BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 8 2005, pages 1479–1486
doi:10.1093/bioinformatics/bti240

Structural bioinformatics

Architecture of basic building blocks in protein and domain structural interaction networks

Hyun S. Moon[1], Jonghwa Bhak[2,3], Kwang H. Lee[2] and Doheon Lee[2,*]

[1]Division of Computer Science and [2]Department of BioSystems, KAIST, Daejeon, Korea and [3]BiO centre, Daejeon, Korea

Received on July 30, 2004; revised on November 28, 2004; accepted on December 15, 2004
Advance Access publication December 21, 2004

ABSTRACT

Motivation: The structural interaction of proteins and their domains in networks is one of the most basic molecular mechanisms for biological cells. Topological analysis of such networks can provide an understanding of and solutions for predicting properties of proteins and their evolution in terms of domains. A single paradigm for the analysis of interactions at different layers, such as domain and protein layers, is needed.

Results: Applying a colored vertex graph model, we integrated two basic interaction layers under a unified model: (1) structural domains and (2) their protein/complex networks. We identified four basic and distinct elements in the model that explains protein interactions at the domain level. We searched for motifs in the networks to detect their topological characteristics using a pruning strategy and a hash table for rapid detection. We obtained the following results: first, compared with a random distribution, a substantial part of the protein interactions could be explained by domain-level structural interaction information. Second, there were distinct kinds of protein interaction patterns classified by specific and distinguishable numbers of domains. The intermolecular domain interaction was the most dominant protein interaction pattern. Third, despite the coverage of the protein interaction information differing among species, the similarity of their networks indicated shared architectures of protein interaction network in living organisms. Remarkably, there were only a few basic architectures in the model (>10 for a 4-node network topology), and we propose that most biological combinations of domains into proteins and complexes can be explained by a small number of key topological motifs.

Contact: doheon@kaist.ac.kr

1 INTRODUCTION

The recent availability of various genome-scale biological networks has enabled the analysis of their topological structures (Uetz *et al.*, 2000; Ito *et al.*, 2000, 2001; Mewes *et al.*, 2002). Some parameters of network topology, such as scale freeness and clustering coefficients, focus on vertex degree. These are useful parameters in explaining how networks were built and evolved, even though they are insufficient for characterizing patterns of interconnections among vertices. More detailed network motif identification, classification, search and analysis can provide a deeper understanding of the nature of interaction networks (Shen-orr *et al.*, 2002). However, the size of network and the diversity of the motifs lead to an almost infinite number of network motifs, even when the target network is non-biological. In biology, the increasing availability of protein interaction databases has made it possible to predict and analyze protein functions (Bolser and Park, 2003, http://bio.cc/Biopaper/Paper/BiOpaper20030901_00001; Lappe *et al.*, 2001; Park *et al.*, 2001; Deng *et al.*, 2002b; Vazquez *et al.*, 2003; Letovsky and Kasif, 2003) through the analysis of interaction data (Deng *et al.*, 2002a; Ju *et al.*, 2003; Kim *et al.*, 2002). One of the sources of large-scale protein-interaction datasets is the underlying three-dimensional (3D) structure that has been available in Protein Data Bank (PDB). Protein structures represented as distinct protein domains are fundamental units in the evolution of genes and proteins (Copley *et al.*, 2003; Ikeo *et al.*, 1995; Ponting and Russell, 2002; Teichmann *et al.*, 1998). The physical interactions of proteins are controlled by their structural domains. There have been numerous studies on individual protein structural interactions such as the coevolution of interacting domains (Pazos and Valencia, 2001; Goh *et al.*, 2000; Goh and Cohen, 2002; Bolser and Park, 2003; Kim *et al.*, 2004). Also, structural interactions at the protein domain level can be mapped into a global protein domain interaction network, such as a protein structural interactome map (PSIMAP) (Park *et al.*, 2001; Dafas *et al.*, 2004). The advantage of such a network is that it is the most definite and conceptually clear category of molecular interaction. Distinct 3D structural domains contain 3D domain partners that exhibit well-characterized interactions. Even the whole human interactome can be represented as a structural interaction network (Human Protein Interaction Database) (Kim *et al.*, 2003; Han *et al.*, 2004) using PSIMAP's homologous interaction protocol. One of its disadvantages is that the relationship representations in proteins (i.e. domain–domain, domain–protein and protein–protein) are very complex in terms of evolution, since proteins evolve at different levels of functional constraints such as domain, multidomain and complex. The presence of higher levels of biological interactions in cells makes a lower level protein–domain interaction network insufficient for understanding the evolution and functions of genes. Therefore, it is necessary to clearly map and represent the interrelationships among all the levels of protein interactions in a coherent way. If feasible, computational methods that can encompass different levels of structural interaction simultaneously will provide the most insight. In the present study, we built Protein and Domain Interaction Network (PaDIN) using a colored vertex graph model to simultaneously analyze interactions at both the protein and domain levels. In the PaDIN, we identified four building blocks showing relations between protein interactions and domain interactions. We searched for network motifs

[*]To whom correspondence should be addressed.

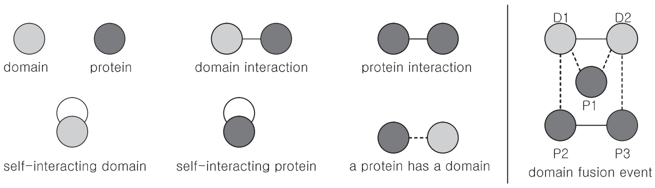


Fig. 1. PaDIN representation and a domain fusion event represented by the model.

