244–247 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

Protein folds and families: sequence and structure alignments

Liisa Holm* and Chris Sander[1]

European Bioinformatics Institute, EMBL-EBI, Genome Campus, Cambridge CB10 1SD, UK and
[1]Whitehead Institute, MIT Genome Center, Cambridge, MA 02138, USA

Received October 29, 1998; Accepted October 30, 1998

ABSTRACT

Dali and HSSP are derived databases organizing protein space in the structurally known regions. We use an automatic structure alignment program (Dali) for the classification of all known 3D structures based on all-against-all comparison of 3D structures in the Protein Data Bank. The HSSP database associates 1D sequences with known 3D structures using a position-weighted dynamic programming method for sequence profile alignment (MaxHom). As a result, the HSSP database not only provides aligned sequence families, but also implies secondary and tertiary structures covering 36% of all sequences in Swiss-Prot. The structure classification by Dali and the sequence families in HSSP can be browsed jointly from a web interface providing a rich network of links between neighbours in fold space, between domains and proteins, and between structures and sequences. In particular, this results in a database of explicit multiple alignments of protein families in the twilight zone of sequence similarity. The organization of protein structures and families 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. The databases are available from http://www.embl-ebi.ac.uk/dali/

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. We use the Dali program for structural alignment (2–4) to automatically and continuously process the new structures released by the PDB (Fig. 1). The information derived as a result in the Dali databases (5) 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.

Complementing Dali, which detects distant evolutionary relationships based on a comparison of the 3D chain traces, HSSP (6) collects sequence families around known 3D structures based on comparison of 1D sequences. Though the range of detection of biologically interesting similarities is much shorter by sequence methods, sequence databases are 2–3 orders of magnitude larger than the dataset of known structures. If the 3D structure of only one family member is known, then by implication one can derive the basic 3D structure, or fold, of all family members using model building by homology. To exploit this principle, the HSSP database is generated by aligning, for each protein of known 3D structure in the PDB (1), all its likely sequence homologues. As a result, HSSP is not only a database of aligned sequence families, but also a database of implied secondary and tertiary structures. Likely secondary structures can be carried over directly from the PDB protein to each homologue. Tertiary models can be built by fitting the sequence of the homologue, as aligned, into the 3D template of the protein of known structures (sequence inserts, however, are very difficult to model in 3D). Relative to experimentally derived information in PDB, HSSP increases the number of effectively known protein structures several fold.

The joint fold/family classification is available on the web. 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. For example, the structural alignment of 11 distantly related myoglobins, hemoglobins and plant leghemoglobins (aligned by Dali) can be expanded with hundreds of sequence neighbours (aligned by HSSP) revealing a structurally conserved core with a heme binding pocket and invariant histidines that are required for heme binding. The capability of merging sequence and structure alignments is useful for analyzing residue conservation in structural context, for defining structurally meaningful sequence patterns, for deriving test sets for fold recognition and, in general, for studying protein evolution, folding and design.

CONTENT AND FORMAT OF THE DATABANKS

Dali: fold space organized in 3D

The PDB is highly redundant in terms of sequence and structure similarities. We use the Dali algorithm for all-on-all comparison

*To whom correspondence should be addressed. Tel: +44 1223 494454; Fax: +44 1223 494471; Email: holm@embl-ebi.ac.uk

