Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W173–W176***
doi:10.1093/nar/gkl113

ArchPRED: a template based loop structure prediction server

Narcis Fernandez-Fuentes, Jun Zhai and András Fiser*

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

Received February 7, 2006; Revised March 1, 2006; Accepted March 9, 2006

ABSTRACT

ArchPRED server (http://www.fiserlab.org/servers/archpred) implements a novel fragment-search based method for predicting loop conformations. The inputs to the server are the atomic coordinates of the query protein and the position of the loop. The algorithm selects candidate loop fragments from a regularly updated loop library (*Search Space*) by matching the length, the types of bracing secondary structures of the query and by satisfying the geometrical restraints imposed by the stem residues. Subsequently, candidate loops are inserted in the query protein framework where their side chains are rebuilt and their fit is assessed by the root mean square deviation (r.m.s.d.) of stem regions and by the number of rigid body clashes with the environment. In the final step remaining candidate loops are ranked by a *Z*-score that combines information on sequence similarity and fit of predicted and observed [φ/ψ] main chain dihedral angle propensities. The final loop conformation is built in the protein structure and annealed in the environment using conjugate gradient minimization. The prediction method was benchmarked on artificially prepared search datasets where all trivial sequence similarities on the SCOP superfamily level were removed. Under these conditions it was possible to predict loops of length 4, 8 and 12 with coverage of 98, 78 and 28% with at least of 0.22, 1.38 and 2.47 of r.m.s.d. accuracy, respectively. In a head to head comparison on loops extracted from freshly deposited new protein folds the current method outperformed in a ∼5:1 ratio an earlier developed database search method.

INTRODUCTION

Functional characterization of a protein sequence is often facilitated by its 3D structure. In the absence of an experimentally determined structure, comparative modeling and threading may be applicable to provide a useful 3D model and fill the growing gap between sequence and structure spaces (1). The accuracy of comparative models can be very high in the core of the model, corresponding to low resolution experimental solution structures, especially if many high resolution structures are available as templates sharing the same general fold. However the loop regions of these structures are often different. For these unique structural segments that are often found on the surface of the proteins, comparative modeling techniques cannot generally be applied. Loop segments in the target may be missing in the template or structurally divergent, resulting in inaccurate parts in the model. Meanwhile loops represent an important part of the protein structure and often determine the functional specificity of a given protein framework, contributing to active and binding sites (2). Functional differences among the members of the same protein family are usually a consequence of the structural differences on loops. Thus, the accuracy of loop modeling is a major factor in determining the usefulness of models in studying interactions between the protein and its ligands and in analyzing active and binding sites. Loop modeling also plays an important role in completing poorly refined experimentally determined three dimensional models. The impact of loop modeling is significant. Currently, ∼60% of all protein sequences can have at least one domain modeled on a related, known protein structure (3). At least two-thirds of the comparative modeling cases are based on <40% sequence identity between the target and the templates, and thus generally require loop modeling.

METHOD

Details of the method and its benchmarking have been described in a recent publication (N. Fernandez-Fuentes,

*To whom correspondence should be addressed. Tel: +1 718 430 3233; Fax: +1 718 430 856; Email: andras@fiserlab.org

