

## Open Access Article

The authors, the publisher, and the right holders grant the right to use, reproduce, and disseminate the work in digital form to all users.



Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102  
Volume 28, Issue Number 3, (2010)  
©Adenine Press (2010)

# The MBLOSUM: A Server for Deriving Mutation Targets and Position-specific Substitution Rates

<http://www.jbsdonline.com>

Bin-Guang Ma  
Igor N. Berezovsky\*

Computational Biology Unit,  
Bergen Center for Computational Science  
University of Bergen,  
Bergen 5008, Norway

## Abstract

To facilitate mutagenesis study, it is necessary to be able to derive mutation targets and associated substitution rates in the sequence of interest regardless of the availability of corresponding structure. It is also important to obtain these data depending on the specific aims of the mutation process. The MBLOSUM server determines candidate positions for mutations and derives position-specific substitution rates given only a protein sequence. Different sets of complete genomes collected according to their phylogeny or specificity of environments along with complete set of non-redundant sequences can be used in calculations depending on the experimental task. MBLOSUM server is available at: <http://apps.cbu.uib.no/mblosum>

## Introduction

Advances in computational hardware and tools have made determination of protein structure and function by modeling from sequence fairly routine as it has been shown in several recent publications in this Journal (1-12). However, rational protein design and focused directed evolution, which create proteins with desired properties, have not yet become a trivial procedure. This technology relies on the knowledge of mutation(s) of selected residue(s) in the sequence that will result in desired changes of the protein when mutation is done. *A priori* knowledge of mutation targets allows one to significantly reduce number of mutants that should be biochemically analyzed, facilitating experimental effort (13, 14). The functional amino acid residues involved into substrate binding, its stabilization in the functional site, and chemical transformations are frequently selected as “hot spots” for modification of enzyme catalytic properties (14), and several computational approaches for predicting them have been developed so far (15, 16). Besides, there has been a series of papers published recently and reviewed in (17), where role of mutations as a major constrain and, at the same time, a driving force in protein evolution was illuminated. Several models for predicting and analyzing stability effects of mutations, and their role in protein evolution have been developed (17), and automated estimators of protein stability (18) have been proposed recently. Currently existing servers such as ‘Rate4Site’ (15) and ‘HotSpot Wizard’ (16) provide reference information for mutagenesis and predict effect of mutations on protein stability (18), but they analyze the mutability of sites based on conservation and require known protein structure as an input. Here we propose a procedure, which calculates BLOSUM-like substitution rates for individual positions in multiple sequence alignments and provides position-specific rates of amino acid replacements. Commonly used BLOSUM matrices provide generic substitution rates between different amino acids, reflecting similarity of their physico-chemical features and pace of replacements between them depending on the sequence identity. Despite its importance for general alignment and homology search procedures, BLOSUM substitution rates may

\*Phone: +47 55 58 47 12  
Fax: +47 55 58 42 95  
E-mail: [Igor.Berezovsky@uni.no](mailto:Igor.Berezovsky@uni.no)

be insufficient in case of specific positions with biased amino acid compositions, as well as position-specific scoring matrices (PSSMs) contain information on amino acid frequencies only and do show probabilities of exchanges in amino acid pairs. It is highly desirable, therefore, to know the substitution rates in individual positions, reflecting their specific structural and/or functional role. The MBLOSUM server uses only a sequence as an input and suggests potential mutation targets in the original sequence along with sets of candidates to be used for replacements. In other words, mutation targets here are positions found in the sequences of interest, which should be mutated in order to gain desired changes in the protein. Importantly, multiple sequence alignments used for the derivation of these substitution rates can be built depending on the experimental task. Therefore, substitution rates will be determined by set of sequences in multiple sequence alignment, and suggested mutation will reflect characteristic of protein sequences used for building this alignment. Though suggested mutation targets can affect both structural and functional characteristic of proteins, distinguishing between them will demand additional analysis.

### **Server: Input, Output and Options**

#### *Multiple Sequence Alignment for Query Sequence*

The first part of MBLOSUM server, seq2Msa, derives multiple sequence alignment for the input query sequence. User can upload a file with one protein sequence in fasta format or input sequence directly. There is a link to example sequence which user can use to learn how procedure works given the sequences. User is asked to provide an e-mail to be notified about results. Figure 1 shows a snapshot of the homepage with all the input options and links. There is also link to “HowTo” page

**mBlosum**  
Webtool for Mutagenesis Target Analysis

Home | Contact

Home About Howto Contact

**Introduction**

The mBlosum server provides information on the choice of potential mutation targets in the sequence of interest by showing the position-specific substitution rates derived on the set of sequences chosen by user.

**Sequence or Multiple Sequence Alignment (FASTA format) ?**

Upload FASTA file  no file selected ?

E-Mail  ?

Generate substitution matrix from:

**Input Description**

- One protein sequence (see an example) in FASTA format
- A multiple sequence alignment (see an example) in FASTA format.

If the input is one protein sequence, please click the 'Blast' button.  
If the input is multiple sequence alignment, please click the 'Substrate' button.  
If you provided an email address, you will receive an email with a link to the result page.

