Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Modularization of Protein Interaction Networks by Incorporating Gene Ontology Annotations

Young-Rae Cho, Woochang Hwang and Aidong Zhang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260, USA
Email: {ycho8,whwang2,azhang}@cse.buffalo.edu

Abstract— **Recent computational analyses of protein interaction networks have attempted to understand cellular organizations, processes and functions. However, they have encountered difficulties due to unreliable interaction data and the complexity of the networks. In this paper, we propose the integration of protein interaction networks with Gene Ontology annotations for assessing the reliability of current protein-protein interaction data. The interaction reliability can be used for building weighted protein interaction networks. We apply an information flow-based modularization algorithm to the weighted protein interaction networks. Our experimental results show that the interaction reliability between two proteins is positively correlated to the likelihood of functional and locational associations. We finally demonstrate that our approach identifies accurate modules in the protein interaction networks with high statistical confidence with respect to biological function and cellular localization. Moreover, this algorithm outperforms our previous method [5] integrating with genetic co-expressional profiles.**

I. Introduction

The complete and systematic analysis of protein-protein interactions is one of the most fundamental challenges in understanding cellular organizations, processes and functions. The interactions potentially provide useful insights into functional associations between proteins [11]. Recent large-scale experimental methods, such as two-hybrid systems [15], [24] and mass spectrometry [8], [12], have led to the accumulation of vast quantities of interaction data, which can build complete protein interaction networks.

Various computational approaches have attempted to discover functional information from protein interaction networks. However, they have encountered difficulties due to unreliable interaction data and the complexity of the networks. It is known that the large-scale experiments have yielded numerous false positives [25], which mean that the interactions do not occur in real living tissues. Most of the previous works had a limitation in accuracy because of a large number of false connections in protein interaction networks.

To avoid inaccuracy resulting from the false connections, we can extend the unweighted protein interaction network to a weighted graph [2] by assigning a weight to each edge. We define the weight of an edge as the reliability of the corresponding interaction, that is, the probability of the interaction being a true positive. Other biological knowledge can be used to verify the protein-protein interactions. For example, we can deal with Microarray expression data for this purpose. The existence of the correlation between protein-protein interactions and mRNA expression profiles has been shown in [16], [17]. Several recent studies [4], [22] have integrated protein interaction networks with genetic co-expressions to evaluate the strength of the experimental interaction data.

To accurately assess the reliability of protein-protein interactions, we explore Gene Ontology (GO) [10] and its annotations. Ontology represents the knowledge in which concepts are described by their meaning and the relationships to each other [1]. Gene Ontology (GO) is currently one of the most comprehensive ontology databases in bioinformatics community. It provides GO terms and their relationships. GO terms are the shared biological concepts across different organisms. The relationships include the specific-to-general and part-to-whole relations between two GO terms.

Genes and gene products have been manually annotated on each GO term. The GO annotations can be useful resources for comparative functional analysis of the genes and gene products. Recently, the GO annotation data has been incorporated into the analysis of Microarray expression data for predicting biological process [14], measuring the distance of genes to detect clusters [13], and estimating missing expression values [23]. It has been also integrated with protein-protein interaction data to predict biological functions of uncharacterized proteins [6].

In our earlier study [5], we proposed the information flow-based approach to identify modules in weighted protein interaction networks. We used the genetic co-expressions for computing the weight for each interaction. In this paper, we integrate a protein interaction network with GO annotations. Based on the connectivity between two proteins in a network and the annotated proteins on each GO term, we provide a novel measurement for computing the interaction reliability, which is assigned as a weight to each edge in a protein interaction network. We then apply the information flow-based approach [5] for modularization of the weighted network. The overlapping modules can be detected by the information flow simulation starting from each essential node. We finally demonstrate that this approach identifies accurate modules with higher statistical confidence than other methods, including our previous work [5], with respect to biological function and cellular localization.

1-4244-0710-9/07/$20.00 ©2007 IEEE

II. METHOD

A. Integration of Protein Interaction Networks with GO Annotations

The protein interaction network is generally represented by an undirected, un-weighted graph $G(V, E)$ with proteins as a set of nodes V and interactions as a set of edges E. $N(v_i)$ denotes the neighbors of v_i, which mean a set of nodes connected to the node v_i. The degree of v_i is then equivalent to the number of neighbors of v_i, $|N(v_i)|$. We assign a weight to each edge for building a weighted network. The weight $w_{i,j}$ indicates the reliability of the interaction between v_i and v_j, which represents the probability of the interaction being a true positive.

