	0%	0%	100%	0%
G	0%	0%	0%	0%	0%	0%	0%	100%

Top: original amino acid
Side: replacement amino acid

Dayhoff's PAM2000 mutation probability matrix: the rules for very distantly related proteins

PAM∞	A	R	N	D	C	Q	E	G
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly
A	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%
R	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%
N	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%
D	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%
C	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%
Q	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%
E	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%
G	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%

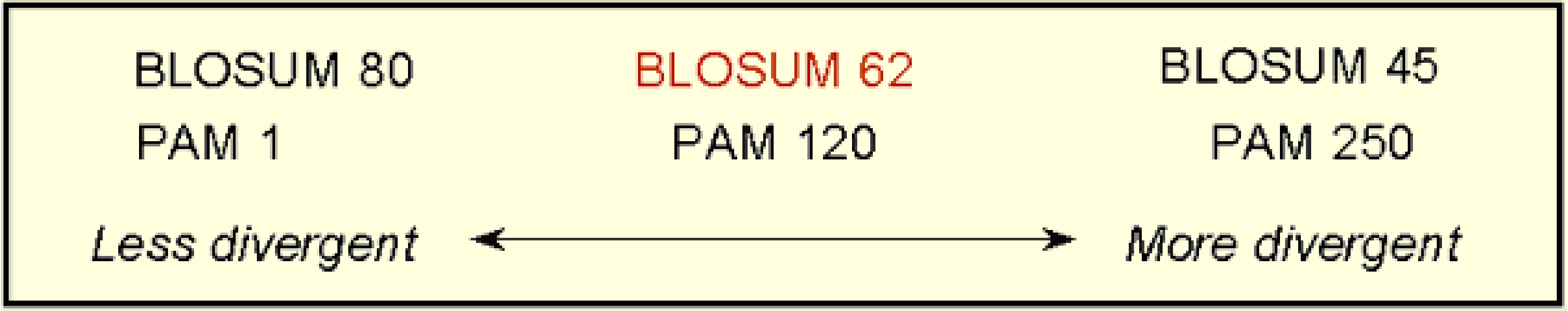
PAM1 matrix is multiplied 2000 times by itself

Top: original amino acid
Side: replacement amino acid

PAM250 mutation probability matrix

	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	5	4	15	10	4	10	5	5	5	7	2	4	17

Top: original amino acid
Side: replacement amino acid



BLOSUM 80 BLOSUM 62 BLOSUM 45

PAM 1 PAM 120 PAM 250

Less divergent ↔ *More divergent*

BLOSUM

BLOSUM matrices are based on local alignments.

BLOSUM stands for blocks substitution matrix.

BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

BLOSUM Scoring Matrices

- In the Dayhoff model, the scoring values are derived from protein sequences with at least 85% identity

- Alignments are, however, most often performed on sequences of less similarity, and the scoring matrices for use in these cases are calculated from the 1 PAM matrix

- Henikoff and Henikoff (1992) have therefore developed scoring matrices based on known alignments of more diverse sequences

BLOSUM Scoring Matrices

- BLOck SUbstitution Matrix

- Based on comparisons of blocks of sequences derived from the Blocks database

- The Blocks database contains multiply aligned un-gapped segments corresponding to the most highly conserved regions of proteins (local alignment versus global alignment)

Conserved blocks in alignments

```
AABCDA...BBCDA
DABCDA.A.BBCBB
BBBCDABA.BCCAA
AAACDAC.DCBCDB
CCBADAB.DBBDCC
AAACAA...BBCCC
```

Constructing BLOSUM *r*

- To avoid bias in favor of a certain protein, first eliminate sequences that are more than *r*% identical

- The elimination is done by either
  - removing sequences from the block, or
  - finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster.

