

Mapping Protein Family Interactions: Intramolecular and Intermolecular Protein Family Interaction Repertoires in the PDB and Yeast

Jong Park¹, Michael Lappe¹ and Sarah A. Teichmann^{2*}

¹*European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridgeshire, CB10 1SD UK*

²*Department of Biochemistry and Molecular Biology University College London Darwin Building, Gower Street London WC1E 6BT, UK*

In the postgenomic era, one of the most interesting and important challenges is to understand protein interactions on a large scale. The physical interactions between protein domains are fundamental to the workings of a cell: in multi-domain polypeptide chains, in multi-subunit proteins and in transient complexes between proteins that also exist independently. To study the large-scale patterns and evolution of interactions between protein domains, we view interactions between protein domains in terms of the interactions between structural families of evolutionarily related domains. This allows us to classify 8151 interactions between individual domains in the Protein Data Bank and the yeast *Saccharomyces cerevisiae* in terms of 664 types of interactions, between protein families. At least 51 interactions do not occur in the Protein Data Bank and can only be derived from the yeast data. The map of interactions between protein families has the form of a scale-free network, meaning that most protein families only interact with one or two other families, while a few families are extremely versatile in their interactions and are connected to many families. We observe that almost half of all known families engage in interactions with domains from their own family. We also see that the repertoires of interactions of domains within and between polypeptide chains overlap mostly for two specific types of protein families: enzymes and same-family interactions. This suggests that different types of protein interaction repertoires exist for structural, functional and regulatory reasons.

© 2001 Academic Press

Keywords: protein interactions; scale-free network; oligomers; protein complexes

*Corresponding author

Introduction

Protein domain interactions are essential to the functioning of individual cells and whole organisms by acting in several ways: domain-domain interactions in multi-domain polypeptide chains, inter-chain protein interactions in obligate complexes such as multimers and in transient complexes between proteins that can also exist independently. Therefore, it is not surprising that protein interactions have been extensively investi-

gated using a variety of methods. The physical and chemical properties of domain interfaces have been studied by computational analysis of the entries in the Protein Data Bank (PDB; Bernstein *et al.*, 1977) by Argos (1988), Chothia and co-workers (Janin *et al.*, 1988; LoConte *et al.*, 1999) and Thornton and co-workers (Jones & Thornton, 1997; Jones *et al.*, 2000). The interactions between individual proteins in cells have been studied using genetic and biochemical methods for many years, and for yeast, the results of the individual experiments have been collected in the MIPS database (Mewes *et al.*, 2000; <http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/index.html>). Recently, protein interactions in yeast have been studied on a large scale by Ito *et al.* (2000) and Uetz *et al.* (2000). There have been various computational methods for predicting protein interactions in whole genomes, for instance by Ouzounis and co-workers

Present address: J. Park, Bioinformatics/Proteomics, MRC-Dunn Human Nutrition Unit, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, UK.

Abbreviations used: PDB, Protein Data Bank; SCOP, Structural Classification of Proteins.

E-mail address of the corresponding author: sat@biochem.ucl.ac.uk

(Enright *et al.*, 1999; Tsoka & Ouzounis, 2000) and Eisenberg and co-workers (Marcotte *et al.*, 1999). The coverage and error rates of these methods vary for prediction of physical interactions. A method based on gene fusion, similar to that of Enright *et al.* (1999), has been compared to other prediction methods by Huynen *et al.* (2000).

While all the work mentioned above is concerned with interactions between individual proteins, we take an entirely new perspective and view protein interactions in terms of whole protein families that interact with each other. We use the protein families that are classified as superfamilies in the Structural Classification of Proteins (SCOP) database (Murzin *et al.*, 1995). This has the advantage that we can classify the 8151 interactions between individual domains in terms of about 664 types of interactions between pairs of protein families. This way, we learn about how the interactions between protein families are organised on a large scale, and obtain an overview of their evolution.

Results and Discussion

The protein family interaction map

The interactions between protein families are derived from classifying the interactions between domains of known three-dimensional structure in the Protein Data Bank, and those of domains assigned to yeast proteins by homology, into types

of interactions between pairs of protein families. We use pairs of yeast polypeptides that are known to interact from experimental data, and the protein domain families are derived from the Structural Classification of Proteins database (Murzin *et al.*, 1995), a hierarchical classification of all domains of known three-dimensional structure. Please refer to the Methods section below for the details of our procedures for determining interactions between domains and families. The resulting graph of family interactions is shown in Figure 1, and the detailed numbers of interactions in Table 1.