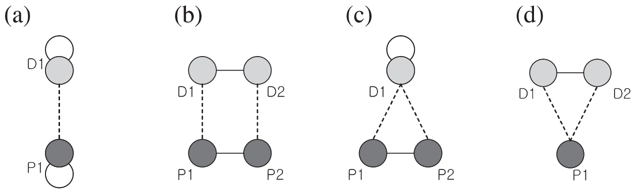


Fig. 2. Four building blocks that support protein interactions at the domain level. (**a**) Homointeractions, (**b**) heterodomain inter-molecular interaction, (**c**) homologue interaction and (**d**) heterodomain intramolecular interaction.

in the PaDIN to characterize their local topological characteristics using an algorithm designed for the fast detection of network motifs. This algorithm allowed us to classify protein molecular interactions into a small number of types that can be explained by the combination of one or more of the four basic building blocks. This classification revealed that domain interactions can explain a substantial proportion of protein interactions, and that there is a relatively small number of basic architectures for interwired interaction networks.

2 SYSTEMS AND METHODS

2.1 Protein and domain interaction network

The class of a vertex is determined by its color, while the class of an edge is determined by the colors of the two vertices at its ends. Reflexive edges on domains or proteins indicate that they are self-interacting. Relationships which indicate that some protein has some domain are indicated by dotted lines, while interactions between domains or proteins are indicated by solid lines. The model is simple yet useful because it can represent biologically meaningful events such as domain fusion (Marcotte *et al*., 1999). For example, on the far right-hand side of Figure 1, P2 and P3 have interacting domains D1 and D2, respectively; D1 and D2 are fused into protein P1.

In the PaDIN, we found that interactions between two proteins can be explained by the combination of the four basic components of domain interactions illustrated in Figure 2:

- *Homointeraction*: Self interaction of a protein can be explained by its self-interacting domain.
- *Heterodomain intermolecular interaction*: If two interacting proteins P1 and P2 have interacting domains D1 and D2 respectively, the protein interaction can be supported by the domain interaction between D1 and D2.
- *Homologue interaction*: If two interacting proteins P1 and P2 consist of the same self-interacting domain D1, this protein interaction can be explained by the self interacting property of D1.
- *Heterodomain intramolecular interaction*: If protein P1 has two interacting domains D1 and D2, we can infer that there is a domain interaction between D1 and D2.

When structures of interacting proteins are given, we can represent the characteristics of the structure through a PaDIN.

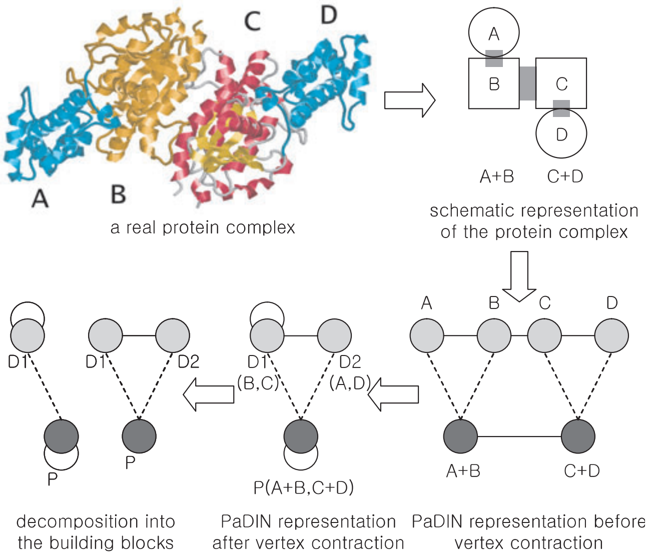

C
D
A
B
a real protein complex
A
B
C
D
A+B
C+D
schematic representation of the protein complex
A
B
C
D
A+B
C+D
PaDIN representation before vertex contraction
D1
D2
(B,C)
(A,D)
P(A+B,C+D)
PaDIN representation after vertex contraction
D1
P
D1
D2
P
decomposition into the building blocks

Fig. 3. An example of protein and domain interaction using a PaDIN. The 3D structure is the PDB entry 1azt.

An example using the structure 1azt in the PDB is shown in Figure 3. The two cyan-colored domains (A and D) at both ends have the same structure [a.66.1 in the SCOP (Structural Classification of Proteins) superfamily], as are the two central domains (B and C) (c.37.1 in the SCOP superfamily). The two left-hand domains (A and B) form a protein in which the two domains interact intramolecularly. The proteins A + B and C + D are homodimers. Interactions between domains are represented as shaded areas in the figure. The second step shows the schematic representation. In the third step, domain interaction information and protein interaction information are represented through the PaDIN model. Since protein A + B consists of two domains A and B, there is a dotted line between A + B and A, and between A + B and B. Since three domain interactions (between A and B, B and C, and C and D) are also evident, solid lines represent these interactions. There is also a solid line between A + B and C + D, since these two proteins are interacting. In the fourth step, since domains B and C are the same domain, they are combined into a singular node D1, and the edge between B and C becomes a reflexive edge on D1. Similarly, A + B and C + D are combined into P, and A and D are combined to D2. This is a complete representation through the PaDIN. The last step shows that this pattern of protein interaction can be broken down into two building blocks, illustrated in Figure 2a and d. This is unique in the PaDIN model when a structure of protein complex is given.

