SUPFAM - Database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: Implications for structural genomics and function annotation in genomes

Shashi B. Pandit[1], Dilip Gosar[1], S. Abhiman[1,2], S. Sujatha[1], Sayali S. Dixit[1,2,3], Natasha S. Mhatre[1], R. Sowdhamini[2] and N. Srinivasan[1*]

[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

[3]Biotechnology Centre
Indian Institute of Technology - Bombay
Powai
Mumbai 400 076
India

**Author for correspondence*
Phone: +91-80-309 2837; Fax: +91-80-360 0535
E-mail: ns@mbu.iisc.ernet.in

ABSTRACT

Members of a superfamily of proteins could result from divergent evolution of homologues with insignificant similarity in the amino acid sequences. A superfamily relationship is detected commonly after the three-dimensional structures of the proteins concerned are determined using X-ray analysis or NMR. The SUPFAM database described here relates two homologous protein families in a multiple sequence alignment database of either known or unknown structure. The present release (1.1) which is the first version of the SUPFAM database has been derived by analysing Pfam which is one of the commonly used database of multiple sequence alignments of homologous proteins. The first step in establishing SUPFAM is to relate Pfam families with the families in PALI which is an alignment database of homologous proteins of known structure that is derived largely from SCOP. The second step involves relating Pfam families which could not be associated reliably with a protein superfamily of known structure. The profile matching procedure, IMPALA, has been used in these steps. The first step resulted in identification of 1280 Pfam families (out of 2697 - 47%) which are related, either by close homologous connection, to a SCOP family or by distant relationship to a SCOP family potentially forming new superfamily connections. Using the profiles of 1417 Pfam families with apparently no structural information an all-against-all comparison involving sequence-profile match using IMPALA resulted in clustering of 67 homologous protein families of Pfam in to 28 potential new superfamilies. Expansion of groups of related proteins of yet unknown structural information, as proposed in SUPFAM, should help in identifying "priority proteins" for structure determination in structural genomics initiatives to expand the coverage of structural information in the protein sequence space. For example we could assign 858 distinct Pfam domains in 2203 of the gene products in the genome of *Mycobacterium tubercolosis*. 51 of these Pfam families of unknown structure could be clustered into 17 potentially new superfamilies forming good targets for structural genomics. SUPFAM database can be accessed at http://pauling.mbu.iisc.ernet.in/~supfam.

INTRODUCTION:

The number of proteins of known amino acid sequences is known to be overwhelmingly higher than the number of proteins of known three-dimensional (3-D) structures (e.g. see 1). This problem is compounded by rapid accumulation of very large number of sequences of putative protein products from the genome sequencing projects. As one of the benefits of structural genomics initiatives, the gap between the number of proteins of known amino acid sequence and those with known 3-D structures is expected to progressively come down with time (2-5). Structural genomics efforts carry the promise of the availability of structural information for all the globular proteins encoded in the genomes of various organisms by determining X-ray analysis or NMR structures of either all the proteins or a set of proteins whose structures serve as model systems for other proteins.

It is now well known that while the homologous proteins are characterised by high degree of structural resemblance (e.g. 6), many proteins with no detectable similarity in the amino acid sequences can also adopt similar structures (7). It is hoped that the experimental structures of a number of proteins will serve as templates for generating reliable 3-D models of other proteins (3). Therefore the structural genomics efforts could be effective if, among other factors, the list of proteins for experimental structure determination is carefully identified and by prioritizing for experimental determination of these "representative" proteins (2,5,8).

Identification of proteins for such experimental structure determination involves, as one of the first steps, relating all the proteins of known structure with those of unknown structure (e.g. 2, 9-12). In general the set of proteins of unknown structure can be clustered into families of homologous proteins based on similarities in their amino acid sequences and a representative from each of these families could serve as reasonable targets for experimental structure determination. The present database is the result of an effort along this direction as well as to progress one step further in grouping homologous protein families of unknown structure in forming potential new superfamilies. It is hoped

that this approach results in more effective set of priority proteins for structure determination in structural genomics. Deriving such superfamilies of homologous families also widens-up the scope of function annotation of a gene product with clear, but remote, similarity with a known homologous family which is a member of a superfamily.