- BLOSUM *r* is the matrix built from blocks with no more the *r*% of similarity
  - E.g., BLOSUM62 is the matrix built using sequences with no more than 62% similarity.
  - Note: BLOSUM 62 is the default matrix for protein BLAST

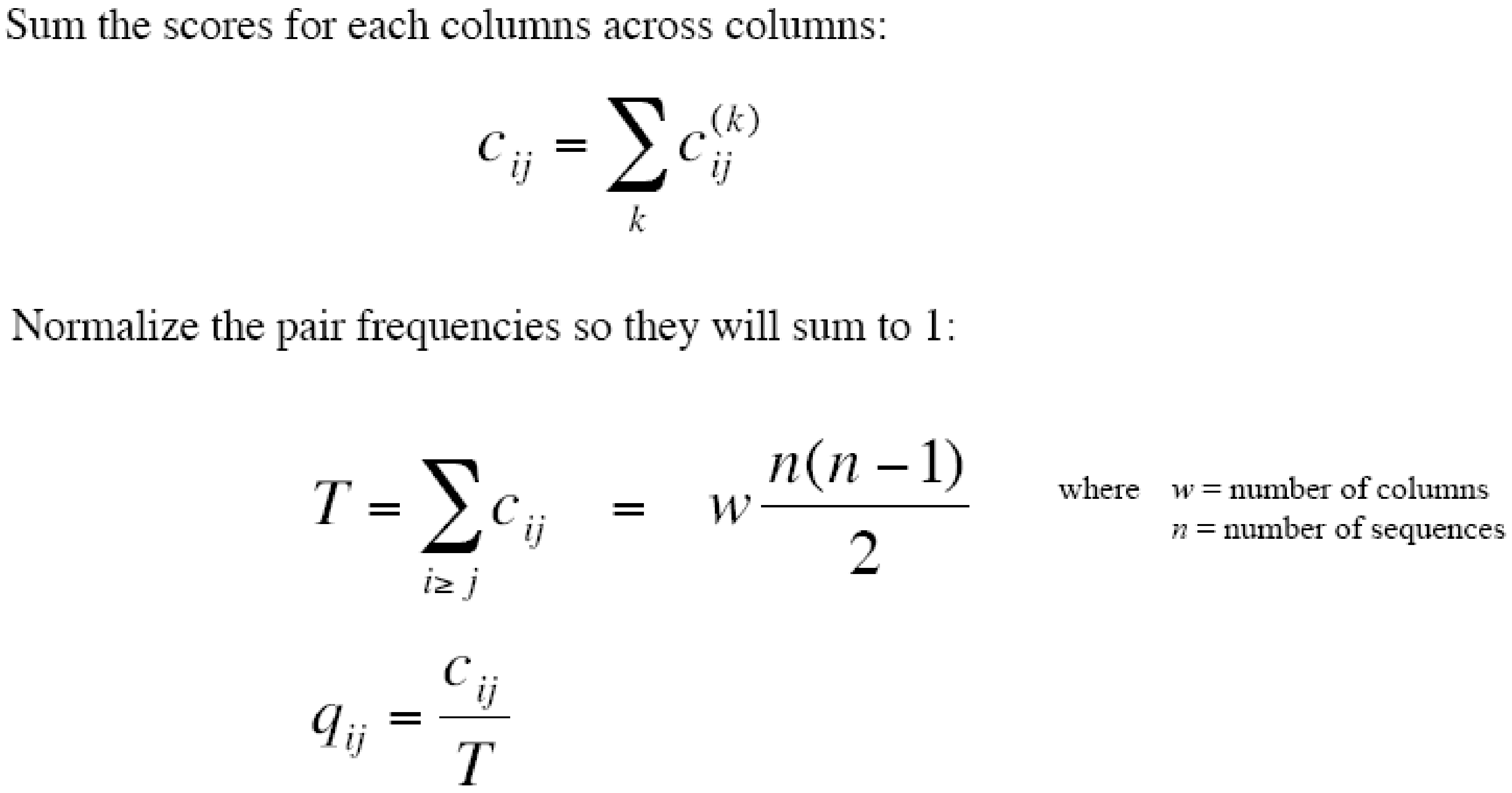
Collecting substitution statistics

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

    $s(X,Y) = \log (q_{XY} / (p_X\, p_y))$

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

Computing probabilities



Sum the scores for each columns across columns:

$$c_{ij} = \sum_{k} c_{ij}^{(k)}$$

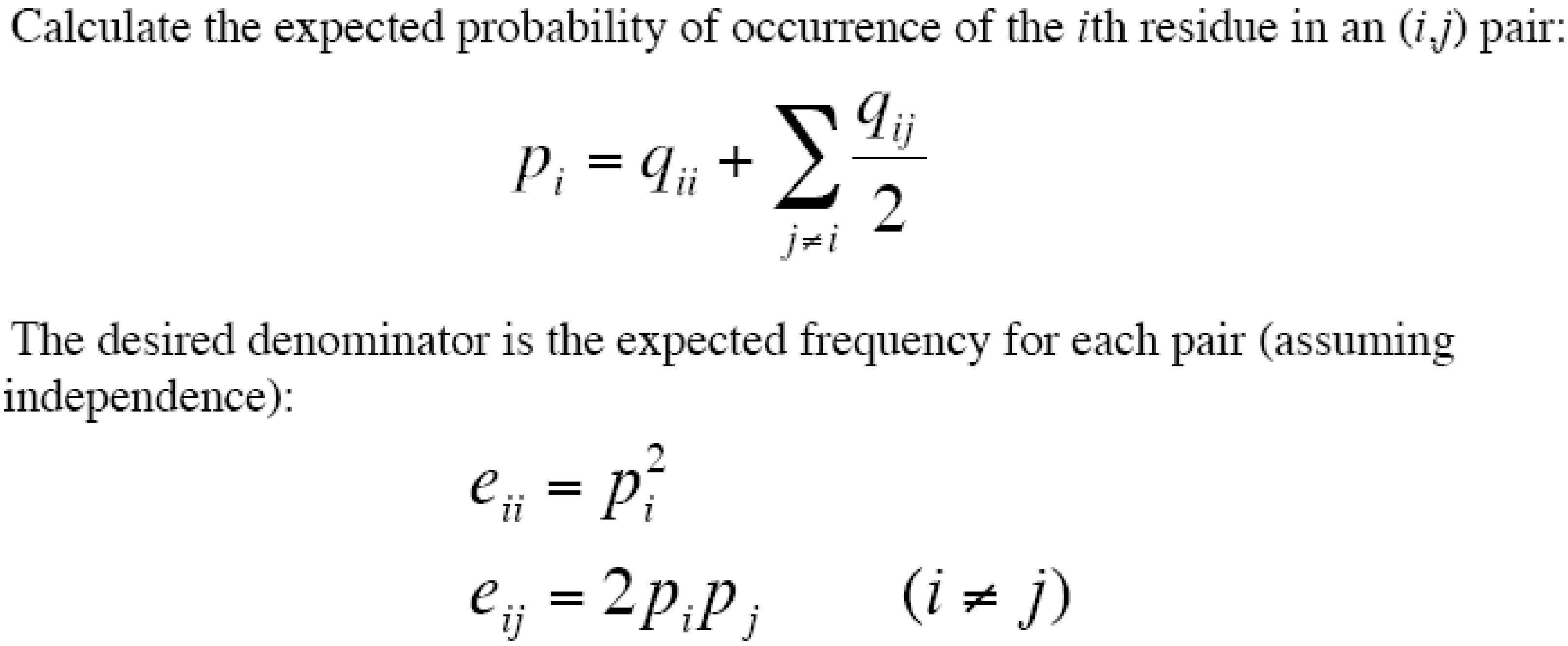
Normalize the pair frequencies so they will sum to 1:

$$T = \sum_{i \geq j} c_{ij} \quad = \quad w\frac{n(n-1)}{2}$$

where w = number of columns
n = number of sequences

$$q_{ij} = \frac{c_{ij}}{T}$$

Computing probabilities



Calculate the expected probability of occurrence of the *i*th residue in an (*i*,*j*) pair:

