UNIT 4: SEQUENCE ALIGNMENT PART 4

Rubina Chongtham
Department of Botany
Deshbandhu College, University of Delhi

TOPIC

- BLOcks substitution matrix (BLOSUM)
- Comparison between pam & BLOSUM

This material is being circulated, as supplement to my online class, in pursuance of the official notification on Covid-19, for continued class engagement via e-resources.

Rubina Chongtham, Dept. of Botany

BLOCKS SUBSTITUTION MATRIX (BLOSUM)

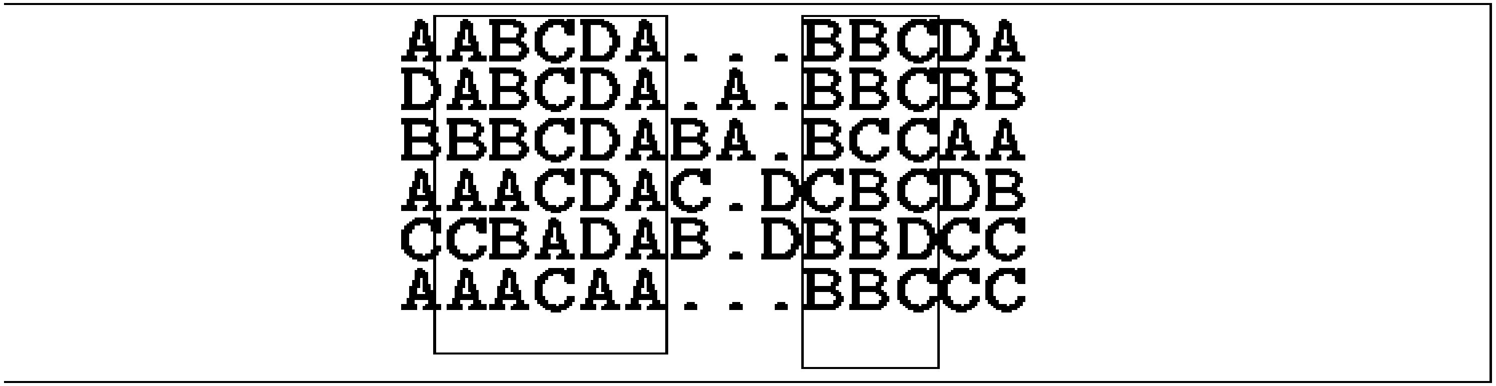
- This approach was initiated by Henikoff and Henikoff, 1992.
- BLOSUM addresses a major shortcoming of PAM approach that assumes a uniform evolutionary rate over the entire protein sequence.
- It is based on a much larger dataset, 500 Prosite families identified by Bairoch using conserved amino acid blocks that define each family.
- Typically it is used for multiple sequence alignment.
- AA substitutions are noted & log odds ratios derived.
- Eg. Block patterns that are 60% identical give rise to Blosum60 matrix, etc. i.e. conservation of functional blocks.
- It is not based on explicit evolutionary model.

BLOSUM CONTD.

- For its calculation only blocks of amino acid sequences with small change between them are considered. These blocks are called *conserved blocks (Fig 1)*.
- One reason for this is that one needs to find a multiple alignment between all these sequences and it is easier to construct such an alignment with more similar sequences.
- Another reason is that the purpose of the matrix is to measure the probability of one amino acid to change into another, and the change between distant sequences may include also insertions and deletions of amino acids.
- also, we are more interested in conservation of regions inside protein families, where sequences are quite similar, and therefore we restrict our examination to such.

Rubina Chongtham, Dept. of Botany

- Fig1: Alignment of several sequences. The conserved blocks are marked.