2.2 Source data

The interaction information for domains and proteins was derived from the PSIMAP (http://psimap.org/, Park *et al*., 2001) and DIP (Database of Interacting Proteins) (Xenarios *et al*., 2002) respectively (Fig. 4). PSIMAP is based on SCOP (Murzin *et al*., 1995), which is a database of hierarchical structures. Protein domains in SCOP are classified into families, superfamilies, folds and classes, on the basis of PDB data (Andreeva *et al*., 2004; Berman *et al*., 2000). We used the superfamily as the domain abstraction level, taking under consideration not only the sequence homology of the proteins but also their structural features and, most importantly, their evolutionary relationships. Information on protein–protein interactions was obtained from DIP. Currently this database holds protein-interaction information for various species, including human and yeast, even though the quality of results from two-hybrid studies is still debated. The sequences of proteins from DIP are matched with the structures of PSIMAP using the SUPERFAMILY database whose core is a protein sequence library based on a hidden Markov model. The current version covers 60% of all proteins (Madera *et al*., 2004).

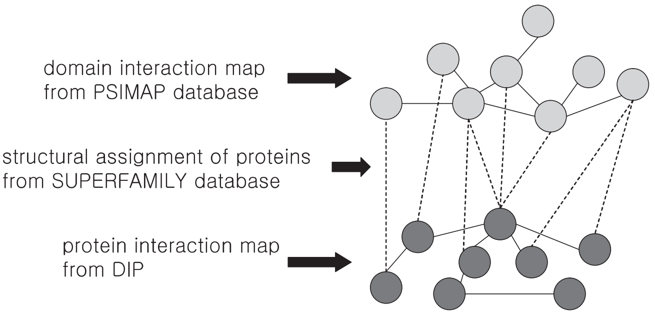

domain interaction map from PSIMAP database
structural assignment of proteins from SUPERFAMILY database
protein interaction map from DIP

Fig. 4. The conceptual integration of protein-interaction and domain-interaction maps into a singular representation. The bottom network represents a protein-interaction map (obtained from experimental data), and the top network represents the PSIMAP, from PDB and SCOP.

3 ALGORITHM FOR DETECTING NETWORK MOTIFS

In graph theory, a network motif is a subgraph of a given graph whose frequency is substantially higher than that of randomized networks. Therefore, in order to find network motifs from a graph, one can (1) enumerate all of its possible subgraphs, (2) count the frequency of each subgraph and (3) compare their frequencies with those in randomized graphs. In order to count the frequency of each subgraph efficiently, we used a canonical labeling of a graph. Two graphs have the same canonically relabeled graph if and only if they are isomorphic to each other. An overview of the algorithm for detection of the network motifs is given in Algorithm 1.

Algorithm 1. Find network motifs in a graph G.

```
 1: Let $G$ be the given network.
 2: Let $H$ be a set of network motifs and $C$ be its counter, initially
    empty
 3: $\phi$ = the set of all possible subgraphs of $G$
 4: for each $s \in \phi$ do
 5:    if s is not biconnected then
 6:       skip $s$
 7:    else
 8:       Convert $s$ into a canonical form.
 9:       if $s \in H$ then
10:          $C(s) \leftarrow C(s) + 1$
11:       else
12:          $H \leftarrow H \cup \{s\}, C(s) \leftarrow 0$
13:       end if
14:    end if
15: end for
16: Generate randomized graphs and repeat above.
17: Compare frequencies.
```

The main difficulty in the procedure is that there can be too many subgraphs in a given graph to detect within a practical timescale. In order to reduce the complexity, we applied a biconnectedness heuristic that rapidly detects biologically less-meaningful network motifs (lines 5 and 6 in Algorithm 1). Another problem is the efficient determination of the corresponding network motif (line 9). To solve this computation problem, we used canonical labels for rapidly detecting differences between two graphs based on hashing (line 8). The proposed algorithm reduced the computational complexity of the problem. The details of each step are discussed in a later part of this section.

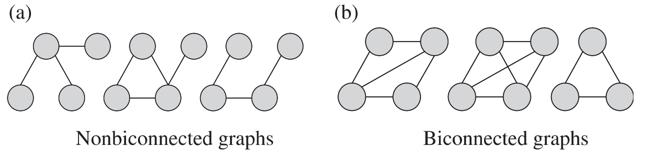


Fig. 5. Examples of (**a**) non-biconnected and (**b**) biconnected graphs. Biconnected graphs always have two paths at least between every pair of vertices, while in non-biconnected graphs there is always more than one articulation point whose deletion partitions the graph.

