The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 75–88

Establishing Protein Functional Linkage in a Systematic Way*

Yong Wang[1] Rui-Sheng Wang[2,3] Xiang-Sun Zhang[1,†]
Luonan Chen[3,4,‡]

[1] Academy of Mathematics and Systems Science, CAS, Beijing 100080, China
[2] Renmin University of China, Beijing 100872, China.
[3] Osaka Sangyo University, Osaka 574-8530, Japan.
[4] Institute of Systems Biology, Shanghai University, Shanghai 200444, China.

Abstract Most gene products facilitate their functions within complex interconnected networks by interacting with other biomolecules. Thus elucidating protein functional relationships from their interacting neighbors is one of the challenging problems of the post-genomic era. High-throughput experiments such as genome-wide protein-protein interaction networks are expected to be fertile sources of information for deriving their functional relationships. However, a high rate of false positives and the sheer volume of the data are making reliable interpretation of these experiments difficult. In this work, we overcome these difficulties by using a network-based statistical significance analysis method that forms reliable functional associations between proteins. The basic mechanism is if two proteins share similar neighbors globally than random, they have close functional associations. Our method tries to establish a framework to explore the protein relationships by analyzing statistical significance of sharing similar global partnerships for all protein pairs in the interaction network. In this framework, many methods can be integrated to globally define and construct protein neighborhood from protein interaction data. Furthermore our framework can be applied directly to binary data, experimental strength data and integration data. Applying our framework in yeast protein interaction datasets and the shortest path to form protein neighborhood, our method is shown to be able to infer reliable functional linkages from experimental data which are verified by GO functions.

Keywords Functional linkage map; protein-protein interaction; GO function; statistical significance.

1 Introduction

Annotating protein functions is one of the most challenging problems of the post-genomic era. Traditionally, functional annotation of proteins can be summarized as sequence based approaches, structure-based approaches, motif-based approaches and "guilt-by-association" based approaches. The basic idea of them is that if protein P_a has function X and protein P_b is often "associated" by possessing sequence

*This work is partly supported by Grant No. 5039052006CB from the Ministry of Science and Technology, China, and National Natural Science Foundation of China under Grant No. 10471141, 60503004.
†Corresponding author. E-mails: zxs@amt.ac.cn.
‡Corresponding author. E-mails: chen@eic.osaka-sandai.ac.jp.

similarity, structure similarity or common motifs with protein P_a, the protein P_b might have a similar function related to X.

Generally, a biological function is facilitated not by the individual proteins but by the interactions or a concerted effect of those proteins. Furthermore the possibility to access protein's interaction patterns has attracted the attention from the study of single proteins or small complexes to that of the entire proteome, which makes it possible to annotate protein function by their neighbors in protein interaction network [1, 2, 3], with nodes representing proteins and edges representing the detected PPIs (protein-protein interactions). In this context, the research on reliable methods for revealing proteins' function relationships from experimental data is of uttermost importance. The reason lies in that there are many new proteins whose biological functions remain a mystery.

The framework of a 'functional linkage network' [4] is a promising step toward obtaining a detailed understanding of the functional relationships between proteins. In a typical functional-linkage network, each node corresponds to a protein, and an edge connects two proteins if some experimental or computational procedure suggests that these proteins might share the same function. Though such links reveal the important clues to relate proteins by their function similarity, they usually do not provide the detailed information on which specific functional annotation the proteins share. The single function annotation can be achieved based on the local or global neighbor information in functional linkage network [4, 5, 6, 7, 8].

Commonly there are two ways to form the functional linkage between proteins. One way is to integrate various biological experimental data and then to extract the information contributed to functional relationships. For instance two proteins might be linked if they share similar sequence, structure or gene ontology, test positive in a yeast two-hybrid screen, in the same protein complex or if their gene-expression patterns are correlated in several experimental conditions [7]. The other way is discovering reliable protein interactions from high-throughput experimental protein interaction data using network topology. For example in [9] a network-based statistical algorithm allows us to derive functions of unannotated proteins from large-scale interaction data. They hypothesize that if two proteins have significantly larger numbers of common interaction partners in the measured data set than what is expected from a random network, it would suggest close functional links between them. But in their method only local neighbors are considered thus it suffers from the highly errorneous high-throughput experimental methods, such as yeast-two-hybrid or tandem affinity purification, which have been reported high false positive rates. It will lead to potentially costly spurious discoveries.

