Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W499–W502***
doi:10.1093/nar/gkm367

FRalanyzer: a tool for functional analysis of fold-recognition sequence–structure alignments

Harpreet Kaur Saini* and Daniel Fischer

Computer Science and Engineering Department, 201 Bell Hall University at Buffalo, Buffalo, NY 14260, USA

Received January 4, 2007; Revised and Accepted April 26, 2007

ABSTRACT

We describe FRalanyzer (Fold Recognition alignment analyzer), a new web tool to visually inspect sequence–structure alignments in order to predict functionally important residues in a query sequence of unknown function. This tool is aimed at helping to infer functional relationships between a query sequence and a template structure, and is particularly useful in analyzing fold recognition (FR) results. Because similar folds do not necessarily share the same function, it is not always straightforward to infer a function from an FR result alone. Manual inspection of the FR sequence-structure alignment is often required in order to search for conservation of functionally important residues. FRalanyzer automates parts of this time-consuming process. FRalanyzer takes as input a sequence–structure alignment, automatically searches annotated databases, displays functionally significant residues and highlights the functionally important positions that are identical in the alignment. FRalanyzer can also be used with sequence-structure alignments obtained by other methods, and with structure–structure alignments obtained from structural comparison of newly determined 3D-structures of unknown function. Fralanyzer is available at http://fralanyzer.cse.buffalo.edu/.

INTRODUCTION

As the genome sequencing and metagenomics projects are continuing to deliver huge numbers of new protein sequences, the percentage of sequences characterized as ‘unknown function’ is growing rapidly (1,2). In addition, as a result of Structural Genomics, there is also an increasing number of proteins with known structure but with no functional information (3). Functional characterization of a new protein can be based on sequence similarity to a family of well-characterized proteins (4,5). In most cases, when the sequence similarity between the query protein and a well-characterized protein sequence is high, an automatic transfer of function can be made. However, when the similarity is lower, this automatic ‘annotation’ can be misleading and other methods are needed. One method is based on template searching, where the query sequence is searched for sequence motifs, patterns or fingerprints as defined in databases such as Prosite (6) and InterPro (7). Other methods attempt to transfer Gene Ontology terms (8–10) or EC terms (11) to a new sequence. However, such a transfer is only reliable when there is a clear match to a homolog of known function. Although 3D-structural information is often very valuable in understanding protein function, there is an increasing number of proteins with known structures but with unknown functions. Recently, new methods for functional inference for such 3D-structures have been developed (12). FR methods lie in the midpoint between sequence-homology and 3D-structural similarity approaches. In the lack of significant sequence similarity and of an experimental 3D-structure, FR methods are often able to identify compatible 3D-templates with relatively accurate sequence–structure alignments (13). Because the compatible template is often characterized functionally, a FR result can be a rich source of functional information (14,15). Functional transfer from a template to the query sequence can be attempted using the functionally significant residues in the template that are conserved in the FR sequence–structure alignment.

METHOD

FRalanyzer (Fold Recognition alignment analyzer) is a web tool, aimed at visualizing and annotating sequence–structure alignments. Functionally important residues are displayed and highlighted if they are conserved in the alignment. Functional residues in a query sequence of unknown function are identified based on the conservation of functionally significant residues in sequence–structure and structure–structure alignments. Thus, the input to FRalanyzer is an alignment, obtained independently (i.e. not a part of FRalanyzer). To facilitate the human interface and to avoid the need of cutting and

*To whom correspondence should be addressed. Tel: 716 645 3180 (Ext. 163); Fax: 716 645 3464; Email: hksaini@cse.buffalo.edu

pasting of alignments, FRalanyzer can directly retrieve the alignments from other tools. For FR sequence–structure alignments, a user can enter the job ID of the (independently generated) FR results from the Bioinfo Meta server (16) or the FR Inub server (17). FRalanyzer displays a copy of the FR server results with an additional column labeled as 'Fralanyzer' which consists of the PDB codes of the templates identified by the FR server. A sequence–structure alignment is defined by selecting one of the PDB codes. For structure–structure alignments, FRalanyzer accepts the results of the VAST server, which are displayed so that the user selects one of the templates as before.

