The FSSP database: Fold Classification based on Structure–Structure alignment of Proteins

Manish Kumar[1], Kapil Govil[2]

[1]PhD Scholar, Department of Computer Science, Shri Venkateshwara University, Uttar Pradesh, India
[2]Assistant Professor, TMIMT, TMU, Moradabad, India

Abstract: ***FSSP (families of structurally similar proteins) is a database of structural alignments of proteins in the Protein Data Bank (PDB) (1). The current release has 2860 entries. The FSSP (http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-id+5Ti2u1RffMj+-lib+FSSP) database presently contains an extended structural family for every of 330 representative protein chains. Each information set contains structural alignments of one search structure with all different structurally significantly similar proteins within the representative set, for remote homologs, below 30% sequence identity, Further as all structures within the Protein Data Bank with 70-30% sequence identity relative to the search structure (medium homologs). For Very close homologs, which are above 70% sequence identity, they are excluded as they rarely have marked structural variations. The alignments of remote homologs are the result of pair wise all-against-all structural comparisons in the set of 330 representative protein chains. All such comparisons are based strictly on the 3D co-ordinates of the proteins and are derived by automatic (objective) structure comparison programs. The importance of structural similarity is estimated based on statistical criteria. The FSSP info is obtainable electronically from the EMBL file server and by anonymous ftp (file transfer protocol).***

Keywords: PDB, FSSP, DALI, families, Z-score.

1. Introduction

Fold classification based on the structure-structure alignment of proteins and families of structurally similar proteins (FSSP) is a database based on the structural alignment of pair wise combinations of proteins in the Protein Data Bank. The Alignments and classification of proteins are done automatically and are updated continuously by the DALI search engine. The similarities can be detected by structural comparisons that merge protein families of known 3-D structure into structural classes, the members of which can or might not be evolutionarily related (2–5). The FSSP (Figure 1) database presents a continuously updated structural classification of three-dimensional protein folds. It is derived victimization an automatic structure comparison program (DALI) for the all-against-all-comparison of three-dimensional coordinates sets in the Protein Data Bank. Sequence-related protein families are covered by a representative set of 330 protein chains. (Figure 2) shows homologous domains having similar structures. It shows a comparison between PH domains of human plekstrin (a major substrate of protein kinase C in platelets) and dynamic (a large GTPase involved in the scission of nascent vesicles from parent membranes).

Protein families are known to retain the shape of the fold even when sequences have diverged below the limit of detection of significant similarities at the sequence level unlike that shown in (Figure 2). Structural comparisons merge protein families of known 3D structure into structural classes, the members of which may or may not be evolutionary related. To aid investigation in the database, the 330 protein chains contained in the representative set have been clustered into fold families. A dendogram of the families was produced by average linkage clustering based on structural similarity scores. Chain length effects were corrected for by transforming the pair wise similarities into statistical significance scores (Z-scores). Families and subfamilies result from truncating the tree at different cut levels of the Z-score. The higher the cut, the larger the resulting number of distinct folds families.

2. Hierarchical clustering in FSSP

Hierarchical clustering supported structural similarities yields a fold tree that defines **253 folds** classes. For every representative protein chain, there's a information entry containing structure-structure alignments with its structural neighbours among the PDB. The information is accessible online through the World Wide Web browsers and by anonymous ftp (File Transfer Protocol). The outline of fold space and therefore the individual datasets offer an upscale supply of data for the study of each divergent and oblique aspects of molecular evolution and for outlining helpful check sets and a regular of truth for assessing the correctness of sequence-sequence or sequence-structure alignments.

3. DALI

The DALI (Figure 3) database is accessible over the www addressing URL http://ekhidna.biocenter.helsinki.fi/dali/start/. The DALI or **D**istance m**A**trix a**LI**gnment server is a network service used to compare three-dimensional protein structures. The query sequence coordinates are compared against those inside the PDB. A multiple alignment of structural neighbors is that the output. The DALI server is helpful to compare 3D structures wherever similarities don't seem to be detectable by comparing sequences directly. The comparison uses Max Sprout program to generate backbone and side-chain coordinates if these are not submitted along with the query sequence. Secondary structure elements and domains are defined using the DSSP and PUU programs. It is additionally attainable to understand the structural neighbours of a protein already within the Protein Data bank from the FSSP database. (Figure 4) summarizes the activity that the DALI server undertakes before arriving at the output. Details about the DALI method used to derive the database are given in

refs (6) and (7).

