Nucleic Acids Research, 1997, Vol. 25, No. 1 231–234

Dali/FSSP classification of three-dimensional protein folds

Liisa Holm and Chris Sander

European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Received October 8, 1996; Accepted October 9, 1996

ABSRACT

The FSSP database presents a continuously updated structural classification of three-dimensional protein folds. It is derived using an automatic structure comparison program (Dali) for the all-against-all comparison of over 6000 three-dimensional coordinate sets in the Protein Data Bank (PDB). Sequence-related protein families are covered by a representative set of 813 protein chains. Hierachical clustering based on structural similarities yields a fold tree that defines 253 fold classes. For each representative protein chain, there is a database entry containing structure–structure alignments with its structural neighbours in the PDB. The database is accessible online through World Wide Web browsers and by anonymous ftp (file transfer protocol). The overview of fold space and the individual data sets provide a rich source of information for the study of both divergent and convergent aspects of molecular evolution, and define useful test sets and a standard of truth for assessing the correctness of sequence–sequence or sequence–structure alignments.

INTRODUCTION

A large fraction of newly determined genes can be classified into families based on similarities of their deduced amino acid sequences. It is then inferred, without actually performing the biochemical experiments, that the members of such families have been structurally and functionally conserved during the process of evolutionary divergence (1,2). Importantly, however, detectable global sequence similarity in a protein family is not required for retention of the three-dimensional fold, and only a very small number of invariantly conserved functional residues are required for biochemical activity. The globins, cytochrome *c*, immunoglobulin domains and NAD-binding domains are textbook examples of strongly diverged protein families for which three-dimensional, atomic-resolution structures of several members are available (3). As the detection limit of homology by standard sequence database search methods lies ~25% sequence identity, structure comparisons are required to illuminate the more distant evolutionary past. For example, 11 distantly related metal-dependent amidohydrolase families can be merged into a structurally and functionally conserved superfamily based on a seed alignment of the structures of urease, phosphotriesterase and adenosine deaminase (4).

Improved methods of protein engineering, crystallography and NMR spectroscopy have led to a surge of new three-dimensional protein structures deposited in the Protein Data Bank (PDB) (5), and a number of derived databases that organize these data into hierarchical classification schemes or in terms of structural neighbourhoods have appeared on the World Wide Web in recent years (6–10). Our FSSP database presents a fully automatic protein fold classification based on exhaustive structural alignments of known structures using the Dali method (see refs 11,12 for details). Briefly, common structural cores are delineated by optimizing the agreement of intramolecular Cα–Cα distances, i.e. using purely geometrical criteria. Each pair of structures in the PDB is related by a similarity score which is expressed in terms of statistical significance. A general overview of the spectrum of known structures is obtained by hierarchical clustering from the table of all-against-all structural similarities which leads to the definition of fold classes or subclasses at varying levels of similarity. The structural alignments of a query structure with all its structural neighbours are provided in the FSSP database for detailed analyses of individual protein families.

LATEST DEVELOPMENTS

The FSSP database continuosly monitors the flow of new structures as they are released into the public domain by the PDB. 732 PDB entries were released or updated from January 1, 1996 to the time of writing of this article (October 1, 1996), an average rate of four entries per workday. Designating this set as ‘new’, let us examine how it mapped to the two main levels of classification in the FSSP database, i.e. sequence-related protein families and structurally related fold classes. The 732 new entries contained 835 protein chains with >30 amino acids, an increase of 16% since the end of 1995. Fifteen per cent of the new structures represented a new protein family, i.e. the sequences of these proteins were <25% identical to that of any structure previously in the PDB. The 130 new families increased the coverage of protein space by 19% since the end of 1995, to a total of 813 protein families with at least one known structure. This structural

Email: surname@embl-ebi.ac.uk

knowledge can be extended to all family members for which the sequence is known (see HSSP database) (13).

