Vol. 26 no. 20 2010, pages 2564–2570
doi:10.1093/bioinformatics/btq495

Advance Access publication August 27, 2010

Preferential use of protein domain pairs as interaction mediators: order and transitivity

Zohar Itzhaki, Eyal Akiva and Hanah Margalit*

Department of Microbiology and Molecular Genetics, The Institute for Medical Research – Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91120, Israel

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Many protein–protein interactions (PPIs) are mediated by protein domains. The structural data of multi-domain PPIs reveal the domain pair (or pairs) that mediate a PPI, and implicitly also the domain pairs that are not involved in the interaction. By analyzing such data, preference relations between domain pairs as interaction mediators may be revealed.

Results: Here, we analyze the differential use of domain pairs as mediators of stable interactions based on structurally solved multi-domain protein complexes. Our analysis revealed domain pairs that are preferentially used as interaction mediators and domain pairs that rarely or never mediate interaction, independent of the proteins' context. Between these extremes, there are domain pairs that mediate protein interaction in some protein contexts, while in other contexts different domain pairs predominate over them. By describing the preference relations between domain pairs as a network, we uncovered partial order and transitivity in these relations, which we further exploited for predicting interaction-mediating domains. The preferred domain pairs and the ones over which they predominate differ in several properties, but these differences cannot yet determine explicitly what underlies the differential use of domain pairs as interaction mediators. One property that stood up was the over-abundance of homotypic interactions among the preferred domain pairs, supporting previous suggestions on the advantages in the use of domain self-interaction for mediating protein interactions. Finally, we show a possible association between the preferred domain pairs and the function of the complex where they reside.

Contact: hanahm@ekmd.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 7, 2010; revised on August 23, 2010; accepted on August 25, 2010

1 INTRODUCTION

Domains are considered as the fundamental building blocks determining the structure and function of proteins. Many domains were specialized to mediate the interaction of proteins with other molecules. For example, there are DNA-binding domains, RNA-binding domains and domains that mediate interactions between proteins. Protein–protein interaction (PPI) is usually achieved either via interactions between domains and short motifs (Pawson *et al*., 2002), often in transient interactions, or by domain–domain interactions (DDIs) (Pawson and Nash, 2003), often in stable interactions. Analyses of structurally solved PPIs determined the domain pairs that are used as interaction mediators (Finn *et al*., 2005; Stein *et al*., 2005, 2009). Further analysis of these DDIs has revealed distinct domain pairs that are repeatedly used as interaction mediators in different protein contexts. Moreover, it was shown that a statistically significant fraction of the interactome of various organisms can be attributed to these interacting domain pairs and that they are evolutionarily conserved (Itzhaki *et al*., 2006; Schuster-Bockler and Bateman, 2007).

It is conceivable that protein domains that mediate interactions with other molecules were optimized through evolution to perform this function. Therefore, while the domain repertoire of a proteome defines thousands of theoretically possible interacting domain pairs, only a fraction of these pairs seem to actually mediate protein interactions. Furthermore, certain domain pairs mediate interactions in some protein contexts but not in others, in which there are other domain pairs opted for this task (as exemplified in Fig. 1). The data of multi-domain PPIs solved by crystallography reveal many examples of such variability in the utilization of domain pairs as interaction mediators. For example, the self-interaction of the domain Hpt (Pfam accession PF01627, Finn *et al*., 2008) is used for mediating the interaction in all solved complexes containing it, independent of other domains. On the other hand, the GHMP kinase N-terminal domain (PF00288) mediates dimerization by self-interaction in some protein contexts (e.g. Mevalonate kinase), but not in others (e.g. Galactokinase), where other domains mediate the dimerization. Notably, the structural data of multi-domain PPIs reveal the domain pair (or pairs) that mediate a PPI, and implicitly also the domain pairs that are not involved in the interaction, enabling a systematic analysis of such preference relations. Here we carry out such an analysis, using reliable DDI and PPI data from solved structures of multi-domain protein complexes. We report on a partial order in the preferential use of domain pairs as interaction mediators, compare properties of the domains that may underlie this order and show by a few examples possible functional implications of our findings. We also demonstrate the utilization of this order for predicting the domains mediating experimentally determined PPIs.

2 METHODS

2.1 Compilation of PPI and DDI data

We retrieved the DDI data from the 3DID database (March 2008 version, Stein *et al*., 2009). Only domains involved in PPIs were included, and of those

*To whom correspondence should be addressed.

