J Struct Funct Genomics (2009) 10:95–99
DOI 10.1007/s10969-008-9044-9

Improved scoring function for comparative modeling using the M4T method

Dmitry Rykunov · Elliot Steinberger · Carlos J. Madrid-Aliste · András Fiser

Received: 14 August 2008 / Accepted: 16 October 2008 / Published online: 5 November 2008

Abstract Improvements in comparative protein structure modeling for the remote target-template sequence similarity cases are possible through the optimal combination of multiple template structures and by improving the quality of target-template alignment. Recently developed MMM and M4T methods were designed to address these problems. Here we describe new developments in both the alignment generation and the template selection parts of the modeling algorithms. We set up a new scoring function in MMM to deliver more accurate target-template alignments. This was achieved by developing and incorporating into the composite scoring function a novel statistical pairwise potential that combines local and non-local terms. The non-local term of the statistical potential utilizes a shuffled reference state definition that helped to eliminate most of the false positive signal from the background distribution of pairwise contacts. The accuracy of the scoring function was further increased by using BLOSUM mutation table scores.

Keywords Homology modeling · Comparative modeling · Multiple mapping method · Target-template alignment · Template selection

D. Rykunov · E. Steinberger · C. J. Madrid-Aliste · A. Fiser (✉)
Department of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461, USA
e-mail: afiser@aecom.yu.edu

D. Rykunov · E. Steinberger · C. J. Madrid-Aliste · A. Fiser
Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461, USA

Abbreviations

MMM	Multiple mapping method
M4T	Multiple mapping method with multiple templates

Introduction

A key aim of the worldwide structural genomics efforts is the experimental solution of the three dimensional structures of a carefully selected few thousand target sequences of structurally uncharacterized proteins. Subsequently these newly solved structures will be used as templates for computational modeling of about 10–100 times more proteins whose sequences are related [1]. These efforts further underline the importance of theoretical approaches to structure modeling, since consequently more than 99% of all three dimensional models will be obtained computationally [2]. Homology modeling proved to be the most accurate approach for protein structure prediction provided that a three dimensional structure of a sequentially similar template protein exist [3–9]. Two key steps in homology modeling for the remote target-template cases are the identification and optimal combination of the template structures and the construction of a correct target-template alignment.

We recently introduced a new approach, called multiple mapping method (MMM) to address the target-template alignment problem [10, 11]. Depending on the overall target-template similarity, various sequence alignment methods can result in solutions that share identically aligned (constant regions) and differently aligned fragments (variable regions) between the target and template sequences. MMM constructs an optimal alignment by identifying and combining the more accurate solutions of

alternatively aligned segments with the constant parts of the input alignments. It is achieved by mapping the alternatively aligned segments to the corresponding environment of the template and by using a scoring function that evaluates the fit of each alternatively aligned segment. MMM has two advantages over the input, sequence based alignments. First, it can identify additional residues that may be far in the sequence but close in space to the alternatively aligned segments, therefore it can increase the sensitivity of the alignment by comparing not only a pair of residue positions, but also the fit of all those positions that compose the environment of the segment. Second, by incorporating structural information, MMM can take advantage of such scoring function terms that depend on the structural environment and not only on the compared sequential positions.

It has been demonstrated that combining multiple template structures results in better quality models as compared to models built on single templates [12]. This conclusion has been confirmed at the meetings on Critical Assessment of Techniques for Protein Structure Prediction [13]. While this idea seems to be obvious, its automated implementation is not straightforward. It has been shown that a trivial combination of available templates does not provide any advantage [14]. We established an approach, multiple mapping method with multiple templates (M4T) [15, 16], to optimally select and combine the most suitable templates. The method uses an iterative clustering approach, which takes into account information on the template sequence similarity to the target, sequence similarity among the templates, completeness of structural domain of the template, its experimental quality and the unique contribution of each template to the target. Here we report new algorithmic developments that improved template detection and the introduction our new pairwise statistical potential that increased the accuracy of MMM approach for target-template alignments.

