Conservation of Protein Interaction Network in Evolution

Jong Park **Dan Bolser**
jong@mrc-dunn.cam.ac.uk dmb@mrc-dunn.cam.ac.uk
MRC-DUNN Human Nutrition Unit, Hills Road, Cambirdge, CB22XY, England, UK

Abstract

A functional analysis of protein fold interaction suggested that structural fold familieshave gradually acquired more diverse interacting partners while maintaining central biochemical interactions and functions. This means thatthe protein interaction network (map) maintains its robust architecture due to the functional constraints associated with the interactions.

Keywords: protein structural interaction map, PSIMAP, protein fold antiquity

1 Introduction

It is probable that there are no more than 10,000 distinct protein structural foldsin nature [3, 1, 14, 15]. At present, around 1,000 distinctprotein fold typesare classified as superfamilies with clear evolutionary boundaries (SCOP [8].) This is a sufficient number to make a global map or network of protein structural interactions with a phylogenetic (evolutionary) context. PSIMAP (Protein Structural Interaction MAP [10]) is such a map, covering the whole known protein fold space at the superfamily level. The analysis of this map allowed us to examine the relationships between protein folds (superfamilies) with regard to the evolutionary conservation of protein interactions.

Protein structures themselves contain limited amount of information about molecular evolution as it is not yet possible to reconstruct the whole 'phylogenic tree' of protein structures. However, the occurrence of specific folds and fold homologues within different species can provide us with additional information about the evolution and spread of fold types. We define 'basal folds' for certain superfamilies found throughout the whole phylogenetic tree. We observed that 'terminal' folds (those thatbranched off later in evolution) tend to occur with fewer interactions as shown by PSIMAP than the basal folds. Simply counting the number of superfamilies that belong to specific levels of the phylogenetic tree enables us to predict whether they are more ancient than others. It is an important question if the basal folds are relatively more ancient than the terminal folds. Here, we introduce protein interaction information to taxonomic diversity of folds to try to answer the question. Even a rapidly propagating protein fold must synchronize its evolution with its interacting partners in the network. Based on this phenomenon of 'network evolution', we attempted to examine 1) if the degree of interaction between folds is related to the functional diversity of the interacting partners, and 2) if the functions of proteins with different numbers of interaction partners (interactability) correspond to the importance and antiquity of the fold.

2 Protein Structure Interaction Map

The interaction map used, PSIMAP, is a compact, low-resolution map of protein folds, which has fourkey aspects: (1) it is structural (derived from PDB), (2) it reflects the functional importance of proteins (especially for the basal nodes in the map), (3) its interactions areevolutionarily meaningful

