Exploring Cross-Function Domain Interaction Map

Xiao-Li Li, Soon-Heng Tan, and See-Kiong Ng
Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Email : xlli@i2r.a-star.edu.sg, soonheng@i2r.a-star.edu.sg, skng@i2r.a-star.edu.sg

ABSTRACT: Living cells are sustained not by individual activities but rather by coordinated summative efforts of different biological functional modules. While recent research works have focused largely on finding individual functional modules, this paper attempts to explore the connections or relationships between different cellular functions through cross-function domain interaction maps. Exploring such a domain interaction map can help understand the underlying inter-function communication mechanisms. To construct a cross-function domain interaction map from existing genome-wide protein-protein interaction datasets, we propose a two-step procedure. First, we infer conserved domain-domain interactions from genome-wide protein-protein interactions of yeast, worm and fly. We then build a cross-function domain interaction map that shows the connections of different functions through various conserved domain interactions. The domain interaction maps reveal that conserved domain-domain interactions can be found in most detected cross-functional relationships and a few domains play pivotal roles in these relationships. Another important discovery in the paper is that conserved domains correspond to highly connected protein hubs that connect different functional modules together.

1 INTRODUCION

Understanding complex protein interaction networks is a major challenge in post-genomic biology. Particularly, discovering protein functional organization is a key to understand biological processes, and much recent research works have focused on finding individual functional modules in biological interaction networks [2, 3, 6, 8]. While the unraveling of such functional modules reveal biological groupings of proteins with similar functions, it is also very important to study the interactions among different functional modules. This is because living cells are sustained not by individual activities but rather by coordinated summative efforts of different functional modules. Therefore, in addition to studying the internal wiring mechanisms of biological functional modules, we also need to understand how these modules communicate with each other to synchronize their activities, for example, in response to extra-cellular signals and environmental changes.

This paper attempts to explore the cross-function communication mechanisms through a domain interaction map. Protein domains are evolutionarily-conserved structural or functional subunits found across different proteins. They can be thought of as "building blocks" of proteins — a vast amount of proteins with diverse functionalities are created from assembly of different protein domains. It is thus natural to analyze protein interactions at the level of domains, and in this work, we will derive a cross-function domain interaction map instead of a protein map to unravel how different functional modules co-operate or interact.

To construct a cross-function domain interaction map, our proposed technique consists of the two steps.

1. **Inferring conserved domain interactions.** As current genomic-wide data contained much spurious protein interactions [10], we mine *conserved* domain-domain interaction using genomic-wide protein interaction data from multiple species. This step aims to obtain accurate domain interactions through aligning domain pairs in multiple species, based on an assumption that domain interactions observed across multiple species are more likely to be real. In addition, interacting domain pairs that are conserved across species are more likely to correspond to the core essential interactions networks found across multiple species [7]. In this work, a domain pair is regarded as an interacting pair if and only if it is detected in all the three species.

2. **Constructing a two-level domain interaction map.** Using the protein-protein interaction datasets together with the conserved domain interactions that we have mined in the previous step, we then construct a cross-function interaction map that consists of two levels: the first level shows the interactions between various biological functional modules, while the second level shows the domain interactions that mediate such cross-functional biological interactions. The resulting cross-function domain interaction map can help biologists develop useful insights about the underlying communication mechanisms between the various biological functional modules.

Using the cross-function domain interaction map that we have constructed from the cross-species interaction datasets from yeast, worm and fly, we performed various detailed analysis to explore the following topics: (a) what functions co-operated to perform biological processes; (b) which domains (domain pairs) are often used to bridge the protein functions, and (c) what are the relationships between protein hubs and conserved domains.

2 METHODOLOGY

In this section, we introduce our two-step proposed technique. First, we infer conserved domain-domain interactions through aligning domain pairs in species yeast, worm and fly. We then build the cross-function domain interaction map which shows how the different functions are connected and mediated by various domains. There are two levels in this interaction map: the first level is the functional level interaction map; it shows the overall functional relationships mediated by conserved domain interactions. In this map, the vertexes are the various

function modules and an edge connects two functions if conserved domain interactions are found in the interacting proteins from the corresponding two functions. The second level of the map is the detailed cross-function interaction map in terms of the actual domain-domain interactions. While the functional level interaction map provides insights to the relationships among the functions, this second level can show how the different functions are coordinated at the protein domain level.