Materials and methods

Template database for M4T

A local database for protein sequences of known structure was compiled. First, all protein sequences from the Protein Data Bank (PDB) [17] were clustered with CD-HIT [18] at 99.9% sequence identity level. Each cluster was restricted to a subset of hits that are less than eight residues different in length. The rest of the cluster was recursively re-clustered. Finally, one entry with the highest resolution was selected from each cluster as a representative. We refer the resulting database as "aapdb" in order to distinguish it from NCBI "pdbaa" database. Current version of the "aapdb" database can be downloaded from the M4T server web page (http://www.fiserlab.org/servers//M4T/aapdb.fasta).

MMM testing set

The previously described set of 1,624 structurally aligned protein pairs [10] was used in this study for analysis of the scoring of alternative alignments. MMM alignments produced from CLUSTALW [19] and ALIGN2D [20, 21] or by CLUSTALW and MUSCLE [22] alignments were compared to structural alignments obtained with STAMP [23]. The testing set was additionally filtered in the following way: (a) pairs with nearly identical alignments (less than 10 residues in variable regions) were excluded from the set; (b) pairs with little agreement between any of participating alignments and the "true" (structural) alignment were also excluded from further consideration. We required that between the variable regions of the input alignments and the corresponding structural alignment at least 20% or a minimum five residues (whichever is bigger) must be shared. This filtering reduced testing set to 1,397 pairs for CLUSTALW-LIGN2D test and to 1,216 pairs for CLUSTALW-MUSCLE test.

MMM scoring function

A composite scoring function is employed in MMM to assess the fitness of alternative alignments of variable regions within the structural context of the template [10]. Originally, it combined three scoring methods of different nature: environment specific substitution matrices from FUGUE [24], secondary structure based 3D–1D substitution matrix H3P2 [25] and pairwise residue-based contact potential [26]. The latter one we will further refer as "MJ". The scores in these terms were converted to *Z*-scores and combined with weights 0.4, 0.4 and 0.2 to calculate a pseudo-energy score for the fit of he aligned segment in the given structural neighborhood [10].

In addition to these scoring methods a direct sequence similarity scoring method implemented in BLOSUM62 matrix [27] as well as our recently developed pairwise statistical potential were explored in the current work.

The latter one (referred as "RF") is C_β-based residue–residue pair-potential (Rykunov and Fiser, in preparation) derived in a similar way to the all-atom potential described previously [28]. Briefly, this potential was generated from distances between C_β-atoms measured on the same set of proteins as described for all-atom potential. Artificial C_β-pseudo-atoms were generated for Glycine residues. To improve the signal to noise ratio a shuffled reference state was used [28]. Our earlier designed shuffled reference state

was further improved in this work by introducing environmental dependence. This was achieved by choosing a random position for a given atom only from that subset of atoms that make a similar number of inter-residue contacts. Atom pairs were counted within 1 Å bins, with first bin accumulating all pairs shorter than 4 Å. Pairs with sequence separation less than 3 (i.e. $i{:}i+1$, $i{:}i+2$) were excluded.

Resulting MMM scoring function can be expressed as

$$S = Z_{\mathrm{BS62}} + Z_{\mathrm{FUGUE}} + Z_{\mathrm{H3P2}} + Z_{\mathrm{RF}} \quad (1)$$

where corresponding Z-scores are calculated from scores for test alignment and those calculated for randomized environments as described in [10]. BLOSUM62 score is obtained as

$$S_{\mathrm{BS62}} = \sum_i s\left(A_i^t, A_i^q\right) \quad (2)$$

here summation is taken over all residues in the variable region(s), upper index "t" designates template sequence at the position i, and upper index "q" stays for query sequence. RF potential is calculated as:

$$S_{\mathrm{RF}} = \sum_i \sum_{j,|i-j|>2} \varepsilon_{ij}\left(A_i^q, A_j^q\right) \quad (3)$$

here, first summation is taken over all residues in the variable region(s), while second summation is taken over all residues that are within 12 Å from the residue i.

