The FSSP database: fold classification based on structure–structure alignment of proteins

Liisa Holm and Chris Sander

European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

Received October 2, 1995; Accepted October 4, 1995

ABSTRACT

The FSSP database presents a continuously updated classification of 3-D protein folds based on an all-against-all comparison of structures currently in the Protein Data Bank (PDB) [Bernstein *et al.* (1977) *J. Mol. Biol.*, 112, 535–542]. The database currently contains an extended structural family for each of 600 representative protein chains which have <25% mutual sequence identity. The results of the exhaustive pairwise structure comparisons are reported in the form of a fold tree generated by hierachical clustering and as a series of structurally representative sets of folds at varying levels of uniqueness. For each query structure from the representative set, there is a database entry containing structure–structure alignments with its structural neighbours in the representative set and its sequence homologs in the PDB. All alignments are based purely on the 3-D co-ordinates of the proteins and are derived by an automatic structure comparison program (Dali). The FSSP database is accessible electronically on the World Wide Web and by anonymous ftp.

INTRODUCTION

Most newly determined protein sequences can be classified into families by sequence homology. However, 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. These similarities can be detected by structural comparisons that merge protein families of known 3-D structure into structural classes, the members of which may or may not be evolutionarily related (1–4). The FSSP database contains a fold classification based on exhaustive structural alignments of known structures. The database provides a rich source of information for the study of both divergent and convergent aspects of the evolution of protein folds and defines useful test sets and a standard of truth for assessing the correctness of sequence–sequence or sequence–structure alignments.

The major new developments since last year (5) are continuous updates of the database and easy access to the data using browsers on the World Wide Web (WWW).

FORM AND CONTENT OF THE DATABASE

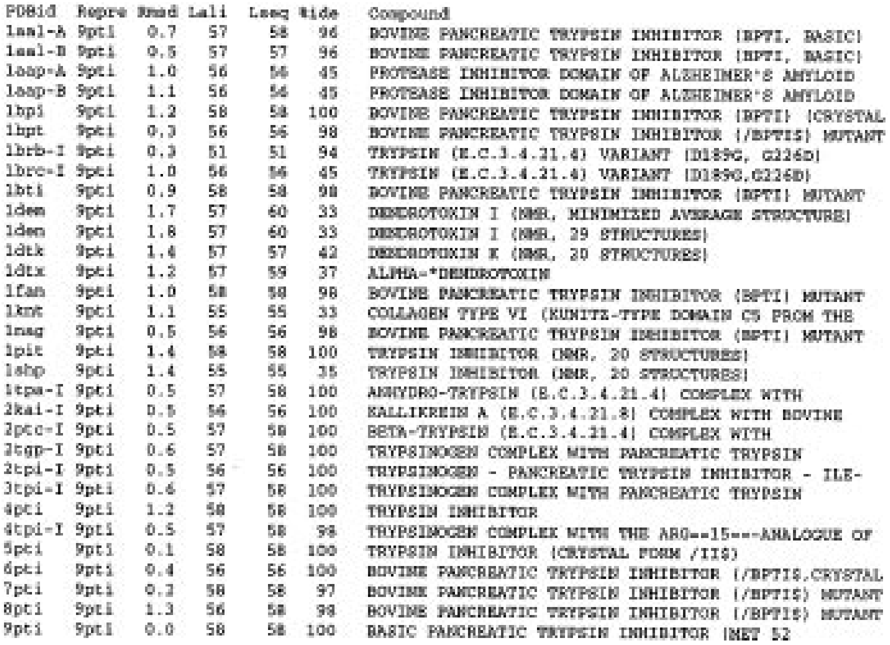
Fold classification

The basic structural entity used currently in the FSSP database are protein chains, which are identified by the Protein Data Bank (PDB) entry code plus chain identifier. All protein chains in the PDB entries that are >30 residues are listed alphabetically in PROTEIN INDEX which gives the pointer to the representative structure of the protein family and short summary information about the strength of similarity to the representative. The sequence-representative set is derived using algorithm #1 of ref. 6 so that all pairwise sequence identities within this set are <25%. For example, PROTEIN INDEX (Fig. 1) tells you that the protease inhibitor domain of Alzheimer’s amyloid beta-protein precursor is deposited in the PDB as entry 1AAP which has two chains, A and B. Both the A and B chain are 45% sequence identical to the representative structure of the family, which is bovine pancreatic trypsin inhibitor (PDB entry 9PTI). As expected from the high sequence identity, the folds of both of the 1AAP chains and that of 9PTI are as good as identical (1.0–1.1 Å root-mean-square deviation of CA positions).