FRalanyzer displays the selected alignment along with a number of options. A sample output is shown in Figure 1. We submitted the conserved hypothetical protein EF3048 to Bioinfo, which returned the Bioinfo ID '49416'. We used this ID as input to FRalanyzer (see http://fralanyzer.cse.buffalo.edu/samples/ for the initial output window as well as our online documentation files), and selected the first template (PDB 1w1a). Figure 1 shows the main FRalanyzer output. Along with the alignment, the sequence identity % and the secondary structure identity of the aligned query and template are shown (14 and 70%, respectively). Identical residues and identical secondary structure states in the query and template sequences are highlighted. The secondary structure of the query and template are obtained from the FR server and/or PDBSum (18).

Further automatic annotation of the alignment can be obtained from functional features of the template from the PDBSum and the SwissProt databases (19). PDBSum functional features include ConSurf residue conservation, active site residues, Prosite motifs, contacts to DNA, ligands, metals and H-bonds to ligands. Users can select any of the available features or all. Matched Prosite motifs are highlighted in the target sequence in bold red fonts. There are two options to obtain PDBSum features: 'To Bring' (access the PDBSum site) and 'Local Copy' (access local copies instead, if they exist). In the present example, the PDBSum features, active site residues, contacts to ligand, metal and Hbonds to ligand are retrieved and the respective checkboxes are checked (Figure 1). Each feature is displayed in a separate row with 1-letter amino acid of the template (if present in the template) or '-'s (if present in a structural homolog). In addition, the functionally important positions that are conserved in the alignment (aligned to identical residues in the query) are further highlighted with vertical bars. For example, see the template active site residues (D^{73}, G^{75}, H^{124} and H^{128}) highlighted in Figure 1. Other possible annotations include: catalytic CSA, Prosite patterns, ligand type or residue conservation. For instance, the ligand types in the present example are GOL and NDG. The residue conservation of the template is displayed as colored bars according to the Consurf (20) coloring scheme, where the most variable positions are colored turquoise, intermediately conserved positions are colored white and the most conserved positions are colored maroon.

The user can access further functional information for the template or the query from SwissProt. If selected, a BLAST search is executed to identify the SwissProt sequences that are most similar (if any). Features are extracted from the SwissProt FT lines from entries having sequence identity >90%.