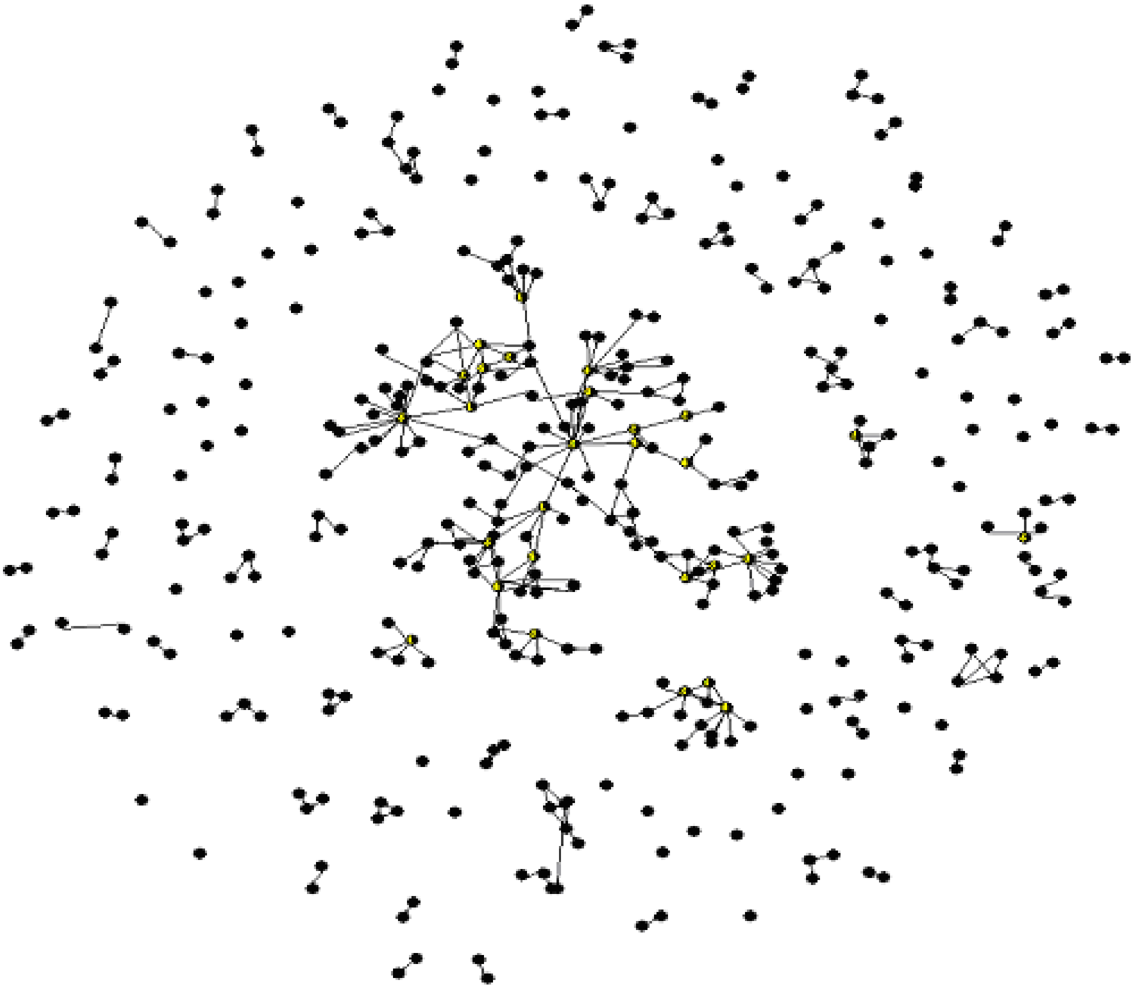

Figure 1: The Protein Structural Interaction MAP (PSIMAP version 1.53) showing distinctly different types of interactability. The central network is an example of highly connected interaction.

(derived from the homology-basedsuperfamily level SCOP), and (4) it is global, i.e., containing all the known protein folds. Figure 1 shows a representation of PSIMAP drawn by a computer graph layout algorithm. Note that there are both highly and sparsely connected superfamilies. In this paper the diversity of interaction was compared to both function and taxonomic diversity.

The maincriterion for the SCOP classification is structural similarity based on secondary structure content and fold topology. Each dot in Figure 1 represents an evolutionarily distinct superfamily, classified by experts after comparing all the 3D structures from PDB. PSIMAP's interaction criterion is also strictly structural, denoting distinct pairs of protein domains as interacting if there is a set of contacts between them inside a PDB entry by an empirical rule called 5-5 rule (5 contacts within 5 angstrom distance rule between definable domains of SCOP for all the PDB structures. Other rules ranging from 3 to 10 contacts with different distance resulted in similar results). As shown in Figure 1, superfamilies have various degrees of interactability withindistinctly shaped sub-networks.

To assess the functional and evolutionary differences between the most interactive and the least interactive folds, two extreme superfamily groups were chosen: high interaction folds (HIINFOLD, see table 1) and low interaction folds (LOINFOLD, accessible at: www.biointeraction.net/ProteinAge/LOINFOLD and www.bio.cc/ProteinAge/LOINFOLD). The elevenHIINFOLD superfamilies (with at least nineother interacting partners) contained domain structures which were often functionally critical and found widely in central biochemical pathways. They were also found to occur widely in nature (in at

Table 1: The most highly Interactive SCOP superfamilies.

Number of Interacting Superfamilies	Taxonomic Diversity	SCOP Superfamily (v. 1.53)
32	4	2.1.1 (IG domain)
28	4	3.32.1 (P-loop containing nucleotide triphosphate hydrolase and kinases)
16	4	2.44.1 (Trypsin-like serine proteinase, thromibin, viral proteases)
13	4	3.1.8 (Glycosyltransferases, TIM barrel, alpha amylases, type 2 chitinase)
12	4	3.3.1 (FAD/NAD binding dom, C-term adrenodoxin reductase-like)
12	3	1.41.1 (EF hand, calmodulin like)
11	3	4.14.7 (2Fe-2S ferredoxin)
10	4	3.2.1 (NAD binding Rossman, alcohol/glucose dehydrogenase, C-term)
9	4	4.130.1 (MAP kinases, serine/threonine kinase, tryosine kinases)
9	3	4.128.1 (Glutathine synthetase ATP-binding domain-like)
9	3	4.92.1 (tRNA synthetase, biotin synthetases)

least three out of 4 superkingdoms). Each HIINFOLD superfamily showed extensive structural diversity within its member domains. By contrast the 81 LOINFOLD superfamilies (each with only one interacting partner) contained functionally specific and structurally distinct, non-centrally interacting protein folds, which are unique to a single kingdom. The secondary structure of the HIINFOLD group was mostly alpha and beta, with one all-alpha superfamily (EF hand, calmodulin like domain) and two all-beta superfamilies (IG: Immunoglobin and trypsin-like serine protease, SCOP classification 2.1.1. and 2.44.1 respectively). The 81 LOINFOLD superfamilies showed a different trend in secondary structure distribution, having 38 all-alpha protein superfamilies (47%), fifteen all-beta (18%), four alpha &beta (5%) and 24 alpha+beta class superfamilies (30%).

