Published in final edited form as:

J Mol Biol. 2008 September 5; 381(3): 785–802. doi:10.1016/j.jmb.2008.04.071.

Architectures and Functional Coverage of Protein-Protein Interfaces

Nurcan Tuncbag¹, Attila Gursoy¹, Emre Guney¹, Ruth Nussinov^{2,3,*}, and Ozlem Keskin^{1,*}

¹Koc University, Center for Computational Biology and Bioinformatics and College of Engineering, Rumelifeneri Yolu, 34450 Sariyer Istanbul, Turkey

²Basic Research Program, SAIC-Frederick, Inc., Center for Cancer Research Nanobiology Program, NCI-Frederick, Frederick, MD 21702

³Sackler Inst. of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Abstract

The diverse range of cellular functions is performed by a limited number of protein folds existing in nature. One may similarly expect that cellular functional diversity would be covered by a limited number of protein-protein interface architectures. Here, we present 8205 interface clusters, each representing unique interface architecture. This dataset of protein-protein interfaces is analyzed and compared with older datasets. We observe that the number of both biological and crystal interfaces increase significantly compared to the number of PDB entries. Further, we find that the number of distinct interface architectures grows at a much faster rate than the number of folds and is yet to level off. We further analyze the growth trend of the functional coverage by constructing functional interaction networks from interfaces. The functional coverage is also found to steadily increase. Interestingly, we also observe that despite the diversity of interface architectures, some are more favorable and frequently used, and of particular interest, those are the ones which are also preferred in single chains.

INTRODUCTION

Most molecular and cellular processes are controlled by protein-protein interactions. Deciphering the mechanism of protein interactions is crucial for the comprehension of the large scale organization of the cells and their biological pathways (recently reviewed in ¹). Recent studies have discovered thousands of novel protein interactions, changing our perception of the cellular organization. Protein interactions change during development and in response to external stimuli. Therefore, the interaction networks they form are dynamic, regulating and supporting each other. Antigen-antibody recognition, enzyme substrate binding, hormone receptor binding, RNA splicing, DNA replication, transcription, and signaling pathways are some examples of the diverse and complex biological processes dominated by protein-protein interactions. Pioneering studies on the mechanism of protein-protein recognition provided insights into the properties of different types of protein-protein complexes and the principles

*Correspondence should be addressed to O. Keskin (okeskin@ku.edu.tr) or R. Nussinov (ruthn@ncifcrf.gov). The publisher or recipient acknowledges right of the U.S. Government to retain a nonexclusive, royalty-free license in and to any copyright covering the article.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

of the interactions^{2; 3; 4}. Protein-protein interactions can take place between identical and non-identical chains (homo- or hetero-complexes) and can be obligate or non-obligate³. The life time of the interaction can be classified into transient or permanent⁵. Such classifications assist in studies of the relationship between protein recognition and their biological functions^{3; 5; 6}.

It was proposed that there are approximately 1000 types of protein folds in nature⁷. It is remarkable how, with such a relatively small number, nature can still perform the immense functional diversity. In a similar vein, it was stated that currently there are 700–800 known protein folds and despite an exponential increase in the number of structures, the increase in folds is leveling off which makes the 1000 fold estimate still hold⁸. Since proteins generally associate with each other to carry out their functions, the next obvious questions to ask are (i) how many different protein-protein interaction types exist in nature that will allow the diversity in biological processes; and (ii) what fraction of the interaction types is already known toward elucidation of the cell's organization. Aloy and Russell estimated that there should be 10,000 protein-protein interaction types based on distinct structural types of the folds that associate⁸. They defined the 'interaction type' as the interaction equivalent of a protein fold. This number does not include antibody-antigen interactions, membrane proteins, protein-peptide complexes and 'very special' interactions. As a result, they estimated that in 2004, 1800 of the 10000 types of interactions would already be known with the number expected to increase at a rate of 250 new interactions per year.

Interactions between proteins take place through their interfaces. The number of distinct protein interaction types should be related to the number of distinct protein interfaces. Thus, identification of different protein interfaces can provide information regarding how close we are to the limit in the number of interaction types. Protein interfaces have long been studied at both the protein and the domain levels. They have been represented as interface datasets^{9; 10} and deposited into databases such as PiBASE¹¹, InterPare¹², SCOPPI¹³, SCOWLP¹⁴, 3did¹⁵ and iPfam¹⁶. Interfaces are extracted from multi-chain X-ray or NMR structures. However, not all the complex structures in the PDB¹⁷ are biologically relevant. Some are crystal packing interactions, and may cause noise in analyses. A number of studies addressed the problem of distinguishing between biological and crystal packing interfaces. In PQS, Henrick and Thornton define crystal packing interfaces by assigning a cutoff value (400 Å²) in the buried surface area¹⁸. Based on the assumption that the biological interfaces are more conserved than non-biological interfaces, Valdar and Thornton¹⁹ suggested that interfaces can be distinguished by residue conservation. Zhu et al.²⁰ proposed that the amino acid and chemical composition of the interfaces are also useful in identification of crystal packing interfaces. Using a combination of interface properties in a support vector machine application, NOXclass distinguishes biological and non-biological interfaces reaching an accuracy of 91.8 % based on three parameters (interface area, interface area ratio and area based amino acid composition)²⁰.

Large scale protein-protein interface and interaction datasets allow analyses of interface properties, such as sequential and structural conservation, residue propensities, interface size, shape, and complementarity (reviewed in^{21; 22}). These provide insight into the types and evolutionary history of protein interactions. Permanent interfaces are more hydrophobic than transient interfaces and evolve more slowly²³. Binding site residues are closer to the center of mass than the non-binding surface residues²⁴ and can be predicted by local statistical properties of the protein surface²⁵. Interfaces can be used as a framework for prediction of interactions providing clues to protein recognition mechanisms^{26; 27}. Phylogenetic trees, which incorporate information on the evolutionary history of the species, also were used to distinguish between protein interactions²⁸.