The Gene Ontology (GO) project [10] is a collaborative effort to provide consistent description of genes and gene products. GO is a repository of biological knowledge in a computational format. GO provides a collection of well-defined biological terms, which is called GO terms. The GO terms are shared across different organisms. They comprise three categories as the most general concepts: biological processes, molecular functions and cellular components. The GO terms are structured by the relationships to each other, such as "is-a" and "part-of". For example, if a GO term g_i has a relationship "is-a" to g_j, then the term g_i is more specific than g_j. A DAG (Directed Acyclic Graph) is then formed with the GO terms as a set of nodes and their relationships as a set of directed edges. For the relationship "is-a" or "part-of" of g_i to g_j, g_i is a child node of g_j and g_j is a parent node of g_i in a DAG structure.

Genes and gene products are annotated on each GO term. The GO annotations follow the transitivity property, which means if a gene is annotated on a GO term, then it is also annotated on more general GO terms on the path from the GO term to the root. Thus, the set of the annotated genes on a GO term g_i is a subset of the annotated genes on a parent node of g_i. Also, the root GO term has the largest number of annotated genes, while a leaf GO term has the smallest number of annotated genes among the GO terms on the path from the root to the leaf.

Suppose a protein v_j is annotated on k different GO terms. $S_t(v_j)$ in the range of $1 \leq t \leq k$ denotes a set of annotated proteins on the GO term g_t, whose annotation includes v_j. We define the *interaction coverage* of a protein v_i to $S_t(v_j)$, based on the connectivity of v_i in a protein interaction network.

Definition 1. The *interaction coverage* of v_i to a set $S_t(v_j)$ is the intersection between $S_t(v_j)$ and a set of neighbors of v_i including v_i itself:

$$C_t(v_i, v_j) = S_t(v_j) \cap N'(v_i), \tag{1}$$

where $N'(v_i) = N(v_i) \cup \{v_i\}$.

We then compute the probability $P(v_i, v_j)$ of a protein v_i interacting with v_j as:

$$P(v_i, v_j) = \max_{1 \leq t \leq k} |C_t(v_i, v_j)|/|N'(v_i)|, \tag{2}$$

where k is the number of the GO terms, on which v_j is annotated. $P(v_i, v_j)$ indicates the probability of v_i interacting with the annotated proteins on the GO terms. Since v_j can be annotated on several different GO terms, we use the maximum size out of k possible *interaction coverage*. If v_i and its neighbors are not annotated in $S_t(v_j)$ with any t, then $P(v_i, v_j)$ is 0. If all of them are annotated in $S_t(v_j)$, then $P(v_i, v_j)$ is 1. Equation 2 thus satisfies the range of $0 \leq P(v_i, v_j) \leq 1$.

We finally compute the reliability of the interaction between v_i and v_j using the geometric mean of $P(v_i, v_j)$ and $P(v_j, v_i)$. The reliability can be assigned to the corresponding edge as the weight $w_{i,j}$.

$$w_{i,j} = \sqrt{P(v_i, v_j) \times P(v_j, v_i)}. \tag{3}$$

Since we consider the interaction reliability between two proteins, v_i and v_j, having experimental evidence of interactions, v_i is one of the neighbors of v_j and v_j is one of the neighbors of v_i in a network. Then, both $P(v_i, v_j)$ and $P(v_j, v_i)$ are always greater than 0. The weight $w_{i,j}$ is thus calculated in the range of $0 < w_{i,j} \leq 1$.

This method measures the interaction reliability by the integration of the connectivity in a protein interaction network with GO annotations. Since GO terms are described in terms of biological processes and functions, the interaction reliability can properly quantify the functional associations between two proteins.

B. Modularization of Weighted Protein Interaction Networks

In order to identify modules from the weighted protein interaction network, we apply the information flow-based modularization approach from our previous study [5]. A module represents a maximal set of proteins that have correlated behaviors with respect to a specific biological feature. For example, if a group of proteins has the same functional behavior, then they form a functional module. The module in a graph G can be identified as a sub-graph G' that includes the correlated proteins. Our modularization process consists of three phases as follows:

Phase 1. Selecting informative nodes.
Phase 2. Detecting preliminary modules by simulating information flow starting on each informative node.
Phase 3. Merging preliminary modules.