2.1 Infer domain-domain interactions

In recent years, in order to understand the protein interaction mechanisms and reveal potential protein interactions, researchers have inferred domain-domain interactions from protein interaction data. Sprinzak et al. [9] were amongst the first to attempt to characterize protein interactions using domains in InterPro. Deng et al. [1] described scoring techniques for inferring domain-domain interactions from interaction data, while Ng et al. [5] devised an integrative approach to infer domain interactions from multiple interaction sources. All these methods inferred the domain interactions from the model organism yeast.

In this study, we focus on conserved domain-domain interactions across multiple species, unlike the previous works above. Multi-species genome-wide protein interaction datasets are exploited to infer conserved domain-domain interactions. As it has been noted that there are ~50% false (noisy) protein interactions in current protein interaction data [10], many domain interactions predicted by the noisy protein interactions from a single species may be false positives. As such, we propose to infer accurate domain interactions from domain interactions observed across multiple species. A domain pair is regarded as an interacting pair in our study if and only if it is detected in protein interaction data of the various species. By requiring that the domains must be conserved as well as their interactions across multiple species, the possibility that the inferred domain interaction is a false positive is much reduced. In this study, we use the protein interaction data from yeast, fly, and worm.

Given a multi-species protein-protein interaction (PPI) set, i.e. yeast PPI (YPPI), fly PPI (FPPI) and worm PPI (WPPI), our algorithm attempts to infer accurate conserved domain-domain interactions as follows:

Algorithm for inferring domain interactions

1. $DDI = \Phi$;
2. Construct domain set D_y, D_f, and D_w for proteins in YPPI, FPPI and WPPI respectively;
3. Generate all domain pair sets for the three species (here domain $d_i, d_j \in D_y \cup D_f \cup D_w$)
   $DP_y=\{(d_i, d_j) | d_i \in p_a, d_j \in p_b, (p_a, p_b) \in YPPI\}$;
   $DP_f=\{(d_i, d_j) | d_i \in p_a, d_j \in p_b, (p_a, p_b) \in FPPI\}$;
   $DP_w=\{(d_i, d_j) | d_i \in p_a, d_j \in p_b, (p_a, p_b) \in WPPI\}$;
4. **For** all $(d_i, d_j) \in DP_y \cup DP_f \cup DP_w$
5. **If** (d_i, d_j) occurs in $DP_y \cap DP_f \cap DP_w$
6. $DDI = DDI \cup \{(d_i, d_j)\}$;
7. **Endif**
8. **Endfor**

In the algorithm, Step 1 first initializes the inferred domain interaction set DDI as empty set. Then, Step 2 obtains the domain information for all the proteins in the three species. Step 3 generates three domain pair sets from the YPPI, FPPI and WPPI respectively. Steps 4 to 8 validate if the domain pairs generated are shared by all the three species. If so, then we store them into our inferred domain interaction set (DDI). The inferred domain interactions in DDI will then be used to build cross-function domain interaction map described in the next section.

2.2 Cross-function domain interaction map

In this section we build a 2-level domain interaction map to show (a) the overall functional relationships in the biological system (the functional-level interaction map), and (b) how the different functions are connected and mediated by various protein domains through the conserved domain interaction (the cross-function domain interaction map).

2.2.1 Functional-level interaction map

First, we build a functional-level interaction map using the conserved domain-domain interactions inferred above (DDI) and the multi-species protein interaction data (PPI). We use the following assumption: if proteins with different functions interact with each other, then the corresponding functional modules are biologically related. In other words, we construct a functional interaction map by connect two biological function modules if there are two proteins, each from one of the two functions, interact in PPI.

