*316–319* *Nucleic Acids Research, 1998, Vol. 26, No. 1* 

# Touring protein fold space with Dali/FSSP

**Liisa Holm and Chris Sander**

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

Received October 10, 1997; Accepted October 15, 1997

## ABSTRACT

**The FSSP database and its new supplement, the Dali Domain Dictionary, present a continuously updated classification of all known 3D protein structures. The classification is derived using an automatic structure alignment program (Dali) for the all-against-all comparison of structures in the Protein Data Bank. From the resulting enumeration of structural neighbours (which form a surprisingly continuous distribution in fold space) we derive a discrete fold classification in three steps: (i) sequence-related families are covered by a representative set of protein chains; (ii) protein chains are decomposed into structural domains based on the recurrence of structural motifs; (iii) folds are defined as tight clusters of domains in fold space. The fold classification, domain definitions and test sets for sequence-structure alignment (threading) are accessible on the web at www.embl-ebi.ac.uk/dali . The web interface provides a rich network of links between neighbours in fold space, between domains and proteins, and between structures and sequences leading, for example, to a database of explicit multiple alignments of protein families in the twilight zone of sequence similarity. The Dali/FSSP organization of protein structures provides a map of the currently known regions of the protein universe that is useful for the analysis of folding principles, for the evolutionary unification of protein families and for maximizing the information return from experimental structure determination.**

## INTRODUCTION

The number of three-dimensional protein structures in the Protein Data Bank (PDB; 1) has been doubling approximately every 18 months. This acceleration means that automatic methods are increasingly important for efforts to organize the data. The FSSP database (2), established in 1992, and its new supplement, the Dali Domain Dictionary, are produced using the Dali program for structural alignment (3) to automatically and continuously process the new structures released by the Protein Data Bank (Fig. 1). The information derived as a result includes the description of protein domain architecture, the definition of structural neighbours around each known structure, the definition of structurally conserved cores and explicit multiple alignments of distantly related protein families; these are made available on the web.

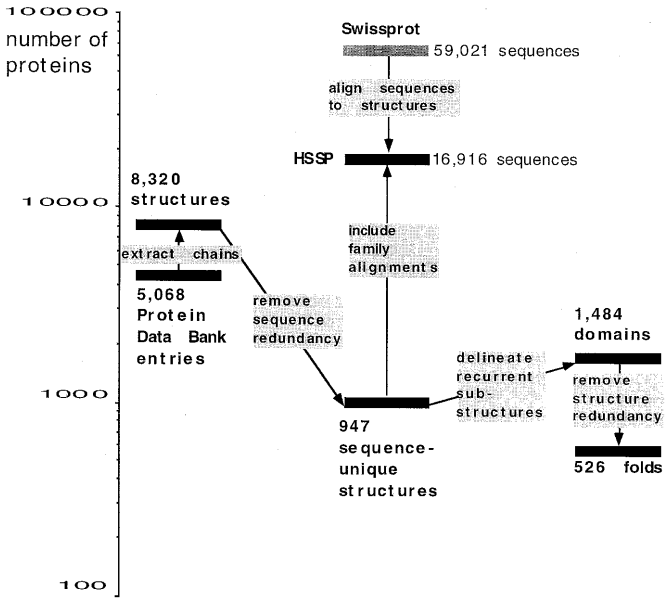


**Figure 1.** Flowchart of the processing of protein structures in PDB. The high redundancy of biological databases presents a number of problems in practical use. To overcome these problems, it is useful and essential to derive representative subsets and/or classify the data. Our structural classification starts from extracting all structures (chains) from the PDB (left). Based on all-on-all structure comparison, we define a representative set of structures which is free of sequence redundancy (middle bottom). Each structure is decomposed into domains (upper right). Folds are defined by clustering domains based on structural similarities. As a result, all known protein structures can be completely described in terms of 526 fold types (bottom right; the numbers refer to April 1997). The arrows in the middle column put the fold classification in context with the world of sequence analysis via the HSSP database of structure-sequence alignments (15). About one quarter of all sequences in the SWISS-PROT database (13) are clearly homologous to proteins of known structure.