To utilize the global information in the protein interaction data to get reliable functional relationships, many methods are proposed to redefine the reliable interaction relationship between two proteins, such as shortest path, alternative path or diffusion kernel [2, 5]. They all show that global information in protein-protein interaction network is useful to discovering reliable protein interaction. But no statistical significance analysis is preformed based on the interaction from global information.

In this paper, we establish a unified framework to explore the protein relationships by analyzing statistical significance of sharing similar global partnerships for all protein pairs in the interaction network. In this framework, many methods to globally bridge protein pairs by protein interaction data can be integrated, such as shortest path, domain level inference[10, 11] and diffusion kernel. Furthermore we deal with the protein-protein interaction data as real numbers instead of binary values. It means that the interaction relationship of two proteins are assessed by a probabilistic score. The reason lies in that strength data have more information and it is more reasonable to consider the noise in biological data.

2 Methods and materials

We represent the evidence from protein-protein functional relationships using a graphical formalism called a functional linkage graph in which an edge or link between two nodes (proteins) represents evidence that they might share the same function. The translation of experimental data containing information of protein relationships into a functional linkage graph is straightforward based on the observation that if two proteins share similar neighbors globally than random, they have close functional associations. In this section we will introduce the shortest path method as an example to form the global neighborhood for every protein by protein interaction data and then describe the statistical significance analysis framework.

Mathematically, the functional linkage network is an undirected graph $G = (V,E)$, where the node set $V = \{P_1, P_2, \cdots, P_n\}$ is all the proteins concerned. The edge set is $E = \{e_{ij}, \quad i, j = 1, 2, \cdots, n\}$ and e_{ij} denotes the strength of functional relationship between proteins P_i and P_j. To construct such a functional linkage network, our strategy is firstly to represent every protein (node of G) by a vector considering all its neighbor relationships with other proteins, then a statistical method is applied to derive the score to assess the strength of functional relationship between two proteins.

As an example, we deal with protein interaction data to show how to build functional linkage by utilizing global information. Similarly protein interaction network is expressed by $G = (V, \Phi)$, here $\Phi = \{\rho_{ij}, \quad i, j = 1, 2, \cdots, n\}$ and ρ_{ij} denotes the interaction strength of proteins P_i and P_j ($\rho_{ij} = 0, 1$ for the binary data) [12]. Then the key is the translation of protein-protein interaction data (real type or binary type) into a functional linkage score for every protein pair.

2.1 Representing protein by its neighborhood information

With the protein interaction network $G = (V, \Phi)$, global information is extracted from the network topology to represent protein by the relationships with other proteins. Given an appointed protein, it is naturally to consider its neighbors to get the functional information because proteins facilitate their functions by interconnected macromolecules. The problem is how to define the neighborhood relationship between proteins. Protein interaction network provides natural neighborhood by its graph representation $G = (V, \Phi)$. In such a graph, a set of proteins connected

to the appointed protein (physically interacting) are defined as 'neighbors'. Previous methods for protein function prediction are almost based on this neighborhood [7, 8, 13, 14]. Recently instead of using direct interaction neighbor, non-directly interacting neighbors at different levels are also incorporated to give more clues of functional linkage [2, 6, 15].