The following section details the functions of the two groups that accounts for the differences in interactability.

3 HIINFOLD Group

Most of the domains in the HIINFOLD group are from functionally important enzymes, with only two main exceptions: 1) the IG domain, which often functions as a linker which interacts with many different fold types and 2) the EF hand all-alpha protein (1.41.1). The EF hand is regarded as a structural motif (with an average size of around 40 amino acids) involved in calcium binding and the diverse regulatory functions associated with Ca. Of the enzymatic superfamilies, 2.44.1 is a serine protease family found ubiquitously in viruses, archae, bacteria and eukaryotes [12]. They include a wide range of peptidase activity, including exopeptidase, endopeptidase, oligopeptidase and omega-peptidase activity. The TIM barrel fold (3.1.8) includes hydrolases and is widely found in all superkingdoms. Due to their distinct architecture, it is possible that the superfamily consist of some non-homologous members (convergent evolution, [4, 13]. Pyridine nucleotide-disulphide oxidoreductase (3.3.1) includes both class I and II oxidoreductases and also NADH oxidases and peroxidases. Ferredoxins (4.14.7) are iron-sulfur proteins which transfer electrons in a wide variety of metabolic reactions, indicating a very early evolutionary origin. The short-chain dehydrogenases/reductases (3.2.1) superfamily is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. They typically exhibit residue identities only at the 15-30% level, indicating early duplicatory origins and extensive divergence [5]. The 4.130.1 superfamily encompasses enzymes that belong to a

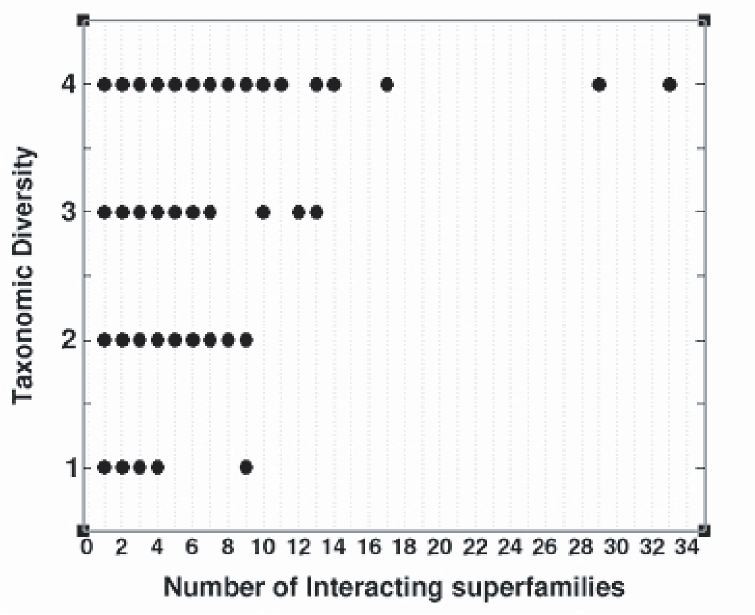

Figure 2: The number of interaction partners for each superfamily compared to its taxonomic diversity.

very extensive family of proteins [6]. They share a conserved catalytic core common with both serine/threonine and tyrosine protein kinases. 4.128.1 represents enzymes which catalyse the reversible conversion of ATP to AMP, pyrophosphate and phosphoenolpyruvate (PEP), a critical function in cells. 4.92.1 is the asparagine and biotin synthetases superfamily.

4 LOINFOLD Group

A careful examination of all the 81 least interactive superfamilies showed that the functions of 47 of them (58 %) did not require extensive interactions. They belong to inhibitors, various factors, toxins, ligand binding domains (such as DNA binding domains), effectors, co-factors, repressors, viral capsid domains, structurally repeating domains and surface domains. Also, many of the LOINFOLD superfamilies mediate organism specific functions such as light harvesting and plant toxins.

This comparison highlights a clear distinction between the HIIFOLD and LOIFOLD groups in terms of their functions that is reflected in their interactability.

5 Protein Interactability and Taxonomic Diversity

Figure 2 shows the number of interacting partners foreach superfamily againstthe number of different superkingdoms to which they belong. On average, a superfamily that occurred in one kingdom only (e.g., only in prokaryotes) has 1.5 interaction partners, while a superfamily which occurred in all foursuperkingdoms has 3.3 interaction partners. It showed that the taxonomic diversity is proportional to the degree of interaction among the superfamilies.