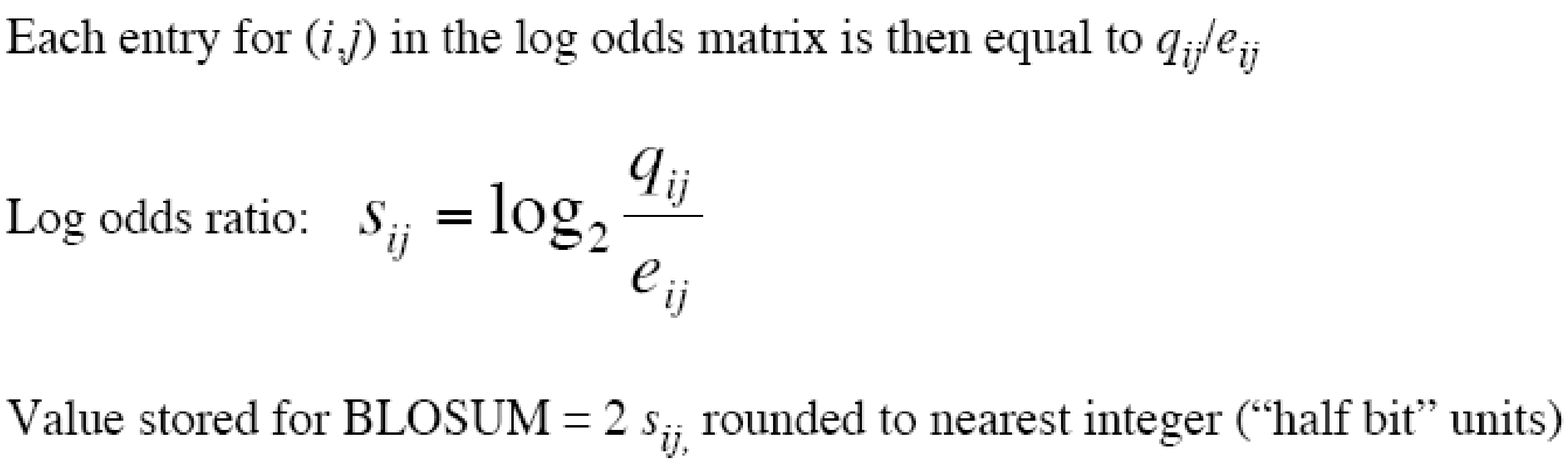
$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2}$$

The desired denominator is the expected frequency for each pair (assuming independence):