Other FRalanyzer links

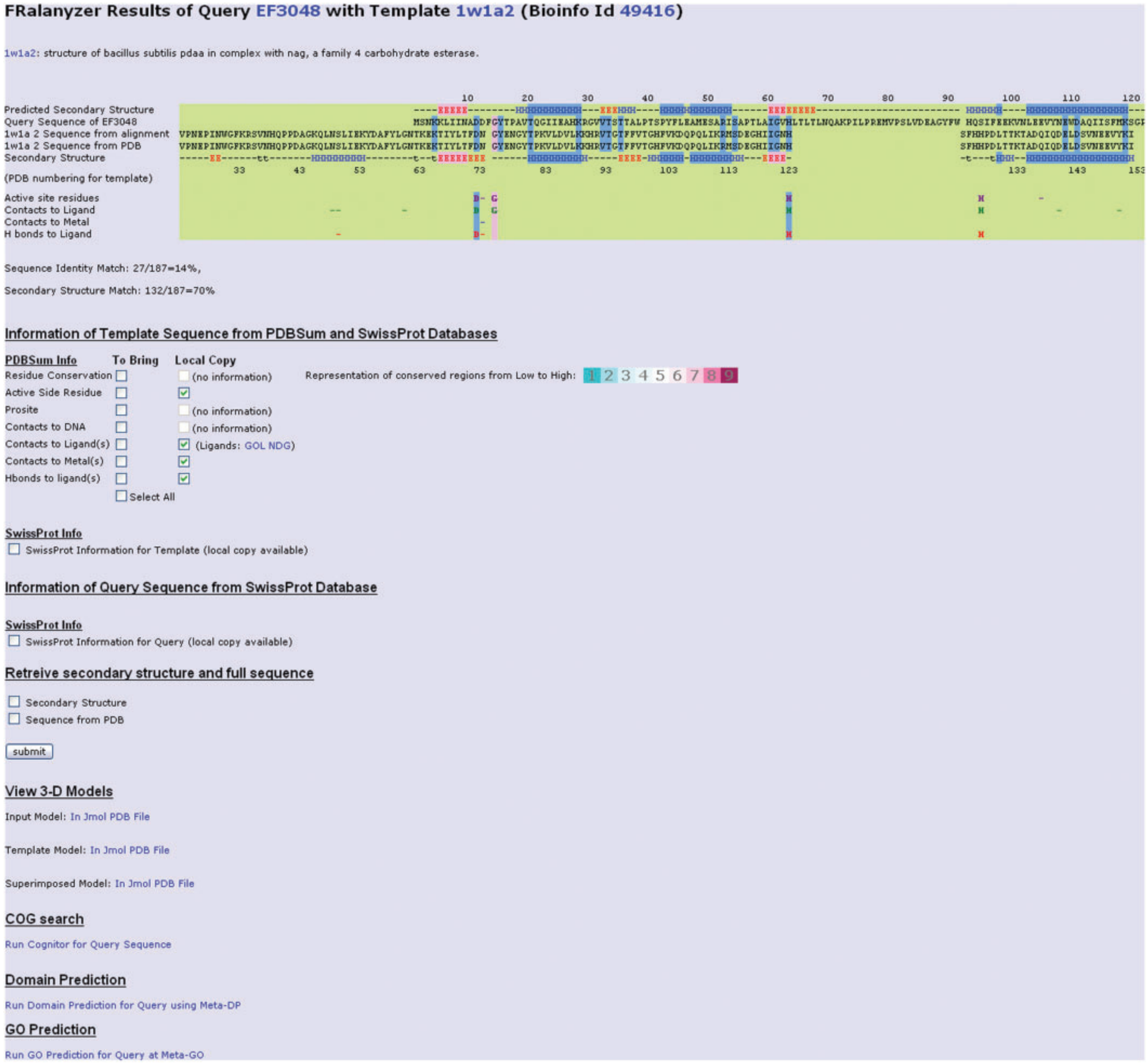
Currently, FRalanyzer has links to COGNITOR (21), to Meta-DP for domain prediction (22) and to Meta-GO for functional inference (unpublished).

Two sample predictions for proteins of unknown function

Our example shown in Figure 1 corresponds to a prediction for the Hypothetical UPF0249 protein EF3048, a protein of unknown 3D-structure. To identify a compatible template and to generate a sequence–structure alignment, we submitted this sequence to Bioinfo, which identified the structure of the hydrolase PDAA (PDB code 1w1a; a carbohydrate esterase) as a compatible template (with a very reliable score; see our 'samples' url). To determine whether any functional inference can be obtained, we analyzed the sequence–structure alignment. The alignment shows (Figure 1) that a number of the active site residues of 1w1a annotated in PDBSum are aligned to identical residues in the query. Based on this conservation of functionally important residues, the following hypothesis may be worthwhile to analyze in further detail: the conserved residues of EF3048 correspond to active site residues, and EF3048 is functionally related to the carbohydrate esterase family (see legend of Figure 1). The second example corresponds to a functional prediction obtained for the recently determined structure of a hypothetical protein of unknown function (PDB code 1ylo). We first identified structural homologs using VAST. FRalanyzer was used to visualize the structural alignment and to identify possibly functional important residues in the query (see our 'samples' url). Based on the conservation of functionally important residues, a plausible hypothesis locating the active site residues of 1ylo and its possible function as a peptidase can be postulated.

