PAM-1 Matrix × 10,000

	From:	Ala	Arg	Asn	Asp	Cys	Gln	Glu	
		A	R	N	D	C	Q	E	
To:									
Ala	A	9867	2	9	10	3	8	17	···
Arg	R	1	9913	1	0	1	10	0	···
Asn	N	4	1	9822	36	0	4	6	···
Asp	D	6	0	42	9859	0	6	53	···
Cys	C	1	1	0	0	9973	0	0	···
Gln	Q	3	9	4	5	0	9876	27	···
Glu	E	10	0	7	56	0	35	9865	···

PAM1 is the expectation after approximately 1% of the sequence has been substituted.

PAM2 is calculated as PAM1 $\times$ PAM1

PAMx is calculated as PAM(x-1) $\times$ PAM1

PAM250 is generally used for distant comparisons. It corresponds to 2.5 differences per site ($\sim$20% identity).

NOTE: These measure divergence _not_ time.

PAM-250 Matrix $\times$ 100

To:	From:	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	
Ala	A	13	6	9	9	5	8	9	$\cdots$
Arg	R	3	17	4	3	2	5	3	$\cdots$
Asn	N	4	4	6	7	2	5	6	$\cdots$
Asp	D	5	4	8	11	1	7	10	$\cdots$
Cys	C	2	1	1	1	52	1	1	$\cdots$
Gln	Q	3	5	5	6	1	10	7	$\cdots$
Glu	E	5	4	7	11	1	9	12	$\cdots$

PAM scoring matrix

The PAM scoring values are generally shown as a symmetric "*log odds ratio*" matrix.

Odds (for those who do not gamble) are $\frac{p}{1-p}$ where p is the probability of an event and $1-p$ is the probability of some other event. For example if $p = 0.5$ then the odds are $50/50$ or 1 to 1 ($\frac{0.5}{0.5} = 1$). While if $p = 0.75$ then the odds are 3 to 1 ($\frac{0.75}{0.25} = 3$).

The odds ratio is the ratio of the odds for and against.