In phase 1, the informative nodes [5] are selected by the weighted degree, which is defined as the sum of the weights of the edges between the node of interest and its neighbors. The weighted degree d_i of a node v_i is:

$$d_i = \sum_{v_j \in N(v_i)} w_{i,j}. \tag{4}$$

Since the weights are computed on the basis of biological knowledge such as Gene Ontology annotations, the weighted degree is both topologically and biologically meaningful. The nodes with high weighted degrees correspond to not only hubs in a network but also biologically essential proteins.

Phase 2 simulates the information flow [5] starting from each informative node. A walk is a sequence of nodes such that each node is linked to its succeeding node. A path is a walk such that each node in the walk is distinct. The flow simulation is based on the concept that the information of a node v_s flows through every possible walk in a weighted network. We thus quantify the amount of information of v_s, which is called the *information rate* of v_s, flowing on the other nodes. $inf_s(v_i)$ denotes the *information rate* of v_s that is flowing on v_i. The $inf_s(v_i)$ implies how much v_s biologically influences v_i. The influence of v_s on v_i is a major factor to determine how likely it is that v_s and v_i are included in the same module.

We first assign the weighted degree d_s to each informative node v_s as an initial information rate $inf_s(v_s)$, whereas 0 to all non-informative nodes. The initial information is delivered into all neighbors v_i with being reduced by the weight of the edge.

$$f_s(v_s \rightarrow v_i) = w_{s,i} \times d_s, \tag{5}$$

where the edge $\langle v_s, v_i \rangle \in E$ and $0 < w_{s,i} \leq 1$. The information rate of v_s on v_i, $inf_s(v_i)$, is then updated by adding the sum of all incoming flow to v_i.

$$inf_s(v_i) = inf'_s(v_i) + \sum_{v_k \in N(v_i)} f_s(v_k \rightarrow v_i), \tag{6}$$

where $inf'_s(v_i)$ is the old information rate of v_s on v_i. The information of v_s then traverses all connected edges in the network by the formula defined as:

$$f_s(v_i \rightarrow v_j) = w_{i,j} \cdot \frac{inf_s(v_i)}{|N(v_i)|}, \tag{7}$$

where the edge $\langle v_i, v_j \rangle \in E$ and $0 < w_{i,j} \leq 1$. During the flow, the amount of information on each edge is repeatedly updated by Formula 6 and traverses the connected edges by Formula 7. The flow simulation algorithm from an informative node v_s is described in Algorithm 1.

As information flows through an edge, the information rate is reduced according to the weight of the edge, which is represented as the reliability of the corresponding interaction. If the weight is close to 0, then it is quickly reduced. However, if an edge $\langle v_i, v_j \rangle$ is fully reliable, that is $w_{i,j} = 1$, then the information rate of v_s on v_i can be transferred to v_j without being reduced. Since information visits all the nodes through every possible walk, densely connected areas generally have higher information rates than sparsely connected areas.

The flow in a walk stops if the information rate on a node reaches a minimum threshold, θ_{inf}. That means the amount of information is small enough for the influence to the node to be ignored. The flow from an informative node v_s terminates when there is no more information of v_s flowing in the network. A preliminary module M_s is then created with a set of proteins under the influence of v_s.

$$M_s = \{v_i | inf_s(v_i) > \theta_{inf}\}. \tag{8}$$

Simulating information flow from all informative nodes generates a set of preliminary modules, which allow overlapping.

Phase 3 is a post-processing step of merging preliminary modules to produce final modules. Each informative node v_s is a representative for a preliminary module. Two preliminary modules may be similar if two informative nodes are included into the same module. The similarity $S(M_s, M_t)$ between two modules M_s and M_t can be measured based on the weighted interconnectivity defined as:

$$S(M_s, M_t) = \frac{\sum_{v_i \in M_s, v_j \in M_t} c(v_i, v_j)}{\min(|M_s|, |M_t|)}, \tag{9}$$

where

$$c(v_i, v_j) = \begin{cases} 1 & \text{if } v_i = v_j \\ w_{i,j} & \text{if } v_i \neq v_j \text{ and } \langle v_i, v_j \rangle \in E \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

and $|M_s|$ is the number of nodes in M_s. The modules with the highest similarity are iteratively merged in this phase.

