BMC Bioinformatics

BioMed Central

Database

Open Access

PASS2: an automated database of protein alignments organised as structural superfamilies

Anirban Bhaduri, Ganesan Pugalenthi and Ramanathan Sowdhamini*

Address: National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK campus, Bellary Road, Bangalore, Karnataka 560 065, India

Email: Anirban Bhaduri - anirban@ncbs.res.in; Ganesan Pugalenthi - pugal@ncbs.res.in; Ramanathan Sowdhamini* - mini@ncbs.res.in

* Corresponding author

Published: 02 April 2004

BMC Bioinformatics 2004, **5**:35

Received: 28 November 2003

Accepted: 02 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/35

Abstract

Background: The functional selection and three-dimensional structural constraints of proteins in nature often relates to the retention of significant sequence similarity between proteins of similar fold and function despite poor sequence identity. Organization of structure-based sequence alignments for distantly related proteins, provides a map of the conserved and critical 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 Protein Alignment organised as Structural Superfamily (PASS2) database represents continuously updated, structural alignments for evolutionary related, sequentially distant proteins.

Description: An automated and updated version of PASS2 is, in direct correspondence with SCOP 1.63, consisting of sequences having identity below 40% among themselves. Protein domains have been grouped into 628 multi-member superfamilies and 566 single member superfamilies. Structure-based sequence alignments for the superfamilies have been obtained using COMPARER, while initial equivalencies have been derived from a preliminary superposition using LSQMAN or STAMP 4.0. The final sequence alignments have been annotated for structural features using JOY4.0. The database is supplemented with sequence relatives belonging to different genomes, conserved spatially interacting and structural motifs, probabilistic hidden markov models of superfamilies based on the alignments and useful links to other databases. Probabilistic models and sensitive position specific profiles obtained from reliable superfamily alignments aid annotation of remote homologues and are useful tools in structural and functional genomics. PASS2 presents the phylogeny of its members both based on sequence and structural dissimilarities. Clustering of members allows us to understand diversification of the family members. The search engine has been improved for simpler browsing of the database.

Conclusions: The database resolves alignments among the structural domains consisting of evolutionarily diverged set of sequences. Availability of reliable sequence alignments of distantly related proteins despite poor sequence identity and single-member superfamilies permit better sampling of structures in libraries for fold recognition of new sequences and for the understanding of protein structure-function relationships of individual superfamilies. PASS2 is accessible at http://www.ncbs.res.in/~faculty/mini/campass/pass2.html

Background

Classification of proteins into families is performed on the basis of the similarity of sequences to the family members [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 conserved functional residues are required for biochemical activity amongst proteins belonging to a superfamily [3]. Establishing evolutionary relationships between superfamily members having similar structure and function but sequentially diverged is challenging. Over 49,000 domains deposited in the Protein Data Bank (PDB) [4] are organized in different databases by hierarchical classification schemes or in terms of structural neighbourhood distances [5-7]. SCOP (1.63 release) records 49,497 protein domains, grouped into merely 765 folds, suggesting a strong structural convergence of proteins. Homologous families can be easily grouped by simple sequence searches whereas superfamily members, adopting the same fold and performing similar biological roles [8-13] can often be identified by sensitive fold prediction algorithms followed by a careful alignment of sequences.

Availability of reliable sequence alignments for distantly related proteins despite poor sequence identity permits better sampling of structures in libraries for fold recognition of new sequences and for the understanding of protein structure-function relationships of individual superfamilies. In addition, the construction of three-dimensional models using homology modelling techniques are usually reliable where the sequence identity between query and the structural homologues (templates) are 30% or above. Analyses of structural and sequence differences amongst known superfamily members can hopefully provide useful guidelines for modelling distantly related proteins. PASS2 database [14,15] presents alignments of sequentially distant proteins related at the superfamily level. We report an automated, updated version of the superfamily alignment database that is in direct correspondence with SCOP (1.63) database.