In this paper, we present our current dataset of protein-protein interfaces, and its comparison with the older datasets (1994 and 2002). We examine whether (i) the increase in the number of unique interface architectures is leveling off and (ii) whether the functional coverage and the functional interaction maps in the PDB are complete. The increase from 1629 interfaces and 351 non-redundant clustered architectures in 1994¹⁰; to 21686 interfaces and 3799 clusters in 2002⁹; to currently 34817 PDB structures, which contain 15268 multimeric structures, with 49512 protein interfaces leading to 8205 structurally distinct clusters allows us to carry out such a study. We further annotate the interfaces as biological or crystal and compare the yearly growth of each type of protein interfaces. The number of interface clusters is found to increase at a much faster rate compared to the number of protein folds and is yet to reach a plateau. The functional coverage of the interactions is also found to increase continuously in the PDB, leveling off during the last years though there is a significant number of missing interactions among these functions.

RESULTS and DISCUSSION

In our current dataset, we started with 49512 two-chain interfaces extracted from the Feb, 2006 version of PDB. The hierarchical clustering of the interface architectures resulted in 8205 interface clusters. One of the members is designated as the representative interface of the cluster. The dataset is available at <http://prism.ccbb.ku.edu.tr/interface>.

Biological and crystal interfaces

Protein interactions can occur between homo- or hetero-dimeric chains, and the interactions can be obligate or non-obligate, depending on whether the monomers are stable on their own. In order to perform meaningful analysis, one needs to be sure that the interfaces studied are obtained from biological complexes. To characterize the interfaces in our dataset, we separate them into biological versus crystal; and obligatory versus nonobligatory by using NOXclass²⁰. Table 1 summarizes the results for all datasets (current, 2002, and 1994): the first row gives the number of structures available in the PDB on the date studied; the second row provides the number of PDB entries containing multiple chains. The total number of interfaces and interfaces coming from homo-, hetero- dimers are also presented. The number of characterized interfaces by Noxclass is given. Note that this number is smaller than the total (since some complexes do not have conservation scores from CONSURF²⁹, gap volume indices; or the ASA of some large proteins can not be calculated by NACCESS, thus cannot be classified by NOXclass). In the current dataset, 17210 and 10545 interfaces are annotated as biological or non-biological, correspondingly. Of the biological ones, 14501 and 2709 are found to be obligate and non-obligate, respectively. The relatively high number of crystal interfaces suggests that, one needs to be careful when using large scale interface datasets. Almost half of the interfaces might be artifacts of crystallography or the high concentrations of the proteins studied. The characteristics of interfaces are available at <http://prism.ccbb.ku.edu.tr/interface>. Domain classifications of the constituent chains are provided in this link.

We further note that we eliminate interfaces formed by non-globular folds, antigen-antibody complexes, membrane proteins, synthetic and theoretical proteins and protein-peptide complexes (chains under 100 residues with no SCOP domains). Out of those 17210 biological interfaces, following filtering we obtain 9873 biologically relevant interfaces between globular folds, which form 1901 interface clusters. We call this remaining dataset the filtered set. In the 2002 and 1994 datasets, the numbers of filtered clusters were 1006 and 84, respectively.

The growth trend of protein-protein interfaces

The old datasets extracted in 1994 and 2002^{9; 10} and the new dataset (2006) are compared. The substantial increase in interfaces makes the newly generated dataset more diverse. Table

2 shows the change in the number of PDB structures, interfaces, clusters and SCOP families³⁰. The number of interface clusters increases from 351 (in 1994) to 3799 (in 2002) to 8205 in 2006. The number of PDB entries increases from 2814 to 18687 to 34817.

The generation of the current dataset with new PDB structures illustrates both (i) an increase in the number of interface clusters, i.e. newly discovered interface architectures; and (ii) an increase in the population of clusters i.e. more interfaces with a given architecture. Fig 1 presents the increases in the number of different SCOP families, interfaces, distinct interface clusters and PDB structures over the years. Since three points are insufficient for obtaining reliable growth rates, we map the three points to annual data points. In order to find these annual growth points, we assign clusters to a particular year if the representative or any of the member structures in the clusters were available in the corresponding year in the PDB. Fig 1A displays the growth trends of the PDB structures, as well as overall, biological and crystal interfaces. In all cases, the growth trends are significant with the interfaces following a faster increase. The same result was obtained in a previous study based on domain-domain interfaces³¹. Note that some of the interfaces could not be annotated as crystal or biological, therefore the number of interfaces is higher than the sum of the biological and crystal interfaces in the present study. Fig 1B presents the increase ratio in protein families from SCOP, protein interface clusters and PDB entries. Clearly, there is a rapid increase in the number of interfaces and interface clusters when compared with the increase in the number of SCOP families and PDB structures. Thus, the currently available clusters suggest a broad diversity. In the 12 year period from 1994 to 2006, while there is a ~13 fold increase in the number of PDB entries, the number of biological interfaces has increased ~30 folds. Considering the number of PDB entries and the number of interfaces (or, similarly SCOP families compared to the interface clusters) of the three datasets, we observe an exponential increase; however, the increase in the number of interface clusters is much faster than in the number of SCOP families. This rapid increase in the number of interface clusters may also be the result of the rapid growth in the determination of the large multi-chain complexes, which contain more than one interface. Fig 1C illustrates the annual growth rate of the filtered interface set. The exponential increase in the cumulative number of biological interface clusters (and the annual increase of biological interface clusters) is also observed when only globular proteins are considered.

If we consider that these unique representative interface architectures (clusters) are related to the interaction types in the PDB as proposed by Aloy and Russell⁸, our results are in agreement with theirs. They proposed that eventually 10000 interaction types would be observed and the interaction types would grow by 200–300 per year since midnineties. Our results show that starting from 1999, the average annual growth rate of the filtered interfaces has been around 250/year.

The chain level interfaces are also mapped to their domains to extract domain-domain interfaces in the inter-chain interactions. In Fig 1D, the domain-domain interface growth rate is illustrated for inter-chain interfaces for both the biological and overall interface dataset. A growth trend similar to the increase of chain-level interfaces is observed between the years 1999 and 2005. We further analyze the increase in the number of interfaces with increase in the number of folds between 1999 and 2006 by comparing the annual increase in biological new interface clusters and SCOP folds (Fig 1E). We observe that interface clusters grow faster than the folds.

In addition to the substantial increase in the number of clusters, not surprisingly there is also an increase in the cluster populations (Fig 2). Currently the cluster populations are larger than in 2002. The largest cluster in 2002 had 281 members which increased to 421 in 2006, suggesting frequent usage of the same favorable interface motifs. In 2002, 2946 clusters (out of 3799) had less than 5 members; in 2006 dataset 6424 clusters (out of 8205) have less than

5 members. Further, in the 2006 set there are 6 clusters with more than 300 members. No such cluster was present in 2002.