Figure 1 and Table 1 show that most families only engage in a few types of interactions: they only have one, two or three other families with which they interact. A few families are very versatile in their interactions and combine with many other families, so that they are hubs in the graph of family interactions. This is not because there are more instances of proteins in these families in the database: there is no correlation between the number of individual domain interactions in the PDB and yeast and the number of interactions the family as a whole engages in, as shown in the graph in Figure 2. The versatile families are special for functional reasons: the P-loop nucleotide triphosphate hydrolases because they are kinases and transferases as well as providing energy for reactions or motion, the immunoglobulins due to their role in the immune system and in cell-adhesion proteins, the protein kinases due to their diversity in substrates, and so forth.

Table 1. Family interactions within and between polypeptide chains in the PDB and yeast

No. of interaction partner families	No. of families within chains in the PDB (intramolecular)	No. of families between chains in the PDB (intermolecular)	No. of families within chains in yeast (intramolecular)	No. of families between chains in yeast (intermolecular)
1	253	350	123	41
2	52	41	25	6
3	13	12	5	2
4	4	4	2	
5	3	4	2	
6	6	1 Membrane all-alpha	1 NAD(P)-binding Rossmann fold domains	
7	1 NAD(P)-binding Rossmann fold domains	1 NAD(P)-binding Rossmann fold domains		
8				1 Protein kinases (PK), catalytic core
9				1 P-loop nucleotide triphosphate hydrolases
13	1 P-loop nucleotide triphosphate hydrolases	1 Trypsin-like serine proteases		
14	1 Immunoglobulins	1 P-loop nucleotide triphosphate hydrolases		1 ARM repeat
15			1 P-loop nucleotide triphosphate hydrolases	
18		1 Immunoglobulins		
Total number of non-redundant pairwise interactions:	278	405	136	55

In this Table, each family in the PDB and yeast is viewed separately, and the number of different interaction partner families is summed for all the domains in each family. For instance, if one family F has two domains, a and b, and each of these interacts with a domain from a different family, the number of interaction partner families for F will be two. All such families will be entered into the second row of Table 1. The SCOP names of the most versatile families are given. The number of non-redundant interactions in the bottom row of the Table is the total number of different pairs of families seen to interact.

It is worth noting that the shape of the graph of family interactions in Figure 1, and the form of the distribution of family interactions in Table 1, are those of a scale-free network, where the probability of a node being connected to k other nodes is given by $P(k) = Ak^{-\gamma}$. The fit to the power law with $A = 0.34$ and $\gamma = 1.6$ is shown in Figure 3, and has a coefficient of determination of 0.88. Distributions of this form, which is similar to Zipf's law, have been found to apply in a variety of other protein-related situations, such as the number of proteins with increasing numbers of transmembrane helices in genomes (Gerstein, 1998), the occurrence of oligonucleotide words (Konopka & Martindale, 1995; Flam, 1996) and the occurrence of sets of protein structures (Bornberg-Bauer, 1997). In contrast to these cases, the power law distribution in our network refers to interactions rather than the occurrence of individual biological elements. The network presented here will expand as new structural families enter the Protein Data Bank, but there is no reason to expect the form of the network to change.

This type of distribution, a scale-free network, has recently been described for metabolic pathways (Jeong *et al.*, 2000) and the World-Wide Web (Albert *et al.*, 2000). Indeed, when we look at the interactions between individual yeast proteins rather than families, the resulting network has a very similar appearance to that of protein family interactions. In the same way that exponential distributions characterise the sequence and structure family distributions of individual proteins (Teichmann, 1999; Wolf *et al.*, 2000), many different types of networks of proteins may turn out to be scale-free. In our family interaction network, this must be due to the evolutionary pressures on the different types of families. In other biological networks, such as metabolic pathways or individual protein interactions, the form of the scale-free network means it is robust and tolerant to error, because many of the nodes that are not highly connected can be removed without damaging the network very much.

The interactions common to the intermolecular and intramolecular family interaction repertoires

The family interactions in the PDB and in yeast can be divided into two repertoires: the set of domain interactions within polypeptide chains (intramolecular) and the set between polypeptide chains (intermolecular). (These are distinguished by the thick or thin broken or continuous lines connecting the families in Figure 1, and the different types of interactions are listed separately in Table 1.) Comparing the intramolecular and intermolecular repertoires is of interest to see whether types of domains that interact within polypeptide chains also do so when on different polypeptide chains. Gene fusion (Enright *et al.*, 1999) and the co-existence of

domains within proteins (Marcotte *et al.*, 1999) have been used by other authors as a means of predicting interactions between proteins. Tsoka & Ouzounis (2000) point out that gene fusion events most frequently take place between metabolic enzymes. We can shed light on the extent and the functional category of the domain interactions that take place within as well as between polypeptide chains, by investigating the overlap between the intramolecular and intermolecular repertoires of protein family interactions.

