BLOSUM Scoring Matrices

- BLOck SUbstitution Matrix
- Based on comparisons of Blocks of sequences derived from the Blocks database
- The Blocks database contains multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins (local alignment versus global alignment)
- BLOSUM matrices are derived from blocks whose alignment corresponds to the BLOSUM-,matrix number (*e.g.* BLOSUM 62 is derived from Blocks containing >62% identity in ungapped sequence alignment)
- BLOSUM 62 is the default matrix for the standard protein BLAST program

BLOSUM Background

- Prosite data base: “dictionary of sites and patterns in proteins”; linked to Swiss-Prot database
  - Goal is to identify “biologically significant” patterns in protein families (with special emphasis on those regions thought to be important to protein function)
  - Tries to find good “discriminators” that emphasize reliable identification of known family members while excluding known non-members
  - Prosite patterns: signature “motifs”
- Example: Helicase proteins
  - involved in unwinding and opening of DNA strands in preparation for transcription
  - “Werner’s syndrome”: mutation in helicase causes affected individuals to age at a an accelerated rate
  - Hundreds of helicases from different organisms have been sequenced; much of what we know about how they work comes from computer-assisted analysis of these sequences

BLOSUM Background (continued)

- **Motifs**: features conserved across all sequences from a family (e.g., helicases) or across different subsets of them
- These motifs can be used to search protein/DNA databases to discover previously unknown members
- "Family" typically defined by function: helicases share in common the property of helping to unwind DNA

  By finding new helicases and asking what they have in common, we can better understand their mechanics

- One helicase pattern motif:

  [&H][&A]D[DE]x_n[TSN]x_4[QK]Gx_7[&A] ("regular expression")

  where & = any aa from I L V M F Y W
  x = anything
  x_n= any sequence of *n* amino acids

BLOSUM Background (continued)

- Patterns may also be represented as **profiles**:

E.g., consider the multiple alignment:

sequence 1	a	b	c	-	a
sequence 2	a	b	a	b	a
sequence 3	a	c	c	b	-
sequence 4	c	b	-	b	c

Corresponding profile:

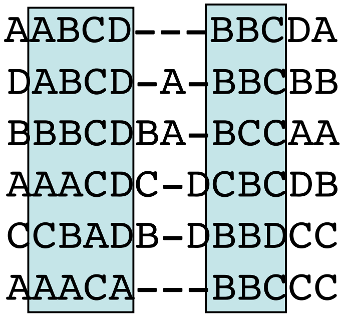
	1	2	3	4	5
a	0.75		0.25		0.50
b		0.75		0.75	
c	0.25	0.25	0.50		0.25
-			0.25	0.25	0.25

Profile has higher “resolution” (reflects different frequencies of representation by amino acids at a site)

BLOSUM Background (continued)

Henikoff and Henikoff (1991) developed a database of “blocks” based on sequences with shared motifs (>2,000 blocks of aligned sequence segments from >500 groups of related proteins)

E.g.:



```
AABCD---BBCDA
DABCD-A-BBCBB
BBBCDBA-BCCAA
AAACDC-DCBCDB
CCBADB-DBBDCC
AAACA---BBCCC
```

Why *blocks*?

- Need to have a multiple alignment; easier to align with similar sequences
- Don’t want insertions and deletions to complicate estimation of substitution probabilities
- Interested in detecting *conserved* regions of protein sequences, so restrict attention to these regions when computing the scoring matrix

$c_{ij}^{(k)}$

Calculating a BLOSUM Matrix

Just as with the PAM matrix, we will compute the BLOSUM score as the (log) ratio of the observed probability of substitution of one amino acid by another divided by the probability expected purely due to chance. First the numerator:

1. Count pair frequencies $c_{ij}^{(k)}$ for each pair of amino acids *i* and *j*, for each column *k* of each block:

E.g., 1st column is AACABA

AA	4	4	4(4-1)/2 = 6
AB	4	1	(4)(1) = 4
AC	4	1	(4)(1) = 4
BB	1	1	(1)(1-1)/2 = 0
BC	1	1	(1)(1) = 1
CC	1	1	(1)(1-1)/2 = 0

i.e., for "like" comparisons, $$c_{ii}^{(k)} = \binom{n_i}{2}$$

for "unlike" comparisons, $$c_{ij}^{(k)} = n_i n_j$$

where n_i = the number of times residue *i* was observed in the column

Calculating a BLOSUM Matrix (continued)

2. Sum the scores for each columns across columns:

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

3. Normalize the pair frequencies so they will sum to 1:

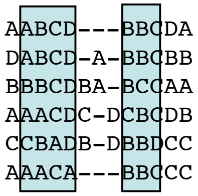
$$T = \sum_{i \geq j} c_{ij} \quad = \quad w \frac{n(n-1)}{2}$$

where w = number of columns
n = number of sequences

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

For previous example, q_{AB} calculation across columns is:

$$q_{\mathrm{AB}} = \frac{4+8+0+0+0+0+0}{7\frac{(6)(5)}{2}} = \frac{12}{105}$$



```
AABCD---BBCDA
DABCD-A-BBCBB
BBBCDBA-BCCAA
AAACDC-DCBCDB
CCBADB-DBBDCC
AAACA---BBCCC
```

Calculating a BLOSUM Matrix (continued)

Now, we will calculate the denominator of the odds ratio.

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

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

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

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

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

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

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

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

Example BLOSUM matrix calculation

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

Matrix of c_{ij} values:

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

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

Example BLOSUM matrix calculation (continued)

Matrix of q_{ij} values:

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

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

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 BLOSUM matrix calculation (continued)

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 BLOSUM matrix calculation (continued)

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 = $\mathit{round}(2 \cdot 0.7516) = 2$

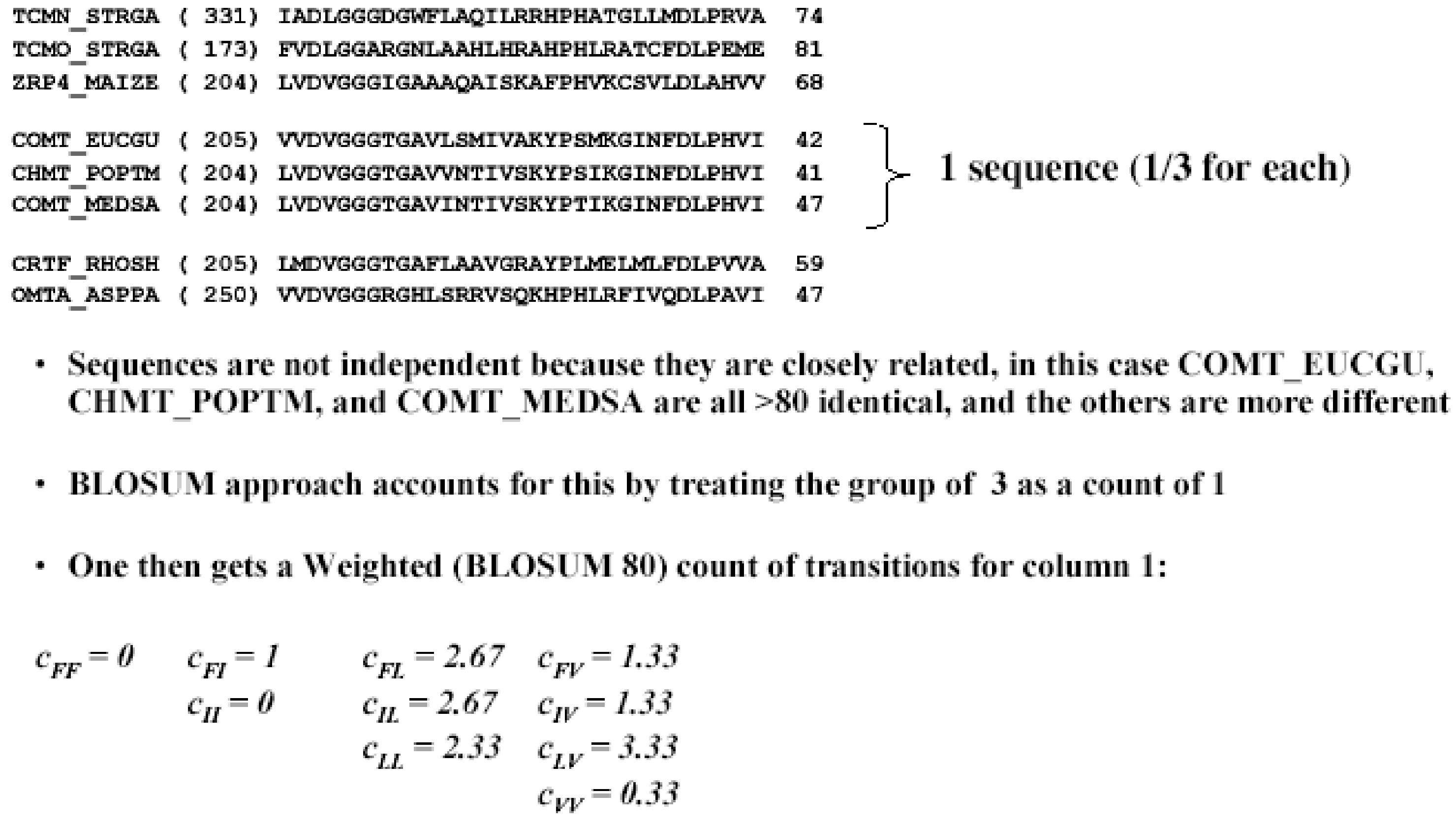
Full matrix:

	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)

Dealing with sequence redundancy

E.g., for BLOSUM-80, group sequences that are >80% similar



```
TCMN_STRGA  ( 331)  IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA   74
TCMO_STRGA  ( 173)  FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME   81
ZRP4_MAIZE  ( 204)  LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV   68

COMT_EUCGU  ( 205)  VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI   42
CHMT_POPTM  ( 204)  LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI   41
COMT_MEDSA  ( 204)  LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI   47

CRTF_RHOSH  ( 205)  LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA   59
OMTA_ASPPA  ( 250)  VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI   47
```

} 1 sequence (1/3 for each)

- **Sequences are not independent because they are closely related, in this case COMT_EUCGU, CHMT_POPTM, and COMT_MEDSA are all >80 identical, and the others are more different**
- **BLOSUM approach accounts for this by treating the group of 3 as a count of 1**
- **One then gets a Weighted (BLOSUM 80) count of transitions for column 1:**

$$\begin{array}{llll}
c_{FF} = 0 & c_{FI} = 1 & c_{FL} = 2.67 & c_{FV} = 1.33 \\
 & c_{II} = 0 & c_{IL} = 2.67 & c_{IV} = 1.33 \\
 & & c_{LL} = 2.33 & c_{LV} = 3.33 \\
 & & & c_{VV} = 0.33
\end{array}$$

(slide from Michael Gribskov)