RELATING SEQUENCE AND STRUCTURE-BASED DOMAIN FAMILIES OF HOMOLOGOUS PROTEINS:

One of the important components of SUPFAM is the relationship between sequence and structural alignment databases of homologous proteins. In the present release of SUPFAM we have used PALI (13) and Pfam (14,15) which are homologous structure and sequence alignment databases respectively. PALI database is essentially derived by aligning non-identical homologous proteins of known 3-D structure derived from the SCOP (7) database and by generating structure-based phylogenetic relationships. PALI contains structure-based sequence alignments of multiple proteins in a family apart from all possible pairwise alignments. Pfam database has multiple sequence alignment of homologous proteins irrespective of the availability of 3-D structural information. Relating SCOP and Pfam families has also been performed by Elofsson and Sonnhammer (9) about couple of years ago using Pfam and SCOP releases available at that time.

The program IMPALA (16) which performs search for the match of a queried sequence with the Position-Specific Score Matrices (PSSM) of families (also referred to as profiles) has been used to associate sequence and structure-based families. PSI-BLAST (17) has been used to construct the PSSMs of all the families in PALI. Sequences of protein domains in PALI have been embedded in the Non-Redundant sequence Data Base (NRDB) before employing PSI-BLAST to generate the PSSMs. Integration of PALI and NRDB is hoped to enable use of homologous proteins of unknown experimental structure in the generation of PSSM of a family. Multiple structure-based alignment of the proteins in a PALI family has been supplied as an input to the PSI-BLAST in order to maintain the high quality (structure-based) alignment in the generation of PSSM of the family. An e-value cut-off of 10^{-4} has been used in the profile generation in the

successive cycles as well as in the homologue detection using PSI-BLAST. A PSI-BLAST run is made until no more new homologues are detected or up to 20 cycles whichever is earlier. The output of the PSI-BLAST for almost every family in PALI has been subjected to manual scrutiny to eliminate false positives if present. Profiles of all the 1249 PALI families have been generated. PSSM of every family of Pfam has been generated using the multiple sequence alignment of "seed proteins" given for every family in the Pfam database. The present release of SUPFAM uses profiles generated for 2697 families in a recent release of PfamA database. A search tool has been linked to the SUPFAM web site that can enable a user to search for a queried sequence, using IMPALA, in the PALI or Pfam profile databases. In our IMPALA searches a very stringent e-value cut off of 3×10^{-5} has been used in order to minimise the chances of picking-up false positives at the cost of missing a few true positives. A detailed manual analysis of information about each of the Pfam families was also performed to identify cases with direct relationship with a known structure that are not picked-up by IMPALA. Further potentially true relationships have been obtained using a less-stringent e-value cut-off of 3×10^{-4}. However the additional connections thus obtained are tentative and await confirmation using different sequence-structure analysis methods. We had already benchmarked IMPALA against the "SCOP genome" in order to understand the effectiveness of IMPALA in identifying proteins at the level of family, superfamily and common fold (N.S.M. and N.S., manuscript in preparation). Several superfamilies in SCOP have been recognised by IMPALA very successfully without using any structural information as also noted earlier by Schaffer et al (16). Information from this exercise was also useful in manual decision making about the relatedness or otherwise of a Pfam family to a SCOP/PALI family.

Using extensive analysis of forward (search for Pfam sequences against PALI profile database) and reverse (search for PALI sequences against Pfam profile database) searches a total of 1280 families of Pfam (47%) could be associated with families of SCOP/PALI. Most of these Pfam families are also documented to have 3-D structural information. 1097 of the PALI families (88%) could be associated to one or more Pfam family. It is important to note that the basic objectives and philosophy behind the

generation of PALI and Pfam databases are distinct and hence exact one-to-one family correspondence can not be expected. For example, α-subunit of heterotrimeric G-proteins, Ras proteins, ADP ribosylation factors and elongation factors form same homologous family in SCOP and PALI. In Pfam these proteins are separated into different families. These 1097 PALI families exist in 719 superfamilies of SCOP. Thus the Pfam families associated with 1097 PALI families automatically become members of 719 SCOP superfamilies.

At least 119 of the 1280 Pfam families connected to a structural family or superfamily using IMPALA are not documented to have a relationship with 3-D structures. Sequences in number of these Pfam families have been subjected to fold recognition using GENTHREADER (18) and the results are highly consistent with the relationships established using IMPALA. For example, Pfam families with description "Protein of unknown functon DUF 114" and "Protein of unknown function DUF 107" could be reliably associated with the superfamily of ClpP/crotanase in SCOP using IMPALA as well as GENTHREADER. This raises the possibility of DUF 114 and DUF 107 Pfam families to have a fold and gross biochemical property similar to those of ClpP/crotanases. Similarly, the families of replication proteins E1 and parvovirus non-structural protein NS1 could be reliably associated with the superfamily of NAD(P) binding Rossmann fold domains.

CLUSTERING OF HOMOLOGOUS PROTEIN FAMILIES OF UNKNOWN STRUCTURE TO FORM POTENTIAL NEW SUPERFAMILIES.