In this paper, we will focus on the transitive functional relationship by defining remote neighborhood. The reason is that the neighborhood in protein interaction network is not so reliable due to many false positive data. As indicated by a rigorous comparative analysis and performance assessment [12], common technologies like yeast two-hybrid may experience high rates of false positive detection and it is necessary to associate confidence scores with protein interactions. Also from viewpoint of network topology, protein interaction network is sparse and found to be small-world, scale-free and modular [1, 16]. Thus it is often hard to build long-range relationships with other proteins and can only provide limited information to infer function linkage. Therefore we define new neighborhood relationships from protein interaction data by using the shortest path in this paper. In the same manner, other methods like domain level inference and diffusion kernel can be easily incorporated to provide extensive relationships between proteins and we will report these results in another paper.

2.2 The shortest path analysis

In [5], the shortest path analysis is used to identify transitive genes between two given genes from the same biological process. By computing shortest path (SP), not only functionally related genes with correlated expression profiles were identified but also those without.

We use the following Floyd-warshall algorithm to identify the SPs between a source protein to all other proteins in the protein-protein interaction network. The Floyd-warshall algorithm is an algorithm for solving the all-pairs shortest path problem on weighted, directed graphs in cubic time. i.e. the time complexity is $O(n^3)$ where n is the node number of the network. A detailed introduction to Floyd-warshall algorithm can be found in [17] and other algorithm textbooks. It should be noticed that the SP model here is scalable to larger graphs. Because for a given graph, the computation needs to be done only once and is easily distributed over multiple processors for parallel implementation.

In our computation, $D_{ij} = 1/\rho_{ij}$ can be replaced by any decreasing function of the protein interaction strength. For the negative score of protein-protein interaction strength, we simply set them to zero in the shortest path computation. $\forall(P_i, P_j) \in V(G) \times V(G)$, the new similarity measure is $1/D_{ij}, 0 < D_{ij} < 1$. By computing the shortest paths of all pairs of two proteins, we can represent a protein by its shortest path neighborhood. For example, protein P_i is represented by a vector of its shortest paths to other proteins in the network as $P_i = (D_{i1}, D_{i2}, \cdots, D_{i,i-1}, D_{i,i+1}, D_{in})$.

Algorithm 1 Floyd-warshall algorithm for PPI network

$D_{ij} = 1/\rho_{ij} \quad (i,j) \in \phi(G) \quad n = |V(G)|$
$D_{ij} = \infty \quad (i,j) \in (V(G) \times V(G)) \backslash \phi(G), i \neq j$
$D_{ii} = 0 \quad i = 1,2,\cdots n$

```
for j = 1 to n do
    for i = 1 to n do
        if i ≠ j then
            for k = 1 to n do
                if k ≠ j then
                    if D_ik > (D_ij + D_jk) then
                        D_ik = (D_ij + D_jk)
                    end if
                end if
            end for
        end if
    end for
end for
```

2.3 Statistical significance analysis

In the traditional model, only direct interactions of proteins are considered. In this study, we further incorporate all neighborhood information. The basic idea is to understand a protein's function based on information on all the neighbor proteins. As the first step, we build the new global representation by the shortest path, diffusion kernel, domain interaction or other methods. Thus every protein is denoted by a vector in which every element is the relationship with other proteins. For a pair of proteins $(P_i, P_j) \in V(G) \times V(G), i < j$, we have global representations

$$(p_{i,1}, p_{i,2}, \cdots, p_{i,i-1}, p_{i,i+1}, p_{i,j-1}, p_{i,j+1}, p_{i,n}, p_{ij})$$
$$(p_{j,1}, p_{j,2}, \cdots, p_{j,i-1}, p_{j,i+1}, p_{j,j-1}, p_{j,j+1}, p_{j,n}, p_{ij})$$