Classifying proteins into sequence families yields a reduction from nearly 5000 protein chains in the PDB to ~600 representatives. This set includes many pairs of remote homologs that have completely superimposable 3-D structures despite low sequence similarity and pairs with recurrent common folding motifs. The sequence-representative set is clustered further based on all-against-all structure comparison within the sequence-representative set.

FOLDTREE is a tree representation of the sequence-representative set produced by hierarchical clustering. The tree gives a simple overview of protein families, grouping together remote homologs 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 200 fold classes. For example, Figure 2 shows how the first C2 domain of synaptotagmin I (PDB entry 1RSY), which presented a new calcium-binding fold (7), is firmly anchored in a large structural class that contains beta-sandwich proteins with topological similarity to immunoglobulin-like domains and blue copper proteins.

An alternative way of defining clusters in protein fold space is used to derive the PDBfolds series of structurally representative sets using algorithm #2 of ref. 6. The sets of representative folds contain a maximal number of protein folds where no pair is allowed to have a larger fraction of structurally equivalent residues



```
PDBid  Repre Rmsd Lali  Lseq %ide   Compound
1aal-A 9pti   0.7   57    58   96   BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI, BASIC)
1aal-B 9pti   0.5   57    57   96   BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI, BASIC)
1aap-A 9pti   1.0   56    56   45   PROTEASE INHIBITOR DOMAIN OF ALZHEIMER'S AMYLOID
1aap-B 9pti   1.1   56    56   45   PROTEASE INHIBITOR DOMAIN OF ALZHEIMER'S AMYLOID
1bpi   9pti   1.2   58    58  100   BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI) (CRYSTAL
1bpt   9pti   0.3   56    56   98   BOVINE PANCREATIC TRYPSIN INHIBITOR (/BPTI$) MUTANT
1brb-I 9pti   0.3   51    51   94   TRYPSIN (E.C.3.4.21.4) VARIANT (D189G, G226D)
1brc-I 9pti   1.0   56    56   45   TRYPSIN (E.C.3.4.21.4) VARIANT (D189G,G226D)
1bti   9pti   0.9   58    58   98   BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI) MUTANT
1dem   9pti   1.7   57    60   33   DENDROTOXIN I (NMR, MINIMIZED AVERAGE STRUCTURE)
1den   9pti   1.8   57    60   33   DENDROTOXIN I (NMR, 29 STRUCTURES)
1dtk   9pti   1.4   57    57   42   DENDROTOXIN K (NMR, 20 STRUCTURES)
1dtx   9pti   1.2   57    59   37   ALPHA-*DENDROTOXIN
1fan   9pti   1.0   58    58   98   BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI) MUTANT
1knt   9pti   1.1   55    55   33   COLLAGEN TYPE VI (KUNITZ-TYPE DOMAIN C5 FROM THE
1nag   9pti   0.5   56    56   98   BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI) MUTANT
1pit   9pti   1.4   58    58  100   TRYPSIN INHIBITOR (NMR, 20 STRUCTURES)
1shp   9pti   1.4   55    55   35   TRYPSIN INHIBITOR (NMR, 20 STRUCTURES)
1tpa-I 9pti   0.5   57    58  100   ANHYDRO-TRYPSIN (E.C.3.4.21.4) COMPLEX WITH
2kai-I 9pti   0.5   56    56  100   KALLIKREIN A (E.C.3.4.21.8) COMPLEX WITH BOVINE
2ptc-I 9pti   0.5   57    58  100   BETA-TRYPSIN (E.C.3.4.21.4) COMPLEX WITH
2tgp-I 9pti   0.6   57    58  100   TRYPSINOGEN COMPLEX WITH PANCREATIC TRYPSIN
2tpi-I 9pti   0.5   56    56  100   TRYPSINOGEN - PANCREATIC TRYPSIN INHIBITOR - ILE-
3tpi-I 9pti   0.6   57    58  100   TRYPSINOGEN COMPLEX WITH PANCREATIC TRYPSIN
4pti   9pti   1.2   58    58  100   TRYPSIN INHIBITOR
4tpi-I 9pti   0.5   57    58   98   TRYPSINOGEN COMPLEX WITH THE ARG==15===ANALOGUE OF
5pti   9pti   0.1   58    58  100   TRYPSIN INHIBITOR (CRYSTAL FORM /II$)
6pti   9pti   0.4   56    56  100   BOVINE PANCREATIC TRYPSIN INHIBITOR (/BPTI$,CRYSTAL
7pti   9pti   0.3   58    58   97   BOVINE PANCREATIC TRYPSIN INHIBITOR (/BPTI$) MUTANT
8pti   9pti   1.3   56    58   98   BOVINE PANCREATIC TRYPSIN INHIBITOR (/BPTI$) MUTANT
9pti   9pti   0.0   58    58  100   BASIC PANCREATIC TRYPSIN INHIBITOR (MET 52
```

