ANN Based Protein Function Prediction Using Integrated Protein-Protein Interaction Data

Lei Shi, Young-Rae Cho, Aidong Zhang
Computer Science & Engineering Department
State University of New York at Buffalo
Buffalo, NY 14260, U.S.A.
Email: {lshi2,ycho8,azhang}@cse.buffalo.edu

Abstract

A major challenge in the post-genomic era is to determine protein function on a proteomic scale. There are only less than half of the actual functional annotations available for a typical proteome. The recent high-throughput bio-techniques have provided us large-scale protein-protein interaction data,and many studies have shown that function prediction from protein-protein interaction data is a promising way since proteins are likely to collaborate for a common purpose. However, the protein interaction data is very noisy, which makes the task very challenging.

In this paper, we propose a distance matrix based on the small world property of the protein-protein interaction network. It measures the reliability of edges and filter the noise in the network. In addition, an ANN (Artificial Neural Network) based method was designed to predict protein functions on the basis of integration of several protein interaction data sets. We tested our approach with MIPS functional categories. The experiential results show that our approach has better performance than other existing methods in terms of precision and recall.

1. Introduction

The classical way to predict protein functions is to find homologies between an unannotated protein and other proteins using sequence similarity algorithms, such as FASTA [16] and PSI-BLAST [12]. The function of the unannotated protein can then be assigned according to the annotated proteins with similar sequences. In addition, several computational approaches are proposed based on correlated evolution mechanisms of genes. For example, the domain fusion analysis infers that a pair of proteins interacts with each other and thus performs related functions [13].

In recent years, the high-throughput bio-techniques have provided additional opportunities for inference of protein functions. Protein-protein interaction (PPI) data, enriched by high-throughput experiments including yeast two-hybrid analysis [9][26], mass spectrometry [5][8] and synthetic lethality screen [25], have provided the important clues of functional associations between proteins. Proteins are likely to collaborate for a common purpose. Therefore, the functions of an unannotated protein can be deduced when the functions of its binding partners are known.

Many approaches have been proposed to predict protein functions from protein interaction networks. The neighbor counting method [22] used the majority-rule to label a protein with the functions that occur most frequently in its interaction partners. Hishigaki et al. [7] used a chi-square statistics to calculate the significance of the functions of neighbors. However, these methods can only predict the proteins which have at least one interaction partner. Moreover, the predicted annotations for an unknown protein are limited by the annotations of its interacting partners.

To avoid those limitations, several other approaches are based on the global topology of protein interaction networks. Vazquez et al. [27] and Karaoz et al. [10] attempted to maximize the functional consistency through neighboring in the whole network. Nabieva et al. [14] applied the concept of functional flow that is propagated from an annotated protein to unannotated proteins. Deng et al. [4] adopted the Markov Random Field (MRF) model to simulate the protein interaction network with functional annotations. Lee et al. [11] developed a kernel logistic regression (KLR) method, which used diffusion kernels and incorporated all indirect neighbors in the networks. While these approaches demonstrated that using machine learning and statistical methods can improve prediction performance, they bank on the same functional concept that the interaction partners of a protein are likely to share similar functions with it [1].

Although previous methods have proved to be useful to predict protein functions from PPI networks, they have a critical challenge. Protein interaction data derived from the high-throughput techniques are typically very noisy. The data may include many false negatives (true interactions which remain undetected) and false positives (putative interactions that in fact do not occur). Sprinzak et al. [23] reported that the reliability of high-throughput yeast two-hybrid assays is about 50%.

978-0-7695-3739-9/09 $25.00 © 2009 IEEE
DOI 10.1109/IJCBS.2009.98

Table 1: The percentage of function-relevant interactions in three protein interaction data sets

Data Set	Total number of interactions	Number of Functional-relevant interactions	Percentage
DIP	14162	5216	36.83%
MIPS	13877	4189	30.18%
BioGrids	117675	36446	30.97 %

Table 2: The percentage of function-consistent protein pairs which interact in protein interaction data sets

Data Set	Total number of interactions	Number of Functional-relevant interactions	Percentage
DIP	20,099	1283	6.38%
MIPS	21,795	898	4.12%
BioGrids	21,499	2718	12.64%

A protein interaction network is normally represented as an unweighted graph. However, since it includes a large number of unreliable interactions, the unweighted graph is far from optimal in representing the data. In this paper, we build a weighted graph model of protein interaction networks. Based on that, we propose a topological measurement to reflect our knowledge of small world network property to filter the protein interaction network and get a more reliable interaction network. Since one protein may have multiple functions, it is a typical multi-label problem. In a weighted graph, it is adequate to use Artificial Neural Network (ANN) method to predict protein functions. In this paper, we investigate the problem of predicting protein functions from protein interaction data and make the following contributions:

- We analyze the reliability of connections in several protein interaction networks.
- We propose a novel topological measurement to calculate the interaction reliability between two proteins and filter the protein-protein interaction networks.
- We propose an ANN based method to predict the functions of proteins.

