W24–W28 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 20 May 2011*
doi:10.1093/nar/gkr393

CLICK—topology-independent comparison of biomolecular 3D structures

M. N. Nguyen[1], K. P. Tan[1] and M. S. Madhusudhan[1,2,3,*]

[1]Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671, [2]Department of Biological Sciences, National University of Singapore and [3]School of Biological Sciences, Nanyang Technological University, Singapore

Received February 26, 2011; Revised April 19, 2011; Accepted May 3, 2011

ABSTRACT

Our server, CLICK: http://mspc.bii.a-star.edu.sg/click, is capable of superimposing the 3D structures of any pair of biomolecules (proteins, DNA, RNA, etc.). The server makes use of the Cartesian coordinates of the molecules with the option of using other structural features such as secondary structure, solvent accessible surface area and residue depth to guide the alignment. CLICK first looks for cliques of points (3–7 residues) that are structurally similar in the pair of structures to be aligned. Using these local similarities, a one-to-one equivalence is charted between the residues of the two structures. A least square fit then superimposes the two structures. Our method is especially powerful in establishing protein relationships by detecting similarities in structural subdomains, domains and topological variants. CLICK has been extensively benchmarked and compared with other popular methods for protein and RNA structural alignments. In most cases, CLICK alignments were statistically significantly better in terms of structure overlap. The method also recognizes conformational changes that may have occurred in structural domains or subdomains in one structure with respect to the other. For this purpose, the server produces complementary alignments to maximize the extent of detectable similarity. Various examples showcase the utility of our web server.

INTRODUCTION

The 3D structures of biomolecules at near atomic-level resolution often give us unique insights into their evolution and function. This has been extensively studied for molecules such as proteins, where similarity in 3D structure often implies homology (1–4). Given the rapid pace with which new structures are deposited in the PDB (5), it is crucial to have tools to classify and categorize these structures and investigate them for similarities at different levels. In the case of proteins, it has been beneficial to have categorization based on 3D-fold types that follow the primary sequence order (2–4). There are, however, several structural features whose similarities do not follow primary sequence order. Detecting these similarities in topologically different structures establishes new relationships between proteins in the different categories mentioned above. Frequently, these new relationships are of functional significance (6).

The functionality of a biomolecule depends on the spatial orientation of its chemically various atoms. Sometimes different topologies result in similar/same functionality (6–8). Methods (1,9–20) that align a pair of structures imposing constraints on sequence order and topology may be inadequate to establish such functional similarities. To establish these relationships one needs to make use of non-sequential and non-topological protein structure matching programs (21–23). Here, we report on a web server that uses the CLICK algorithm (24) to align the 3D structures of any pair of biomolecules, independent of topology.

The CLICK algorithm aligns 3D structures by matching cliques of points. Cliques are groupings of representative atoms of the biomolecules within a certain spatial proximity. Any pairwise distance among clique members is less than a set threshold. Points in a clique could also be assigned values for features such as secondary structure, solvent accessibility and depth. A pair of biomolecules is structurally aligned by matching cliques with similar features. In general, any pair of constellation of points can be aligned using CLICK and thus comparison between different types of biomolecules (for example, DNA with RNA) is also possible. To the best of our knowledge, this is the first web server equipped with this capability. We hope that our server is useful for a wide range of biomolecule structural analysis, especially in detecting conformational change, similarity of structural motifs (both local and global) and evolutionary relationships.

*To whom correspondence should be addressed. Tel: (65) 6478 8500; Fax: (65) 6478 9047; Email: madhusudhan@bii.a-star.edu.sg

METHOD

Briefly, the CLICK algorithm consists of four sequential steps (for a comprehensive description of the method, see Supplementary Material).

Extracting features

Residues of a biomolecule are featured by the Cartesian coordinates of one or more representative atoms. If the biomolecules in question are proteins, additional structural features such as side-chain solvent accessibility, secondary structure and residue depth (25–26) are computed.

Forming cliques

We define a *n*-body clique as a subset of *n* points, where the Euclidean distance between any pair within the clique is within a predefined threshold. For each of the two structures to be compared, all possible *n*-body (*n* in the range of 3–7) cliques are computed from the representative atoms.

Clique matching

Cliques are matched by the superimposition of their Cartesian coordinates subject to the matching of other features. The objective here is to establish local structural similarities.

Alignment

Clique matching identifies structurally equivalent residues in the two structures. Using these equivalences, a final 3D least squares fit is performed to superimpose the two structures. The matching of cliques is not necessarily unique, *i.e.* multiple structural alignments are possible to be generated. The chosen alignment is the one that maximizes structure overlap.

