Structural bioinformatics

**MMM: a sequence-to-structure alignment protocol**

Brajesh K. Rai, Carlos J. Madrid-Aliste, J. Eduardo Fajardo and András Fiser*

Department of Biochemistry and Seaver Center for Bioinformatics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Received on June 2, 2006; revised on July 17, 2006; accepted on August 16, 2006
Advance Access publication August 23, 2006

ABSTRACT

Motivation: Accurate alignment of a target sequence to a template structure continues to be a bottleneck in producing good quality comparative protein structure models.

Results: Multiple Mapping Method (MMM) is a comparative protein structure modeling server with an emphasis on a novel alignment optimization protocol. MMM takes inputs from five profile-to-profile based alignment methods. The alternatively aligned regions from the input alignment set are combined according to their fit in the structural environment of the template structure. The resulting, optimally spliced MMM alignment is used as input to an automated comparative modeling module to produce a full atom model.

Availability: The MMM server is freely accessible at http://www.fiserlab.org/servers/mmm

Contact: andras@fiserlab.org

1 INTRODUCTION

Comparative protein structure modeling relies on detectable similarity spanning most of the modeled sequence and at least one known structure (Fiser, 2004). When the structure of one protein in the family has been determined by experiment, the other members of the family can be modeled based on their alignment to the known structure. Accurate alignment of a target sequence to a template structure continues to be a bottleneck in producing good quality homology models. A number of alignment methods have been developed and are publicly available (Edgar, 2004; Madhusudhan et al., 2006). However, none of these alignment methods consistently produces better solution for all cases (Prasad et al., 2003; Rai and Fiser, 2006). Furthermore, alignments produced by different methods are often better in some regions and worse in others when compared with each other. One possible solution to this problem is to consider several alignment methods and combine better-aligned parts into a unique solution (Kosinski et al., 2005).

Multiple Mapping Method (MMM) has been developed to minimize errors associated with input alignments. The details of the method and its performance have been recently described elsewhere (Rai and Fiser, 2006). In brief, MMM constructs an optimal and often unique solution from an arbitrary set of input alignments. The two principal components of this method are (i) Sampling and (ii) Scoring. MMM efficiently explores a limited, but biologically relevant, sampling space, which is defined by the observed differences in a set of alternative alignments of the same template and target sequence using different methods (or using the same method but different parameters). A composite environment-specific scoring function is used to evaluate various sampling scenarios that is composed of (i) environment-specific substitution matrices from FUGUE (Shi et al., 2001); (ii) a 3D–1D substitution matrix, H3P2 (Rice and Eisenberg, 1997), that scores the matches of predicted secondary structure of the target sequence to the observed secondary structures and accessibility types of the template residues; and (iii) a statistically derived residue–residue contact energy term, which determines the compatibility of alternative variable segments in the protein environment (Miyazawa and Jernigan, 1996).

MMM combines the better-aligned parts into a unique solution, which, on average, is more accurate than any of the input alignments alone (Rai and Fiser, 2006).

2 IMPLEMENTATION

The current implementation of the server has several new additions as compared with the original publication. One of them is the expansion of the set of alignment tools that are implemented; another one is the automated building of multiple sequence profiles, to perform profile-to-profile alignments instead of simple pairwise alignments.

The server takes as input the target sequence (to be modeled) and the Protein Data Bank (PDB) code of the template structure or a user uploaded coordinate file that serves as template in a subsequent modeling exercise. The server currently provides five competitive alternative alignment approaches of which at least two need to be selected: ClustalW (Thompson et al., 1994), ClustalW with modified gap penalty function, Align2D (Sali and Blundell, 1993), MUSCLE (Edgar, 2004) and T-Coffee (Notredame et al., 2000). In addition any number of other alignments can be added from other sources, e.g. manually edited ones.

Next, a newly developed module, BlastProfiler is run to build sequence profiles for both the target and template sequences. BlastProfiler initiates a PSI-BLAST search on a locally installed and frequently updated NR (Boeckmann et al., 2003) database. The program then parses all iterations of PSI-BLAST outputs and locates and stores those pairwise alignments between the query and database sequences that meet the filtering criteria. The values specified for filtering are as follows: (i) Lower and upper cutoffs for percent sequence identities between the hit and the query, as reported in the pairwise Blast alignment; 30 and 90%, respectively. (ii) Lower bound for alignment length is 30 residues. (iii) Maximal E-value for each hit is 1E−4. (iv) Minimal required coverage...
of the query in the alignment, in percentage; default: 30%. Typically the PSI-BLAST output contains more than one alignment for the same hit sequence, especially when multiple iterations are performed. Such alternative alignments may include either the same or different regions of the hit sequence. Alignments to different regions of the target are kept as separate entries. Two alignments that involve the same hit sequence are considered redundant if the overlap is >50%. Because alignments produced in later iterations contain more specific information about the sequence profile, these alignments are preferred over earlier ones in case of overlapping cases.

The second major step in the selection of a set of representative hit sequences is removing sequence redundancy by the CD-HIT clustering program (Li et al., 2002) at 40% identity level. At the end of this step, alternative profile-to-profile-based sequence alignments are available, which are used as input to the MMM module (Rai and Fiser, 2006). MMM samples all alternatives and splices together an optimal, consensus alignment, which is then used as input to MODELLER (Sali and Blundell, 1993) to generate an all-atom comparative model for the target sequence.

The performance of the original MMM method is discussed in detail in a recent publication (Rai and Fiser, 2006). The algorithm was tested on a dataset of 1400 protein pairs using 11 combinations of 2–5 alignment methods. In all cases MMM showed statistically significant improvement by reducing alignment errors in the range of 3–17%. MMM also compared favorably over two alignment meta-servers tested (Lambert et al., 2002; Prasad et al., 2003). Figure 1 illustrates that the current implementation of MMM using sequence profiles and an optimized set of input alignments further improves performance over the previous version of the program.

### 3 USING THE SERVER

The MMM server has a straightforward interface. The user only needs to provide a target sequence, which can be entered in a text box, or can be uploaded as a text file and a PDB code for the template structure or upload a coordinate file in the PDB format. The target sequence must either be in pure text containing one letter amino acid codes (without any header), in the FASTA (Pearson, 1990) or in the PIR format. The user also needs to supply a return e-mail address.

The MMM server returns a full atom model in the PDB format as output. In addition the alignment that is used for modeling is sent to the user by e-mail.

### ACKNOWLEDGEMENTS

Financial support for this work was provided by NIH GM62519-04 and the Seaver Foundation.

Conflict of Interest: none declared.

### REFERENCES