Results

Improved MMM scoring function

As we pointed out in the Methods section, the MMM scoring function was originally composed of FUGUE, H3P2 and MJ terms [10]. The present work explored the contribution of two other possible components, BLOSUM62 substitution matrix and a distance-dependent statistical "RF" potential described in the Methods section. Alignments obtained with MMM were compared to structural alignments. Test cases were split into two categories of difficulty: target-template pairs sharing less than 30%, and more than 30% sequence identity, respectively. Results of these tests obtained with different scoring function terms and their combinations are shown in Fig. 1. Only the variable regions of the participating MMM alignments were used in the analysis, since the rest was identically aligned by both compared methods by definition and therefore would result in identical scores. "Ideal" performance refers to the number of positions properly identified by at least one participating methods. It is the maximal possible agreement with a structural alignment that can be achieved by MMM with the given set of input sequence alignments. In other words, if a certain—correct—alignment is not sampled by any of the input alignments then it will not be possible to identify through MMM. The lower value for "ideal" performance for the CLUSTALW-MUSCLE alignment methods as compared to CLUSTALW-ALIGN2D "ideal" performance is due to the fact that variable regions in the former pairs tend to be shorter and more difficult to score. "Ideal" values calculated for whole alignments are $78.66 \pm 0.29\%$ and $83.57 \pm 0.25\%$ for CLUSTALW-ALIGN2D and CLUSTALW-MUSCLE alignment pairs, respectively. When comparing the scoring function terms individually in order to identify the best performing ones, MJ potential turns out to be the least selective in all test sets. BLOSUM62 is the most selective individual scoring method for CLUSTALW-ALIGN2D alignments and FUGUE is the second best one (Fig. 1a), while their ranks are reversed in the case of CLUSTALW-MUSCLE alignments (Fig. 1b). The "RF" pairwise potential introduced in this work is significantly more selective than "MJ" potential, especially in the difficult and most frequently occurring cases when less than 30% of the residues are shared by target and template sequences.

Template detection protocol

In addition to improvement in the MMM scoring function we also added a new module to the template search protocol of M4T. In the original implementation of M4T template candidates were detected with three iterations of PSI-BLAST search on PDB [29]. This search is quick and results in suitable templates for almost 50% cases as we observed during ongoing CASP8 experiment. In the improved version, if the PSI-BLAST search does not result a hit against PDB or results in a hit that covers less than 60% of the target, a sequence profile is built by searching the target sequence against the "nr" database, which is used then to run a profile-to-profile alignment against possible targets in PDB. This improvement allowed us to model additional 17% of CASP8 targets. A third, rather time consuming step, is a standard PSI-BLAST search against "nr" database for a maximum of 10 iterations. It is invoked only if the first two fail to result in any suitable templates and in our experience is activated in about 4% of the cases only.

Availability of methods

MMM and M4T methods are publicly accessible via web servers at http://www.fiserlab.org/servers.