USER PERSPECTIVE

Input

The web server is freely accessible at http://mspc.bii.a-star.edu.sg/click without login requirements.

Input biomolecular structures can be submitted by specifying the four-letter code for existing structures deposited in the PDB, or by uploading structures in PDB format. In addition to the four-letter code, users can specify the identity of particular chains from the two structures. This specification is however optional as CLICK produces alignments irrespective of the number of chains in the PDB structure. By default, it considers all residues in the PDB file.

Users are given the options to select one or more representative atoms (default: C^{α} atom) for individual residues of the molecule. For protein structure alignments, residue solvent-accessible surface area, secondary structure and depth information can be included as structural features to guide the alignment. These additional features are not yet operable for aligning other biomolecular structures. When four-letter PDB codes are specified, by default solvent-accessible surface area and secondary structure features are selected.

Output

The structure alignment is shown on two lines, one line per structure. The number and chain identifier of the first aligned residue on both structures precedes the listing of the residue one-letter codes. Each time the alignment fragments (probably because of topological differences), the number and chain identifier of the last residue in the fragments are also listed. Accompanying each alignment is a 3D rendition of the structural superimposition using Jmol. (Figures 1 and 2)

Should there be conformational changes in the proteins being compared, CLICK first reports the largest alignment, in terms of number of residues aligned. The method then seeks to compare the regions of the proteins that were not aligned first. The detection of further structural similarity results in additional output alignments, shown one below the other in the aforementioned format.

The alignments are downloadable in PIR (Protein Information Resource) format and in CLICK format that shows one equivalent representative atom match per line. Also downloadable are the coordinates of the superimposed structures in PDB format.

Statistics relevant to the alignment including structure overlap, root mean square deviation (RMSD), fragment score, topology score, number of representative atoms in the two structures, length of the match and the number of identical residue matches are displayed in a table. Detailed help pages explain the significance of the different alignment measures.

Users can also download an executable file of the CLICK program along with associated library files used in the server.

Examples

We demonstrate the versatility of the server with five different types of alignments. (i) A conventional structural alignment between a pair of proteins (PDB codes 1ANG and 7RSA) that are ~35% identical in sequence and similar in overall topology. (ii) An alignment between two protein structures (1UBP chains A, B and C; and 1E9Y chains A and B) with multiple chains. Chains A and B of 1UBP superimpose onto chain A of 1E9Y, and chain C of 1UBP superimposes onto chain B of 1E9Y. This showcases the feature of CLICK to align structures regardless of chain breaks. (iii) An alignment between two proteins (PDB codes 1IWM chain A and 1OXE chain A) that are structurally similar and topologically different. The fragmented alignment is a result of the different topology. (iv) Multidomain proteins, where individual domains in one structure (PDB 1AIV) have a structural equivalence in the other (PDB 1OVT), but the relative orientations between domains differ in the two structures. Domain swapping and rigid body shift detection belong to this category. (v) An alignment showcasing the general utility of CLICK. Two DNA double-helical fragments bound to proteins (PDB codes 1YSA and 2AYG) are aligned with one another. The representative atom used in this instance was C3′. In principle, for such examples, representative atoms could have been used for amino acid and nucleotides residues at the same time.