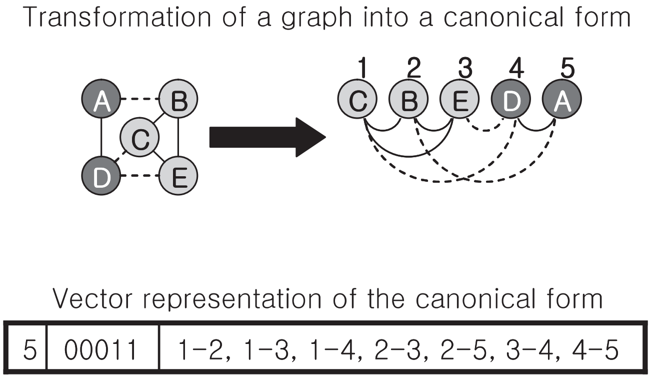
3.1 Biconnectivity in network motifs

A graph is biconnected for every pair of its vertices if and only if there are at least two linking paths (Fig. 5). In network motif analysis, the most difficult problem is to explain why such network motifs, statistically overrepresented subgraphs, exist. In some cases, we found self-descriptive network motifs such as feed-forward networks in a gene regulatory network or cliques in a protein interaction network. However, even in that case, we have many other network motifs whose biological meaning is not clear. In our analysis, we focus on network motifs with biconnected structure for the following two reasons. First, in case of the PaDIN, network motifs should be at least biconnected to properly describe interactions which are consistent at both protein and domain levels. Second, it can reduce the size of search space without missing statistically significant subgraphs, since a linear time algorithm is known for testing whether a graph is biconnected or not (Horowitz *et al*., 1993), and most of network motifs discussed in several previous reports have biconnected structures (Shen-orr *et al*., 2002; Milo *et al*., 2002; Wuchty *et al*., 2003). How far-reaching this property is in molecular interactions requires further investigation. However, it can represent an efficient filtering rule for analyzing the relationship for both domain–domain and protein–protein interactions. This tighter association rule between proteins and domains is also relevant to functional clusters of proteins and domains that form complexes.

Applying this heuristic to the protein domain network dramatically decreased the number of subgraphs which require detailed analysis and interpretation. For example, the human interaction map comprised about 0.2 million subgraphs with vertex size 3 and 4 before the heuristic, and application of our heuristic approach resulted in around 2000 subgraphs, a reduction of 99%.

3.2 Subgraph comparison

Comparing the structures of two graphs is a graph-isomorphism problem. Solving this problem involves finding a homomorphism function h, if it exists. For two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, a homomorphism function $h : V_1 \rightarrow V_2$ is a bijection mapping (both a one-to-one and an onto function) that satisfies the following condition: $(v, w) \in E_1$ if and only if $(h(v), h(w)) \in E_2$. This problem is known to be equivalent to the canonical labeling problem. Canonical labeling of a graph is a permutation of the vertices of the graph which guarantee that canonically relabeled forms of two graphs are the same if and only if they are isomorphic to each other. Even though it requires subexponential time in the worst case, the algorithm devised by McKay (1978) is efficient for addressing the problem of canonical labeling.


Transformation of a graph into a canonical form
A
B
C
D
E
1
2
3
4
5
C
B
E
D
A
Vector representation of the canonical form

5	00011	1–2, 1–3, 1–4, 2–3, 2–5, 3–4, 4–5

Fig. 6. Example of transformation of a graph into the canonical form. The labels on the vertices are labeling of the input and the digits above the vertices are canonical labelings. A vector representation of the canonical form is shown at the bottom of the figure.

In the problem of network motif detection, we have a set of candidate network motifs. For each generated subgraph, we have to determine which of the candidates is isomorphic to the generated subgraph. We used a hash table for this process. Since it is useless to store graphs in a hash table, we stored canonically relabeled graphs in a hash table instead of the graphs themselves. Once a canonical labeling of a graph is found, we can represent the graph as a vector of integers. The vector representation of a graph includes following information:

- the number of the vertices of the graph,
- the colors of the vertices in canonical order and
- sorted list of the edges in the canonical form.

Figure 6 shows how a graph is represented as a vector form. The first part of the vector representation is the number of vertices (in this example, five). The second part of the vector is the list of colors of the vertices. In the example, 0 means light vertices while 1 means dark vertices. The third part of the vector consists of all edges in the canonical form. For example, 1–2 means that there is an edge between vertex 1 and vertex 2 where 1 and 2 are canonical labels.

For efficient implementation, we built a hash table of vector representations of candidate network motifs. A hash table is a useful data structure in determining whether a value is in a set or not. The steps are:

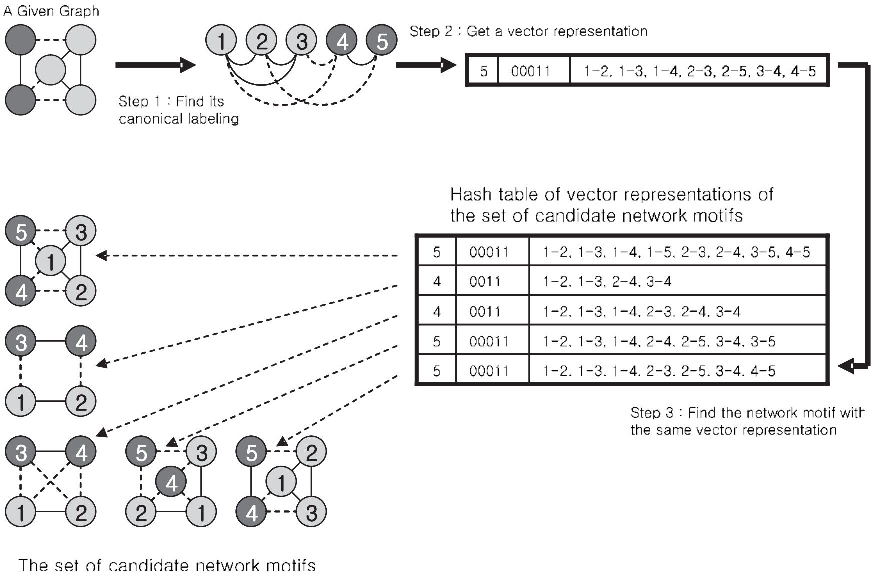
(1) Find canonical labelling of the given graph.
(2) Get a vector representation.
(3) Find the network motif isomorphic to the given graph.
(4) If found, increase the counter. Otherwise, insert the given graph to the set of candidate network motifs. The overall procedure is described in Figure 7.