The remainder of the paper is organized as follows. In Section 2, we present our weighted graph model and topological measurement to rebuild protein interaction networks. In Section 3, we present our ANN based prediction model. Extensive experimental results are reported in Section 4. The paper is concluded in Section 5.

2. Weighted Graph Model of Protein Interaction Networks

Many methods are based on the assumption that interacting proteins should share common functions. Table 1 shows the percentage of *function-relevant* interactions in three protein-protein interaction data sets, namely, DIP, MIPS and BioGrid (see Result Section for detail description). An interaction is considered to be *function-relevant* if the two proteins involved in the interaction have at least one function in common. In this test, we adopt FunCat(version 20070316) [15] in the MIPS database as our annotation categories. From Table 1, we can see that only $30\% - 40\%$ observed interactions are relevant in functions. In other words, most of the observed interactions do not share functions. Among those sharing function pairs, some of them share more functions than the others. Table 2 shows the percentage of *function-consistent* protein pairs which are observed to interact in the three data sets. Formally, we define two proteins $P1$ and $P2$ to be *funtion-consistent* if $|\frac{F(P_1)\cap F(P_2)}{F(P_1)\cup F(P_2)}| \geq \frac{1}{2}$, where $F(P_1)$ and $F(P_2)$ are functions of P_2 and P_2, respectively. As shown, only a small percentage of *function-consistent* protein pairs are observed to interact in the interaction data sets. These observations suggest two things: the protein interaction data has a high false-positive rate and a false-negative rate, so false interactions need to be removed from protein interaction data; a weighted graph needs to be built to show the reliability between two proteins and to show the *functional similarity* between two proteins. Since proteins with similar functions are likely to interact with each other in cells, we assume that the more reliable two proteins are, the more chance that they share common proteins.

We define a *weighted protein interaction network* [18] as follows: A weighted protein interaction network is a weighted undirected graph $G = (P, I, W)$, where P is a set of vertices, I is a set of edges between the vertices $(I \subseteq (u, v)|u, v \in P)$ and W is a function making each edge in I to a real value in the range of [0...1]. Each vertex $v \in P$ in the graph represents a protein. Each edge $(u, v) \in I$ represents an interaction between proteins u and v. For each edge (u, v), $w(u, v)$ is the weight of (u, v) which represents the probability of this interaction being a true positive. Figure 1 shows our weighted protein interaction network model. The nodes represent the proteins, the edges between nodes represent the interactions between proteins, and the numbers on the edges represent the weights between interacted proteins.

In this paper, we use the following additional terminologies: A *neighbor* of a vertex v is a vertex adjacent to v, also called *direct neighbor*. *Level-k neighbor* of vertex v is a vertex having k edges or steps to reach vertex v. The degree of a vertex v, denoted as $D(v)$, is the sum of weights of the edges connecting v: $D(v) = \sum_{(u,v)\in I} w(u, v)$. A *walk* is an alternating sequence of vertices and edges, with each edge being incident to the vertices immediately preceding and succeeding it in the sequence. A *path* is a walk with no repeated vertices.

Generally, there are two approaches to give a probability estimate for each interactions: We can use either the probability estimates of single interactions or the reliability esti-

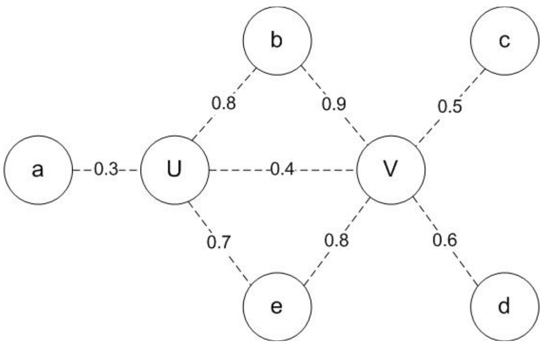

Figure 1: a weighted protein interaction network model

mates of interaction data sets. In this paper, we just simply use the latter one. Now we need to combine several different protein interaction data sets $S = \{S_1, S_2, ..., S_n\}$, where each set S_i includes many interactions. If an interaction (u, v) appears only in one data set, we will set its probability as the reliability of this data set:

$$w(u, v) = r_k \qquad \text{for each } (u, v) \in S_k, \tag{1}$$

where r_k is the estimated reliability of the protein interaction data set S_k. The interaction (u, v) may appear in several data sets, i.e.,

$$(u, v) \in S_1 \bigcap S_2 ... \bigcap S_n, \tag{2}$$

where $n > 1$. In this case, its probability is set to :

$$w(u, v) = 1 - (1 - r_1) * (1 - r_2) ... * (1 - r_n), \tag{3}$$

where r_i is the estimated reliability of S_i. This formula reflects the fact that interactions detected in multiple experiments are generally more reliable than those detected by only one experiment.

Then the reliability between two proteins is calculated by the following formulas:

$$PR(A, B) = \sum_{k=2}^{|P|-2} PR^k(A, B) \tag{4}$$

$$PR^k(A, B) = \frac{PS^k(A, B)}{\mathit{MaxPS}^k(A, B)} \tag{5}$$

$$PS^k(A, B) = \sum_{p=<v_0=A, v_1, ..., v_k=B>} PS(p) \tag{6}$$

$$PS(p) = \prod_{i=1}^{l} w(v_{i-1}, v_i) \tag{7}$$

$$\mathit{MaxPS}^k(A, B) = \begin{cases} \sqrt{D(A) * D(B)} & \text{if } k = 2 \\ D(A) * D(B) & \text{if } k = 3 \\ \sum_{P_i \in N(A), P_j \in N(B)} \mathit{MaxPS}^{k-2}(P_i, P_j) & \text{if } k > 3 \end{cases} \tag{8}$$

Table 3: Fraction of annotated yeast proteins that share function with 1) level-1 neighbors exclusively; 2) level-2 neighbors exclusively; 3) level-3 neighbors exclusively; 4) level-4 neighbors exclusively

Shared functions with	Number of corresponding neighbors	Number of sharing common functions	Fraction
Level-1 neighbors exclusively	4812	1136	23.61%
Level-2 neighbors exclusively	203574	40275	19.78%
Level-3 neighbors exclusively	1381525	185182	13.45%
Level-4 neighbors exclusively	913742	49068	5.17%

where $PS(p)$ is the *PathStrength* of a path $p = < v_0, v_1, ..., v_l >$, $PS^k(A, B)$ is $k-$*length PathStrength between two vertices* A and B, which is the sum of the *PathStrength* of all k-length paths between vertices A and B, $\mathit{MaxPS}^k(A, B)$ is the $k-$*length MaxPathStrength between two vertices between* A and B, $PR^k(A, B)$ is the $k-$*length PathRatio* between two vertices A and B, $PR(A, B)$ is the *PathRatio* between two vertices A and B, and $|P|$ is the number of vertices in the graph.

The value of PathRatio reflects the reliability between two proteins in the network. In a protein interaction network, the closer two proteins, the more chances they should share common functions. Table 3 shows clearly that there are quite strong functional influence between level-1 and level-2 neighbors, still some influence for level-3 neighbors, but weaker influence for neighbors higher than 3. So in this paper, when we calculate the PathRatio between two vertices, we just calculate the PathStrength up to the third level. The bigger the value of PathRatio, the more reliable these two proteins are, i.e, the higher probability that these two proteins share common proteins. In this way, we changed this weighted graph to a *functional similarity* interaction network.

Fig. 2 shows that our assumption works well for some simple protein function prediction methods, such as *neighbor counting* method [22]. In this test, function prediction performance from the weighted interaction network was assessed comparing to an unweighted interaction network. The result shows how significantly the weighted graph can improve the predict performance.

This new interaction network is different from the former one in the following ways:

- The neighbors of one annotated protein include both direct neighbors and indirect neighbors up to level-3.
- These functional similarity based weighted networks have the weight for each pair of neighbors, which indicates the chance that they may have similar functions.