Figure 1. Finding proteins in FSSP. All protein structures in the PDB are listed alphabetically in the PROTEIN INDEX table. The index can be used for searching by protein name or PDB code. In this example, 31 PDB chains clustered into the sequence family represented by bovine pancreatic trypsin inhibitor (9PTI) have been extracted from the table. These include multiple determinations of the same protein in different crystallographic conditions (chains with 100% sequence identity to the representative) and homologs from other species with sequence identity down to 33% relative to the representative. Notation: PDBid, PDB entry name, chain identifier appended; Repre, representative structure of the family; Rmsd, root-mean-square deviation of CA atoms in 3-D superimposition; Lali, number of structurally equivalent residues; Lseq, number of residues in PDBid; %ide, percentage of identical residues between PDBid and Repre in structural alignment; Compound, protein name echoed from the PDB entry.

than a given threshold percentage. This reduces the number of unique folds to consider for structural analysis, depending on the threshold chosen. For example, the common structural core covers >90% of the chain in all globin–globin pairs and >70% in any phycocyanin–globin pair. Accordingly, there is only one globin structure in the 90% list and only one representative for the phycocyanin-globin fold in the 70% list of PDBfolds.

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 a few or tens of proteins structurally similar to the search structure, alongside the secondary structure and solvent accessibility extracted from the 3-D coordinates of the search structure (8). 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 ($Z \geq 2$). Details about the Dali method used to derive the database are given in refs 9 and 10**.**

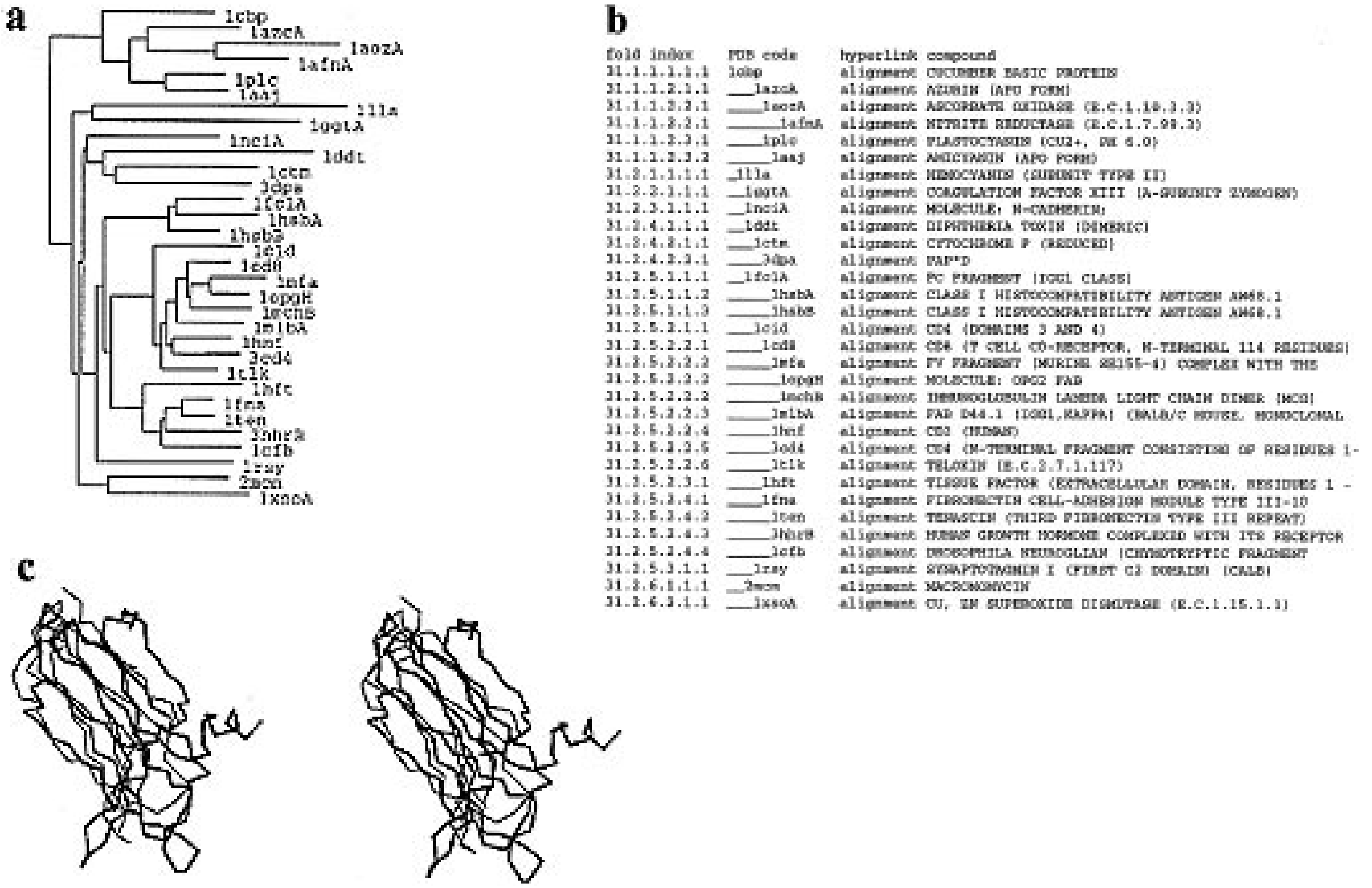
An FSSP file is divided in five formatted blocks and a free text footer which explains the format. (i) The header block identifies the query structure, database and structural alignment method used and gives the number of structural neighbours. (ii) The summary block gives a one-line summary for each 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 3-D coordinates in the respective PDB entries, yield the least-squares superimposition of the matched protein onto the query structure. See below for automatic parsing of FSSP entries.