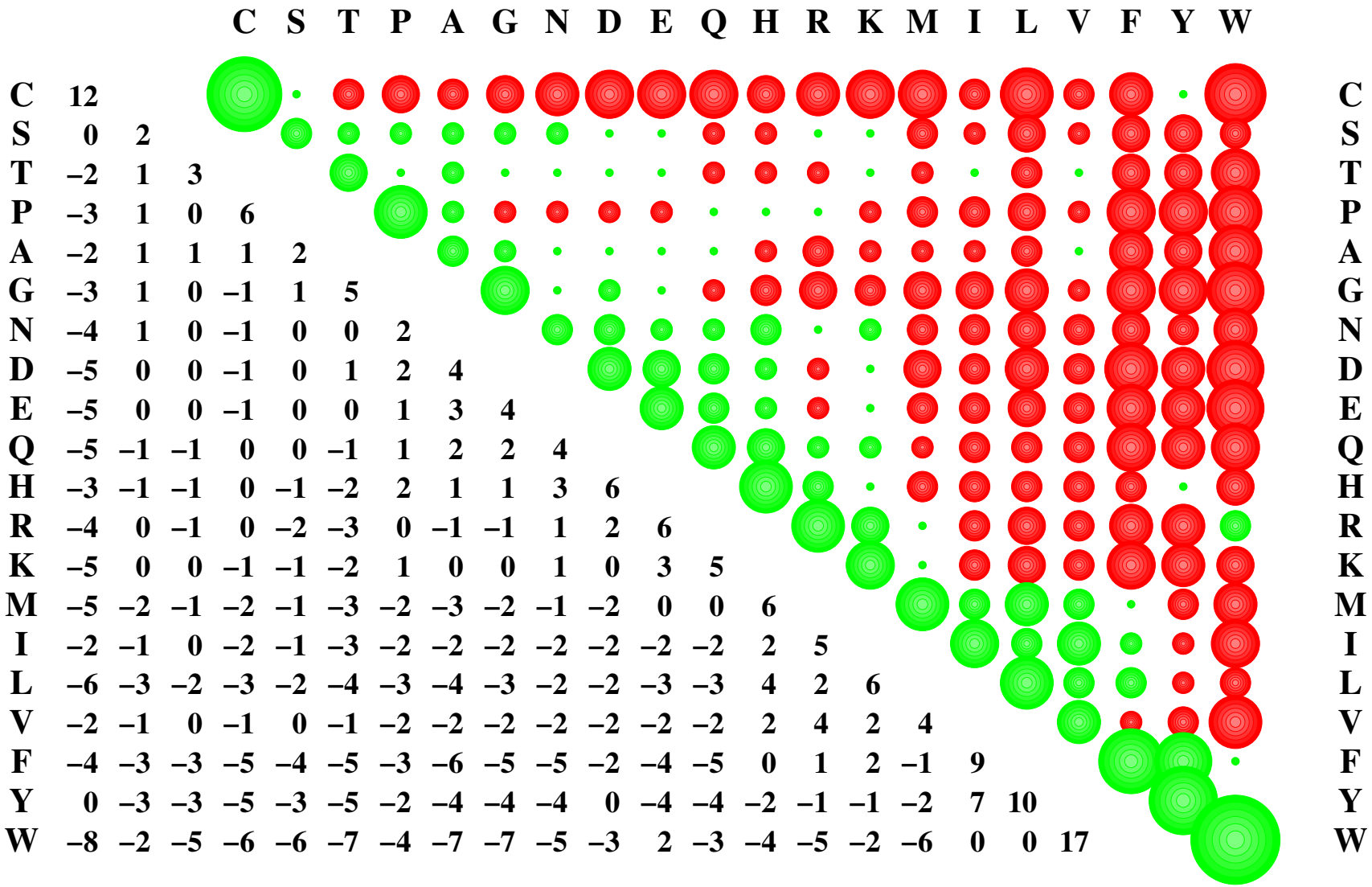
PAM scoring matrix

Generally the odds are presented as log values. For PAM matrices it is generally $\log_{10}$ that is used and so each integer value represents an order of magnitude. For example if $p = 0.08$, odds are $0.08/0.92 = 0.087$ (11 to 1) and log odds are $\log_{10}(0.087) = -1.06$ while if $p = 0.996$, odds are $0.996/0.004 = 249$ (order magnitude larger and opposite direction), the log odds are $\log_{10}(249) = +2.40$.

For a PAM scoring matrix

$$S_{ij} = \log \frac{p_i \cdot M_{ij}}{p_i \cdot p_j} = \log \frac{M_{ij}}{p_j} = \log \frac{\textit{observed frequency}}{\textit{expected frequency}}$$

This matrix will be symmetric.



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

Values multiplied by 10.

A **log odds of zero** implies the two amino acids are found across from each in an alignment as often as expected by chance (given their mutabilities and frequencies of occurrence).

A **log odds greater than zero** implies the two amino acids are found across from each in an alignment more often than expected by chance (given their mutabilities and frequencies of occurrence).

A **log odds less than zero** implies the two amino acids are found across from each in an alignment less often than expected by chance (given their mutabilities and frequencies of occurrence).

