

Biological Network Evolution Hypothesis Applied to Protein Structural Interactome

Dan M. Bolser¹ and Jong Hwa Park^{1,2*}

¹MRC-DUNN Human Nutrition Unit, Hills Road, Cambridge, CB2 2XY, England, UK

²Object Interaction Technologies Inc. (OITEK), Seoul, Korea

Abstract

The latest measure of the relative evolutionary age of protein structure families was applied (based on taxonomic diversity) using the protein structural interactome map (PSIMAP). It confirms that, in general, protein domains, which are hubs in this interaction network, are older than protein domains with fewer interaction partners. We apply a hypothesis of 'biological network evolution' to explain the positive correlation between interaction and age. It agrees to the previous suggestions that proteins have acquired an increasing number of interaction partners over time via the stepwise addition of new interactions. This hypothesis is shown to be consistent with the scale-free interaction network topologies proposed by other groups. Closely co-evolved structural interaction and the dynamics of network evolution are used to explain the highly conserved core of protein interaction pathways, which exist across all divisions of life.

Keywords: Network Evolution, Structurefamily Evolution, Protein Interaction, Protein Structural Interactome, PSIMAP, Interactomics.

Introduction

There are around 300 distinct classification schemes used to relate over 140,000 species and sub-species in the NCBI Taxonomy database (Wheeler *et al.*, 2000) (July 2002). This 'tree of life' classifies species into four superkingdoms, namely: eukaryota, eubacteria, archaea and viruses. The huge diversity of life is the result of billions

of years of evolution on Earth. However, the basic core of protein mediated metabolic pathways in all these species is relatively homogeneous (Benner *et al.*, 1989; Morowitz, 1992; Morowitz, 1999). Furthermore, despite the continuing growth in the quantity of determined protein structures, sequences and even whole genomes, the rate of finding novel protein topologies is decreasing (Fig. 1). It is probable that there are no more than 2,000 distinct protein topologies in nature (Chothia, 1992; Orengo *et al.*, 1994; Alexandrov *et al.*, 1995; Wang, 1996; Zhang, 1997). One can ask how such an ancient and diverse evolutionary history could maintain such a homogenous biochemical backbone, supported by so few protein topologies. What constraints prevent life from using unique biochemical pathways and discovering new protein folds? The proposed scale free topology of the interaction network (Jeong *et al.*, 2000), the structural interaction network (Park *et al.*, 2001), the closely co-evolved nature of protein interactions (Bennett *et al.*, 1994; Marcotte *et al.*, 1999; Fraser *et al.*, 2002) and the rate of network evolution (Kauffman *et al.*, 1993) contribute significantly to an account of these observations. We have proposed that protein interaction networks are conserved in evolution and highly interacting groups are relatively old and functionally important (Park and Bolser, 2001). Here, we explain it further with the latest data by using the old concept of biological Network Evolution applied to protein structural interactome.

Currently the structural classification of proteins database (SCOP) (Murzin *et al.*, 1995) defines around 1,000 distinct protein fold types as Superfamilies (termed as structurefamilies in this paper), denoting homology between their representative protein structural domains (domain) members. A fold in SCOP is defined solely on the basis of structural similarity between domains. Structurefamilies therefore divide folds into evolutionary groups, using sequence and functional similarities. However, around 90% of the folds defined in SCOP are thought to have a single evolutionary origin, constituting a single superfamily. Structurefamilies (superfamilies) are the most useful domain classification for comparing structures and functions in bioinformatics by virtue of their structural and phylogenetic classification.

Although the rate at which new protein structures are deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000) is increasing (Fig. 1c), the rate of discovering new structurefamilies is decreasing (Fig. 1b). The recent

* Corresponding author: E-mail j@bio.cc, <http://networkevolution.org>

Accepted 23 February 2003

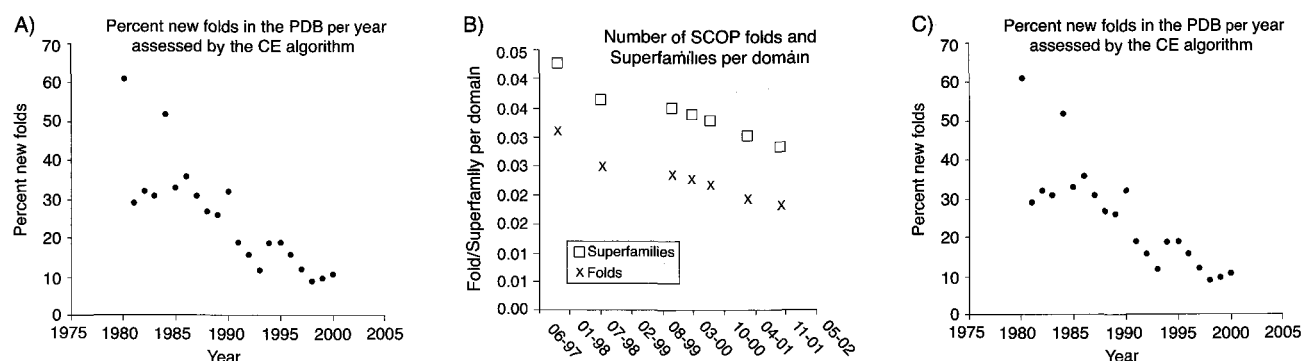


Fig. 1. (A) The number of new folds added to the PDB (Berman *et al.*, 2000) since 1980. The trend shows a general decrease, with a slight recent upturn. New folds are assessed by the combinatorial extension (CE) algorithm (Shindyalov & Bourne, 1998). (B) The number of SCOP folds, superfamilies and families per domain (Murzin *et al.*, 1995) in the PDB. The trend shows increasing domain redundancy within the groups. (C) The total number of new structures deposited in the PDB per year. Both the trends in A and B oppose the increasing number of structures deposited in the PDB each year.

conservative structurefamily assignment of 56 genomes covered between 40-67% of the total detected genes in eukaryotes and eubacteria (~ 100,000 genes) and between 31-54% of the total detected genes in archaeobacteria (~ 10,000 genes) (Gough *et al.*, 2001). Given that a significant portion of the unassigned genes may represent trans-membrane and other proteins, not assigned to structures due to experimental difficulty in structure determination, it is reasonable to suggest that there are now enough soluble protein structurefamily data in the PDB to make a global map of structurally observed structurefamily interactions. PSI-MAP (Protein Structural Interactome MAP) (Park *et al.*, 2001) is the first such map (Fig. 2). It also compared for the first time the protein experimental interaction information such as yeast two hybrid system (Uetz *et al.*, 2000) with structural interaction information.