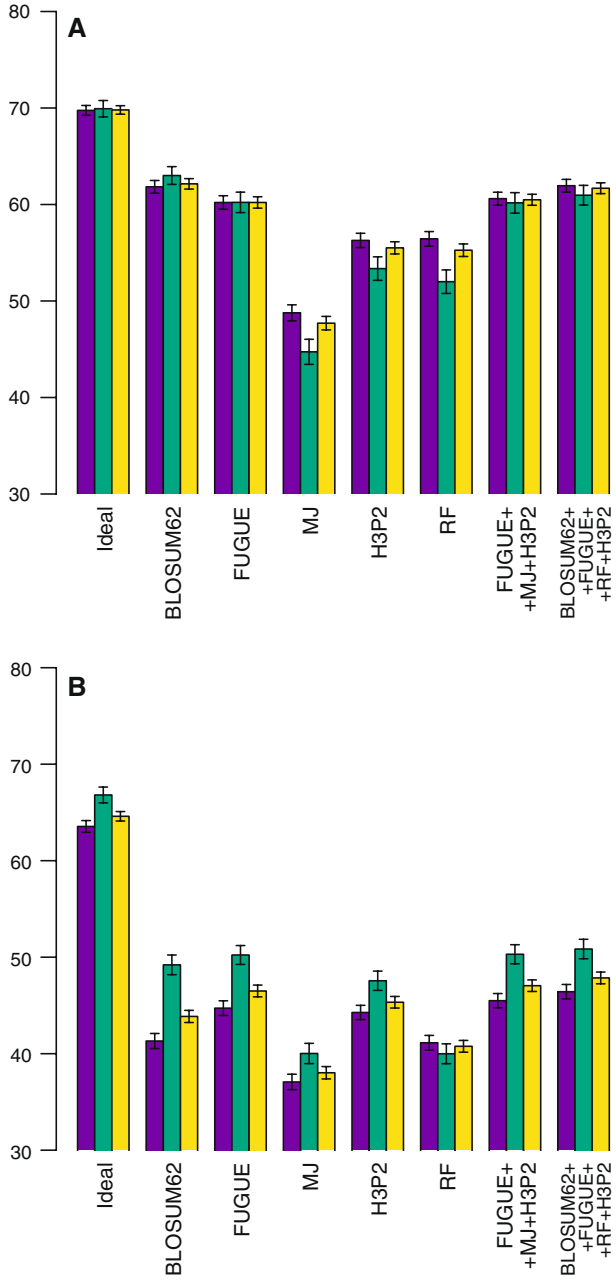

Fig. 1 Agreement between structural and MMM alignments determined with different components of the MMM scoring function for **a** CLUSTALW-ALIGN2D and **b** CLUSTALW-MUSCLE MMM alignments. *Blue bars* represent protein pairs sharing less than 30% identical residues, *green bars*—more than 30%, and *yellow bars* show overall performance. *Error bars* represent corresponding standard errors. Values obtained with "Ideal", "FUGUE", "MJ", "H3P2" and "FUGUE + MJ + H3P2" methods are shown for reference

Discussion

We introduced a new scoring function for the MMM. The composite nature of the MMM scoring function is important since different components demonstrate different performances in "easy" and "difficult" modeling cases, and their combination performs most selectively over the whole range of that target difficulties. For instance, BLOSUM62 scoring term is more selective than FUGUE if MMM alignments are constructed from CLUSTALW and ALIGN2D (which is not surprising because CLUSTALW is based on the BLOSUM62 matrix and it generally produces more accurate alignment than ALIGN2D), while FUGUE performs better when the more accurate MUSCLE method replaces ALIGN2D. We introduced a distance-dependent statistical potential (RF term) that demonstrates substantially superior selectivity as compared to the earlier applied MJ contact potential. In case of CLUSTALW-MUSCLE-based MMM alignments the RF pairwise statistical potential shows higher selectivity than BLOSUM62 term for low sequence identity cases (Fig. 1b, blue bars). As a result, the new, combined scoring function (BLOSUM62 + FUGUE + H3P2 + RF) is superior to the old one (FUGUE + H3P2 + MJ) and to any scoring terms when those are used individually. After converting scores from BLOSUM, FUGUE, RF and H3P2 terms into *Z*-scores over randomized samples and combining them into a scoring function we obtained a superior performance of MMM over the previously used combination of FUGUE, MJ and H3P2 scores.

Although MMM significantly improves alignment accuracy in difficult cases when comparing to individual alignment methods, but it is still unable to produce alignments identical or really close to the one obtained from structural superposition. As one can see in Fig. 1, "ideal" performance for the variable regions using MMM with the current alignment sampling technique and scoring function terms is about 65–70%. Meanwhile the overall performance for the entire length of the alignments are 78–83%. Thus, the biggest improvement in the overall alignment quality arises from the use of more accurate MUSCLE alignment method instead of ALIGN2D. However, even in that case 17% of residues remain misaligned and cannot be recovered with MMM, simply because MMM will not be presented with a correct alignment solution to identify. Further improvement in the alignment accuracy can be achieved when better alignment methods will become available that sample the sequence alignment space more accurately.

