

FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web

Jessica Shapiro¹ and Douglas Brutlag^{1,2,*}

¹Biophysics Program and ²Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305-5307, USA

Received February 13, 2004; Revised and Accepted March 24, 2004

ABSTRACT

The FoldMiner web server (<http://foldminer.stanford.edu>) provides remote access to methods for protein structure alignment and unsupervised motif discovery. FoldMiner is unique among such algorithms in that it improves both the motif definition and the sensitivity of a structural similarity search by combining the search and motif discovery methods and using information from each process to enhance the other. In a typical run, a query structure is aligned to all structures in one of several databases of single domain targets in order to identify its structural neighbors and to discover a motif that is the basis for the similarity among the query and statistically significant targets. This process is fully automated, but options for manual refinement of the results are available as well. The server uses the Chime plugin and customized controls to allow for visualization of the motif and of structural superpositions. In addition, we provide an interface to the LOCK 2 algorithm for rapid alignments of a query structure to smaller numbers of user-specified targets.

INTRODUCTION

Despite the ever-increasing size of the Protein Data Bank (PDB) (1), detection of structural similarity remains only a partially solved problem. While many existing superposition algorithms are able to align closely related proteins, they often provide differing and contradictory results in the more interesting cases of distantly related structures (2–4). Protein structural superposition is an essential tool in the process of garnering information about both individual proteins and protein structure in general, and is used in fields such as the selection of structural genomics targets, protein fold prediction and functional annotation (5,6). In cases of low sequence similarity, structural alignment provides an

alternative to sequence alignment for obtaining residue correspondences required for detection of remote homologies and for studies regarding the relationship between sequence and structure (5,7). Structural motifs can provide guidance to *ab initio* fold prediction methods and may serve as templates in homology modeling and threading applications (8–10), but few fully automated and unsupervised methods for motif discovery exist. As the scientific community continues to make progress in addressing such fundamental issues as the relationship between global structural similarity and functional similarity (11–14), the role of structural superposition in functional annotation, particularly of protein structures solved through structural genomics efforts, will increase as well.

Curators of structural classification schemes have noted several issues that complicate the analysis of structural similarities among groups of proteins. While a protein fold can be described by a motif that is common to all members of the fold, individual structures may have insertions consisting of entire secondary structure elements (SSEs) whose presence must be accounted for when assessing the statistical significance of structural similarities. Simply allowing for local alignments is not an ideal solution due to the presence of small structural motifs consisting of a handful of secondary structure elements that are common to many globally dissimilar folds (15). When these small motifs appear in two larger structures, a highly local alignment may seem to be of greater statistical significance than an alignment that is global with respect to both structures. In many applications, this is not a desirable result. Requiring that a superposition be global with respect to both the query and target structures, however, can cause alignments of structures known to have the same fold to be considered insignificant when the proteins share a common core fold but have other insertions and deletions with respect to one another.

We have previously described an algorithm, FoldMiner, which addresses this issue by performing unsupervised motif discovery in the context of a structural similarity search (16). Given a query structure and a database of targets, FoldMiner first discovers the core fold shared among the query and its structural neighbors and then identifies additional targets that

*To whom correspondence should be addressed. Tel: +1 650 723 6593; Fax: +1 650 723 6783; Email: brutlag@stanford.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

align to this core fold even when the query and/or the targets have additional secondary structure elements that may not align. Pairwise structural alignments are performed by LOCK 2 (16). Here, we describe the FoldMiner web server, which allows remote access to FoldMiner and LOCK 2. Both of these methods use only structural information, making them ideal tools in the analyses of the relationships among sequence, function and structure. While other automated structural alignment servers do exist (17–26), few, if any, duplicate either FoldMiner's options for visual analysis of alignments or its ability to change search parameters and/or exclude portions of the query structure from analysis without redoing alignments.

METHODS

LOCK 2 parameters

While the FoldMiner server is primarily intended for database searches and motif discovery, it also provides a direct interface to LOCK 2, the algorithm used to perform structural superpositions. LOCK 2, which has been described in detail elsewhere (16), first aligns the secondary structure elements of the query and target and then aligns individual residues. Residues are not considered to be aligned if the alpha carbons are further apart than a user-specified threshold distance, typically 3 Å. The user may also opt to force LOCK 2 not to violate sequence order constraints, or may alternatively allow residues to be aligned out of order, as would be the case in a circular permutation. The final parameter influences the speed of the algorithm; while it has little effect on speed for most single domain proteins and often a significant impact on alignment accuracy in these cases, it can significantly decrease the amount of computation time required to obtain accurate alignments of larger proteins or structures with internal repeats.