DISTRIBUTION

World Wide Web

The FSSP database is accessible over the WWW addressing URL http://www.embl-heidelberg.de/dali/fssp/.

The most convenient starting point for a walk in fold space is via clicking the ‘alignment’ link in the FOLDTREE table. FSSP entries are parsed on the fly to display structural neighbours of individual proteins in the form of structure alignments laid out horizontally, multiple structure alignments (known structures) combined with multiple sequence alignments [sequences homologous to a known structure: HSSP database (11)] or superimposed coordinates [retrieved from PDB (12)] for viewing with molecular graphics programs such as Rasmol (13). There are further hypertext links to functional annotations and literature references via SRS (14). For example, a study of the p21 *ras* family could start from the FOLDTREE table, which immediately shows transducin alpha, the ADP-ribosylation factor 1 and elongation factor G as the closest structural neighbours. From the structural alignment of these remote homologs one can identify the conserved sequence motifs GxxxxGKS and NKxD (15). These



fold index	PDB code	hyperlink	compound
31.1.1.1.1.1	1cbp	alignment	CUCUMBER BASIC PROTEIN
31.1.1.2.1.1	___1aznA	alignment	AZURIN (APO FORM)
31.1.1.2.2.1	____1aozA	alignment	ASCORBATE OXIDASE (E.C.1.10.3.3)
31.1.1.2.2.1	______1afnA	alignment	NITRITE REDUCTASE (E.C.1.7.99.3)
31.1.1.2.3.1	____1plc	alignment	PLASTOCYANIN (CU2+, PH 6.0)
31.1.1.2.3.2	_____1aaj	alignment	AMICYANIN (APO FORM)
31.2.1.1.1.1	_1lla	alignment	HEMOCYANIN (SUBUNIT TYPE II)
31.2.2.1.1.1	__1ggtA	alignment	COAGULATION FACTOR XIII (A-SUBUNIT ZYMOGEN)
31.2.3.1.1.1	__1nciA	alignment	MOLECULE: N-CADHERIN;
31.2.4.1.1.1	__1ddt	alignment	DIPHTHERIA TOXIN (DIMERIC)
31.2.4.2.1.1	___1ctm	alignment	CYTOCHROME F (REDUCED)
31.2.4.2.2.1	____3dpa	alignment	PAP'D
31.2.5.1.1.1	__1fc1A	alignment	FC FRAGMENT (IGG1 CLASS)
31.2.5.1.1.2	_____1hsbA	alignment	CLASS I HISTOCOMPATIBILITY ANTIGEN AW68.1
31.2.5.1.1.3	_____1hsbB	alignment	CLASS I HISTOCOMPATIBILITY ANTIGEN AW68.1
31.2.5.2.1.1	___1cid	alignment	CD4 (DOMAINS 3 AND 4)
31.2.5.2.2.1	____1cd8	alignment	CD8 (T CELL CO-RECEPTOR, N-TERMINAL 114 RESIDUES)
31.2.5.2.2.2	_____1mfa	alignment	FV FRAGMENT (MURINE SE155-4) COMPLEX WITH THE
31.2.5.2.2.2	______1opgH	alignment	MOLECULE: OPG2 FAB
31.2.5.2.2.2	______1mcnB	alignment	IMMUNOGLOBULIN LAMBDA LIGHT CHAIN DIMER (MCG)
31.2.5.2.2.3	_____1mlbA	alignment	FAB D44.1 (IGG1,KAPPA) (BALB/C MOUSE, MONOCLONAL
31.2.5.2.2.4	_____1hnf	alignment	CD2 (HUMAN)
31.2.5.2.2.5	_____1cd4	alignment	CD4 (N-TERMINAL FRAGMENT CONSISTING OF RESIDUES 1-
31.2.5.2.2.6	_____1tlk	alignment	TELOKIN (E.C.2.7.1.117)
31.2.5.2.3.1	____1hft	alignment	TISSUE FACTOR (EXTRACELLULAR DOMAIN, RESIDUES 1 -
31.2.5.2.4.1	____1fna	alignment	FIBRONECTIN CELL-ADHESION MODULE TYPE III-10
31.2.5.2.4.2	_____1ten	alignment	TENASCIN (THIRD FIBRONECTIN TYPE III REPEAT)
31.2.5.2.4.3	_____3hhrB	alignment	HUMAN GROWTH HORMONE COMPLEXED WITH ITS RECEPTOR
31.2.5.2.4.4	_____1cfb	alignment	DROSOPHILA NEUROGLIAN (CHYMOTRYPTIC FRAGMENT
31.2.5.3.1.1	___1rsy	alignment	SYNAPTOTAGMIN I (FIRST C2 DOMAIN) (CALB)
31.2.6.1.1.1	__2mcm	alignment	MACROMOMYCIN
31.2.6.2.1.1	___1xsoA	alignment	CU, ZN SUPEROXIDE DISMUTASE (E.C.1.15.1.1)

Figure 2. Overview of protein fold space. (**a**) Part of fold tree obtained by hierarchical clustering based on structural similarities between proteins in the representative set (<25% pairwise sequence identity). (**b**) The same part of the fold tree as it appears in the FOLDTREE table. A fold index is constructed by cutting an average linkage clustering tree at a similarity level of two standard deviations above expected (Z = 2), for example 31 in 31.2.5.3.1.1. for synaptotagmin. Subfamilies are defined and indexed according to cuts at similarity levels of Z = 3, 4, 5, 6 and 10, that is increasing levels of stringency. For example, the cut at Z = 4 (31.2.*) separates between blue copper proteins, hemocyanin, coagulation factor, cadherin, bacterial and eukaryotic immunoglobulin-like domains and superoxide dismutases. Indentation in the 'PDB code' column corresponds to the fold indices and means that a protein belongs to the same structural family/subfamily as the protein above. **(c)** Stereo view of superimposition between synaptotagmin I (PDB entry 1RSY, thick line) and a fibronectin type III domain (PDB entry 1FNA, thin line) reveals the common topological arrangement of strands in the beta sandwich (cf. ref. 7). Plotted with WhatIf (17).

patterns are conserved in all members of the protein families as seen by extending the structure alignment with the results from a sequence database search (11). The number of sequence relatives displayed can be reduced from several hundred to a few tens using a cutoff of 50% identity between any pair that is displayed (Fig. 3). Clicking on the sequence identifier (e.g. rash_rat) pops up the Swissprot (16) annotation for this sequence.

Anonymous ftp

The FSSP data sets can be obtained by anonymous ftp from ftp.embl-heidelberg.de in the directory: /pub/databases/protein_extras/fssp.

Conditions

Academic redistribution of single files or of the entire database is permitted. No inclusion in other databases or database services, academic or other, without explicit permission of the authors. All rights reserved. Not to be used for classified research. Users are asked to refer to ref. 9 and this paper in reporting results obtained using the database.