4. Difference between FSSP and DALI

The major difference between the two classification schemes, relevant to our work, is their degree of automation. FSSP relies on a fully automated structure comparison algorithm, DALI, that calculates a structural similarity measure (represented in terms of Z-score) between pairs of structures of protein chains taken from the PDB. A tree is then created by average linkage bunch of the structural similarity score. The tree is cut at DALI Z-score levels 2,4,8,16,32, and 64. The primary level (Z > 2) can be used as an operational definition of folds.

FSSP first selects a subset of representative structures from the PDB and then applies the DALI algorithm to calculate the Z-scores or all pairs of representatives. The domain number is appended as _n. Domains are numbered 1,2,3,…. Proteins with a domain numbered 0 are not appointed by the structural domain decomposition algorithm but are appointed as single-domain structures by default.

Next, it calculates the Z-scores between every representative and also the PDB. Being totally automatic, FSSP may be updated fairly often. FSSP was recently extended by replacement info, referred to as the DALI database at http://ekhidna.biocenter.helsinki.fi/dali/start/.

5. Anonymous Ftp

The FSSP data sets can be obtained by anonymous ftp from ftp://ftp.ebi.ac.uk/pub/databases/fssp/ in the directory.

5.1 Size of the current release

The size of the FSSP database is tightly coupled to it of the PDB from that it's derived. The FSSP database is updated with unleash of latest structures by the PDB. The current release has 2860 entries and was indexed 20-Aug-2012.