Other interface databases

Table 3 compares three recent interface databases and the one presented here. The first row lists the name of the databases. The next rows show the attributes of the databases, such as interface level (where the interfaces are extracted from, either chains or domains), and some key aspects of the databases. All the databases (SCOPPI, CBM, PIBASE) in the table represent interfaces between domains. Protein structures in SCOPPI¹³ are taken from PQS¹⁸ to generate domain-domain interfaces. In SCOPPI, interfaces are clustered geometrically. The resulting domain-domain interfaces can be filtered explicitly according to their Δ ASA and sequence identity. The Conserved Binding Mode (CBM) dataset is constructed by using structural alignment³² of domains. Interacting domain pairs from different organisms are clustered to obtain unique interface geometries. If two or more members in the resulting clusters have similar interface locations, they are defined as conserved binding modes. By using the CBM, the authors distinguish between biological interactions and crystal packing interactions. In PiBase¹¹ interfaces are generated by using the PDB and PQS¹⁸ databases. To remove redundancy, interfaces are further eliminated according to an ASA threshold ($300\text{\AA}^2$) and duplicated domain-domain interactions. Protein complexes, interfaces and binding sites are clustered by their topologies.

Examples of similar architectures in interfaces and folds

Previous studies have indicated that protein binding and protein folding are similar processes^{33; 34; 35; 36}. Forces that stabilize a protein fold are also responsible for protein-protein associations. Tsai *et al.* showed that the global features of the architectural motifs, present in monomers, recur in the interfaces. They further found that interfaces derived from two-state complexes, where the chains fold cooperatively, display higher similarity to architectures in protein cores compared to the three-state complexes, where the folds first fold and then associate³⁶.

When the two sides of the ten most crowded interface clusters (twenty binding sides) in the current dataset are investigated, we see that six of these are helical, eleven are composed of both helices and strands whereas three are formed by strands. Similarly in the 2002 dataset, when the most crowded ten clusters are studied, seven of the twenty sides are solely helical; eight are composed of both helices and strands whereas five are formed by strands. Thus, combinations of helices and strands are commonly used in the interfaces and there is a slight preference for helices compared to strands. When the top ten clusters from the two datasets are compared, clusters with an RNA binding protein, a ferritin and an antibody representatives are found to be absent in the 2002 dataset but emerged in the current dataset.

To observe similarities between single chain folds and interface architectures, we compare the representative interfaces of the top 10 populated clusters with the most populated folds in SCOP. A recent study emphasized that the most populated folds in SCOP do not necessarily have high sequence similarity³⁷. Here, the same holds for interface clusters. In SCOP, the most populated folds are the immunoglobulin-like (represented by PDB id: 1fna), Rossman (represented by PDB id: 3chy), TIM barrel (PDB id: 1ypi), jelly roll (PDB id: 1sac), α - β plait (PDB id: 1ris), 3 - helix bundle (PDB id: 1enh), globin (PDB id: 1a6n), and β - grasp (PDB id: 1pgb) folds³⁸. MultiProt³⁹ is used to compare the structures of these single-chain folds with the interface architectures. Interestingly, the most popular folds in single chains are those architectures which are most populated in the interfaces. Table 4 gives the RMSD values and the number of matching residues between the folds and the interfaces of the five best matches. For example, 1o1pAD is a representative interface composed of helices. When it is compared

with the 3-helix bundle fold (1enh), 38 out of 54 residues in 1enh match structurally with the residues in 1o1pAD. The RMSD value is 1.86 Å (considering C^α atoms) for this match which is an appropriate value to declare that the structure of the interface 1o1pAD is similar to the 3-helix bundle fold. In Fig 3, the ribbon diagrams and the matching parts of 1enh and 1o1pAD are shown. The red parts are structurally matched residues; the yellow parts are unmatched residues. The figure highlights this similarity visually. Additional data for this match is provided in the Supp. Mat. (Fig S5). Another example is the similarity between 1djrDE and 1iris, as shown in Fig 4. The RMSD value for this match (Table 4) is 1.92 Å for the 58 residue match.

Fig 5 also illustrates two more examples of the correspondence between folding and binding. In these figures, the same coloring scheme is used as in previous examples with the red parts being the structurally matched residues. The RMSD value for the structural match between 1z6oAE and 1a6n, shown in Fig 5A, is 1.84 Å for 62 residues. In Fig 5B, the similarity between the upper part of the 1ncwLH and the immunoglobulin fold 1fna can be distinguished visually. The result of the structural matching between 1ncwLH and 1fna supports this similarity. 52 out of the 91 residues of 1fna match with the residues 1ncwLH. The RMSD between these 52 matched residues is 1.70 Å. These examples illustrate that there exist structural resemblances between protein folds and interfaces with slight variances.

Functional interaction networks constructed from biological interfaces

It is the functional and modular organization of interactomes that controls the complexity of the organisms⁴⁰. To analyze the functional coverage of the PDB, we generate the functional interaction network. The network is constructed by assigning first level GO functions⁴¹ to proteins that constitute the interface representatives. The functional networks of the previous datasets are also constructed to detect the progress in the functional coverage of the datasets and to compare them with each other.

The interaction network of the functional classes of the filtered dataset are formed by using the first level GO functions⁴¹. First level molecular functions with their function IDs are given in Fig 6. The comparison of the putative functional interaction networks of three interface datasets (1994, 2002 and the current) are shown in Fig 7 (upper panel). As expected, we observe that pair-wise functional interactions between binding and other classes, and also between catalytic activity and other classes are high in the networks since most of the proteins are annotated with these two GO functions. To show the development of the functional coverage more clearly, binding and catalytic activity and their connections are removed in the figures (the interaction networks with catalytic activity and binding are provided in the Supp. Mat. Fig S3). The chronological progress of functional interaction networks in Fig 7 demonstrates the growth of the PDB-based functional network in the 12 year- period. We observe that functional interactions are increasingly being discovered in the last few years. In 1994, five function nodes interacted with themselves resulting in a total of 5 interactions. There was no connection among them. In 2002, the number of nodes (functional classes) and interaction numbers (edges between functions) increased to 10 and 18, respectively. In 2006, the number of nodes and the number of edges increase to 11 and 20, respectively. The chemo-attractant and -repellant activities are still not present. On the other hand, in the 2002 dataset, the pair-wise interactions between the antioxidant activity (Node ID: 1) and transporter activity (Node ID: 19); motor activity (Node ID: 11) with itself were not yet formed. In these networks, some functional interactions are more frequent as indicated by a high number of interface clusters used for the corresponding functional interaction (the numbers on the edges); for example, in the current network, two interface clusters are observed between structural molecular activity (Node ID: 15) and translational regulator activity (Node ID: 18) functions, whereas 19 different clusters are involved between molecular transducer activity (Node ID: 10) and transcriptional regulator

activity (Node ID: 17). Overall, the number of interface clusters used in different functional interactions is not uniform and vary significantly resulting in different weights for edges.