Two uses for PAM matrices,

Scoring matrix	–	**PAM250 (very distant)**
	–	**PAM160 (distant)**
	–	**PAM70 (less distant)**
	–	**PAM30 (more similar)**
	–	**etc**
Transition matrix	–	**PAM1**

PAM-1 Matrix × 10,000

	From:	Ala	Arg	Asn	Asp	Cys	Gln	Glu	
		A	R	N	D	C	Q	E	
To:									
Ala	A	9867	2	9	10	3	8	17	···
Arg	R	1	9913	1	0	1	10	0	···
Asn	N	4	1	9822	36	0	4	6	···
Asp	D	6	0	42	9859	0	6	53	···
Cys	C	1	1	0	0	9973	0	0	···
Gln	Q	3	9	4	5	0	9876	27	···
Glu	E	10	0	7	56	0	35	9865	···

PAM - strange (?) patterns

Lots of interesting properties

Many exchanges between amino acids D and E

Far more double codon substitutions than expected

Fewer of some single codon substitutions; e.g. G and W

PAM - scoring an amino acid alignment

Consider an alignment . . .

Seq1	C	G	N	G
Seq2	C	G	D	R
PAM250	12	5	2	-3

Total score is $12 + 5 + 2 - 3 = 16$

The chances of getting an alignment this good by chance is given by the odds. Normally one would multiply the odds at each site (assuming independence) but since log's have been taken we can add the log odds.

The $\log_{10}$ odds of 1.6 corresponds to odds of 39.8. So this is an unusual similarity between these two peptides despite their length (in large part due to rare cysteines across from each other).

The PAM matrix was computed on globular proteins and may therefore not be a good representation of the substitution matrix for membrane or other non-globular proteins.

It assumes that all sites are equally mutable (but not all residues).

Only a limited number of proteins were available in comparison to the huge numbers today.

The JTT matrix (Jones, Taylor, Thornton 1992) was an update of the PAM matrix.

It is mostly used as a transition matrix rather than as a scoring matrix (for the later purpose PAM250 still seems the method of choice).

A matrix of BLOCKS

BLOcks **SU**bstitution **M**atrix

Based on the analysis of conserved proteins regions from the BLOCKS database.

More reliable than the PAM matrix for distantly related proteins

Default for BLAST searches

Used in many other programs including FASTA

BLOSUM matrix

1. Find the frequency of occurrence of one amino acid

$$p_i = q_{ii} + \sum q_{ij}/2$$

2. Expected frequencies

$$e_{ij} = p_i^2 \qquad \text{if} \quad i = j$$

$$e_{ij} = 2p_i p_j \qquad \text{if} \quad i \neq j$$

3. Score

$$s_{ij} = 2\log_2(q_{ij}/e_{ij})$$

The matrix consist of the scores ...

