Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W85–W88***
doi:10.1093/nar/gki353

Fragment Finder: a web-based software to identify similar three-dimensional structural motif

P. Ananthalakshmi[1], Ch. Kiran Kumar[1], M. Jeyasimhan[1], K. Sumathi[1] and K. Sekar[1,2,*]

[1]Bioinformatics Centre and [2]Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India

Received November 19, 2004; Revised and Accepted December 30, 2004

ABSTRACT

FF (Fragment Finder) is a web-based interactive search engine developed to retrieve the user-desired similar 3D structural fragments from the selected subset of 25 or 90% non-homologous protein chains. The search is based on the comparison of the main chain backbone conformational angles (ϕ and φ). Additionally, the queried motifs can be superimposed to find out how similar the structural fragments are, so that the information can be effectively used in molecular modeling. The engine has facilities to view the resultant superposed or individual 3D structure(s) on the client machine. The proposed web server is made freely accessible at the following URL: http://cluster.physics.iisc.ernet.in/ff/ or http://144.16.71.148/ff/.

INTRODUCTION

The general problem in protein designing and modeling is to investigate the relationship between sequence and its functional specificities. The key to this problem lies in addressing and understanding how the amino acid sequence determines its corresponding 3D structure. The approach on the use of sequence homologies showed that short peptides having similar sequences from different protein chains exhibit different 3D structures (1). Further to add, similar conclusions were drawn on hexa- and hepta- peptides (2). Later on, Argos (3) showed that sequence identical penta-peptide pairs in unrelated protein structures maintain same structural conformation approximately 20% of the time. It is known that substructures or short 3D structural fragments of a protein molecule are closely related to the biological function of the protein molecule (4). Hence, it is highly indispensable to retrieve a reasonable 3D structural motif from the solved 3D structures housed in the PDB (5,6).

Since the introduction of the 3D data archive (PDB), there has been a tremendous growth in the number of available protein and nucleic acid structure entries. This is further augmented due to the recent advances in Crystallography, such as high-intensity synchrotron beam lines and significant methodological progress. In effect, ∼29 000 protein and nucleic acid structures are presently available in this entity. Distilling useful information from the available 3D structures and its related amino acid sequences is highly beneficial for the scientific community in the post-genomics era. Hence, an analysis requires an efficient search engine equipped with the curated knowledge base to cull the useful information from the massive data archival. The proposed software facilitates the user to fetch the user interested, exact or similar 3D structural fragment (using the main chain conformation angles) from the non-homologous (25 and 90%) protein chains (7).

MATERIALS AND METHODS

The backbone conformational angles or Ramachandran angles (ϕ and φ) of all the non-homologous protein chains (25 and 90%) are computed and housed in a locally maintained database using MySQL, a Relational Data Base Management System (RDBMS). RDBMS package allows more complex queries and addresses efficient maintenance issues. The structures solved using X-ray crystallography and NMR spectroscopy are considered in the present study. For structures solved using NMR spectroscopy, only the first model is used to compute the required backbone torsion angles. In-house developed PERL scripts are deployed to calculate the conformation angles and to store the corresponding values directly in MySQL database without any human intervention. Thus, the present knowledge base contains the main chain conformation angles of 2216 and 6254 protein chains from 25 and 90% non-homologous datasets, respectively. The 3D structural superposition programs, such as STAMP (8) and ProFit (http://www.bioinf.org.uk/software/profit/) are deployed for superposition. The user-friendly molecular visualization tool RASMOL (9) is interfaced with the search engine to view the individual or superposed fragments in the client machine.

*To whom correspondence should be addressed. Tel: +91 080 23601409/22923059; Fax: +91 080 23600085/23600551; Email: sekar@physics.iisc.ernet.in

In the output display, options are provided for the users to store the 3D atomic coordinates of the resultant fragments in the local disk of the client machine. The database related to non-homologous sequences will be updated as and when it is available (Hobohm and Sander anonymous FTP server, Heidelberg, Germany).