Three in four new protein families were structurally similar to known fold classes, and one in four defined a new fold class (see http://www.ebi.ac.uk/dali/newfold) (12). The coverage of fold space increased by 18% to a total of 253 fold classes. The population of fold classes is highly skewed, so that one-quarter of all protein families map to the five most populated fold classes in the FSSP classification. On the other hand, 57% of all fold classes are singletons, i.e. known from only one protein family. A second protein family was added to five known singleton fold classes. Most of the new fold classes, except for four cases, are singletons. The most highly populated new fold class comprises the triplet 4-oxalocrotonate tautomerase, 5-carboxymethyl-2-hydroxymuconate isomerase and macrophage migration inhibitory factor (Fig. 1). Other rapidly expanding fold classes include SH3-like domains (DNA-binding domain of HIV-1 integrase), lysozymes (transglycosylase, chitosanase), and a group of nucleic acid binding domains with similarity to acylphosphatase (arginine repressor, DCOH, KH domain and T4 RegA). The already most populated fold classes received the largest absolute increments: immunoglobulin-like domains got five new families to a total of 34, TIM barrels got eight new families to a total of 42, and α/β domains got 14 new families to a total of 85.

FORM AND CONTENT OF THE DATABASE

Families and folds

Different protein families have very unequal representation in the PDB. For example, there are >230 structures of engineered mutants of T4 lysozyme, the folds of which are minor variations of that of the wild type. In producing the FSSP database, sequence redundancy is removed by selecting (14) a single representative for each family of homologous proteins (>25% sequence identity). The WWW page PROTEIN INDEX is a relational table that maps each of the >6000 chains in the PDB on to the closest of ~800 representatives. Precalculated all-against-all structure alignments within the representative set and structure alignments of homologs to their representative are retrievable directly (the structural alignment of any pair of structures is implicit, and may be retrieved via the Dali server). FOLD TREE is a tree representation of the sequence-representative set generated by hierarchical clustering. The tree gives a simple overview of protein families, grouping together remote homologs, with very strong structural similarity despite low sequence identity, and joining topologically similar but not necessarily evolutionarily related proteins in the lower branches. Cutting the tree at a level of $Z = 2$ (i.e. structural similarity scores two standard deviations above database average, taking domain size into acccount) yields 253 fold classes.

Structural alignments

For each protein chain in the representative set, with PDB identifier Nxxx (like: 1PPT, 5PCY) and chain identifier Y (omitted if blank), there is an ASCII (text) file Nxxx.FSSP or NxxxY.FSSP which contains the alignments of a few or tens of proteins structurally similar to the search structure. The structural neighbours that are reported include any sequence homologs to the query structure that have a structure in the PDB, and all structurally similar chains from the representative set ($Z \geq 2$). The current database contains at most one alignment per pair of full length proteins. The alignments are constrained to be sequential (i.e. reconnections of loops are disallowed) as this is biologically meaningful, although not imposed by the Dali method (11).

An FSSP file is divided in five formatted blocks and a free text footer which explains the format: (i) the header block identifies the query structure, database and structural alignment method used and gives the number of structural neighbours; (ii) the summary block gives a one-line summary for each neighbour, including the statistical significance of the similarity (Z score), positional root-mean-square deviation of superimposed CA coordinates, total number of equivalent residues, and the percentage of sequence identity over structurally equivalent positions; (iii) the alignments block is a multiple structural alignment, printed vertically and showing the sequence and secondary structure of matched residues; (iv) the equivalences block is a machine readable listing that gives the residue numbers of the structurally equivalent segments; (v) the matrices block gives the rotation–translation matrices that, when applied to the three-dimensional coordinates in the respective PDB entries, yield the least-squares superimposition of the matched protein on to the query structure.

World Wide Web interface

There are currently two starting points for a walk in fold space, either by querying the PROTEIN INDEX for protein name or PDB identifier, or via clicking the ‘alignment’ link in the FOLDTREE table. FSSP entries are parsed on the fly to display selected alignments in one and three dimensions (Fig. 1). In one dimension, multiple sequence alignments are phased on structure–structure comparison and can be combined with sequence neighbours (sequences homologous to a known structure: HSSP database) (13) and viewed with the Belvu program (Sonnhammer E., unpublished). In three dimensions, superimposed coordinates can be viewed with molecular graphics programs such as Rasmol (15). There are further hypertext links to other structural classifications (7–9) via the PDB identifier and to functional annotations and literature references via Swissprot sequence identifiers (16).

Distribution

The FSSP database is accessible addressing URL http://www.embl-heidelberg.de/dali/fssp/ or by anonymous ftp (file transfer protocol) from ftp.embl-ebi.ac.uk in the directory /pub/databases/fssp. Users are asked to refer to ref. (11) and this paper in reporting results obtained using the database.