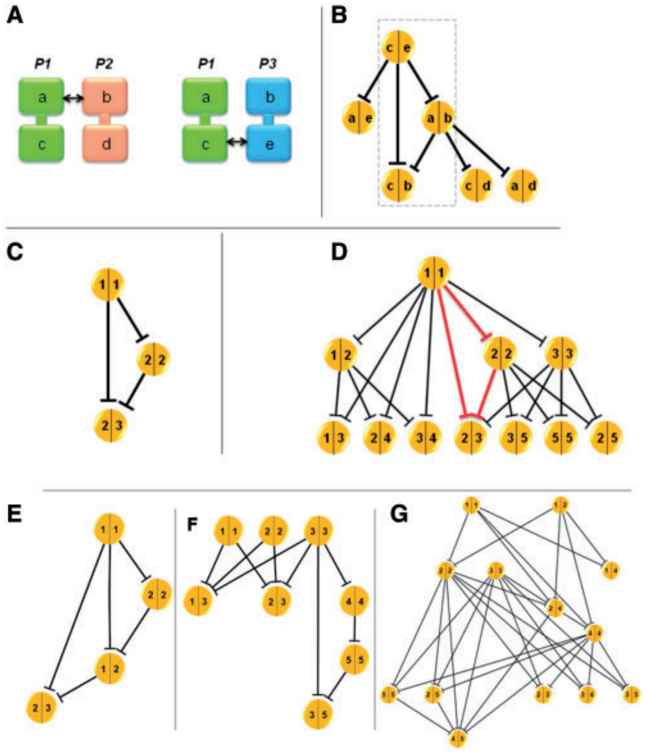

Fig. 1. Domain pair relations. (**A**) An example of context-dependent DDI. Protein P1 contains domains a and c, and protein P2 contains domains b and d. Domain a in P1 and domain b in P2 mediate the interaction between P1 and P2. In principle, there are four combinations of domain pairs that can mediate the interaction: (a,b), (a,d), (c,b) and (c,d), but actually three combinations do not take place and only (a,b) is responsible for the interaction. However (a,b) might not mediate the interaction in a different context, for example, when P1 interacts with another protein, P3, which contains the domains b and e. For the P1–P3 interaction, the possible domain pairs are (a,b), (a,e), (c,b) and (c,e), but the interaction is mediated only by (c,e). Thus, in the context of the domains of P1 and P2, (a,b) mediates the interaction, but in the context of the domains of P1 and P3 (a,b) does not mediate the interaction and another domain pair, (c,e), does. (**B**) Description of the relations between the domain pairs described in (A) as a network, where nodes represent the domain pairs and edges point from the domain pairs mediating an interaction to the domain pairs over which they predominate. The domain pair (c,e) is preferred over the domain pairs (a,b), (a,e) and (c,b), as derived from P1–P3 interaction. The domain pair (a,b) is preferred over the domain pairs (c,b), (a,d) and (c,d), as derived from P1–P2 interaction. Hence, the domain pair (c,e) predominates over the domain pair (c,b) both directly and indirectly. Such a pattern in the domain pair relation network implies transitivity and would suggest preferential order of the domain pairs when used as interaction mediators. (**C**) A transitive pattern in the actual domain pair relation network. For clarity, the domains in the network are numbered: 1. 3’ exoribonuclease family, domain 1 (PF01138); 2. OB fold (CL0021); 3. K-Homology (KH) domain Superfamily (CL0007). (**D**) A sub-network containing a transitive pattern, as described in (C) (highlighted in red). The additional domain numbers represent: 4. 3’ exoribonuclease family, domain 2 (PF03725); 5. NusA N-terminal domain (PF08529). (**E**-**G**) Other examples of sub-networks based on the actual data demonstrating a partial order, in which the preference relations between domain pairs are consistent and transitive: (**E**) 1. Tim barrel glycosyl hydrolase superfamily (CL0058); 2. Galactose-binding domain-like superfamily (CL0202); 3. Glycosyl hydrolases family 2, immunoglobulin-like beta-sandwich domain (PF00703). (**F**) 1. Receptor L domain (PF01030); 2. Furin-like cysteine rich region (PF00757); 3. Protein kinase superfamily (CL0016); 4. Ig-like fold superfamily (E-set) (CL0159); 5. Immunoglobulin superfamily (CL0011). (**G**) 1. Globin-like (CL0090); 2. Ferredoxin/Ferric reductase-like NAD binding (CL0091); 3. Nitric oxide synthase, oxygenase domain (PF02898); 4. Riboflavin synthase/Ferredoxin reductase FAD binding domain (CL0076); 5. Flavoprotein (CL0042).

Table 1. Network and motif analyses of the various datasets

The dataset	No. of PPIs	No. of DDIs	Network type	Network edges representing DD relations (nodes)	No. of transitive motifs	No. of cyclic motifs
Basic[a]	1206	1339	All data	6895 (3608)	4310	27
Ext[b]	3267	1953	All data	19 797 (11 021)	7301	26
			Stringent[c]	2916 (2461)	107	0
SCOP[d]	366	478	All data	1183 (1060)	104	5

[a] Basic—the dataset used in the main analysis, of multi-domain protein complexes with all domains solved by crystallography.

[b] Ext (extended)—the extended dataset is based on solved PPIs with at least one multi-domain protein, as in the main analysis, relieving the requirement that all domains in addition to the ones mediating the interaction were solved by crystallography.

[c] The stringent network was generated by keeping relations that appeared at least twice in the network based on the extended dataset.