From a physical perspective, there is little difference between multi-domain and multi-subunit interactions (Argos, 1988; Janin *et al.*, 1988; Jones *et al.*, 2000). The protein interfaces of complexes of proteins that can also exist independently differ somewhat from multi-domain and multi-subunit proteins in that they are less hydrophobic (LoConte *et al.*, 1999; Jones & Thornton, 1997). The main difference between intramolecular and intermolecular interactions between domains is functional: domains on the same chain are always expressed together as well as being localized together. Co-regulation and co-localisation can be advantageous for certain independent polypeptides that are part of the same enzyme, oligomer or protein, or some enzymes in a multienzyme complex. However, being encoded by separate genes has the advantage that they can then be independently regulated, for instance to modulate flow through a metabolic pathway.

The extent to which the two repertoires overlap for the combined set of yeast and PDB interactions encompasses 103 types of interactions: this is 31% of the 330 family interactions within and 24% of the 435 family interactions between polypeptide chains. As shown in Figure 4, of these 103 interactions, 60 are interactions between domains from the same family, as discussed in the next section, and 33 are between domains of enzymes involved in small-molecule or macromolecular metabolism. The remaining ten family interactions can be classified as follows: two are between domains in transcription factors, two are between domains in electron transfer chain proteins, four are miscellaneous and only two are in signal transduction proteins. Both of the family interactions in signal transduction proteins involve the very common P-loop nucleotide triphosphate hydrolase family. These results suggest that the functional constraints on domains within polypeptide chains have divided the repertoire of intramolecular and intermolecular interactions into distinct sets. The overlap is heavily biased to specific types of family: interactions between domains from the same family (as in oligomers) and enzyme domains, where the domains can exist independently or as part of a large multi-domain protein.

This bias is not due to a lack of families involved in signal transduction in our data set. For instance the following three families in the top right quarter