The criteria for assigning interactions in PSI-MAP is strictly structural and exhaustive; distinct pairs of domains in the PDB are denoted as interacting if they share 5 or more residue - residue contacts within 6 angstroms or less (5-5 rule of protein structure interaction). These criteria were chosen as being the most discriminative within a range of other criteria (Fig. 3). Different contact algorithms yield qualitatively similar results (Park *et al.*, 2001). By using the SCOP domain definition (version 1.59 unless otherwise stated) it is possible that these criteria will denote covalently linked domains as interacting. These interactions (intra-interaction) are in the minority, accounting for 30% of the (11281) domain-domain interactions observed. For a breakdown of the 651 observed structurefamily-structurefamily interactions see

Table 1. The number of structurefamilies displaying interaction through both covalently and non-covalently linked domains (73 interacting structurefamilies representing 3,337 interacting domains) indicates that observed domain fusion events in the PDB are extensive. The validity of assigning the only intra-interacting structurefamily pairs as interacting is two fold. Firstly, domain proximity is a result of the selective pressure to associate genes that physically interact (Marcotte *et al.*, 1999; Dandekar *et al.*, 1998; Doolittle 1999; Enright *et al.*, 1999). Secondly, domain proximity is more generally indicative of indirect functional associations between domains (Marcotte *et al.*, 1999; Overbeek *et al.*, 1999; Enright & Ouzounis, 2001). Domain fusion has been successfully used to predict protein interaction from sequence information alone (see Huynen *et al.*, 2000) and as a hypothesis for the evolution of homo (Bennett *et al.*, 1994) and hetero (Marcotte *et al.*, 1999) dimers (see Table 1 for PSIMAP multimer information). In addition, it has been observed that intra-domain interfaces have strong similarities to inter-domain interfaces within multi-domain proteins (Miller, 1989; Tsai *et al.*, 1996; Jones *et al.*, 2000).

Using the expert SCOP domain and superfamily definitions to predict superfamily-superfamily interactions from observed domain fusion events overcomes some of the technical problems associated with the identification of homology and fusion encountered using other computational methods to predict interaction (Overbeek *et al.*, 1999; Enright & Ouzounis, 2001). PSI-MAP therefore represents a robust and reliable method of computationally predicting protein interaction.

It has been argued that the PDB is a fair representation

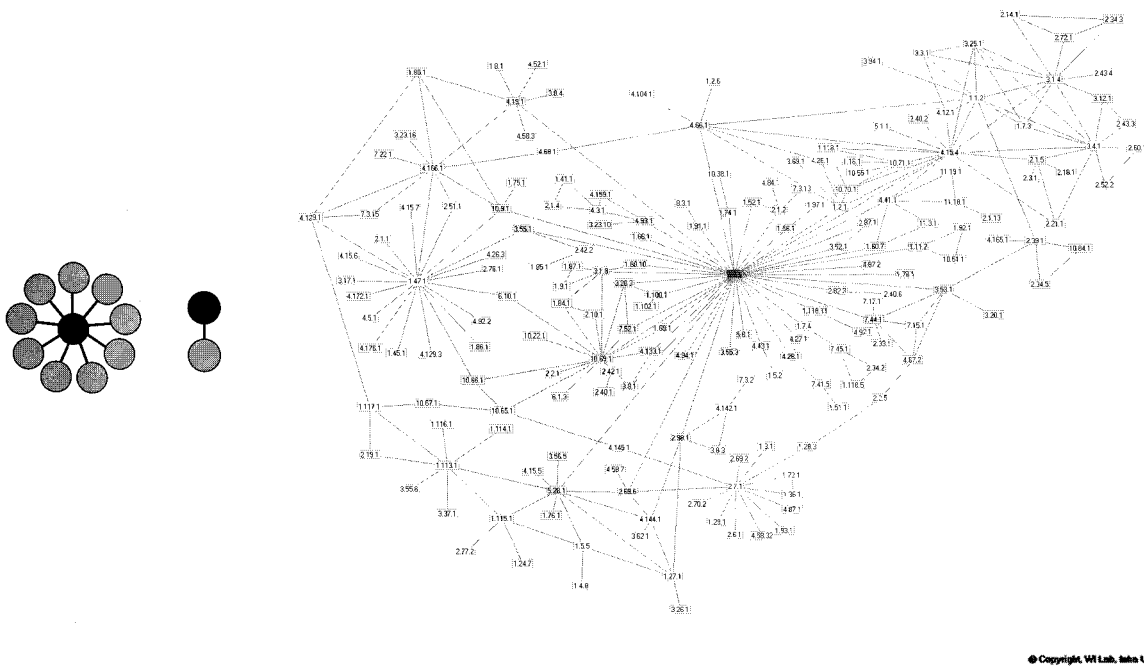


Fig. 2. (Right; one of the big sub-networks in PSI-MAP shown as visualised by the layout drawing program INTERVIEWER (Ju, et al., 2003). Protein structurefamilies directly interacting with the NAD(P)-binding Rossmann-fold superfamily (c.2.1) shown in blue are highlighted in yellow. Left; two contrasting structurefamilies in PSI-MAP are represented. The red circle on the left has nine interaction partners, for example the Protein kinase-like (d.144.1) superfamily. The red circle on the right has only one interaction partner (a pairwise interaction), like many structurefamilies in PSI-MAP.

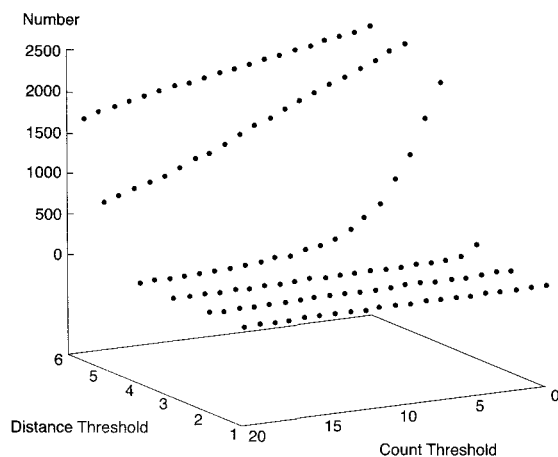


Fig. 3. Shows the different number of structurefamily - structurefamily interactions (Y-axis) observed at different mean residue centre distance thresholds (Z-axis) and different number of residue - residue contact count thresholds for defining two domains as interacting (X-axis). Requiring 5 residue - residue contacts or more at 6 angstroms or less gives a good cut-off for structurefamily - structurefamily interactions (highlighted above). Different criteria give qualitatively similar results.