[d] The SCOP-based network of domain pair relations based on the SCOP domain definitions (Stein *et al.*, 2009), rather than Pfam definitions.

we kept DDIs with 3DID Z-score >0.3 and with at least six residue–residue contacts, which we consider more reliable. Notably, 98% of the interactions had Z-score >1 [for the Z-score description, see Supplementary Material (Stein *et al.*, 2009)]. In order to use the correct biological unit, we used also data from SNAPPI-DB (Jefferson *et al.*, 2007), which is based on the Protein Quaternary Structure database (Henrick and Thornton, 1998) and retrieved the SNAPPI-DB’s DDIs. We combined the information from SNAPPI-DB and 3DID to generate the database of PPIs and DDIs used in the analysis. In order to avoid bias in the analysis due to overrepresentation of evolutionarily related structures or of identical and similar structures reported by different groups, we compared the protein sequences in our database using BLAST (Altschul *et al.*, 1990). When sequences were found by BLAST to be similar with E-value $\leq 10^{-3}$, only one representative of them was kept in the data. In addition, viral proteins were excluded from the database. We labelled the proteins by their Pfam (Finn *et al.*, 2008) domains, or, when relevant, by domain clans, where a clan contains several similar domains clustered by the Pfam database. Throughout the article, domain and clan names contain the prefix PF and CL, respectively. Finally, for each PPI we extracted all possible domain–domain, clan–clan or clan–domain combinations. The basic units in our analysis are these pairs and we term them throughout the article domain pairs, even though some of them consist of clans.

2.2 Analysis of connected patterns in the network of domain pair relations

We used the data of all domain pairs and the information on their use as interaction mediators, and described these relations between domain pairs as a network, in which every node is a domain pair and edges point from domain pairs that are used to mediate an interaction to the domain pairs over which they predominate. In the main analysis, we did not distinguish between edges based on their weight, and all edges were included. In a more stringent analysis that followed, we included only edges which appeared at least twice in our data (Table 1). When a protein interaction was mediated by two or more domain pairs, no relations were defined between them. In cases where one domain pair predominates over a second domain pair in the context of one PPI, and the second domain pair predominates over the first one in another PPI, a bi-directional edge was assigned between these nodes.

Table 2. Comparison of properties between preferred and non-preferred domain pairs (basic dataset)[a]

The property	Median difference between preferred and non-preferred domain pairs[a] (standard error)	Total number of edges included in the analysis[b]	P-value (Wilcoxon paired test)
Length (No. of amino acids)	18.25 (0.99)	8783	4.8×10^{-175}
Phylogenetic age	0 (0.0076)	8279	1.22×10^{-30}
Number of contacts	3.28 (1.02)	969	5.57×10^{-09}
Estimated binding free energy (kcal/mol)	−0.7 (0.66)	713	2.23×10^{-04}
Interface area (Å2)	662.3 (281.9)	957	1.20×10^{-03}
Plasticity (average RMSD)	0.043 (0.01)	2201	1.39×10^{-04}
Sequence conservation	0.05 (0.0001)	7141	2.46×10^{-12}

[a] For each property we carried out a paired Wilcoxon test, analyzing the differences in the property value between all domain pair nodes connected by an edge in the network of domain pair relations. For description of the properties studied and the computation of their values, see Supplementary Figure 4. When a preference relation (edge in the network) appeared in more than one PPI instance, we computed the average value of the property difference over all instances. When the relations involved a clan, we regarded the actual domains leading to the preference relations. In these cases, we computed the averages of the differences between the properties of the corresponding domain pairs.

[b] The numbers of compared couples of domain pairs depend on the property computed. For example, for the length and phylogenetic age computations we used the actual domain pairs included in the clans, and therefore the number is higher than the total number of edges. On the other hand, for some properties, such as number of contacts, values could be obtained only for domain pairs with solved structures.

To find the connected patterns we used the Cytoscape NetMatch plugin (Ferro *et al.*, 2007; Shannon *et al.*, 2003). We also created random networks that maintain the number of nodes and the degree of each node as in the actual network (Shen-Orr *et al.*, 2002; Yeger-Lotem *et al.*, 2004), and compared the numbers of connected patterns in the actual network to those in the random networks.

2.3 Comparison of properties between preferred and non-preferred domain pairs

For each property we carried out a paired Wilcoxon test, analyzing the differences in the property value between all domain pair nodes connected by an edge in the network. For properties that are computed per domain (such as length) the value of a domain pair was computed as the average of values of the individual domains comprising it. When a preference relation (edge in the network) appeared in more than one PPI instance, we computed the average value of the property difference over all instances. When the relations involved a clan, we regarded the actual domains leading to the preference relations. In these cases, we computed the averages of the differences between the properties of the corresponding domain pairs (Table 2).

