

MASIA: recognition of common patterns and properties in multiple aligned protein sequences

H. Zhu, C. H. Schein and W. Braun*

Sealy Center for Structural Biology, University of Texas Medical Branch, Galveston, TX 77555-1157, USA

Received on March 29, 2000; revised on May 26, 2000; accepted on June 15, 2000

Abstract

Summary: MASIA is a software tool for pattern recognition in multiple aligned protein sequences. MASIA converts a sequence to a properties matrix that can be scanned in both vertical and horizontal steps. Consistent patterns are recognized based on the statistical significance of their occurrence. Preset macros can be altered on-line to seek any combination of amino acid properties or sequence characteristics. MASIA output can be used directly by our programs to predict the 3D structure of proteins.

Availability: Access MASIA at <http://www.scsb.utmb.edu/masia/masia.html>.

Contact: werner@newton.utmb.edu

Proteins with a high degree of sequence identity (> 30%, depending on length, Abagyan and Batalov, 1997) have a similar fold (Wood and Pearson, 1999; Koehl and Levitt, 1999). Detecting patterns characteristic for a family in aligned sequences is the most accurate way to determine whether sequences with low similarity are truly related. Our new bioinformatics tool, MASIA, scans the individual columns of a matrix of aligned sequences vertically to reveal conserved positions in the family, and horizontally to reveal patterns (signatures, motifs, characteristics) that may indicate structurally or functionally important areas (Cuff and Barton, 1999; Zimmermann and Gibrat, 1998).

MASIA converts a sequence alignment matrix to a property matrix, whose contents and interpretation can be directed by the user. New motifs are identified using a preset macro for determining consensus at any conservation level. The user can alter the property library and define macros to search for specific patterns. MASIA differs from other programs now available on the web in the transparency of its method and the degree of control the user has over the detection criteria. Unlike programs such as SAS (Milburn *et al.*, 1998), MASIA analysis is based on properties of the amino acids and does not rely on a pre-determined 3D structure.

MASIA accepts optimized alignments (Krogh *et al.*, 1994) of the query sequence with related protein sequences (identified with FASTA or BLAST searches of the protein databases, di Francesco *et al.*, 1996) in the formats from ClustalW (Thompson *et al.*, 1999) or Pile-up (GCG package), with file extension names .aln or .msf, respectively. Commands are read via macros generated in the GUI of the stand alone version, or specified on-line in the web server of MASIA, as detailed in the manual (Zhu *et al.*, 1999). First, sequences are translated into significant conserved properties according to a pre-defined *property library*. The *matrix of properties* is scanned in vertical and horizontal steps to search for combinations of conserved properties (*groups*). A *characteristic* is a user-defined collection of groups. The three major components of the program are the property library, statistical evaluation methods for determining property conservation, and the command-line macro system.

Three methods, based on statistical expectation, probability entropy, or simple dominance, were implemented in MASIA to determine whether a property is significantly conserved in a column (Hänggi and Braun, 1994). Standard macros for determining various properties or a consensus sequences can be selected with buttons in the web server and altered on-line.

The biologically significant patterns and sequence characteristics of families summarized in PROSITE (Hofmann *et al.*, 1999) can be translated into MASIA macros. For example, a 'Greek key' motif, present in two domains of the β - and γ -crystallins, has the consensus pattern: [LIVMFYWA]-x-{DEHRKSTP}-[FY]-[DEQHKY]-x(3)-[FY]-x-G-x(4)-[LIVMFCST]. Figure 1 shows a MASIA macro for detecting this motif, displayed in the command block of the GUI in the stand-alone version. The amino acid groups in brackets were defined as properties (2bb2_225_1-4) in the property library. The resulting property matrix has symbols in the 7 rows where members of the group occur in a given column at a statistically significant frequency. The property matrix is scanned in 2 horizontal steps (hstep=2) and 1 vertical step (default), then two steps with 1×1, then 4×1, 2×1 and

*To whom correspondence should be addressed.

