BMC Bioinformatics

BioMed Central

Database



# SUPFAM: A database of sequence superfamilies of protein domains

Shashi B Pandit[1], Rana Bhadra[1], VS Gowri[1], S Balaji[1], B Anand[1,2,3] and N Srinivasan*[1]

Address: [1]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India, [2]National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK campus, Bangalore 560 065, India and [3]Present address: Department of Biosciences and Bioengineering, Indian Institute of Technology, Kanpur, Kanpur – 208 016, India

Email: Shashi B Pandit - shashi@mbu.iisc.ernet.in; Rana Bhadra - rana@mbu.iisc.ernet.in; VS Gowri - gowri@mbu.iisc.ernet.in; S Balaji - sbalaji@mbu.iisc.ernet.in; B Anand - banand@iitk.ac.in; N Srinivasan* - ns@mbu.iisc.ernet.in

* Corresponding author

Published: 15 March 2004

*BMC Bioinformatics* 2004, **5**:28

Received: 30 December 2003
Accepted: 15 March 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/28



## Abstract

**Background:** SUPFAM database is a compilation of superfamily relationships between protein domain families of either known or unknown 3-D structure. In SUPFAM, sequence families from Pfam and structural families from SCOP are associated, using profile matching, to result in sequence superfamilies of known structure. Subsequently all-against-all family profile matches are made to deduce a list of new potential superfamilies of yet unknown structure.

**Description:** The current version of SUPFAM (release 1.4) corresponds to significant enhancements and major developments compared to the earlier and basic version. In the present version we have used RPS-BLAST, which is robust and sensitive, for profile matching. The reliability of connections between protein families is ensured better than before by use of benchmarked criteria involving strict e-value cut-off and a minimal alignment length condition. An e-value based indication of reliability of connections is now presented in the database. Web access to a RPS-BLAST-based tool to associate a query sequence to one of the family profiles in SUPFAM is available with the current release. In terms of the scientific content the present release of SUPFAM is entirely reorganized with the use of 6190 Pfam families and 2317 structural families derived from SCOP. Due to a steep increase in the number of sequence and structural families used in SUPFAM the details of scientific content in the present release are almost entirely complementary to previous basic version. Of the 2286 families, we could relate 245 Pfam families with apparently no structural information to families of known 3-D structures, thus resulting in the identification of new families in the existing superfamilies. Using the profiles of 3904 Pfam families of yet unknown structure, an all-against-all comparison involving sequence-profile match resulted in clustering of 96 Pfam families into 39 new potential superfamilies.

**Conclusion:** SUPFAM presents many non-trivial superfamily relationships of sequence families involved in a variety of functions and hence the information content is of interest to a wide scientific community. The grouping of related proteins without a known structure in SUPFAM is useful in identifying priority targets for structural genomics initiatives and in the assignment of putative functions. Database URL: http://pauling.mbu.iisc.ernet.in/~supfam.

## Background

Divergent evolution of proteins gives rise to homologues of similar three-dimensional (3-D) structure and, often, similar biochemical function [1,2]. However, extent of similarity in their amino acid sequences can be widely varying depending upon extent of divergent evolution [3]. A group of homologous proteins with high sequence similarity (typically above about 30% of sequence identity) is usually detectable easily from a large pool of sequences of diverse proteins using common sequence search tools and such proteins are suggested to form a family. However, homologous proteins with poor sequence similarity (less than about 25% of sequence identity), referred to as a superfamily, are more difficult to identify from sequence search methods [4]. Such remotely related homologues are often detected after the 3-D structures of proteins concerned are determined using X-ray analysis or Nuclear Magnetic Resonance (NMR).

Detection of remote homologues from sequence information alone, for the proteins without experimental structures, is a useful step in structural genomics initiatives. Moreover, proposing a reliable relationship between two proteins, one with known function and the other with unknown function, may enable prediction of putative function for the second protein if the functional residues are conserved. If one of these two proteins has already an experimental structure determined it could serve as a template for building 3-D structure of the remote homologue of unknown experimental structure [5,6]. Clustering of closely related and remotely related homologues with none of the proteins in the cluster with a known structure result in a new potential superfamily [7,8]. Such a collection of new potential superfamilies suggests priority targets for the structural genomics with a view to enhance the coverage of structural knowledge of proteins and fold space. Determination of experimental structure of at least one member per new potential superfamily can be useful in arriving at a framework structural model of all the members in the superfamily.