The similarity between two n dimensional vectors is computed by their inner product as follows:

$$S_{ij} = \sum_{k=1,k\neq i,j}^{n} p_{ik} p_{jk}$$

As stated by the central limit theorem, if $X_1, X_2, \cdots, X_N$ are a set of N independent random variables and each X_i has an arbitrary probability distribution $p(x_1, x_2, \cdots, x_N)$ with mean μ_i and a finite variance σ_i^2, then the normal form variable

$$X_{norm} = \frac{\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \mu_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}}$$

has a limiting cumulative distribution function which approaches a normal distribution when N is sufficiently large. Since protein interaction networks in living organ-

isms are generally in large scale and involve a lot of proteins, for example, there are about 6000 proteins in yeast database and 50000 proteins in human database. Furthermore for the newly defined neighborhood, two neighbors of an appointed protein can be assumed to be independent. Thus score S_{ij} approximately obeys the normal distribution (which will be shown by a real example with about 1507 proteins in Figure 1) and its mean and standard variance are μ and σ respectively. Therefor we can define a Z-score for the assessment of statistical significance of function similarity between two proteins by

$$Z_{ij} = \frac{S_{ij} - \mu}{\sigma}$$

where Z_{ij} is the Z-score for protein P_i and protein P_j sharing the same function. S_{ij} is the score computed by 2.3. μ and σ are the average and standard deviation respectively from the normal distribution obtained by central limit theorem. A negative Z-score indicates that share of common function of a particular protein pair is less possible than expected by random chance. A positive Z-score indicates that two proteins are more likely to share similar function than expected at random. A score near zero indicates that the possibility is at the level near that expected by random.

3 Experimental Results

We proposed to establish functional relationships of proteins from their neighbors in the network of physical interactions considering their strengths, by assessing the statistical significance of interacting proteins with similar functions. The function similarity is based on a global scale and depends on the entire connectivity pattern of the protein network. In this section, our method is applied to the yeast Saccharomices Cerevisiae protein-protein interaction network. Effectiveness and efficiency have been tested in real biological dataset and the accuracy is assessed by showing the correlation with GO function similarity and by discovering meaningful functional module.

To benchmark our method, we use the recent protein interaction data [21] which is the first genome-wide screen for complexes in a model organism, budding yeast, using affinity purification and mass spectrometry. Their approaches explicitly avoided to define protein relationships from binary interactions which are not deemed appropriate as these are not directly inferable from purifications. They derived a 'socio-affinity' score that quantifies the propensity of proteins to form partnerships. It measures the log-odds of the number of times that two proteins are observed together, relative to what would be expected from their frequency in the data set. Generally, pairs with socio-affinity indices below 5 should be considered with caution and protein pairs with high socio-affinity indices are more likely to be in direct contact as measured either by three-dimensional structures or the yeast two-hybrid system.

The Gavin's core dataset has 1507 nodes and 70647 strength links. We established functional linkages by our systematic method. If we set the Z-score threshold as 1.65 which means that the P-value is 0.05, then there are 40419 functional links. If the Z-score is 3.08 and the P-value is 0.01, we have 5909 links. To show the central

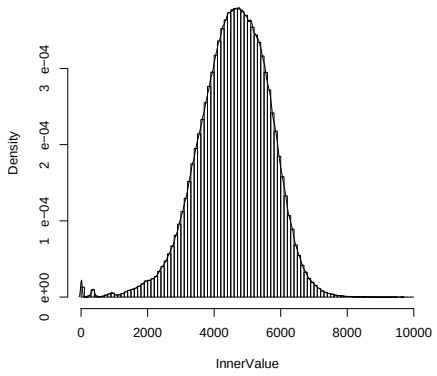

Figure 1: The distribution of score S_{ij} for Gavin's core dataset

limit theorem is properly used, we present the distribution of the score S_{ij} of Gavin's core dataset in Figure 1, which clearly verified the assumption of normal distribution of the score S_{ij}.

Next we show that the Z-score derived from Gavin's protein interaction data correlates well with GO function similarity and can aid to discover meaningful functional modules.

3.1 Correlation with the GO function

The purpose of this paper is to establish the protein functional linkage network by exploring the functional similarity of protein pairs by a statistical score. It is straightforward to test if or not the defined Z-score can correlate well with the protein function similarity score. A particular gene product can be characterized with different types of functions, including molecular function at the biochemical level (e.g. cyclase or kinase, whose annotation is often more related to sequence similarity and protein structure) and the biological process at the cellular level (e.g. pyrimidine metabolism or signal transduction, which is often revealed in the high-throughput data of protein interaction and gene expression profiles). In our study, function annotation of protein is defined by the GO biological process. The GO biological process ontology which is widely used as function benchmark and is available at `http://www.geneontology.org`. It has a hierarchical structure with multiple inheritances.