© 2009 CBU, BCCS, Unifob AS | Design by: styleshout

**Figure 1:** Overview of the homepage for MBLOSUM server. User can start from its own sequence or multiple sequences alignment. Examples of input sequence and MSA are provided, as well as link to “HowTo” page containing detail description of computations performed by the server.

in the homepage (Figure 1), where all steps of the algorithm and corresponding parameters are described, their meanings are explained, and default values along with recommendations for usage are given. For the input sequence (minimal length is 50 residues), the server will perform homology detection against the selected database by using blast program. In the server's database, there is a list of most used phylogenetic groups of organisms (such as Archaea, Bacteria, Prokaryotes, Eukaryotes) and the dataset of non-redundant protein sequences. In addition, there are sets of organisms grouped according to their specific features, *e.g.* habitat temperature (psychrophiles, mesophiles, thermophiles, and hyperthermophiles).

First, the homology detection (blast) and the multiple sequence alignment (clustalW) are performed. For the blast procedure, there are several options, as E-values (standard for blast), coverage of query sequence by the longest matched segment, and the sequence identity of the longest match. These parameters are supposed to be the most useful ones in the blast procedure. If users need to have a full control on blast procedure, they may do blast independently and then directly use the second part of the server. For the query sequence, the output is a set of homologues derived from the selected database. The server uses homologues found in the blast step as the input for the clustalW program. Default parameters of clustalW are used in the multiple sequence alignment procedure. Multiple sequence alignment is generated for the query sequence and is used in the calculation of substitution rates.

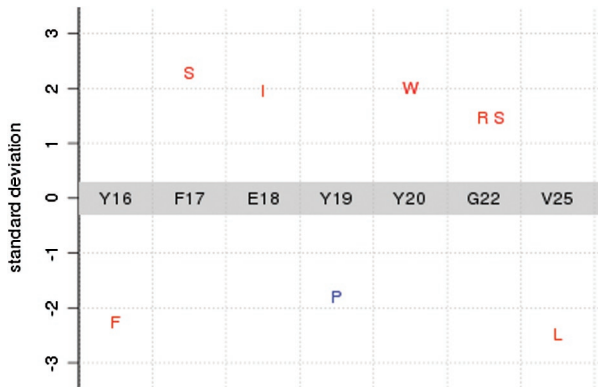
#### *Position-Specific Substitution Rates*

The second part of the MBLOSUM sever, msa2SubRate, uses multiple sequence alignment as an input, which can be can be the output of the first part (seq2Msa) or can be provided directly by the user. There is a link to example MSA (see Figure 1), which user can use to learn how procedure works given the MSA provided by the user. The calculation of the substitution rates follows the standard BLOSUM procedure. Similar to the BLOSUM procedure (19), the amino acid substitution rates are represented as the Lod (logarithm of odds) ratios  $s_{ij}$  (see the above paper for details);  $s_{ij} = 0$  means the observed frequency of a residue pair  $i,j$  is just as expected, and  $s_{ij} < 0$ , less than expected, and  $s_{ij} > 0$ , more than expected. The major difference is that they are *position-specific substitution rates*, which are calculated for each position separately rather than over all positions in the block. While in standard BLOSUM procedure there are no gaps allowed, a small fraction of gaps is allowed in our procedure to maximize the usage of available information in the multiple sequence alignments. The amount of allowed gaps can be set interactively. User is also suggested to select level of conservatism for the matrix depending on the degree of conservation in sequences used for its derivation. Similar to BLOSUM procedure, there is a choice of types from 30 to 100 with step 5, where number reflects degree of sequence identity used for clustering sequences in blocks. There are also two additional thresholds, 62% (the default corresponding to BLOSUM62) and 'no' (all sequences are used).

The output contains position-specific substitution rates for each occurring pair of amino acids in the multiple sequence alignment. The output is organized in a two-dimensional table, where the row contains results obtained for particular position in the query sequence. Detailed description of the output is presented in the on-line example, which is accessible through the link in the home page (Figure 1).

#### *Statistical Analysis of Position-Specific Substitution Rates*

Positions with particularly high or low variability are indicated by the sequence entropies (columns 4 and 5 in the output table), similar to outputs of "Rate4Site" (15) or "Hotspot Wizard" (16). Additionally, based on the substitution rates of all the occurring amino acid pairs in positions selected as mutation targets, the average and the standard deviation of the positional substitution rate can be calculated,



**Figure 2:** An example of the output for potential mutation targets. Mutations targets in the original sequences are lined up along the x-axis. Y-axis shows deviation of the substitution rate from the average for the outlier(s) in this position (F17S, more than two STD). Red – observed substitution rate is higher than expected, blue – lower.

and outliers are defined as the pairs that have substitution rates highly deviated (at least one standard deviation) from the average. Figure 2 shows an example of the output where positions with outliers are arranged in the center of the figure along the x-axis. The y-axis shows how different a particular substitution rate (in number of standard deviations) compared to the average in that position. Residue types for the outliers are shown in the figure. The larger the deviation in positive or negative direction of y-axis, the higher or lower the substitution rate to this residue type observed in the position compared to average substitution rate in this position. The candidates for replacement of the original residue are colored red or blue, corresponding to the positive or negative sign of the substitution rates (less or higher than expected), respectively. Finally, we compare the overall substitution rates (across all positions in the multiple sequence alignment) obtained in calculations and the ones of the standard *BLOSUM matrix*. The correlation between them can be regarded as an assessment of the quality of the multiple sequence alignment used for the calculation of position-specific substitution rates: high correlation with standard BLOSUM matrix means that the quality of the multiple sequence alignment is satisfactory for getting relevant substitution rates.

