BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 13 2004, pages 2150–2152
doi:10.1093/bioinformatics/bth208

F2CS: FSSP to CATH and SCOP prediction server

Gad Getz[1], Alina Starovolsky[2] and Eytan Domany[1,*]

[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel and [2]Computer Science Department, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Received on October 23, 2003; revised on January 6, 2004; accepted on January 27, 2004
Advance Access publication April 1, 2004

ABSTRACT

Summary: The F2CS server provides access to the software, F2CS2.00, which implements an automated prediction method of SCOP and CATH classifications of proteins, based on their FSSP *Z*-scores.

Availability: Free at http://www.weizmann.ac.il/physics/complex/compphys/f2cs/

Contact: eytan.domany@weizmann.ac.il

Supplementary information: The site contains links to additional figures and tables.

INTRODUCTION

Since during evolution protein structures are much more conserved than sequences and even functions (Holm and Sander, 1996), proteins are usually classified first by their structural similarity. Newly solved structures of proteins are regularly stored in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). Many research groups study the diversity of protein structures and maintain their web-accessible hierarchical classifications. Three widely used databases are FSSP (Holm and Sander, 1997), CATH (Orengo *et al.*, 1997) and SCOP (Conte *et al.*, 2000); although each has its own way to compare and classify proteins, the resulting classification schemes are, largely, consistent with each other (Getz *et al.*, 2002; Getz, 1998; Hadley and Jones, 1999).

The major difference between these three classification schemes, relevant to this work, is their degree of automation. FSSP is based on a fully automated structure comparison algorithm, DALI (Holm and Sander, 1994; Dietmann *et al.*, 2001), which calculates a structural similarity measure (represented in terms of *Z*-scores) between pairs of structures of protein chains taken from the PDB. FSSP first selects a subset of representative structures from the PDB and then applies the DALI algorithm to calculate the *Z*-scores for all pairs of representatives. Next, it calculates the *Z*-scores between each representative and the PDB structures it represents. Being fully automated, FSSP can be updated fairly often. FSSP was recently extended by a new database, called Dali [Holm (2003), Dali database at http://www.bioinfo.biocenter.helsinki.fi:8080/dali, private communication] which contains all-against-all *Z*-scores between chains and domains of a larger representative set, PDB90 (Hubbard *et al.*, 1999), in which no two chains are $>90\%$ sequence identical. In contrast, CATH and SCOP use manual classification at certain levels of their hierarchy, which slows down the classification process and makes it more subjective and error-prone.

CATH arranges protein domains in a four-level hierarchy according to their Class (secondary structure composition), Architecture (shape formed by the secondary structures), Topology (connectivity order of the secondary structures) and Homologous superfamily (structural and functional similarity). Classification of Architecture is done by visual inspection; hence CATH is partially manual.

The top level (Class) of the SCOP database also describes the secondary structure content of a protein domain. The next level (Fold) groups together structurally similar domains. The lower two levels (superfamily and family) describe near and distant evolutionary relationships (Levitt and Chothia, 1976). 'Fold' largely corresponds to CATH's topology level (Getz *et al.*, 2002). SCOP is constructed manually, based on visual examination and comparison of structures, sequences and functions.

We present here a web-based server, available at http://www.weizmann.ac.il/physics/complex/compphys/f2cs/, whose aim is to predict, without human intervention, using a protein's FSSP (or DALI) *Z*-scores, its full SCOP and CATH classifications. This can help classify proteins of known structure that have not yet been processed by SCOP or CATH (whose new releases are provided about every six months), and call attention to yet unseen structural classes.

If a protein appears in FSSP, the server returns our prediction. If it is not in FSSP, the user can submit the new structure to the DALI server, insert the resulting *Z*-scores into our server and obtain its predicted classification. In both cases

*To whom correspondence should be addressed.

Table 1. Results obtained by submitting '1dowb' to the F2CS server

Chain id	CATH v2.5				CATH prediction			SCOP 1.63			SCOP prediction	
	No.	C	A	T	**C**	**A**	**T**	No.	C	F	**C**	**F**
1dowb	−1				**1**	**20**	**5**	−1			**8**	**1**
Success%					**100**	**99**	**100**				**97**	**100**

This protein was classified neither by CATH v2.5 nor SCOP 1.63 (indicated by −1 in the 'number of domains' columns). We predict the following classifications: 1.20.5 for CATH and 8.1 for SCOP, both at 100% confidence level.