As mentioned earlier, the fractions of false positives in genome-wide interactions detected by high-throughput methods are high [10]. In this work, we therefore use two strategies to remove potential false functional relationships inferred from PPI: 1) first, we filter away functional interactions with low connectivity strengths between the functions in terms of the corresponding interactions in PPI; and 2) keep the functional relationships only when the corresponding cross-function protein interactions from PPI can also be explained by conserved domain interactions in the set DDI inferred above.

We build the functional interaction map as follows: for each protein-protein interaction in PPI, we exclude it if it cannot be explained by some conserved domain interaction in DDI. Then, we search the two proteins' functions for each protein interaction. If they belong to different functions, we connect the corresponding function modules with an edge. We compute the connectivity strength for each of these edges as an indication of how reliable of these functional relationships based on the following assumption: if there exists many interacting proteins connecting two protein functions and the interactions can be explained by our conserved domain interactions, then it is more reliable to infer that the two functions perform corresponding biological processes in a co-operative way.

Given two functions f_u and f_v, their connectivity strength $cs(f_u, f_v)$ is defined by using the following formula:

$$cs(f_u, f_v) = |\{(p_a, p_b) | (f(p_a) = f_u) \wedge (f(p_b) = f_v), \exists d_i, d_j, (d_i, d_j) \in DDI, (d_i \in p_a) \wedge (d_j \in p_b)\}|$$

where $cs(f_u, f_v)$ measures the degree of affinity between the function pair f_u and f_v. We count the number of interacting protein pairs that support a functional link between f_u and f_v if 1) one protein has function f_u while the other has function f_v and 2) this protein pair contains a domain pair which is in our conserved domain interaction set *DDI*. Basically, the more protein interactions (that can be explained by conserved domain interactions) connect the two functions f_u and f_v, the more likely the function f_u and f_v co-operate to perform biological processes, which is indicated by a bigger value of $cs(f_u, f_v)$. In practice, we only keep those edges with $cs(f_u, f_v) > \delta$, where δ is a user-defined threshold.

Based on the functional level interaction map, we then try to unravel the underlying inter-functional communication mechanisms mediated by various protein domains. Below, we show how to construct our cross-function domain interaction map.

2.2.2 Cross-function domain interaction map

To understand how different functions corrdinate and interact at the domain level, we expand the functional-level interaction map into a cross-function domain interaction map using proteins' functional information obtained from MIPS[1].

Overall algorithm for constructing cross-function domain interaction map

1. **For** all function pair (f_u, f_v)
2. **Next If** ($f_u == f_v$)
3. **For** all protein pair (p_a, p_b) in *PPI*
4. **If** ((fun(p_a)=f_u) & (fun(p_b)=f_v))
5. connect function f_u and f_v
6. update the weight $cs(f_u, f_v)$
7. **Endif**
8. **Endfor**
9. **Endfor**
10. construct functional level interaction map with $cs(f_u, f_v) > \delta$
11. **For** all function pair (f_u, f_v)
12. construct functional region for f_u and f_v;
13. **For** all protein pair (p_a, p_b) in *PPI*
14. **If** ((fun(p_a)=f_u) & (fun(p_b)=f_v))
15. Generate all domain pairs (d_e, d_f) from (p_a, p_b);
16. **If** (d_e, d_f) $\in DDI$
17. Add d_e in function region of f_u;
18. Add d_f in function region of f_v;
19. Link d_e and d_f with a edge;
20. **Endif**
21. **Endif**
22. **Endfor**
23. **Endfor**

[1] ftp://ftpmips.gsf.de/yeast/catalogues/funcat/

Our overall algorithm consists of two main loops. Steps 1 to 10 try to construct the functional level interaction map (section 2.2.1). Steps 11 to 23 try to find both cross-function protein interactions and the underlying domain interactions (section 2.1) that connect the two functions. The final graph constructed has a two-level structure. The first level (functional-level interaction map, see Figure 2 for an example) shows the possible relations between functional modules. Using this map, we can find the overall functional relationships and which functions are related though checking the edge and corresponding connectivity strength. In the second level, we expand the functional-level interaction map into a cross-function domain interaction map (see Figure 3 for an example) where each functional node is expanded into their corresponding domain vertices and edges that represent domain-domain interactions are unraveled across the different functional modules. These functional and domain interactions can help biologists to understand how the protein functions are connected and mediated by various domains in the biological systems.