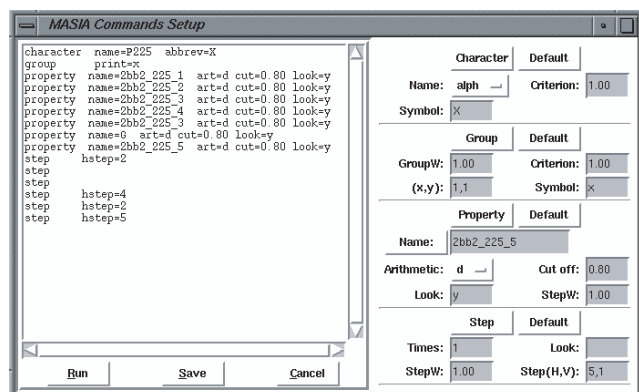


Fig. 1. The GUI of MASIA, with an illustration of a macro to detect a Greek key motif.

5×1 steps. The program prints an X under the residues that fulfill the motif in the output file. MASIA detected the motif in an alignment of 58 highly diverse sequences for $\beta 2$ -crystallin proteins. See the MASIA manual or the on-line server for other macros.

MASIA can be used to rapidly define motifs characteristic for novel sequences. For example, the consensus macro of the MASIA web server (Soman *et al.*, 2000) identified two blocks of conserved amino acids, HGLWP and KHGX_C, in an alignment of protein sequences from the RNase Rh/T2/stylar family. A macro similar to the one in the figure was used to search alignments of other family members. The motifs could be further defined by analyzing more general properties, such as hydrophobicity, in the surrounding residues. As MASIA is fast, many different pattern combinations can be searched to determine the most reliable motifs and the degree of deviation among members of a family. The method is also transparent, as the property matrix and decision tree are printed out.

The simplified WEB version is the fastest way to test MASIA. The stand-alone program for the SGI, with GUI, has other features, such as automatic coloring-coding of all 20 amino acids using a color setting library and easy selection of several macros by pushing buttons in the middle frame of the interface.

In conclusion, our new software tool, MASIA, rapidly determines patterns in multiple aligned sequences and searches for the presence of any user-specified pattern. The user can alter, on-line, the input macros and the

property library and search for distant variants of a given motif by altering the criterion values. This flexibility, combined with programming directly on the web, makes MASIA a useful tool for bioinformatics.

Acknowledgements

This work was supported by the NSF (DBI-9714937), the US DOE (DE-FG03-96ER62267), the Sealy and Smith Foundation and an ARP grant (004952-0084-1999) of the Texas Higher Education Coordinating Board.

References

- Abagyan, R.A. and Batalov, S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.
- Cuff, J.A. and Barton, G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Struct. Funct. Genet.*, **34**, 508–519.
- di Francesco, V., Garnier, J. and Munson, P.J. (1996) Improving the accuracy of protein secondary structure prediction with aligned homologous sequences. *Protein Sci.*, **5**, 106–113.
- Hänggi, G. and Braun, W. (1994) Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin. *FEBS Lett.*, **344**, 147–153.
- Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
- Koehl, P. and Levitt, M. (1999) A brighter future for protein structure prediction. *Nature Struct. Biol.*, **6**, 108–110.
- Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
- Milburn, D., Laskowski, R.A. and Thornton, J.M. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.*, **11**, 855–859.
- Soman, K.V., Schein, C.H., Zhu, H. and Braun, W. (2000) Homology modeling and simulations of nuclease structures. In Schein, C.H. (ed.), *Nuclease Methods and Protocols* Humana Press, New Jersey, in press.
- Thompson, J.D., Plewniak, F. and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
- Wood, T.C. and Pearson, W.R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.*, **291**, 977–995.
- Zhu, H., Schein, C.H. and Braun, W. (1999) User's Manual for MASIA.
- Zimmermann, K. and Gibrat, J.F. (1998) In unison: regularization of protein secondary structure predictions that makes use of multiple sequence alignments. *Protein Eng.*, **11**, 861–865.