Algorithm 1 FlowSimulation($G(V, E)$, v_s)

```
1: inf_s(v_s) and C_s(v_s) ← initial rate of v_s
2: for each v_i ∈ N(v_s) do
3:    inf_s(v_i) and C_s(v_i) ← w_{s,i} · inf_s(v_s)
4: end for
5: Create a list S with all neighbors of v_s
6: inf_s(v_s) ← 0
7: while |S| > 0 do
8:    for each v_j ∈ S do
9:       Compute sum(v_j) = Σ f_s(v_i → v_j) for all v_i ∈
         N(v_j) and f_s(v_i → v_j) > θ_inf
10:   end for
11:   inf_s( neighbors of v_j) ← 0 for all v_j ∈ S
12:   for each v_j ∈ S do
13:      if sum(v_j) > 0 then
14:         inf_s(v_j) ← sum(v_j)
15:         Increment C_s(v_j) by inf_s(v_j)
16:      end if
17:   end for
18:   Replace S with all distinct neighbors of v_j for all v_j ∈
      S and sum(v_j) > 0
19: end while
20: return C_s
```

III. Experiments and Results

A. Reliability of Protein-Protein Interactions

The experiments for assessing the reliability of protein-protein interactions were performed on a full protein interaction data of Saccharomyces cerevisiae from February 2006

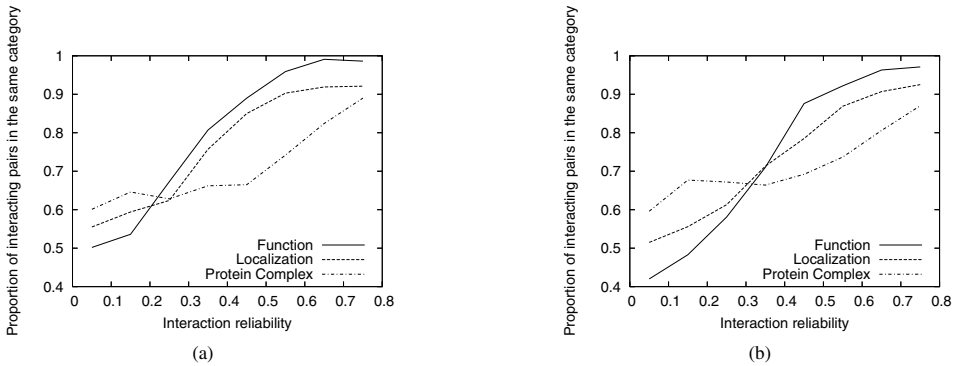

Fig. 1. The proportions of the interacting pairs, each of which is appearing in the same functions, locations and protein complexes from MIPS. The reliability was measured by the integration of interaction evidence with Gene Ontology (GO) annotations from (a) specific GO terms, which have more than 50 annotating proteins, and (b) general GO terms, which have more than 100 annotating proteins.

version of DIP, the database of interacting proteins [19]. It contains 4823 distinct proteins and 17471 interactions. We also extracted the annotated proteins for Saccharomyces cerevisiae from Gene Ontology (GO) Consortium database [10]. 33964 proteins are annotated in September 2006 version.

By the transitivity property of GO annotations, more proteins are annotated on a general term than a specific term. Similar to previous studies [6], we first filtered out excessively specific GO terms, on which a small number of proteins are annotated. Next, we selected terminal GO terms, which mean the leaf nodes in the DAG structure of GO. In this study, we experimented with two groups of GO terms. One is the terminal GO terms with more than 50 annotated proteins. The other is those with more than 100 annotated proteins. The first group has more specific GO terms than the second. The first group has 129 GO terms with 73.89 of the average size of annotations, while the second has 81 GO terms with the average size 150.95 of annotations.

To validate the measured reliability for each interaction, we employed the functional and locational categories and protein complex data with real physical interactions from MIPS database [18]. MIPS categorizes 359 different functions and 50 sub-cellular locations. It also provides the list of total 1063 protein complexes, each of which is composed of more than one protein. The protein complexes have been detected by various experiments in literature, including two large-scale experiments [8], [12].