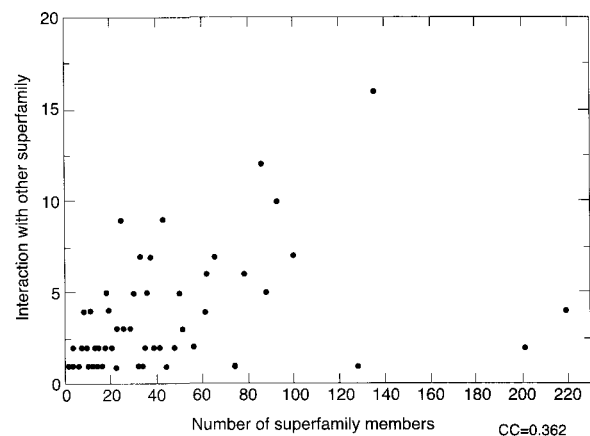


Fig. 4. Graph of the number interacting structurefamily domains versus the number of structurefamily interactions for each structurefamily produced using. The observed correlation is weak (0.36), and is reduced by the removal of outliers.

Table 1. Number of hetero and homo multimer interactions in PSIMAP broken down into inter chain, intra chain and structurefamilies showing both inter and intra chain interactions in the PDB.

	HOMO MULTIMER	HETERO MULTIMER	TOTAL
INTRA	26 (4)	147 (23)	173 (27)
INTER	287 (44)	118 (18)	405 (62)
BOTH	44 (7)	29 (4)	73 (11)
TOTAL	357 (55)	294 (45)	651 (100)

Values in parentheses give percent of the total number of structurefamily interactions. The large proportion of inter chain homo multimer contacts in the PDB could be the result of unrecognised crystal contact structures. In future the Protein Quaternary Structure (PQS) server (Henrick & Thornton, 1998) will be used to construct PSIMAP to alleviate this problem

of all the soluble protein structurefamilies which may exist. However, given the combinatorial effect, it is unlikely that the PDB covers a representative set of pair-wise structurefamily interactions. Importantly, extending the repertoire of predicted structurefamily-structurefamily interactions by using structurally annotated genomic sequence data does not alter the distribution of observed interactions (Park *et al.*, 2001; Apic *et al.*, 2001a; Apic *et al.*, 2001b). Thus it is likely that the relative distribution of interactions in PSIMAP (which forms the basis of our results and discussion) will reflect the distribution of a hypothetical 'complete' map.

Another criticism which has been levelled at PSI-MAP is that the variance in the number interactions assigned to a structurefamily could be biased by the number domains in the PDB assigned to that structurefamily. Fig. 4 shows that only a weak correlation exists between the number of domain interactions for a structurefamily and the number of unique structurefamily interactions it has. This correlation coefficient falls to 0.16 upon removal of the four most prominent outliers. A similar correlation is measured between the number of structurefamily interactions and the absolute number of domains for that structurefamily (data not shown).

It has been suggested that artificial structures in the PDB may affect the overall distribution of structurefamily interactions discussed here. PSI-MAP is constructed using only structurefamilies from SCOP class 1 to 4. In all, there are only 12 multi-domain synthetic proteins in these classes.

Results and Discussion

High and low interaction structurefamilies

PSI-MAP was used to identify all the structurally observed

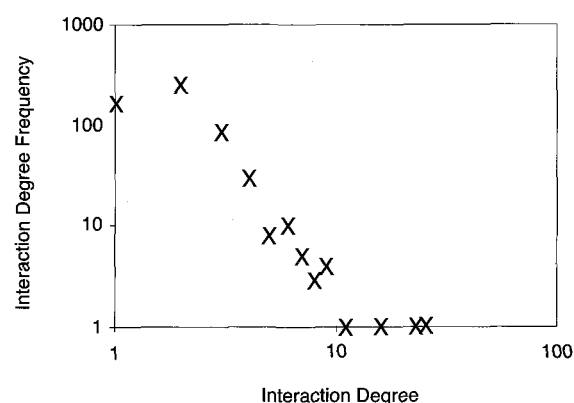


Fig. 5. A log-log plot of the number of structurefamily interactions against the frequency of structurefamilies with this number of interactions. The linear fit to a log-log plot indicates a power law frequency distribution, which is broken by the 'one interaction partner' class of structurefamilies. Traces of the familiar dove tale distribution can be seen from left to right, caused by an increased variance associated with lower frequency events. The upper left hand side, however, is usually quite linear, suggesting that the lower than expected number of 'one interaction partner' structurefamilies may be significant. Linear regression of the log-log transformed data gives a correlation coefficient of 0.9.

interactions at the structurefamily level. Structurefamilies have various degrees of 'interactability', and the interaction frequency distribution obeys a power law (Fig. 5).

To assess the functional and evolutionary differences between the most interactive and the least interactive folds, we use the latest HIINFOLD and LOINFOLD comparison sets (Park and Bolser, 2001): high interaction structurefamilies (HIINFOLD, see supplement Table A) and low interaction structurefamilies (LOINFOLD, see supplement Table B). The sixteen HIINFOLD structurefamilies (with at least seven other interacting partners) have functions related to glycolysis; oxidative phosphorylation; catabolism and nucleotide syntheses, well as DNA binding, replication and metabolic regulatory processes. The group contains functionally important domains, often and found in core biochemical pathways (Park *et al.*, 2001; Apic *et al.*, 2001a; Apic *et al.*, 2001b). By contrast the 160 LOINFOLD structurefamilies (each with only one structurefamily interaction) contains only 91 (57%) structurefamilies with at least one assigned enzyme classification (see methods section for details of the functional assignment), covering a total of ~130 distinct enzyme reactions.