Because LOCK 2 places a great deal of emphasis on the SSE alignment phase and derives its alignment score from the secondary structure alignment, different SSE definitions may produce different results. If no SHEET or HELIX records are present in a PDB file, LOCK 2 uses DSSP to identify secondary structure elements (27). One may use different SSE definitions by changing or adding the SHEET and/or HELIX records in an uploaded PDB file.

The query structure may be specified by uploading a PDB file or by entering a PDB or SCOP (28) identifier, and a chain identifier may be supplied to limit the alignment to one chain. The target is specified in an identical fashion except when aligning the query to multiple targets, in which case file uploads are not currently permitted. The server accepts up to 10 targets in a single run; FoldMiner should be used to perform extensive database searches.

FoldMiner parameters

FoldMiner aligns a query structure to one of several databases of targets and detects a structural motif shared by the query and high-scoring targets. The query is specified in the same manner as for LOCK 2 runs. We currently offer target databases obtained from the ASTRAL server, which produces subsets of SCOP domains based on sequence similarity or structural similarity as determined by the SCOP hierarchy (29,30). We recommend the ASTRAL subset consisting of

domains that share less than 25% sequence similarity, but also offer sequence similarity thresholds ranging from 10% to 70% and target databases containing a single representative from each SCOP family, superfamily or fold.

FoldMiner uses LOCK 2 to align the query to each structure in the selected target database, with LOCK 2 parameters set by the user as described above. The server automatically calculates a statistical significance cutoff that will produce results containing on average no more than the number of false positives specified by the expectation parameter. After aligning the query structure to each of the targets in the database and selecting only those that meet the statistical significance cutoff, FoldMiner examines the alignment of the query to each target in order to determine which regions of the query tend to align well to structurally similar proteins. This analysis is possible because LOCK 2 assigns a score to each aligned secondary structure element based on its relative orientations in the query and target structures. Secondary structure elements whose orientations are conserved among the query and its structural neighbors will therefore tend to attain high alignment scores. FoldMiner then filters out false positives by requiring not only that targets achieve high scores overall, but also that they align to a specific, conserved region of the query. By adapting the scoring system in a way that places less emphasis on poorly conserved regions of the query structure, FoldMiner also identifies true positives that were not recognized in the first pass. This process iterates until the motif definition converges.

One can control the extent to which the scoring system is adapted to favor conserved regions of the query by varying a parameter along the interval [0, 1]. A value of zero turns off this process entirely, while intermediate values alter the balance between the influences of the motif and of the query structure itself on the scoring system.

EXAMPLE FoldMiner ALIGNMENTS AND MOTIFS

Annotation in SCOP indicates that the immunoglobulin-like core fold consists of a seven-stranded Greek key, but many immunoglobulin domains have additional strands that are not part of this core fold. In order to achieve high sensitivity without sacrificing specificity when performing a structural similarity search, FoldMiner learns to place greater emphasis on the quality of the alignment within the core fold. This allows for greater deviations from the query in structurally variable regions and penalizes targets that align poorly to relatively invariant regions, which are considered to be part of the core fold. No prior knowledge of the topology, location or even the size of the core fold is used.

The extracellular domain of the rat myelin adhesion membrane protein P0, which is assigned the identifier 1neu in the PDB and d1neu_ in SCOP, is a member of SCOP's immunoglobulin-like fold and has a total of 11 beta strands as defined by DSSP. Using default parameters and a target database consisting of a subset of SCOP domains created such that no two targets have >25% sequence similarity, FoldMiner iteratively refines both the motif definition and the search results by using the motif to update the list of statistically significant targets and then using this refined list of targets

to update the motif definition. This refinement process, which does not require additional structural superpositions to be performed, executes rapidly. Secondary structure elements whose positions among Ineu and its structural neighbors tend not to vary are considered to be conserved, while SSEs that are deleted in some of these structures or whose orientations tend to vary receive lower conservation values. The motif is defined probabilistically by the structural conservations of Ineu's SSEs.