4 RESULTS

The PSIMAP database contains structural interaction information on 1125 domains defined as superfamilies in the SCOP database. Out of 1125 superfamilies, 859 domains (76.4%) were self-interacting domains, while the number of non-self-interacting domains was only 266 (23.6%). The total number of pairwise interactions was 1212, of which 698 (57.6%) were interactions between two self-interacting domains (including both of homointeractions and homologue interactions), 402 (33.2%) were between self-interacting and non-self-interacting domains, and 112 (9.2%) were between non-self-interacting domains. The high ratio of self-interaction (homointeraction) is due to a superfamily being a high-level classification. Interaction between two superfamilies is established when there is at least a pair of interacting domains from each of the superfamilies. Nevertheless, the ratio of homointeraction is significantly high, a feature commonly observed in protein interactions. Also, we predict that the ratio of homointeraction will continue to increase since the number of interacting domains will increase with further structure determinations and experiments. We interpret that the very high ratio of homointeraction in protein domains reflects (1) that it is a characteristic of the interactions of proteins in life forms on Earth, (2) the perfect symmetry achieved in dimerization that leads to more stable protein-interaction architecture and (3) some degree of artifact by the crystallization process if the protein interaction information is derived from structures determined by X-ray diffraction.

If protein interactions are correlated with the domain constitution of the interacting proteins, the frequencies of the network motifs that show correlation between protein interactions and domain interactions should be substantially higher than those in the randomized networks. In our experiments, we divided the sets of protein interactions according to their species. We generated PaDINs for both human and yeast protein interactions, and their network motifs with three or four nodes were searched. Although hundreds of biconnected network motifs were found from the protein and domain networks, we carefully selected six of them which described types of protein interactions. Some of the other network motifs which do not describe types of protein interactions are shown in Figure 8.

Our analysis of the human and yeast PaDINs revealed that a substantial part of the protein interaction information can be explained by domain-level structural interactions. A total of 591 out of 715 human protein interactions and 7469 out of 15 060 yeast protein interactions had their structures assigned in SCOP. We considered only these protein interactions because we could not determine the other structures or whether they were supported by structure-level interactions. We found that 340 out of 591 (57.5%) human protein interactions in DIP can be explained by structure-level interactions. In the case of yeast protein interactions, ignoring those proteins which are not structurally assigned, there were 1908 out of 7469 (25.5%) protein interactions that could be explained by at least one of the building blocks in Figure 2. The low coverage is not surprising considering previous studies on overlaps between MIPS complexes and genome-wide Y2H interaction datasets in which it was shown that less than 50% of high-throughput protein interactions were overlapped with MIPS complex categories (Edwards *et al*., 2002). This is mainly due to the following three reasons. First, domain-level structural interaction information may not yet be complete. By November 2004, there are about 28 000 known protein structures and their complexes in the PDB, while sequences of more than 1.6 million proteins are known in the UniProt database. There were a few approaches to bridge the gap in knowledge between complexes of known 3D structures and those known from other experimental methods such as two-hybrid system (Aloy *et al*., 2003). However, low coverage of structural interaction support for high-throughput protein interaction data imply that it is still questionable whether structural domain


A Given Graph
Step 1 : Find its canonical labeling
Step 2 : Get a vector representation
Hash table of vector representations of the set of candidate network motifs
Step 3 : Find the network motif with the same vector representation
The set of candidate network motifs

5	00011	1–2, 1–3, 1–4, 2–3, 2–5, 3–4, 4–5

5	00011	1–2, 1–3, 1–4, 1–5, 2–3, 2–4, 3–5, 4–5
4	0011	1–2, 1–3, 2–4, 3–4
4	0011	1–2, 1–3, 1–4, 2–3, 2–4, 3–4
5	00011	1–2, 1–3, 1–4, 2–4, 2–5, 3–4, 3–5
5	00011	1–2, 1–3, 1–4, 2–3, 2–5, 3–4, 4–5

Fig. 7. Determination of the isomorphic network motif for a given graph. The digits on the vertices are canonical labeling of the graphs.

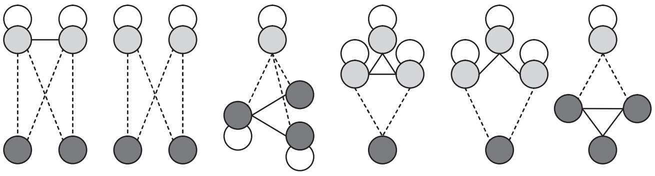


Fig. 8. Arbitrary examples of network motifs found in PaDIN that do not describe types of protein interactions. These network motifs can also have biological significance.

interaction information from 3D protein complex structures covers all of the domain interactions in nature. Second, the protein interaction maps have many false positives. In case of more reliable dataset of yeast protein interaction, the CORE dataset of the DIP, 1425/3755 (37.9%) assigned protein interactions are supported by structural interaction. This is higher compared with that of raw yeast protein interaction data (25.5%). Third, the coverage of genome assignment is not high. For example, in case of the SUPERFAMILY database, its coverage is <60% in both protein and amino acid sequences. One-domain proteins are less common than multidomain proteins in most organisms, which reduces the amount of protein interaction that can be accounted for by domain interaction. However, even in the case of yeast, the coverage was significant compared with randomized networks of domain and protein interactions. The detailed statistics will be discussed later.