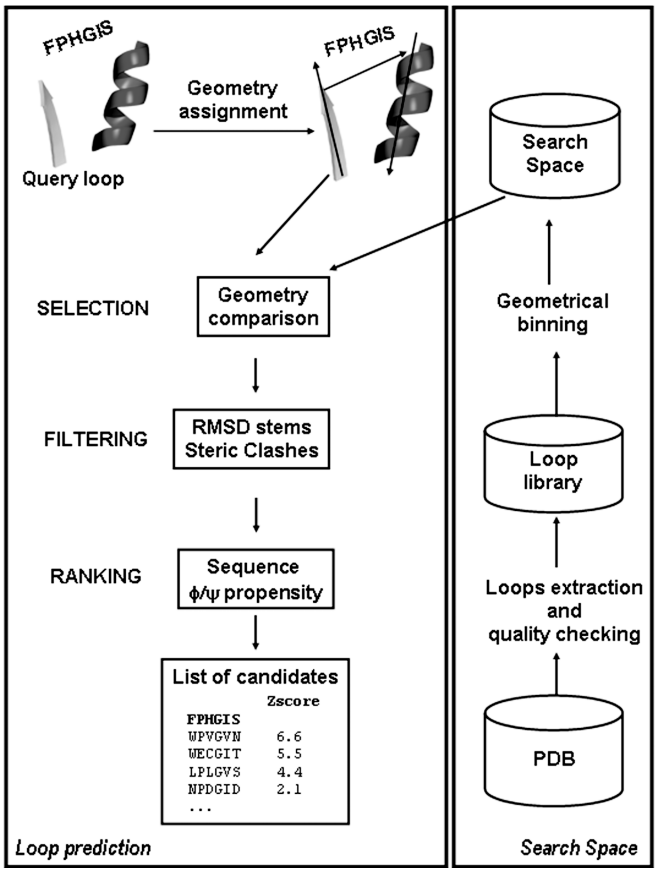

Figure 1. Schematic representation of the two components of ArchPRED: the loop database (*Search Space*) and the *loop prediction* algorithm.

B. Oliva and A. Fiser, manuscript submitted) (Figure 1). Briefly, the method relies on an exhaustive conformational fragment library that is organized in a hierarchical and multidimensional database, called *Search Space*. The *Search Space* is a multidimensional library of loops of known structures organized into a three level hierarchy: (i) at the top, loops are identified according to the type of the bracing secondary structures: αα loops βα loops, αβ loops and ββ loops; (ii) at the next level, loops are grouped according to their length, and finally (iii) loops are grouped according to the geometry of the bracing secondary structures. This geometry is defined by a distance, *D*, and three angles, a hoist (δ), a packing (θ) and a meridian (ρ) (4). The *Search Space* is regularly updated by analyzing all the available structures in Protein Databank (PDB) (5) and extracting the loop segments [defining loops as the region that connect two secondary structures, beta strands or helices as defined by DSSP (6)]. Only those loops that satisfy several quality rules (i.e. crystal resolution, no missing main chain atoms and, low B-factors) are incorporated to the *Search Space* that currently contains about 240 000 fragments.

The prediction algorithm includes three steps (i) *Selection*, (ii) *Filtering* and (iii) *Ranking*. During *Selection* step the *Search Space* is queried by the length of the loop, the type of secondary structures that span the query loop and by the geometry of the motif. If this information is missing (i.e. poorly defined secondary structures) the *Search Space* can be queried by the distance of the ending points (i.e. stem residues). In the *Filtering* step the algorithm discards unfavorable candidates by assessing the fit of stem regions and by steric fitting in the new protein framework. Finally, in the *Ranking* step the remaining set of candidate loops is ranked by a composite *Z*-score that combines a sequence similarity score (7) and [ϕ/φ] main chain dihedral angle propensities (8).

Performance of the method

We tested the performance of ArchPRED by (i) benchmarking it against known structures (ii) directly comparing it with an earlier developed, publicly available fragment search based method (9).

The prediction method was tested on artificially prepared search datasets where all trivial sequence similarities on the SCOP superfamily level were removed. Under these conditions it is possible to predict loops of length 4, 8 and 12 with coverage of 98, 78 and 28% with at least of 0.22, 1.38 and 2.47 Å of root mean square deviation (r.m.s.d.) accuracy, respectively. We also performed a head-to-head comparison of performances between the current ArchPRED and the FREAD methods (9). To avoid a trivial exercise we used only new structural releases from PDB (5), which could not yet enter the classification schemes of either methods and we tracked these new PDB structures for two weeks. Among the new structures we identified new folds by removing all proteins with sequence (>40% sequence identity) and structural similarity [DALI (10) *Z*-score >3] to any known PDB structures. From the remaining six novel fold structures we located 35 loop regions and submitted the sequences of these fragments to our method and to the FREAD server. The predicted loops were superposed with the experimental solution and r.m.s.d. values obtained. The current method, ArchPRED not only provides a higher coverage (it predicted all segments, while FREAD did not return answer for four cases) but on average it returned more accurate predictions in 23 out of 28 cases, while in three cases they returned identical solutions.