The SUPFAM database [7], first set up a couple of years ago, attempts to include sequence families of yet unknown structure in a known superfamily of known structures. It also groups sequence families, without a detectable relationship with a family of known structure, into new potential superfamilies. Both these steps are performed using sophisticated tools, which are capable of matching a sequence with sequence profiles generated from aligned sequences in protein families. An important feature of SUPFAM is that every periodic revision of the database results in many new non-trivial relationships between protein families. This feature is particularly due to the rapid growth in the sequence domain families apart from increase in the number of known domain superfamilies of known structure. Hence a visitor to SUPFAM site would find every release of SUPFAM containing complementary and radically reorganised information compared to earlier releases.

## Construction, content, enhancement and improvements

Compared to the earlier basic release of SUPFAM [7] couple of years ago, SUPFAM 1.4 has been generated using more improved sets of programs, more robust criteria for identifying hits and using a much larger datasets of sequence and structural families. A new web tool for profile matching is also introduced from SUPFAM 1.4. The general steps and strategies employed in the development of SUPFAM 1.4 are described in this section.

### *Databases*

We have used the Pfam database [9,10] that consists of a collection of large number of curated protein domain families with sequences of homologues in each family aligned accurately. In the current release of SUPFAM Pfam version 10.0 consisting of 6190 sequence domain families in the group of Pfam-A have been used. The protein domain families of known structure have been used from PALI database [11,12] that is derived largely from SCOP [4]. Every family in PALI consists of homologues of known X-ray or NMR structure and their amino acid sequences are aligned on the basis of their 3-D structures. The PALI release 2.2 used in the current version of SUPFAM contains 1029 multi-member protein domain families and 1288 orphans (single member families) with more than 35000 structure-based alignments and more than 1300 phylogenetic trees. The size of the sequence and structural family databases used in the construction of SUPFAM 1.4 is much higher than that used in the previous release.

### *Generation of position specific scoring matrices*

RPS-BLAST searchable profiles or Position Specific Scoring Matrices (PSSMs) were generated using PSI-BLAST [13] for every Pfam and PALI family. The Pfam family profiles were generated using PSI-BLAST by querying one of the family members against its own family members, with an additional input of multiple sequence alignment as provided in Pfam. For PALI family profile generation, as a first step PALI family sequences were integrated with homologous sequences from corresponding Pfam families or Non-Redundant sequence Database (NRDB). In the next step, profile is generated for every PALI family using PSI-BLAST, by a querying reference sequence from the PALI family against its own integrated structure sequence database. A high quality of profile is ensured by providing structure-based multiple sequence alignment as an input to PSI-BLAST.

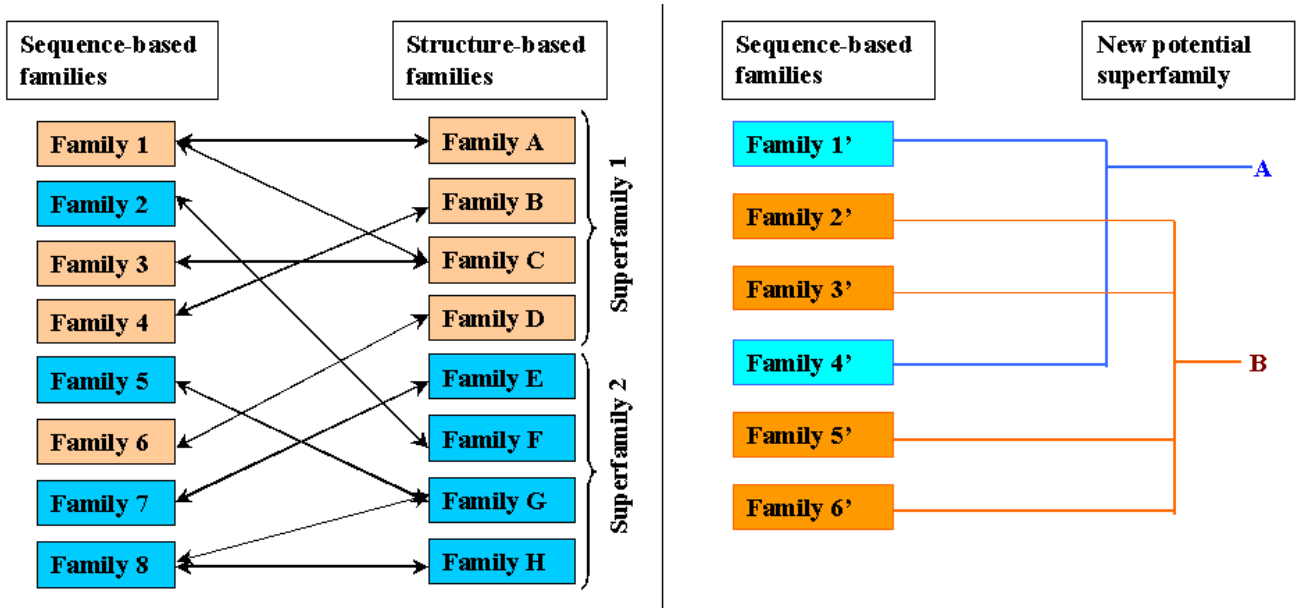


**Figure 1**
Left: Approach to relating sequence-based families and structure-based families with grouping of sequence based families into superfamilies. Right: Grouping of sequence-based families of yet unknown structure into new potential superfamilies.