Acknowledgements This work was supported by NIH GM62519-04. This work is dedicated to the memory of Elliot Steinberger, who passed away recently.

References

1. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S (1999) Structural genomics: beyond the human genome project. Nat Genet 23:151

2. Manjasetty BA, Shi W, Zhan C, Fiser A, Chance MR (2007) A high-throughput approach to protein structure analysis. Genet Eng (N Y) 28:105–128
3. Cardozo T, Totrov M, Abagyan R (1995) Homology modeling by the ICM method. Proteins 23:403. doi:10.1002/prot.340230314
4. Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips SE, Poljak RJ (1986) The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. Science 233:755. doi:10.1126/science.3090684
5. Fiser A (2004) Protein structure modeling in the proteomics era. Expert Rev Proteomics 1:97–110. doi:10.1586/14789450.1.1.97
6. Greer J (1981) Comparative model-building of the mammalian serine proteases. J Mol Biol 153:1027. doi:10.1016/0022-2836(81)90465-4
7. Levitt M (1992) Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 226:507. doi:10.1016/0022-2836(92)90964-L
8. Sutcliffe MJ, Haneef I, Carney D, Blundell TL (1987) Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1:377. doi:10.1093/protein/1.5.377
9. Yang AS, Honig B (1999) Sequence to structure alignment in comparative modeling using PrISM. Proteins 37:66. doi:10.1002/(SICI)1097-0134(1999)37:3+<66::AID-PROT10>3.0.CO;2-K
10. Rai BK, Fiser A (2006) Multiple mapping method: a novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. Proteins Struct Funct Bioinform 63:644–661. doi:10.1002/prot.20835
11. Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A (2006) MMM: a sequence-to-structure alignment protocol. Bioinformatics 22:2691–2692. doi:10.1093/bioinformatics/btl449
12. Sanchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins (Suppl 1):50–58. doi:10.1002/(SICI)1097-0134(1997)1+<50::AID-PROT8>3.0.CO;2-S
13. Venclovas C, Margelevicius M (2005) Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. Proteins 61(Suppl 7):99–105. doi:10.1002/prot.20725
14. Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA (2003) Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins 53(Suppl 6):424–429. doi:10.1002/prot.10549
15. Fernandez-Fuentes N, Madrid-Aliste CJ, Rai BK, Fajardo JE, Fiser A (2007) M4T: a comparative protein structure modeling server. Nucleic Acids Res 35:W363–W368
16. Fernandez-Fuentes N, Rai BK, Madrid-Aliste CJ, Eduardo Fajardo J, Fiser A (2007) Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics 23:2558–2565. doi:10.1093/bioinformatics/btm377
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235. doi:10.1093/nar/28.1.235
18. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658. doi:10.1093/bioinformatics/btl158
19. Thompson JD, Higgins DG, Gibson TJ (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. doi:10.1093/nar/22.22.4673
20. Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A (2006) Variable gap penalty for protein sequence-structure alignment. Protein Eng Des Sel 19:129–133. doi:10.1093/protein/gzj005
21. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 95:13597–13602. doi:10.1073/pnas.95.23.13597
22. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi:10.1093/nar/gkh340
23. Russell RB, Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14:309–323. doi:10.1002/prot.340140216
24. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243. doi:10.1006/jmbi.2001.4762
25. Rice DW, Eisenberg D (1997) A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 267:1026–1038. doi:10.1006/jmbi.1997.0924
26. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256:623–644. doi:10.1006/jmbi.1996.0114
27. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919. doi:10.1073/pnas.89.22.10915
28. Rykunov D, Fiser A (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins Struct Funct Bioinform 67:559–568. doi:10.1002/prot.21279
29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. doi:10.1093/nar/25.17.3389