BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 4 2006, pages 504–506
doi:10.1093/bioinformatics/bti825

Structural bioinformatics

Multiple Alignment of protein structures and sequences for VMD

John Eargle[1], Dan Wright[2,3] and Zaida Luthey-Schulten[1,3,*]

[1]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, 607 South Mathews Avenue, Urbana, IL 61801, USA, [2]Graduate School of Library and Information Sciences, University of Illinois at Urbana-Champaign, 501 Daniel Street, Champaign, IL 61820, USA and [3]Department of Chemistry, University of Illinois at Urbana-Champaign, 505 South Mathews Avenue, Urbana, IL 61801, USA

Received on July 29, 2005; revised on December 1, 2005; accepted on December 6, 2005
Advance Access publication December 8, 2005
Associate Editor: Anna Tramontano

ABSTRACT

Multiple Alignment is a new interface for performing and analyzing multiple protein structure alignments. It enables viewing levels of sequence and structure similarity on the aligned structures and performing a variety of evolutionary and bioinformatic tasks, including the construction of structure-based phylogenetic trees and minimal basis sets of structures that best represent the topology of the phylogenetic tree. It is implemented as a plugin for VMD (Visual Molecular Dynamics), which is distributed by the NIH Resource for Macromolecular Modeling and Bioinformatics at the University of Illinois.

Availability: Both binary and source code downloads for VMD 1.83, which includes *Multiple Alignment*, are available from http://www.ks.uiuc.edu/Research/vmd/. The Multidimensional QR factorization algorithm is available at http://www.scs.uiuc.edu/~schulten/software.html

Contact: multiseq@scs.uiuc.edu

1 INTRODUCTION

Multiple Alignment is an extension to VMD (Visual Molecular Dynamics) (Humphrey *et al*., 1996) for structural phylogenetic analysis. It carries out the multiple structural alignment of homologous proteins and expresses the comparisons in terms of a distance-based phylogenetic tree using two different structural metrics. It also includes a QR factorization of the alignment matrix to identify a non-redundant set of structures that represent the variation observed in the structural phylogenetic tree. The extension was primarily designed to facilitate evolutionary and bioinformatic investigations of the type described in O'Donoghue and Luthey-Schulten (2003, 2005) on the evolution of structure in the aminoacyl-tRNA synthetases (AARSs). To this end, it provides a number of tools for analyzing conservation among the multiple structures and their associated sequences. These tools allow users to correlate changes in structure with corresponding changes in the sequences of homologous proteins.

2 EVOLUTIONARY ANALYSIS OF STRUCTURES

Evolutionary analysis of multiple structures is most meaningful when carried out on single, functional domains or the longest homologous regions among a set of related proteins. The ASTRAL compendium (Brenner *et al*., 2000) (available at http://astral.berkeley.edu) provides PDB format structure files broken down by protein domain as determined by the SCOP database (Murzin *et al*., 1995) and is strongly recommended for use in this type of analysis.

3 MULTIPLE STRUCTURAL ALIGNMENTS

The multiple structural alignments are carried out using the program STAMP (Russell and Barton, 1992), which minimizes the C_α distance between aligned residues in each molecule through globally optimal rigid-body transformations. STAMP is used by the HOMSTRAD (Mizuguchi *et al*., 1998) and PALI (Sujatha *et al*., 2001) databases to generate the protein structure alignments, and more recently it has been used in a series of papers to construct structure-based phylogenetic trees (O'Donoghue and Luthey-Schulten, 2003, 2005; Sethi *et al*., 2005; O'Donoghue *et al*., 2006). In our implementation, the STAMP algorithm first applies a rough alignment derived simply by aligning the N-terminal ends of the first protein to all the others.

From this initial superposition, the multiple structure alignment follows a procedure similar to tree-based multiple sequence alignment. Each pair of structures is aligned, and similarity scores are calculated to derive a dendrogram that guides the progressive structural alignment algorithm. Structures and structure sets are superimposed in order of similarity by following the dendrogram from the branches to the root. At each node, a dynamic programming algorithm obtains a tentative alignment of the two structure sets and generates a superposition of the structures. This process is iterated to maximize the number of aligned residues.

This structural alignment method does require that the molecules being aligned have similar structural signatures. For this reason, *Multiple Alignment* should be used for a single domain evolutionary analysis and not for alignment of larger multi-domain proteins, unless the multi-domain proteins display homology over all domains. Attempting to align unrelated multi-domain structures or dissimilar proteins with STAMP may result in STAMP's failure to align the proteins. Before failure, STAMP may spend several minutes attempting to compute an alignment. For this reason, users are warned and given an option to abort the alignment before attempting to align multi-domain structures. If the user does decide to attempt this, structures should be loaded in order from shortest to longest.