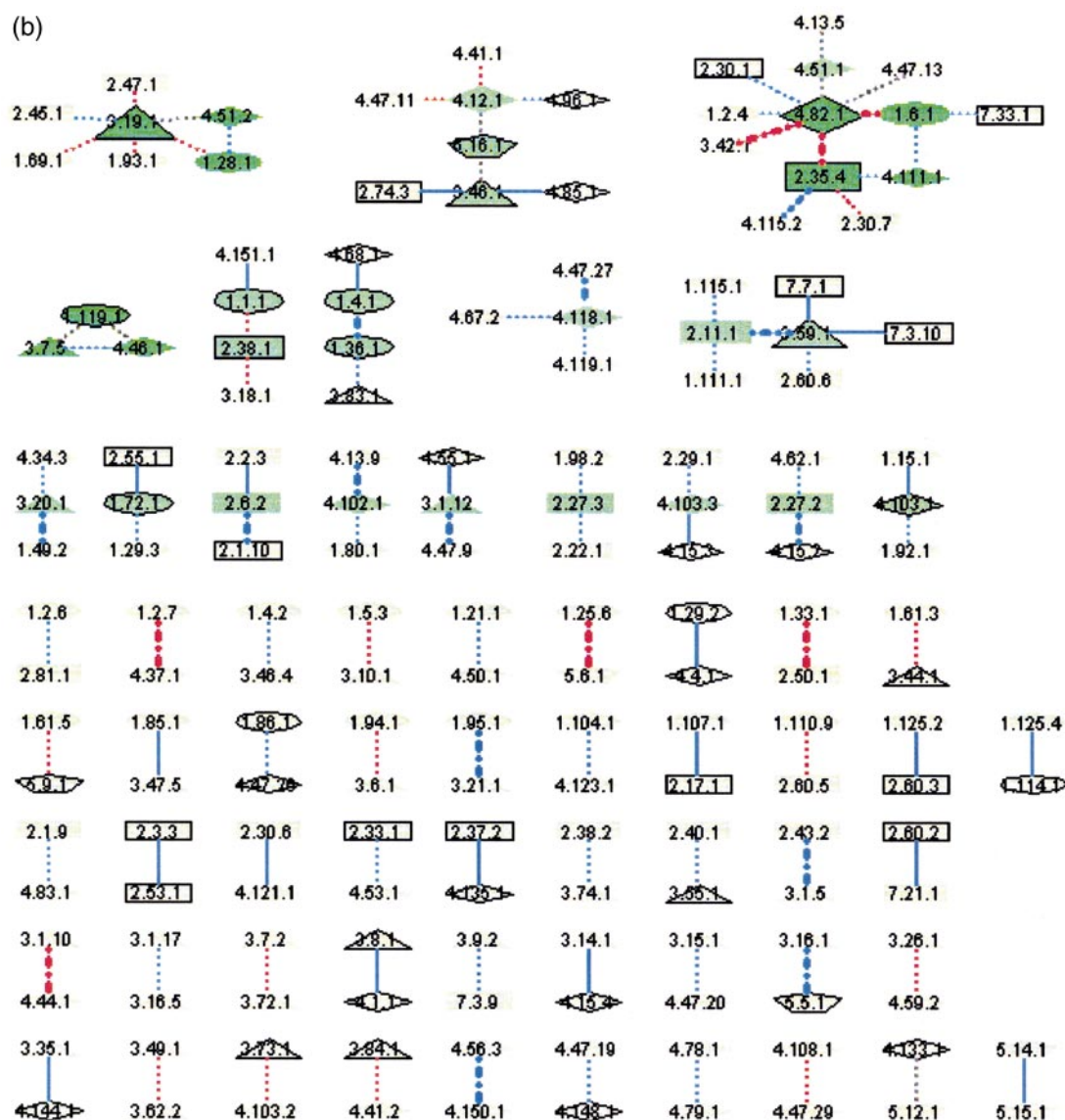


Figure 1. Family interaction maps. Each family is represented by a shape according to its class in the SCOP database and identification number in version 1.48 of SCOP. (Ellipses: all- α ; rectangles: all- β and small proteins; triangles: α/β ; diamonds: $\alpha + \beta$; pentagons: multidomain proteins; hexagons: membrane and cell-surface proteins.) The size of the shape of a family is proportional to the number of family interactions it undergoes. Black edges to a shape means some members of the family interact with each other. The interactions are colour-coded as follows: blue, interactions in the PDB only; red, interactions in the PDB and yeast; grey, interactions in yeast only. Inter and intramolecular interactions can be distinguished by the type of line connecting two families: continuous line, intermolecular; thin broken line, intramolecular; thick broken line, both intra and intermolecular. More than one-third of the families are part of the big connected cluster shown in (a), centring around the immunoglobulin superfamily (2.1.1) and the P-loop nucleotide triphosphate hydrolase superfamily (3.30.1). Other highly connected nodes are 1.110.1 (armadillo repeat), 4.117.1 (protein kinases, catalytic core) and 2.41.1 (trypsin-like serine proteases). The names of all other families, shown here only by their SCOP identification number, can be obtained from the Supplementary Material web page. Graph (a) shows 229 out of at least 572 protein families that interact in the PDB and yeast. The remaining families are part of smaller isolated clusters in (b).

of Figure 1 are classical signal-transduction domains: SH3 domains (SCOP identifier 2.30.2), SH2 domains (4.72.1) and PH domains (2.49.1). As can be seen from Figure 1, the interactions for these domains are either intramolecular (thin bro-

ken lines) or intermolecular (continuous lines), but never both (thick broken lines). Furthermore, since our data set consists of data from yeast as well as from the PDB, the interactions should not be biased towards domains for which the structure

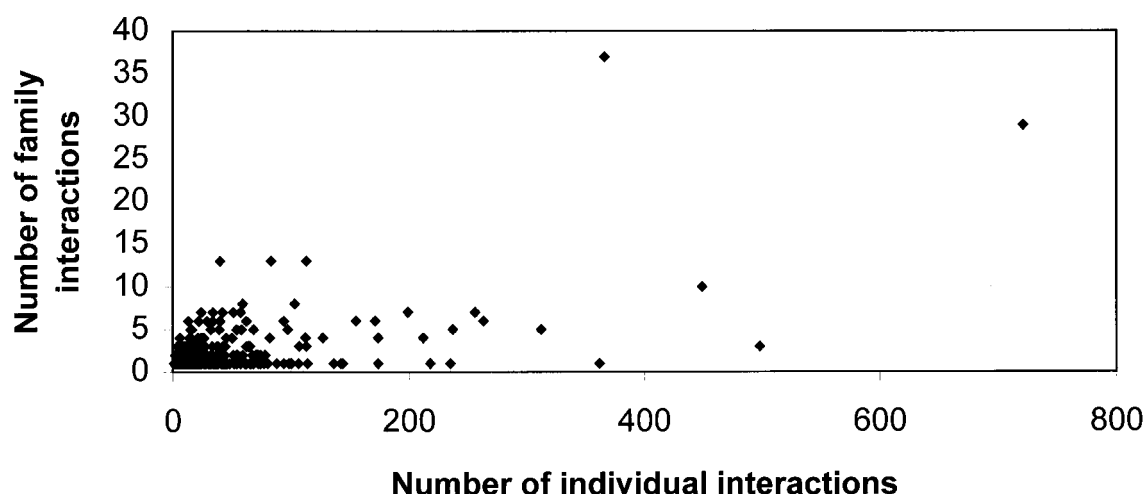


Figure 2. Graph of the number of individual interactions against the number of family interactions. On the x -axis, the number of individual interactions that are observed for one particular family in the PDB and yeast, both intra and inter-chain, is plotted. On the y -axis, the number of family interactions for that family is shown. There is no correlation between the number of individual interactions and the number of family interactions.

was solved individually instead of the whole multi-domain protein, or towards high-affinity complexes.