We use GO biological process classification, as of Oct. 2006, to assign function to unannotated proteins in the study. There are several novel measures [18] that can be used to assess the the similarity of two gene products based on the GO terms describing them. In this paper we adopt a simple and easy method used successfully in [7, 19].

When quantifying the similarity between two GO terms, it is desired that both their commonality and individual specificities can be captured simultaneously. Let G_s and G_t be the subgraphs induced from two GO terms T_s and T_t, and R_s and R_t be

the set of GO items which form the paths of G_s and G_t, respectively. We define the similarity of two GO terms $s(T_s, T_t)$ as:

$$s(T_s, T_t) \equiv \max_{R_s \in G_s, R_t \in G_t} |R_s \cap R_t|$$

.

The score means that the higher the number of terms shared by R_s and R_t, the more similar for the two GO terms to describe proteins. Since a protein may be involved in more than one biological process, it may be assigned with multiple GO terms. Let $T(P_i)$ denote the set of all the GO terms assigned to a protein P_i. Thus the functional similarity of two proteins are defined by the GO similarity for a pair of proteins P_i and P_j as the maximum similarity of all possible combinations of $T(P_i)$ and $T(P_j)$, i.e.

$$S_{GO}(P_i, P_j) \equiv \max_{T_s \in T(P_i), T_t \in T(P_j)} s(T_s, T_t)$$

.

To make the measurement of protein function similarity practical, we assign the biological process functional annotation for the known proteins along by a GO Identification (ID). we generated a numerical GO INDEX, which represents the hierarchical structure of the classification. The more detailed level of the GO INDEX, the more specific is the function assigned to a protein. The maximum level of GO INDEX is 14. In general, the function similarity between proteins P_x and P_y is defined by the maximum number of index levels from the top shared by P_x and P_y. The smaller the value of function similarity, the broader is the functional category shared by the two proteins.

In Figure 2 we show correlation relationship between our Z-score with the above defined protein GO similarity score. To highlight the advantage of using global interaction instead of local interaction information in the protein interaction network, we also compute the correlation of Gavin's socio-affinity score with the GO function similarity score for Gavin's dataset and compare with our result. The tendency is clearly that our score correlates better with functional similarity score in general. Especially when our Z-score increases from -0.5 to 4.32, the GO functional similarity score increases also from 8 to 14 linearly. On the other hand, the Gavin's socio-affinity score does not correlate well when it is higher than 5, which demonstrates the effect of false positive data in high-throughput protein interaction data. The detailed pearson correlation coefficient of Gavin's socio-affinity score with GO similarity is 0.541078 (1408 random samples). By using our statistical framework to integrate the non-direct neighborhood information in protein interaction network, we can improve the pearson correlation coefficient between Z-score and GO similarity to 0.6696 (1000 random samples).

3.2 Identifying Functional Modules

A functional module is defined as a group of genes or their products which are related by one or more genetic or cellular interactions, e.g. co-regulation, co-expression or membership of a protein complex, of a metabolic or signaling pathway

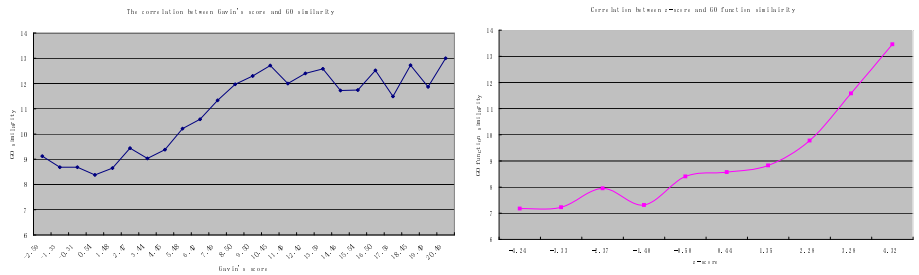

Figure 2: Correlation analysis. SubFig1: the correlation of Gavin's socio-affinity score with the GO function similarity score for Gavin's dataset. The pearson correlation coefficient is 0.541078 (1408 random samples). SubFig2: The correlation of the Z-score with the GO function similarity score for Gavin's dataset. The pearson correlation coefficient is improved to 0.6696 (1000 random samples).