***Relating protein families of unknown structure with families of known structure***

In the formation of SUPFAM database, the initial step involves relating the sequence families from Pfam with protein domain families of known structure from PALI [14]. The sensitive profile matching method RPS-BLAST available from NCBI has been used in querying for PALI proteins in the Pfam PSSM database and in querying Pfam proteins in the PALI PSSM database (see the left panel of Figure 1). An e-value cutoff of $3 \times 10^{-5}$ has been used to relate families, on the basis of the information given in the paper by Schaffer et al. [15] and using our benchmarking (B. Anand, N. Mhatre & N. Srinivasan, unpublished). We have listed all the connections, which pass the e-value cutoff of $10^{-4}$. Both these e-value cutoffs may be considered stringent although a few borderline cases of connections may exist with an e-value between $3 \times 10^{-5}$ and $10^{-4}$. We have imposed length of the alignment as a further constraint in order to increase the reliability of connections made between protein families. The region of alignment should be more than or equal to 40%, with respect to query or profile lengths. The connections arrived at by the slight relaxation of e-value cutoff are indicated by asterix in the SUPFAM web site. We have also consulted the curated information available in the Pfam flat files for the connectivity with a protein of known structure. These search conditions and criteria to detect hits are improved enormously compared to the basic version of SUPFAM released couple of years ago.

This exercise resulted in 2286 Pfam families with proposed relationship with a PALI/SCOP family. Interestingly 245 of these Pfam families, of yet unknown structure according to curation available in Pfam files, could be related to family of known structure. In an earlier version of SUPFAM (release 1.2) there were 315 Pfam families with no structural information documented in Pfam but could be related to a PALI version. Subsequent to the release of that version of SUPFAM, structures became available for the members of 110 of these Pfam families. Interestingly, the fold observed from the experimental structures for the members in 106 of these Pfam families matched with the predicted relationship derived earlier with the PALI resulting in ~96% of prediction accuracy (S. Namboori, N. Srinivasan, S.B. Pandit, manuscript communicated). The remaining 4 cases of wrong prediction corresponded often to the short alignments between the query and PSSM. In the current version of SUPFAM we have made the domain definition more robust by basing it more on structural domains in PALI rather than the sequence domains. Thus, it is expected to further increase the reliability of the associations between families.

### *Clustering of related protein families of unknown structure to form new potential superfamilies*

The next step in setting up SUPFAM involves clustering of protein families of yet unknown structure. There are 3904 Pfam families with no detectable relationship with a structural family in PALI. Sequences from each of these Pfam families have been queried in a database of PSSMs of all these Pfam families using RPS-BLAST (see the right panel of Figure 1). The same e-value cutoff of $3 \times 10^{-5}$ has been used in this step in order to relate families as well. However additional hits obtained by using the e-value cut-off of $10^{-4}$ are shown marked by an asterix.

This step clustered 96 of the 3904 Pfam families into 39 new potential superfamilies. It is expected that members of all the families in each new potential superfamily would share the same fold and might have gross similarity in their functional properties.

## Utility and discussion

The web interface of SUPFAM has search option for key words in the database and also profile searching against PALI and Pfam family profiles as described in Pandit et al [7]. Links to pages which list related sequence and structural families and new potential superfamilies are also provided. A separate link of Pfam families of yet unknown structure that are related to structural family is provided. From the current release of SUPFAM a new web-based tool is provided, which employs RPS-BLAST, in associating a query sequence with one of the protein domain families in SUPFAM. In the future updates it is envisaged that SUPFAM would integrate other sequence family databases, in addition to Pfam. The database of protein domains is growing rapidly as a result of genome sequencing projects. The SUPFAM is updated regularly to keep abreast with the growing number of known protein sequences. However, every release of SUPFAM database is complete reorganization of superfamilies, hence, the new information present in each release has insignificant overlap with the previous release.

There are several putative relationships proposed in the current release of SUPFAM between Pfam family, without structural information, and a structural family in PALI. For example, two Pfam domain families of unknown function, DUF227 and DUF954 are related to eukaryotic protein kinase-like superfamily in SCOP. Arabidopsis protein AIG1 which is suggested to be involved in plant resistance to bacteria in the Pfam database could be related to the PALI family of G-proteins which contains a variety of GTPases such as Ras, ARF and elongation factors.