Oligomers: interactions between domains from the same family

A special subset of the protein family interactions are those between members of the same family. Domains of the same family can interact with each other within a protein in cases of internal duplication, or between proteins in homo-

multimers or multimers of different proteins from the same family. As described below, this occurs so frequently that the evolution of protein-protein interactions between members of the same family must be especially favourable, perhaps for reasons of symmetry. One way multimers can potentially evolve is by domain swapping, as described in detail by Bennett *et al.* (1995).

In the PDB, there are 70 families in which two domains from the same family contact each other within a single polypeptide chain, and 307 families in which two domains from the family interact

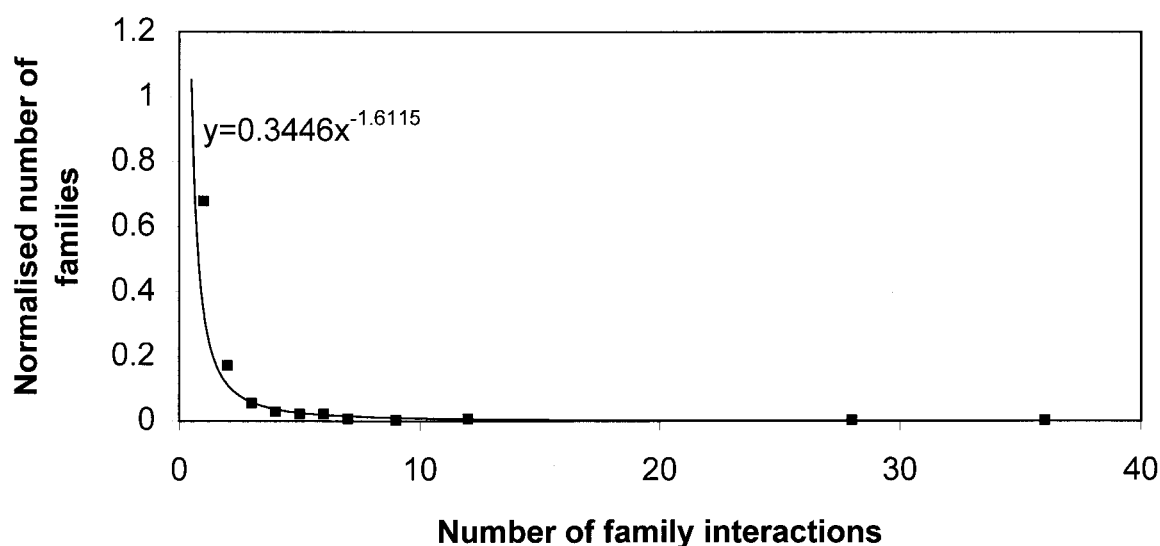


Figure 3. Power law fit for number of family interactions. The number of family interactions is given on the x -axis, while the number of families that have that many interactions is on the y -axis in normalised form. The number of interactions excludes interactions between members of the same family, or self-interactions, as they are not counted as edges of the graph. The observed data, taking the PDB and yeast data together, are represented by the diamonds. The fit to the power law is represented by the continuous line and is significant at the 1% level as determined by a chi-square test.

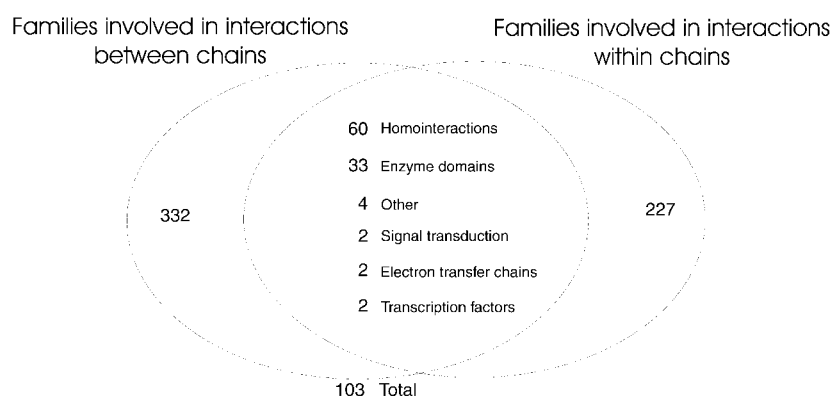


Figure 4. Overlap of intramolecular and intermolecular family interactions. This Venn diagram shows the types of interactions in the overlap between intramolecular and intermolecular interactions in the complete data set of the PDB and yeast interactions. It is clear that most interactions are either homointeractions or enzymes involved in small-molecule or macromolecular metabolism.

with each other between chains. Ten additional families are found to interact from the yeast data. All together, domains from 347 families interact with another domain from the same family, which is 45% of all the families in the PDB. These families have black borders in Figure 1. Of these families, 60 are found both in the intramolecular and intermolecular form, as discussed in the section above.