or of a cellular aggregate (e.g. chaperone, ribosome, protein transport facilitator, etc.). An important property of a module is that its function is separable from other modules and that its members have more relations among themselves than with members of other modules, which is reflected in the network topology. The separability may stem from, for example, cellular localization or special interaction of proteins or special regulation of genes. Modules can be understood as a separated substructure of a network or pathway, e.g. the complex of fatty acid synthetase subunits may serve as an example of a module of the fatty acid biosynthesis pathway and the protein complex is a module of a protein interaction network [20].

To find out the underlying modular structure in networks, i.e., structural subunits (communities or clusters) characterized by highly interconnected nodes, the modularity score Q has been introduced as a measure to assess the quality of clusterizations [22]. Highly effective approach is proposed in [22] to optimize the quality of modularity Q over all possible divisions of a network. We use this method to decompose the constructed network into modules and reveal the patterns of the modularity in the predicted functional linkage network.

As a result we found eight modules in the functional linkage network which means groups of cellular components and their dense relationships that can be attributed to a specific biological function. Two of them are shown in Figure 3, where different protein functions are indicated by different vertex colors. The module shown in the left subfigure contains 18 proteins with 69 functional links. All the proteins are annotated by the protein biosynthesis function. On the other hand, the module in the right subfigure contains 75 proteins with 255 functional links. The proteins in this group are annotated mostly by transcription and DNA repair function.

Next we use the biological functional modules to predict or annotate the functions of unknown proteins. With the detected modules, simple methods are usually used for function prediction within the modules. For example, every function shared

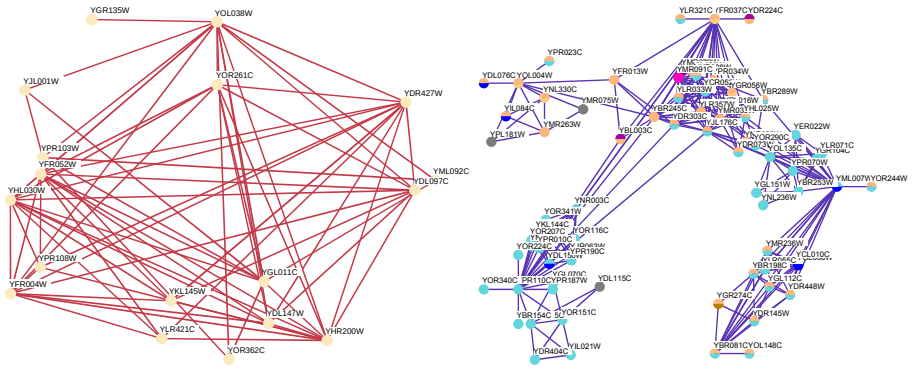

Figure 3: Two functional modules in the predicted functional linkage network revealed by the statistical significance Z-score.

by the majority of the module's genes is assigned to all the genes in the module. Alternatively a hypergeometric enrichment P-value is computed for every function. The functions enriched within the module (i.e. obtaining P-value below some threshold) are then predicted for all the genes in the module. In the right subfigure of Figure 3, the protein YDL115C is predicted to have the transcription function because it belongs to a 25-protein submodule and other group members within this submodule all have transcription function. In the same manner, proteins YPL181W and YMR075W are annotated by DNA repair function.

Here we stress that the established protein functional linkage network provides a good and solid basis for functional annotation which is a fundamental problem in the post-genomic era. The availability of functional linkage network instead of raw interaction network will spur on the development of computational methods to elucidate protein functions in a more accurate way. Current computational approaches for the task based on protein interaction network are able to include direct methods, which propagate functional information through the network and module-assisted methods and infer functional modules within the network and use those for the annotation task.