In summary, with the new PDB structures and new interactions, the functional network has started to grow. The growth rates are shown in the lower panel of Fig 7. In this figure, the 2006 dataset is mapped to different years to get annual growth rates. The open squares represent the functions (nodes in the network) including catalytic activity and binding covered annually, whereas the black diamonds show the number of interactions between the functions (edges of the network). The figure shows that functional coverage has started to grow slowly; yet there are still unconnected functional classes which indicate that this network will continue to develop in the coming years. However, the number of interactions between the functions increases continuously; there is a jump in 2002 which corresponds to the completion of several genomes or the starting outcome of structural genomics^{42; 43}.

In order to assess the coverage of the functional interactions from PDB interfaces, we compare it with functional interaction networks from gold standard protein-protein interaction data which do not have structural constraints. A previous study showed that the top level GO molecular functions are similarly distributed in the structural database (PDB) and in the proteins of the human genome⁴⁴. We construct similar functional interaction networks starting with the gold-standard protein-protein interaction data of Jansen et al.⁴⁵ and HPRD^{46; 47}. In the gold standard networks (Supp. Mat. Fig S1A) nine of the essential functions are mostly connected to each other. In the HPRD network (Supp. Mat. Fig S1B) 12 functional classes with 71 pair-wise interactions are covered. We can assume that the PDB will eventually reach the same level of connectivity, with all functional nodes covered. Nevertheless, it is difficult to estimate when the completed network of structural protein function interactions will be reached. Clearly, the growth rate will increase with the progress in structural genomics and completion of new genomes. If we assume that the growth trend will not change in the future, by extrapolating the growth rate we speculate that the complete function interaction coverage in the PDB will be attained in about ten years leading to 17000 different interface architectures (see Supp. Mat.). However, we should note that each edge of the functional interaction network can be represented by an unpredictable number of interface clusters, so even if the structural interactome already covers all edges of the function interaction network, new interface clusters can still be discovered.

Gene fusion events observed in the dataset

Gene or domain fusions are observed when two separate proteins in one organism appear as a single homologous fusion protein in another organism. Similarly, two domain pairs that interact may belong to a single chain in one species, whereas the same domain pairs belong to different proteins in another species¹³. It was stated that the dominant mechanisms to have a diverse repertoire of proteins are gene duplication, divergence and recombination⁴⁸. Gene fusion was probably preferred since it was much faster and easier to create new proteins with different functions by combining existing ones. Even the new domains are observed to be related to old ones that have been modified so that their origin cannot be easily recognized⁴⁹. It is difficult to find distant protein family relations based on sequences only; on the other hand three dimensional structures can help reveal such relations. Close to 50 percent of the sequences in genomes are homologous to proteins of known structures⁵⁰. For instance, previously, Kim et al.³¹ found that two-thirds of the domain fusions associate in the same orientation, while the rest interact differently using the interfaces among domains. Shoemaker et al. found that usually the binding modes are conserved among different organisms³².

Our structure-based clusters allow efficient detection of domain fusion events in the PDB. Given an interface with multiple domains, one can detect physically interacting domain pairs through the interface they utilize. Instead of considering all two-chain complexes, only the

cluster representative interfaces can be used to determine fusion events. An example of gene fusion is seen in the enzyme urease as shown in Fig 8. SCOP version 1.69³⁰ was used to identify the domain classifications. The enzyme is composed of four domains, a urease gamma-subunit (d.8.1.1), a urease beta-subunit (b.85.3.1), an alpha-subunit of urease catalytic domain (c.1.9.2), and an alpha-subunit of urease (b.92.1.1). The numbers in the parentheses are the SCOP domain IDs. We compare the structures of the enzyme from three different organisms: *Klebsiella aerogenes* (pdb ID: 1a5k) and *Bacillus pasteurii* (pdb ID: 1ubp), *Helicobacter pylori* (pdb ID: 1e9y). The beta and gamma subunit domains are located on separate chains in 1ubp and 1a5k but in 1e9y are fused, and located in the same chain. As a result of this possible fusion event, 1e9y produces only one interface (1e9yAB), 1a5k produces two interfaces (1a5kAC, 1a5kBC) and 1ubp produces three interfaces (1ubpAB, 1ubpBC, 1ubpAC). Although the interactions between the gamma- (d.8.1.1) and beta- subunits (b.85.3.1) are conserved in *H. pylori* and *B. pasteurii*, there is no interaction between these domains in *K. aerogenes* in PDB. However, the absence of this interaction does not affect the function of urease in the organism. As expected, the surface area is small between the two mentioned domains, thus it is highly possible that the interface between them is non-biological and results from crystal contacts. Our structure-based clustering allows straightforward detection of such cases. Interfaces 1a5kAC and 1ubpAC are in the same cluster; 1e9yAB is a member of another cluster, although 1e9yAB contains a similar interaction between domains c.1.9.1, b.92.1.1 and d.8.1.1.

Another example for gene fusion in the current dataset is imidazole glycerophosphate synthase enzyme (Fig 9); observed also in other studies^{13; 51}. This enzyme consists of two domains, the histidine biosynthesis (c.1.2.1) and class I glutamine amidotransferase (c.23.16.1). In *Thermotoga maritime* (1gpwAB, 1gpwAD, 1gpwBE, 1gpwCD, 1gpwEF) and *Thermus thermophilus* (1ka9FH), these two domains interact while belonging to separate chains; however, in *Saccharomyces cerevisiae* (1jvnA, 1jvnB, 1ox4A, 1ox4B, 1ox5A, and 1ox5B) these two domains are fused and interact in the same chain. In Fig 9, the proteins are colored by their chains. Despite the *Saccharomyces cerevisiae* gene fusion, these domains interact in a way similar to the domains in other species (*Thermotoga maritime*, *Thermus thermophilus*). 1gpw has six chains which form 5 interfaces whereas 1ka9 has 2 chains forming a single interface in our dataset.