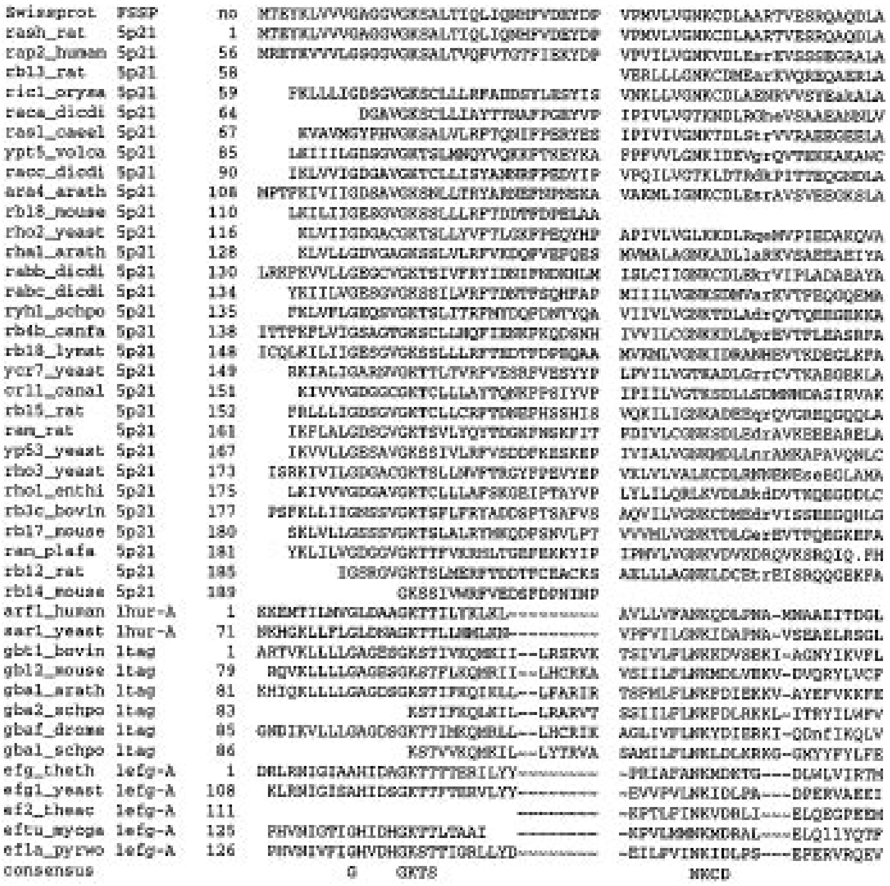
SIZE OF THE CURRENT RELEASE

The size of the FSSP database is tightly coupled to that of the PDB from which it is derived. The FSSP database is updated with each release of new structures by the PDB. The size of the sequence-representative set of chains was 600 in August 1995, an 80% increase from June 1994. The complete set of result files requires ~60 Mb of disk storage.

LIMITATIONS

The current database contains at most one alignment per pair of full length proteins. The alignments are constrained to be sequential as this is biologically meaningful though not imposed by the Dali method. Different chains in one PDB entry are compared separately; chains with <30 residues or unknown sequence are excluded.