CONCLUSION

Predicting functionally important residues of a protein of unknown function is a useful step in characterizing it. An effective way to is to identify conserved functional key residues in an alignment with a template of known 3D-structure. If an experimental structure exists for the query protein, the alignment can be obtained from structural alignments. Otherwise, it can be obtained by e.g. sequence searches or FR. Relevant functional features of the query can be predicted based on the functional features of the template. FRalanyzer was developed to help in this process. FRalanyzer automatically extracts important functional residues from annotated databases and displays them along with the sequence–structure alignment, highlighting the conserved positions. This helps the user to quickly locate functionally important residues



FRalanyzer Results of Query EF3048 with Template 1w1a2 (Bioinfo Id 49416)

1w1a2: structure of bacillus subtilis pdaa in complex with nag, a family 4 carbohydrate esterase.

```
                                                                              10        20        30        40        50        60        70        80        90        100       110       120
Predicted Secondary Structure                                          ----EEEEE--------HHHHHHHHHHH---EEEHHH----HHHHHHHHHHHH------EEEEEEEEE----------------------- HHHHHHH----HHHHHHHHHHHHHHHH---
Query Sequence of EF3048                                               MSNKKLIINADDFGYTPAVTQGIIEAHKRGVVTSTTALPTSPYFLEAMESARISAPTLAIGVHLTLTLNQAKPILPREMVPSLVDEAGYFW HQSIFEEKVNLEEVYNEWDAQIISFMKSGP
1w1a 2 Sequence from alignment VPNEPINWGFKRSVNHQPPDAGKQLNSLIEKYDAFYLGNTKEKTIYLTFDN GYENGYTPKVLDVLKKHRVTGTFFVTGHFVKDQPQLIKRMSDEGHIIGNH                            SFHHPDLTTKTADQIQDELDSVNEEVYKI
1w1a 2 Sequence from PDB       VPNEPINWGFKRSVNHQPPDAGKQLNSLIEKYDAFYLGNTKEKTIYLTFDN GYENGYTPKVLDVLKKHRVTGTFFVTGHFVKDQPQLIKRMSDEGHIIGNH                            SFHHPDLTTKTADQIQDELDSVNEEVYKI
Secondary Structure            -----EE------tt-------HHHHHHHHHH-------t--tEEEEEEEE ------HHHHHHHHHHHH-----EEEE-HHHHH-HHHHHHHH---EEEE-                            -t---tHHH--HHHHHHHHHHHHHHHHHH
(PDB numbering for template)           33        43        53        63        73        83        93       103       113       123                                133       143       153

Active site residues                                                            D- G                                                H                                 H      -
Contacts to Ligand                                            --     -           D  G                                                H                                 H         -         -
Contacts to Metal                                                                 -
H bonds to Ligand                                              -                 D-                                                  H                                 H
```

Sequence Identity Match: 27/187=14%,

Secondary Structure Match: 132/187=70%

Information of Template Sequence from PDBSum and SwissProt Databases

PDBSum Info	To Bring	Local Copy
Residue Conservation	☐	☐ (no information)
Active Side Residue	☐	☑
Prosite	☐	☐ (no information)
Contacts to DNA	☐	☐ (no information)
Contacts to Ligand(s)	☐	☑ (Ligands: GOL NDG)
Contacts to Metal(s)	☐	☑
Hbonds to ligand(s)	☐	☑
	☐ Select All	

Representation of conserved regions from Low to High: 1 2 3 4 5 6 7 8 9

SwissProt Info

☐ SwissProt Information for Template (local copy available)

Information of Query Sequence from SwissProt Database

SwissProt Info

☐ SwissProt Information for Query (local copy available)