FEATURES

The primary goal of this project is to maintain a high quality knowledge base and an efficient search engine to get the user interested 3D structural fragments present in the non-homologous protein chains. We have developed a user-friendly interactive web interface to access the information in the MySQL database while querying the exact or similar structural fragments. As of December 2003 release, a total of 2216 polypeptide chains from 2105 protein structures are available under 25% non-homologous subset and the corresponding numbers in the 90% subset are 6254 and 5602, respectively. Users can select the input fragment of interest using three options. For the first option, user needs to provide the PDB-ID and the complete chain information will be displayed on the resultant output page so that users can select the fragment of interest. Users need to upload the 3D atomic coordinates of the fragment (PDB file format) for the second option. For the third option, the main chain conformation angles (ϕ and φ) of the interested structural fragment are required and the user needs to provide the same.

Once the query fragment is chosen, the user has several fine-tuning parameters (described in the subsequent sections) to control the quality of the output fragments.

(i) The users can opt to pickup the structural fragments with identical (as given in the input) or similar residues or any residue pattern as long as the backbone conformation angles match.
(ii) Users can select a particular experiment (X-ray diffraction or NMR technique) method so that the appropriate information in the knowledge base will be used during search process.
(iii) In addition to the experiment method, users can also select the required non-homologous (25 and 90%) protein chains on which the search has to be performed.
(iv) There is a provision to select the tolerance level (by default 5°) on the conformation angles (on either side of ϕ and φ). During validation of this software, we experienced that 5° tolerance level is reasonable for α-helical fragments. However, >10–15° is required in the case of β-strands. Users are therefore advised to use the appropriate values to get the hidden structural fragments.
(v) If the queried structural motif is not present in the knowledge base, facilities are provided in the search engine for the users to truncate the residues (by default one at a time) and repeat the search from the N- or C-terminal end. This option facilitates the user to have freedom in selecting the required information.
(vi) Finally, in the output display, user has the freedom to see either the detailed output (example not shown) or simple output (see case study for details) of the queried motif and the resultant motifs from the knowledge base for better understanding of the agreement.

After completing the above fine-tuning options, the output page displays the structural fragments that match the input fragment. Furthermore, options are there to superpose the fragments displayed in the output page. To facilitate this, user needs to select the fragments by clicking the radio button provided against them. To avoid delay in displaying the superposition results, the program is coded in such a way that a maximum of 20 fragments can be superposed at any given time. For superposition, two programs, STAMP and ProFit, are interfaced with the search engine. The program STAMP looks for overall topological similarity for superposing the structures, whereas ProFit is based on least squares fit of proteins. Thus, the user has the option to choose a suitable program for superposition. The output shows details like root-mean-square (rms) deviations between the fixed and the superposed fragments. Most importantly, the users can view the superposed fragments using the molecular visualization tool RASMOL. Additionally, the users can save the atomic coordinates of the superposed fragments in the local machine for further analysis. The users can get parameters like rotation matrix and translation vector applied to the individual mobile fragments used in the superposition by clicking the option 'Detailed report' (see Figure 2 for details). In addition, it shows the sequence identity and stamp score (only if the program STAMP is used for superposition). The proposed search engine allows the users to view the structural fragments, which has low rms deviations and the deviation of the individual Cα atoms between the fixed and the mobile fragments.

CASE STUDY

A sample output of a typical search using a part of a helix containing eight residues [PDB-ID: 1UNE (10), residues 2–9] is shown in Figure 1. The top panel of Figure 1 shows the main chain conformation angles computed using the input structural fragment stated above. The bottom panel shows the matched hits available in 25% non-redundant protein chains. The output page displays the simple output of 13 fragments match with the input structural motif (using default values provided in the search engine) from various protein chains available in 25% non-redundant data set. The top left Rasmol panel of Figure 2 shows the nature and the location of the input fragment (green colored ribbon) with respect to the entire protein molecule (backbone trace). The right bottom panel shows the superposed structural fragments of all 13 hits listed in Figure 1. The first and the second columns of the adjacent panel show the sequence identity (between the fixed and the corresponding mobile molecule) and the STAMP score, respectively. The root mean square deviations of various fragments with respect to the fixed molecule are listed in the third column. The last column shows the coloring scheme adopted in the Rasmol display.