The structure comparison program Dali (9) defines the extent of the common structural core by maximizing the agreement of *intra*molecular CA–CA distances. The scoring function was deliberately designed to allow inter-domain conformational flexibility; hence, positional root mean square deviations for the



```
Swissprot  FSSP      no  MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDP  VPMVLVGNKCDLAARTVESRQAQDLA
rash_rat   5p21       1  MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDP  VPMVLVGNKCDLAARTVESRQAQDLA
rap2_human 5p21      56  MREYKVVVLGSGGVGKSALTVQFVTGTFIEKYDP  VPVILVGNKVDLEsrEVSSSEGRALA
rb11_rat   5p21      58                                      VERLLLGNKCDMEarKVQREQAERLA
ric1_orysa 5p21      59     FKLLLIGDSGVGKSCLLLRFADDSYLESYIS  VNKLLVGNKCDLAENRVVSYEaKALA
raca_dicdi 5p21      64            DGAVGKSCLLIAYTTNAFPGEYVP  IPIVLVGTKNDLRQheVSAAEANNLV
ras1_caeel 5p21      67      KVAVMGYPHVGKSALVLRFTQNIFPESYES  IPIVIVGNKTDLStrVVRAEEGEELA
ypt5_volca 5p21      85     LKIIILGDSGVGKTSLMNQYVQKKFTKEYKA  PPFVVLGNKIDEVgrQVTEEKAKAWC
racc_dicdi 5p21      90     IKLVVIGDGAVGKTCLLISYANNRFPEDYIP  VPQILVGTKLDTRdRPITTEQGNDLA
ara4_arath 5p21     108  MPTFKIVIIGDSAVGKSNLLTRYARNEFNPNSKA  VAKMLIGNKCDLEsrAVSVEEGKSLA
rb18_mouse 5p21     110     LKILIIGESGVGKSSLLLRFTDDTFDPELAA
rho2_yeast 5p21     116      KLVIIGDGACGKTSLLYVFTLGKFPEQYHP  APIVLVGLKKDLRqeMVPIEDAKQVA
rha1_arath 5p21     128      KLVLLGDVGAGKSSLVLRFVKDQFVEFQES  MVMALAGNKADLlaRKVSAEEAEIYA
rabb_dicdi 5p21     130  LRKFKVVLLGEGCVGKTSIVFRYIDNIFNDKHLM  ISLCIIGNKCDLEkrVIPLADAEAYA
rabc_dicdi 5p21     134     YKIILVGESGVGKSSILVRFTDNTFSQHFAP  MIIILVGNKSDMVarKVTPEQGQEMA
ryh1_schpo 5p21     135     FKLVFLGEQSVGKTSLITRFMYDQFDNTYQA  VIIVLVGNKTDLAdrQVTQEEGEKKA
rb4b_canfa 5p21     138  ITTFKFLVIGSAGTGKSCLLHQFIENKFKQDSNH  IVVILCGNKKDLDprEVTFLEASRFA
rb18_lymst 5p21     148  ICQLKILIIGESGVGKSSLLLRFTEDTFDPEQAA  MVKMLVGNKIDRANHEVTKDEGLKFA
ycr7_yeast 5p21     149     RKIALIGARSVGKTTLTVRFVESRFVESYYP  LPVILVGTKADLGrrCVTKAEGEKLA
crl1_canal 5p21     151      KIVVVGDGGCGKTCLLLAYTQNKFPSIYVP  IPIILVGTKSDLLSDMNNDASIRVAK