Possible applications of structural assignments to interacting proteins for target selection for structural genomics, prediction and homology modelling

Target selection for structural genomics

Structural genomics projects have as their aim the solution of the three-dimensional structures of all soluble proteins in genomes. One result of these projects will be a complete fold library of individual protein domains. If multi-domain and multi-chain structures are solved, another result will be a library of domain-domain interfaces. Our analysis of interactions between yeast proteins with structurally assigned domains can point out which pairs of proteins are of interest from this point of view, as described in the next paragraph.

In addition to the interactions between completely assigned single-domain proteins, there are 66 other interactions between completely assigned multi-domain proteins, as shown in part 3 of Figure 5. Because with multi-domain proteins we cannot resolve exactly which domains are interacting between the two polypeptides, the families interacting in these pairs are ambiguous and were not used in our calculations on family interactions. Only five of these 66 interactions have analogous crystal structures in the PDB, i.e. structures that contain two chains with the same combinations of domains that interact with each other. All the other interactions are novel, between multi-domain polypeptides with domain combinations which may or may not have been observed in the PDB. Therefore, we suggest that priority be given to structure elucidation or further experimental investigation of

interacting polypeptides that do not yet have analogues in the PDB.

Prediction

There are 371 interactions between yeast polypeptides where one chain is completely assigned and one chain has no structural assignment, as shown in part 4 of Figure 5. (Again, these could not be used for deriving family interactions.) In total, 282 of these interactions are ones in which the polypeptide chain with a structural assignment is a single-domain protein, like the second pair in part 4 of Figure 5, and these 282 domains belong to 47 families. Most of these domains are seen to interact with other families either in the PDB or in the yeast data (domains within chains or completely assigned single-domain pairs). However, eight new families are known to be involved in protein-protein interactions from this set. A total of 11 families are known to interact with one other family, so a careful prediction could be made for these interactions. The reliability of such a domain prediction, which would endeavour to provide information on the structure and function of the unknown polypeptide, depends on the pair of families involved.

Homology modelling

The yeast intermolecular family interactions discussed above are derived from pairs of yeast polypeptides known to interact from experimental data, where each polypeptide has a single-domain structural assignments (see Methods for details). These pairs can be modelled as complexes in three dimensions if there is an example of the two types of domain interacting in the PDB. A caveat to this type of modelling is that the members of two families can sometimes interact in different ways, using different types of interface, for example the different modes of oligomerisation of nucleoside diphosphate kinases (Giartosio *et al.*, 1996). However, the putative interface from a homology model could be tested experimentally, and a correct three-dimensional model can be helpful in