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

STEPS OF BUILDING THE BLOSUM MATRIX

- The first stage is eliminating sequences, which are identical in more than *x*% of their amino acid sequence. This is done to avoid bias of the result in favor of a certain protein. The elimination is done either by removing sequences from the block, or by finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster. The matrix built from blocks with no more the *x*% of similarity is called BLOSUM-*x* (e.g. the matrix built using sequences with no more then 50% similarity is called BLOSUM-50.)
- The second stage is counting the pairs of amino acids in each column of the multiple alignment. For example in a column with the acids AABACA (as in the first column in the block in fig 1), there are 6 AA pairs, 4 AB pairs, 4 AC, and one BC. The probability $q_{i,j}$ for a pair of amino acids in the same column to be A_i and A_j is calculated, as well as the probability p_i of a certain amino acid to be A_i.
- In the third stage the *log odd ratio* is calculated as $s_{i,j} = \log_2 \frac{q_i}{p_i}$ as as discussed earlier. As final result we consider the rounded $2s_{i,j}$, this value is stored in the (*i*,*j*) entry of the BLOSUM-*x* matrix.

Rubina Chongtham, Dept. of Botany

BLOSUM CONTD.

- BLOSUM matrices are based on local alignments.
- **BLOSUM x is a matrix calculated from comparisons of sequences with no less than x% divergence. Eg. BLOSUM62 is based on comparisons of sequences with no less than 62% divergence.**
- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Rubina Chongtham, Dept. of Botany

EG. BLOSUM62 MATRIX

- It’s a 20X20 matrix.
- Every possible identity and substitution is assigned a score based on the observed frequencies of such occurrences in alignments of related proteins.
- Identities are assigned the most positive score.
- Frequently observed substitutions are also scored positive, whereas rarely observed substitutions are given negative scores.

Rubina Chongtham, Dept. of Botany

BLOSUM 62 scoring matrix

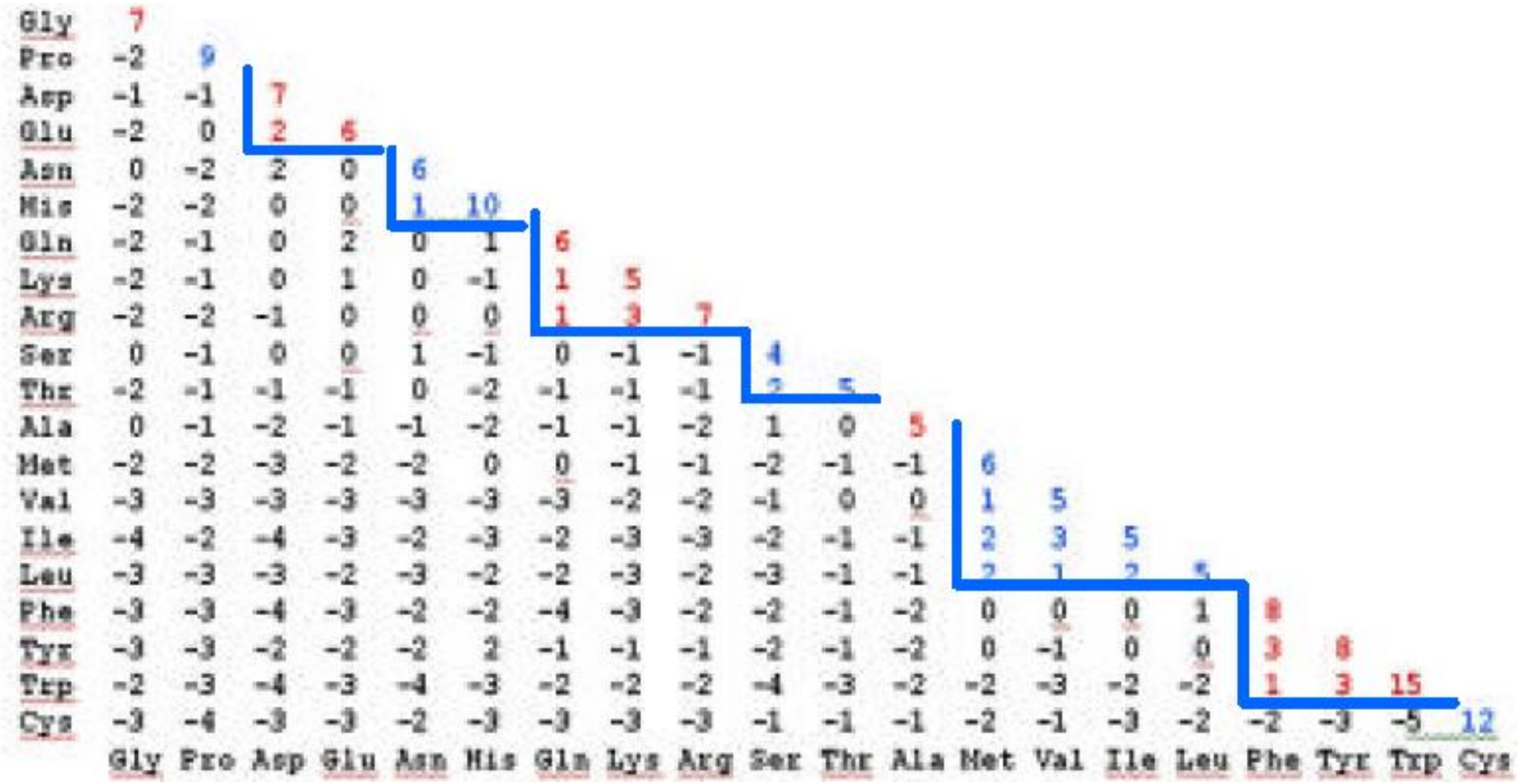
(positive values are shaded)