The search engine is written in Perl. To meet the increasing demand and to drastically improve the efficiency, the search engine is designed for a high-end processor, Intel based Solaris operating environment (a 3.06 GHz Pentium IV processor with 1 GB of main memory). The software has been validated and the response time is very fast. However, the response time varies depending upon the network speed. The front end of this

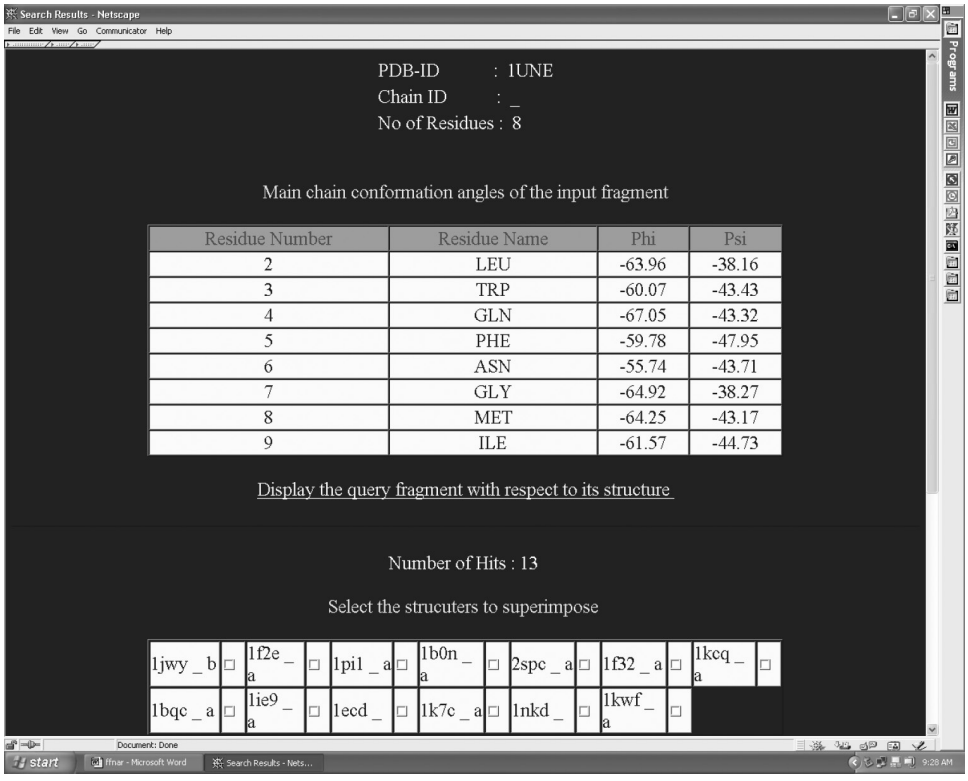

Figure 1. The top panel of the output page displays the main-chain conformation angles for the input fragment (residues 2–9 from the PDB-ID code 1UNE). The bottom panel shows the output of the matched structural fragments (only PDB-IDs) found in 25% non-homologous protein chains.