$$e_{ii} = p_i^2$$

$$e_{ij} = 2p_i p_j \qquad (i \neq j)$$

Computing probabilities

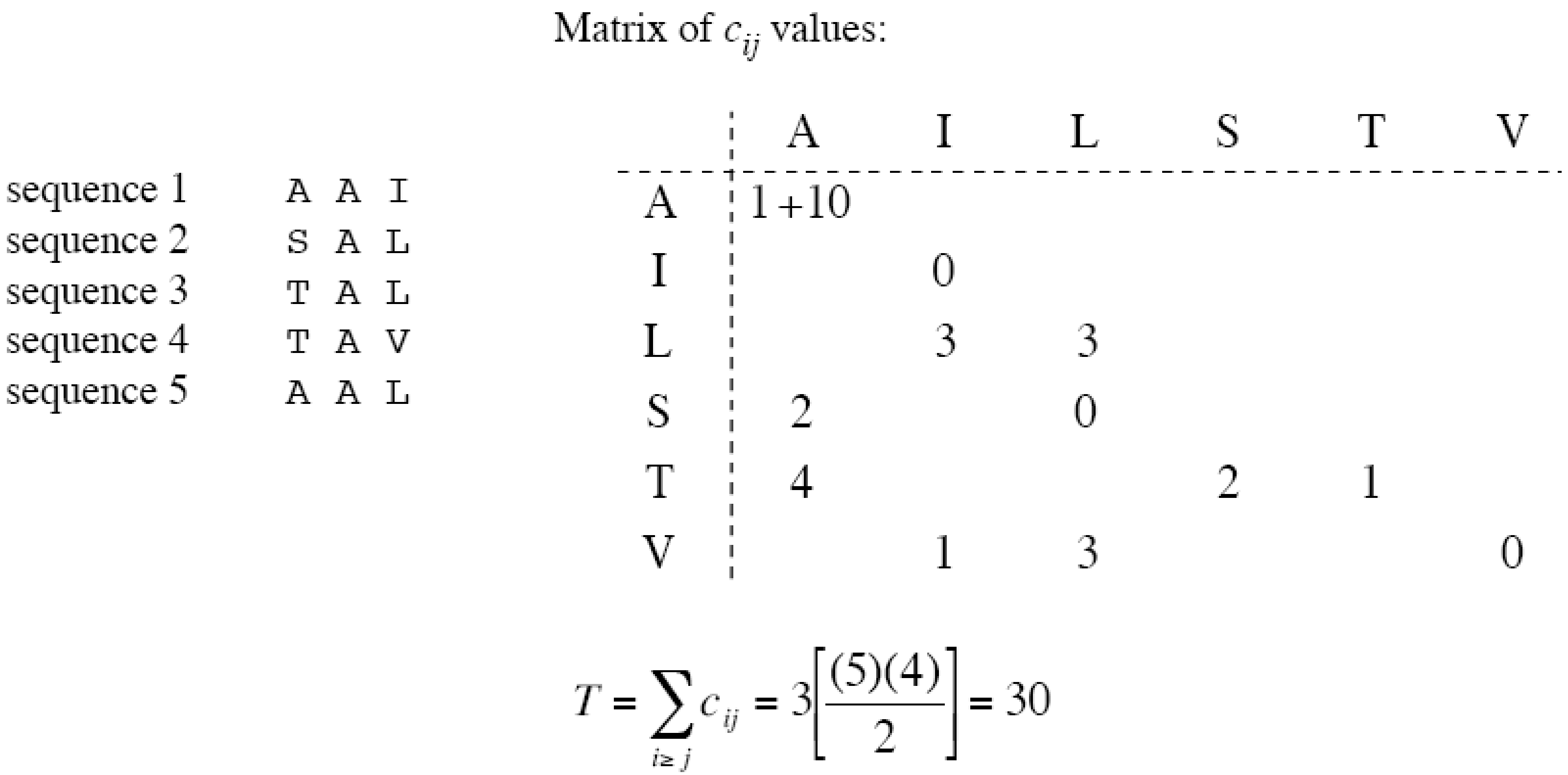


Each entry for (i,j) in the log odds matrix is then equal to q_{ij}/e_{ij}

Log odds ratio: $$s_{ij} = \log_2 \frac{q_{ij}}{e_{ij}}$$

Value stored for BLOSUM = 2 $s_{ij,}$ rounded to nearest integer (“half bit” units)

Example



Matrix of c_{ij} values:

sequence 1	A A I
sequence 2	S A L
sequence 3	T A L
sequence 4	T A V
sequence 5	A A L

$$\begin{array}{c|cccccc}
 & A & I & L & S & T & V \\
\hline
A & 1+10 & & & & & \\
I & & 0 & & & & \\
L & & 3 & 3 & & & \\
S & 2 & & 0 & & & \\
T & 4 & & & 2 & 1 & \\
V & & 1 & 3 & & & 0
\end{array}$$

$$T = \sum_{i \geq j} c_{ij} = 3\left[\frac{(5)(4)}{2}\right] = 30$$

From http://www.csit.fsu.edu/~swofford/bioinformatics_spring05/lectures/lecture03-blosum.pdf

Example

Matrix of q_{ij} values:

$$
\begin{array}{c|cccccc}
 & \mathrm{A} & \mathrm{I} & \mathrm{L} & \mathrm{S} & \mathrm{T} & \mathrm{V} \\
\hline
\mathrm{A} & 11/30 & & & & & \\
\mathrm{I} & & 0 & & & & \\
\mathrm{L} & & 3/30 & 3/30 & & & \\
\mathrm{S} & 2/30 & & 0 & 0 & & \\
\mathrm{T} & 4/30 & & & 2/30 & 1/30 & \\
\mathrm{V} & & 1/30 & 3/30 & & & 0
\end{array}
$$

$$
=
\begin{array}{c|cccccc}
 & \mathrm{A} & \mathrm{I} & \mathrm{L} & \mathrm{S} & \mathrm{T} & \mathrm{V} \\
\hline
\mathrm{A} & 0.36\overline{6} & & & & & \\
\mathrm{I} & 0 & 0 & & & & \\
\mathrm{L} & 0 & 0.1 & 0.1 & & & \\
\mathrm{S} & 0.06\overline{6} & 0 & 0 & & & \\
\mathrm{T} & 0.13\overline{3} & 0 & 0 & 0.06\overline{6} & 0.03\overline{3} & \\
\mathrm{V} & 0 & 0.03\overline{3} & 0.1 & 0 & 0 & 0
\end{array}
$$

Vector of p_i values:

$$p_{\mathrm{A}} = \left(11 + \frac{6}{2}\right)\Big/30 = 14/30 = 0.46\overline{6}$$

$$p_{\mathrm{I}} = \left(0 + \frac{4}{2}\right)\Big/30 = 2/30 = 0.06\overline{6}$$