The FoldMiner server presents the structural similarity search results for Ineu in three browser windows: one contains the search results, one contains the immunoglobulin motif as manifested in Ineu and options for manual refinement and the third is a small control panel that keeps track of the session history and allows the user to navigate smoothly among multiple sets of results. The motif window uses the Chime plugin (<http://www.mdlchime.com/>) to display the query structure with its secondary structure elements colored according to their calculated conservations. The more strongly conserved a secondary structure element is, the brighter it will appear, allowing the user to identify the conserved portions of a fold by eye.

In the case of Ineu, six strands are brightly colored and are clearly part of the conserved Greek key fold (Figure 1). Three strands are darkly colored and are insertions to the core fold. In this particular structure, the first strand of the core fold is split into two strands. Since some targets align to the first of these two strands and others align to the second, neither one is as strongly conserved as the rest of the core fold. The actual conservation values, which lie on the interval [0, 1], are displayed in a table as well. The user may label SSEs in the Chime display with these values and may also press a button next to each conservation value in a text-based description of the motif in order to cause the corresponding secondary structure element in the Chime display to blink, allowing for visual correlation of the text-based results with the motif visualization.

The table in the search results window (not shown) indicates that Ineu has a total of 63 structural neighbors in the target database (excluding the query itself). Each entry in the table gives the target's SCOP identifier and a link to its page in the SCOP hierarchy, the LOCK 2 alignment score, the *P*-value, the number of SSEs aligned, the number of residues aligned, the RMSD, the header from the corresponding PDB entry and