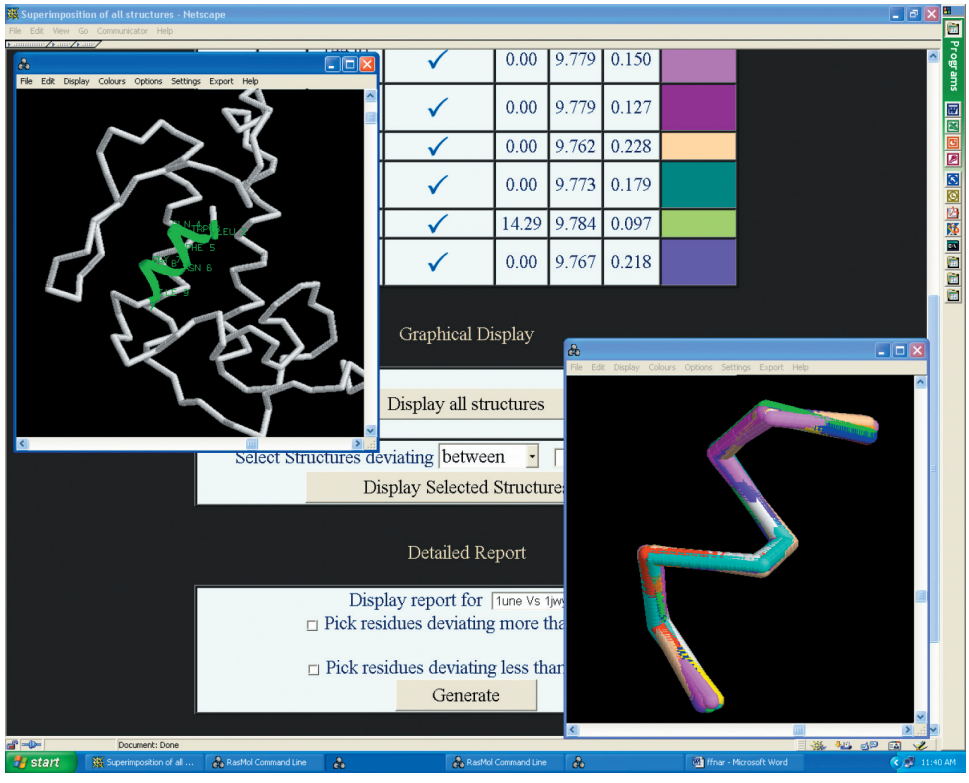

Figure 2. The output page depicts the superposition (top middle panel) of the 13 hits listed at the bottom of Figure 1. The left panel displays the location of the input fragment (ribbon colored green) with respect to the entire molecule (backbone trace). The right bottom panel displays the superposition (using the program STAMP) of all the 13 hits found by the search engine. This panel can be invoked by clicking the option 'Display all the structures'.

tool is designed in HTML and JavaScript. The search engine is very user friendly and can be accessed using Windows 95/98/2000, Windows NT server, Linux and Silicon Graphics (SGI) platforms with the NETSCAPE (version 4.7) browser. The users need to interface the graphics freeware, RASMOL when they use it for the first time (see the help: how to configure RASMOL).

CITATION OF FRAGMENT FINDER (FF)

The users of FF are requested to cite this article and the URL of the search engine in their scientific reports and investigations. General comments and suggestions for additional options are welcome and should be addressed to Dr K. Sekar at sekar@physics.iisc.ernet.in.

CONCLUSIONS

The described search engine is best optimized to identify exact or similar structural fragments from the non-homologous protein chains to better support researches that investigate the relationship between the amino acid sequences and the 3D structures. Hence, we strongly believe that the software is very useful especially for those practicing in the area of modern bioinformatics or computational biology.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the use of the Bioinformatics Centre; the interactive graphics based molecular modeling facility and the Supercomputer Education and Research Centre. The development of this software is fully supported by an individual research grant to Dr K. Sekar from the Department of Biotechnology (Ministry of Science and Technology), Government of India. Finally, the authors thank Dr Geoff Barton and Dr Andrew Martin for permitting to use their superposition programs STAMP and ProFit, respectively, in the proposed search engine. The authors thank Ms P. Mridula for critical reading of the manuscript. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Kabsch,W. and Sander,C. (1984) On the use of sequence homologies to predict protein structure: identical penta-peptides can have completely different conformations. *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.
2. Wilson,I.A., Haft,D.H., Getzoff,E.D., Tainer,J.A., Lerner,R.A. and Brenner,S. (1985) Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition. *Proc. Natl Acad. Sci. USA*, **82**, 5255–5259.
3. Argos,P. (1987) Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. *J. Mol. Biol*., **197**, 331–348.
4. Guo,T., Hua,S., Ji,X. and Sun,Z. (2004) DBsubLoc: database of protein sub cellular localization. *Nucleic Acids Res*., **32**, D122–D124.
5. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M.J. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535–542.
6. 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.
7. Hobohm,U. and Sander,C. (1994) Enlarged representative set protein structures. *Protein Sci*., **3**, 522–524.
8. Russell,R.B. and Barton,G.J. (1992) STAMP: multiple protein sequence alignment from tertiary structure comparison. *Proteins*, **14**, 309–323.
9. Sayle,R.A. and Milner-Whilte,E.J. (1995) RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374–382.
10. Sekar,K. and Sundaralingam,M. (1999) High resolution refinement of the orthorhombic form of bovine pancreatic phospholipase A_2. *Acta Crystallogr. D Biol. Crystallogr*., **D55**, 46–50.