Figure 1 shows how many interacting pairs, each of which appears in the same function, localization and protein complex, are in every range of the reliability scores. For the reliability measurement, we used the annotations from specific GO terms, which have more than 50 annotated proteins in Figure 1 (a), and general GO terms, which have more than 100 annotated proteins in Figure 1 (b). Among the interacting pairs with the reliability of greater than 0.7, more than 97% appeared in the same function, more than 92% appeared in the same localization, and more than 86% appeared in the same protein complex, in both cases of Figure 1 (a) and (b). As the reliability rises, the proportion of the identically categorized pairs is also increased. The results imply that the interaction reliability between two proteins is positively correlated to their functional, locational and physical associations. Moreover, the proportion of the interacting pairs in Figure 1 (a) is a bit higher in functional comparison than that in Figure 1 (b). It is recognized that the reliability can be specifically measured by our approach in terms of biological function.

B. Significance of Informative Nodes

Our previous study [5] investigated that the weighted degree integrated with the genetic co-expressional profiles can select topologically and biologically essential proteins in a protein interaction network. It was shown that the informative nodes selected by the weighted degree have higher average clustering coefficient and higher lethality than the nodes selected by un-weighted degree and Betweenness centrality. Since the clustering coefficient of a node represents the effect of the node on modularity, the selected nodes have topological significance when we modularize the protein interaction network. The weighted degree is also an appropriate index for selecting biologically essential proteins, which are lethal in the protein disruption test. However, due to the noisy expression data generated from Microarray experiments, the results from the previous study may still include false information.

In this study, we integrated the connectivity in a protein interaction network with GO annotations to compute the weight of each interaction. We assessed the topological and biological essentiality of the informative nodes with high weighted degree in the same way to the previous study [5]. In this experiment, we used a core protein interaction data of Saccharomyces cerevisiae from DIP [19], which includes 2526 distinct proteins and 5949 interactions. As terminal GO terms, we chose the terms with more than 50 annotated proteins.

The average clustering coefficients and lethality of the informative nodes, which are selected by GO annotation based weighted degree, are compared with the results from our previous study [5] in Figure 2. Figure 2 (a) shows that the GO annotation based weighted degree selects more modular nodes than the co-expression based weighted degree. As for

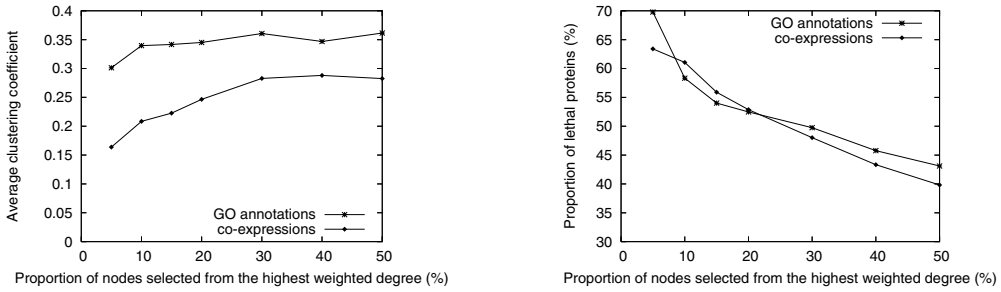

Fig. 2. The patterns of (a) the average clustering coefficients and (b) the lethality of the core nodes that are selected by weighted degree in Formula 4. The weights of interactions were calculated by two different methods: the integration of the network connectivity with Gene Ontology (GO) annotations and gene co-expressional profiles by Pearson correlation.

the lethality of proteins in Figure 2 (b), top 5% of nodes from the GO annotation based weighted degree contains more lethal proteins than those from the co-expression based weighted degree. When we select top 20% of nodes as informative nodes, the proportions of lethal proteins are similar in both cases. However, more lethal proteins are included in top 5% when we use the GO annotation based weighted degree. Overall, the modularity and essentiality of proteins in a protein interaction network can be measured by the co-expressional weighted degree, and even more precisely quantified by the weighted degree integrated with GO annotations.