There are a number of other classification schemes for protein structures available on the web. Although they are based on the same data, the presentations differ in their basic philosophy regarding automation and organization (4–9). For example, MMDB from NCBI (US National Center for Biotechnology Information) provides a fish-eye view of structural neighbours around any PDB structure based on precalculated all-on-all

Email: holm@embl-ebi.ac.uk

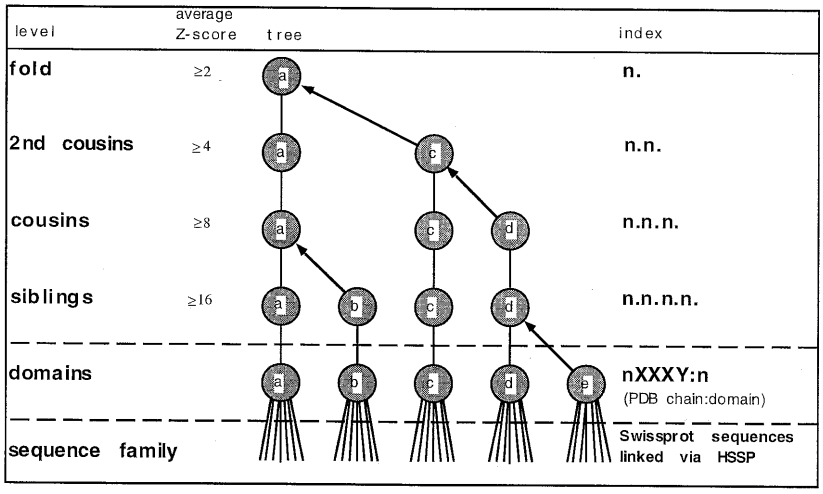


**Figure 2.** Hierarchical clustering of folds. Hierarchical clustering yields a convenient view (dendrogram) of fold neighbours at different level of structural similarity (Z-scores). In this example, five domains (columns a–e) belong to the same fold class. Based on the topology of the dendrogram, domains d and e are siblings (same parent node), domains a and b are cousins (same grandparent node), domains c and d are second cousins (same greatgrandparent node), and so on. To ease navigation, the user is presented with a uniform summary for each node in the dendrogram. The idea is to choose a central member of the cluster as a representative (3D template) onto which structural or sequence variability can be mapped based on the multiple alignment of cluster members. For example, domain a represents the whole class {a,b,c,d,e}, and the link d→c means that domain c is used to represents the set {c,d,e}. The fold...domain levels are based on structure similarity. Sequence families around proteins of known structure (bottom row) are defined by sequence similarity (14). Exploiting links involving structure alignments leads to accurate multiple alignments of distantly related protein families. Currently, the naming of structural similarity levels is not a statement about evolutionary relationships. However, we regularly observe that remote relatives are more similar to each other than to other proteins in the database, so in favourable cases examination of the fold dendrogram can lead to biological discoveries. For example, {a,b} and {d,e} including their associated sequence families are likely candidates for unification into a functionally conserved superfamily.

structure comparisons using the VAST algorithm (4). Scop (5) and CATH (6) are strictly hierarchical classifications based on the abstractions of class (4–10 categories at the top of the hierarchy), architecture/topology or fold, and superfamily (519 in scop). Both classifications are curated by experts, with emphasis in scop on the definition of functionally related superfamilies and in CATH on the definition of architectural types. Dali/FSSP is a fully automatic classification based on the concept of neighbourhoods in fold space, of which it aims to provide useful views at both coarse-grained and fine-grained resolution. In the near neighbour range, the quantitative structural relationships between domains are described in terms of hierarchical clustering (dendrograms, similar to scop and CATH) and in terms of neighbour lists (similar to VAST). In recognition of the continuous rather than discrete distribution of domains in fold space, the global overview of structural relationships between domains is presented in terms of 2D 'roadmaps' of fold space. At all levels, representative sets are used for clarity, removing obvious redundancy of information. Many of the finer branches of the fold dendrograms correspond to evolutionarily related, functionally conserved superfamilies. We are currently developing tools for automatically annotating functional evidence of plausible evolutionary relationships (10).

The structural classification is explicitly linked (11) to sequence families with associated functional annotation, resulting in a rich network of biologically interesting relationships that can be browsed online. In particular, structure-based alignments increase our understanding of the more distant evolutionary relationships. For example, the discovery of remarkable structural similarity between histidine triad (HIT) proteins and galactose-1-phosphate uridylyltransferase (GalT) pointed to a conserved biochemical function in an emerging superfamily (12). The interconnection of structural classification with sequence families also opens the door to studies of structure–sequence–function relationships from a global perspective, for example: 'which folds support function X?', 'which functions have evolved on the framework of fold Y?', 'do protein families in region Z of fold space diverge faster/more slowly than average?'.