A total of 1417 families in Pfam could not be associated with a known structural family in PALI using the profile matching methods. Thus in principle a good representative from each of these families could be potential priority targets in structural genomics projects. A total of 21382 seed alignment proteins present in these families will have a structural neighbour if 1417 structures are determined using experimental methods. Figure 1 shows the distribution of number of members in 67 of such families which are subsequently clustered into 28 superfamilies. The most populated of the families (corresponding to the

family numbers 12, 17, 43 and 16 in figure 1) are tetraspanin family, glycosyl transferases group 1, moaA / nifB / pqqE family and nucleotidyl transferase family with 61, 78, 59 and 66 seed members respectively. Determination of the structure of one representative protein from each of these four families could form templates for 264 (=61+78+59+66) proteins.

However it may be possible to cluster the Pfam families with no connection to a structural family and these may correspond to new superfamilies of yet unknown structure. The commonly used methods for this purpose involve PSSM or Hidden Markov Models (HMM). These approaches are basically equivalent and are shown to be very powerful (15,16). The PSSM matching procedure, IMPALA (16), is capable of rapid matching of large volume of sequences with the profiles of protein families as it takes significantly less computer time. Hence we used IMPALA in making all-against-all sequence-profile matching for 1417 Pfam families of unknown structure. 67 families of Pfam could be clustered into 28 potential new superfamilies. Figure 2 shows the distribution of number of proteins present in the seed alignments of 67 families that make these superfamilies. It is possible that some of these superfamilies have adopted already known folds although the connection at the level of common fold may not have been recognised using the methods used in this work. The most populated families which are numbered as 12, 17, 43 and 16 in figure 1, belong to potential superfamilies which are numbered as 5, 7, 18 and 6 respectively in figure 2. The number of seed members in these superfamilies are 115, 104, 120 and 92 respectively. Thus determination of experimental structures of 4 proteins from each of these four families form templates for 431 (=115+104+120+92) proteins as opposed to 264 proteins proposed before the clustering of Pfam families into potential superfamilies.

An example of superfamilies of yet unknown structural information and recognised by IMPALA is described here: The two Pfam families (a) Dolichyl-phosphate-mannose-protein mannosyltransferases and (b) oligosaccharyl transferase STT3 subunits could be related by profile-matching techniques. The Drosophila rotated abdomen protein belonging to the first of the families above is suggested to be a putative

mannosyltransferase although, in general, this family of proteins are known to be involved in the O-linked glycosylation of proteins. The second family above corresponds to oligosaccharyl transferase STT3 subunit and related proteins. The STT3 subunit is part of the oligosaccharyl transferase (OTase) complex of proteins and is known to be required for its activity. OTase transfers a lipid-linked core-oligosaccharide to selected asparagine residues in the ER. Thus, interestingly, these two Pfam families of unknown structure could not only be related by sequence-profile matching technique, but, also by high similarity in their suggested biochemical properties.

Related integral membrane protein families could also be recognised and documented in SUPFAM. For example, the Pfam families (a) Patched family whose members are known to associate with the smoothened protein to transduce hedgehog signals, (b) AcrB/AcrD/AcrF family of proteins which are known to be integral membrane proteins and involved in drug resistance and (c) Protein export membrane protein family which consists of prokaryotic SecD and SecF proteins, are grouped together as a potential superfamily.

The 28 potential new superfamilies encompass 67 families of yet unknown structure and 1157 proteins that make seed alignments in Pfam. Choosing a good representative from each one of these 28 superfamilies as high priority targets in structural genomics seems to be more effective than choosing representatives from 67 families as both these sets of experimental structures will serve as templates for same number of 1157 proteins.

Function annotation to gene products in genomes by matching the amino acid sequence of the gene products with Pfam families is common. If a gene product is remotely related to all the members of a Pfam family it is not conclusive that the gene product has same or similar function as that of the Pfam family members. Conservation of critical residues important for the function, in the gene product, points to the strong possibility of the same function as the Pfam family members. However, in practice, the mapping of functionally important residues is unavailable especially in several of the

families of homologous proteins with yet unknown 3-D structural information. Under these circumstances it is possible that the gene product under examination may have inherited the function of any one of the homologous protein families that make a superfamily.