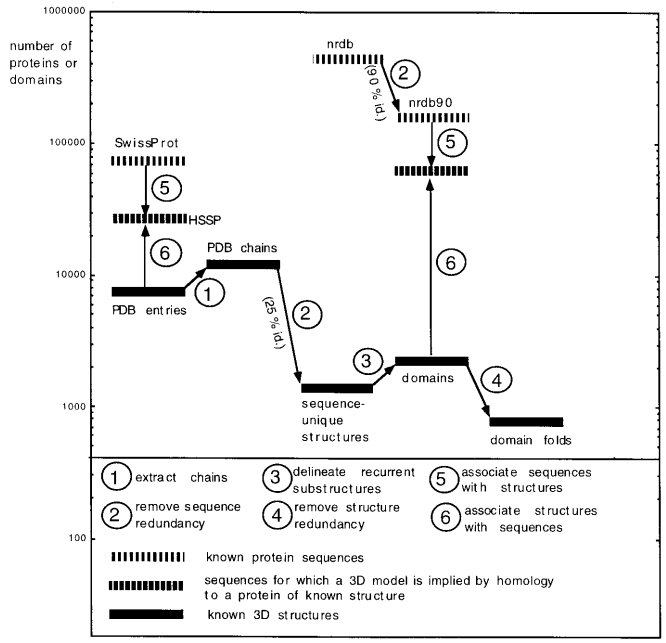

Figure 1. Organizing protein structure space. The high redundancy of biological databases presents a number of problems in practical use. To overcome these problems, it is useful and essential to derive representative subsets and/or classify the data. The lower part of the graph illustrates the reduction steps in the Dali structural classification based on all-on-all structure comparison. Starting from 7373 PDB entries, a representative set of structures which is free of sequence redundancy comprises 1378 chains (bottom middle). Domain folds represent domains, i.e., a level of clustering that groups together compact substructures with similar packing and topological arrangement of secondary structure elements (4). In September 1998, all known protein structures were completely described in terms of 771 fold types (bottom right). The upper part of the graph connects the organization of structure space to the space of all known sequences. The vertical arrows represent sequence–structure associations implied by multiple sequence alignment of putative homologues. For example, nearly 40% of all sequences in nrdb90 (a non-redundant sequence database) are putative homologues to proteins of known structure. Protein families are centred on each PDB entry (HSSP database) or on each structural domain in a representative subset (DaliDD) leading to a database of explicit multiple alignments of protein families in the twilight zone of sequence similarity.

of structures in the PDB, with the aim of a 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 Dali database has two subdivisions. DaliFSSP has one entry per representative PDB structure, and reports the structural alignments with other members of the representative set (related families, relationship difficult or impossible to detect by sequence methods) and with homologous PDB structures (same family, membership detectable by sequence methods). The Dali Domain Dictionary (DaliDD) has the same format but one entry per structural domain. In other words, DaliFSSP is about proteins, or protein chains, while DaliDD is about structural domains.

To provide useful overviews of neighbourhoods in fold space at both coarse-grained and fine-grained resolution, the quantitative structural relationships between domains in the near neighbour range are described in terms of hierarchical clustering (dendrograms, similar to popular fold classifications; 7–9). 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 ‘road maps’ of fold space. At all levels, representative sets are used for clarity, removing obvious redundancy of information.

HSSP: sequence families centred on known 3D structures

HSSP (homology-derived structures of proteins) groups sequence-similar proteins into families of structural homogues that merges information from 3D structures and 1D sequences of proteins. Each HSSP family is centred on a known 3D structure. Thus, for each protein in PDB, with identifier xxxx (such as: 1PPT, 5PCY), there is an ASCII (text) file xxxx.HSSP which contains: (i) the sequence of the protein of known structure, along with the derived secondary structure and solvent accessibility calculated from the coordinates using DSSP (10); (ii) aligned sequences of a few or tens or hundreds of sequences from the Swiss-Prot database deemed structurally homologous to this protein; (iii) sequence variability, using two different measures, at each position in the multiple sequence alignment; and (iv) occupancy, i.e., the number of sequences that span this position. Alignments are produced using a position-weighted dynamic programming method for sequence profile alignment (Maxhom), and likely homologues are selected applying a well-tested threshold for structural homology. Currently, the threshold is in terms of percent of identical residues; we plan to switch to a more discriminative threshold in the future. Some details of the methods are given elsewhere (11).

Separately from HSSP but conceptually similar, DaliDD also provides sequence families around the representative domain structures. Putative sequence homologues are collected using the PSI-BLAST program (12) with the following parameters: 10 iterations of profile searching and an expectation value cutoff of 1e–5 (J.Park, personal communication). Low-complexity regions are masked in the search database. Patterns of conserved residues are highlighted in the web display (13) of the resulting multiple alignments.

DISTRIBUTION

Anonymous ftp

The databases can be obtained over the Internet by anonymous ftp (File Transfer Protocol) from ftp.embl-ebi.ac.uk in directory /pub/databases/[fh]ssp or using a World Wide Web browser from ftp://ftp.embl-ebi.ac.uk/pub/databases/

World Wide Web (WWW)

The DaliFSSP database and the Dali Domain Dictionary can be browsed from http://www2.embl-ebi.ac.uk/dali/fssp/ and http://www2.embl-ebi.ac.uk/dali/domain/2.0/ , respectively. The web browser script is available for sites wishing to mirror the server (local installation of the HSSP and PDB databases is also required).

The HSSP database and HSSP-related information and data are accessible from http://www.sander.embl-ebi.ac.uk/hssp

The programs that generate the alignments (MaxHom for sequences and Dali for 3D structures) are currently not available for distribution, but are accessible via the PredictProtein and Dali servers (see below).

Conditions

Academic redistribution of single files or the entire databases is permitted, provided the dataset integrity is strictly maintained. No inclusion in other databases or datasets, academic or other, without explicit permission of the authors. All commercial rights reserved. Not to be used for classified research. Users are asked to refer to this paper and references 3 and 11 in reporting results based on the use of the databases.

SIZE OF THE CURRENT RELEASE

The content and size of the derived databases is of course tightly coupled to the development of the source databases of protein 3D structures (PDB) and sequences (e.g., Swiss-Prot; 14). An overview of the increase in size of PDB, HSSP and Swiss-Prot is given in Table 1. Interestingly, ~26 000 of 74 000 known sequences (Swiss-Prot relase 35) are putative homologues of known structures and therefore have an implied known 3D structure.

Table 1.