The association of domain of unknown function (DUF) with a family of known functional property suggests a clue for the function of DUF. For example DUF1008 is related to HemS family which is a Haemin-degrading family. Among the various proposed interesting relationships the families of sarcosine oxidase, γ-subunit and aminomethyl transferase are suggested to be related.

The clustering of protein families of unknown function into new potential superfamilies would help in prioritizing the proteins for structural determination. The 96 families that are clustered into 39 new potential superfamilies consist of 14,595 members. The experimental determination of the structure of one representative member in each superfamily would result in 39 structures that can serve as framework models for 14,595 protein members. Hence, these new potential superfamily could form the priority target for structural genomics.

## Conclusion

The association of sequence families into superfamilies using 3-D structures enriches the sequence information within superfamilies. Such connections also enable the prediction of structures and functions of less well studied sequence families. Proposition of many new potential superfamily of proteins enables arriving at clues for the functions of constituent families within superfamilies. Hint for the functions of domains of yet unknown function is expected to be particularly useful. Further more, the list of new potential superfamilies suggest priority targets for structural genomics projects. For example, determination of the structure of a representative from each of the highly populated new potential superfamilies may be priority projects as these structures are expected to serve as templates for a large numbers of proteins.

With growing numbers of known 3-D structures of proteins many sequence families are being moved, in a rapid pace, to the category of structural families. Due to unprecedented growth in the sizes of the sequence databases, as genome sequencing projects progress, the size of sequence families are also growing rapidly. Due to rapid and constant enhancements in the size of structural and sequence families every release of SUPFAM is a complete reorganization of data. The list of sequence families apparently with no structural information, but could be related to known structures, and the list of new potential superfamilies have very little in common with the earlier release of SUPFAM.

## Availability and requirements

SUPFAM is publicly accessible without any restriction at the URL: http://pauling.mbu.iisc.ernet.in/~supfam.

## List of abbreviations

SCOP– Structural Classification Of Proteins. PALI– Phylogeny and Alignment of homologous protein structures. PSI-BLAST– Position Specific Iterative Basic Local Align-

ment Search Tool. PSSM– Position Specific Scoring Matrix. RPS-BLAST– Reversed Positions Specific BLAST. NCBI– National center of Biotechnology and Information. NRDB– Non-redundant database.

## Authors' contributions

S.B.P was involved in database development with R.B. who was involved in relating Pfam families. V.S.G, B.A and S.B were involved in providing the structure-based alignment of protein structures from PALI. NS conceived the idea and design of the database.

## Acknowledgements

S.B.P. is supported by a fellowship from Council of Scientific and Industrial Research (CSIR), New Delhi. R.B. is supported by the NMITLI project sponsored by CSIR. S.B. is supported by CSIR and the Wellcome Trust, London. B.A. is supported by the Wellcome Trust. This research is supported by the award of International Senior Fellowship in Biomedical Sciences to N.S from the Wellcome Trust, London, a computational genomics project sponsored by the Department of Biotechnology, New Delhi and a NMITLI project supported by CSIR.

## References

1. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5:**823-826.
2. Chothia C, Gerstein M: **Protein evolution. How far can sequences diverge?** *Nature* 1997, **385:**579-581.
3. Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214:**149-159.
4. 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.
5. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL: **Knowledge-based protein modeling.** *Crit Rev Biochem Mol Biol* 1994, **29:**1-68.
6. Sanchez R, Sali A: **ModBase: a database of comparative protein structure models.** *Bioinformatics* 1999, **15:**1060-1061.
7. Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R, Srinivasan N: **SUPFAM – a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes.** *Nucleic Acids Res* 2002, **30:**289-293.
8. Aloy P, Oliva B, Querol E, Aviles FX, Russell RB: **Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics.** *Protein Sci* 2002, **11:**1101-1116.
9. Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments.** *Proteins* 1997, **28:**405-420.
10. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
11. Balaji S, Sujatha S, Kumar SSC, Srinivasan N: **PALI: A database of Phylogeny and ALIgnment of homologous protein structures.** *Nucleic Acids Res* 2001, **29:**61-65.
12. Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S: **Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database.** *Nucleic Acids Res* 2003, **31:**486-488.
13. Altschul SF, Madden TL, Schaffer AA, Zhang JH, 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.
14. Elofsson A, Sonnhammer ELL: **A comparison of sequence and structure of protein domain families as a basis for structural genomics.** *Bioinformatics* 1999, **15:**480-500.
15. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF: **IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices.** *Bioinformatics* 1999, **15:**1000-1011.