POTENTIALLY NEW SUPERFAMILIES OF UNKNOWN STRUCTURE ENCODED IN THE GENOME OF *MYCOBACTERIUM TUBERCOLOSIS*:

We have analysed all the gene products coded in the genome of *Mycobacterium tubercolosis* (MTb) (19), using IMPALA, for the match with the profiles of families in Pfam (S.S and N.S, unpublished results). 2203 of the gene products of MTb could be associated with at least one PfamA family. There are 2702 regions of gene products that match with 858 distinct Pfam families. A total of 418 of these Pfam families could not be associated with a known structural family/superfamily. Interestingly 51 of these Pfam families form 17 of the 28 new superfamilies derived. These 17 superfamilies (or 51 families) are coded in about 200 regions of gene products in MTb genome suggesting that experimental structures for 17 representative proteins could form templates for about 200 domains coded in MTb genome.

One of the uses of clustering of sequence families into potential superfamilies could be demonstrated by an example from MTb genome. The gene product with the identification code Rv0823c has significant match with the Pfam family with the description "Uncharacterized protein family UPF0034". However this Pfam family shows a good match in the sequence-profile comparisons with another Pfam family with the description "Histidine biosynthesis protein" enabling grouping of these two families to form a potential new superfamily. Thus it is possible that an "Uncharacterized protein family" in Pfam might perform a function which is closely or remotely similar to that of "Histidine biosynthesis protein". This also raises the possibility of the biochemical properties of Rv0823c being somewhat similar to that of Histidine biosynthesis proteins.

OUTLOOK:

Recognition of relationships between sequence families and the structural families/superfamilies is increasingly becoming effective with the power of profile matching, HMM-based methods and fold recognition techniques. A careful use of these techniques can reduce the list of priority targets in structural genomics. Also, clustering of sequence families with yet unknown structure can enable recognition of potential new superfamilies and also contribute in the further reduction of priority targets in structural genomics.

It is also common in the large-scale genome analysis that a significant and reliable match, but with low sequence similarity, is obtained for a gene product with a Pfam family. Under such circumstances it is difficult to conclude that the function of the new gene product will be same as that of the matched Pfam family. If this Pfam family is clustered with other families to form a superfamily the functions of proteins making different families in the superfamily could be possible functions of the gene product.

The first version of SUPFAM is an effort towards aiding function association in genome analysis by remote homology detection and in choosing priority proteins for structure determination in structural genomics initiatives. A wealth of new connections made between a Pfam family and a structural family/superfamily and clustering of some of the Pfam families into potentially new superfamilies are expected to provide new insights in the evolution of the proteins involved and will undoubtedly influence genome annotations. Updates of SUPFAM will happen periodically by using not only Pfam, but, also other similar sequence alignment databases and their relationships to SCOP.

SUPFAM can be accessed at http://pauling.mbu.iisc.ernet.in/~supfam.

ACKNOWLEDGEMENTS:

DG, SA. and SS. are supported by The Wellcome Trust, U.K. This research is supported by the award of International Senior Fellowships in Biomedical Sciences to RS and NS from the Wellcome Trust, U.K.

LEGENDS TO FIGURES

Figure 1: Number of proteins in the seed alignment of 67 Pfam families.

Figure 2: Number of proteins in the seed alignment of 67 Pfam families in figure 1 that are clustered into 28 potential new superfamilies.

REFERENCES:

1. Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science*, **273,** 595-603.
2. Brenner, S.E. and Levitt, M. (2000) Expectations of structural genomics. *Protein Sci*., **9,** 197-200.
3. Sanchez, R., Piper, U., Melo, F., Eswar, N., Marti-Renom, M.A., Madhusudhan, M.S., Mirkovie, N. and Sali, A. (2000) Protein structure modeling for structural genomics. *Nature Strl. Biol*., **7,** 986-990.
4. Thornton, J.M., Todd, A.E., Milburn, D., Borakoti, N. and Orengo, C.A. (2000) From structure to function: Approaches and limitations. *Nature Strl. Biol*., **7,** 991-994.
5. Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) Completeness in structural genomics. *Nature Strl. Biol*., **8,** 559-566.
6. Balaji, S. and Srinivasan, N. (2001) Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Engng.,* **14,** 219-226.
7. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536-540.
8. Brenner, S.E., Barken, D. and Levitt, M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Res.,* **27,** 251-253.
9. Elofsson, A. and Sonnhammer, E.L.L. (1999) A comparison of sequence and structure of protein domain families as a basis for structural genomics. *Bioinformatics,* **15,** 480-500.
10. Bray, J.E., Todd, A.E., Pearl, F.M.G., Thornton, J.M. and Orengo, C.A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Engng.,* **13,** 153-165.
11. Lindahl, E. and Elofsson, A. (2000) Identification of Related Proteins on Family, Superfamily and Fold Level. *J. Mol. Biol.,* **295,** 613-625.
12. de Bakker, P. I. W., Bateman, A., Burke, D.F., Miguel, R.N., Mizuguchi, K., Shi, J, Shirai, H. and Blundell, T.L. (2001) HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics,* **17,** 748-749.
13. Balaji, S., Sujatha, S., Kumar, S.S.C. and Srinivasan, N. (2001) PALI: A database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.* **29,** 61-65.
14. Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998) Pfam multiple sequence alignment and HMM-profiles of protein domains. *Nucleic Acids Res.,* **26,** 322-325.
15. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L.L. (2000) The Pfam Protein Families Database *Nucleic Acids Res.,* **28,** 263-266.
16. Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics,* **15,** 1000-1011.

17. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25,** 3389-3402.
18. Jones, D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol*., **287,** 797-815.
19. Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D.,Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell,T., Gentles, S., Hamlin, N.,Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, S., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, S., Squares, S., Sqares, R., Sulston, J.E., Taylor, K., Whitehead, S. and Barrell, B.G. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393,** 537-544.

Figure 1:

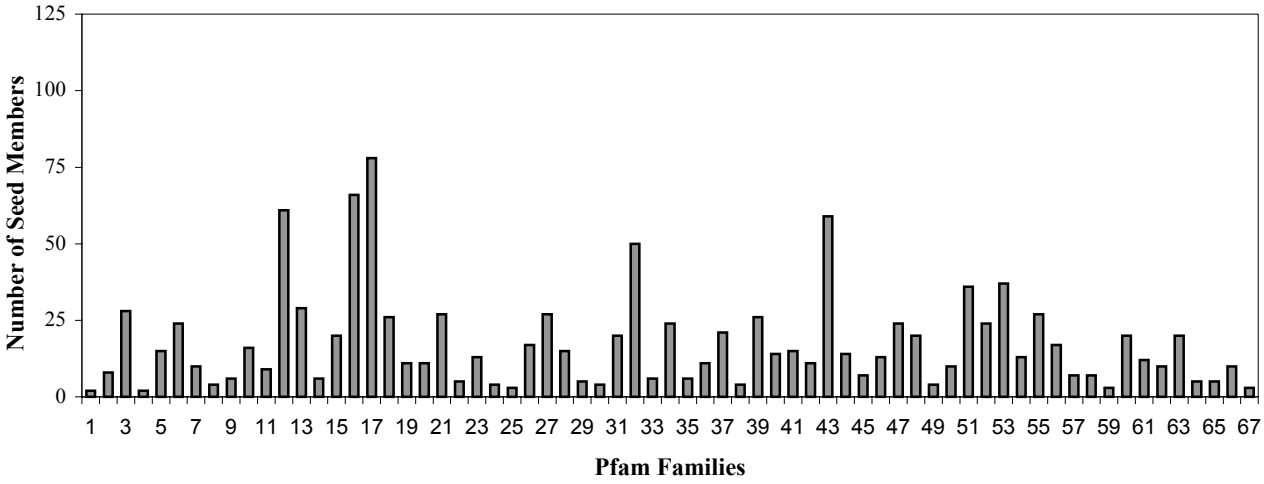
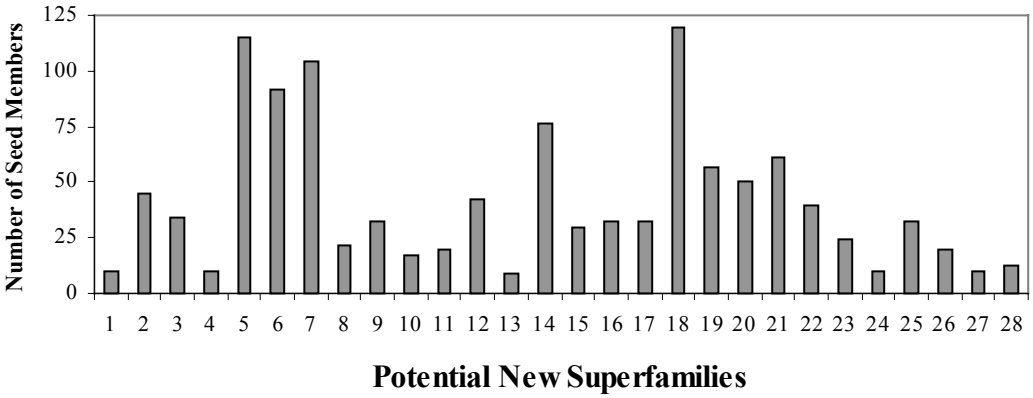

Number of Seed Members
125
100
75
50
25
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67
Pfam Families

Figure 2:


Number of Seed Members
125
100
75
50
25
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Potential New Superfamilies