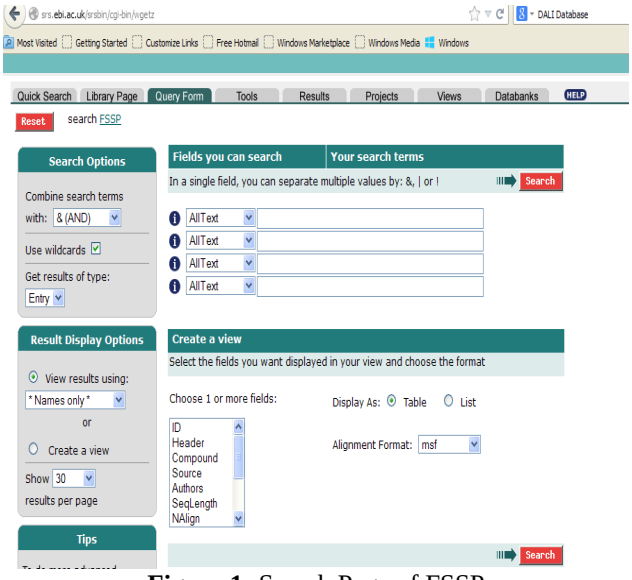

Figure 1: Search Page of FSSP

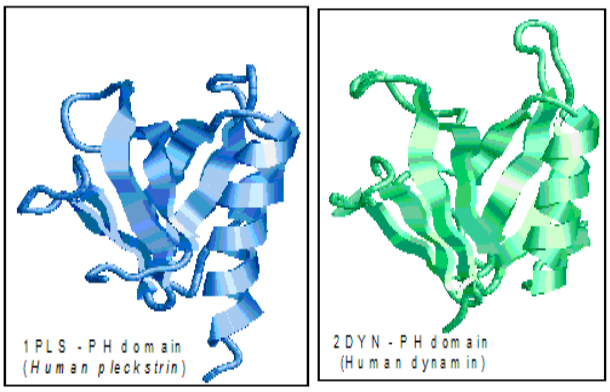

Figure 2: Similar structures of homologous domains

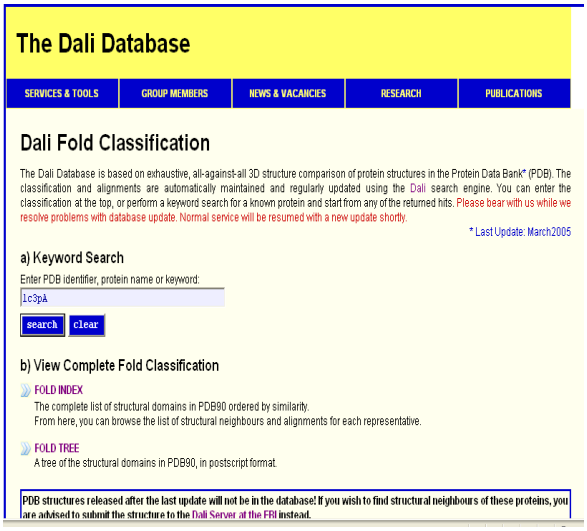

Figure 3: Search Page of DALI

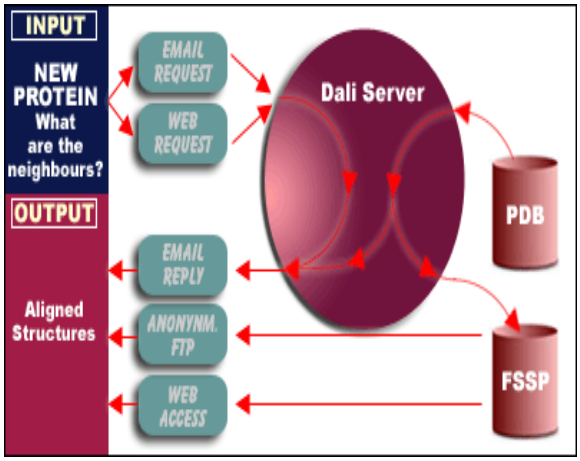

Figure 4: Activity of DALI server

6. Conclusion

As proteins evolve, their structures modify. Among the refined details that evolution has powerfully attended conserve are the patterns of contacts between residues. If two residues are in contact in one protein, the residues may also be aligned with these two in a very connected protein also are possible to be in reality. Mutations that modification the sizes of packed buried residues turn out changes within the packing of the helices and sheets against each other. This program runs quick enough to hold out routine screens of the entire Protein Data Bank for structures like a freshly

determined structure, associated even to perform a classification of protein domain structures from an all-against-all comparison. But during this info many surprising similarities not detectable at the extent of Pairwise sequence alignment.

References

[1] L. Holm, and C. Sander, Touring protein fold space with Dali/FSSP. Nucleic Acids Res. **26(1)**: 316-9 PubMedId: 9399863 DOI: 10.1093/nar/26.1.316, 1998.
[2] A.G. Murzin, S.E. Brenner, Hubbard, T., and C. Chotia, J. Mol. Biology., **247**, 536-540, 1994.
[3] J. Overington, M.S. Johnson, A. Sali, and T.L. Blundell, Proc. Royal Soc. Lond., **B241**, 132-145, 1990.
[4] C.A. Orengo, T.P. Flores, J.M Thomton, and W.R.Taylor , Protein Eng., **6**, 485-500, 1993.
[5] L. Holm and C. Sander, Proteins, **19**, 165-173, 1994.
[6] U. Hobohm, R. Schneider, M. Scharf, C and R. Sander, Protein Sci, **1**, 409-417, 1992.
[7] R.B. Sutton, B.A. Davletov, T.C. Sudhof, A.M. Berghuis, and S.R. Sprang, Cell, **80**, 929-935, 1995.

Author Profile

Manish Kumar is pursuing PhD in Bioinformatics, from Shri Venkateshwara University, Uttar Pradesh. He has also completed M. Sc (Bioinformatics) and B. Sc (Biosciences) from Jamia Millia Islamia University, New Delhi. He has three years of teaching and research experience. He has been earlier associated with Guru Nanak Dev University, Amritsar, in area of Computer Aided Drug Design and Sequence Analysis. He has published number of research papers in national and international journals. He has also attended number of conferences, workshops and refresher course within India. His areas of interest are Computer Aided Drug Design, Sequence Analysis and Computational & Structural Biology.