The latest functional analysis of HIINFOLD and LOINFOLD supports the previous observation that the absolute number of protein-protein interactions correlates

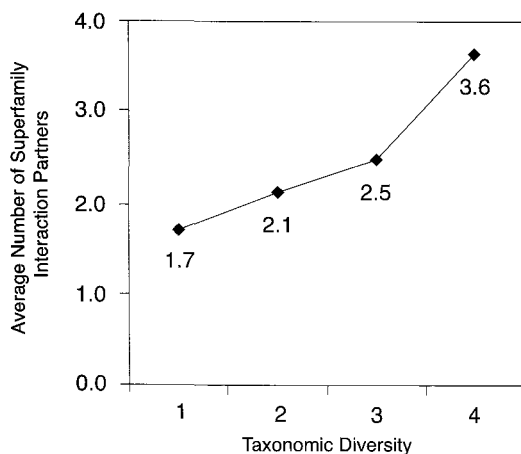


Fig. 6. The average number of superfamily interaction partners is plotted for superfamilies in different 'taxonomic diversity' groups. The taxonomic diversity of a superfamily is simply the number of superkingdoms in which that superfamily has been identified.

with the lethality of knock out mutation (Jeong *et al.*, 2001). Thus PSI-MAP reflects the functional importance of superfamilies by showing number of interactions they have.

Protein phylogeny, age and interaction

The occurrence of specific superfamilies within different branches of the tree of life gives us information on superfamily evolution and spread. By inference, this information also gives us the relative age of those superfamilies (Ponting *et al.*, 1999; Anantharaman *et al.*, 2001; and Snel *et al.*, 2002 for recent examples of this general approach, also suggested by authors, Park and Bolser, 2001). We used the NCBI taxonomic database and the Swissprot (Bairoch A. & Apweiler R., 2000) taxonomic annotation to collect this information (see methods). Simply counting the occurrence of a superfamily at the highest level of the taxonomic tree (the superkingdom) allowed us to infer an evolutionary age (Table 2). This measure of 'taxonomic diversity' gives each superfamily an approximate relative rank age.

In general, superfamilies with low taxonomic diversity are less likely to have interactions than those found throughout the tree of life. In combination with this observation, the average number of superfamily interaction partners also increases with diversity (Fig. 6).

As the super-kingdom level is very high, it is necessary to verify this trend at higher resolution in the future work.

Similar age-interaction correlations have been reported for metabolic networks (Jeong *et al.*, 2000; Wagner & Fell,

Table 2. Shows the number of superfamilies allocated to each 'taxonomic diversity' group. The number of superfamilies in each group with at least one observed interaction is also given, along with the percentage of interacting superfamilies for the group.

Super-kingdoms	Number of superfamilies	Number with interactions	Percent (%)
1	363	163	45
2	207	117	57
3	300	221	74
4	57	53	93

2001). Jeong *et al.* analyse the metabolic networks of 43 organisms, representing eubacteria, eukaryota and archaea. In this analysis 4% of all substrates are found to be present in all 43 organisms. These ubiquitous metabolites also represent the most highly connected substrates in the individual metabolic networks. Similarly, the "less connected substrates ... serve as educts or products of species-specific enzymatic activities" (Jeong *et al.*, 2000). Wagner and Fell concentrate on the analysis of the metabolic network of *Escherichia coli*. They ranked metabolites according to local and global network connectivity. The authors state that "many of the most highly connected metabolites ... have a proposed early evolutionary origin" (Wagner & Fell, 2001).

Network evolution hypothesis

Recently, it was discovered that many 'non-centralised' networks, including protein interaction networks, have a statistically similar connection topology (Barabasi & Albert, 1999). In these networks low and intermediate numbers of connections are common, while highly connected nodes in the network are rare but statistically significant (Dorogovtsev & Mendes, 2001). Typically, the connection distribution is described by a power law and the network is said to be 'scale free' (Barabasi & Albert, 1999). Such networks also have the 'small world' property, whereby the network diameter is significantly smaller than a random network with the same number of nodes (Watts & Strogatz, 1998). Scale free networks are optimized for the small world property, as randomly removing nodes has a very small effect on the network diameter (Albert *et al.*, 2000). Such networks are said to be robust, as they can tolerate random deletions without changing overall connectivity (Albert *et al.*, 2000). The structural interaction network produced by PSI-MAP has such a scale free topology (Fig. 5) (Park *et al.*, 2001).

Using models of genetic 'network evolution' it has been shown that as the allosteric interactions between alleles increases, the rate of finding fitter 'genotypes'

decreases (Kauffman, 1993). In these models 'interactions' limit the ability of a network to evolve. This rate of network evolution suggests why early, functionally important and interconnected life processes are slow to change at evolutionary time scales. Core metabolic pathways can display permutations (for example loss of specific pathways (Huynen *et al.*, 1999)), however, the overall network does not change radically. Ancient, fundamental biochemical pathways such as the TCA cycle and glycolysis are fixed in their basic architecture early in evolution.

The conclusion that the rate of network evolution combined with the scale free network topology can account for the structurefamily age-interaction correlation is somewhat at odds with Wagner, 2001. Here gene duplication events are identified in the yeast genome, and they are used to measure of the rate of interaction formation and loss between paralogous genes. A high rate of 'interaction flux' is estimated, suggesting 50% of all the network interactions change every 300 million years. This estimate is based on the assumption that the rate of interaction flux after gene duplication is indicative of the overall rate. However, there is evidence to suggest that this rate could be specifically accelerated after duplication (Long & Langley, 1993; Benton *et al.*, 1997; Cirera & Aguade, 1998; Tsaur *et al.*, 1998), leading to an overestimate of total interaction flux.

The results and conclusions in this paper corroborate the results of Fraser *et al.*, 2002. Here, exactly the same principals of network evolution are used to explain an observed negative correlation between connectivity and evolutionary rate. The principals are interpreted in the biological context of reciprocal mutations and the coevolution of proteins in the interaction network.