4 Discussion

Many approaches aimed to deduce the unknown function of a class of proteins have exploited sequence similarities or clustering of co-regulated genes, phylogenetic profiles, protein-protein interactions and protein complexes. In this paper we established protein functional relationships in a systematic way. We extract functional information from the physical interactions between proteins from the viewpoint of systems biology. Although most of the existing methods assume that protein-protein interaction data are given as binary data (i.e. whether or not each protein pair interaction is given), multiple experiments are performed for the same protein pairs in practice and thus the ratio of the number of observed interactions to the number of

experiments is available for each protein pair. For example, Ito et al [23] performed multiple experiments for each of protein-protein pairs. But, the results are not always the same for the same pair. Therefore, it is reasonable to use the ratio of the number of observed interactions to the number of experiments as input data, where the ratio is also referred to as the strength in this paper. Clearly in the procedure of converting strength data to binary data, certain information is lost in addition to man-made noise.

Another source of the protein interaction comes from the integration of many existing protein interaction datasets. It is well known that there are many kinds of experimental methods for physical protein interactions. They overlap or conflict in some cases. Also other high-throughput experimental data such as microarray and ChIP-chip can provide genetic interaction relationships. To discover the true topology of protein interaction network, these datasets are often integrated by all kinds of methodologies in the systems biology framework. As a result every protein interaction relationship is assigned with a probability score. These scores indicate the reliability of the protein interactions and can be viewed as strength or ratio of interaction in our computation.

Functional associations that are derived from a genomic context do not necessarily imply a direct physical interaction between two molecules. Proteins at opposite ends of a single pathway or complex can give the same signal as those in tight, direct, physical contact. Moreover, errors in the underlying genome or expression data can also lead to false prediction or to interaction being missed. To overcome these problems, several groups are developing methods to combine several types of interaction data quantitatively, which also consider the accuracy of each dataset. The result is an overall confidence score for each interaction, and higher scores are more likely to indicate direct physical contacts. We can incorporate the integrated confidence score into our statistical framework to find more meaningful biological functional linkages. The challenge for the future is how to incorporate more biological information from such diverse data sources.

We also found that the newly established protein functional linkage map is denser than raw protein interaction network. This result is quite reasonable. Except the direct physical protein interaction, functional links can be formed by genetic interaction or co-member relationship in the same protein complex. Furthermore we found transitive phenomenon in the functional linkage network. Intuitively this refers to situations where two proteins do not interact, but they carry out similar functions in biological process. Deep exploration indicates that both proteins strongly interact with the same set of other proteins. In this simplest case, proteins P_a and P_b strongly interact as same as proteins P_b and P_c do. However proteins P_a and P_c do not have strong interaction. Hence, for such a case, we say that proteins P_a and P_c transitively interact and the protein P_b serves as the transitive protein. In such as way, this phenomenon makes protein functional linkages more abundant than directly interaction.

In this paper, we use only experimental protein interaction data to reveal the protein functional relationship. Our method can be easily extended to other kind

of experimental data, such as sequence or structure similarity, ChIP-chip interaction data, microarray gene expression data, protein subcellar localization data and integration of heterogeneous data sources. Since the emphasis of this paper is not data integration, we only focus on the effectiveness and efficiency of the statistical framework. Furthermore other neighborhood discovering methods such as domain level inference[10, 11] and diffusion kernel can be utilized to provide remote neighborhood information in the framework besides the shortest path, we will study this topic in another paper.

5 Conclusion

Interpreting data from large-scale protein interaction experiments has been a challenging task because of the widespread presence of random false positives. Here, we presented a network-based statistical algorithm that overcomes this difficulty and allows us to derive functions of unannotated proteins from large-scale interaction data. The algorithm uses the insight that if two proteins share significantly strong interaction and larger number of common interaction partners (including long-range partners) than random, they have close functional associations. Analysis of publicly available data from Saccharomyces cerevisiae reveals that our method can find reliable functional associations.