Construction and content

The present version of PASS2 consider domains as assigned in SCOP 1.63 [6]. Domains within a superfamily, no more than 40% identical with each other, have been considered for curating the database. The choice of 40% cut-off in percentage sequence identity, as compared to the previous version of PASS2 that works at 25% identity level, was to reduce the number of single-member superfamilies. The 4,001 protein domains were assigned 1,194 superfamilies spanning the seven classes of proteins and were thus chosen for structure based sequence alignments.

Curation of alignments

Structural domains, obtained consulting SCOP [6] definitions, have been grouped at the superfamily level and superposed by rigid-body superposition (Figure 1). An initial superposition for all the structural domains belonging to each non-redundant superfamily was performed using LSQMAN [16] or STAMP 4.0 [17]. LSQMAN [16] was used for superposing two member superfamilies while STAMP 4.0 [17] was utilised in multi-member superfamilies. From the coarse alignment, equivalent regions were identified using JOY [18]. COMPARER [19] was employed to derive a refined alignment and superposition for the structures. Superposition was achieved by the choice of 'initial equivalencies' that served as seeds for pairwise rigid-body superposition using PMNFC, a modified form of MNYFIT [20] (Figure 2). The final alignment was presented using the three-dimensional structural features of JOY [18] (Figure 3).

Utility and discussion

Assigning new structural entries to pre-existing superfamilies

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. PASS2 allows classification of three-dimensional domains into respective superfamilies based on sequential and structural properties. Sequence of the uploaded structure is compared to the hidden markov models of PASS2 and assigned to superfamilies on the basis of liberal expectation values (E = 1.0). Representative structures of the putative superfamilies have been superposed with the query using LSQMAN [16], thus associating the query to a particular superfamily. Alternatively, the user can superpose an uploaded structure to specific superfamilies.

Predicting superfamilies and alignment for sequences

Links have been provided to popular sequence search methods like PSI-BLAST [21] and PHI-BLAST [22], which may be employed to associate unannotated sequences to PASS2 superfamilies. A sequence to probabilistic profile match method Hmmpfam [23] can also be used for similar assignment. Sequence alignments for a query sequence can be obtained with superfamily members using MALIGN [24]. 3-dimensional features can also be attributed to the sequence alignment using JOY [18].

Hidden markov models for PASS2

During search for sequence homologues and sequence assignment, profile-based methods perform better compared to those that use pairwise comparisons [25]. Family profiles based on hidden markov models are popular probabilistic models applied for sequence annotations and searches [26,27]. Structure-based sequence alignment