Network topology

Although the scale free topology is said to be robust to the effects of random deletion, conversely, the non random removal of the most highly connected nodes in the network rapidly fragments the network (Albert *et al.*, 2000). Why then do such vulnerable network topologies exist in nature? Two models of network growth have been used to account for the prevalence of scale free networks. The first, called preferential attachment (Barabasi & Albert, 1999), models an attachment bias towards already connected nodes. The second model assumes a constrained network diameter (Puniyani & Lukose, 2001) and the random attachment of nodes. The diameter of the metabolic networks from a total of 43 prokaryotes, eukaryotes and archae are all very similar (around 4), despite the varying number of metabolites and complexity of these organisms (Jeong *et al.*, 2000). This observation is not predicted by

preferential attachment, but is implicit in the second model. It implies that the metabolic network diameter is a limiting factor in evolution. The same constraint has been suggested of protein interaction networks (Jeong *et al.*, 2001). Both models of network growth result in the scale free topology, where old nodes accumulate more links over time (without the specific treatment of age as in Dorogovtsev & Mendes, 2000).

Protein topology

The secondary structure of the HIINFOLD group was mostly alpha and beta (81%, alpha&beta and alpha+beta) with only one all-alpha structurefamily (ARM repeat a.118.1) and two all-beta superfamilies (Immunoglobulin b.1.1 and Trypsin-like serine proteases b.47.1). The 160 LOINFOLD structurefamilies show a more even distribution among the classes (Fig. 7).

Methods

Taxonomic diversity

Superkingdoms were assigned to SCOP domains via the species identification codes of SWISS-PROT Protein Sequence Database (Bairoch & Apweiler, 2000; Release 39.0, May 2000). Each SCOP domain sequence in PDB90D (non-redundant SCOP domain sequences at 90% mutual sequence identity) was searched against a non redundant SWISS-PROT (90% mutual sequence identity) database. The search was done using the PDB-ISL protocol (Park *et al.*, 1997; Teichmann *et al.*, 2000) for reliable structural assignments, implemented to integrate with a relational database for easy analysis. Briefly, the PSI-BLAST search algorithm (Altschul *et al.*, 1997) is used with e-value 0.0005 and up to 10 iterations. These values

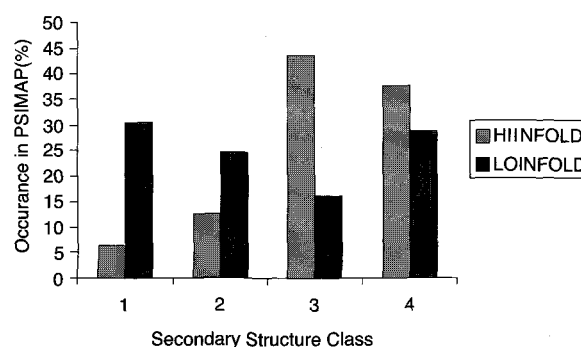


Fig. 7. Percentage of the different structural classes in the 160 most highly interacting structurefamilies and the 160 least interacting structurefamilies. Constructed using the SCOP structural classification database.

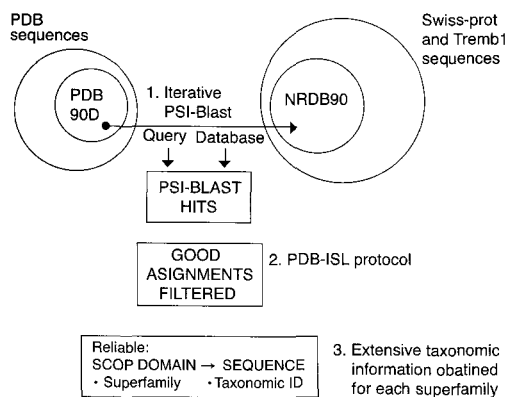


Fig. 8. Flow chart of the taxonomic assignment of SCOP superfamilies. Not to scale. For details see accompanying text.

have been previously verified and are known to give less than 1% false positives (Park *et al.*, 1998). Each statistically significant match (e-value below 0.0005) is checked for overlap with matches from other PDB90D domain sequence with different structurefamily classifications, and these 'classification collisions' are removed. Further filtering reduces the error rate even further (Park *et al.*, 1997; Teichmann *et al.*, 2000). The resulting structural assignments between representative SCOP domains and proteins in SWISS-PROT give the structurefamily - superkingdom correspondence. These data were used to reliably derive the taxonomic diversity for each structurefamily (Fig. 6).

Functional assignment

Using the same method as above each superkingdom was assigned to a list of SWISS-PROT accession numbers. These numbers give links to entries in the enzyme database via the ENZYME number. A very low e-value threshold was used to select the most reliable enzyme classifications for each structurefamily.

Summary

The latest functional analysis of high and low interaction groups showed most highly interacting structurefamilies in PSI-MAP represent functionally important enzymatic protein domains with homologues in an average of 3.6 superkingdoms. The least interacting structurefamilies in represent fewer enzymatic protein domains, occurring in an average of 2 superkingdoms.

In all, the correlation between the relative age and the interactability of protein structurefamilies is consistent with a hypothesis of network growth that proceeds via random

'add-on' interactions with constraints (after Puniyani & Lukose, 2001). New, specialised functions are attached to the existing network of protein interactions, and structurefamilies gradually acquire an increasing number of interaction partners throughout the course of evolution.

We attribute the extremely conserved nature of core biochemical pathways to a mechanism of 'network evolution' where relatively ancient components are under strong optimization constraints through multiple interactions (Kauffman, 1993). Thus, in general, protein structurefamilies in central positions in the structural interaction network are more ancient than peripheral structurefamilies.

Acknowledgments

This work was supported by MRC (MRC-DUNN) UK. This work was also supported by the Ministry of Information and Communication of South Korea under grant number IMT2000-C3-4. We thank Dr. Kyung#Sook Han on partial collaboration.

