C. Accuracy of Identified Modules

We implemented the information flow-based approach for identifying modules in the yeast protein interaction network. The terminal GO terms with more than 50 annotated proteins were used to estimate the reliability of interactions. After calculating the weighted degree for each node, we selected top 5% of the nodes as informative nodes. We then simulated the information flow starting from each informative node with the minimum information rate threshold of 0.1.

To statistically assess the accuracy of the generated modules by our algorithm, we compared them with the functional and sub-cellular locational categories from MIPS database [18]. Similar to our previous work [5], we calculate p-value on the hypergeometric distribution such that

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{|X|}{i}\binom{|V|-|X|}{n-i}}{\binom{|V|}{n}}, \qquad (11)$$

where $|V|$ is the total number of nodes, $|X|$ is the number of nodes in a category, n is the number of nodes in a module, and k is the number of common nodes between the category and the module. Low P in Formula 11 indicates that the module is similar to the category. After computing P with each category, we assigned one major function and localization to the module by finding the lowest P, which means the best match. We define p-score for each module as the negative of $log(P)$ with the assigned function. We then evaluate the overall accuracy using the average p-score of all modules.

We first compared the accuracy of this work with that of our previous study [5]. Table I shows that the average p-score of modules was remarkably improved by the integration of GO annotations. Node discard rate in Table I represents the proportion of the nodes that are not included in any modules. After implementation, we selected the modules whose size is greater than or equals to 5 as final modules. Thus, the small-sized modules with less than 5 nodes have been discarded.

The modules resulted from a GO annotation based weighted network have higher accuracy by more than 100% in function and more than 70% in localization than those from the co-expression based weighted network. As a post-process, we merged similar modules to increase the accuracy of generated modules. The post-process in GO annotation based method improved the accuracy by 4% in function and 15% in localization. However, the post-process did not work well for the co-expression based method. These results indicate that the GO annotation based method identified more accurate modules by information flow simulation, and more effectively merged them than the co-expression based method.

We finally compared the performance of our algorithm with several other methods, such as the maximal clique algorithm, the quasi-clique algorithm for finding clusters that maximize an optimization function [3], [21], and the statistical method for combining clusters with the significance of neighborhood [20], the minimum cut algorithm, the interconnection cut algorithm for iteratively cutting the edge with the highest Betweenness centrality [9], and the Markov clustering algorithm [7]. Table I demonstrates our approach outperforms the others in terms of accuracy.

IV. Conclusion

Unreliable interaction data is one of the most critical issues in current research of protein interaction networks. To resolve this problem, we have studied the integration of protein interaction networks with other biological knowledge. In this

TABLE I

COMPARISON OF MODULARIZATION RESULTS.

Methods	Number of Modules	Average Size of Modules	Node Discard Rate (%)	**Average p-Score** Functions	Localization
Information Flow (GO annotation based)					
Before Merging	125	45.79	23.2	38.52	22.61
After Merging	**99**	**50.56**	**23.2**	**40.06**	**25.94**
Information Flow (co-expression based)					
Before Merging	125	33.36	34.8	18.24	15.42
After Merging	115	34.70	34.8	18.27	15.09
Maximal Clique	120	5.65	98.4	10.61	7.93
Quasi-Clique	103	11.17	80.8	11.50	6.58
Neighbor Merging	64	7.91	79.9	9.16	4.89
Minimum Cut	114	13.46	35.0	8.36	4.75
Interconnection Cut	180	10.26	21.0	8.19	4.18
Markov Clustering	163	9.79	36.7	8.18	3.97

paper, we use the annotations in Gene Ontology (GO) database for assessing the reliability of interactions.

Our experimental results signify that GO annotations help identifying more accurate modules than Microarray expressional data. The GO annotations are explicitly a useful resource to evaluate the functional and locational associations between two proteins because GO terms are precisely described in biological processes, molecular functions and cellular components. Also, it is known that the high-throughput Microarray experiments typically generate large amounts of noisy data. The unreliable data in protein interactions cannot be cured by other noises in Microarray expressions.

Based on this study, we can discover valuable information from protein interaction networks. For example, we can predict functions of uncharacterized proteins from the identified modules. There are still a large number of functionally unknown proteins in yeast database even though the yeast is one of the most well-studied organisms. Discovering complete knowledge of molecular functions may be an ultimate goal in the field of Bioinformatics research.