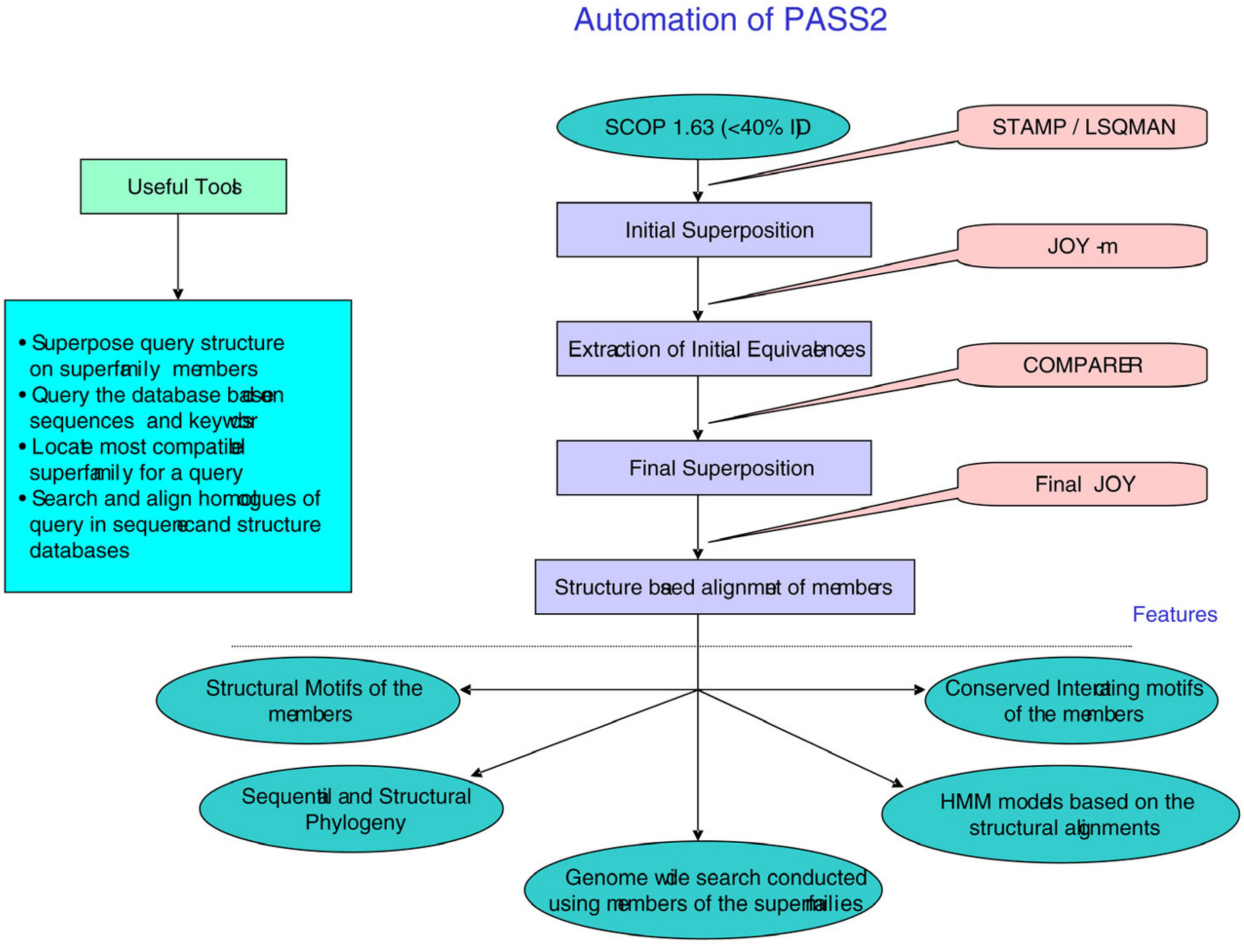

Automation of PASS2
SCOP 1.63 (<40% ID)
STAMP / LSQMAN
Useful Tools
Initial Superposition
JOY -m
• Superpose query structure on superfamily members
• Query the database based on sequences and keywords
• Locate most compatible superfamily for a query
• Search and align homologues of query in sequence and structure databases
Extraction of Initial Equivalences
COMPARER
Final Superposition
Final JOY
Structure based alignment of members
Features
Structural Motifs of the members
Conserved Interacting motifs of the members
Sequential and Structural Phylogeny
HMM models based on the structural alignments
Genome wide search conducted using members of the superfamilies

Figure 1
Flowchart representation of the steps involved in the curation of PASS2 database. Listed are useful tools and additional derived information that may be obtained from PASS2.

of respective superfamilies in PASS2 provides a reliable basis for building hidden markov models. We provide HMMs, built using HMM suite [23], for superfamily alignments corresponding to the latest version of PASS2. The performances of these HMMs have been compared with models built using their structural homologues present in the PDB [28]. Search for homologues have been performed on the non-redundant sequence database using both sets of models. Higher coverage has been obtained (Table 1) for superfamilies using PASS2 HMMs suggesting their value in sensitive sequence searches. Hidden markov models for both the structure-based sequence alignments and the sequence enriched superfamily alignments can be downloaded from the World Wide Web.