3 RESULTS and DISCUSSIONS

In this work, we construct the cross-function domain interaction map in the species yeast. First, we discovered conserved domain interactions from protein interaction data of species yeast, worm and fly obtained from DIP (http://dip.doe-mbi.ucla.edu/). We obtained the corresponding domain information from the Pfam database (http://www.sanger.ac.uk/Software/Pfam). Figure 1 details the domain pair distributions in all the three species. In all, 29476 domain pairs occurred in at least one species (yeast 13052, worm 3839, fly 14511). The 214 conserved domain pairs that occurred in all the three species are used to build cross-function domain interaction map.

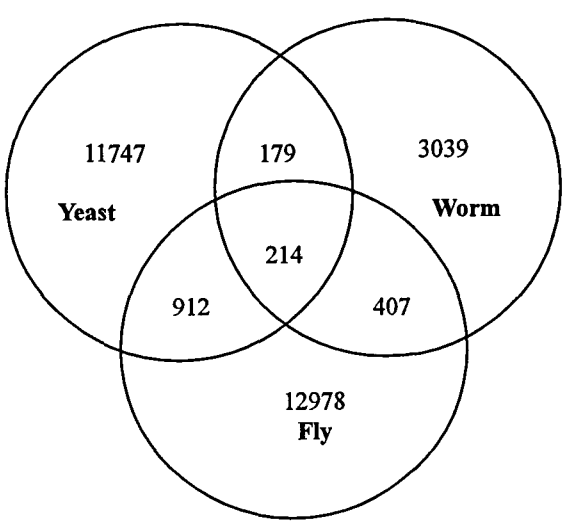

Figure 1. Domain pair distributions in three species yeast, fly and worm.

Next, we construct the functional domain interaction map to reveal the overall functional relationships shown in Figure 2. In this map, the vertices are the functional modules and the edges represent the functional relationships. As discussed in Section 2.2.1, we retain only the edges supported by our domain interaction set and the