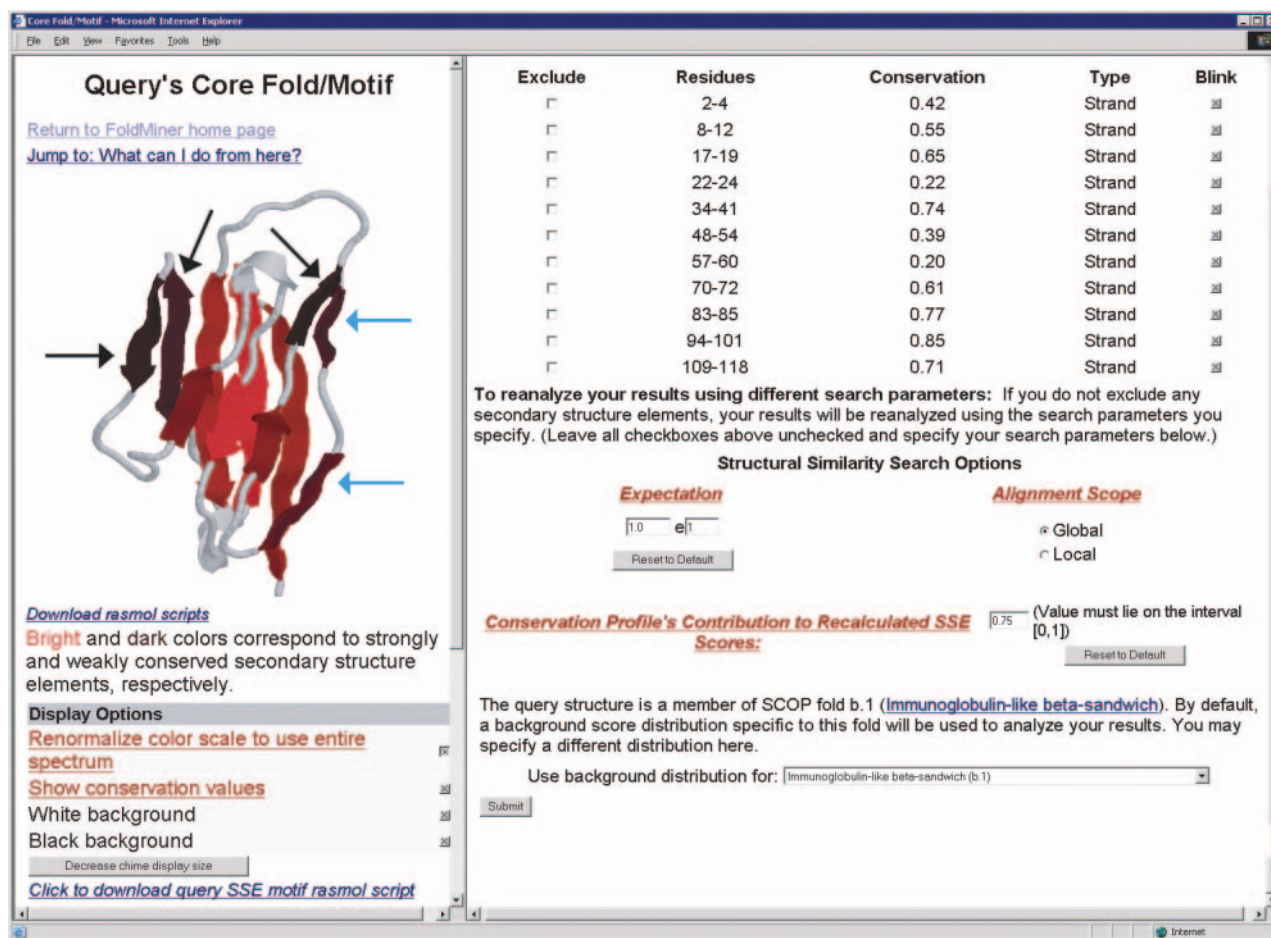


Figure 1. The motif window. The motif window displays the core of SCOP's immunoglobulin-like fold as represented in Ineu, the query structure for this search. The right-hand portion of the window shows both the motif definition in the form of conservation values and options for manual refinement as described in the text. Black arrows indicate the three SSEs that are insertions with respect to the core fold, and blue arrows indicate the two strands detected by DSSP that both correspond to the first strand of the core fold.

three options for viewing alignment results. Results are initially listed in order of decreasing score, but can be sorted by any of the other numeric values, by SCOP fold, or by the PDB header. A second table at the bottom of the page shows the number of target domains belonging to each SCOP fold represented among the results. A checkbox next to each fold in this table highlights the members of that fold in the search results table.

In this case, 55 of the 63 domains are members of the same fold as 1neu, the immunoglobulin-like fold. The remaining eight domains are members of two additional SCOP folds: the cupredoxin-like fold and the diphtheria toxin/transcription factors/cytochrome *f* fold. SCOP annotation indicates that both of these folds have cores consisting of Greek keys with between seven and nine strands and that the diphtheria toxin/transcription factors/cytochrome *f* fold is topologically similar to the immunoglobulin-like fold. Thus, all of the targets in the table appear to be structurally similar to 1neu.

In order to determine whether the iterative refinement process described above improves the search results, the parameter that controls the extent to which the alignment scoring system is adapted to focus on conserved regions of 1neu may be set to zero. Without the motif discovery and iterative refinement phase of FoldMiner, only 42 statistically significant targets are identified, 38 of which are members of the immunoglobulin-like fold. Clearly, the core fold definition identifies regions of the query that are most useful for assessing structural similarity and allows FoldMiner to improve the sensitivity of the search without sacrificing specificity.

There are three ways to view the LOCK 2 alignment for any target listed in the search results table. The text-based result page shows the residue alignment and the alignment of secondary structure elements, and also contains the

transformations used to superimpose the two structures. The secondary structure definitions used in the alignment are given at the bottom of this page. One may click the 'PDB' link to download a PDB file that shows the superposition of the query and target in which the query is chain A and the target is chain B. Finally, the 'Chime' link opens a new window in which Chime is used to display and analyze the superposition.

Sorting by the number of aligned secondary structure elements shows that all targets listed in the results table have at least six secondary structure elements aligned, and examination of some of the top-ranking results reveals that these structures have both the core fold and other regions in common with the query. The FoldMiner server's alignment viewer provides a number of options to improve visualization of these superpositions. Clicking on the 'Chime' link for d1ycsa_, one of the nine structural neighbors of 1neu that are not in the immunoglobulin-like SCOP fold, opens a new window showing a cartoon diagram of the alignment in which 1neu is shown in blue and d1ycsa_ in red.

Four other visualization methods facilitate more detailed analyses of alignments. These options allow one to view aligned secondary structure elements or residues, the secondary structure element alignment scores that FoldMiner uses in the calculation of structural conservations (Figure 2), or the query motif. In the first two cases, the aligned secondary structure elements or aligned residues are shown in blue (query) and red (target), while the rest of the structures are shown in gray. These options more clearly show the topology of the aligned region. The SSE scores option changes the brightness of the aligned secondary structure elements according to their scores to improve visualization of well-aligned regions, while the motif option colors query SSEs and aligned target SSEs according to the query SSEs' conservation values.