2.4 Prediction of DDIs using the order of domain pairs

We used the order in the network of domain pair relations to predict which domain pair is most likely to mediate a given PPI. Given a pair of interacting proteins, we determined all possible domain pair combinations and checked the network for direct or indirect paths from each domain pair to all other pairs. The domain pair(s) with paths to most of the other domain pairs was (were) predicted as mediating the interaction. We confirmed this approach by 5-fold cross-validation: we generated a network using 4/5 of the PPIs in our data and tested the predictions on the rest 1/5. This process was repeated five times. We also carried out 3-fold cross-validation. We checked the statistical significance of the results by performing 10 000 iterations assigning randomly the interacting domain pair for each PPI and checking whether the assigned domain pair indeed mediates the interaction. The statistical significance was defined as the fraction of random iterations in which the rate of successfully assigned domain pairs was equal to or exceeded the rate of successful assignments of domain pairs according to the network path analysis.

3 RESULTS

3.1 PPI and DDI data

Our analysis is based on the 3DID (Stein *et al.*, 2009) and SNAPPI-DB (Jefferson *et al.*, 2007) databases of interacting domain pairs from a variety of organisms (from bacteria to human), derived from stable protein complexes solved by crystallography (Berman *et al.*, 2000). Uniting and filtering these databases resulted in 9633 PPIs mediated by 3977 DDIs. Removal of homologous PPIs and clustering of similar domains following Pfam definitions of domain clans further reduced the data to 5024 PPIs mediated by 3005 DDIs (see Section 2). The distribution of domain pairs by the number of PPIs they mediate followed a power law ($y=3184x^{-2.1}$): many of the domain pairs mediated only one PPI, while only a few domain pairs mediated many PPIs (Supplementary Fig. 1). This observation suggests that domain pairs differ in their use as interaction mediators.

3.2 Preference-relations between domain pairs

To study the relative use of domain pairs as interaction mediators and their inter-relations, we focused on PPIs that involve at least one multi-domain protein. We included only multi-domain proteins for which all domains were solved by crystallography. These requirements reduced substantially our data to 1206 PPIs. The domain content of the interacting proteins defines all the possible domain pairs that could theoretically mediate the interaction. The structure of the protein complex determines the actual domain pair(s) that mediate the interaction and implicitly also the domain pairs that are not involved in the interaction. We used this information to define preference relations between domain pairs: the domain pair(s) that actually mediate(s) the interaction 'predominate(s)' over domain pairs that do not mediate the interaction. To systematically analyze these relations, we described them as a network, in which every node represents a domain pair and directed edges point from domain pairs that are used to mediate an interaction to the domain pairs over which they predominate, as exemplified in Figure 1B. Proteins P1, P2 and P3 in the example presented in Figure 1A define a network of six nodes: (a,b), (a,d), (c,b), (c,d), (a,e) and (c,e). The interaction between P1 and P2 is mediated by domains a and b, and the interaction between P1 and P3 is mediated by domains c and e. Therefore, edges point from the node (a,b) to nodes (a,d), (c,b), (c,d) and from the node (c,e) to nodes (a,b) ,(a,e), (c,b) (Fig. 1B). Examination of this exemplary network reveals a transitive pattern in which (c,e) predominates over (c,b) both directly and indirectly, through (a,b). Thus, in this example the preference relations of domain pairs as interaction mediators are consistent, implying an order among domain pairs. Our goal is to test whether we can identify such consistent relations between domain pairs

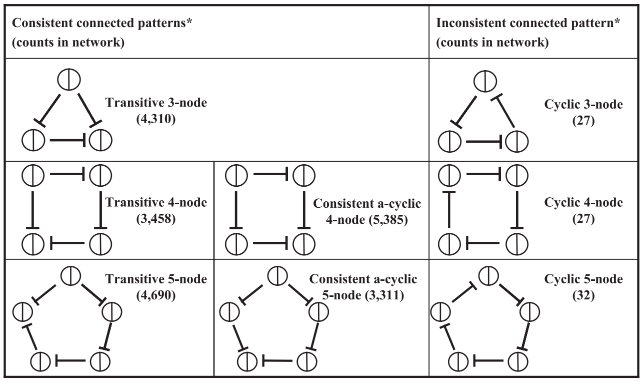

Fig. 2. Types of connected patterns and their counts in the network of domain pair relations. *Nodes represent domain pairs and edges point from domain pairs to domain pairs over which they predominate. Patterns are represented with uni-directional edges but their counts include all occurrences of the pattern, derived from corresponding patterns with uni- and bi-directional edges (see Section 2 and Supplementary Table 1 for details).

in the actual network. This could be revealed by searching the network of domain pair relations for transitive and cyclic connected patterns. Transitive patterns are consistent with an order among domain pairs, while cyclic patterns imply conflicts in the preference relations. High abundance of transitive patterns and avoidance of cyclic patterns would support an order between domain pairs as interaction mediators.

We defined the network of domain pair relations based on the data in the final set of 1206 PPIs. This network included 3608 nodes (domain pairs) and 6895 edges (pointing from domain pairs to domain pairs over which they predominate, see Supplementary Material). Of note, 393 of the 2545 domain pairs that were predominated by other domain pairs were shown to mediate interactions in other solved structures. 142 of the edges were bi-directional. Systematic analysis of connected patterns in this network showed that transitive and consistent acyclic patterns exceeded the cyclic patterns by two orders of magnitude (Fig. 1C–G, Fig. 2 and Supplementary Table 1). The ratio of transitive to cyclic patterns was substantially higher in the actual network (159.6) than in random networks (27.9, see Section 2). We verified that the over-abundance of the transitive patterns did not result simply from the node degrees (i.e. the transitivity was observed because the highest degree node predominated over the other nodes). We tested this hypothesis and found that the degrees are consistent with only less than a third of the transitive patterns and therefore cannot account for the discovered phenomenon.