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**
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

Rubina Chongtham, Dept. of Botany

EG: BLOSUM45 AMINO ACID SIMILARITY MATRIX



Gly	7																			
Pro	-2	9																		
Asp	-1	-1	7																	
Glu	-2	0	2	6																
Asn	0	-2	2	0	6															
His	-2	-2	0	0	1	10														
Gln	-2	-1	0	2	0	1	6													
Lys	-2	-1	0	1	0	-1	1	5												
Arg	-2	-2	-1	0	0	0	1	3	7											
Ser	0	-1	0	0	1	-1	0	-1	-1	4										
Thr	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
Ala	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
Met	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
Val	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
Ile	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
Leu	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
Phe	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Tyr	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
Trp	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
Cys	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12
	Gly	Pro	Asp	Glu	Asn	His	Gln	Lys	Arg	Ser	Thr	Ala	Met	Val	Ile	Leu	Phe	Tyr	Trp	Cys

Rubina Chongtham, Dept. of Botany

COMPARISON BETWEEN PAM & BLOSUM:

PAM

1) They’re based on global alignments of closely related proteins.
2) PAM1 is calculated from comparisons of sequences with no more than 1% divergence.
3) Other PAM matrices are extrapolated from PAM1.
4) **They’re based on an explicit evolutionary model (that is,** replacements are counted on the branches of a phylogenetic tree).
5) They’re based on mutations observed throughout a **global alignment, this** includes both highly conserved and highly mutable regions.

BLOSUM

1) They’re based on local alignments.
2) BLOSUM62 is calculated from comparisons of sequnces with no less than 62% divergence.
3) All matrices are based on observed alignments & are not extrapolated from comparisons of closely related proteins.
4) **Based on an implicit rather than explicit model of evolution.**
5) **They’re based only on highly conserved regions in series of alignments** forbidden to contain gaps.

Rubina Chongtham, Dept. of Botany

COMPARISON BETWEEN PAM AND BLOSUM

BLOSUM80	BLOSUM62	BLOSUM45
PAM1	PAM120	PAM250

Less divergent ↔ More divergent

- BLOSUM matrices with higher nos. and PAM matrices with lower nos. are both designed for comparisons of closely related sequences.
- BLOSUM matrices with lower nos. and PAM matrices with higher nos. are designed for comparisons of distantly related sequences.

REFERENCES

- http://hackert.cm.utexas.edu/courses/bch370/fall2014/Sequence%20Alignment%2009.pdf
- http://www.math.lsa.umich.edu/~dburns/seqalmit2.pdf
- http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec03/node10.html
- Ghosh & Mallick (2012). *"Sequence Alignment"* in Bioinformatics Principles and applications (Oxford University Press).