References

- Alexandrov, N. N. & Go, N. (1995). Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci.* 3, 866-875.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., 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
- Anantharaman, V., Koonin, E. V. & Aravind, L. (2001). Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.*, 307, 1271-1292.
- Apic, G., Gough, J. & Teichmann, S. A. (2001a). An Insight into Domain Combinations. *Bioinformatics.* 17, 83S-89S.
- Apic, G., Gough, J. & Teichmann, S. A. (2001b). Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes. *J. Mol. Biol.* 310, 311-325.
- Bairoch A. & Apweiler R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45-48.
- Bairoch A. (2000). The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304-305.
- Barabasi, A. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286, 509-512.
- Benner, S. A., Ellington, A. D. & Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. *Proc. Natl Acad. Sci. USA*, 86, 7054-7058.
- Bennett, M. J., Choe, S. & Eisenberg, D. (1994). Domain swapping: Entangling alliances between proteins. *Proc. Natl. Acad. Sci. U.S.A.* 91, 3127-3131.
- Benton, BK., Tinkelenberg, Gonzalez, I. and Cross, FR. (1997). Cla4p, a *Saccharomyces-Cerevisiae* Cdc42p-activated kinase involved in cytokinesis is activated at mitosis. *Mol. Cell. Biol.*

- 17, 5067-5076.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucl. Acids Res.*, 28, 235-241.
- Cirera, S., and Aguade M. (1998) Molecular evolution of a duplication: the sex-peptide (Acp70A) gene region of *Drosophila subobscura* and *Drosophila madeirensis*. *Mol. Biol. Evol.* 15, 988-996.
- Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357, 543-544.
- Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a finger-print of proteins that physically interact. *Trends Biochem. Sci.* 23, 324-328.
- Doolittle, R. F. (1999). Do you dig my groove? *Nat Genet.* 23, 6-8.
- Dorogovtsev, S. N. & Mendes, J. F. F. (2000). Evolution of reference networks with ageing. <http://xxx.lanl.gov/abs/cond-mat/0001419>.
- Dorogovtsev, S. N. & Mendes, J. F. F. (2001). Evolution of networks. <http://xxx.lanl.gov/abs/cond-mat/0106144>.
- Enright, A. J. & Ouzounis, C. A. (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusion. *Genome Biology*, 2(9), research0034.1-7.
- Enright, A. J., Iliopoulos, I., Kyripides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402, 86-90.
- Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002). Evolutionary Rate in the Protein Interaction Network. *Science*, 296, 750-752.
- Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genomes sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, 313, 903-919.
- Henrick, K & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* 23(9), 358-361.
- Huynen, M. A., Dandekar, T. & Bork, P. (1999). Variation and evolution of the citric acid cycle: a genomic perspective. *Trends Microbiol.* 7, 281-291.
- Huynen, M., Snel, B., Lathe, W. & Bork P. (2000). Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Res.* 10, 1204-1210.
- Jeong, H., Mason, S., Barabasi, A. & Oltvai, Z. (2001). Lethality and centrality in protein networks. *Nature*, 411, 41-42.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407, 651-654.
- Jones, S., Marin, A. & Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* 13, 77-82.
- Ju, BH, Park, B, Park, JH, and Han, K, (2003) Visualization and analysis of protein interactions. *Bioinformatics* 2003, 19, 317-318
- Kauffman, SA. (1993). *The Origins of Order*, New York, Oxford, Oxford University Press, pp. 39-67.
- Long, M and Langley, CH. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science*, 260, 91-95.
- Marcotte, E. M., Pellegrini, M., Ng, H., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, 285, 751-753.
- Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng* 3, 77-83.
- Morowitz, H. J. (1992). *Beginnings of cellular life: metabolism recapitulates biogenesis*, New Haven, Yale University Press.
- Morowitz, H. J. (1999). A theory of biochemical organization, metabolic pathways and evolution. *Complexity*, 4, 39.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
- Orengo, C. A., Jones, D. T., & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, 372, 631-634.
- Overbeek, R., Fonstien, M., D' Souza, M., Pusch, G. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2896- 2901.
- Park, J., Lappe, M. & Teichmann, S. A. (2001). Mapping Protein Family Interactions: Intramolecular and Intermolecular Protein Family Interaction Repertoires in the PDB and Yeast. *J. Mol. Biol.* 307, 929-938.
- Park, J and Bolser, D, (2001). Conservation of protein interaction network in evolution. *Genome Informatics*, 12, 135-140.
- Park, J., Teichmann, S. A., Hubbard, T., and Chothia, C., (1997) Intermediate sequences increase the detection of distant sequence homologies. *J. Mol. Biol.* 273, 349-354.
- Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T and Chothia, C (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, 284, 1201-1210.
- Ponting, C. P., Aravind, L., Schultz, J., Bork, P. & Koonin, E. V. (1999). Eukaryotic signaling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* 289, 729-745.
- Puniyani, A. R. & Lukose, R. M. (2001). Growing random networks under constraints. <http://xxx.lanl.gov/abs/cond-mat/0107391>.
- Shindyalov I. N. & Bourne P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9) 739-747.
- Snel B., Bork P. & Huynen, M. A. Genomes in flux: The evolution of archaeal and proteobacterial gene content. (2002). *Genome Res.* 12, 17-25.
- Teichmann, SA., Chothia, C., Church, GM., and Park, J. (2000) Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics*, 16, 117-124.
- Tsai, C., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996). A Dataset of Protein-Protein Interfaces Generated with a Sequence-order-independent Comparison Technique. *J. Mol. Biol.*, 260, 604-620.
- Tsaur SC, Ting CT, and Wu CI. (1998) Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of *Drosophila*: II. Divergence versus polymorphism. *Mol. Biol. Evol.* 15, 1040-1046.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T,

- Vijayadamodar G, Yang M, Johnston M, Fields S, and Rothberg JM. (2000), A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403, 623-627.
- Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18, 1283-1292.
- Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. R. Soc. Lond.* 268, 1803-1810.
- Wang, Z. X. (1996). How many fold types of protein are there in nature? *Proteins*. 26. 186-191.
- Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440-442.
- Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., & Rapp, B. A. (2000). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* 28, 10-14. (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomy_home.html/index.cgi).
- Zhang, C. T. (1997). Relations of the numbers of protein sequences, families and folds. *Protein Engineering*, 10, 757-761.

Appendix A. HIINFOLD: The sixteen most Interactive superfamilies.