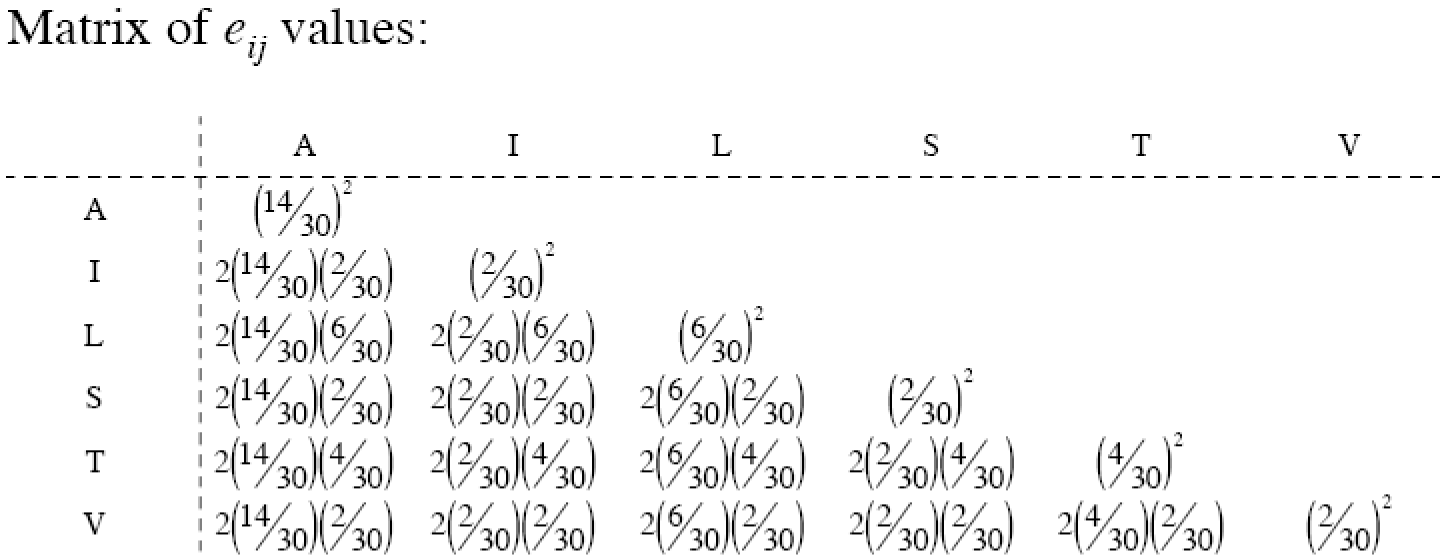
$$p_{\mathrm{L}} = \left(3 + \frac{6}{2}\right)\Big/30 = 6/30 = 0.2$$

$$p_{\mathrm{S}} = \left(0 + \frac{4}{2}\right)\Big/30 = 2/30 = 0.06\overline{6}$$

$$p_{\mathrm{T}} = \left(1 + \frac{6}{2}\right)\Big/30 = 4/30 = 0.13\overline{3}$$

$$p_{\mathrm{V}} = \left(0 + \frac{4}{2}\right)\Big/30 = 2/30 = 0.06\overline{6}$$

Example



Matrix of e_{ij} values:

	A	I	L	S	T	V
A	$\left(\frac{14}{30}\right)^2$					
I	$2\left(\frac{14}{30}\right)\left(\frac{2}{30}\right)$	$\left(\frac{2}{30}\right)^2$				
L	$2\left(\frac{14}{30}\right)\left(\frac{6}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{6}{30}\right)$	$\left(\frac{6}{30}\right)^2$			
S	$2\left(\frac{14}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{6}{30}\right)\left(\frac{2}{30}\right)$	$\left(\frac{2}{30}\right)^2$		
T	$2\left(\frac{14}{30}\right)\left(\frac{4}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{4}{30}\right)$	$2\left(\frac{6}{30}\right)\left(\frac{4}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{4}{30}\right)$	$\left(\frac{4}{30}\right)^2$	
V	$2\left(\frac{14}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{6}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{4}{30}\right)\left(\frac{2}{30}\right)$	$\left(\frac{2}{30}\right)^2$

Example

Log odds ratio:

$$\text{e.g.,}\quad s_{\text{AA}} = \log_2 \frac{0.36\overline{6}}{\left({}^{14}\!/\!{}_{30}\right)^2} = \log_2 1.6837 = 0.7516$$