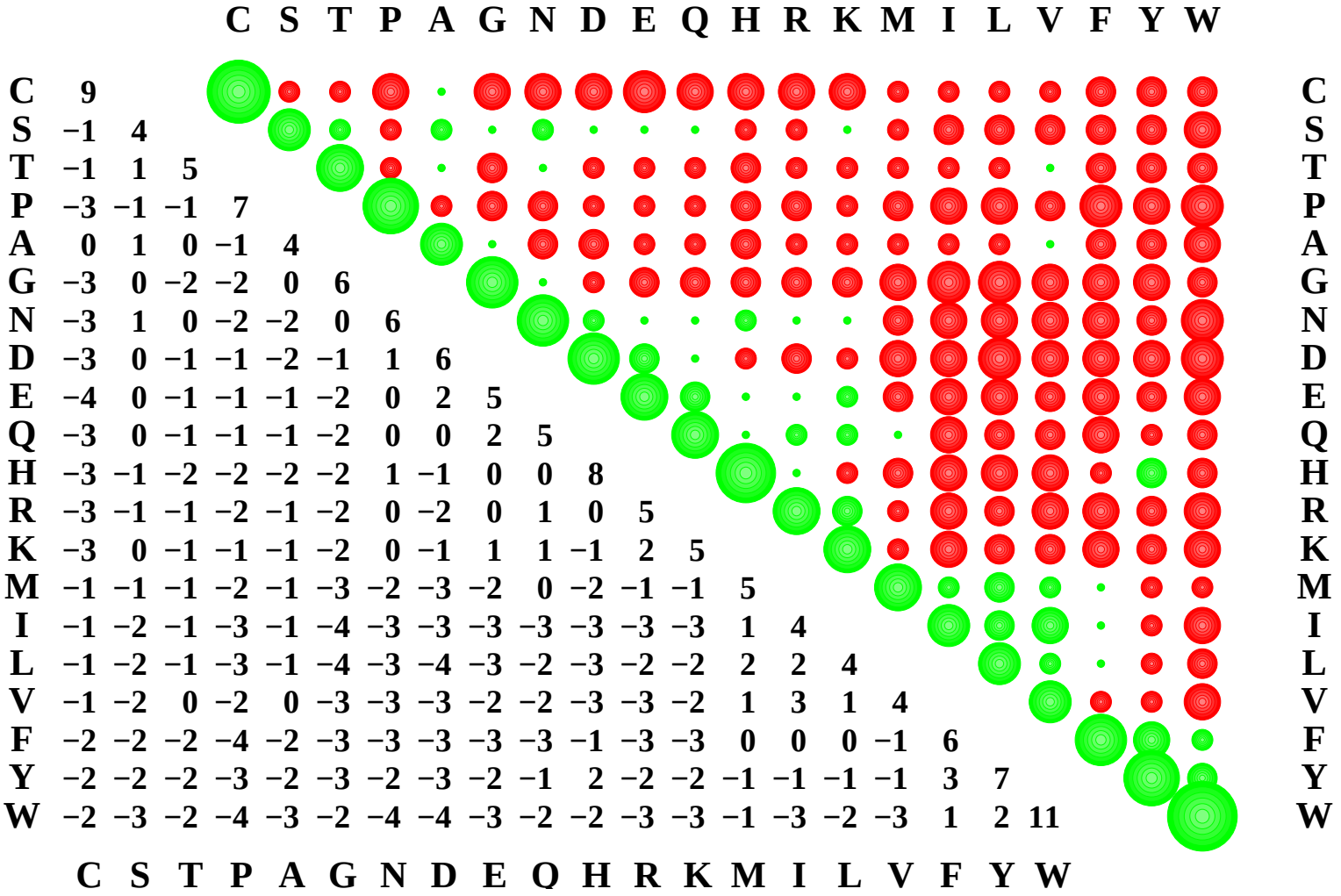
$$s_{ij} = 2 \log_2(q_{ij}/e_{ij}).$$

If the observed number of differences between a pair of amino acids is equal to the expected number then $s_{ij} = 0$