DESIGN, IMPLEMENTATION AND USE

ArchPRED is implemented on an Apache server running Fedora core 3 operating system. The server is interfaced with CGI Perl and javascript coded web interface. The loop database (*Search Space*) is stored in a MySQL relational database. DBI–DBD (DataBase Interface–DataBase Driver) and related modules are used for communication between the scripts and the MySQL database. Results are either displayed in html format or sent by email to the user as a hyperlink. Users need to use a visualization program of their choice to display the atomic coordinates of the predicted loop.

Submitting a query

Users need to provide and define data on the submission web page (Figure 2):

- Atomic coordinates: users have to upload the atomic coordinates of the protein structure where the missing loop is going to be predicted. If the structure contains more than

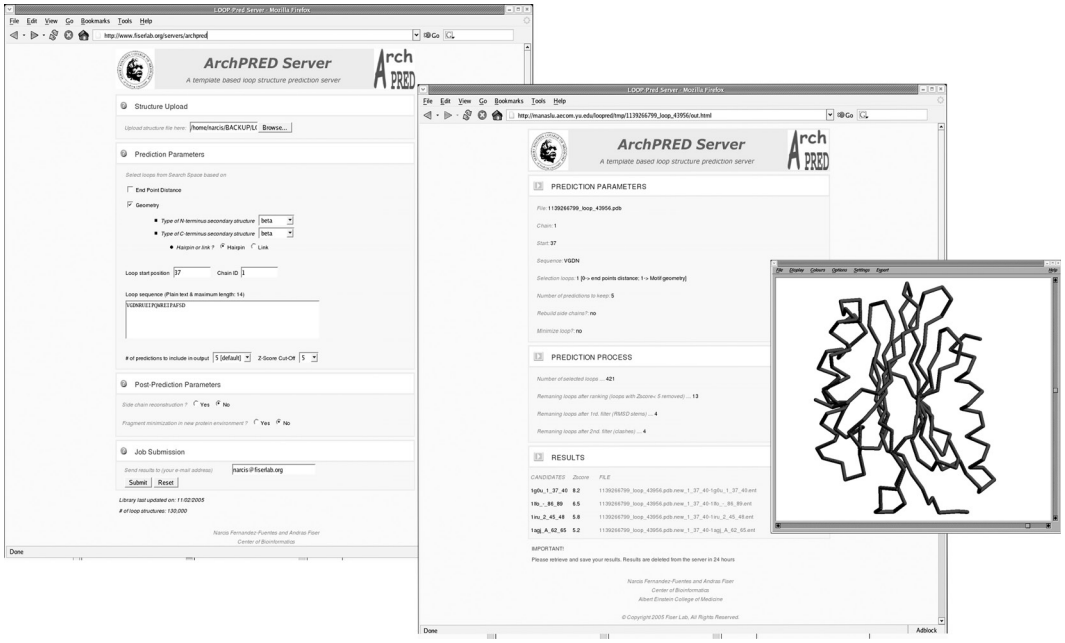

Figure 2. Screenshot of the submission and results web pages. All parameters have a links to a help web page in order to provide further information.

one chain, only the chain that includes the missing loop will be considered. The format of the atomic coordinates must be the default PDB and must contain at least all main chain atoms for all residues.

- Prediction parameters: users have to choose prediction parameters and define the location of the missing loop. Users have to choose how to query the *Search Space* (selection of candidate loops); whether using the geometry of the motif or the distance of the end points only. If geometry is selected users are prompted to define the type of the motif (i.e. type of flanking secondary structures; in case of β−β motif, it must be refined if the motif is a β-hairpin or a β-link). If end point distance is selected, all candidate loops with $(|\text{Distance}_{C\alpha,\text{ stems query}} - \text{Distance}_{C\alpha,\text{ stems candidate}}|) \leqslant 1$ Å. are selected. Users have to define the start position and sequence (in one letter amino acid codes) of the missing loop. Although the loop is missing from the coordinate file, the numbering must be consistent with the missing loop. If not null, users must define chain identification where the missing loop is located. Users might want to keep more than one possible prediction; by default the server returns the top five predictions. Users can also select an appropriate *Z*-score cut-off.
- Post-prediction parameters: For each predicted loop only the coordinate for the main chain atoms are provided. By default the predicted loop is fitted in the new protein environment without any optimization. Users can request refinements, such as side chain construction and energy minimization. Side chain building is done by SCWRL3 program (11). For energy minimization a short conjugate gradient minimization [using minimization procedures embedded in MODELLER package (12)] is applied to anneal the stem residues but preserving the overall conformation of the loop structure.