Superfamily members in the genome database

PASS2 has several new features to associate the structure-based sequences to their homologues in various genome databases. Sequence homologues of the superfamilies have been searched in the non-redundant sequence database using PSI-BLAST [21] and Hmmsearch [23]. For the PSI-BLAST searches, individual member for each superfamily was queried against the non-redundant sequence database. The expectation value was set to 0.001 with 20 iterations. Hidden Markov Models for every superfamily was built using structural alignments (as explained above). These models were searched against the non-redundant database to enrich the sequence members using the Hmmsearch program belonging to the HMM suite applying an E-value threshold of 0.1. A third approach has been to employ interacting motifs,

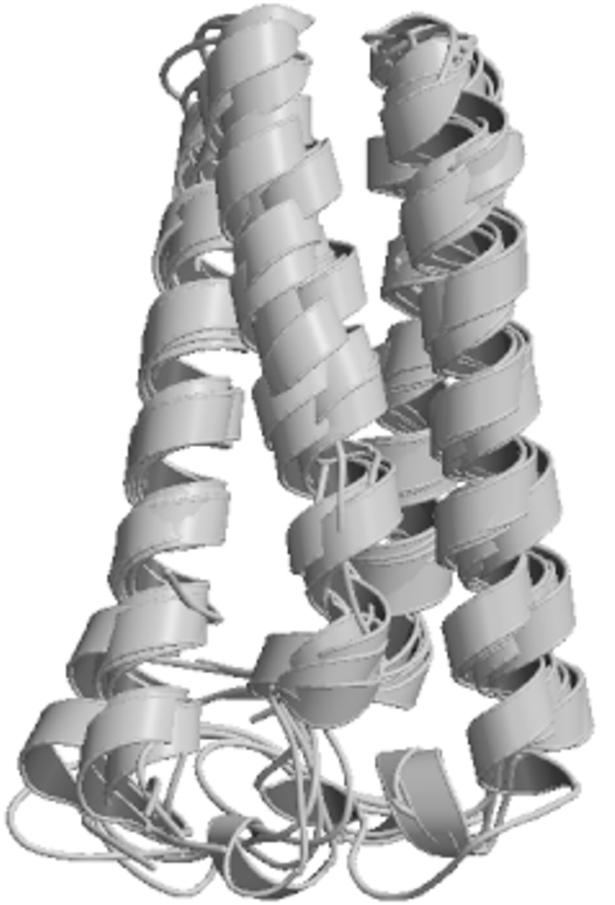


Figure 2
Superposed structures of the cytochrome superfamily representatives: The cytochrome superfamily has six representative members in PASS2 (1a7va-, 1bbha-, 1cpq--, 1e85a-, 256ba-, 2ccya-) which have been superposed as explained (see Curation of Alignments section). The figure has been created using MOLSCRIPT [32].

identified for superfamilies, as constraints in PHI-BLAST against searches in the non-redundant database using an E-value 1.0 as explained elsewhere [29]. Hits obtained by the three approaches belonging to the genomes were aligned using CLUSTALW [30] and presented along with their structural representatives of the superfamilies. The top 10 hits displayed in the web are aligned with PASS2 members. The entire set of hits corresponding to genomes can also be downloaded.

Information about superfamily members

A structure-based sequence alignment for the query with the appropriate superfamily can be obtained. Superposed coordinates for the query with the best ranking superfamily (based on the RMSD value) is also provided. Motifs represent invariant regions of the superfamily and are helpful in protein design, engineering and folding studies. Spatially conserved interacting motifs are identified as described elsewhere [29] for each superfamily and are listed in the current version of the database along with psuedoenergies for their spatial interactions (Bhaduri et al., in press). Corresponding links to the structural motifs of superfamily (SMoS) database [31] can also be accessed.

Phylogenetic analysis aids in the understanding of the diversity among the members. Diversification of structural members may be studied in terms of the dissimilarity of structure or divergence of the sequences. The database has been linked to other useful protein databases as in the previous version of PASS2 [15].