4.1 Domain composition patterns of interacting proteins

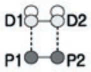
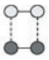
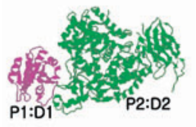
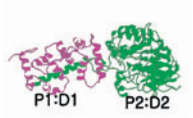
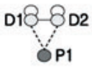
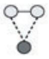
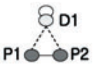
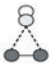
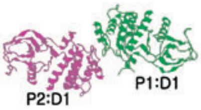
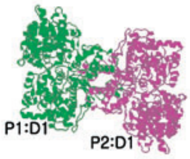
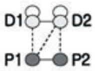
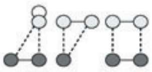
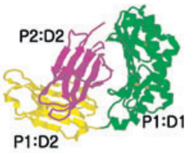
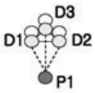
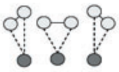
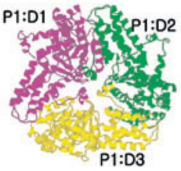
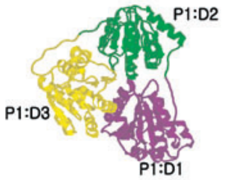
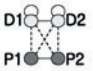
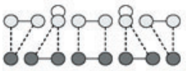
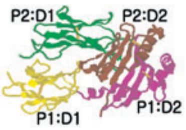
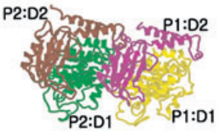
Network motifs in the PaDINs can provide insight into how the domains/proteins interact and form complexes. We could identify a set of network motifs that shows the domain composition patterns of interacting proteins. These network motifs are summarized in Table 1, which includes examples of their 3D structures. All of them were explained by a combination of the basic interaction components (building blocks) described in Figure 2. Detailed statistics of these network motifs are described in Table 2.

- *Intermolecular interaction*: Two different interacting domains D1 and D2 belong to different protein chains P1 and P2.
- *Intramolecular interaction*: Two different interacting domains D1 and D2 belong to a single protein chain.
- *Homologue interaction*: Two proteins P1 and P2 with the same domain D1 interact with each other.
- *2:1 Intermolecular interaction*: Intermolecular interaction between protein P2, which consists of domain D2 (violet), and protein P1, with two domains D1 and D2 (green and yellow). Note that domain D2 is shared by the two interacting proteins P1 and P2.
- *Intramolecular interaction among three domains*: Three different interacting domains (D1, D2 and D3) belong to a single protein chain.
- *2:2 Intermolecular interaction*: Two protein chains P1 and P2 with the same domain composition of D1 and D2 form a complex. The green and brown domains form protein P2, and the violet and yellow domains form protein P1. The structures of the brown and violet domains are similar, and the structures of the yellow and green domains are similar.

Note that motifs representing 2:2 Inter, 2:1 Inter and 3 Intra interaction in Table 1 can be broken down into building blocks in Figure 2.

An advantage of the PaDIN is that it gives more refined and exact classification of protein interactions or protein complexes. For example, it is estimated that the number of protein interaction types is ~10 000 (Aloy and Russell, 2004). In this estimation, two protein interactions (P1,P2) and (P3,P4) are classified into the same type if homology between P1 and P3, and homology between P2 and P4 are found (Aloy and Russell, 2004; Russell *et al*., 2004). However,

Table 1. Network motifs showing interaction patterns between domains and the 3D structures of their protein complexes

Description	Motifs	Decomposition to building blocks	Two examples of corresponding structures from PDB	
Intermolecular				
Intramolecular				
Homologue				
2:1 Intermolecular				
3 Domains Intramolecular				
2:2 Intermolecular				

according to the proposed model, this classification method can classify different protein interactions into the same type. Suppose that two protein interactions (P1,P2) and (P3,P4) are observed. If protein P1 consists of domain D1, P2 consists of D2, P3 consists of D1 and D2, and P4 consists of D2, then these two protein interactions, (P1,P2) and (P3,P4), will be classified into the same interaction type, since P1 and P3 have the same domain D1, and P2 and P4 have the same domain D2. However, under the proposed model, these two interactions are clearly distinguished: the former, (P1,P2), is an intermolecular interaction, and the latter, (P3,P4), is a 2:1 intermolecular interaction. From our network motif analysis, we found that the number of more complicated interaction types, e.g. 2:1, 2:2 intermolecular interaction, was not negligibly small even though it was smaller than the number of basic interaction types. (Table 2) This observation strongly implies that the number of types of protein interactions can be greater than estimated by Russell's group (Aloy and Russell, 2004).

We also found other network motifs in PDINs (some of their arbitrary examples are shown in Fig. 8). These network motifs cannot directly explain the domain composition patterns of complex formation. Although we suspect that these network motifs have some topological and biological meaning, in this paper we restrict ourselves to the complex formation of proteins in Table 1.