## FORM AND CONTENT OF THE DATABASE

The Protein Data Bank (PDB) is highly redundant in terms of sequence and structure similarities. Our aim is complete and economical description of structural data (Fig. 1). The first reduction step is the generation of a sequence-unique set. No pair of proteins within this set is more than 25% identical in sequence and all removed structures are more than 25% identical with a representative. To avoid the removal of unique domains next to more common domains, the percentage used here is calculated as the number of residue identities in the structurally aligned region, divided by the average length of the two proteins (not by the length of the aligned region). The second step is to describe the structural neighbourhood around each sequence-unique representative chain, in the form of structural alignments. The FSSP database (DaliFSSP) has one entry per representative, reporting the structural alignments with the representative's sequence homologs (same family, membership detectable by sequence methods) and with other members of the representative set (related families, relationship difficult or impossible to detect by sequence methods). The Dali Domain Dictionary (DaliDD) is a new complement to the FSSP database that has the same format but one entry per structural

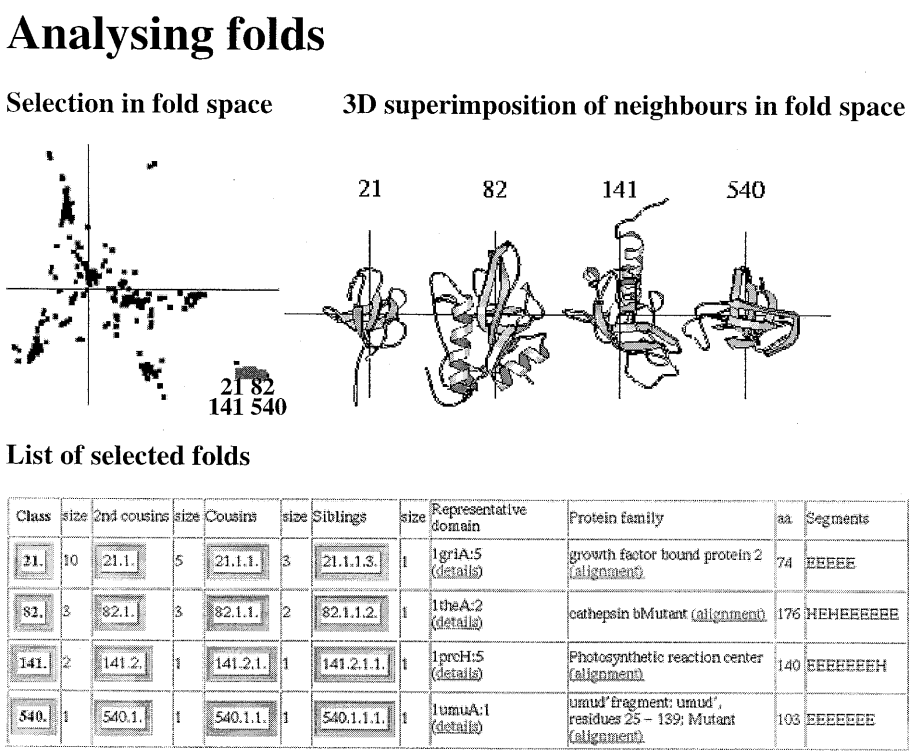


**Figure 3.** Touring fold space.The dictionary is based on the quantification of structural similarities by all-on-all comparison of known structures. Using the pairwise similarities, each structure can be positioned in an abstract high-dimensional fold space. The overall distribution of domains into general architectural types is visualized using 2D projections of fold space ('roadmaps') generated by multivariate scaling methods (3). Within fold space, there are tight clusters of domains that have the same fold, i.e., similar overall arrangement of secondary structure elements. The structural relationships between instances (member domains) of a fold are visualized using dendrograms (explained in Fig. 2). The WWW interface allows the database of structural neighbours to be queried in a variety of ways with dynamic views generated on the fly. In this example, clicking in the lower right corner of the 2D map (top left) leads to a table view (middle) of folds occupying this region of fold space. Click on 'details' for a representative domain to identify structural neighbours that form bridges between the fold clusters and can be used for 3D superimposition. In this case, superimposition reveals a shared motif consisting of two crossed β-hairpins (upper right, the numbers above the ribbon diagrams refer to fold class). To analyse a fold cluster in more detail, the user can expand or contract the fold tree (click on a node, e.g., 21.1.1.) and invoke different graphical views of selected subsets that highlight conserved sequence features and structural elements (bottom).

domain. In other words, DaliFSSP is about proteins, or protein chains, while DaliDD is about structural domains.