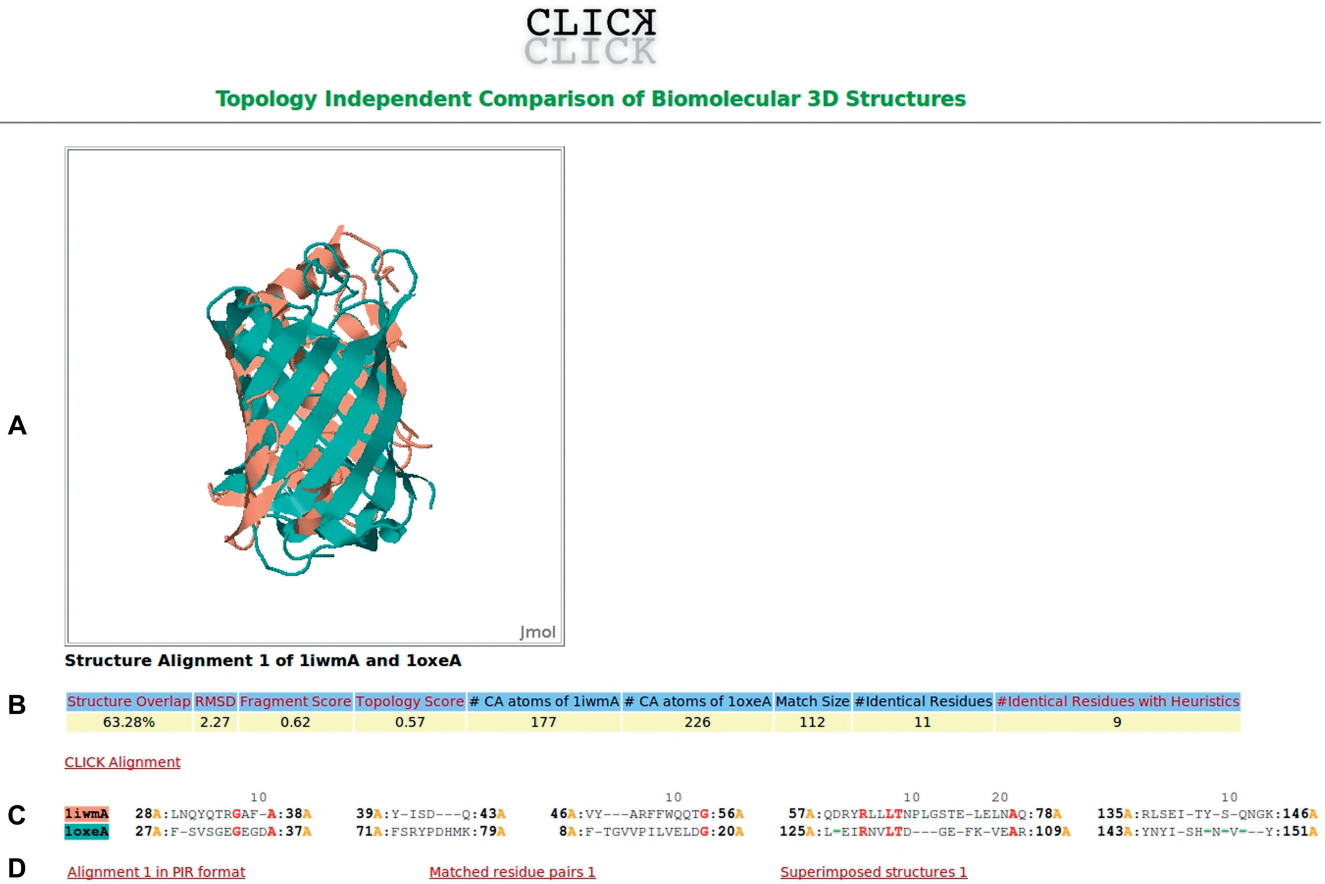

Figure 1. A snapshot of the server showing the output of a structural alignment of two topologically different yet structurally similar proteins that belong to a different SCOP families, PDB codes 1iwm:A (salmon and SCOP entry: b.125.1.2) and 1oxe:A (green and SCOP entry: d.22.1.1), according to CLICK. (**A**) 3D representation of the superimposition is shown in an embedded JMol viewer. (**B**) The measures of alignment accuracy such as structure overlap (coverage), RMSD, sequence identity, topology score and fragment score. (**C**) The sequence alignment between the two proteins as inferred from the structural alignment. The conserved residues are shown in bold and red lettering. (**D**) Download links to the resulting sequence alignment in PIR format, the detailed alignment and matched residue (atom) pairs in text format, as well as the link to download the superimposed coordinates of the two structures, in PDB format.

IMPLEMENTATION

The program run time increases with both sizes of input structures and number of best matched cliques. On average CLICK took 1 s to perform a single alignment of a pair of proteins each of size ~150 residues on an Ubuntu 8.04 Linux platform with 3.00 GHz CPU (Core 2 Duo E8400) and 3.5 GB primary memory.

PERFORMANCE

The performance of CLICK was compared with other popular structural alignment methods. For protein structure comparisons, three different data sets consisting of 9357, 64 and 89 pairs of structures were used. For details on each of these data sets, refer to Supplementary Tables S1a and b, S2a and b, and S3a and b. Over each of these datasets, the alignment accuracy of CLICK was compared with other popular protein structure alignment programs including MUSTANG (21), Geometric Hashing (C-alpha Match) (27–29), SALIGN (19), DALI (22,30) and alignments from the HOMSTRAD database (4). All these programs were run using default parameters, and no effort was made to adjust the parameters for specific cases. With the exception of SALIGN, CLICK alignments are statistically significantly better than those of the other methods compared in terms of structure overlap. In all of these datasets, the structure overlap from CLICK alignments was never below 40%.

