228–229 Nucleic Acids Research, 2001, Vol. 29, No. 1

A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm

Ilya N. Shindyalov and Philip E. Bourne[1,*]

San Diego Supercomputer Center and [1]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received September 5, 2000; Revised and Accepted October 25, 2000

ABSTRACT

The database reported here is derived using the Combinatorial Extension (CE) algorithm which compares pairs of protein polypeptide chains and provides a list of structurally similar proteins along with their structure alignments. Using CE, structure–structure alignments can provide insights into biological function. When a protein of known function is shown to be structurally similar to a protein of unknown function, a relationship might be inferred; a relationship not necessarily detectable from sequence comparison alone. Establishing structure–structure relationships in this way is of great importance as we enter an era of structural genomics where there is a likelihood of an increasing number of structures with unknown functions being determined. Thus the CE database is an example of a useful tool in the annotation of protein structures of unknown function. Comparisons can be performed on the complete PDB or on a structurally representative subset of proteins. The source protein(s) can be from the PDB (updated monthly) or uploaded by the user. CE provides sequence alignments resulting from structural alignments and Cartesian coordinates for the aligned structures, which may be analyzed using the supplied Compare3D Java applet, or downloaded for further local analysis. Searches can be run from the CE web site, http://cl.sdsc.edu/ce.html, or the database and software downloaded from the site for local use.

INTRODUCTION

The number of protein structures is increasing rapidly. This increase was near exponential in the early 1990s and has become linear over the past several years, with over 2500 structures deposited with the PDB during 1999 (1). Individually these structures provide new functional insights. Collectively they define our current knowledge of protein fold space (2,3) and provide a framework for comparative (homology) modeling (4). Several structure classification schemes (2,5,6) have been derived from the current contents of the PDB which when taken with other information, for example a demonstrated evolutionary relationship from sequence homology, lead to classifications by protein family. The database reported here does not provide this level of classification, but rather reports close and distant structure similarities for all protein structures in the PDB, or allows the user to compute such similarities for a protein structure not in the PDB.

Structure comparison and detailed structure alignment will become increasingly important in the era of structural genomics (7), as a tool in deciphering possible biological function. To this end we have developed, and operate through, the San Diego Supercomputer Center, a database of structure alignments based on the previously published Combinatorial Extension (CE) algorithm (8). It is the features of this database that are discussed here.

It has previously been pointed out (9) that providing accurate structure alignments is not a solved problem. Given the computational intractability of the problem when applied to a large data set like the PDB, each method of alignment makes simplifying assumptions. Different assumptions and the lack of a good statistical foundation for qualifying these assumptions leads to different detailed alignments even though the same fold may be recognized as being similar by a variety of methods. While these alignment methods do significantly better than sequence alignments at low sequence identity (10), the most accurate alignments come from using a variety of automated methods and then manually applying biological knowledge to optimize the alignment. The database presented here provides that automated starting point. The database takes no account of protein domains, but simply works on complete polypeptide chains. Since the algorithm searches for local alignments most similarities at the domain level are captured automatically.

FEATURES

The following features are available as specific options selected from the CE web page, http://cl.sdsc.edu/ce.html.

Find similar structures already in the PDB

The simplest operation is to find all similar structures to a starting polypeptide chain that is >30 residues in length and already in the PDB. The user can filter the list of hits based upon statistical significance (Z-score; 8), root mean square deviation (RMSD), length difference, allowable gaps (given as a percentage of the total number of residues without a matching partner relative to the complete alignment) and

*To whom correspondence should be addressed. Tel: +1 858 534 8301; Fax: +1 858 822 0873; Email: bourne@sdsc.edu

sequence identity. This search may be conducted on all polypeptide chains contained in the PDB or on a representative set defined with respect to structure and not sequence. The full definition of a representative structures is given in Shindyalov and Bourne (3) and only a synopsis is given here.