Query: d1neu__ Target: d1ycsa_

Structural Alignment and Motif Options

Show me:

- Aligned Secondary Structure Elements
- Aligned Residues
- Query's motif/core fold
- Secondary structure element alignment scores

Residue Selection Options

Select all residues in:

- Query
- Target
- Query and Target
- Strands
- Helices

Select aligned:

- Query Residues
- Target Residues
- Query and Target Residues
- Query secondary structure elements
- Target secondary structure elements
- Query and target secondary structure elements

Enter a chime script (separate commands with carriage returns or semicolons) and press "Execute":

Color Options

Color selected residues:

- Blue Red
- Cyan Magenta
- Green Yellow
- Purple Gray
- Default By 2^o structure

Display Options

Display selection in:

- Cartoons Backbone
- Spacefill Wireframe

Other Options

Labels:

Background:

Black

White

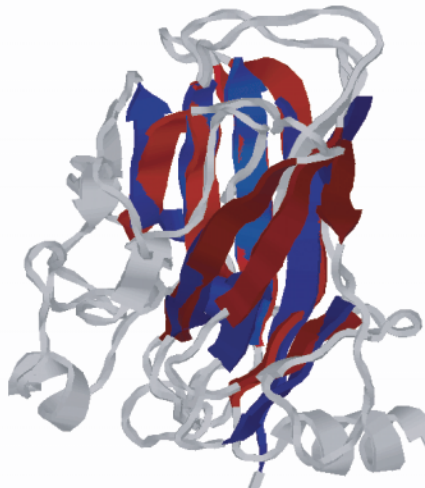


Figure 2. The FoldMiner server alignment viewer. The alignment of the query, 1neu, to a cupredoxin target (SCOP domain d1ycsa_) is colored by SSE alignment scores. The alignment viewer provides a number of customized options for viewing and analyzing superpositions and an interface that allows for the execution of arbitrary sequences of Chime commands.

The latter of these two options shows that much of dlycsa_'s alignment score is derived from alignments to query strands that are part of the core fold; five of the six strongly conserved strands are aligned, only one of the query strands in the insertion region is aligned, and both of the strands comprising the first SSE of the core fold are aligned (not shown).

The remaining controls provide options for selecting certain subsets of residues (e.g. aligned residues or helical residues) in order to change their rendering independently of the rest of the query and target structures using the provided color and display options. We also provide a command line interface for Chime that will execute an arbitrary sequence of commands.

One might wonder whether eliminating poorly conserved query SSEs from analysis would allow for the identification of additional structural neighbors by removing any advantage or disadvantage a target receives in the scoring process due to the quality of the alignment in regions that are not part of the core fold. This process is conceptually similar to searching the target database using the core fold as a query. The three insertions to the immunoglobulin-core fold seen in 1neu can be excluded from analysis in the portion of the motif window devoted to manual refinement of the results. All FoldMiner parameters may be adjusted as well, and a new set of results appears in the motif and search windows. The control panel is updated to reflect the second search; the user switches back and forth between the two results sets using buttons on the control panel.

Excluding 1neu's three SSEs that are insertions with respect to the immunoglobulin core fold yields an additional 44 targets, 29 of which are members of the immunoglobulin-like SCOP fold (results not shown). Interestingly, as dlycsa_ derived some of its score from an inserted region, it is no longer among the results. All other new results are members of folds that SCOP describes as Greek keys or beta sandwiches. By clicking the 'Chime' link for the highest-ranking non-immunoglobulin and coloring secondary structure elements first according to the query SSEs' conservation values and then according to their alignment scores, it is readily apparent that this target aligns well to all six strongly conserved strands of 1neu's core fold and to the more highly conserved of the two 1neu strands that both correspond to the first SSE of the core fold (not shown). This target is a cupredoxin with the SCOP identifier d1kzqa2 and ranks 16th in the list of statistically significant alignments.

If there are no statistically significant global alignments, as is the case for *Escherichia coli*'s type I DNA topoisomerase (SCOP domain d1d6ma_ or chain A of PDB structure 1d6m), FoldMiner can discover a high-quality local motif that is abundant in the target database. We have provided preliminary evaluation of the statistical significance of local alignments, and at this point in time we recommend that users focus their attention on higher-ranking results. In this example, many winged helix DNA binding domains appear at the top of the list. This DNA binding domain is located in close proximity to the region of the topoisomerase believed to be the DNA binding groove (31), suggesting that this is a functionally relevant result (figure 5 of reference 16).