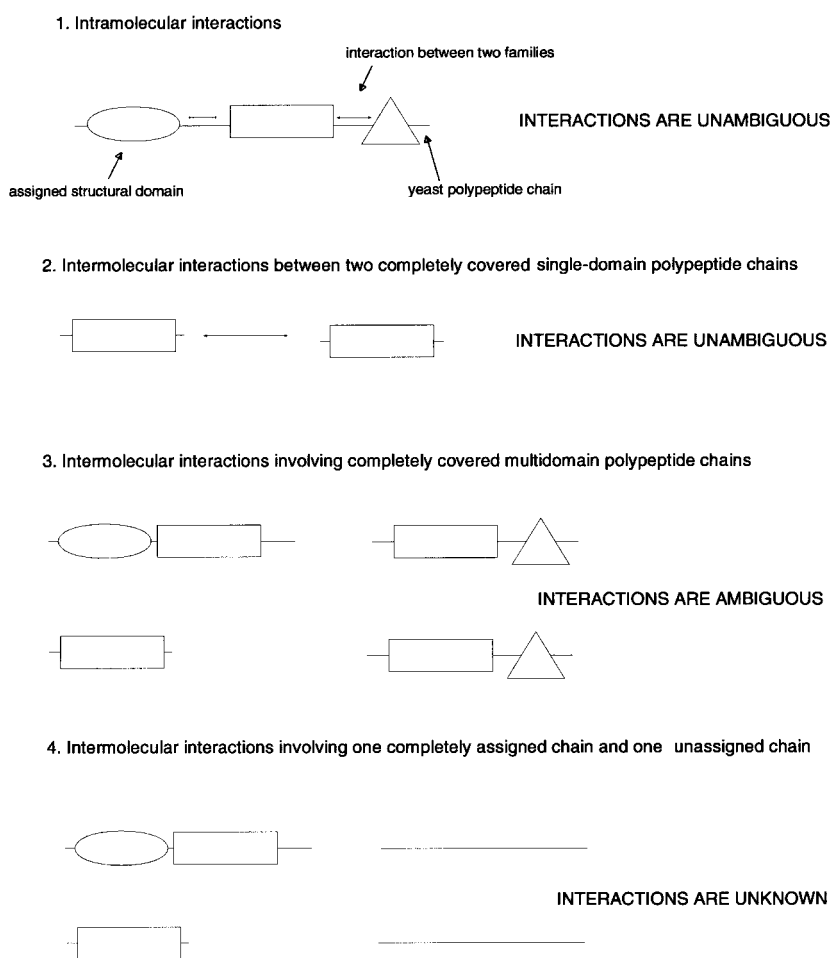


Figure 5. Types of interactions in yeast polypeptide chains. These diagrams show how we define intra- and intermolecular interactions in yeast polypeptides. In part 1, the intramolecular interactions are shown. In part 2, the set of unambiguous intermolecular interactions for yeast are derived from the pairs of yeast polypeptides that each have a single assigned domain covering most of the polypeptide. Parts 3 and 4 show types of interactions that are not unambiguous, but may have other uses.

understanding interactions and guiding further experiments.

Conclusions

This survey is the first attempt at classifying interactions between all the known structural protein domains according to their families. For the SCOP families, we have surveyed the interactions that exist within and between proteins by analysing the whole PDB and the complete yeast genome. The analysis of experimentally derived interactions between yeast proteins identified at least 51 reliable new interactions between superfamilies, which are potential targets for crystallisation and experimental investigation of the domain interfaces.

The library of interactions between families built up in this analysis also has some predictive value: complexes between genome sequences with structural assignments can be modelled in three dimensions. In addition, for interactions in which one partner does not have a structural assignment, possible structures can be picked from the set of known family interactions.

Several interesting results with respect to the evolution of protein families have also emerged from this survey. The graph of family interactions takes the form of a scale-free network. This means that there are a few pivotal families that interact with many different families, while the large majority of families have only one, two or three partner families. Each family has its own spectrum of partners, with few connections between the families. Furthermore, 45% of the families are found to interact with members of the same family, indicating that internal duplication and oligomerisation of domains is very favourable, perhaps due to symmetry of the interaction interface.

An intriguing result of this work is that the repertoire of family interactions between domains within single proteins only overlaps with the repertoire for interactions between separate proteins for specific types of families. Evidently functional and regulatory constraints in evolution have separated the domains that are brought together by recombination and gene fusion within proteins, and the domains that interact between proteins. The result is that the pairs of families that interact both within and between polypeptide chains belong mostly to

two types of domains: enzyme domains and domains from the same family.

Materials and Methods

Protein families in the SCOP database

The domain definitions and evolutionary families in the Structural Classification of Proteins (SCOP) database (Murzin *et al.*, 1995) version 1.48 were used. This database contains 21,828 structural domains in 764 superfamilies from 9580 PDB entries. In SCOP, domains are defined as evolutionary units, and so each domain has either been observed as a single polypeptide, or has been seen in combination with at least two other domains. Domains belonging to the same SCOP superfamily are related as detected by a combination of sequence, structural and functional clues. Here, we used the 764 superfamilies in SCOP to build maps of interacting families, so that the family-family interactions have an evolutionary meaning. Therefore, when the term "family" is used here, a SCOP superfamily is meant.

Interactions between families in the Protein Data Bank

To determine which families of domains interact with one another in the 9580 PDB entries in SCOP, the coordinates of each domain were parsed to check whether there are five or more contacts within 5 Å to another domain. The distance of 5 Å was chosen as this is a conservative threshold for interaction between two atoms, where the atoms are either C α atoms or atoms in side-chains. The five-contact threshold was chosen to make sure the contact between the domains was reasonably extensive. (In fact, the number of domains identified as contacting each other hardly changed for thresholds between one and ten contacts and 3 to 6 Å.) The program for parsing PDB coordinates will become available upon publication at: <http://www.biointeraction.net/PSIPFI/>

The results obtained for contacts between domains with this method are physiological, except for the small number of interactions that are due to crystal packing. Distinguishing between crystal packing and oligomer interfaces is also difficult in some cases when analysing possibilities for symmetric homooligomers, as described by Henrick & Thornton (1998) and Ponstingl *et al.* (2000). Our method does not take account of symmetric homooligomers for which only one of the monomers is in the PDB entry, so the number of homomultimeric family interactions, presently at over 45% of all families, might be slightly underestimated in this survey.

Interactions between families in the yeast genome

The protein families in the yeast genome were obtained by assigning protein structures to the yeast proteins using the domains in SCOP (version 1.48) as queries for PSI-BLAST searches, as described by Park *et al.* (1998) and Teichmann *et al.* (2000). In addition, the yeast sequences were compared to the PDB intermediate sequence library (PDB-ISL) with FASTA, as described by Teichmann *et al.* (2000). In total, 40% of the polypeptides of the yeast genome were assigned a structure with this procedure in 368 SCOP superfamilies.