If the observed is less than expected then $s_{ij} < 0$

If the observed is greater than expected $s_{ij} > 0$

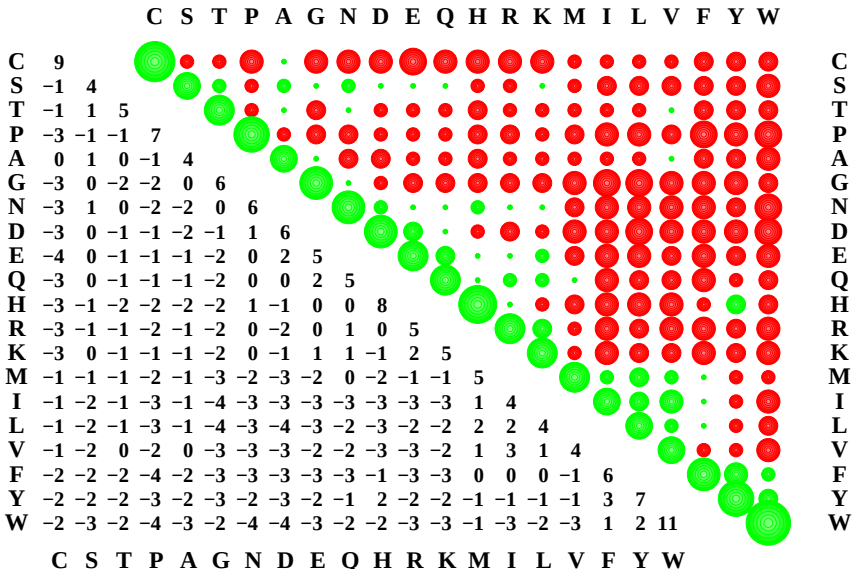
BLOSUM matrix



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

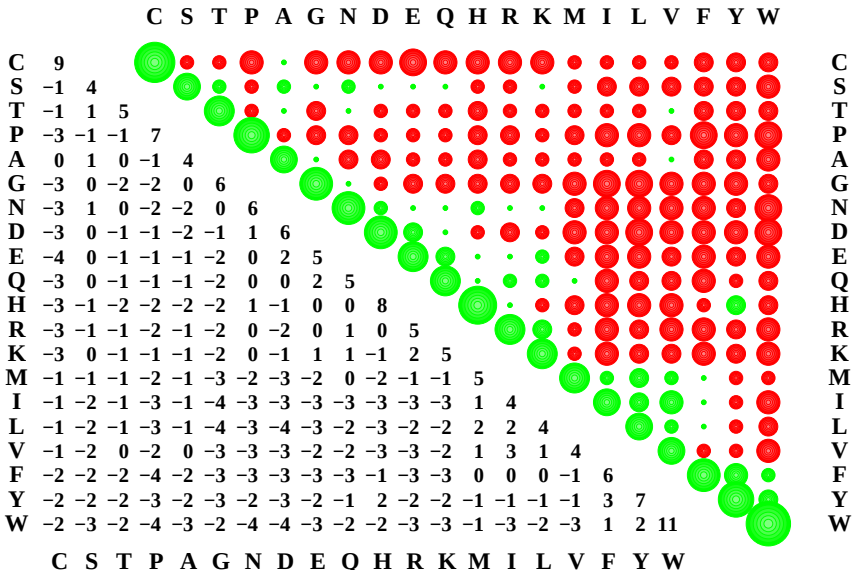
The lower left gives the log odds matrix (BLOSUM62).

BLOSUM matrix



The BLOSUM matrix is less tolerant of substitutions to or from hydrophilic amino acids, but more tolerant of hydrophobic changes, cysteine, and tryptophan mismatches than a similar level PAM matrix.

BLOSUM matrix



This is a BLOSUM62 matrix. It is roughly equivalent to a PAM160 matrix. The levels come from weighting different entries. In this case all proteins within 62% identity sum to a weight of 1.

NCBI's recommendations

Query length	Substitution matrix	Gap costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
>85	BLOSUM-62	(11,1)

Empirical measures still seem to work best despite many advances.

GONNET matrix

- Uses classical distance measures to produce protein alignments
- Given the alignments it computes a new distance matrix
- Align again using the new distance matrix
- Repeat this process many times