We repeated the analysis of connected patterns of domain pairs using three additional networks of domain pair relations, as summarized in Table 1. Analysis of each of these networks revealed an order in the use of domain pairs as interaction mediators, and a high abundance of transitive compared to cyclic patterns. This emphasizes the robustness of our findings regardless of the dataset and the domain definitions that are used.

3.3 Comparison between preferred and non-preferred domain pairs

It is intriguing to find out what determines the preference relations between domain pairs and why certain domain pairs are used for mediating interaction in one context but other domain pairs predominate over them in another context. One possible explanation would be that domains capable of interaction, each residing in one interacting partner, are not engaged in certain PPIs because they are involved in intra-protein interactions. Therefore, other domain pairs predominate over them. However, careful examination of the 3891 PDB structures, that are involved in the complexes on which our analysis was based, revealed that this explanation could hold for only 98 structures. Another possibility is that the preference relations stem from certain properties of the domain pairs. To this end, we carried out a pairwise comparison of certain features between the preferred domain pairs and the domain pairs over which they predominate (in every PPI context). We analyzed seven properties (listed in Table 2), including properties of the domains comprising the pair, such as their length or phylogenetic age, and properties related to the interaction, such as the number of contacts in the interface. For a given property, we computed for each domain pair in the network (a node) the property value (see Section 2). Next, for all couples of domain pairs connected by an edge, we tested by a paired Wilcoxon test whether the differences between their property values were statistically significant. Notably, a domain pair in one context may be preferred while in another context another domain pair can predominate over it. Comparison of the various properties between the preferred domain pairs and the ones over which they predominate showed highly statistically significant differences (Table 2). However, the biological meaning of these differences is less clear, as for some properties (such as phylogenetic age, evolutionary conservation and plasticity) the differences were very modest, and for others (such as the estimated binding free energy) the average difference was in the range of the variation among various PPIs mediated by the same domain pair.

Remarkably, preferred domain pairs exhibit a high tendency towards homotypic interactions (self-interactions of domains), when such interactions are possible. Out of the 1206 multi-domain PPIs in our data, 680 are of homodimers. A homodimer made of a protein of at least two different domains can dimerize through interactions that involve the same domain in the interacting proteins (homotypic interaction) or through interactions that involve different domains (heterotypic interaction). We found that 630 of the 680 homodimers (93%) dimerize through at least one homotypic interaction, high above random expectation (which is 44% based on the domain content of the homodimers). Furthermore, there are 79 heterodimers that have the potential of homotypic interactions (the proteins share at least one common domain), and 50 of those (63%) demonstrate such interactions, high above random expectation (28%).

3.4 Functional implications of domain pair preferences

The preference of certain domain pairs over others as interaction mediators is consistent also with the functionality of the complex. For instance, there are examples in which the functional activity is embedded in the interface of the preferred domains. This is exemplified by the prokaryotic acetaldehyde dehydrogenase dimerization domain (PF09290) and the HMGL-like domain (PF00682, clan CL0152), present in the DmpG and DmpF proteins from *Pseudomonas sp.* Each domain of the preferred pair includes a separate active site. Upon their interaction, a thin tunnel that originates in one domain, and ends in the other, allows the transfer of substrates from one active site to the other (Manjasetty *et al.*, 2003,

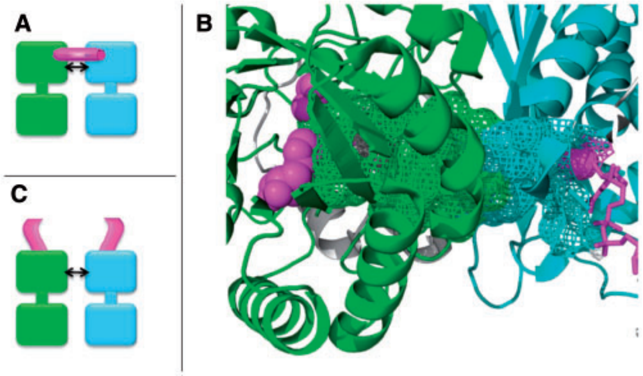

Fig. 3. The preference of certain domain pairs over others has functional implications. (**A**) Schematic illustration of the interaction between two double-domain proteins, protein A (green) and protein B (cyan), where the top domain pair is preferred and mediates the interaction. Here, the interaction mediated by the preferred domain pair yields a passage between the domains, colored in magenta, which is related to the functionality of the complex (e.g. channeling an electron or tunneling a substrate from a domain in protein A to a domain in protein B). (**B**) An actual example for the scenario depicted in (A): The interaction between DmpG and DmpF (PDB code: 1NVM) is mediated by an HMGL-like domain (green, chain B) and the dimerization domain of prokaryotic acetaldehyde dehydrogenase (cyan, chain A). The two proteins include separate active sites (magenta), and a tunnel through which substrates transverse from one to another (residues that compose this tunnel are represented as a mesh). The picture was created using PyMOL http://pymol.sourceforge.net/. The identities of the residues that form the tunnel were calculated using the CASTp program (Dundas *et al.*, 2006). (**C**) The homotypic interaction of the AsnC domain (the preferred domain pair) mediates the protein interaction and also positions two helix-turn-helix motifs in the right distance for binding the DNA site.

