BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001
Pages 746–747

DBAli: a database of protein structure alignments

*Marc A. Martí-Renom, Valentin A. Ilyin and Andrej Sali**

Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Ave, New York, NY 10021, USA

Received on February 14, 2001; revised and accepted on April 25, 2001

ABSTRACT

Summary: The DBAli database includes approximately 35 000 alignments of pairs of protein structures from SCOP (Lo Conte *et al.*, *Nucleic Acids Res.*, **28**, 257–259, 2000) and CE (Shindyalov and Bourne, *Protein Eng.*, **11**, 739–747, 1998). DBAli is linked to several resources, including Compare3D (Shindyalov and Bourne, http://www.sdsc.edu/pb/software.htm, 1999) and ModView (Ilyin and Sali, http://guitar.rockefeller.edu/ModView/, 2001) for visualizing sequence alignments and structure superpositions. A flexible search of DBAli by protein sequence and structure properties allows construction of subsets of alignments suitable for a number of applications, such as benchmarking of sequence–sequence and sequence–structure alignment methods under a variety of conditions.

Availability: http://guitar.rockefeller.edu/DBAli/

Contact: sali@rockefeller.edu; http://guitar.rockefeller.edu

Accurate reference alignments of protein sequences are necessary for improving methods for aligning protein sequences with each other and with proteins of known structure. Such reference alignments are invariably based on comparison of protein structures because structure is more conserved in evolution than sequence (Pascarella and Argos, 1992; Sali and Overington, 1994; Fischer *et al.*, 1996; Shindyalov and Bourne, 1998; Mizuguchi *et al.*, 1998; Thompson *et al.*, 1999b; Lo Conte *et al.*, 2000; Holm and Sander, 1999; Thompson *et al.*, 1999a; Domingues *et al.*, 2000; Bateman *et al.*, 2000). Although the structure-based alignments are the best alignments for most applications, different structure comparison methods sometimes result in different alignments; many different methods are reviewed in Swindells *et al.* (1998).

DBAli is a new database of alignments of pairs of related known protein structures. Its main distinction with respect to other sources of protein structure alignments is the flexibility of selecting different sets of alignments according to a variety of criteria. One of the aims of DBAli is to provide a large set of reference alignments for comprehensive evaluation of alignment methods over the whole spectrum of sequence and structure similarity. Currently, DBAli includes all 1843 pairwise alignments from SCOP 1.38 (Lo Conte *et al.*, 2000) and 33 920 non-redundant alignments from CE with the Z-score higher than 3.8 (Shindyalov and Bourne, 1998); the non-redundant CE alignments were obtained by aligning a structure with only one other structure from the same 'H' level in the CATH classification (Orengo *et al.*, 1999). The CE database was selected as the main source of pairwise structure alignments because of its availability, completeness, currency, and automatic construction. DBAli will continue to be updated to reflect the growth of the CE database.

DBAli can be queried by MySQL (version 3.23) (DuBois, 2000). The database is searchable by sequence and structure properties of the proteins and their alignments (Figure 1). Sequence properties that can be queried include Protein Data Bank (PDB) code (Berman *et al.*, 2000), sequence similarity to entries in the database, percentage sequence identity of the alignment, length of the alignment and the number of aligned residues. Structure properties include C_α RMSD and number of residues with C_α atoms within 3.5 Å after superimposition. Each query produces a table of alignments satisfying the search criteria. The table lists the alignment name with a link to the alignment description page, a link to the original alignment file in the MODELLER format, PDB codes corresponding to the aligned structures, percentage identity between the sequences in the alignment, C_α RMSD, a link to a schematic representation of the alignment and links to the Compare3D applet (Shindyalov and Bourne, 1999) and the ModView plugin (Ilyin and Sali, 2001) for structural visualization of the alignment. The user can download the selected alignments.

All alignments were processed, but not changed, by MODELLER (Sali and Blundell, 1993). Sequence and structure properties of an alignment are shown on its main page. The main alignment page is divided in three sections: (i) the sequence alignment with general data; (ii) sequence properties; and (iii) structure properties. Links to

*To whom correspondence should be addressed.