GONNET matrix

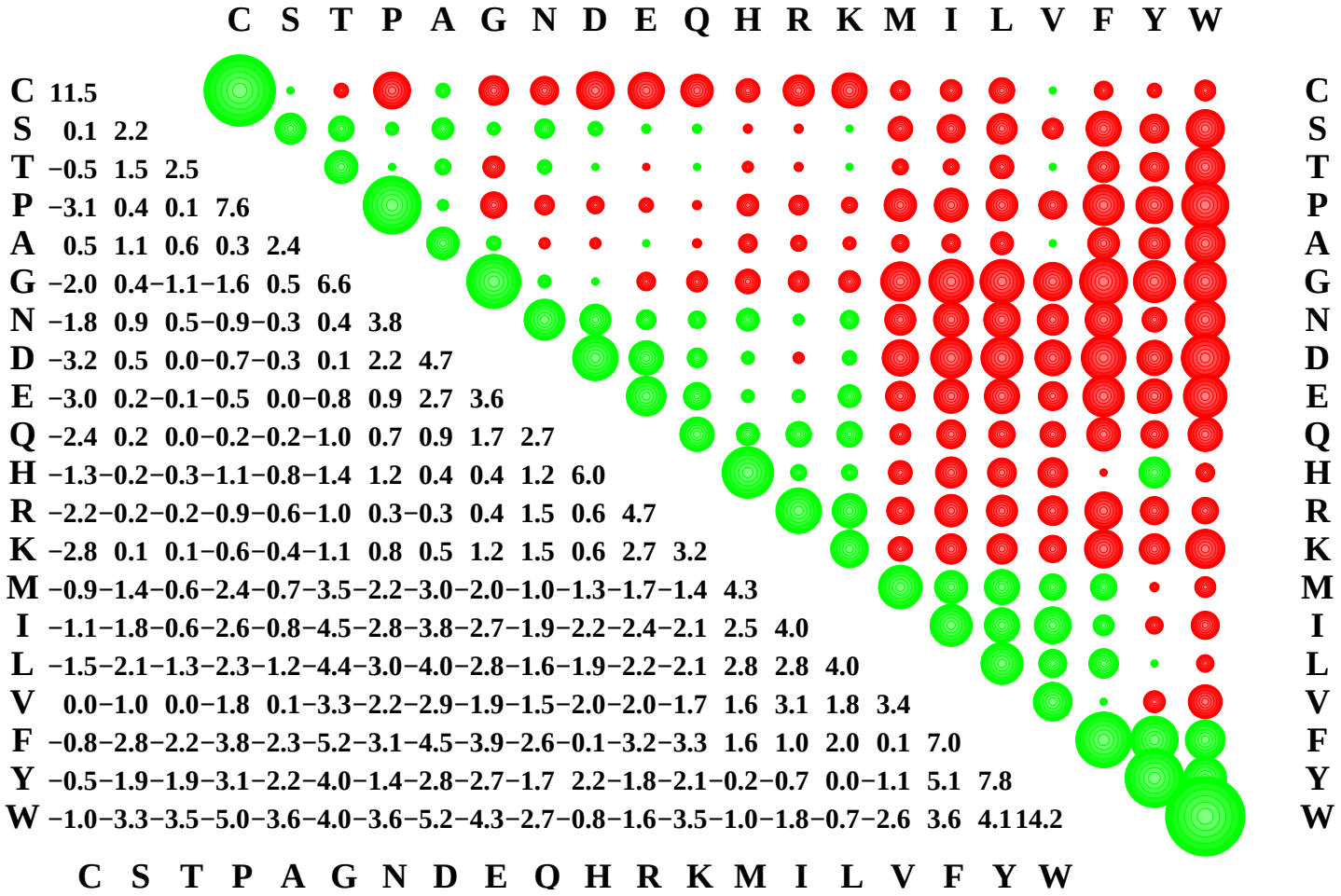
- Uses classical distance measures to produce protein alignments
- Given the alignments it computes a new distance matrix
- Align again using the new distance matrix
- Repeat this process many times
- In addition, they computed empirical measures for gap penalties. They suggest
  - For a probability P of a gap of length k

$$10 \ln(P) = -36.31 + 7.44 \ln(\text{PAM distance}) - 14.93 \ln(k)$$

  - If a PAM distance is not available

$$10 \ln(P) = -20.63 - 1.65 ln(k - 1)$$

GONNET matrix



The log odds matrix is lower left. It is 10 times the log of the prob these aa are aligned / prob of chance alignment.

Specialized matrices

Some matrices also incorporate additional information - STR matrix includes information about protein structure and can be used with very distantly related sequences

Other matrices are specific for different types of proteins - SLIM (ScoreMatrix Leading to Intra-Membrane) and PHAT (Predicted Hydrophobic and Transmembrane matrix) are designed from/for membrane proteins (not soluble proteins)

As of 2006, 94 matrices in GenomeNet