Relative entropy

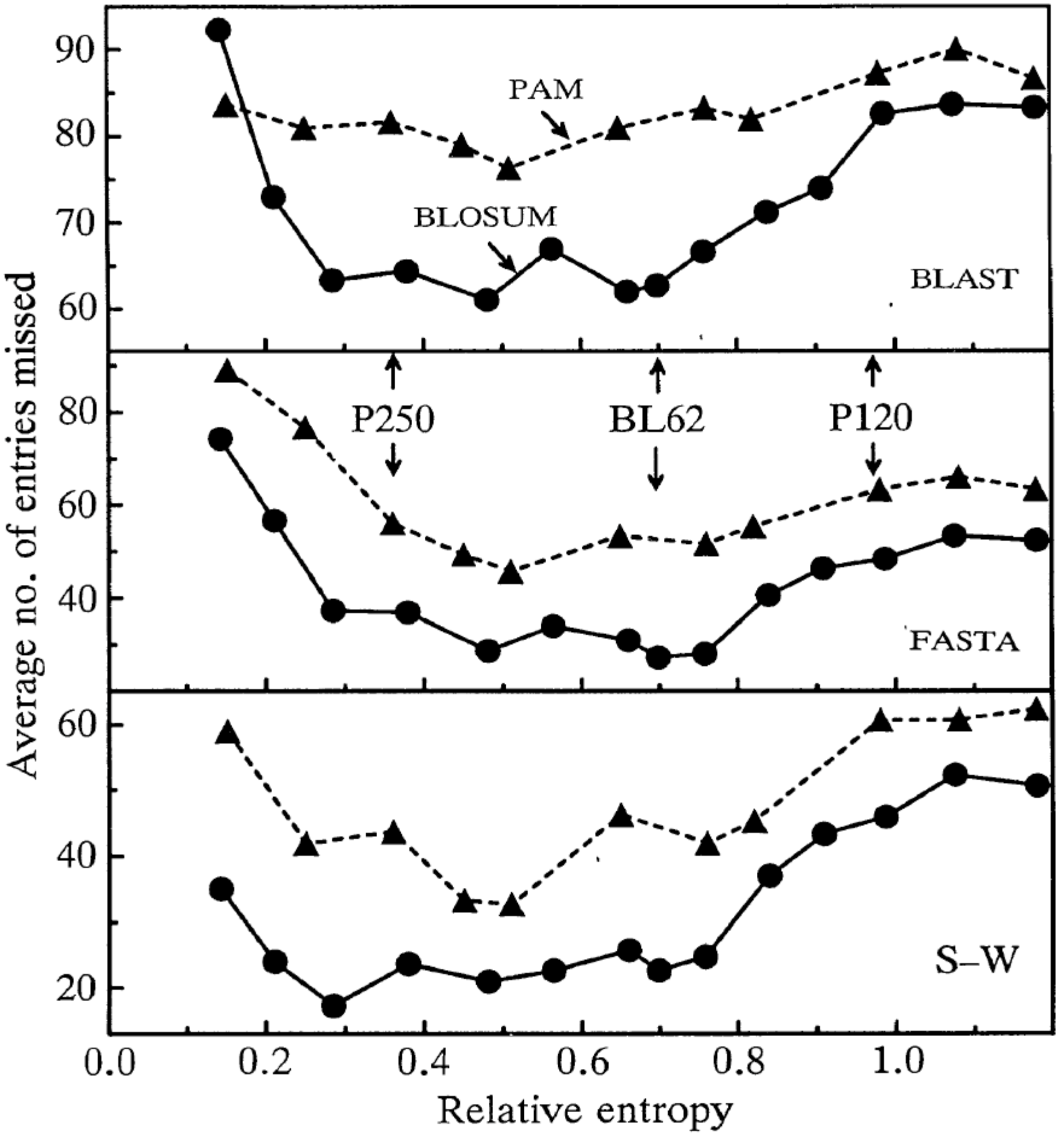
$$H = \sum_{i \geq j} q_{ij} s_{ij}$$

“Average information per residue pair”

Equivalent PAM and BLOSUM matrices based on relative entropy

PAM100	==>	Blosum90
PAM120	==>	Blosum80
PAM160	==>	Blosum60
PAM200	==>	Blosum52
PAM250	==>	Blosum45

Superiority of BLOSUM for database searches (according to Henikoff and Henikoff)



PAM versus BLOSUM

PAM properties:

- Based on an explicit evolutionary model
- Assumes that more distant changes are reflection of repeated short-term changes, and therefore can work over a wide range of divergences

PAM limitations:

- Assumptions of model clearly violated
- Each position is context dependent
  - Rates of substitution vary across and within proteins
  - Local 3-D environments vary
- Rare changes more prone to sampling error (changes in similar sequences occur at sites that are less constrained)

PAM versus BLOSUM

BLOSUM properties:

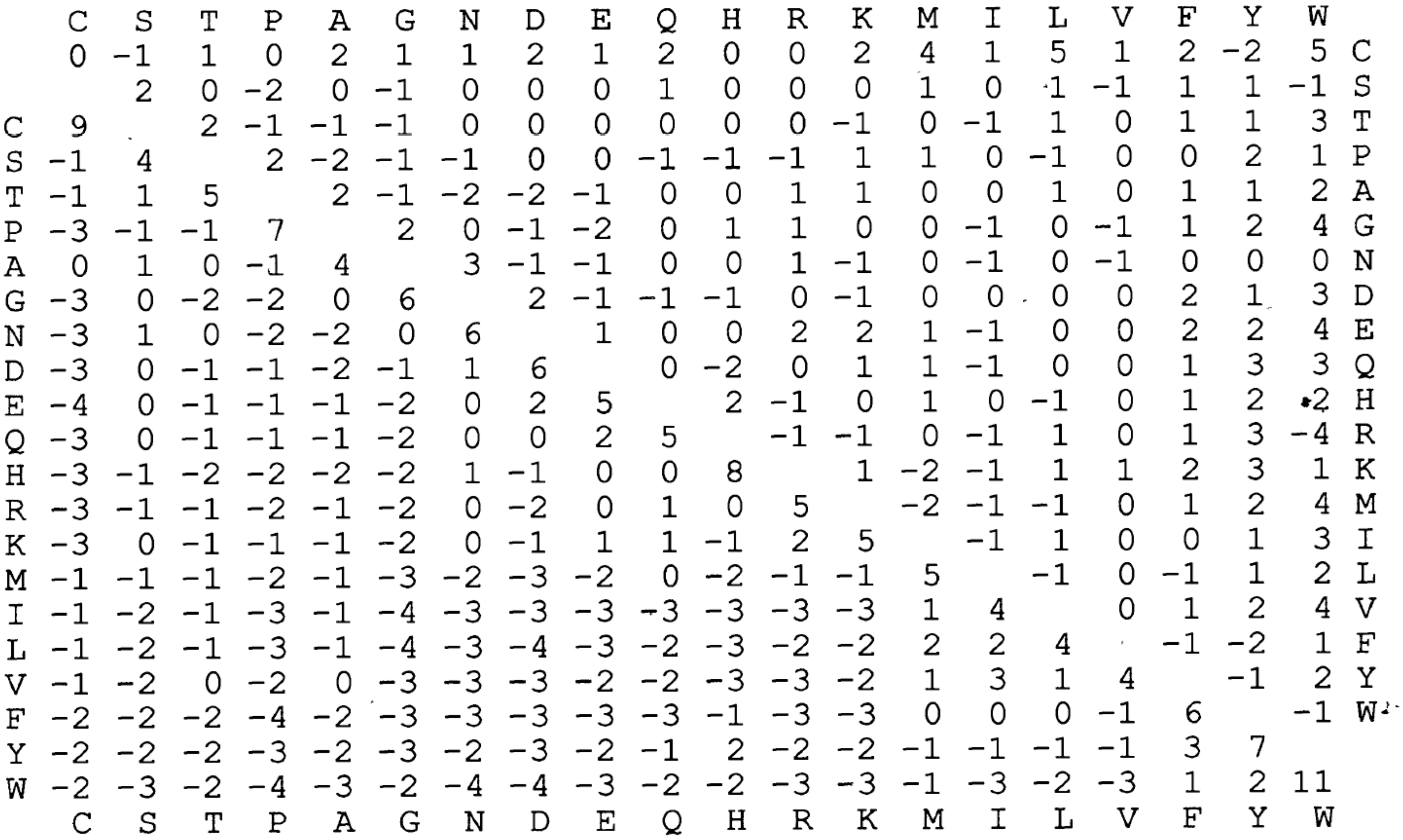
- Not based on an explicit evolutionary model; purely empirically derived
- Based on sequence comparisons covering a broad range of divergences

BLOSUM limitations:

- Restricted to a subset of conserved domains
- Implied "star-tree" model of evolution: closeness of relationship ignored

PAM versus BLOSUM

Below diagonal: BLOSUM 62
Above diagonal: BLOSUM 62 - PAM 160



	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	