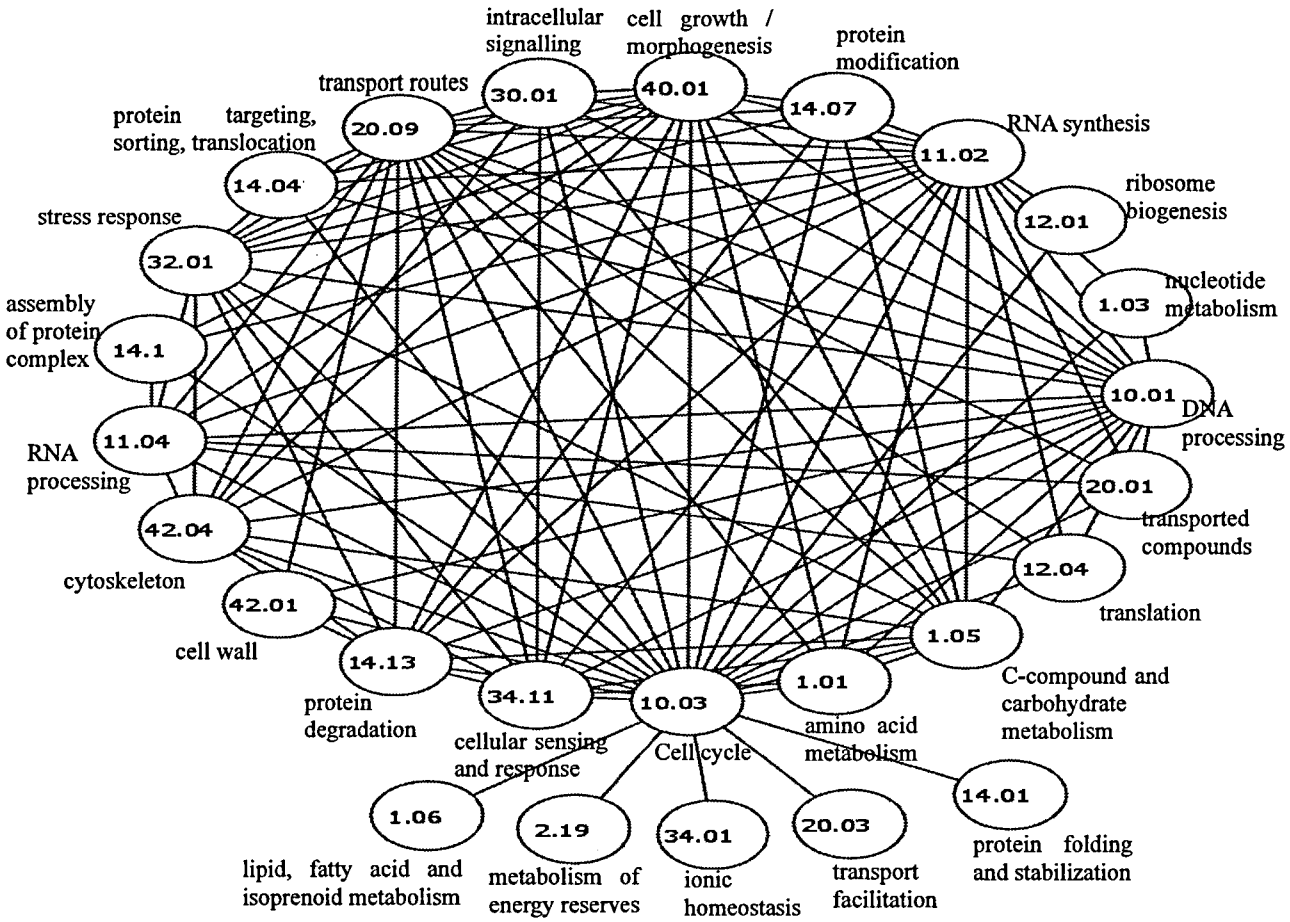

Figure 2. Functional level interaction map revealing the overall functional relationships. Numbers in the ovals are the MIPS functional ID (top two levels of MIPS "ftp://ftpmips.gsf.de/yeast/catalogues/funcat/")

connectivity strength $cs(f_u, f_v)$ is larger than δ (in this paper δ =50). This resulted in a total of 137 cross-function relationships among 26 MIPS function categories. Our inferred conserved domain interaction supported 110 (out of 137 or 80.29%) cross-function relationships. The resulting graph is highly connected with an average degree of 8.46 per node (MIPS function). Such a high connectivity suggested that the cellular functional organization is a highly compact one in which the biological activities of many functions are closely connected. Biologically, this makes sense as many cellular activities are indeed the results of complex multi-facetted interactions among different cellular components. Such a tightly-connected network facilitates efficient communication and synchronization of activities between different functional modules.

Among the cross-functional relationships, there are some with high connectivity strength. We listed the top 10 cross-functional relationships in Table 1 according to their connectivity strength (See Methodology section). For example, the function *Cell Cycle* (MIPS: 10.03) was strongly connected to *RNA Synthesis* (MIPS: 11.02) with 554 protein-protein interactions, followed by *Cell Cycle* (MIPS: 10.03) to *Cell Growth/Morphogenesis* (MIPS: 40.01) with 504 protein-protein interactions. The high connectivity strength implied that these functions are tightly related or connected in biological processes.

In addition to the high connectivity of the functional map, the map in Figure 2 also reveals that some functions may play a more central role as "*function hubs*". Such functions are characterized by its high connectivity to other functions such as the *Cell Cycle* function that is connected to all other functions. The top 10 most connected function hubs are listed in Table 2.

