BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 8 2002
Pages 1141–1142

Domain Fishing: a first step in protein comparative modelling

*B. Contreras-Moreira and P. A. Bates**

Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

Received on January 17, 2002; revised on February 27, 2002; accepted on March 11, 2002

ABSTRACT

Summary: To optimize the search for structural templates in protein comparative modelling, the query sequence is split into domains. The initial list of templates for each domain, extracted from PFAM plus PDB and SCOP, is then ranked according to sequence identity (%ID), coverage and resolution. If %ID is less than 30, secondary structure matching is used to filter out false templates.

Availability: http://www.bmm.icnet.uk/~3djigsaw/dom_fish

Contact: b.contreras-moreira@cancer.org.uk; paul.bates@cancer.org.uk

INTRODUCTION

In the process of comparative modelling the structure of a protein (query), the first step must be a search for candidate templates. The simplest approach is a sequence similarity search against the database of protein structures, PDB (Berman *et al.*, 2000). To increase the sensitivity, a non-redundant sequence database is usually added to the PDB sequences (Bates and Sternberg, 1999). Nevertheless, query proteins may have several domains and if the closest template is selected, as reported by a local sequence similarity search such as PSI-BLAST (Schaffer *et al.*, 2001), only some of the domains would be modelled. Furthermore, it is necessary to identify automatically false templates found by sequence similarity.

To address these problems Domain Fishing uses a combination of the PDB, the protein families database PFAM (Bateman *et al.*, 2000) and the structural classification of proteins SCOP (Murzin *et al.*, 1995). Templates are still taken from the PDB but PFAM and SCOP are then used to rationally split them and the query sequence into single domains and to add remote homologous templates, generating a sequence profile for the query. In addition, functional annotations for each domain are taken from PFAM. Finally we use secondary structure matching to reject false templates.

*To whom correspondence should be addressed.

SYSTEMS AND METHODS

Two sequence databases are built: dPFAM_PDB and dSCOP, the fasta file of the latest release of SCOP, taken from http://astral.stanford.edu. To construct dPFAM_PDB, fasta files from the original PFAM A and B (Stockholm format, http://www.sanger.ac.uk/Pfam) are generated and family identifiers added to sequence headers. These two files are added to a weekly updated PDB sequence file and low complexity regions are masked with SEG (Wootton and Federhen, 1993).

Each search consists of two iterations of PSI-BLAST-2.2.1 against dPFAM_PDB, keeping the position-specific scoring matrix (pssm) file generated and all the alignments (e-value cut-off 0.05). The secondary structure of the query is predicted using PsiPred-2 (Jones, 1999). The secondary structure for the templates is assigned using the DSSP program (Kabsch and Sander, 1983).

ALGORITHM

For each query (see Figure 1) a search is performed against dPFAM_PDB. Domain boundaries are assigned according to the more significant PFAM hits spanning each part of the sequence.

For each domain a list of overlapping PDB hits is generated. The list may be extended by taking additional templates from the corresponding PFAM and/or SCOP families. dSCOP is also used to define the starting and ending residues for each template. This list is initially filtered to reject Cα only proteins and chains with missing atoms in the backbone. The remaining templates are then aligned to the query domain using our own procedure, a global sequence alignment using the pssm as scoring matrix and taking into account secondary structure matching (SSM). It has been reported that this combination of sequence profiles and secondary-structure information yields better alignments than using just sequence (Elofsson, 2002), particularly with low similarity sequences.

Finally, they are sorted by *coverID* , using crystallographic resolution to discriminate between identical tem-

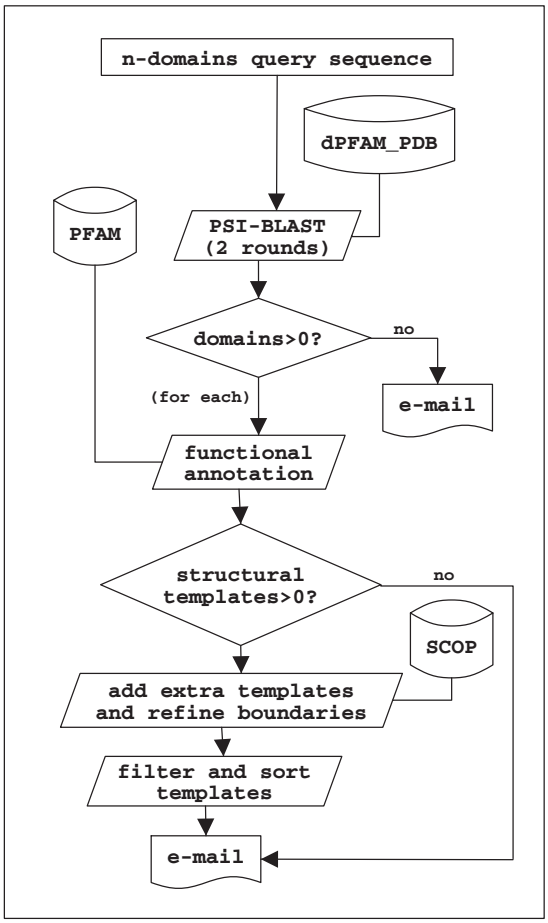

Fig. 1. Flow diagram for Domain Fishing

plates, where:

$$coverID = \%ID \times \frac{alignment\ length}{domain\ length}$$

For templates under 30%ID, so called low-accuracy templates (Baker and Sali, 2001), a minimum of 60% SSM is expected (based on Sander and Schneider, 1991), otherwise they are rejected. This calculation is done without considering coil regions.

IMPLEMENTATION

The Domain Fishing code is written in Perl, except the query-to-template alignment program, which is written in C++. Each query takes an average runtime of 100 s. When a job is finished the user is sent an e-mail with a link to the results, which are kept for up to a week. The results page allows the user to see a graphical display of the query domains and access PFAM, PDB and SWISS-PROT (Bairoch and Apweiler, 2000) annotations for each. Alignments between domains and templates can also be displayed along with the reported %ID and secondary structure mismatches.

DISCUSSION AND CONCLUSION

As seen in the CASP blind trials for comparative protein structure prediction (Jones and Kleywegt, 1999), two major sources of errors are selection of incorrect templates and bad alignments. This server addresses both problems, by looking for high scoring templates at the domain level and using secondary structure information to improve alignments. Even if there are no sufficiently good candidate templates for a domain or protein, the predicted secondary structure and the functional annotation reported to the user may be useful to characterize the protein.

REFERENCES

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Bates,P.A. and Sternberg,M.J.E. (1999) Model building by comparison at CASP3: using expert knowledge and computer automation. *Proteins: Struct. Funct. Genet.*, **Suppl 3**, 47–54.

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.

Elofsson,A. (2002) A study on protein sequence alignment quality. *Proteins*, **46**, 330–339.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,T.A. and Kleywegt,G.J. (1999) CASP3 Comparative modeling evaluation. *Proteins: Struct. Funct. Genet.*, **Suppl 3**, 30–46.

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.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, **17**, 149–163.