Table 2. Statistics of network motifs showing the relationship between protein- and domain-level interactions

Interaction class	Human Real	Randomized	*P*-value	Yeast Real	Randomized	*P*-value
Inter	129	20.43 ± 8.3	10^{-39}	712	306.41 ± 95.9	10^{-7}
Intra	97	12.7 ± 5.8	10^{-47}	335	45.06 ± 12.8	10^{-113}
Homologue	34	6.83 ± 2.8	10^{-22}	460	123.23 ± 36.6	10^{-21}
2:1 Inter	20	1.7 ± 2.1	10^{-18}	83	42.27 ± 16.7	10^{-3}
3 Domains intra	10	∼0	∼0	35	∼0	∼0
2:2 Inter	3	∼0	∼0	23	∼0	∼0

The numbers in the 'Real' columns are the frequencies of the patterns in the real networks, and the numbers in 'Randomized' columns are means and standard deviations in randomized networks. Randomized networks which have the same degree distribution with the original graph were computed by Monte-Carlo simulations as follows. Starting from the original PaDIN, we performed a long series of random edge crosses, each time picking random two edges (a,b), (c,d), and replacing them with (a,c), (b,d) without allowing self-loops. This is a typical method to generate random networks (Sharan *et al*., 2004). In case of PaDINs, one more constraint was given to edge crosses since each of the vertices has its own type: when we pick two edges (a,b) and (c,d) the type of a and c should be the same and the same holds for b and d.

4.2 Evolution of domain interaction pattern

If the patterns of protein interaction do evolve as a distinct object or entity, we can hypothesize that the frequencies of some motifs are higher than others, and that some types of interaction are more frequent than others. As a simple test, we have compared the frequencies of network motifs and types of interaction in two genomes. Table 2 lists the intermolecular domain interactions that are the most common in human and yeast protein-interaction networks, from which we suggest that the intermolecular domain interaction is the most common type of interaction in the two genomes. Biologically, this implies that the majority of domains are found as complexes (as well as the constituents of multidomain proteins). In the human-interaction map, the second most common pattern is a heterodomain intramolecular interaction, whereas it is a homologue interaction pattern in yeast. It is not clear why this is the case, but one hypothesis is that higher organisms contain more proteins comprising different types of domains interacting intramolecularly. Assuming this is generally true, we can predict that the protein domain interaction map will become more heterogeneous—in terms of heterodomain combinations—as complex organisms evolve.

4.3 Cross-species conservation of network motifs

We can compare the topological characteristics of networks based on their motif constituents. However, it is difficult to compare the similarity of topological structures directly between two networks because of their different sizes. We measured the similarity of networks by comparing the frequencies of common network motifs. If network motifs whose frequencies are high in one network appear frequently in another network, the structures of the two networks may be similar. The network motifs were considered when they were biconnected and had at least one protein and one domain simultaneously. As indicated in Figure 9, there was a strong correlation (correlation coefficient 0.88) between the frequencies of the network motifs in human and yeast. We suggest that the interaction maps of other species will also show similar trends, since protein-interaction networks are tightly conserved (Park *et al*., 2001), nearly enough to be used as molecular clock for estimating the age of species by comparing protein interactomes (Bolser and Park, 2003).

5 CONCLUDING REMARKS

Despite continuous improvements in the analysis of network motifs, such analyses have been restricted mostly to small networks because of the associated computational complexity. In this paper we have presented an implementation strategy for the rapid detection of network motifs that employs a pruning technique using the connectivity of graphs. The vertex colored graph model can be easily adapted to non-biological networks, and it can be used to detect and analyze interactions present in networks comprised of different layers. The PaDIN representation can be used to compare, search and detect network motif patterns to concurrently classify and analyze the evolutionary relationship observed in interactions among proteins and domains. One merit of the PaDIN representation is that it can be used to classify and track the evolution of specific molecular interactions in both individual protein complexes and multidomain proteins.

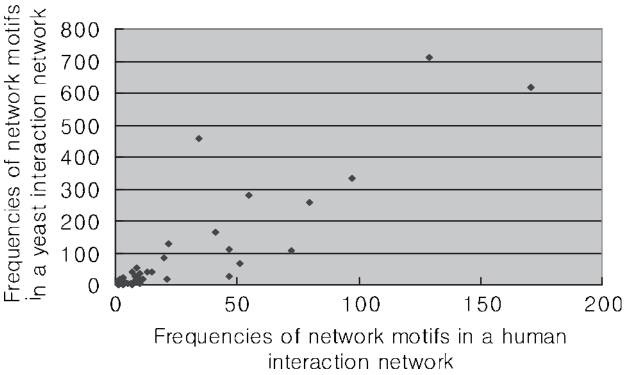


Fig. 9. Comparison of the frequencies of the network motifs in human- and yeast-interaction networks. Each dot is a network motif found in both yeast and human. Although there are differences between the two networks in terms of size and species, the topological components of the networks contain common motifs.

We analyzed both protein–protein and domain–domain interaction networks. To detect the network patterns between proteins and their domains, we built a unified model for both protein interactions and domain interactions. From the model, we derived statistically significant network patterns of interactions that are basic building blocks in domain and protein interactions. The architecture of the interactions discovered in the complicated interactions between proteins and their component domains indicates that life forms can be shown by specific network topologies. Some of the popular network motifs were simple and robust, and comprised only a small number of different types. Statistical analyses of the network motifs showed that a substantial portion of the protein-level interactions could be explained by information on domain-level structural interactions. This confirms the suggestion that in interacting proteins, the domain level is the most important to evolutionary selection. Overall, protein structural domains seem to be the most distinct and important biological entities for interaction, function and evolution. We observed a very high degree of homointeraction (self-interaction) in protein domains, consistent with the results from various previous studies. Moreover, we discovered that the intermolecular domain interaction is the most basic and common pattern of protein interaction. Despite the different coverage of the protein-interaction information between yeast and human species, the patterns of the protein/domain integrated interaction maps were quite similar indicating the commonality and consistency of protein interactome evolution.