[*]To whom correspondence should be addressed.

$$Q_{res,ni} = \aleph \sum_{\substack{m \\ m \neq n}} \sum_{\substack{j \\ j<(i-1) \\ j>(i+1)}} \exp\left[-\frac{(r_{nij} - r_{mi'j'})^2}{2\sigma_{nij}^2}\right]$$

$$\sigma_{ij}^2 = |i-j|^{0.15}$$

$$\aleph = \frac{1}{(N_{seq}-1)(N_{res}-k)}$$

Fig. 1. i and j are residue numbers of sequences n; i' and j' are residues in sequences m that are aligned to ni and nj, respectively. r is the distance between residues i and j in a particular alignment. N_{seq} is the number of sequences in the alignment. N_{res} is the number of residues in the current sequence. k is 2 when the residue is on the end, 3 otherwise because the metric ignores comparisons of the residue with itself and with its immediate neighbors in the sequence.

4 USAGE FEATURES

Multiple Alignment provides a number of features for protein structure analysis. After aligning multiple protein structures, users can:

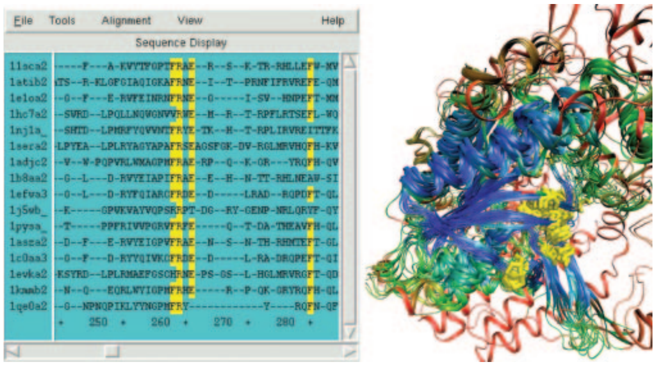
- Color the 3D display of structures by structure (Q_{res}, defined in Fig. 1), sequence conservation or RMSD per residue (Fig. 2).
- Display a UPGMA phylogenetic tree based on one of two structural measures: Q_H or RMSD (Fig. 3). Structural trees based on Q_H have been shown to be congruent to sequence-derived trees (O'Donoghue and Luthey-Schulten, 2003).
- Export structural alignments and secondary structure information in FASTA format.
- Export transformed coordinates of aligned proteins.

A tutorial demonstrating the above features is available through our website at http://www.scs.uiuc.edu/~schulten/tutorial-aars.pdf.

5 IMPLEMENTATION

VMD is available for Windows and MacOS X, in addition to Linux and most UNIX platforms. VMD includes a Tcl/Tk scripting interface allowing for the implementation of plugins extending the functionality of the core program. *Multiple Alignment* is implemented as a set of Tcl/Tk plugins for VMD, which facilitates rapid implementation and debugging of the interface while taking advantage of the powerful, fast scripting language interface to VMD's core functions.

Several of the features of *Multiple Alignment* are implemented as Tcl wrappers around C or C++ programs. STAMP (Russell and Barton, 1992), which is used to align the protein structures, is a C program compiled separately for each of the platforms VMD runs on. The Tcl wrapper provides an interface for calling the C program from *Multiple Alignment* using a uniform method on all platforms. Structural similarity is evaluated using the metric Q_H that accounts for contributions from both aligned and gapped portions of the pairwise alignments (O'Donoghue and Luthey-Schulten, 2005). Both it and RMSD can be used to constructed the distance-based phylogenetic trees. Q_{res} (Fig. 1) determines the degree of structural similarity per residue for only the aligned portions of the complete multiple structure alignment.



```
11sca2  -----F---A-KVVTFOPTFRAE--R--S--K-TR-RHLLEFW-MV
1atib2  ATS--R-KLGFGIAQIGKAFRNE--I--T--PRNFIFRVREFE-QM
1eloa2  --G--F---E-RVFEINRNFRNE--G-----I-SV--HNPEFT-MM
1hc7a2  --SWRD--LPQLLNQWGNVVRWE--M--R--T-RPFLRTSEFL-WQ
1nj1a_  --SHTD--LPMRFYQVVMTFRYE-TK--H--T-RPLIRVREITTFK
1sera2  -LPYEA--LPLRYAGYAPAFRSEAGSFGK-DV-RGLMRVHQFH-KV
1adjc2  --V--W-PQPVRLWMAGPMFRAE-RP--Q--K-GR---YRQFH-QV
1b8aa2  --G--L---D-RVYEIAPIFRAE--E--H--N-TT-RHLNEAW-SI
1efwa3  --G--L---D-RYFQIARCFRDE--D-----LRAD--RQPDFT-QL
1j5wb_  --K-----GPVKVAYVQPSRRPT-DG--RY-GENP-NRLQRYF-QY
1pysa_  --T-----PPFRIVVPGRVFRPE-----Q--T-DA-THEAVFH-QL
1asza2  --D--F---E-RVYEIGPVFRAE--N--S--N-TH-RHMTEFT-GL
1c0aa3  --G--F---D-RYYQIVKCFRDE--D-----L-RA-DRQPEFT-QI
1evka2  -KSYRD--LPLRMAEFGSCHRNE-PS-GS--L-HGLMRVRGFT-QD
1kmmb2  --N--Q---EQRLWVIGPMFRHE-----R--P-QK-GRYRQFH-QL
1qe0a2  --G--NPNQPIKLYYNGPMFRY-----------Y----RQFN-QF
         •   250   •   260   •   270   •   280   •
```

Fig. 2. Multiple structural alignment of a representative set of the class II aminoacyl-tRNA synthetases. Structures are colored by structural conservation Q_{res}. User-selected residues are highlighted on both the sequence and structure displays.