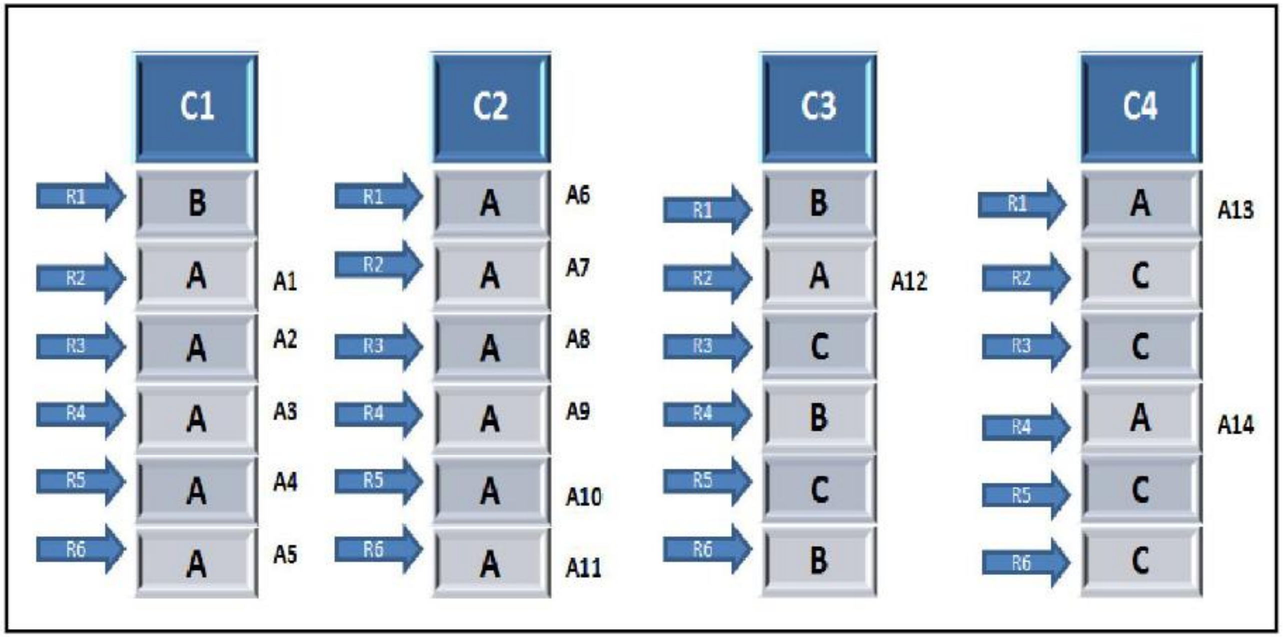
BLOSUM value for AA = $round(2 \cdot 0.7516) = 2$

Full matrix:

	A	I	L	S	T	V
A	2					
I	?	?				
L	?	4	3			
S	0	?	?	?		
T	0	?	?	4	2	
V	?	4	4	?	?	?

Note: undefined values result from unobserved pairs (would ordinarily not happen with real data)

Example



	C1		C2		C3		C4	
R1	B		A	A6	B		A	A13
R2	A	A1	A	A7	A	A12	C	
R3	A	A2	A	A8	C		C	
R4	A	A3	A	A9	B		A	A14
R5	A	A4	A	A10	C		C	
R6	A	A5	A	A11	B		C	

https://comp.utm.my/saberi/files/2014/12/blosum.pdf

Example

Since it is getting 4 column in this sample, so we times 4 to gain the total of frequency of all pairs.

=15*4

=60

A aligning with another A (AA)

P(E) of AA = A*A

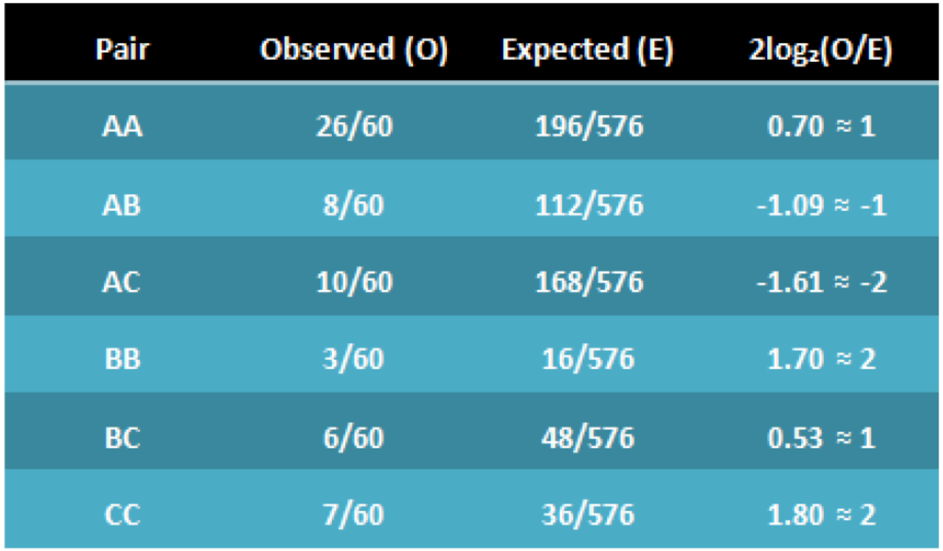
= (14/24)*(14/24)