CLICK was also compared with programs that align RNA 3D structures, including ARTS (31,32) and SARA (33,34). See Supplementary Tables S4a and b for details. The structure overlap of CLICK in this comparison was also statistically significantly better.

CONCLUSION

The CLICK method formalizes the biomolecular structure superimposition problem as one of feature-point matching. The features include Cartesian coordinates, solvent-accessible surface area and residue depth. The method is flexible and can be easily implemented with

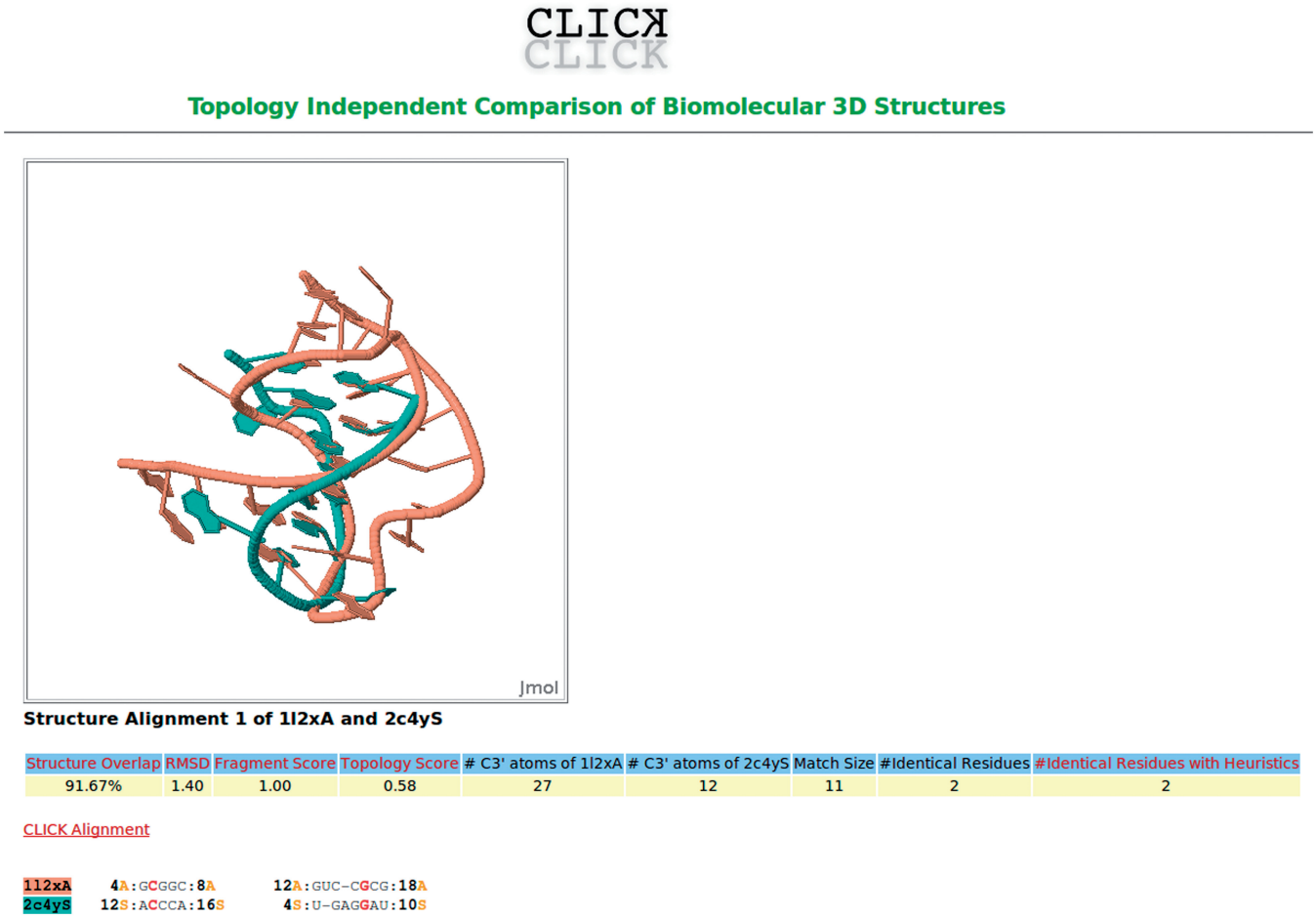

Figure 2. Another snapshot of the server showing RNA structure alignments (PDB codes 1l2x:A color salmon and 2c4y:S, color green).