Results remain available on the FoldMiner server for at least three days after the search finishes. All alignment results, a single text file containing the same information shown in the search results table, and a RasMol script (32) that produces the same view of the query as is shown in the motif window can be

downloaded as well. Links to relevant help topics are provided on each page the user views.

AVAILABILITY

FoldMiner is available on the Internet at <http://foldminer.stanford.edu/> and LOCK 2 is available at <http://lock2.stanford.edu/>. Source code is available royalty-free to academic and not-for-profit institutions at <http://motif.stanford.edu/software/> and from Stanford's Office of Technology Transfer (<http://otl.stanford.edu/>) for for-profit institutions.

ACKNOWLEDGEMENTS

We thank Amit Singh for his contribution to the design of the FoldMiner homepage. This work was supported by NIGMS grant number GM63495.

REFERENCES

- 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.
- Feng, Z.K. and Sippl, M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold Des.*, **1**, 123–132.
- Godzik, A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
- Grishin, N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
- Tang, C.L., Xie, L., Koh, J.Y., Posy, S., Alexov, E. and Honig, B. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.*, **334**, 1043–1062.
- Goldsmith-Fischman, S. and Honig, B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci.*, **12**, 1813–1821.
- Yang, A.S. and Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.*, **301**, 691–711.
- Byströf, C. and Shao, Y. (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics*, **18** (Suppl. 1), S54–S61.
- Madej, T., Gibrat, J.F. and Bryant, S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
- Panchenko, A., Marchler-Bauer, A. and Bryant, S.H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, **37** (Suppl. 3), 133–140.
- Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N. and Orengo, C.A. (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.*, **7** (Suppl.), 991–994.
- Orengo, C.A., Todd, A.E. and Thornton, J.M. (1999) From protein structure to function. *Curr. Opin. Struct. Biol.*, **9**, 374–382.
- Huang, C.C., Novak, W.R., Babbitt, P.C., Jewett, A.I., Ferrin, T.E. and Klein, T.E. (2000) Integrated tools for structural and sequence alignment and analysis. *Pac. Symp. Biocomput.*, 230–241.
- Balaji, S. and Srinivasan, N. (2001) Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng.*, **14**, 219–226.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
- Shapiro, J. and Brutlag, D. (2004) FoldMiner: structural motif discovery using an improved superposition algorithm. *Protein Sci.*, **13**, 278–294.
- Holm, L. and Sander, C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.

18. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
19. Kleywegt,G.J. and Jones,T.A. (1997) Detecting folding motifs and similarities in protein structures. *Methods Enzymol.*, **277**, 525–545.
20. 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.
21. Martin,A.C. (2000) The ups and downs of protein topology; rapid comparison of protein structure. *Protein Eng.*, **13**, 829–837.
22. Gilbert,D., Westhead,D., Viksna,J. and Thornton,J. (2001) A computer system to perform structure comparison using TOPS representations of protein structure. *Comput. Chem.*, **26**, 23–30.
23. Carugo,O. and Pongor,S. (2002) Protein fold similarity estimated by a probabilistic approach based on C([alpha])-C([alpha]) distance comparison. *J. Mol. Biol.*, **315**, 887–898.
24. Harrison,A., Pearl,F., Mott,R., Thornton,J. and Orengo,C. (2002) Quantifying the similarities within fold space. *J. Mol. Biol.*, **323**, 909–926.
25. Kawabata,T. (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res.*, **31**, 3367–3369.
26. Krissinel,E. and Henrick,K. (2003) Protein structure comparison in 3D based on secondary structure matching (SSM) followed by C_α alignment, scored by a new structural similarity function. In Kungl,A.J. and Kungl,P.J. (eds), *Proceedings of the 5th International Conference on Molecular Structural Biology*, Vienna, Austria, p. 88.
27. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
28. 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.
29. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.
30. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**(Database issue), D189–D192.
31. Mondragon,A. and DiGate,R. (1999) The structure of *Escherichia coli* DNA topoisomerase III. *Struct. Fold Des.*, **7**, 1373–1383.
32. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.