SCOP ID 1.59	Name	Number of Interacting Superfamilies	Taxonomic Diversity	Super-kingdoms	Distribution
c.37.1	P-loop containing nucleotide triphosphate hydrolases	25	4	ABEV	
b.1.1	Immunoglobulin	23	4	ABEV	
c.1.8	(Trans) glycosidases	16	4	ABEV	
c.3.1	FAD/NAD(P)-binding domain	11	3	ABE	
b.47.1	Trypsin-like serine proteases	9	4	ABEV	
c.1.4	FMN-linked oxidoreductases	9	3	ABE	
c.2.1	NAD(P)-binding Rossmann-fold domains	9	4	ABEV	
d.142.1	Glutathione synthetase ATP-binding domain-like	9	3	ABE	
d.3.1	Cysteine proteinases	8	4	ABEV	
d.15.1	Ubiquitin-like	8	2	EV	
d.144.1	Protein kinase-like (PK-like)	8	4	ABEV	
a.118.1	ARM repeat	7	3	BEV	
c.23.16	Class I glutamine amidotransferase-like	7	3	ABE	
c.56.5	Zn-dependent exopeptidases	7	3	ABE	
d.58.1	4Fe-4S ferredoxins	7	3	ABE	
d.92.1	Metalloproteases (zincins), catalytic domain	7	4	ABEV	
		AVG	TOTAL		
		3.4	80483		

Appendix B. LOINFOLD: The 160 least interactive superfamilies

SCOP ID	Name	Taxonomic Diversity	Superkingdoms
1.59			
a.2.7	A class II aminoacyl-tRNA synthetase N-domain	3	ABE
a.2.1	Epsilon subunit of F1F0-ATP synthase C-terminal domain	2	BE
a.2.11	Fe, Mn superoxide dismutase (SOD), N-terminal domain	3	ABE
a.4.2	Methylated DNA-protein cysteine methyltransferase, C-terminal domain	3	ABE
a.4.6	C-terminal, effector domain of the bipartite response regulators	2	BE
a.4.11	RNA polymerase subunit RPB10	2	AE
a.5.3	N-terminal domain of phosphatidylinositol transfer protein sec14p	1	E
a.6.1	Putative DNA-binding domain	3	ABE
a.7.3	Succinate dehydrogenase/fumarate reductase C-terminal domain	3	ABE
a.8.1	Bacterial immunoglobulin/albumin-binding domains	1	B
a.11.2	Second domain of FERM	1	E
a.15.1	TAF(II)230 TBP-binding fragment	1	E
a.23.2	Diol dehydratase, gamma subunit	1	B
a.23.3	Methane monooxygenase hydrolase, gamma subunit	1	B
a.24.11	Bacterial GAP domain	1	B
a.24.13	Domain of the SRP/SRP receptor G-proteins	3	ABE
a.29.5	alpha-ketoacid dehydrogenase kinase, N-terminal domain	1	E
a.41.1	Domain of poly(ADP-ribose) polymerase	1	E
a.44.1	Disulphide-bond formation facilitator (DSBA), insertion domain	1	B
a.45.1	Glutathione S-transferases, C-terminal domain	2	BE
a.47.1	STAT	1	E
a.48.2	Transferrin receptor ectodomain, C-terminal domain	1	E
a.51.1	Cytochrome c oxidase subunit h	1	E
a.6.7	5' to 3' exonuclease, C-terminal subdomain	4	ABEV
a.6.8	HRDC-like	2	AE
a.6.1	Enzyme I of the PEP:sugar phosphotransferase system HPr-binding (sub)domain	1	B
a.69.1	C-terminal domain of alpha and beta subunits of F1 ATP synthase	3	ABE
a.85.1	Hemocyanin, N-terminal domain	1	E
a.86.1	Di-copper centre-containing domain	2	BE
a.87.1	DBL homology domain	1	E
a.88.1	LigA subunit of an aromatic-ring-opening dioxygenase LigAB	1	B
a.96.1	DNA-glycosylase	3	ABE
a.98.1	R1 subunit of ribonucleotide reductase, N-terminal domain	3	BEV
a.99.1	FAD-binding (C-terminal) domain of DNA photolyase	4	ABEV
a.12.3	Chondroitin AC/alginate lyase	2	BV
a.114.1	Interferon-induced guanylate-binding protein 1 (GBP1), C-terminal domain	1	E
a.116.1	GTPase activation domain, GAP	1	E
a.117.1	Ras GEF	1	E
a.118.2	Ankyrin repeat	4	ABEV
a.118.5	Bacterial muramidases	1	B
a.118.6	Protein prenyltransferase	2	AE
a.118.7	14-3-3 protein	1	E
a.119.1	Lipoxygenase	2	BE
a.124.1	Phospholipase C/P1 nuclease	3	ABE
a.137.2	Quinoprotein alcohol dehydrogenase	1	B
a.137.3	Transducin (heterotrimeric G protein), gamma chain	1	E
a.137.4	Fe-only hydrogenase smaller subunit	2	BE
a.137.7	Proteinase A inhibitor IA3	0	NULL
a.137.8	Epsilon subunit of mitochondrial F1F0-ATP synthase	1	E
b.1.5	Transglutaminase, two C-terminal domains	1	E
b.1.1	Clathrin adaptor appendage domain	1	E
b.1.12	Purple acid phosphatase, N-terminal domain	1	E
b.2.1	Diphtheria toxin, C-terminal domain	1	V
b.3.1	Starch-binding domain	3	ABE
b.3.2	Carboxypeptidase D, a regulatory domain	1	E
b.3.3	VHL	1	E
b.5.1	alpha-Amylase inhibitor tendamistat	1	B
b.14.1	Calpain large subunit, middle domain (domain III)	1	E
b.24.1	Hyaluronate lyase-like, C-terminal domain	1	B
b.3.1	beta-Galactosidase, domain 5	2	BE
b.3.4	Lactobacillus maltose phosphorylase, N-terminal domain	2	AB