Fig. 3A and B). The functional contribution of substrate channeling to the complex may explain the preference of this particular domain pair over other alternative domain pairs in mediating the interaction. This principle is also exemplified in the FAD binding (PF00890, clan CL0063) and Fer2 (PF00111) domain pair that mediates the interaction between the FrdA and FrdB proteins. These two proteins serve as the soluble component of the quinol-fumarate reductase respiratory complex in *Escherichia coli*. The catalytic activity of this complex requires the channeling of electrons from the Fer2 domain to the FAD-binding domain (Iverson *et al.*, 2002).

A similar scenario is exhibited in the APS reductase complex of *Archaeoglobus fulgidus* (Fritz *et al.*, 2002), which is another instance of this domain pair. Collectively, these examples show that the preference of a certain domain pair in mediating the interaction in different protein complexes can be explained by the coupling between this interaction and the function of the protein complex, as these functions cannot be achieved by any other domain pair combinations. In other examples, the functional reasoning for the domain pair preference does not reside in the interaction interface, but rather in the correct positioning of complex components. Such a scenario is found, for instance, in LrpA from *Pyrococcus furiosus*, which dimerizes through the interaction of two AsnC domains (PF01037, clan CL0032). This protein includes also a helix-turn-helix motif that interacts with the DNA (Leonard *et al.*, 2001). The interacting domains place these DNA recognition motifs in a precise distance from each other, which is required for accurate recognition of the DNA binding site (Fig. 3C). No other domain combination can place the DNA recognition motifs in the precise distance and orientation, and hence this homotypic domain pair interaction is preferred over all other possible domain pairs in this complex. These functional implications are consistent also in other complexes where these domain pairs are found and preferred. Thus, the association between the preferred domain pairs and functional traits provides yet another explanation for their repeated use as interaction mediators, since their interaction aids in defining and positioning the functional modules of different complexes.

3.5 Prediction of DDIs using the network of domain pair relations

The analysis of the domain pair relation network revealed transitivity, implying that the domain pairs present a consistent order. This suggests that if there are data indicating that (c,e) predominates over (a,b) and (a,b) predominates over (b,c), it is most probable that (c,e) predominates also over (b,c), even if there is no such indication in the current network. This finding has potential implications for prediction of the domains that actually mediate an experimentally verified PPI. While prediction of interacting domains was not the aim of this work, we briefly demonstrate how the network of domain pair relations can be exploited for such predictions. Given a pair of interacting proteins, we can determine all possible domain pairs and use the domain pair relation network to check whether there are direct or indirect paths from one domain pair to all other possible domain pairs. Such a pair will be predicted as the one mediating the interaction. We confirmed this approach by analyzing the data of 1206 PPIs using cross-validation. When determining the network of domain pair relations based on 4/5 of the protein interactions in our data and testing the predictions on the remaining 1/5, we could show that by tracing the paths between domain pairs in this network, we can correctly predict the domain pairs mediating the interactions for 73% of the PPIs. This regards PPIs in the test set with domains that are found in the network based on the training set. This result is highly statistically significant ($P < 0.0001$, see Section 2). Similar results were obtained when the analysis was carried out with 3-fold cross-validation. We also examined the accuracy of the prediction in comparison to random success when the PPIs were grouped by the numbers of competing domain pairs: We correctly predicted 85, 71 and 90% of the DDIs when two, three, or four domain pair combinations, respectively, were possible. This success rate is high above random expectation (Fig. 4 and Supplementary Table 2). Similar success rates were obtained for PPIs with more competing domain pairs.

4 DISCUSSION

The recent accumulation of a vast amount of PPI data has enabled the study of PPI networks [for review see (Cho *et al.*, 2004; Kiel *et al.*, 2008; Levy and Pereira-Leal, 2008; Sharan and Ideker, 2006)], as well as of the inter-relations between PPIs and DDIs (e.g. Chen and Liu, 2005; Deng *et al.*, 2002; Jothi *et al.*, 2006; Liu *et al.*, 2005; Martin *et al.*, 2005; Qi *et al.*, 2006; Riley *et al.*, 2005; Schlicker *et al.*, 2007; Sprinzak and Margalit, 2001; Sprinzak *et al.*, 2006), where we employed network analysis tools to study the preference relations between domain pairs as interaction mediators. In this network, the