For many types of analysis, it is useful to work within a discrete classification framework, although the data does not easily lend itself to disjoint clustering. To produce a discrete classification of domains, the all-on-all structure comparison is used to derive a fold tree (dendrogram) by a simple hierarchical clustering procedure using average linkage. Folds are then defined by cutting the fold tree at an empirically chosen cutoff such that most secondary structure elements are structurally equivalent between members of a cluster, i.e, they have the same fold. To ease navigation, subclusters that group together domains with similarities of architectural detail are obtained by cutting the tree at higher levels of structural similarity (Fig. 2).

The distribution of representative structures in folds is highly uneven. The largest fold has >100 member domains, and the four dominant folds [αβ domains, immunoglobulin-like domains, $(\alpha\beta)_8$ barrels, helical bundles] comprise one quarter of the number of secondary structure elements in the representative set. For book-keeping purposes, we have chosen to index folds in

order of decreasing population; these indices have no intrinsic meaning and may change as more structures are solved.

## USES OF THE DATABASE

The web service provides graphical and tabular views of the data so that the user can take a tour of fold space while sitting and clicking (Fig. 3). A tour of fold space can start from a region of fold space seen in 2D projection, from a structure selected automatically at random, from a node in the fold dendrograms, or from a string (text) search in structure or sequence databases (13–15). Hyperlinks connect structures to structural neighbours allowing 'walking' through neighborhoods of structural motifs.

Strong structural similarity despite low overall sequence similarity hints at a possible distant evolutionary relationship. The web server provides powerful tools for analysing superfamilies because the structural alignments are linked with protein families and functional annotation in sequence databases. Particularly informative (and rarely available) are the explicit multiple alignments of distantly related representatives with their sequence neighbours which often reveal a signature of invariantly conserved residues. Although such invariant residues may be widely dispersed along the 1D sequence, mapping these residues onto a structural template typically shows that they cluster together in 3D to form an active site (16). Such sets of residues are an excellent starting point for the crafting of far-reaching search profiles.

In the context of fold recognition, the structural classification thus leads to sequence models (profiles) that more accurately model the evolutionary variation within a superfamily, provides core templates with information about structurally conserved or variable parts, and reduces the size of the target structure database. See http://www2.embl-ebi.ac.uk/dali/testset for proposed test sets.

## DISTRIBUTION

The FSSP database and Dali Domain Dictionary are accessible at http://www.embl-ebi.ac.uk/dali and by anonymous ftp (file transfer protocol) from ftp.embl-ebi.ac.uk in the directory /pub/databases/fssp. The complete set of database files requires ~140 Mb of disk storage. The web browser script is available for sites wishing to mirror the server [local installation of the HSSP (15) and PDB databases is also required].



## RELATED SERVICES

The Dali server (3) is the 'BLAST server' of protein 3D structures. Dali performs a database similarity search of a new structure solved by crystallography or NMR against the 3D co-ordinates of structures in the Protein Data Bank. Requests must contain at least the $C^{\alpha}$ co-ordinates of the new structure and may be sent by e-mail to dali@embl-ebi.ac.uk or submitted interactively through http://www.embl-ebi.ac.uk/dali . Please report any problems to the authors by electronic mail.

## REFERENCES

1 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.
2 Holm,L, Ouzounis,C., Sander,C., Tuparev,G. and Vriend,G. (1992) *Protein Sci.* **1**, 1691–1698.
3 Holm,L. and Sander,C. (1996) *Science* **273**, 595–602.
4 Gibrat,J.-F., Madej,T. and Bryant,S.H. (1996) *Curr. Opin. Struct. Biol.* **6**, 377–385.
5 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.* **247**, 536–540.
6 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure* **5**, 1093–1108.
7 Islam,S.A., Luo,J. and Sternberg,M.J.E. (1995) *Protein Engng* **8**, 513–525.
8 Siddiqui,A.S. and Barton,G.J. (1995) *Protein Sci.* **4**, 872–884.
9 Sowdhamini,R., Rufino,S.D. and Blundell,T.L. (1996) *Fold Des.* **1**, 209–220.
10 Holm,L. and Sander,C. (1997) *ISMB* **5**, 140–146.
11 Etzold,T. Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.* **266**, 114–128.
12 Holm,L. and Sander,C. (1997) *Trends Biochem. Sci.* **22**, 116–117.
13 Bairoch,A. (1992) *Nucleic Acids Res.* **20**, 2013–2018.
14 Sander,C. and Schneider,R. (1991) *Proteins* **9**, 56–68.
15 Schneider,R., de Daruvar,A. and Sander,C. (1997) *Nucleic Acids Res.* **25**, 226–230 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 313–315].
16 Holm,L. and Sander,C. (1997) *Proteins* **28**, 72–82.