Retreive secondary structure and full sequence

☐ Secondary Structure

☐ Sequence from PDB

View 3-D Models

Input Model: In Jmol PDB File

Template Model: In Jmol PDB File

Superimposed Model: In Jmol PDB File

COG search

Run Cognitor for Query Sequence

Domain Prediction

Run Domain Prediction for Query using Meta-DP

GO Prediction

Run GO Prediction for Query at Meta-GO

Figure 1. FRalanyzer results of the alignment of the conserved hypothetical protein EF3048 with the 3D-structure of pdaa protein (PDB code 1w1a). The alignment was obtained from the Bioinfo FR meta-server, (job ID 49416). For clarity, only the first 120 residues are shown (see our 'samples' url for the full alignment). The predicted secondary structures (H and E) of EF3048 are shown above its sequence. The experimental secondary structure of 1w1a is shown below its sequence. Identical residues are highlighted. PDBSum features of the template such as active site residues, contacts to metal and ligands are shown in separate rows and are marked as asterisks or hyphens. 1w1a belongs to the carbohydrate esterase family 4, and contains a conserved D and 3 H residues (D^{73}, H^{124}, H^{128} and H^{222}), which interact with the substrate. Other active site residues (e.g. G^{75}) annotated by PDBSum line the active site groove. The FRalanyzer results highlight the conservation of the residues D^{73}, G^{75}, H^{124} and H^{222}. The alignment also shows that H^{128} is shifted 2 residues from H^{93} in the query, probably due to a misalignment in the FR result. Thus, by quickly visualizing the FR alignment, we are able to derive a verifiable hypothesis that identifies EF3048's catalytic residues D^{12}, H^{64}, H^{93} and H^{215}, suggesting that EF3048 is also a member of the carbohydrate esterase family 4.

in the query, which in turn can help in its functional characterization. FRalanyzer can also be useful in the visualization of structural alignments to known templates of newly determined proteins of unknown function. Future versions may include links and automatic accesses to other annotated databases and servers, as well as graphical enhancements, faster response times and distinction between different types of ligands.

ACKNOWLEDGEMENTS

The Fischer lab is supported by grants from the NIH and Genefun. We thank the developers of PDBSum and Roman Laskowski for his support. Funding to pay the Open Access publication charges for this article was provided by a grant from the NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Friedberg,I., Jambon,M. and Godzik,A. (2006) New avenues in protein function prediction. *Protein Sci.*, **15**, 1527–1529.
2. Siew,N. and Fischer,D. (2003) Twenty thousand ORFan microbial protein families for the biologist? *Structure*, **11**, 7–9.
3. Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
4. Thornton,J.M. (2001) From genome to function. *Science*, **292**, 2095–2097.
5. Whisstock,J.C. and Lesk,A.M. (2003) Prediction of protein function from protein sequence and structure. *Q Rev Biophys*, **36**, 307–340.
6. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
7. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
8. Groth,D., Lehrach,H. and Hennig,S. (2004) GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res.*, **32**, W313–W317.
9. Martin,D.M., Berriman,M. and Barton,G.J. (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, **5**, 178.
10. Khan,S., Situ,G., Decker,K. and Schmidt,C.J. (2003) GoFigure: automated Gene Ontology annotation. *Bioinformatics*, **19**, 2484–2485.
11. Tian,W., Arakaki,A.K. and Skolnick,J. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.*, **32**, 6226–6239.
12. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.
13. Fischer,D., Rice,D., Bowie,J.U. and Eisenberg,D. (1996) Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.*, **10**, 126–136.
14. Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.
15. Siew,N., Saini,H.K. and Fischer,D. (2005) A putative novel alpha/beta hydrolase ORFan family in Bacillus. *FEBS Lett.*, **579**, 3175–3182.
16. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.
17. Fischer,D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.
18. Laskowski,R.A., Chistyakov,V.V. and Thornton,J.M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.
19. Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.*, **25**, 31–36.
20. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
21. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
22. Saini,H.K. and Fischer,D. (2005) Meta-DP: domain prediction meta-server. *Bioinformatics*, **21**, 2917–2920.