As described in the paper, our main contributions in this paper are summarized as follows:

1. A new statistical analysis framework to build protein functional linkage is introduced. It is a general framework which can be further utilized to deal with integration of heterogeneous data sources.
2. New neighborhood relationships can be defined and global information is utilized to analyze the statistical significance of similar function.
3. The method can deal with the experimental protein interaction data by considering the strength of interactions. Also the binary data also can be dealt as a special case.

References

[1] Albert-László Barabási, Network biology: Understanding the cell’s functional organization. Nature Reviews Genetics, 5, 101–113, 2004.

[2] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[3] Roded Sharan, Igor ULitsky and Ron Shamir. Network-based prediction of protein function. Molcular Systems Biology, 3:88, 2007.

[4] Ulas Karaoz, T. M. Murali, Stan Letovsky, Yu Zheng, Chunming Ding, Charles R. Cantor, and Simon Kasif. Whole-genome annotation by using evidence integration in functional-linkage networks. *PNAS*, 101(9):2888–2893, 2004.

[5] Xianghong Zhou, Ming-Chih J. Kao, and Wing Hung Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *PNAS*, 99(20):12783–12788, 2002.

[6] A. Vazquez, A. Flammini, A. Maritan, and Vespignani A. Global protein function prediction in protein-protein interaction networks. *Nature Biothech.*, 21:697, 2003.

[7] Yu Chen and Dong Xu. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucl. Acids Res.*, 32(21):6414–6424, 2004.

[8] Mustafa Kirac, Gultekin Ozsoyoglu, and Jiong Yang. Annotating proteins by mining protein interaction networks. *Bioinformatics*, 22(14):e260–270, 2006.

[9] Manoj Pratim Samanta and Shoudan Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *PNAS*, 100(22):12579–12583, 2003.

[10] Luonan Chen, Ling-Yun Wu, Yong Wang, and Xiang-Sun Zhang. Inferring protein interactions from experimental data by association probabilistic method. *Proteins*, 62:833–837, 2006.

[11] Shihua Zhang, Guangxu Jin, Xiang-Sun Zhang, Luonan Chen. Discovering functions and revealing mechanisms at molecular level from biological networks, *Proteomics*, in press, 2007.

[12] Silpa Suthram, Tomer Shlomi, Eytan Ruppin, Roded Sharan, and Trey Ideker. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7(1):360, 2006.

[13] Stanley Letovsky and Simon Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(suppl1):i197–204, 2003.

[14] Jordi Espadaler, Ramon Aragues, Narayanan Eswar, Marc A. Marti-Renom, Enrique Querol, Francesc X. Aviles, Andrej Sali, and Baldomero Oliva. Detecting remotely related proteins by their interactions and sequence similarity. *PNAS*, 102(20):7151–7156, 2005.

[15] Hyunju Lee, Zhidong Tu, Minhua Deng, Fengzhu Sun, and Ting chen. Diffusiton kernel based logistic regression models for protein function prediction. *OMICS: A Journal of Integrative Biology*, 10(1):40–55, 2006.

[16] András Szilágyi, Vera Grimm, Adrián K Arakaki, and Jeffrey Skolnick. Prediction of physical protein-protein interactions. *Physical Biology*, 2(2):S1–S16, 2005.

[17] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency, Volume A*. Springer, 2003.

[18] Mihail Popescu, James M. Keller, and Joyce A. Mitchell. Fuzzy Measures on the Gene Ontology for Gene Product Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 03(3):263–274, 2006.

[19] Hongwei Wu, Zhengchang Su, Fenglou Mao, Victor Olman, and Ying Xu. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucl. Acids Res.*, 33(9):2822–2837, 2005.

[20] Sabine Tornow and H. W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucl. Acids Res.*, 31(21):6283–6289, 2003.

[21] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.

[22] Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**(23), 8577–8582, 2006.

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