F2CS outputs a table showing the prediction, along with its confidence level.

THE SERVER

The current predictions are based on the latest versions of the databases; FSSP (Jun 16, 2002 update), combined with the Dali database (preliminary version, May 2003); CATH version 2.5 (July 2003) and SCOP 1.63 release (May 2003). The FSSP database contains 27 182 chains, of which 2860 are representatives. We superimposed on these the Z-scores from the Dali database, which were calculated for 6433 PDB90 chains; we refer to the combined database as FSSP/DD. Only significant Z-scores are reported ($\geq$2) and used; all other Z-scores are assumed to be zero.

The server implements our method (Getz *et al*., 2002), Classification by Optimization (CO), an optimization procedure that searches for that class assignment of proteins (that have not yet been processed by CATH or SCOP), which attains a minimal cost. The cost of an assignment is the sum of Z-scores between all pairs of proteins that were not assigned to the same class. This is a 'semi-supervised' algorithm, since it utilizes for its prediction the labels of the proteins with known classification and also the Z-scores among the training and predicted sets. We cannot classify 'isolated' proteins, which are not connected by a path of neighboring chains (i.e. $Z \geq 2$) to a chain of known classification.

We generate a prediction database of chains that appear in FSSP/DD but not in SCOP or CATH by applying our algorithm for each classification scheme. The FSSP/DD version we are using contains 4014 chains which do not appear in CATH v2.5 (we supply a prediction for 3170 of these) and 511 which are not in SCOP 1.63 (for 403 of these we have a prediction); 272 chains appear neither in CATH nor in SCOP. Since CATH and SCOP handle protein domains, whereas FSSP/DD entries are protein chains (consisting of one or more domains[1]), we use the single domain chains that are of known classification as a training set. Note that SCOP and CATH do not always agree on their separation of proteins into domains.

[1]We do not classify the few cases, when a single domain contains several different chains or a combination of their parts.

Our prediction's success rate was estimated using a blind test in which we hid the assignments of 3605 proteins from CATH and 4570 proteins from SCOP and tested our predictions against the known classifications. The success rate was tested for each class separately (see website for details). Due to a larger number of training examples and more stringent criteria for attempted classification, the success rate has improved over our previous work.

With every new release of the databases, F2CS can be updated; the newly released CATH/SCOP classifications are added to the training set, while predictions are made for proteins contained in a new FSSP/DD release which are not yet classified by CATH or SCOP.

USAGE

In order to retrieve our prediction for CATH's class, architecture and topology or SCOP's class and fold of a protein, enter the protein chain's identifier in the search box and submit the query. If the protein appears in our database, a table will be returned containing both the known and the predicted SCOP and CATH classifications. For example, submission of the chain identifier '1dowb', which was classified by neither CATH v2.5 nor SCOP v1.63, returns Table 1. We predict CATH classification 1.20.5 and SCOP 8.1, both near 100% confidence level. The 'Success%' link points to a table with the exact numbers by which the success rates were estimated.

In case the queried protein is not in our database, the user can obtain its predicted classification by following these two steps: (a) submit the protein's PDB file to the DALI server (the engine behind FSSP), which calculates its structural similarity to the FSSP representatives and returns a list of the representatives and Z-scores for which $Z \geq 2$; and (b) paste DALI's reply in the appropriate query box in our server.

ACKNOWLEDGEMENTS

We thank L.Holm for directing us to her new Dali database, and M.Vendruscolo for his advice and active involvement in the initial stages of this project, which was partially supported by the German–Israel Science Foundation (GIF). G.G. is supported by the Sir Charles Clore Doctoral Scholarship.

REFERENCES

Bernstein,F., Koetzle,T., Williams,G., Meyer,E.J., Brice,M., Rodgers,J., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Conte,L.L., Ailey,B., Hubbard,T.J.P., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.

Dietmann,S., Park,J., Notredame,C., Heger,A., Lappe,M. and Holm,L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.*, **29**, 55–57.

Getz,G. (1998) *Clustering and Classification of Protein Structures*. M.Sc. Thesis, Tel-Aviv University.

Getz,G., Vendruscolo,M., Sachs,D. and Domany,E. (2002) Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. *Proteins*, **46**, 405–415.

Hadley,C. and Jones,D.T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, **7**, 1099–1112.

Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600–3609.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.

Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

Holm,L. (2003).

Hubbard,T.J., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) SCOP: a structural classification of protein database. *Nucleic Acids Res.*, **27**, 254–256.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

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.