Function1	Function2	CS
RNA synthesis	Cell cycle	554
Cell cycle	Cell growth / morphogenesis	504
Transport	Cell cycle	478
Cell cycle	DNA processing	470
RNA synthesis	DNA processing	404
Cytoskeleton	Cell cycle	393
RNA synthesis	C-compound, carbohydrate Metabolism	388
Cell cycle	Cellular sensing and response	386
Cell cycle	C-compound, carbohydrate metabolism	385
Cell cycle	protein degradation	341

Table 1. Functional pairs with high connectivity strength (CS)

The functional interaction map suggests that the most central function is *Cell Cycle*, followed by *RNA Synthesis*, *DNA Processing* and *Transport Routes*, etc. Intuitively, such observations are coherent to current biological knowledge. With the exception of *Stress Response* and *Biogenesis of Cytoskeleton*, all functions are crucial for cell

MIPS	Descrption	*d*
10.03	Cell Cycle	25
11.02	RNA Synthesis	19
10.01	DNA Processing	18
20.09	Transport Routes	17
40.01	Cell Growth / Morphogenesis	15
32.01	Stress Response	13
1.05	C-Compound and Carbohydrate Metabolism	12
14.13	protein degradation	11
34.11	Cellular Sensing and Response	11
42.04	Biogenesis of Cytoskeleton	11

Table 2. The top 10 most connected function hubs.

growth and development which need the integrated efforts of many parts and functional components of a cell. *Cell Cycle* is a process where an active growing and dividing cell is constantly undergoing. Such process requires the participation of many other cellular processes such as synthesis of polynucleotide (*RNA Synthesis*, *RNA Processing*, *DNA Processing*) and the production of energy (*C-Compound and Carbohydrate Metabolism*), which are also listed in Table 2. *Cellular Sensing and Response* is also expected to be a central functional hub for biological processes as this function facilitates the many dynamic responses of a cell to the external environment by relaying signal to the necessary functions for response.

Finally, we constructed the entire cross-function domain interaction map. We present a sub-graph between 5 functional modules (*Cell Cycle*, *Intracellular Signaling*, *Ionic Homeostasis*, and *Ribosome Biogenesis* and *Protein Degradation*) in Figure 3. As expected, *Cell Cycle*, *Intracellular Signaling* and *Protein Degradation* functional modules are highly connected by multiple different domain-domain interactions. In the figure, we also observed that the domains PF07714 and PF00069 are network hubs that connect *Cell Cycle* to all other functional modules. These two domains are also found in the intracellular signaling module.

To get a better understanding on the domain interaction mechanisms of the cross-functional relationships, we also check out which domain pairs are often used to bridge the protein functions. Figure 4 shows the top domain pairs that are involved in multiple (i.e. more than 15 times) cross-functional interactions. Here the thickness of an edge is correlated with the number of cross-function relations which the domain pair is involved. Notice that the previously mentioned domains PF07714 (Protein tyrosine kinase) and PF00069 (Protein kinase domain) are shown to be mediating most cross-functional interactions. This analysis suggests that many functional relationships between different cellular functions are mediated by a core set of conserved domain interactions, especially through interacting with the domain PF07714 and PF00069.

As many of these domains are highly involved in cross-function interactions, we also hypothesize that they may be involved in the activities of highly connected protein hubs [4], since the hub proteins typically play

