Alignment IV
BLOSUM Matrices

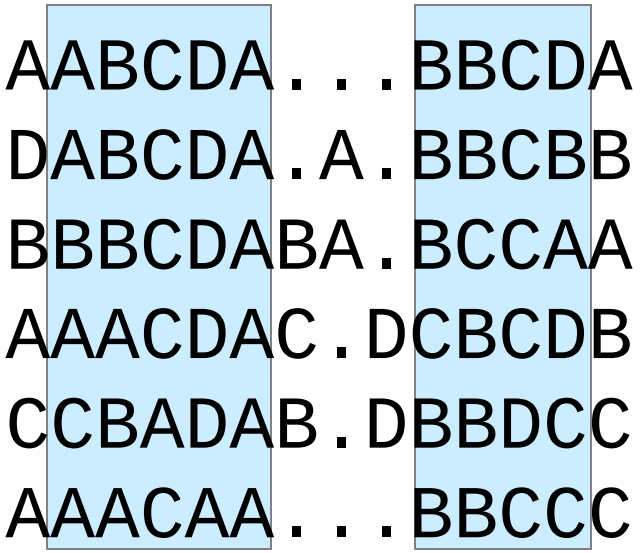
BLOSUM matrices

- **Blo**cks **Su**bstitution **M**atrix. Scores for each position are obtained frequencies of substitutions in blocks of local alignments of protein sequences [Henikoff & Henikoff92].
- For example BLOSUM62 is derived from sequence alignments with no more than 62% identity.

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

Conserved blocks in alignments



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
1. Normalize results to obtain probabilities (p_X's and q_{XY}'s)
2. 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:

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

where w = number of columns
n = number of sequences

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

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

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):

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

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

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

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:

	A	I	L	S	T	V
A	1+10					
I		0				
L		3	3			
S	2		0			
T	4			2	1	
V		1	3			0

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

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

Example

Matrix of q_{ij} values:

	A	I	L	S	T	V
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

$=$

	A	I	L	S	T	V
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

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	$2\left(\frac{14}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{6}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{2}{30}\right)\left(\frac{2}{30}\right)$	$2\left(\frac{4}{30}\right)\left(\frac{2}{30}\right)$	$\left(\frac{2}{30}\right)^2$

Log odds ratio:

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

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

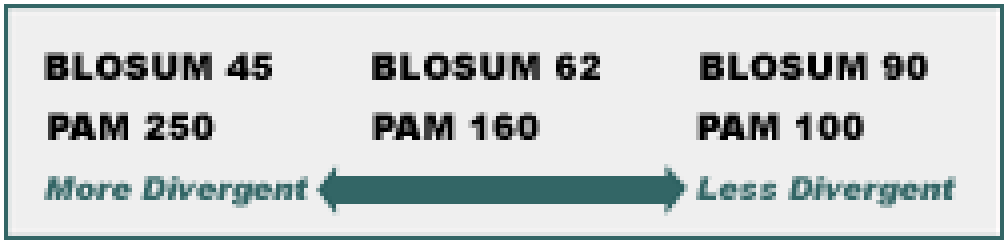
Full matrix:

	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)

Comparison

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



Relative Entropy

$$H = \sum_{i=1}^{20} \sum_{j=1}^{i} p_i p_j s(i, j)$$

- Indicates power of scoring scheme to distinguish from “background noise” (i.e., randomness)
- Relative entropy of a random alignment should be negative
- Can use *H* to compare different scoring matrices

Equivalent PAM and Blossum matrices (according to *H*)

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

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	

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

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