PASS2 and its applications

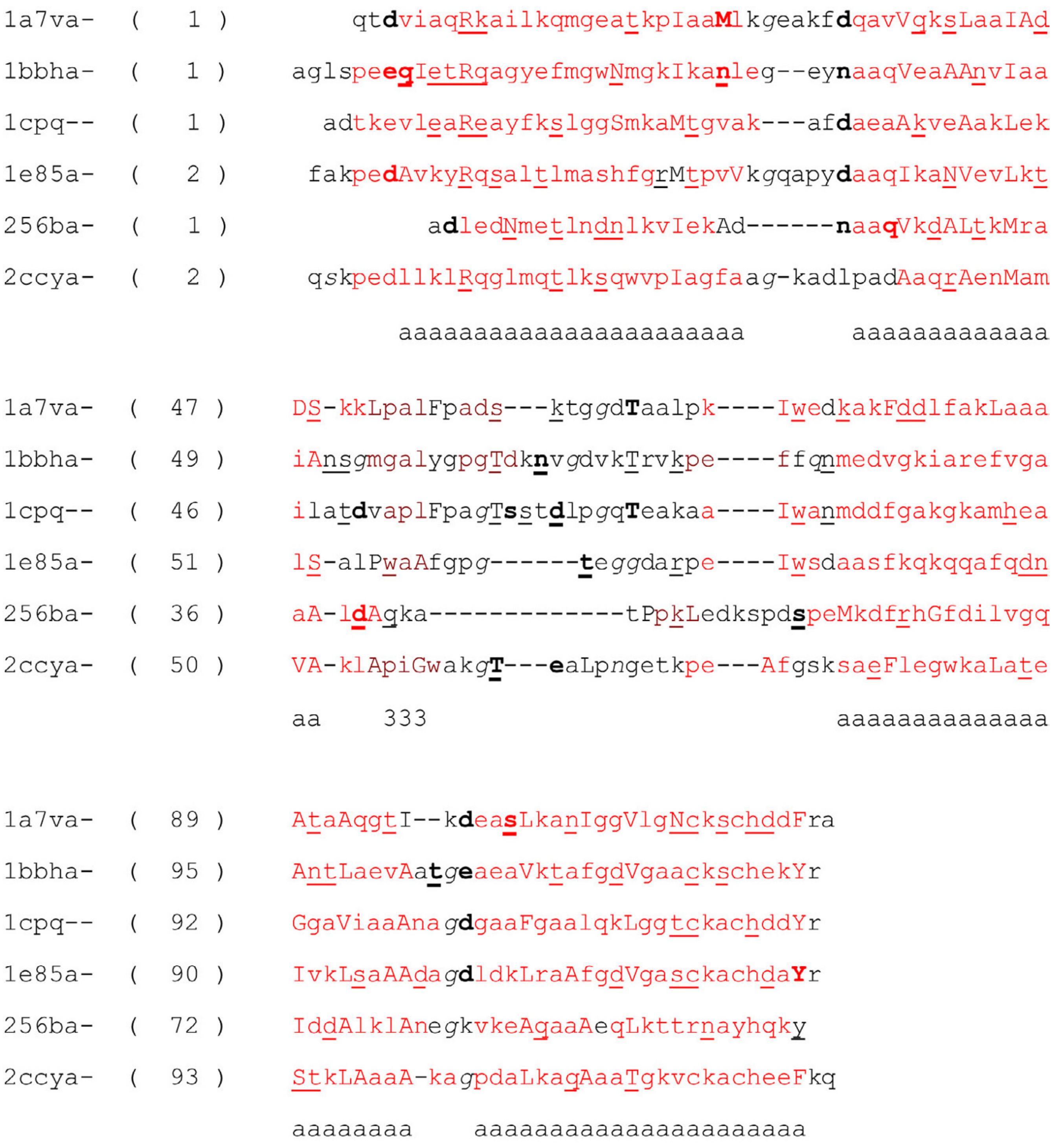
PASS2 is a compendium of structure-based sequence alignments of distantly related proteins grouped at the superfamily level in direct correspondence with SCOP definitions. Furthermore, PASS2 acts as a 'junction' point to obtain links of representative superfamily members to genome, sequence and structural databases. Phylogenies of superfamily members provide a crude but quantitative estimate of evolutionary relationships among the members. Motifs explain the invariant regions of proteins acting as descriptors for the superfamily. HMM models can be useful in identifying more members. Availability of such alignment databases over the World Wide Web facilitates the study and design of experiments on specific superfamilies. They also enable systematic survey and analysis of various structural properties for performing fold predictions. The database may be accessed and downloaded across the World Wide Web.