HSSP release (month / year)	No of HSSP data sets	No of SwissProt entries	Total no of alignments in the HSSP database	No of aligned sequences in the HSSP database	Fraction of SwissProt in the HSSP database [%]
5/91	488	20,024	37,715	3,065	15.3
9/92	736	25,044	49,784	4,825	19.2
10/93	1,532	31,088	123,810	7,642	24.0
8/94	2,158	38,303	154,590	10,136	26.5
8/95	3,158	43,470	241,518	11,762	27.0
9/96	4,189	52,205	317,213	15,140	29.0
10/97	5,745	59,021	485,527	20,025	33.9
9/98	7,373	74,019	689,880	26,370	35.6

The complete sets of data files of HSSP currently require ~900 Mb and those of FSSP ~300 Mb of disk storage.

LIMITATIONS

Accuracy of reported alignments

In general, alignments based on sequence may deviate from alignments based on comparison of known 3D structures in local detail, especially in terms of placement of gaps. In these cases, the sequence alignment may correctly represent conservation in the evolutionary chain of events connecting the two sequences while structural alignment may reflect a local structural rearrangement as a result of mutations in sequence positions spatially near the conserved residues. Alignments, whether based on sequences or structures, are often uncertain in loop regions.

Definition of variability

In using sequence variability scores reported in HSSP, the user should be aware that low occupancy positions (few alignments span that position) have ill-determined variability values; in the limit of zero occupancy, the variability is undefined and set to zero. For some purposes, the user may choose to use only positions with occupancy larger than, say, five proteins.

Multichain biological units

In some cases, a biologically functional unit is divided over several chains in the PDB entry but we currently treat each chain individually.

Sequence redundancy

The problem of redundancy in biological databases is currently not addressed in the HSSP database. For example, there are separate files for hemoglobin and myoglobin, which have ~30–35% identical residues, so that proteins homologous to both hemoglobin and myoglobin appear in both files. Sequence-identical chains in the PDB entry are removed so that the xxxx.HSSP files only contain sequence-unique chains.

RELATED DATABASES AND INFORMATION SERVICES

The following database and information services are also available from the Holm and former Sander groups at EMBL-EBI.

Dali. An electronic mail/WWW server that performs a 3D similarity search in the PDB, given the atomic coordinates of a 3D protein model as input (3). The analog of a BLAST server for 3D structures. http://www2.embl-ebi.ac.uk/dali/

DSSP. A database of secondary structure, solvent accessibility and other information derived from 3D structures in the PDB. http://www.sander.embl-ebi.ac.uk/dssp/ ; personal Email: sander@embl-ebi.ac.uk

GPCRDB. Information system for G-protein coupled receptors. http://swift.embl-heidelberg.de/7tm/ ; personal Email: vriend@embl-heidelberg.de

nrdb90. A non-redundant sequence database that removes sequences at a redundancy level of 90% amino acid identity (15). Weekly updates. http://www.embl-ebi.ac.uk/~holm/nrdb90

PredictProtein. An electronic mail server that provides a predicted secondary structure and solvent accessibility profile for any protein sequence with homologues in Swiss-Prot. Rated at 72% sustained three-state accuracy (16). http://www.embl-heidelberg.de/predictprotein/ ; personal Email: rost@embl-heidelberg.de

Special software is available to construct 3D models by homology based on the information in HSSP files, such as *WhatIf* (17) or *MaxSprout/Torso* (18).

Please report any problems to the authors by Email: dali-help@embl-ebi.ac.uk

REFERENCES

1 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyers,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. and Sander,C. (1992) *Proteins*, **14**, 213–223.
3 Holm,L. and Sander,C. (1996) *Science*, **273**, 595–602.
4 Holm,L. and Sander,C. (1998) *Proteins*, **33**, 88–96.
5 Holm,L. and Sander,C. (1998) *Nucleic Acids Res.*, **26**, 316–319.
6 Dodge,C., Schneider,R. and Sander,C. (1998) *Nucleic Acids Res.*, **26**, 313–315.
7 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
8 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
9 Sowdhamini,R., Burke,D.F., Huang,J.F., Mizuguchi,K., Nagarajaram,H.A., Srinivasan,N., Steward,R.E. and Blundell,T.L. (1998) *Structure*, **6**, 1087–1094.
10 Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
11 Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
12 Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
13 Brown,N., Leroy,C. and Sander,C. (1998) *Bioinformatics*, **14**, 380–381.
14 Bairoch,A. (1992) *Nucleic Acids Res.*, **20**, 2013–2018.
15 Holm,L. and Sander,C. (1998) *Bioinformatics*, **14**, 423–429.
16 Rost,B., Schneider,R. and Sander,C. (1993) *Trends Biochem. Sci.*, **18**, 120–123.
17 Vriend,G. (1990) *J. Mol. Graphics*, **8**, 52–56.
18 Holm,L. and Sander,C. (1997) *Proteins*, **28**,72–82.