Size of the current release

The size of the FSSP database is tightly coupled to that of the PDB from which it is derived. The database classifies 6001 protein chains into 813 protein families and 253 fold classes (status at October 1, 1996). The complete set of result files requires ~140 Mb of disk storage.

a

```
81.1.1.1.1.1  1eaf        DIHYDROLIPOYL TRANSACETYLASE (E.C.2.3.1.12) (CA
81.1.1.1.1.1  ______3cla  TYPE /III$ CHLORAMPHENICOL ACETYLTRANSFERASE (/
82.1.1.1.1.1  1otfA       4-OXALOCROTONATE TAUTOMERASE
82.1.2.1.1.1  __1otgA     5-CARBOXYMETHYL-2-HYDROXYMUCONATE ISOMERASE
82.1.2.1.2.1  ____1fim    MACROPHAGE MIGRATION INHIBITORY FACTOR
83.1.1.1.1.1  1mrj        ALPHA-TRICOSANTHIN COMPLEXED WITH ADENINE
```

b

```
structural alignment of 1otgA + 1otfA + 1fim; homologs with < 80 % sequence identity added

1_hpcd_ecoli_1otgA     1 PHRIVECSDNIREEADLPGLFAKVNPTLAATGIPLRAGTRSRVHWVDTWQMADGQHDYASVHM 63
2_dmpi_psepu_1otfA     1 PIAQLYIIE.GRTD.QKETLIREVSEAMANSLDAPLERVRVLLTEMPHFGIGGE......... 52
3_xylh_psepu_1otfA     1 PIAQIHILE.GRSD.QKETLIREVSEAISRSLDAPLTSVRVIITEMAHFGIGGE......... 52
4_mif_rat_1fim         1 PMFIVNTNVPSVPE.FFSELTQQLAQATG....KPAQYIAVHVVPDQLMTFSGT...DPCALC 55
8_mif_chick_1fim       1 PMFTIHTNVCAVPD.LLGELTQQLAKATG....KPAQYIAVHIVPDQMMSFGGS...DPCALC 55

1_hpcd_ecoli_1otgA    64 TFKIGAGRSLESRQQAGEMLFFIIKTHFAALMESRLLALSFEIEELHPTLNFKQNNVHALFK 125
4_mif_rat_1fim        56 SFHSIG......GGAQNRNYSKLLCGLSDRLH.....RVYINYYDMNA.............. 92
8_mif_chick_1fim      56 SFYSIG......GGQQNKTYTKFLCDMAKHTH.....RVYINYFDINA.............. 92
```