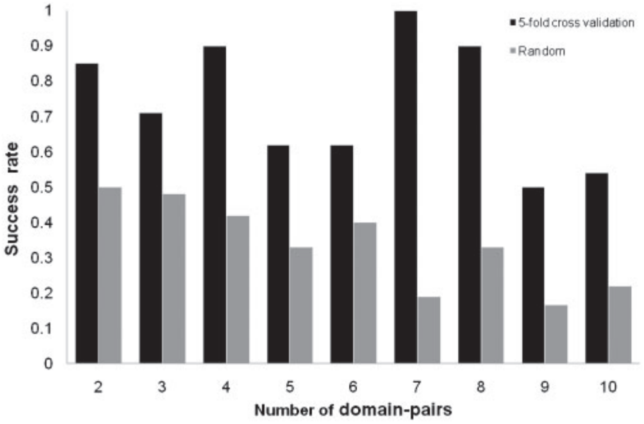

Fig. 4. Prediction of interacting domains. The order in the network of domain pair relations was used to predict which domain pair mediates a given PPI (see text). Presented are rates of success in DDI prediction according to the number of possible domain–domain combinations in interacting protein pairs, in comparison to the expected random success rate (e.g. the random success rate in case of two domain pairs is 50%). The shown data are based on interacting protein pairs with at most 10 possible domain pairs. We confirmed this approach by 3-fold and 5-fold cross-validation (shown here). When a domain pair in the test set does not occur in the network of domain pair relations (training set) the interaction cannot be predicted. Success rates are therefore described for those domain pairs that appear in both the test and training sets. The random success rates were calculated according to the number of potential domain pairs in PPIs, taking into consideration the number of domain pairs that actually mediate the interaction.

nodes are domain pairs and edges point from domain pairs mediating the interaction to all possible domain pairs in the complex, over which they predominate. Analysis of this network revealed an order between domain pairs, which was consistent across various protein pairs carrying the corresponding domains. Of note, the relatively small number of solved complexes implies that the network of domain pair relations is of low connectivity, since many couples of domain pairs may never appear in the same complex, disenabling an edge between them. Accordingly, this network is composed of many subnetworks of connected components, each showing a partial view of the domain pair relations. Thus, the phenomenon that we have identified, of order of the domain pairs by their preferential use as interaction mediators, is re-discovered in many subnetworks of domain pair relation networks. However, in the context of the whole network, due to the missing edges, this order should be regarded as partial order. Thus, our analysis shows that not only there are domain pairs that mediate PPIs and others that do not, but also the ones capable of mediating interaction can be (partially) ordered by their suitability to this task.

The main discovery of our study, that there is an order of domain pairs as interaction mediators, raises additional questions as to the domain pair properties that may underlie the revealed preferences. Conceivably such preferences might develop through evolution, and therefore we asked whether the preferred pairs are more ancient than the ones over which they predominate and whether they show higher plasticity that allows them to accommodate in various protein contexts. These properties showed very modest differences between the compared domain pairs (although highly statistically significant), pointing to tendencies that might be further substantiated with accumulation of more data of interacting proteins. Alternatively, as we found in our analysis of functional implications, it may be that the preference relations are the result of local solutions to the functional constraints of each protein complex. Since these solutions differ between different complexes, specific properties might have implications in only subsets of the data, and therefore we observed only subtle tendencies when applying the analysis of the properties to the whole data.

Our prominent finding—that the preferred domain pairs show high abundance of homotypic interactions—is consistent with previous observations reporting that homotypic interactions are statistically significantly over-represented in the data of DDIs, especially among DDIs conserved in various organisms (Itzhaki *et al.*, 2006). Furthermore, our finding that, when possible, homotypic interactions are preferred also for heterodimerization further supports the advantage suggested for homotypic interactions in stabilizing protein complexes (Andre *et al.*, 2008; Lukatsky *et al.*, 2006). These advantages include a duplicated effect for stabilizing mutations and lower mean energy of the interfaces compared to that of heterotypic interactions (Lukatsky *et al.*, 2006). Another hint at the advantage of homotypic interactions is the relative abundance of functional sites found at the interface of homodimers (Davis and Sali, 2010). In addition, it was demonstrated that early in evolution a bias towards very low energy complexes may have driven the selection of symmetrical structures (Andre *et al.*, 2008). It is possible that following such considerations early in evolution certain domains adjusted themselves for self interaction. Such domains could later recruit other domains to create multi-domain proteins whose interactions are mediated via the homotypic interactions (Bornberg-Bauer *et al.*, 2005). This conjecture is further supported by our finding that the same homotypic domain interactions mediating multi-domain PPIs are also found in homodimers of single domain proteins (data not shown).

The high abundance of homotypic interactions among the preferred domain pairs may raise the concern that the order that we have identified is simply a reflection of the advantage of homotypic interactions. To test this, we repeated the analysis including only heterodimers that do not contain domains capable of self-interaction (data not shown). This analysis again identified a consistent order between domain pairs, implying that it is not due to the preference of homotypic interactions.