continued

SCOP ID 1.59	Name	Taxonomic Diversity	Super- kingdoms
b.34.5	Translation proteins SH3-like domain	3	ABE
b.34.8	Fumarylacetoacetate hydrolase, FAH, N-terminal domain	2	BE
b.4.3	TIMP-like	1	E
b.42.1	Cytokine	2	EV
b.43.2	L-fucose isomerase, C-terminal domain	1	B
b.48.1	mu transposase, C-terminal domain	2	BV
b.49.2	Alanine racemase-like, C-terminal domain	4	ABEV
b.51.1	ValRS/IleRS editing domain	3	ABE
b.53.1	Ribosomal protein L25-like	3	ABE
b.54.1	Core binding factor beta, CBF	1	E
b.58.1	PK beta-barrel domain-like	3	ABE
b.61.2	Metalloprotease inhibitor	1	B
b.69.5	RCC1/BLIP-II	2	BE
b.69.7	Prolyl oligopeptidase, N-terminal domain	3	ABE
b.71.1	alpha-Amylases, C-terminal beta-sheet domain	3	ABE
b.74.1	Carbonic anhydrase	3	BEV
b.77.2	delta-Endotoxin (insecticide), middle domain	1	B
b.79.1	Metalloprotease, C-terminal domain	1	B
b.8.3	Cell-division inhibitor MinC, C-terminal domain	1	B
b.8.4	Alpha subunit of glutamate synthase, C-terminal domain	3	ABE
b.85.3	Urease, beta-subunit	3	ABE
b.85.6	Molybdenum cofactor biosynthesis protein MoeA, C-terminal domain	1	B
b.86.1	Hedgehog/intein (Hint) domain	3	ABE
b.93.1	Epsilon subunit of F1F0-ATP synthase N-terminal domain	2	BE
b.98.1	Leukotriene A4 hydrolase N-terminal domain	3	ABE
b.11.1	Ribonuclease domain of colicin E3	1	B
b.13.1	Molybdenum cofactor biosynthesis protein MoeA, N-terminal and linker domains	3	ABE
c.1.6	PLP-binding barrel	4	ABEV
c.1.17	Quinolinic acid phosphoribosyltransferase, C-terminal domain	3	ABE
c.8.1	Phosphohistidine domain	3	ABE
c.8.2	Aconitase, C-terminal domain	3	ABE
c.8.3	Carbamoyl phosphate synthetase, small subunit N-terminal domain	3	ABE
c.8.4	Transferrin receptor ectodomain, apical domain	2	BE
c.9.1	Barstar (bamase inhibitor)	1	B
c.1.2	L domain-like	3	BEV
c.13.1	C-terminal domain of phosphatidylinositol transfer protein sec14p	1	E
c.23.6	Cobalamin (vitamin B12)-binding domain	3	ABE
c.23.11	Beta-D-glucan exohydrolase, C-terminal domain	3	ABE
c.23.12	Formate/glycerate dehydrogenase catalytic domain-like	4	ABEV
c.26.3	UDP-glucose dehydrogenase (UDPGDH), C-terminal (UDP-binding) domain	3	BEV
c.28.1	N-terminal domain of DNA photolyase	3	ABE
c.32.1	Tubulin, GTPase domain	3	ABE
c.5.1	Leucine aminopeptidase, N-terminal domain	1	E
c.51.1	Anticodon-binding domain of Class II aaRS	3	ABE
c.51.3	Diol dehydratase, beta subunit	1	B
c.53.1	Resolvase-like	4	ABEV
c.55.2	Creatinase/prolidase N-terminal domain	2	BE
c.55.6	DNA repair protein MutS, domain II	2	AB
c.55.7	Methylated DNA-protein cysteine methyltransferase domain	3	ABE
c.83.1	Aconitase, first 3 domains	3	ABE
c.91.1	PEP carboxykinase-like	3	ABE
c.12.1	Cell-division inhibitor MinC, N-terminal domain	1	B
c.19.1	PEP carboxykinase N-terminal domain	3	ABE
d.15.3	MoaD/ThiS	2	AB
d.15.6	Superantigen toxins, C-terminal domain	1	B
d.15.7	Immunoglobulin-binding domains	1	B
d.15.9	Glutamine synthetase, N-terminal domain	2	AB
d.17.1	Cystatin/monellin	1	E
d.17.2	Copper amine oxidase, domains 1 and 2	3	ABE
d.2.1	Ubiquitin conjugating enzyme	2	EV
d.26.1	FKBP-like	3	ABE
d.26.2	Colicin E3 immunity protein	1	B

continued

SCOP ID	Name	Taxonomic Diversity	Super-kingdoms
1.59			
d.26.3	Chitinase insertion domain	4	ABEV
d.41.2	Quinolinic acid phosphoribosyltransferase, N-terminal domain	3	ABE
d.41.3	Pyrimidine nucleoside phosphorylase C-terminal domain	2	AB
d.41.5	Molybdopterin synthase subunit MoaE	3	ABE
d.5.1	dsRNA-binding domain-like	4	ABEV
d.5.2	Porphobilinogen deaminase (hydroxymethylbilane synthase), C-terminal domain	3	ABE
d.56.1	GroEL-like chaperone, intermediate domain	3	ABE
d.58.12	eEF-1beta-like	2	AE
d.58.14	Ribosomal protein S6	1	B
d.58.2	NAD-binding domain of HMG-CoA reductase	3	ABE
d.58.22	TRADD, N-terminal domain	1	E
d.58.32	FAD-linked oxidases, C-terminal domain	2	BE
d.6.1	Probable bacterial effector-binding domain	1	B
d.62.1	Pepsin inhibitor-3	1	E
d.67.2	Arginyl-tRNA synthetase (ArgRS), N-terminal 'additional' domain	3	ABE
d.68.1	Translation initiation factor IF3, C-terminal domain	2	BE
d.69.1	C-terminal domain of TolA	1	B
d.79.2	Tubulin, C-terminal domain	3	ABE
d.82.1	Copper amine oxidase, domain N	1	B
d.94.1	HPr-like	1	B
d.15.1	Clathrin adaptor appendage domain	1	E
d.125.1	Ornithine decarboxylase C-terminal domain	1	B
d.138.1	B3/B4 domain of PheRS, PheT	2	BE
d.139.1	Aminoimidazole ribonucleotide synthetase (PurM) C-terminal domain	3	ABE
d.142.2	DNA ligase/mRNA capping enzyme, catalytic domain	3	BEV
d.146.1	Uridine diphospho-N-Acetylenolpyruvylglucosamine reductase, MurB, C-terminal domain	1	B
d.149.1	Nitrile hydratase alpha chain	1	B
d.151.1	DNase I-like	3	ABE
d.161.1	ADC synthase	3	ABE
d.165.1	Ribosome inactivating proteins (RIP)	3	BEV
d.168.1	Succinate dehydrogenase/fumarate reductase catalytic domain	3	ABE
d.172.1	gp120 core	1	V
d.176.1	Sulfite oxidase, middle catalytic domain	3	ABE
d.178.1	Aromatic aminoacid monooxygenases, catalytic and oligomerization domains	2	BE
d.181.1	Insert subdomain of RNA polymerase alpha subunit	3	ABE
d.184.1	Non-globular alpha+beta subunits of globular proteins	2	BE
d.197.1	Protein-L-isoaspartyl O-methyltransferase, C-terminal domain	1	B
		AVG	TOTAL
		2	19211