We examine the interface dataset and obtain 106 unique domain fusions between 114 domains at the super-family level of SCOP (see Supp. Mat. Fig S4). We further observed a consistent trend that primitive species prefer to have separate domains (inter-chain domain interactions), but evolutionarily advanced species prefer to fuse their related domains (i.e., intra-chain domain interactions)⁴⁹. This trend may be the result of the crowded traffic in the cells of higher organisms.

CONCLUSIONS

Comparison of the new 8205 interface-cluster dataset with the older (1994 and 2002) datasets indicates that the number of interfaces (and clusters) has increased through the 12 years period much more rapidly than the available PDB structures (and SCOP families). This growth largely stems from the larger numbers of multi-chain and high molecular weight proteins in the PDB. The increased cluster diversity implies discovery of new interface architectures; this is in addition to the observed increase in similar interface structures, as indicated by more populated clusters. The number of unique interfaces continues to grow exponentially and has still not reached its upper limit. After elimination of the peptides, membrane proteins, antigen-antibody complexes and ligands, we currently end with 1901 unique interface clusters. Plotted against the older data sets, we detect an increase in the unique type of interface architectures at a rate of around 250 new architectures per year.

The current number of distinct clusters allows analysis of the functional divergence and evolution of the interfaces. We further study the coverage of protein function-function interaction maps. We observe that the functional coverage is still not complete; only 13 first-level GO functions out of 17 are found in the PDB. We also notice that new interface architectures have emerged as well as existing interface clusters have become more populated. We further observe that some interface architectures are more favorable and frequently used in protein-protein associations. Not surprisingly, when these interface architectures are compared with domain folds, we observe a high degree of similarity. Nature appears to use similar preferred fold templates for single chains and for interfaces.

The current protein-protein interface dataset is available at <http://prism.ccbb.ku.edu.tr/interface/>. Accessible surface area (ASA), gap volume, gap volume index of the interfaces, as well as the coordinate files of the interface residues are provided in the dataset. Interfaces are characterized as obligate/transient and biological/crystal. External links to HotSpring (Hot spots in interfaces) ⁵², SCOP and GO are given.

METHODS

The current interface dataset is generated following the flow chart in Fig 10. Initially, binary interfaces are extracted according to the atomic distance constraints between residues (detailed below). Next, they are compared structurally using the Geometric Hashing algorithm and structurally clustered. Detailed information about the clustering algorithm is available in our previous works ^{9; 10}. Using the NOXclass algorithm, the clustered protein interfaces are identified as biological or non-biological. In the second NOXclass step, biological interfaces are characterized as obligate or non-obligate complexes ²⁰. Interface properties are calculated, such as gap volume of the interfaces (using SURFNET) ⁵³, average conservation score of the interfaces (with CONSURF) ⁵⁴ and accessible surface area of the interfaces (NACCESS). At the end of this flowchart, each interface is registered along with its properties.

Construction of the new non-redundant dataset of protein – protein interfaces

Interface definition and extraction of interfaces from the Protein Data Bank—An interface can be defined as the set of amino acids which represents a region that links two polypeptide chains in a protein structure by non-covalent interactions. Residues interacting with each other across the binding region form the interface between two chains. Interface residues are selected according to the closeness of two residues, one from each chain. We defined two types of residues in two-chain interfaces: interacting and nearby. If the distance between any two atoms belonging to two residues, one from each chain, is less than the sum of their van der Waals radii plus a 0.5 Å tolerance, these two residues are defined as interacting. If the distance between the alpha carbons of a non-interacting residue and interacting residue in the same chain is smaller than 6 Å, the non-interacting residue is flagged as a nearby residue. Nearby residues are important for information relating to the architecture of the interface and are convenient in structural alignment of the interfaces ^{9; 10; 21; 55; 56}.

The interfaces are denoted as in our previous works. If the PDB name of a protein structure is 1fwi and there is an interface between the chains A and C, then this interface is named 1fwiAC ^{9; 10}. Fig 11 presents an example of interfaces among three chains of a protein complex (pdb id: 1fwi; chain A colored orange, chain B red, and chain C in blue). Two interfaces are formed between chains A–C and B–C. Chains A and B are not close enough to form an interface. Fig 11 shows both nearby and contacting residues. As stated above, nearby residues are important for the continuity in the interface architectures.

We have extracted all multi (two or more) chain entries from the PDB. On February 16, 2006, there were a total of 34817 structures. Interfaces with less than 10 interface residues were

eliminated since they are assumed to be crystal interfaces. As a result, 49512 two-chain interfaces were obtained.

Structural comparison of interfaces—Since protein interfaces (i) have a discontinuous structure with two separate chains, (ii) may contain isolated residues and (iii) do not have the residue order, their comparison with others necessitates a sequence–order– independent structural comparison technique. Here, the Geometric Hashing^{9; 10; 57} is used for alignment. Each residue is considered as a point in 3-D space. The algorithm uses only the C α coordinates and does not consider connectivity of the C α points in the matching. There are three consecutive steps: (i) hash table construction, (ii) voting and (iii) extension step. Clustering the dataset depends on two principles, homogeneity and separation. Homogeneity implies that elements inside a cluster should be highly similar to each other; while separation implies that elements from different clusters should have low similarity. We have applied iterative clustering. After each clustering cycle, the similarity definition is relaxed. In the first cycle, the first entry in the interface list forms a new cluster. If the similarity between the first and second entries is above the similarity threshold, the second interface entry is a member of the first cluster else it forms a new cluster. Next, the third interface is compared to existing clusters. The first cycle continues until all interfaces are assigned to a cluster. At the end of the cycle, the similarity between the cluster members and their representative should be above the current threshold.