### Server: Implementation

The MBLOSUM server is built on a Linux system using Python and R scripts. The input can be pasted directly into the text field in the form of the web interface or can be uploaded as a file. The job id is generated automatically for each submission. Using the job id, user can check the status of the job and access the output within a fixed period of time (the results will be kept on the server for at least one week). If user provides an e-mail address during the submission, the link to the submitted job will be returned.

### Conclusions and Outlook

There are following major features that make the MBLOSUM server different from others. First, in addition to the mutability of sequence positions in terms of sequence entropy, our server delineates positions in the sequence of interest where substitution rates between particular residues are highly deviated from the average substitution rates in these positions. Second, position-specific substitution rates can be calculated based on the dataset of interest. The database includes non-redundant protein sequence dataset along with sets of proteomes for different taxons (Archaea, Bacteria, Fungi, Mammal) and life styles (psychrophiles, mesophiles, thermophiles, hyperthermophiles). By using dataset-dependant substitution rates, user can obtain set of mutation targets necessary for modifying a protein in the direction of interest. For example, the aim can be to make a protein more adapted to cold conditions (set of psychrophiles is to be used) or to make it more similar to mammalian proteins (Mammals). Finally, the advantage of this server is a minimal requirement on the input data: the MBLOSUM server does not require any structural information and works with only the protein sequence as an input.

By providing dataset-dependant position-specific substitution rates, the present server is supposed to be valuable asset for mutagenesis research, especially for selecting potential mutation targets and obtaining sets of candidate amino acids for the replacement of target residues in the original sequence.

### Acknowledgements

This work is supported by Functional Genomics Programme (FUGE II) from the Norwegian Research Council. Authors thank Kjell Petersen, Svenn Grinhaug, Kidane Tekle, and Yvan Strahm for technical support in incorporating this server

*Conflict of Interest:* None declared.

#### References

1. M. Parthiban, M. B. Rajasekaran, S. Ramakumar, and P. Shanmughavel. *J Biomol Struct Dyn* 26, 535-547 (2009).
2. K. Sujatha, A. Mahalakshmi, D. K. Y. Solaiman, and R. Shenbagarathai. *J Biomol Struct Dyn* 26, 771-779 (2009).
3. R. Chattopadhyaya and A. Pal. *J Biomol Struct Dyn* 25, 357-371 (2008).
4. D. Josa, E. F. F. da Cunha, T. C. Ramalho, T. C. S. Souza, and M. S. Caetano. *J Biomol Struct Dyn* 25, 373-376 (2008).
5. J. Dasgupta and J. K. Dattagupta. *J Biomol Struct Dyn* 25, 495-503 (2008).
6. A. Bagchi and T. C. Ghosh. *J Biomol Struct Dyn* 25, 517-523 (2008).
7. S. Subramaniam, A. Mohammed, and D. Gupta. *J Biomol Struct Dyn* 26, 473-479 (2009).
8. S. S. Mohan, J. J. P. Perry, N. Poullose, B. G. Nair, and G. Anilkumar. *J Biomol Struct Dyn* 26, 455-464 (2009).
9. R. Vinekar and I. Ghosh. *J Biomol Struct Dyn* 26, 741-754 (2009).
10. S. Mishra. *J Biomol Struct Dyn* 27, 283-291 (2009).
11. U. B. Sonavane, S. K. Ramadugu, and R. R. Joshi. *J Biomol Struct Dyn* 26, 203-214 (2008).
12. S. K. Singh, S. R. Choudhury, S. Roy, and D. N. Sengupta. *J Biomol Struct Dyn* 26, 235-245 (2008).
13. R. Chen. *Trends Biotechnol* 19, 13-14(2001).
14. R. A. Chica, N. Doucet, and J. N. Pelletier. *Curr Opin Biotechnol* 16, 378-384 (2005).
15. T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. *Bioinformatics* 18 Suppl 1, S71-77 (2002).
16. A. Pavelka, E. Chovancova, and J. Damborsky. *Nucleic Acids Res* 37, W376-383 (2009).
17. N. Tokuriki and D. S. Tawfik. *Curr Opin Struct Biol* 19, 596-604 (2009).
18. S. Yin, F. Ding, and N. V. Dokholyan. *Nat Methods* 4, 466-467 (2007).
19. S. Henikoff and J. G. Henikoff. *Proc Natl Acad Sci U S A* 89, 10915-10919 (1992).

*Date Received:* July 25, 2010

**Communicated by the Editor Ramaswamy H. Sarma**