ACKNOWLEDGMENT

This work was partly supported by NSF grant DBI-0234895 and NIH grant 1 P20 GM067650-01A1.

REFERENCES

[1] Bard, J. B. L. and Rhee, S. Y., *Ontologies in biology: design, applications and future challenges*, Nature Reviews: Genetics, 5:213-222, 2004.

[2] Barrat, A., Barthelemy, M., Pastor-Satorras, R. and Vespignani, A., *The architecture of complex weighted netowrks*, PNAS, 101(11):3747-3752, 2004.

[3] Bu, D., at al., *Topological structure analysis of the protein-protein interaction network in budding yeast*, Nucleic Acid Research, 31(9):2443-2450, 2003.

[4] Chen, J. and Yuan, B., *Detecting functional modules in the yeast protein-protein interaction network*, Bioinformatics, 22(18):2283-2290, 2006.

[5] Cho, Y.-R., Hwang, W. and Zhang, A., *Identification of overlapping functional modules in protein interaction networks: information flow-based approach*, Proceedings of 6th IEEE International Conference on Data Mining-Workshops, 147-152, 2006.

[6] Deng, M., Tu, Z., Sun, F., and Chen, T., *Mapping gene ontology to proteins based on protein-protein interaction data*, Bioinformatics, 20(6):895-902, 2004.

[7] Enright, A. J., van Dongen, S. and Ouzounis, C. A., *An efficient algorithm for large-scale detection of protein families*, Nucleic Acids Research, 30(7):1575-1584, 2002.

[8] Gavin, A.-C., at al., *Functional organization of the yeast proteome by systematic analysis of protein complexes*, Nature, 415:141-147, 2002.

[9] Girvan, M. and Newman, M. E. J., *Community structure in social and biological networks*, PNAS, 99(12):7821-7826, 2002.

[10] The Gene Ontology Consortium, *The Gene Ontology (GO) project in 2006*, Nucleic Acids Research, 34:D322-D326, 2006.

[11] Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W., *From molecular to modular cell biology*, Nature, 402:c47-c52, 1999.

[12] Ho, Y., at al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*, Nature, 415:180-183, 2002.

[13] Huang, D. and Pan, W., *Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data*, Bioinformatics, 22(10):1259-1268, 2006.

[14] Hvidsten, T. R., Lagreid, A., and Komorowski, J., *Learning rule-based models of biological process from gene expression time profiles using Gene Ontology*, Bioinformatics, 19(9):1116-1123, 2003.

[15] Ito, T., at al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome*, PNAS, 19(4):4569-4574, 2001.

[16] Jansen, R., Greenbaum, D. and Gerstein, M., *Relating whole-genome expression data with protein-protein interactions*, Genome Research, 12:37-46, 2002.

[17] Kemmeren, P., at al., *Protein interaction verification and functional annotation by integrated analysis of genome-scale data*, Molecular Cell, 9:1133-1143, 2002.

[18] Mewes, H. W., at al., *MIPS: analysis and annotation of proteins from whole genome in 2005*, Nucleic Acid Research, 34:D169-D172, 2006.

[19] Salwinski, L., at al., *The database of interacting proteins: 2004 update*, Nucleic Acid Research, 32:D449-D451, 2004.

[20] Samanta, M. P. and Liang, S., *Predicting protein functions from redundancies in large-scale protein interaction networks*, PNAS, 100(22):12579-12583, 2003.

[21] Spirin, V. and Mirny, L. A., *Protein complexes and functional modules in molecular networks*, PNAS, 100(21):12123-12128, 2003.

[22] Tornow, S. and Mewes, H. W., *Functional modules by relating protein interaction networks and gene expression*, Nucleic Acid Research, 31(21):6283-6289, 2003.

[23] Tuikkala, J., Laura, E., Nevalainen, O. S., and Aittokallio, T., *Improving missing value estimation in microarray data with gene ontology*, Bioinformatics, 22(5):566-572, 2006.

[24] Uetz, P., at al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*, Nature, 403:623-627, 2000.

[25] von Mering, C., at al., *Comparative assessment of large-scale data sets of protein-protein interactions*, Nature, 417:399-403, 2002.