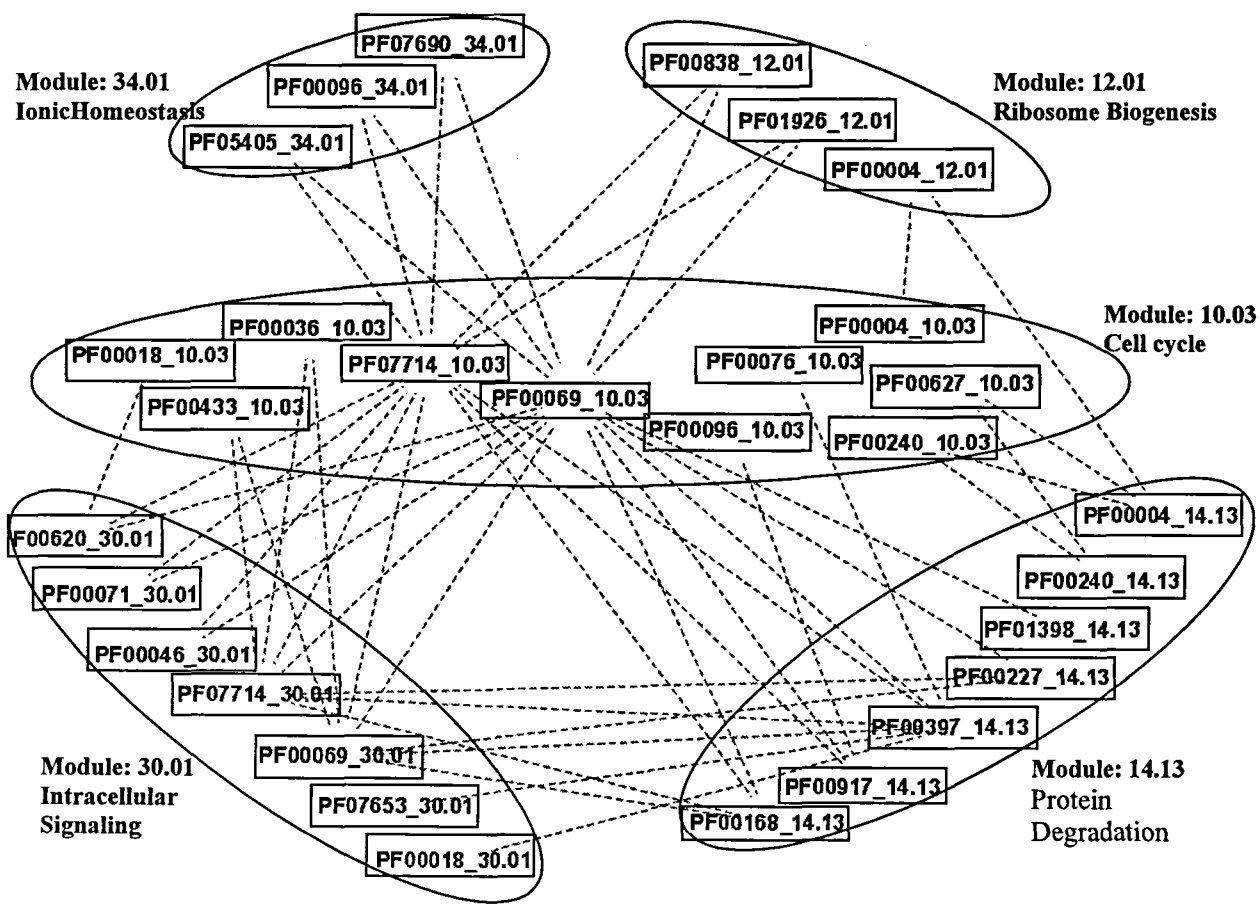

Figure 3. Cross-function domain interaction map. Each oval node represents a functional module and each rectangle node a domain labeled PFXXXXX_XX.XX where the first seven letters are Pfam domain ID and the last five are MIPS functional ID.