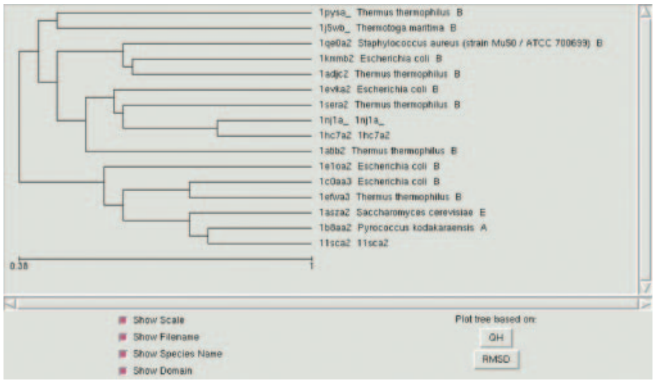

1pysa_ Thermus thermophilus B
1j5wb_ Thermotoga maritima B
1qe0a2 Staphylococcus aureus (strain Mu50 / ATCC 700699) B
1kmmb2 Escherichia coli B
1adjc2 Thermus thermophilus B
1evka2 Escherichia coli B
1sera2 Thermus thermophilus B
1nj1a_ 1nj1a_
1hc7a2 1hc7a2
1atib2 Thermus thermophilus B
1e1oa2 Escherichia coli B
1c0aa3 Escherichia coli B
1efwa3 Thermus thermophilus B
1asza2 Saccharomyces cerevisiae E
1b8aa2 Pyrococcus kodakaraensis A
11sca2 11sca2
0.38
1
Show Scale
Show Filename
Show Species Name
Show Domain
Plot tree based on:
QH
RMSD

Fig. 3. A phylogenetic tree, based on the Q_H structural measure, derived from the structural alignment in Figure 2.

The multiple structure alignment including gaps is orthogonally encoded in an alignment matrix. *Multiple Alignment* includes a C++ implementation of the QR factorization of the alignment matrix that orders the proteins by increasing structural similarity and can be used to identify a non-redundant set of proteins for subsequent analysis (O'Donoghue and Luthey-Schulten, 2005). The multidimensional QR factorization algorithm is also available on our website.

6 FUTURE WORK

Version 2.0 of the *Multiple Alignment* plugin is currently under development; we hope to have this version available early in 2006. It will add many sequence and metadata search features to the application described here. Features of the new version include independent analysis of structure and sequence data, importing and filtering of multiple structures and sequences from BLAST searches, CLUSTALW (Higgins *et al.*, 1994) alignments of sequences, and sequence editing.

7 CONCLUSION

Multiple Alignment represents a first step in adding structural bioinformatics capabilities to VMD, which is already widely used within the community for MD trajectory analysis. By providing researchers with tools to work synergistically with multiple structures and their associated sequences, we hope to help address complex problems relating to evolutionary changes in biological molecules. Our long range goal is to develop an interface that is useful for experimentalists and theoreticians who are working in both the sequence and structural worlds of molecular biology.

ACKNOWLEDGEMENTS

The authors would like to thank Patrick O'Donoghue and Brijeet Dhaliwal for their assistance in the development of this software and accompanying documentation, and John Stone for his help in porting the software and preparing it for final release. The authors would also like to acknowledge the support of the NIH Resource for Macromolecular Modeling and Bioinformatics (grant PHS2P41RR05969) and the NIH Molecular Biophysics Training Grant (grant PHS5T32GM08276).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Brenner,S.E. *et al*. (2000) The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res*., **28**, 254–256.

Higgins,D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Humphrey,W. *et al*. (1996) VMD–Visual Molecular Dynamics. *J. Mol. Graph*., **14**, 33–38.

Mizuguchi,K. *et al*. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*., **7**, 2469–2471.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

O'Donoghue,P. and Luthey-Schulten,Z. (2003) On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev*., **67**, 550–573.

O'Donoghue,P. and Luthey-Schulten,Z. (2005) Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information. *J. Mol. Biol*., **346**, 875–894.

O'Donoghue,P., Sethi,A., Woese,C.R. and Luthey-Schulten,Z. (2006) The Evolutionary History of Cys-tRNA(Cys) formation. *Proc. Natl. Acad. Sci.* USA **102**, 19003–19008.

Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.

Sethi,A. *et al*. (2005) Evolutionary profiles from the QR factorization of multiple sequence alignments. *Proc. Natl Acad. Sci. USA*, **102**, 4045–4050.

Sujatha,S. *et al*. (2001) PALI: a database of alignments and phylogeny of homologous protein structures. *Bioinformatics*, **17**, 375–376.