Table 5 provides the iterative clustering steps and threshold parameters used to calculate the similarities between interfaces in this study. The first column denotes the six consecutive clustering cycles, from A to F. The second column gives the number of interface clusters at the beginning and end of the iteration cycle. At the beginning of the first iteration cycle, the number of interface clusters was the total number of two-chain interfaces in the PDB (49512). Following this cycle, the number decreased to 35744. At the end of the entire clustering process, 8205 interface clusters are obtained. Members of each cluster had at least 0.5 chain connectivity score^{9; 10; 57}, with no threshold on amino acid identity. The maximal amino acid size difference between interfaces is 50 residues.

Elimination of the crystal interfaces—A more rigorous elimination of the crystal complexes is carried out next. NOXclass is used for this purpose. The method uses six interface properties: interface area, ratio of interface area to protein surface area, amino acid composition of the interface, correlation between amino acid compositions of interface and protein surface, interface shape complementarity, and conservation of the interface, calculated by Naccess (for the first four parameters), Surfnet (for shape complementarity), Consurf (for conservation). NOXclass uses a two-stage Support Vector Machine (SVM). A NOXclass threshold of 80% is used to define a biological complex.

Structural comparison of folds and interfaces

Multiprot (<http://bioinfo3d.cs.tau.ac.il/MultiProt/>) is used to structurally superimpose protein interfaces with protein folds. The input files should be given in PDB format. The interface coordinate files are available at <http://prism.cccb.ku.edu.tr/interface/>. The parameter file used during the calculations is provided as a Supp. Mat.

A modified version of Multiprot is used to compare two or more interfaces. Because interfaces are composed of two chains, and most of the time only discontinuous segments from each chain are involved in the binding, Multiprot which is independent of the order of the residues on the chains is a good choice. Thus, let us assume that we have two interfaces: Interface1 and Interface2. The first interface is composed of chains ChainA and ChainB, and the second of ChainC and ChainD. All the segments from ChainA are aligned with segments from ChainC (or ChainD) but not a combination of segments from both ChainC and D. In order to perform

structural alignment of interfaces with Multiprot, ChainEq 0 should be replaced by ChainEq 1 in the parameter file.

Construction of the functional interaction networks

We start with the filtered biological interface representatives (containing 1901, after elimination of the antigen – antibody complexes, membrane proteins, peptides, synthetic proteins etc. over interfaces identified as biological). Each interface chain is annotated by its GO IDs. Then, each GO ID is mapped to the first level molecular functions described in GO annotations. If the function of a representative interface is unknown, the second member in the same cluster is used. The interface is identified as a functional representative interface and the function is used for that cluster. As a summary, Fig 12 illustrates the procedure for the generation of the functional network. Here, if two chains interact with each other through an interface, the functions of the chains should also be related. For example, supposing that the interface I_1 is composed of the partner chains L_1 and R_1 . The functional annotation of R_1 is F_1 and the functional annotation of L_1 is F_2 . Because these two chains interact with each other through the interface I_1 , we propose that there is a putative interaction between functional classes F_1 and F_2 and an edge is put between these two functional classes. By applying this procedure to the whole dataset, the functional interaction network of the dataset is constructed.

Interactions taken from HPRD^{46; 47} and Jansen's gold standard⁴⁵ datasets are also converted to functional networks in a similar way.

Availability

The dataset and the clustering results are available at <http://prism.cccb.ku.edu.tr/interface>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number N01-CO-12400 and TUBITAK (Research Grant No 104T504). O.K. has been granted with Turkish Academy of Sciences Young Investigator Programme (TUBA-GEİP). N.T. has been supported by TUBITAK fellowship. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

REFERENCES

1. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of Protein-Protein Interactions: What are the Preferred Ways For Proteins To Interact? *Chem Rev* 2008;108:1225–1244. [PubMed: 18355092]
2. Chothia C, Janin J. Principles of protein-protein recognition. *Nature* 1975;256:705–708. [PubMed: 1153006]
3. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 1996;93:13–20. [PubMed: 8552589]
4. Tsai CJ, Nussinov R. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci* 1997;6:24–42. [PubMed: 9007974]
5. Nooren IM, Thornton JM. Diversity of protein-protein interactions. *Embo J* 2003;22:3486–3492. [PubMed: 12853464]
6. Ma B, Wolfson HJ, Nussinov R. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr Opin Struct Biol* 2001;11:364–369. [PubMed: 11406388]

7. Chothia C. Proteins One thousand families for the molecular biologist. *Nature* 1992;357:543–544. [PubMed: 1608464]
8. Aloy P, Russell RB. Ten thousand interactions for the molecular biologist. *Nat Biotechnol* 2004;22:1317–1321. [PubMed: 15470473]
9. Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci* 2004;13:1043–1055. [PubMed: 15044734]
10. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol* 1996;260:604–620. [PubMed: 8759323]
11. Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 2005;21:1901–1907. [PubMed: 15657096]
12. Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J. A protein domain interaction interface database: InterPare. *BMC Bioinformatics* 2005;6:207. [PubMed: 16122378]
13. Winter C, Henschel A, Kim WK, Schroeder M. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res* 2006;34:D310–D314. [PubMed: 16381874]
14. Teyra J, Doms A, Schroeder M, Pisabarro MT. SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics* 2006;7:104. [PubMed: 16512892]
15. Stein A, Russell RB, Aloy P. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res* 2005;33:D413–D417. [PubMed: 15608228]
16. Finn RD, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 2005;21:410–412. [PubMed: 15353450]
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242. [PubMed: 10592235]
18. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci* 1998;23:358–361. [PubMed: 9787643]
19. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol* 2001;313:399–416. [PubMed: 11800565]
20. Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* 2006;7:27. [PubMed: 16423290]
21. Keskin O, Ma B, Rogale K, Gunasekaran K, Nussinov R. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys Biol* 2005;2:S24–S35. [PubMed: 16204846]
22. Keskin O, Tuncbag N, Gursay A. Characterization and prediction of protein interfaces to infer protein-protein interaction networks. *Curr Pharm Biotechnol* 2008;9:67–76. [PubMed: 18393863]
23. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 2005;102:10930–10935. [PubMed: 16043700]
24. Nicola G, Vakser IA. A simple shape characteristic of protein-protein recognition. *Bioinformatics* 2007;23:789–792. [PubMed: 17267427]
25. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. PIER: protein interface recognition for structural proteomics. *Proteins* 2007;67:400–417. [PubMed: 17299750]
26. Aytuna AS, Gursay A, Keskin O. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 2005;21:2850–2855. [PubMed: 15855251]
27. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursay A. PRISM: protein interactions by structural matching. *Nucleic Acids Res* 2005;33:W331–W336. [PubMed: 15991339]
28. Pazos F, Ranea JA, Juan D, Sternberg MJ. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol* 2005;352:1002–1015. [PubMed: 16139301]
29. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 2003;19:163–164. [PubMed: 12499312]

30. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540. [PubMed: 7723011]
31. Kim WK, Henschel A, Winter C, Schroeder M. The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput Biol* 2006;2:e124. [PubMed: 17009862]
32. Shoemaker BA, Panchenko AR, Bryant SH. Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci* 2006;15:352–361. [PubMed: 16385001]
33. Argos P. An investigation of protein subunit and domain interfaces. *Protein Eng* 1988;2:101–113. [PubMed: 3244692]
34. Janin J, Chothia C. The structure of protein-protein recognition sites. *J Biol Chem* 1990;265:16027–16030. [PubMed: 2204619]
35. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol* 1988;204:155–164. [PubMed: 3216390]
36. Tsai CJ, Xu D, Nussinov R. Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci* 1997;6:1793–1805. [PubMed: 9300480]
37. Krissinel E. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* 2007;23:717–723. [PubMed: 17242029]
38. Day R, Beck DA, Armen RS, Daggett V. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* 2003;12:2150–2160. [PubMed: 14500873]
39. Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. *Proteins* 2004;56:143–156. [PubMed: 15162494]
40. Tamames J, Moya A, Valencia A. Modular organization in the reductive evolution of protein-protein interaction networks. *Genome Biol* 2007;8:R94. [PubMed: 17532860]
41. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* 2004;32:D262–D266. [PubMed: 14681408]
42. Brenner SE. A tour of structural genomics. *Nat Rev Genet* 2001;2:801–809. [PubMed: 11584296]
43. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science* 2006;311:347–351. [PubMed: 16424331]
44. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res* 2006;34:D302–D305. [PubMed: 16381872]
45. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;302:449–453. [PubMed: 14564010]
46. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A. Human protein reference database--2006 update. *Nucleic Acids Res* 2006;34:D411–D414. [PubMed: 16381900]
47. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjana V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003;13:2363–2371. [PubMed: 14525934]
48. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. *Science* 2003;300:1701–1703. [PubMed: 12805536]
49. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell* 2000;101:573–576. [PubMed: 10892642]

50. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 2001;313:903–919. [PubMed: 11697912]
51. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. *J Mol Biol* 2003;332:989–998. [PubMed: 14499603]
52. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res* 2008;36:D662–D666. [PubMed: 17959648]
53. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 1995;13:323–330. 307–308. [PubMed: 8603061]
54. Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* 2005;58:610–617. [PubMed: 15614759]
55. Keskin O, Nussinov R. Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng Des Sel* 2005;18:11–24. [PubMed: 15790576]
56. Keskin O, Nussinov R. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure* 2007;15:341–354. [PubMed: 17355869]
57. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A* 1991;88:10495–10499. [PubMed: 1961713]

Figure 1.

(A) The increase trend in the number of the PDB structures, and in the number of interfaces - overall, biological and crystal. Note that, some of the interfaces could not be annotated as crystal or biological, therefore the number of interfaces is higher than the sum of the biological and crystal interfaces. (B) Presentation of the increase ratio in protein families (available in SCOP), protein interface clusters and PDB entries. This figure shows the changes in proteins and their interfaces in a 12 year period. (C) Representation of the exponential increase in the number of the filtered biological dataset representatives and the annual increase in filtered biological interface clusters. We eliminate the interfaces belonging to membrane proteins and antibodies, interfaces where one side is a peptide or the number of residues is less than 100;

and call this dataset as the filtered set. Note that 2006 data was obtained in the beginning of the year, so does not represent the complete year. (D) The increase trend of inter-chain domain-domain interfaces for both all structures and for the filtered biological structures. The numbers are obtained by projecting the number of the deposited structures annually onto our interface clusters. (E) The annual increase of biological interface clusters and SCOP folds.

Figure 2.

Representation of the enlargement in the cluster sizes. Our current dataset is colored blue, and the 2002 dataset is colored maroon.

Figure 3.

Comparison of the 1o1pAD interface architecture with the 1enh (3 – helix bundle fold) single chain fold. In (A), the ribbon diagram of the interface named 1o1pAD is shown on the left. Cyan colored part is coming from the chain A, orange colored part is coming from the chain D. On the right, the ribbon diagram of the 3 – helix bundle fold (1enh) is shown. In (B), the matching parts between 1o1pAD and 1enh is shown (38 residues). The red colored parts are the matching parts of the interfaces and folds, the yellow colored parts are the unmatched ones.

Figure 4.

Comparison of the 1djrDE interface architecture with the 1Iris ($\alpha - \beta$ plait :1Iris) single chain fold. In (A), on the left the ribbon diagram of the interface named 1djrDE is shown colored according to the chains. Cyan colored part is coming from chain D, orange colored part is coming from chain E. On the right, the ribbon diagram of the $\alpha - \beta$ plait fold (1Iris) is shown. In (B), the matching parts between 1djrDE and 1Iris are shown (58 residues). The red colored parts are the matching parts of the interfaces and folds, the yellow colored parts are the unmatched parts of the structures.

Figure 5.

Two more examples for similarity between folding and binding architectures. Coloring scheme is as in Fig 3 and Fig 4. **(A)** Similarity between globin fold (1a6n) and the interface 1z6oAE.

(B) Similarity between immunoglobulin-like fold (1fna) and the interface 1ncwLH.

Figure 6.

Functional interaction network constructed from the filtered biological interfaces. The network is built as explained in the methods using the filtered biological interfaces. In the network, each node is a first-level GO function and each edge is an interaction between functions. Labels on the edges represent the number of unique architectures forming this functional interaction. Binding and catalytic activity functional classes are over-represented and are not shown in the network to make the figure visually clearer. Currently, 13 of 19 functional classes are available in PDB. Missing functions are chaperone regulator activity, chemo-attractant and repellent activities (IDs 5, 6 and 7), template for synthesis of G-rich strand of telomere DNA activity (ID 16), together with obsolete molecular function (ID: 13) and protein tag activity (ID: 14).