The 4 superkingdoms used are: archae, eukaryotes, prokaryotes and viruses. The superkingdoms were assigned from the species identification codes of SWISS-PROT Protein Sequence Database (Release 39.0, May 2000 [2]). Each SCOP domain was searched against a large non-redundant protein database, NRDB90 [7]. NRDB90 was used as a source of intermediate sequences for the PSI-BLAST search algorithm [11]. This additional NRDB90 search step was included toproduce the most extensive taxonomic coverage possible. The matched sequences for all the entries of PDB90D SCOP database with significant statistical scores from the NRDB90 were then assigned Swissprot species codes. There were 503 unique superfamilies that have at least one interacting partner. Among them, 130 superfamilies (26%) belonged to one kingdom only, 103 (20%) belonged to 2 superkingdoms, 191 (38%) belonged to 3 superkingdoms and 79 (16 %) belonged to all four superkingdoms. Out of 130 superfamilies in the one kingdom category, only 12 of them (8%) had more than three interacting partners. These exceptions include a eukaryotic superfamily (2.6.1) which can be seen in Figure 2 as an outlier. It is an all beta sheet domain that interacts with 8 different superfamilies. The high interactability and narrow taxonomic diversity of this superfamily can be interpreted in terms of the function of its member domains. They are the phospholipase C isozyme C-terminal domains, the synaptotagmin-like (S variant) domain and the C2 domain (from protein kinase C). The C2 domain (CaLB) is involved in cellular signaling during inflammation, which is specific to multicellular eukaryotes.

6 Conclusion

In summary, protein superfamilies which have a high degree of structural interaction with other superfamilies have central and important functions in cells and are found in all the major superkingdoms of life. This indicates thatthey have a long history of evolution in the core sectors of biochemical pathways. On the other hand, protein superfamilies with a low degree of structural interaction cover functions that are not central to the biochemical pathways, havingspecific, specialised functions. Furthermore, these superfamilies were found less extensively throughout the taxonomic assignments, suggesting that they have occurred relatively recently in evolution in contrast to the highly interactive superfamilies. Throughout all the superkingdoms drastically different organisms maintaina very similar basic backbone ofbiochemical pathways, such as glycolysisand the TCA cycle, with key catalytic enzymes. Our analysis suggests that this is because the central biochemical interaction networks are highly connected and robustto changes. The evolution of life is through 'add-on' interactions of other or newerfolds onto existing ones. In other words, the highly interactive central protein folds will maintain their central positions in biochemistry and propagate continuously throughout time within the biological interaction network.

Acknowledgements

This work was supported by MRC (MRC-DUNN) UK. We thank Dr. John Walker for much help in establishing bioinformatics research group in MRC-DUNN. We thank Maryana Huston for reviewing and correcting the paper. J.P. thanks honest and unselfish scientists.

References

[1] Alexandrov, N.N. and Go, N., Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins, *Protein Sci.*, 3:866–875, 1995.

[2] Bairoch, A. and Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, 28:45–48, 2000.

[3] Chothia, C., One thousand families for the molecular biologist, *Nature*, 357:543–544, 1992.

[4] Farber, G.K., An alpha/beta-barrel full of evolutionary trouble, *Curr. Opin. Struct. Biol.*, 3:409–412, 1993.

[5] Jornvall, H., Persson, B., Krook, M., Atrian, S., Gonzales-Duarte, R., Jeffery, J., and Ghosh, D., Short-chain ehydrogenases/reductases (SDR), *Biochemistry*, 34:6003–6013, 1995.

[6] Hanks, S.K. and Hunter, T., Protein kinases: the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification, *FASEB J.*, 9:576–596, 1995.

[7] Holm, L. and Sander, C., Removing near-neighbour redundancy from large protein sequence collections,*Bioinformatics*, 9:423–429, 1998.

[8] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536–540, 1995.

[9] Orengo, C.A., Jones, D.T., and Thornton, J.M., Protein superfamilies and domain superfolds, *Nature*, 372:631–634, 1994.

[10] Park, J., Lappe, M., and Teichmann, S.A., Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast, *J. Mol. Biol.*, 307:929–938, 2001.

[11] Park, J., Teichmann, S.A., Hubbard, T., and Chothia, C., Intermediate sequences increase the detection of homology between sequences, *J. Mol. Biol.*, 273:349–354, 1997.

[12] Rawlings, N.D. and Farrett, A.J., Families of serine peptidases, *Method Enzymol.*, 244:19–61, 1994.

[13] Reardon, D. and Farber,G.K., Protein motifs: the structure and evolution of alpha/beta barrel proteins, *FASEB J.*, 9:497–503, 1995.

[14] Wang, Z.X., How many fold types of protein are there in nature? *Proteins*, 26:186–191, 1996.

[15] Zhang, C.T., Relations of the numbers of protein sequences, families and folds, *Protein Engineering*, 10:757–761, 1997.