c

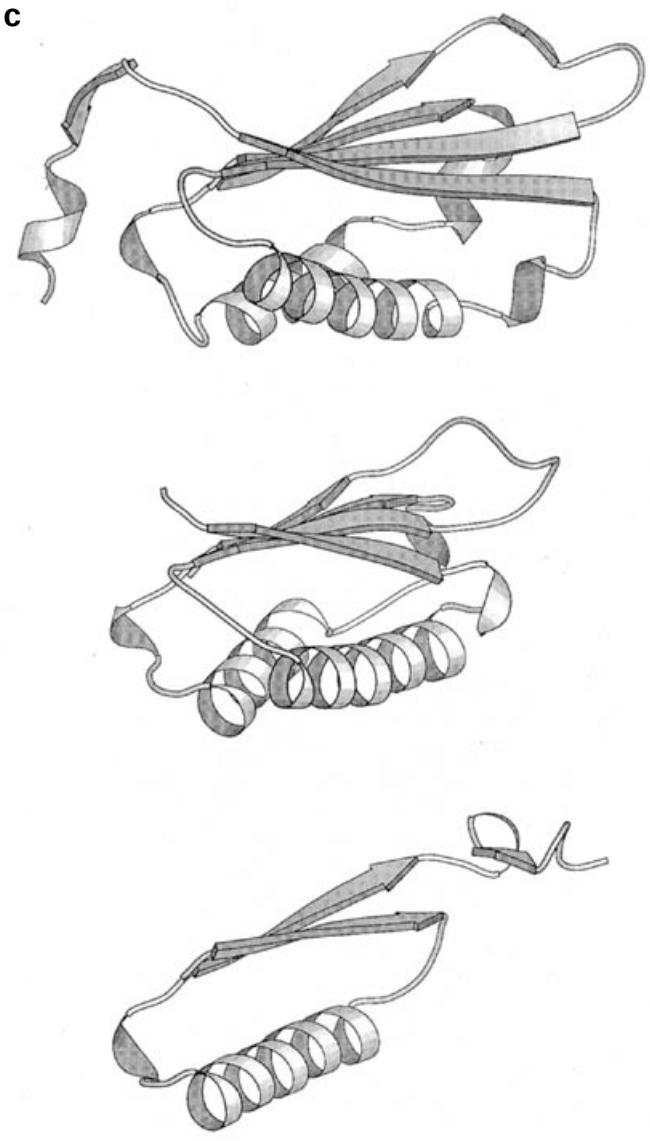

Figure 1. Structural comparison of a cytokine and two kinds of isomerase. The determination of the crystal structure of macrophage inhibitor factor (1fim) revealed unexpected similarities to two kinds of isomerase, i.e. 4-oxalocrotonate tautomerase (1otfA) and 5-carboxymethyl-2-hydroxymuconate isomerase (1otgA) (17). (**a**) Section of the fold classification that groups the three structures in a distinct class (82.*). The indexing tells you that 1otgA and 1fim are in the same subclass with similarity $5<Z<10$, and the similarity of 1otfA to them is $3<Z<4$. No other group of proteins is significantly similar ($Z>2$) to the triplet. (**b**) Combining structure–structure alignment with multiple sequence–sequence alignment (13). The PDB identifier of the known structure is at the right. (**c**) The three-dimensional structures viewed in equivalent orientations (top: 1otgA; middle: 1fim; bottom: 1otfA). 1otfA is a βαβ unit, which is repeated twice in 1otgA and 1fim. Plotted with Molscript (18).

Related services

The Dali server is a computational service for structure comparison using the same machinery used in deriving the FSSP database. Requests for database searches with newly solved crystallographic or solution NMR structures (C^{α} coordinates required) may be sent by e-mail to dali@embl-heidelberg.de (Internet address) or submitted interactively from the World Wide Web addressing http://www.embl-heidelberg.de/dali

New features include preparation of pairwise comparisons and defining coordinate sets by PDB identifier. For example, one may request structural alignments of trypsins relative to the rat enzyme (1brcE) rather than relative to the family representative from the mold *Fusarium oxysporum*.

Please report any problems to the authors by electronic mail.

REFERENCES

1 Scharf M. *et al*. (1994) *ISMB* **2,** 348–353.
2 Bairoch A. and Boeckmann B. (1992) *Nucleic Acids Res.* **20**, 2019–2022.
3 Brändn C-I. and Tooze J. (1991) *Introduction to Protein Structure*. Garland Publishing Inc., New York and London.
4 Holm L. and Sander C. (1997) *Proteins*, in press.
5 Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T. and Tasumi M. (1977) *J. Mol. Biol.* **112**, 535–542.
6 Holm L. and Sander C. (1996) *Nucleic Acids Res.* **24**, 206–210.
7 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
8 Orengo C.A., Flores T.P., Taylor W.R. and Thornton J.M. (1993) *Protein Eng.* **6**, 485–500.
9 Gibrat J-F., Madej T. and Bryant S.H. (1996) *Curr. Opin. Struct. Biol.* **6**, 377–385.
10 Islam S.A., Luo J. and Sternberg M.J.E. (1995) *Protein Eng.* **8**, 513–525.
11 Holm L. and Sander C. (1993) *J. Mol. Biol.* **233**, 123–138.
12 Holm L. and Sander C. (1996) *Science* **273**, 595–602.
13 Sander C. and Schneider R. (1991) *Proteins* **9**, 56–68.
14 Hobohm U., Scharf M., Schneider R. and Sander C. (1992) *Protein Sci.* **3**, 409–417.
15 Sayle R.A. and Milner-White E.J. (1995) *Trends Biochem. Sci.* **20**, 374–376.
16 Bairoch A. (1992) *Nucleic Acids Res.* **20**, 2013–2018.
17 Suzuki M., Sugimoto H., Nakagawa A., Tenaka I., Nishihira J. and Sakai M. (1996) *Nature Struct. Biol.* **3**, 259–266.
18 Kraulis P.J. (1991) *Appl. Crystallogr.* **24**, 946–950.