Retrieving results

When the prediction process is finished an output web page is loaded (Figure 2). Meanwhile, an email is sent to the user with the hyperlink to the output web page. The output page contains a brief report about the prediction process, such as the number of selected candidate loops from the *Search Space*; number of discarded candidate loops through the *Filtering* and *Ranking* steps and the like; and a list of predictions ranked by *Z*-score. Each prediction has its own link to download the corresponding coordinate file of the predicted loop. The new coordinate file is provided in PDB format and can be viewed with default visualization programs.

In case the method does not locate any suitable candidate loops a warning message is shown in the output web page. A full list of possible error messages are listed below:

(i) *Unable to connect*: For some reason (temporary network failure, machine shutdown and the like) the server can not connect to the database. Please, try again later.
(ii) *Something wrong with stem residues*: User defined stem residues do not exist in the coordinates file; in order to predict a loop at least the coordinates for five residues of the stem regions must be known.
(iii) *No selected loops that fulfill your query,* (*geometry*): There is not a single loop in *Search Space* that has the same geometrical definition as the query loop. Try selecting loops by end-point distance only.
(iv) *No selected loops that fulfill your query,* (*end-points*): There is not a single loop in *Search Space* with end-point distance ±1 Å similar to the query loop.
(v) *No suitable loops after r.m.s.d. stem filter*: All candidate loops were discarded because the r.m.s.d. of stems is larger than the applied r.m.s.d. stems cut-off.

(vi) *No suitable loops after filtering by clashes*: After inserting the template loop in the protein environment, all template loops have steric impediments.
(vii) *No suitable loops after Z-score ranking*: All templates loops have a *Z*-score smaller than the selected *Z*-score cut-off.

The prediction process is registered in a log file that users can examine to understand what the problem was during the prediction process. Also, users can contact the authors via email to loopred@fiserlab.org for further information.

DISCUSSION

A webserver for loop structure prediction is described above. The prediction method is fast; all predictions are done in real time, so users can get the results typically within one minute. For additional convenience, for each prediction users receive an email with a hyperlink to a web page where results are shown.

The webserver provides not only a list with the most suitable fragments but their fitting in the query structure. Thus, the result of the prediction is a coordinate file that contains not only the coordinate of the missing loop but its fitting and orientation in the protein structure. Furthermore, if requested users can add the side-chain atoms to the predicted loop and perform an energy minimization in the context of the new protein framework.

ACKNOWLEDGEMENTS

The authors acknowledge all Fiser Lab members for their insightful comments on the work. Financial support was provided by NIH GM62519-04 and the Seaver Foundation. Funding to pay the Open Access publication charges for this article was provided by NIH GM62519-04.

Conflict of interest statement. None declared.

REFERENCES

1. Fiser,A. (2004) Protein structure modeling in the proteomics era. *Expert Rev. Proteomics*, **1**, 97–110.
2. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113.
3. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217.
4. Oliva,B., Bates,P.A., Querol,E., Aviles,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814.
5. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
6. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
7. Kolaskar,A.S. and Kulkarni-Kale,U. (1992) Sequence alignment approach to pick up conformationally similar protein fragments. *J. Mol. Biol.*, **223**, 1053–1061.
8. Shortle,D. (2002) Composites of local structure propensities: evidence for local encoding of long-range structure. *Protein Sci.*, **11**, 18–26.
9. Deane,C.M. and Blundell,T.L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.*, **10**, 599.
10. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123.
11. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.,Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001.
12. Sali,A. (1995) Comparative protein modeling by satisfaction of spatial restraints. *Mol. Med. Today*, **1**, 270.