= 196/576

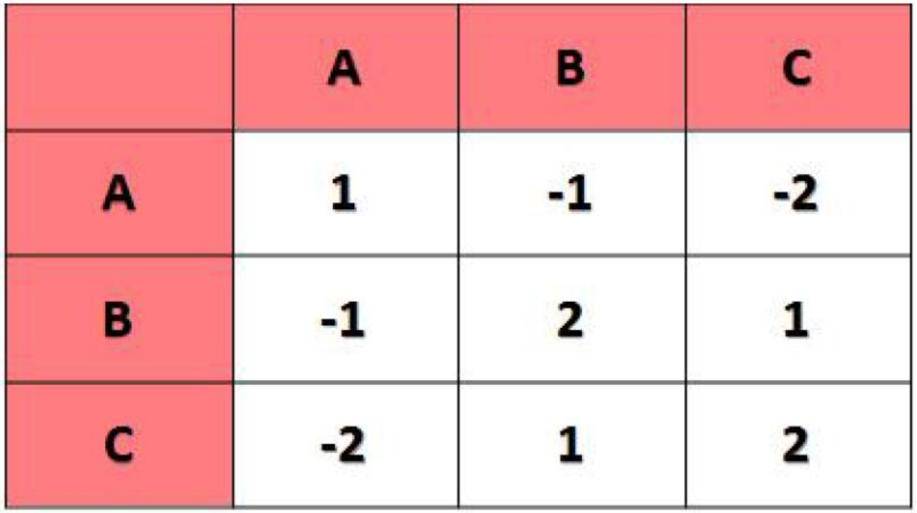
Pair	Observed (O)	Expected (E)	$2\log_2(O/E)$
AA	26/60	196/576	0.70 ≈ 1
AB	8/60	112/576	-1.09 ≈ -1
AC	10/60	168/576	-1.61 ≈ -2
BB	3/60	16/576	1.70 ≈ 2
BC	6/60	48/576	0.53 ≈ 1
CC	7/60	36/576	1.80 ≈ 2

https://comp.utm.my/saberi/files/2014/12/blosum.pdf

Example



Pair	Observed (O)	Expected (E)	$2\log_2(O/E)$
AA	26/60	196/576	0.70 ≈ 1
AB	8/60	112/576	-1.09 ≈ -1
AC	10/60	168/576	-1.61 ≈ -2
BB	3/60	16/576	1.70 ≈ 2
BC	6/60	48/576	0.53 ≈ 1
CC	7/60	36/576	1.80 ≈ 2

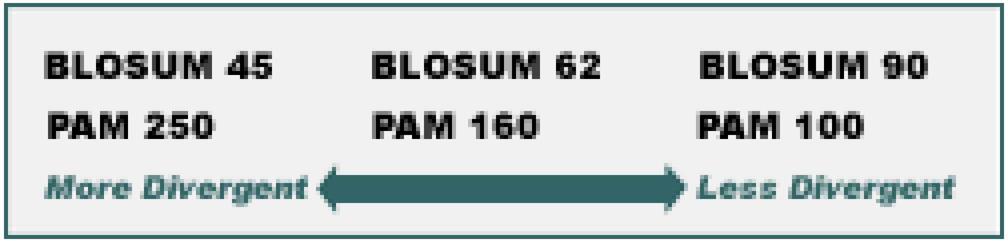


	A	B	C
A	1	-1	-2
B	-1	2	1
C	-2	1	2

https://comp.utm.my/saberi/files/2014/12/blosum.pdf

Comparison

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



BLOSUM 45	BLOSUM 62	BLOSUM 90
PAM 250	PAM 160	PAM 100