ACKNOWLEDGEMENTS

This work was supported by Korea Research Foundation Grant (KRF-2003-041-D20490). We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM SUR program for providing research and computing facilities. BJ is supported under grant IMT-2000-C3-4 of the Ministry of Information and Communication of South Korea and Biogreen21 fund. We thank our colleague, Semin Lee, for providing Figure 3 and Maryana Huston for editing the manuscript.

REFERENCES

Aloy,P., Ceulemans,H., Stark,A. and Russell,R.B. (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, **332**, 989–998.

Aloy,P. and Russell,R.B. (2004) The thousand interactions for the molecular biologist. *Nat. Biotechnol.*, **22**, 1317–1321.

Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bolser,D. and Park,J. (2003) Biological network evolution hypothesis applied to protein structural interactome. *Genomics Inform.*, **1**, 7–19.

Bolser,D., Dafas,P., Harrington,R., Park,J. and Schroeder,M. (2003) Visualisation and graph—theoretic analysis of a large-scale protein structural interactome. *BMC Bioinformatics*, **4**, 45–76.

Copley,P.R., Goodstadt,L. and Ponting,C. (2003) Eukaryotic domain evolution inferred from genome comparisons. *Curr. Opin. Genet. Dev.*, **13**, 623–628.

Dafas,P., Bolser,D., Gomoluch,J., Park,J. and Schroeder,M. (2004) Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics*, **20**, 1486–1490.

Deng,M., Mehta,S., Sun,F. and Chen,T. (2002a) Inferring domain–domain interactions from protein–protein interactions. *Genome Res.*, **12**, 1540–1548.

Deng,M., Zhang,K., Mehta,S., Chen,T. and Sun,F. (2002b) Prediction of protein function using protein–protein interaction data. In *IEEE Computer Society Bioinformatics Conference (CSB'02)*, Stanford, CA, pp. 197–206.

Edwards,A.M., Kus,B., Jansen,R., Greenbaum,D., Greenblatt,J. and Gerstein,M. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes, *Trends Genet.*, **18**, 529–536.

Goh,C. and Cohen,F. (2002) Co-evolutionary analysis reveals insight into protein–protein interactions. *J. Mol. Biol.*, **324**, 177–192.

Goh,C., Bogan,A., Joachimiak,M., Walther,D. and Cohen,F. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Han,K., Park,B., Kim,H., Hong,J. and Park,J. (2004) HPID: the Human Protein Interaction Database. *Bioinformatics*, **20**, 303–305.

Horowitz,E., Sahni,S. and Anderson-Freed,S. (1993) *Fundamentals of Data Structure in C*. Computer Science Press.

Ikeo,K., Takahashi,K. and Gojobori,T. (1995) Different evolutionary histories of kringle and protease domains in serine proteases: a typical example of domain evolution, *J. Mol. Evol.*, **40**, 177–192.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y. (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins, *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.

Ju,B.H., Park,B., Park,J. and Han,K. (2003) Visualization and analysis of protein interactions. *Bioinformatics*, **19**, 317–318.

Kim,W.K., Bolser,D. and Park,J. (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150.

Kim,H., Park,J. and Han,K. (2003) Predicting protein interactions in human by homologous interactions in yeast. *LNCS* **2637**, 159–165.

Kim,W.K., Park,J. and Suh,J.K. (2002) Large scale statistical prediction of protein–protein interaction by Potentially Interacting Domain (PID) Pair. *Genome Inform.*, **13**, 42–50.

Lappe,M., Park,J., Niggemann,O. and Holm,L. (2001) Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics*, (Suppl.), S149–S156.

Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, (Suppl.), I197–I204.

Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.

Marcotte,E., Pellegrini,M., Ng,H., Rice,D., Yeates,T. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

McKay,B. (1978) Computing automorphisms and canonical labelling of graphs. *LNM*, **686**, 223–232.

Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Milo,R., Shen-Orr,S., Itzkovitz,S., Kashtan,N., Chklovskii,D. and Alon,U. (2002) Network motifs: Simple building blocks of complex networks. *Science*, **298**, 824–827.

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

Park,J. and Bolser,D. (2001) Conservation of protein interaction network in evolution. *Genome Inform.*, **12**, 135–140.

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

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, **14**, 609–614.

Ponting,C.P. and Russell,R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 45–71.

Russell,R.B., Alber,F., Aloy,P., Davis,F.P., Korkin,D., Pichaud,M., Topf,M. and Sali,A. (2004) A structural perspective on protein–protein interactions and complexes. *Curr. Opin. Struct. Biol.*, **14**, 313–324.

Sharan,R., Ideker,T., Kelley,B., Shamir,R. and Karp,R. (2004) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *Proceedings of RECOMB2004*, San Diego, CA, pp. 282–289.

Shen-Orr,S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64–68.

Teichmann,S., Park,J. and Chothia,C. (1998) Structural assignments to the proteins of *Mycoplasma genitalium* show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **30**, 14658–14663.

Uetz,P., Gior,L., Cagney,G., Mansfield,T., Judson,R., Knight,J., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.

Wuchty,S., Oltval,Z.N. and Barabasi,A.L. (2003) Evolutionary conservation of motif constituent in the yeast protein interaction network. *Nat. Genet.*, **35**, 176–179.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.