The interactions between protein families within polypeptide chains were calculated by assuming that

assigned structural domains adjacent to each other on one polypeptide chain interact if there are less than thirty amino acid residues between the structural domains, as shown in part 1 of Figure 5. This criterion of 30 residues was chosen as there are only 1.5% of all SCOP domains (version 1.48, filtered at 95% sequence identity) below this threshold, but 3% of all domains are between 30 and 40 residues, for instance. So any stretch of residues longer than 30 has a significant probability of containing a domain undetected by our structural-assignment method, while this is not the case for shorter regions. Using this stringent criterion of only 30 residues between adjacent structurally assigned domains means that there are only three predicted transmembrane helices between any pairs in our set of adjacent domains, as calculated using the program TMHMM (Sonnhammer *et al.*, 1998). Domains with linkers of less than 30 residues always interact with each other in the PDB, with a few exceptions, such as domains in transcription factors like adjacent zinc fingers, or variable and constant immunoglobulin domains. Whether or not this assumption holds in the same way for genome sequences is unclear. However, an overestimate of the interactions in the yeast intramolecular data set will not change our major conclusions, which are the form of the protein family interaction network, the extent of homomultimeric interactions, the overlap of intra and intermolecular interactions and predictions from structural assignments to interacting proteins. The set of interactions encompasses 136 family interactions, as detailed in Table 1, of which 100 overlap with interactions seen in the PDB.

The interactions between pairs of yeast proteins were obtained from three sources: the Munich Information Centre for Protein Sequences (Mewes *et al.*, 2000) web pages on complexes and pairwise interactions (<http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/index.html>, February 2000), the global yeast-two-hybrid experiments by Uetz *et al.* (2000) and large-scale yeast-two-hybrid experiments by Ito *et al.* (2000). Out of 2492 pairwise interactions known experimentally for yeast, 1819 have a structural assignment to at least one of the polypeptide chains of the pair.

Taking the set of completely assigned (to within 100 residues) single-domain intermolecular interactions of the type shown in part 2 of Figure 5, there are 55 non-redundant protein family interactions between 52 families, as shown in Table 1. In total, 28 of these interactions overlap with interactions seen in the PDB.

Acknowledgements

S.A.T. has a Beit Memorial Fellowship for Medical Research. We thank Cyrus Chothia for helpful discussions. We also thank Liisa Holm, Andreas Heger, Sabine Dietmann and Maryana Huston.

References

- Albert, R., Jeong, H. & Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, **406**, 378-382.
- Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101-113.

- Bennett, J. J., Schlunegger, J. P. & Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* **4**, 2455-2468.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
- Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393-2403.
- 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.
- Flam, F. (1996). Hints of a language in junk DNA. *Science*, **266**, 1320.
- Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.
- Giartosio, A., Erent, M., Cervoni, L., Morera, S., Janin, J., Konrad, M. & Lascu, I. (1996). Thermal stability of hexameric and tetrameric nucleoside diphosphate kinases. Effect of subunit interaction. *J. Biol. Chem.* **271**, 17845-17851.
- Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure server. *Trends Biochem. Sci.* **23**, 358-361.
- Huynen, M., Snel, B., Lathe, W. & Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1024-1210.
- Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T. & Nishizawa, M., *et al.* (2000). Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143-1147.
- Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155-164.
- 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. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121-132.
- Jones, S., Marin, A. & Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77-82.
- Konopka, A. K. & Martindale, C. (1995). Noncoding DNA, Zipf's law, and language. *Science*, **268**, 789.
- LoConte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177-2198.
- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753.
- Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S. & Weil, B. (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**, 37-40.
- Murzin, A., Brenner, S. E., Hubbard, T. J. P. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536-540. (see also <http://scop.mrc-lmb.cam.ac.uk/scop>).
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.
- Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins: Struct. Funct. Genet.* **41**, 47-57.
- Sonnhammer, E. L. L., Von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *ISMB*, **6**, 175-182.
- Teichmann, S. A. (1999). Protein families in eighteen genomes. In *Genome Evolution: Analysing Proteomes with New Methods*. PhD Thesis, Cambridge University.
- Teichmann, S. A., Chothia, C., Church, G. M. & Park, J. (2000). Fast assignment of protein structures to sequences using the Intermediate Sequence Library PDB-ISL. *Bioinformatics*, **16**, 117-124.
- Tsoka, S. & Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genet.* **26**, 141-142.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S. & Knight, J. R., *et al.* (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623-627.
- Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897-905.

Edited by J. Karn

(Received 15 November 2000; received in revised form 7 February 2001; accepted 7 February 2001)



<http://www.academicpress.com/jmb>

Supplementary material is available at: http://www.biochem.ucl.ac.uk/~sat/pdb_sc_int.html A related website is: <http://www.biointeraction.net/>