rb15_rat   5p21     152     FRLLLIGDSGVGKTCLLCRFTDNEFHSSHIS  VQKILIGNKADEEqrQVGREQGQQLA
ram_rat    5p21     161     IKFLALGDSGVGKTSVLYQYTDGKFNSKFIT  PDIVLCGNKSDLEdrAVKEEEARELA
yp53_yeast 5p21     167     IKVVLLGESAVGKSSIVLRFVSDDFKESKEP  IVIALVGNKMDLLnrAMKAPAVQNLC
rho3_yeast 5p21     173   ISRKIVILGDGACGKTSLLNVFTRGYFPEVYEP  VKLVLVALKCDLRNNENEseEGLAMA
rho1_enthi 5p21     175     LKIVVVGDGAVGKTCLLLAFSKGEIPTAYVP  LYLILQGLKVDLRkdDVTKQEGDDLC
rb1c_bovin 5p21     177   FSFKLLIIGNSSVGKTSFLFRYADDSFTSAFVS  AQVILVGNKCDMEdrVISSERGQHLG
rb17_mouse 5p21     180     SKLVLLGSSSVGKTSLALRYMKQDFSNVLPT  VVVHLVGNKTDLGerEVTFQEGKEFA
ran_plafa  5p21     181     YKLILVGDGGVGKTTFVKRHLTGEFEKKYIP  IPMVLVGNKVDVKDRQVKSRQIQ.FH
rb12_rat   5p21     185            IGSRGVGKTSLMERFTDDTFCEACKS  AKLLLAGNKLDCEtrEISRQQGEKFA
rb14_mouse 5p21     189                  GKSSIVWRFVEDSFDPNINP
arf1_human 1hur-A     1  KKEMTILMVGLDAAGKTTILYKLKL~~~~~~~~~  AVLLVFANKQDLPNA-MNAAEITDGL
sar1_yeast 1hur-A    71  NKHGKLLFLGLDNAGKTTLLHMLKN---------  VPFVILGNKIDAPNA-VSEAELRSGL
gbt1_bovin 1tag       1  ARTVKLLLLGAGESGKSTIVKQMKII--LRSRVK  TSIVLFLNKKDVSEKI-AGNYIKVFL
gbi2_mouse 1tag      79   RQVKLLLLGAGESGKSTFLKQMRII--LHCRKA  VSIILFLNKMDLVEKV-DVQRYLVCF
gba1_arath 1tag      81  KHIQKLLLLGAGDSGKSTIFKQIKLL--LFARIR  TSFMLFLNKFDIEKKV-AYEFVKKFE
gba2_schpo 1tag      83                 KSTIFKQLKIL--LRARVT  SSIILFLNKFDLAKKL-ITRYILMFV
gbaf_drome 1tag      85  GNDIKVLLLGAGDSGKTTIMKQMRLL--LHCRIK  AGLIVFLNKYDIERKI-QDnfIKQLV
gba1_schpo 1tag      86                 KSTVVKQMKIL--LYTRVA  SAMILFLNKLDLKRKG-GKYYFYLFE
efg_theth  1efg-A     1  DRLRNIGIAAHIDAGKTTTTERILYY~~~~~~~~  -PRIAFANKMDKTG---DLWLVIRTM
efg1_yeast 1efg-A   108   KLRNIGISAHIDSGKTTFTERVLYY---------  -EVVPVLNKIDLPA---DPERVAEEI
ef2_theac  1efg-A   111                           ---------  -KPTLFINKVDRLI---ELQEGPEEM
eftu_mycge 1efg-A   125   PHVNIGTIGHIDHGKTTLTAAI   ---------  -KFVLMMNKMDRAL---ELQllYQTF
ef1a_pyrwo 1efg-A   126   PHVNIVFIGHVDHGKSTTIGRLLYD~~~~~~~~  -EILFVINKIDLPS---EPERVRQEV
consensus                          G    GKTS                          NKCD
```