In establishing a representative set, polypeptide chains are randomly selected from a pool of all chains and compared against a list of structure representatives (which is empty in the beginning). For NMR structures the first member of an ensemble is chosen. If the new chain is similar to one of the structure representatives then it is assigned to that representative and becomes the representing chain. If the new chain is unique it becomes a new representative and is added to the representatives' list. Obviously, in the beginning the first chain examined starts the list of representatives. The following set of criteria are used to define structure representatives:

(i) The RMSD between two aligned chains is <2 Å;
(ii) The difference in chain length is <10%;
(iii) The number of aligned positions is at least two-thirds the length of the represented chain;
(iv) The number of gap positions in the alignment is <20% of the number of aligned positions.

Calculate similar structures starting with a structure not in the PDB

Users may submit their own polypeptide chains (with coordinates formatted according to PDB specifications) to be compared against the representative set of chains within the database derived from the PDB. If the submitted polypeptide chain can be represented by a chain from the representative set (in accordance with criteria above) already in the database, then structural similarities known for this representative polypeptide chain are returned, otherwise a full search against the representative set is performed. Even when the submitted polypeptide chain can be represented the user can request a full comparison against the representative set by unchecking the box labeled 'Use structure representatives' on the form. The calculation can take from 5 min to 10 h depending on the size of the polypeptide chain and the load on the server.

Calculate the alignment between two polypeptide chains

Two polypeptide chains (complete or partial and either in the database or one or both supplied by the user) may be structurally aligned using the same algorithm as for the all-by-all comparison. It is possible to seek shorter alignments with lower RMSD, or longer alignments with higher RMSD. It is also possible to take into account sequence information during the dynamic programming step.

Interpreting results

In a search of the database against a starting chain results are returned as a list of polypeptide chains with their associated annotation taken from the PDB describing compound name and with a hyperlink back to the PDB. The Z-score, RMSD, sequence identity, the number of residues aligned versus total chain length, and gaps are given. Alignments may be selected for further analysis. Those chains selected are aligned pairwise, with the starting (query) chain and a matrix of pairwise scores presented along with the sequence alignment resulting from the structure alignment. The structure alignment may be viewed with the standard tools Rasmol and Protein Explorer (using Chime) or in a specially designed Java applet, Compare3D. The applet allows the user to explore both similarities and differences between the aligned structures both from a sequence and structure perspective. Different features can be mapped onto the aligned structure and the facility exists to perform local superposition of structural fragments. Coordinates for the aligned structures may also be downloaded for further analysis.

FUTURE DEVELOPMENTS

Future plans call for adding structure alignments optimized for specific protein families. This has already been done for protein kinases, esterase and lipases and these are available from the web site. Also available are the alignments for the substructures identified as part of our mapping of protein fold space using the CE algorithm (3). This mapping analysis is on going.

AVAILABILITY AND ACCESS

Access is available from http://cl.sdsc.edu/ce.html. During the period September 1999 to August 2000, 48 720 online queries were conducted. The source code and executables for popular Unix platforms and a database of alignments are available for download and subsequent local use from the same web site. In the same period, 337 copies of the source code or executables were downloaded and 209 copies of the database. The web and downloadable databases are updated once per month from the PDB.

ACKNOWLEDGEMENTS

These databases are maintained by the National Biomedical Computation Resource (NIH P41 RR08605-06) and were developed under a grant from the National Science Foundation DBI 9808706.

REFERENCES

1. Berman,H.M., Bhat,T.N., Bourne,P.E., Gilliland,G., Weissig,H. and Westbrook,J. (2001) The PDB uniformity project. *Nucleic Acids Res.*, **29**, 214–218.
2. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
3. Shindyalov,I.N. and Bourne,P.E. (2000) An alternative view of protein fold space. *Proteins*, **38**, 247–260.
4. Hilbert,M., Bohm,G. and Jaenicke,R. (1993) Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins*, **17**, 138–151.
5. Lo Conte,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.
6. Orengo,C.A., Pearl,F.M., Bray,J.E., Todd,A.E., Martin,A.C., Lo Conte,L. and Thornton,J.M. (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 223–227.
7. Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) Structural genomics: beyond the human genome project. *Nature Genet.*, **23**, 151–157.
8. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension of the optimum path. *Protein Eng.*, **9**, 739–747.
9. Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
10. Sauder,J.M., Arthur,J.W. and Dunbrack,R.L.,Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.