Conclusions

Associating different proteins with structurally similar and evolutionarily related proteins enhance our functional understanding of protein superfamily. The multiple alignments of distantly related representatives are particularly informative and often reveal a signature of invariantly conserved residues. Access to sequence alignments of distantly related proteins over the World Wide Web offers the possibility to study and design experiments on specific superfamilies. They also permit systematic survey and analysis of various structural properties and to perform fold predictions.

Availability of PASS2 database

PASS2 is accessible at http://www.ncbs.res.in/~faculty/mini/campass/pass2.html



```
1a7va-   (   1 )        qtdviaqRkailkqmgeatkpIaaMlkgeakfdqavVgksLaaIAd
1bbha-   (   1 )    aglspeegIetRqagyefmgwNmgkIkanleg--eynaaqVeaAAnvIaa
1cpq--   (   1 )      adtkevleaReayfkslggSmkaMtgvak---afdaeaAkveAakLek
1e85a-   (   2 )     fakpedAvkyRqsaltlmashfgrMtpvVkgqapydaaqIkaNVevLkt
256ba-   (   1 )             adledNmetlndnlkvIekAd------naaqVkdALtkMra
2ccya-   (   2 )     qskpedllklRqglmqtlksqwvpIagfaag-kadlpadAaqrAenMam
                           aaaaaaaaaaaaaaaaaaaaaaa       aaaaaaaaaaaaa

1a7va-   (  47 )    DS-kkLpalFpads---ktggdTaalpk----IwedkakFddlfakLaaa
1bbha-   (  49 )    iAnsgmgalygpgTdknvgdvkTrvkpe----ffqnmedvgkiarefvga
1cpq--   (  46 )    ilatdvaplFpagTsstdlpgqTeakaa----Iwanmddfgakgkamhea
1e85a-   (  51 )    lS-alPwaAfgpg------teggdarpe----Iwsdaasfkqkqqafqdn
256ba-   (  36 )    aA-ldAgka-------------tPpkLedkspdspeMkdfrhGfdilvgq
2ccya-   (  50 )    VA-klApiGwakgT---eaLpngetkpe---AfgsksaeFlegwkaLate
                    aa    333                             aaaaaaaaaaaaaa

1a7va-   (  89 )    AtaAqgtI--kdeasLkanIggVlgNckschddFra
1bbha-   (  95 )    AntLaevAatgeaeaVktafgdVgaackschekYr
1cpq--   (  92 )    GgaViaaAnagdgaaFgaalqkLggtckachddYr
1e85a-   (  90 )    IvkLsaAAdagdldkLraAfgdVgasckachdaYr
256ba-   (  72 )    IddAlklAnegkvkeAqaaAeqLkttrnayhqky
2ccya-   (  93 )    StkLAaaA-kagpdaLkaqAaaTgkvckacheeFkq
                    aaaaaaaa    aaaaaaaaaaaaaaaaaaaaaa
```

Figure 3
Representative structure-based sequence alignment for the cytochrome superfamily. The six members have been aligned and represented incorporating the three-dimensional features of JOY [18].

Table 1: Comparision of the number of hits obtained in HMMSearch using models derived from regular multiple sequence alignments and structure based sequence alignments.

Superfamily name	SCOP code	Hits obtained from PASS2 HMMs	Hits obtained from superfamily HMMs
Superoxide dismutase	46609	152	137
Anticodon-binding domain of class I aminoacyl-tRNA synthetases	47323	220	182
Cyclophilin (peptidylprolyl isomerase)	50891	112	98
Hemopexin-like domain	50923	103	73

Authors' contributions

AB and GP have contributed equally to the curation of the database. RS has supervised the study and provided input both in the design of the study and drafting of the final manuscript.

Acknowledgements

R.S. is a Senior Research Fellow funded by the Wellcome Trust, U.K. GP is supported by the Wellcome Trust. Financial and infrastructural support from NCBS (TIFR) is also acknowledged.

