BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001
Pages 748–749

HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families

P. I. W. de Bakker[1], A. Bateman[2], D. F. Burke[1], R. N. Miguel[1], K. Mizuguchi[1], J. Shi[1], H. Shirai[1] and T. L. Blundell[1]

[1]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK and [2]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received on February 20, 2001; revised on March 30, 2001; accepted on April 6, 2001

ABSTRACT

Summary: We describe an extension to the Homologous Structure Alignment Database (HOMSTRAD; Mizuguchi *et al.*, *Protein Sci.*, **7**, 2469–2471, 1998a) to include homologous sequences derived from the protein families database Pfam (Bateman *et al.*, *Nucleic Acids Res.*, **28**, 263–266, 2000). HOMSTRAD is integrated with the server FUGUE (Shi *et al.*, submitted, 2001) for recognition and alignment of homologues, benefitting from the combination of abundant sequence information and accurate structure-based alignments.

Availability: The HOMSTRAD database is available at: http://www-cryst.bioc.cam.ac.uk/homstrad/. Query sequences can be submitted to the homology recognition/alignment server FUGUE at: http://www-cryst.bioc.cam.ac.uk/fugue/.

Contact: homstrad@cryst.bioc.cam.ac.uk

Recently, we reported the construction of a database containing structural alignment of homologous protein families (HOMSTRAD; Mizuguchi *et al.*, 1998a). These structural alignments have been used to derive environment-specific substitution tables (Overington *et al.*, 1992) for comparative modelling (Burke *et al.*, 1999) and for query-template alignment and recognition (Shi *et al.*, 2001). The rapid growth of the sequence databases has motivated us to augment the structural alignments in HOMSTRAD with sequence information. The Pfam database (Bateman *et al.*, 2000) is a large collection of protein multiple sequence alignments, representing a convenient source to enrich the HOMSTRAD alignments.

We established links between HOMSTRAD and Pfam families by scoring all HOMSTRAD sequences against the Pfam profile hidden Markov models (profile HMMs) with HMMER using a conservative E-value of 0.001. Homologous sequences were collected from the seed alignments of corresponding families in Pfam and aligned against the HOMSTRAD structural profile using the program FUGUE (Shi *et al.*, 2001). Discrepancies between definitions of protein domains due to differing sequence boundaries in HOMSTRAD and Pfam were resolved by replacing Pfam sequences with the original SWISS-PROT/TrEMBL sequences (Bairoch and Apweiler, 2000) and subsequent truncation at the termini of the HOMSTRAD profile. The alignments are displayed according to the JOY format (Mizuguchi *et al.*, 1998b), revealing three-dimensional structural features (such as secondary structure, solvent accessibility and hydrogen bonding) that help explain conservation patterns of key residues within a family. Additionally, the resulting alignments are annotated with matching PROSITE patterns (Hofmann *et al.*, 1999), colour coded according to solvent accessibility and sequence conservation in order to highlight functional residues. Hyperlinks are provided to the corresponding entries in Pfam, PROSITE, SCOP, CATH, FSSP and SWISS-PROT/TrEMBL databases. A link to the FUGUE server is provided that allow query sequences to be aligned against a family profile.

Different protein sequence alignment methods tend to be locally inconsistent, most of the inconsistencies occurring in loop regions where there may not be a single optimal alignment. We should emphasize that Pfam alignments are generated on the basis of sequence alone, while the HOMSTRAD alignments are based on structural superposition and topological equivalences (Šali and Blundell, 1990). It is not surprising that there are some discrepancies between Pfam and HOMSTRAD in their respective domain definitions for a number of families. When these differences are large, they usually involve families of proteins comprising tandem structural repeats. In Pfam, the alignments are generally composed of a single repeat, with the profile–HMM matching a

Table 1. Statistics of HOMSTRAD

Number of families with multiple protein structures	514
Number of families with multiple protein structures enriched with homologous sequences	399
Number of structures in families with multiple protein structures	2049
Average number of protein structures per family	4
Average sequence length of protein structures per family	230
Average sequence identity between protein structures per family (%)	40
Average number of added Swiss-Prot/TrEMBL sequences per family	39
Total number of matching PROSITE patterns	644

protein multiple times, whereas in HOMSTRAD they are usually left intact preserving the entire structure with several structural repeats. For example, the Pfam annexin alignment contains a single structural repeat while the HOMSTRAD alignment contains four. For comparative modelling, it is advantageous to provide a global alignment between the target sequence and multi-repeat template(s), circumventing the problem of the assembly of multiple repeats into a complete structural model. This also applies to proteins comprising multiple domains where their respective orientation cannot otherwise be reliably predicted.

We carried out a comparison between HOMSTRAD and Pfam alignments, in order to provide an indication of the quality of the Pfam alignments. Because the HOMSTRAD and Pfam alignments are built from different sequences, we realigned the HOMSTRAD sequences with the Pfam profile–HMM and counted the number of pairs of residues that were identical between the two alignments. We were able to compare 399 families between the two databases. The majority (221) of the families were identical at 80% of aligned positions with the remaining 20% corresponding to likely inconsistencies at loop regions. However, 59 families shared less than 50% of aligned positions. Eighteen of these are due to repeats being treated differently between Pfam and HOMSTRAD. Other disagreements were due to large differences in domain definitions. In a few families with similar domain definitions, the alignments were very dissimilar. Investigation of these cases showed that, for example, the Pfam histone family is poorly aligned; this and other families are being changed in Pfam to incorporate the structural alignment. In addition, we were able to identify two alignments in HOMSTRAD that were corrupted by poorly resolved structures. Smaller differences between Pfam and HOMSTRAD alignments were evaluated on a case-by-case basis and many are due to the Pfam alignment only containing the conserved core of the domain, as in the immunoglobulin domain.

HOMSTRAD is a unique structural alignment database benefitting from the growth of both sequence and structure databases. The databases is tightly coupled to the homology recognition/alignment server FUGUE which shows significant improvement by the combination of introduced homologous sequence information and existing structural profiles from HOMSTRAD. Such protein databases have proved to be a valuable tool in recent CASP experiments (Burke *et al.*, 1999).

ACKNOWLEDGEMENTS

PIWDB thanks Simon Lovell for valuable comments and the Cambridge European Trust for financial support. RNM thanks the ‘Ministerio Espanol de Educacion y Cultura’ for a post-doctoral grant (ref. EX 99 09274532).

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.

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

Burke,D.F., Deane,C.M., Nagarajaram,H.A., Campillo,N., Martin-Martinez,M., Mendes,J., Molina,F., Perry,J., Reddy,B.V., Soares,C.M., Steward,R.E., Williams,M., Carrondo,M.A., Blundell,T.L. and Mizuguchi,K. (1999) An iterative structure-assisted approach to sequence alignment and comparative modeling. *Proteins*, **3** (Suppl.), 55–60.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998a) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998b) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.

Overington,J., Donnelly,D., Johnson,M.S., Šali,A. and Blundell,T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.

Šali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.

Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, submitted.