Figure 7.

The functional interaction network of the old and current datasets. The growing pattern of the networks is presented in the upper panel. The numbers next to the dataset year in the parentheses represent the number of nodes and edges in the respective datasets (excluding the catalytic and binding activities). Details of the notations are provided in Fig 6 and 12. The graph in the lower panel illustrates the annual increase in the number of functional classes (including catalytic activity and binding) and the interaction between them with open squares and filled diamonds, respectively. The data points corresponding to 2006 are not used in the line fitting.

Figure 8. Gene fusion examples in urease enzymes. Gamma (d.8.1.1) and beta (b.85.3.1) subunits of the urease from *Bacillus pasteurii* (1ubp A and B) and *Helicobacter pylori* (1e9y A) interact in similar ways and function in urease activity. These domains do not interact in *Klebsiella aerogenes* (1a5k A and B). The chains are in different colors for clarity.

Figure 9.

Gene fusion example in imidazole glycerophosphate synthase enzyme. The histidine biosynthesis (c.1.2.1) and class I glutamine amidotransferase (c.23.16.1) domains from *Thermotoga maritima* (1gpw A and B), *Thermus thermophilus* (1ka9 A and B) and *Saccharomyces cerevisiae* (1jvn A, 1ox4 A) are interacting in similar ways and functioning in imidazole glycerophosphate synthase enzyme activity.

Figure 10.
Schematic representation of the methodology to form the protein-protein interface dataset.

Figure 11.

Two chain protein interfaces of the urease (PDB ID: 1fwi). The contacting and nearby residues of the interfaces are shown as spheres. The chains A, B and C are colored yellow, red and cyan. Overall, two interfaces are formed between these three chains. The first interface is between the chains B and C. The second one is between the chains A and C.

Figure 12.

Preparation of the functional network of the interfaces. Each chain which contains the respective interface is annotated, using the GO annotations. The interaction network of these functional classes is formed from representative interface data set. In the scheme, letter I represents an interface, R and L represent the right and left chains of the interface. F represents the functional class of a chain. For example, for the first interface (I1), function 1 (F1) is related to function 2 (F2). Similarly, function 2 relates to function 3 (middle interface, I2), and function 1 relates to function 3 (I3). The resulting functional network is shown at the bottom of the figure where each node represents a function and each edge indicates functional interaction.

Table 1

Content of the current data set and old data sets

	In 2006	In 2002	In 1994
Structures (PDB)	34817	18687	2814
Interface containing PDB structures	15268	7243	747
Number of Interfaces	49512	21686	1629
Number of homo-dimers	31990	13082	766
Number of hetero-dimers	17522	8018	746
Number of Interfaces characterized by NOXclass	27755	11749	1022
Non-Biological Interfaces	10545	4172	340
Biological Interfaces	17210	7577	672
Obligate Interactions	14501	6333	522
Non-Obligate Interactions	2709	1244	150
Number of Clusters	8205	3799	351
Filtering non-globular proteins (considering only 17210 biological interfaces)			
Number of Interfaces	9873	4346	316
Number of Clusters	1901	1006	84

Table 2

Number of PDB structures, interfaces, interface clusters and protein families for each dataset.

Dataset	Number of PDB structures	Two-chain interfaces	Number of interface clusters	SCOP protein families
1994 ¹⁰	2814	1629	351	850 (in 1997)
2002 ⁹	18687	21686	3799	1827 (in 2002)
2006 (current work)	34817	49512	8205	2845 (in 2005)

Table 3

Comparison of the current protein interface datasets.

	SCOPPI	CBM	PIBASE	Current
Reference	Winter et al. 2005	Shoemaker et al. 2005	Davis et al. 2005	
Availability	http://www.scoppi.org/	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	http://alto.compbio.ucsf.edu/pibase/	http://www.ebi.ac.uk/interpro/
Interface Level	Domain level interfaces.	Domain level interfaces	Domain level interfaces	Chain level interfaces
Interface Extraction Method	Atomic Distance	Atomic distance	Inter-Intra Contact	Atomic distance
Query Type	Search by SCOP family, superfamily etc. PDB ID, keywords	Search by PDB ID, sequence, keyword	Atomic distance	Inter-Intra Contact
Redundancy Removal	PQS, the sequence identity and Δ ASA	PQS	Crystal Structures (PQS)	Crystal Structures (PQS)
Clustering	Structural clustering	Structural clustering	Clustering according to complex, binding site and interface topology	Clustering according to complex, binding site and interface topology
Interface Characterization	Δ ASA, interaction type, contact volume, conservation, domain description, etc...	Domain description, links to conserved domains.	Interface area (polar, nonpolar), cluster number, inter or intra domain interaction.	Interface area (polar, nonpolar), cluster number, inter or intra domain interaction.

Comparison of the most populated folds with the representative interfaces of the most populated clusters. "aa" is the number of amino acids. The interface pdb id nomenclature is first the pdb code, followed by the two chains. Thus, for the first interface entry, 1djr is the pdb code for the complex, and the interface is between chains D and E. The number of matched residues is taken from Multiprot's output.

Table 4

Fold pdb id	SCOP code	Fold size (aa)	Interface pdb id	Interface size (aa)	# of Matched residues	RMSD (Å)
Iris	d.58.14.1	97	1djrDE	119	58	1.92
Ienh	a.4.1.1	54	1o1pAD	156	38	1.86
Ia6n	a.1.1.2	151	1z6oAE	131	62	1.84
Ifna	b.1.2.1	91	1newLH	215	52	1.70
Ipgb	d.15.7.1	56	1nh0AB	130	31	2.00

Table 5

The parameters used during the clustering of the interfaces. The connectivity score takes into account the residue connectivity in the polypeptide chains. The score favors a match with consecutive residues. The percent amino acid identity is the count of identical residues in the match divided by the total number of matched pairs. The last column lists the difference in sizes between the interfaces within a cluster.

Cycle	Number of interfaces	Relative connectivity score	Minimal % amino acid identity	Maximal amino acid size difference between interfaces
A	49512 → 35744	0.9	90	0
B	35744 → 20921	0.9	80	3
C	20921 → 14132	0.8	50	10
D	14132 → 11297	0.7	25	20
E	11297 → 9533	0.6	10	40
F	9533 → 8205	0.5	0	50