References

1. Rossmann MG, Moras D, Olsen KW: **Chemical and biological evolution of nucleotide-binding protein.** *Nature* 1974, **250:**194-199.
2. Lesk AM, Chothia C: **How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins.** *J Mol Biol* 1980, **136:**225-270.
3. Reddy BV, Li WW, Shindyalov IN, Bourne PE: **Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins.** *Proteins* 2001, **42:**148-163.
4. Bernstein FC, Koetzle TF, Williams GJ, Meyer Jr EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures.** *J Mol Biol* 1977, **112:**535-542.
5. Holm L, Sander C: **The FSSP database of structurally aligned protein fold families.** *Nucleic Acids Res* 1994, **22:**3600-3609.
6. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
7. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH-a hierarchic classification of protein domain structures.** *Structure* 1997, **5:**1093-1108.
8. Blundell TL, Bedarkar S, Rinderknecht E, Humbel RE: **Insulin-like growth factor I. a model for tertiary structure accounting for immunoreactivity and receptor binding.** *Proc Natl Acad Sci (USA)* 1978, **75:**180-184.
9. Chothia C: **Principles that determine the structures of proteins.** *Ann Rev Biochem* 1984, **53:**537-572.
10. Murthy MRN: **A fast method of comparing protein structure.** *FEBS Letts* 1984, **168:**97-102.
11. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G: **A database of protein-structure families with common folding motifs.** *Protein Sci* 1992, **1:**1691-1698.
12. Russell RB, Barton GJ: **Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility.** *J Mol Biol* 1994, **244:**332-350.
13. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds.** *Nature* 1994, **372:**631-634.
14. Sowdhamini R, Burke DF, Huang JF, Mizuguchi K, Nagarajaram HA, Srinivasan N, Steward RE, Blundell TL: **CAMPASS: a database of structurally aligned protein superfamilies.** *Structure* 1998, **6:**1087-1094.
15. Mallika V, Bhaduri A, Sowdhamini R: **PASS2: a semi-automated database of Protein Alignments Organised as Structural Superfamilies.** *Nucleic Acids Res* 2002, **30:**284-288.
16. Kleywegt GJ, Jones TA: **A super position.** *CCP4/ESF-EACBM Newsletter on Protein Crystallography* 1994, **31:**9-14.
17. Russell RB, Barton GJ: **Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility.** *Proteins* 1992, **14:**309-323.
18. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP: **JOY: protein sequence-structure representation and analysis.** *Bioinformatics* 1998, **14:**617-623.
19. Sali A, Blundell TL: **Definition of general topology equivalence in protein structures-a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming.** *J Mol Biol* 1990, **212:**403-428.
20. Sutcliffe MJ, Haneef I, Carney D, Blundell TL: **Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures.** *Protein Eng* 1987, **1:**377-384.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
22. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF: **Protein sequence similarity searches using patterns as seeds.** *Nucleic Acids Res* 1998, **26:**3986-3990.
23. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
24. Johnson MS, Overington JP, Blundell TL: **Alignment and searching for common protein folds using a data bank of structural templates.** *J Mol Biol* 1993, **231:**735-752.
25. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284:**1201-1210.
26. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology: applications to protein modeling.** *J Mol Biol* 1994, **235:**1501-1531.
27. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: extension and analysis of the basic method.** *CABIOS* 1996, **12:**95-107.
28. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:**903-919.
29. Bhaduri A, Ravishankar R, Sowdhamini R: **Conserved spatially interacting motifs of protein superfamilies: Application to fold recognition and function annotation of genome data.** *Proteins: Structure, Function and Bioinformatics* 2004, **54:**657-670.
30. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31:**3497-3500.
31. Chakrabarti S, Venkatramanan K, Sowdhamini R: **SMoS: a database of structural motifs of superfamilies.** *Prot Engng* 2003, **16:**791-793.

32. Kraulis PJ: **MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures.** *J Appl Cryst* 1991, **24:**946-50.