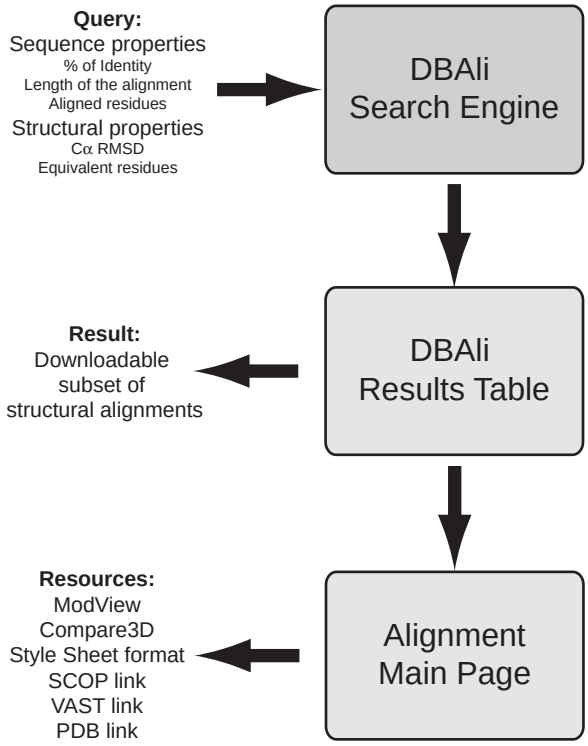

Fig. 1. DBAli flow chart.

the PDB (Berman *et al.*, 2000), SCOP (Lo Conte *et al.*, 2000) and VAST (Madej *et al.*, 1995) databases for each aligned structure are also on the alignment main page.

The usefulness of DBAli is illustrated by the following two examples. When the goal is to benchmark a new alignment method focused on the 'twilight zone' of sequence similarity, a DBAli query asking for high quality alignments (RMSD < 2 Å) of distantly related sequences (less than 30% sequence identity) returns 249 and 1961 pairwise alignments from SCOP and CE, respectively. Similarly, when the goal is to calculate a structure-based residue substitution matrix for remotely related protein sequences, a DBAli query asking for SCOP alignments with less than 2 Å RMSD and at most 25% sequence identity returns 160 pairwise alignments with 55 188 residues and 25 017 residue–residue substitutions.

ACKNOWLEDGEMENTS

The authors are grateful to Drs I.N.Shindyalov and P.E.Bourne for providing the CE alignments and the Compare3D applet. M.A.M-R. is supported by a postdoctoral fellowship awarded by The Burroughs Wellcome Fund. AS is an Irma T. Hirschl Trust Career Scientist. Support by The Merck Genome Research Institute, Mathers Foundation, and NIH is also acknowledged.

REFERENCES

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

Berman,H., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I. and Bourne,P. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Domingues,F., Lackner,P., Andreeva,A. and Sippl,M. (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.

DuBois (2000) *MySQL*. New Riders, Indianapolis, IN.

Fischer,D., Elofsson,A., Rice,D. and Eisenberg,D. (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac. Symp. Biocomput.*, **397**, 300–318.

Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.

Ilyin,V. and Sali,A. (2001) Modview. http://guitar.rockefeller.edu/ModView/.

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

Madej,T., Gibrat,J. and Bryant,S. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

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

Orengo,C., Pearl,F., Bray,J., Todd,A., Martin,A., Lo Conte,L. and Thornton,J. (1999) The CATH database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.

Pascarella,S. and Argos,P. (1992) A data bank merging related protein structures and sequences. *Protein Eng.*, **5**, 121–137.

Sali,A. and Blundell,T. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sali,A. and Overington,J. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.*, **3**, 1582–1596.

Shindyalov,I. and Bourne,P. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Shindyalov,I. and Bourne,P. (1999) Compare3d applet. http://www.sdsc.edu/pb/Software.htm.

Swindells,M., Orengo,C., Jones,D., Hutchinson,E. and Thornton,J. (1998) Contemporary approaches to protein structure classification. *Bioessays*, **20**, 884–891.

Thompson,J., Plewniak,F. and Poch,O. (1999a) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Thompson,J., Plewniak,F. and Poch,O. (1999b) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.