Figure 3. Combining multiple structure–structure alignments with multiple sequence–sequence alignments. A multiple sequence alignment of four protein families: p21 *ras*, transducin alpha, ADP-ribosylation factor 1 and elongation factor G. Only structurally equivalent blocks are shown; the middle part of the alignment has been omitted in order to highlight the conserved sequence signatures near the N- and C-termini. Structural alignment defines the register of each of the families (indicated in the FSSP column) relative to p21 *ras*. In addition to the guide structures, the alignment includes representative sequence homologs (Swissprot column; first sequence corresponds to the known structure) taken from the HSSP database of sequence–sequence alignments (11). The combined multiple alignment is filtered so that any sequence pair displayed has <50% sequence identity. For example, the original HSSP entry for 5p21 lists 189 sequences; here, only 29 representative*ras* sequences are shown. Notation: ~, nonequivalent segments and trailing ends from structure alignment; blanks and dots, gaps and trailing ends from sequence alignment; lowercase, insertions in sequence alignment.

corresponding rigid-body superimpositions are often higher than for comparison methods that put an absolute upper limit on *inter*molecular positional deviations. This, however, is only an apparent disadvantage.

RELATED SERVICE

Requests for alignments of newly solved crystallographic or solution NMR structures (C^{α} co-ordinates required) may be sent to the Dali e-mail server with Internet address:
dali@embl-heidelberg.de.
More information on the Dali server (10) is available on the WWW at:
URL http://www.embl-heidelberg.de/dali/dali.html.

Kindly report any problems to the authors by e-mail.

REFERENCES

1 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1994) *J. Mol. Biol*., **247**, 536–540.

2 Overington J., Johnson M.S., Sali A. and Blundell T.L. (1990) *Proc. Royal Soc. Lond*., **B241**, 132–145.

3 Orengo C.A., Flores T.P., Taylor W.R. and Thornton J.M. (1993) *Protein Eng*., **6**, 485–500.

4 Holm L. and Sander C. (1994) *Proteins*, **19**, 165–173.

5 Holm L. and Sander C. (1994) *Nucleic Acids Res*., **22**, 3600–3609.

6 Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Sci*., **1**, 409–417.

7 Sutton R.B., Davletov B.A., Berghuis A.M., Sudhof T.C. and Sprang S.R. (1995) *Cell*, **80**, 929–935.

8 Kabsch W. and Sander C. (1983) *Biopolymers*, **22**, 2577–2637.

9 Holm L. and Sander C. (1993) *J. Mol. Biol*., **233**,123–138.

10 Holm L. and Sander C. (1995) *Trends Biol. Sci*., **20**, 478–480.

11 Sander C. and Schneider R. (1991) *Proteins*, **9**, 56–68.

12 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.

13 Sayle R.A. and Milner-White, E.J. (1995) *Trends Biol. Sci*., **20**, 374–376.

14 Etzold T. and Argos P. (1993) *CABIOS*, **9**, 49–57.

15 Valencia,A., Kjeldgaard, M., Pai, E.F. and Sander, C. (1991) *Proc. Natl. Acad. Sci*., **88**, 5443–5447.

16 Bairoch, A. (1992) *Nucleic Acids Res*., **20**, 2013–2018.

17 Vriend, G. (1990) *J. Mol. Graphics*, **8**, 52–56.