Our analysis, while based on the currently available limited dataset of structurally derived DDIs in multi-domain PPIs, provides intriguing insights into the preference relations between domain pairs. The order that we identified is robust and is revealed when using datasets based on various systems of domain definitions (Finn *et al.*, 2008; Murzin *et al.*, 1995), and when analyzing extensions or sub-sections of the datasets (Table 1), suggesting that it will persist when more DDI data become available. At present, the order that we identified is partial because many parts of the network of domain pair relations are disconnected due to missing data. As more and more protein complexes of multiple domain combinations will be solved, the repertoire of domain pair relations will be enriched and the preference relation network might become more connected. This will enable improved prediction of DDIs based on the network structure and a deeper understanding as to the molecular principles that make some domain pairs more suitable than others for mediating interactions.

ACKNOWLEDGEMENTS

We thank Sarah Teichmann, Yael Altuvia, Ariel Jaimovich, Dan Reshef and Naama Wald for their helpful comments on the article. We thank Nir London, Barak Raveh, Shmuel Sattath and Ora Schueler-Furman for their useful comments on the manuscript and their help in the analysis. We are grateful to Jim Procter for his very valuable help with the SNAPPI-DB.

Funding: European Union grant 3D-repertoire (contract number LSHG-CT-2005-512028). Israel Science Foundation administered by the Israel Academy of Sciences and Humanities.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215**, 403–410.

Andre,I. *et al.* (2008) Emergence of symmetry in homooligomeric biological assemblies. *Proc. Natl Acad. Sci. USA,* **105**, 16148–16152.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bornberg-Bauer,E. *et al.* (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol. Life Sci.*, **62**, 435–445.

Chen,X.W. and Liu,M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.

Cho,S. *et al.* (2004) Protein-protein interaction networks: from interactions to networks. *J. Biochem. Mol. Biol.*, **37**, 45–52.

Davis,F.P. and Sali,A. (2010) The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput. Biol.*, **6**, e1000668.

Deng,M. *et al.* (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**, 1540–1548.

Dundas,J. *et al.* (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, **34**, W116–W118.

Ferro,A. *et al.* (2007) NetMatch: a Cytoscape plugin for searching biological networks. *Bioinformatics*, **23**, 910–912.

Finn,R.D. *et al.* (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Finn,R.D. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res*., **36**, D281–D288.

Fritz,G. *et al.* (2002) Structure of adenylylsulfate reductase from the hyperthermophilic Archaeoglobus fulgidus at 1.6-A resolution. *Proc. Natl Acad. Sci. USA*, **99**, 1836–1841.

Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.,* **23**, 358–361.

Itzhaki,Z. *et al.* (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol.*, **7**, R125.

Iverson,T.M. *et al.* (2002) Crystallographic studies of the Escherichia coli quinol-fumarate reductase with inhibitors bound to the quinol-binding site. *J. Biol. Chem.*, **277**, 16124–16130.

Jefferson,E.R. *et al.* (2007) SNAPPI-DB: a database and API of structures, iNterfaces and alignments for protein-protein interactions. *Nucleic Acids Res.*, **35**, D580–D589.

Jothi,R. *et al.* (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J. Mol. Biol.*, **362**, 861–875.

Kiel,C. *et al.* (2008) Analyzing protein interaction networks using structural information. *Annu. Rev. Biochem.*, **77**, 415–441.

Leonard,P.M. *et al.* (2001) Crystal structure of the Lrp-like transcriptional regulator from the archaeon Pyrococcus furiosus. *EMBO J*., **20**, 990–997.

Levy,E.D. and Pereira-Leal,J.B. (2008) Evolution and dynamics of protein interactions and networks. *Curr. Opin. Struct. Biol.*, **18**, 349–357.

Liu,Y. *et al.* (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, **21**, 3279–3285.

Lukatsky,D.B. *et al.* (2006) Statistically enhanced self-attraction of random patterns. *Phys. Rev. Lett.*, **97**, 178101.

Manjasetty,B.A. *et al.* (2003) Crystal structure of a bifunctional aldolase-dehydrogenase: sequestering a reactive and volatile intermediate. *Proc. Natl Acad. Sci. USA*, **100**, 6992–6997.

Martin,S. *et al.* (2005) Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218–226.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Pawson,T. *et al.* (2002) Interaction domains: from simple binding events to complex cellular behavior. *FEBS Lett*., **513**, 2–10.

Qi,Y. *et al.* (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**, 490–500.

Riley,R. *et al.* (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**, R89.

Schlicker,A. et al. (2007) Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, **23**, 859–865.

Schuster-Bockler,B. and Bateman,A. (2007) Reuse of structural domain-domain interactions in protein networks. *BMC Bioinformatics*, **8**, 259.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.*, **24**, 427–433.

Shen-Orr,S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, **31**, 64–68.

Sprinzak,E. *et al.* (2006) Characterization and prediction of protein-protein interactions within and between complexes. *Proc. Natl Acad. Sci. USA*, **103**, 14718–14723.

Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.

Stein,A. *et al.* (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.

Stein,A. *et al.* (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.

Yeger-Lotem,E. *et al.* (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc. Natl Acad. Sci. USA*, **101**, 5934–5939.