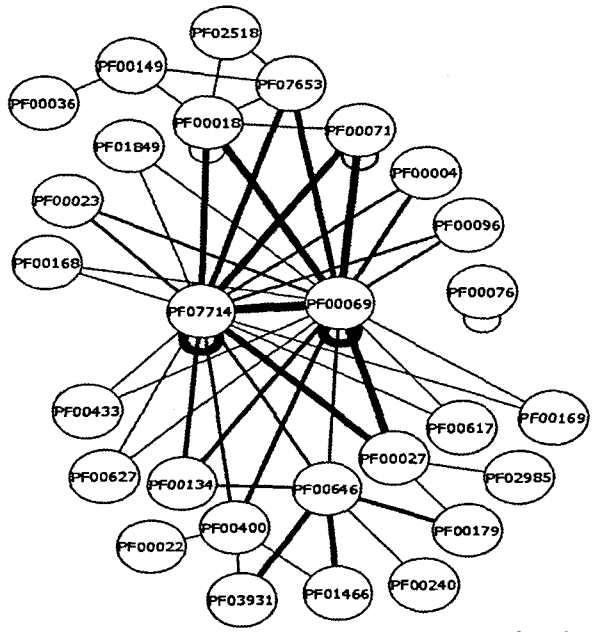

Figure 4. Domain hubs with frequent cross-function domain interactions

multiple roles or functions within a cell. To verify this, we also investigated the propensity for the conserved domain hubs to be found in hub proteins. Here we compared the occurrence of the conserved domain hubs (as shown in Figure 4) with 100 randomly selected non-hub domains in known hub proteins in yeast. Figure 5 depicts the relative proportions of proteins with interaction degree greater than k ($1<=k<=212$) containing the domains from these two groups. Compared to randomly selected domains, the conserved domain hubs in Figure 4 showed a marked tendency to be in hub proteins (proteins with high interaction degrees). Overall, the conserved domain hubs are ~15 times more likely to be found in protein hubs (proteins with at least 20 interactions) than a normal domain.

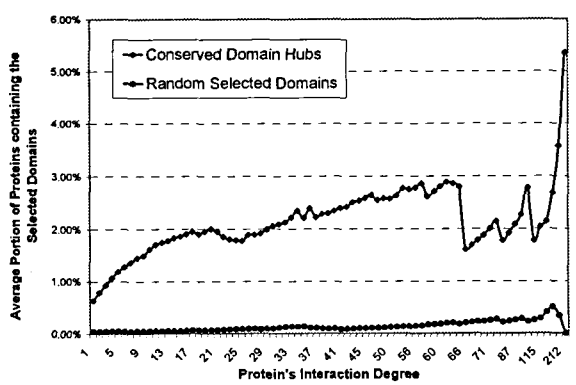

Figure 5. Conserved domains correspond to protein hubs

We further investigated the core domain hubs PF07714 and PF00069 which stood out prominently in Figure 4. We found that 9.21% of protein hubs (proteins with >=20 interactions) contained either PF07714 or PF00069, which is ~145 times than randomly selected domains. In other words, domains PF07714 and PF00069 showed a remarkable propensity to be in proteins hub. The above analysis suggests that the hubness of proteins in the cell could be mediated by a core group of evolutionarily conserved domains.

4 CONLUSIONS

In this paper, we have proposed a novel technique to discover the inter-functional organization of biological activities in a cell using existing protein-protein interaction datasets from multiple species. We have constructed a cross-function domain interaction map that revealed various useful insights about the cellular functional network and the underlying cross-talk mechanisms between different functional modules at the protein domain level. The domain interaction map revealed that most of the detected cross-functional relationships are supported by conserved domain interactions, and that a number of key domains ("domain hubs") played pivotal roles in these relationships. The cross-function domain interaction map also showed that the conserved domain hubs are associated with protein hubs in protein interaction networks. As living cells are sustained not by individual activities but rather by coordinated summative efforts of different biological functional modules, exploring the cross-function domain interaction map can thus help us better understand the many entwining intricacies of cellular processes.

REFERENCES

[1] M. Deng, F. Sun S. Metha, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12:1540–1548, 2002.

[2] J. Gagneur, R. Krause, T. Bouwmeester and G. Casari, Modular decomposition of protein-protein interaction networks, *Genome Biology*, 5 (8): R57, 2004.

[3] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, From molecular to modular cell biology. Nature 1999, 1402(Suppl):C47-C52.

[4] H. Jingdong, B. Nlcolas, H. Tong etc, Evidence for dynamically organized modularity in the yeast protein-protein interaction network, Nature, Vol 430: 88-93, 2004.

[5] S. K. Ng, Z. Zhang, and S.H. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19:923–929, 2003.

[6] Papin JA, Reed JL, Palsson BO: Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem Sci 2004, 29:641-647.

[7] D. Park, S. Lee, D. Bolser, M. Schroeder, M. Lappe, D. Oh, J. Bhak, Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interactome map), Bioinformatics, Vol 21 (15), p. 3234-3240, 2005.

[8] V. Spirin, L.A. Mirny. Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA, 100: 12123—12128, 2003.

[9] E. Sprinzak and H. Margalit, Correlated Sequence-signatures as markers of protein-protein interaction, *Journal of . Mol. Biol.*, 311, 681-692, 2001.

[10] C. Von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.