other features considered important in different contexts. We have extensively benchmarked the method over various protein and RNA structure data sets. The accuracy of CLICK alignments, in terms of structure overlap, is on par or statistically significantly better than several other existing methods for protein and RNA alignments. CLICK performs structural superposition on pairs of structures based on similarity of local structural packing, and thus is capable of aligning structures with dissimilar topologies, conformations or even molecular types. These unique properties make CLICK a utilitarian tool for detecting divergent evolution due to topology changes, convergence evolution where substructures of proteins are similar to one another, and conformational change such as domain swap and rigid-body shift where the relative orientation of domains change. This server now sets the stage for interesting investigations including topology-independent structural motif detection, biomolecular structure design and super-secondary structure classifications in not just proteins but also in molecules such as RNA, DNA, etc.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Chandra Shekar Verma and Dr Raghavan Varadarajan for valuable comments and insights. We also offer special thanks to Yong Taipang for his help in setting up, maintaining and improving the server. We appreciate constructive feedbacks from members of the Biomolecular Molecular Simulation and Design division of the Bioinformatics Institute.

FUNDING

Funding for open access charge: Biomedical Research Council (A*STAR), Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
2. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

3. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
4. Stebbings,L.A. and Mizuguchi,K. (2004) HOMSTRAD: recent developments of the homologous protein structure alignment database. *Nucleic Acids Res.*, **32**, D203–D207.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
7. Hou,J., Sims,G.E., Zhang,C. and Kim,S.H. (2003) A global representation of the protein fold space. *Proc. Natl Acad. Sci. USA*, **100**, 2386–2390.
8. Pascual-Garcia,A., Abia,D., Ortiz,A.R. and Bastolla,U. (2009) Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput. Biol.*, **5**, e1000331.
9. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
10. Friedberg,I., Harder,T., Kolodny,R., Sitbon,E., Li,Z. and Godzik,A. (2007) Using an alignment of fragment strings for comparing protein structures. *Bioinformatics*, **23**, e219–e224.
11. Csaba,G., Birzele,F. and Zimmer,R. (2008) Protein structure alignment considering phenotypic plasticity. *Bioinformatics*, **24**, i98–i104.
12. Leslin,C.M., Abyzov,A. and Ilyin,V.A. (2007) TOPOFIT-DB, a database of protein structural alignments based on the TOPOFIT method. *Nucleic Acids Res.*, **35**, D317–D321.
13. Konagurthu,A.S., Stuckey,P.J. and Lesk,A.M. (2008) Structural search and retrieval using a tableau representation of protein folding patterns. *Bioinformatics*, **24**, 645–651.
14. Abyzov,A. and Ilyin,V.A. (2007) A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Struct. Biol.*, **7**, 78.
15. Veeramalai,M., Ye,Y. and Godzik,A. (2008) TOPS++FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model. *BMC Bioinformatics*, **9**, 358.
16. Andreeva,A., Prlic,A., Hubbard,T.J. and Murzin,A.G. (2007) SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res.*, **35**, D253–D259.
17. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
18. Ye,Y. and Godzik,A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–W585.
19. Madhusudhan,M.S., Webb,B.M., Marti-Renom,M.A., Eswar,N. and Sali,A. (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, **22**, 569–574.
20. Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**, 341–348.
21. Konagurthu,A.S., Whisstock,J.C., Stuckey,P.J. and Lesk,A.M. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
22. Holm,L., Kaariainen,S., Rosenstrom,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
23. Ilyin,V.A., Abyzov,A. and Leslin,C.M. (2004) Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci.*, **13**, 1865–1874.
24. Nguyen,M.N. and Madhusudhan,M.S. (2011) Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res.*, doi:10.1093/nar/gkr348.
25. Chakravarty,S. and Varadarajan,R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, **7**, 723–732.
26. Tan,K.P., Varadarajan,R. and Madhusudhan,M.S. (2011) Depth computation and prediction of small molecule binding site. *Nucleic Acids Res.*, doi:10.1093/nar/gkr356.
27. Nussinov,R. and Wolfson,H.J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.
28. Bachar,O., Fischer,D., Nussinov,R. and Wolfson,H. (1993) A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.*, **6**, 279–288.
29. Tuncbag,N., Gursoy,A., Guney,E., Nussinov,R. and Keskin,O. (2008) Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.*, **381**, 785–802.
30. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
31. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21(Suppl. 2)**, ii47–ii53.
32. Dror,O., Nussinov,R. and Wolfson,H.J. (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res.*, **34**, W412–W415.
33. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.
34. Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, **37**, W260–W265.