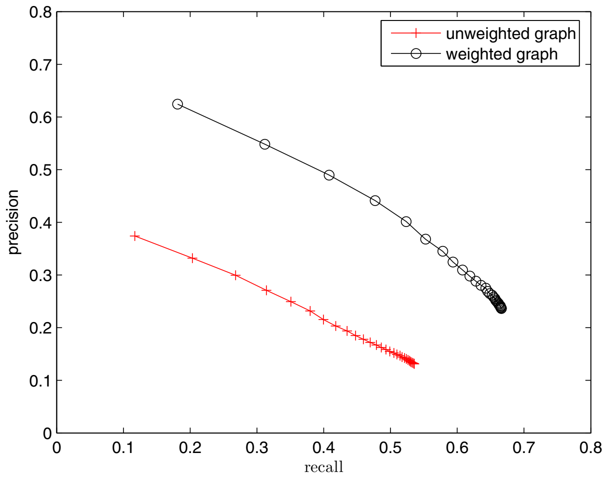

Figure 2: Comparison of Neighbor Counting method with unweighed and weighted protein interaction graphs

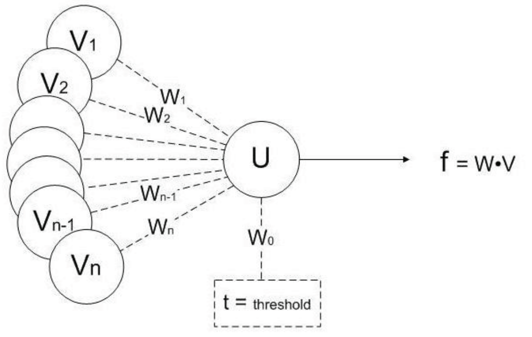

Figure 3: ANN based function prediction model. U is the unannotated protein, $V_{1\sim n}$ are U's neighbors, $W_{1\sim n}$ are the PathRatio value of the edge, t is the user defined threshold, f is the binary vector indicating the functions predicted for U.

3. Function prediction algorithm

Since one protein can have multiple functions, predicting multiple functions of each unannotated protein is a typical multi-label problem with functions as labels and proteins as instances or items. Because the proteins are connected and the protein interaction network has been proved to have small world properties [28], it automatically leads us to use Artificial Neural Network (ANN) to solve this problem and use the PathRatio as the weight W and the neighbors as the nodes. ANN is a computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Here we escape the adaptive part and only use a simple model called *perceptron*. In the network we built from the last section, every annotated protein should have a list of variables indicating if this protein has any particular functions. For example, if there are 4 functions $\{f1, f2, f3, f4\}$, and protein p has $\{f1, f2, f4\}$, the function vector v of protein p will be $(1, 1, 0, 1)$. Then for an unannotated protein u, the set of possible functions f it may have can be predicted using the following formula.

$$\begin{aligned}\hat{f} &= sign(\mathbf{w}\cdot\mathbf{v}) \\ &= sign[w_d v_d + w_{d-1}v_{d-1} + ... + w_1 v_1 + w_0 v_0]\end{aligned} \tag{9}$$

where $w_0 = -t$, $v_0 = 1$, $\mathbf{w}\cdot\mathbf{v}$ is a dot product between the weight vecotr $\mathbf{w}$ and the input attribute matrix $\mathbf{v}$, t is the threshold value to be set, v_i is the function vector of the neighbor i, and w_i is the *functional similarity* weight between neighbor i and protein u, where i is the index of the neighbors of protein u, and d is the number of neighbors that the unannotated proteins has. Function $sign$ outputs 1 if $\mathbf{w}\cdot\mathbf{v}$ is bigger than 0, otherwise, it outputs -1. The result $\mathbf{f}$ is a vector indicating if protein u has the corresponding function. This ANN based model is shown in Fig. 3. The proposed ANN model unites the functional information from the neighbors of the unannotated protein and assigns those information with different weights which represent the functional similarities between the neighbors and the unannotated protein. Since this model fully uses the neighbors up to level-3 and treat different neighbors seperately, it overcomes weak points of previous methods, such as narrow neighborhood and equal effection from neighbors.

4. Experimental Results

4.1. Cross-validation of function prediction

For our experiments, we built protein interaction networks from three different yeast interaction networks. One is MIPS data set [15], which contains 3882 proteins and 13877 interactions. The second one is BioGrid data set [24], which contains 4265 proteins and 117675 interactions. The third one is DIP data set [21], which contains 4935 proteins and 14162 interactions.

To evaluate the effectiveness of our method, we used FunCat as the functional annotations from MIPS database [15]. The scheme of FunCat is a tree-shaped hierarchical structure. To avoid overly specific annotations, we cut the scheme at the third level and obtained 259 functional categories.

We assessed the performance of our function prediction approach by the leave-one-out cross-validation method. For

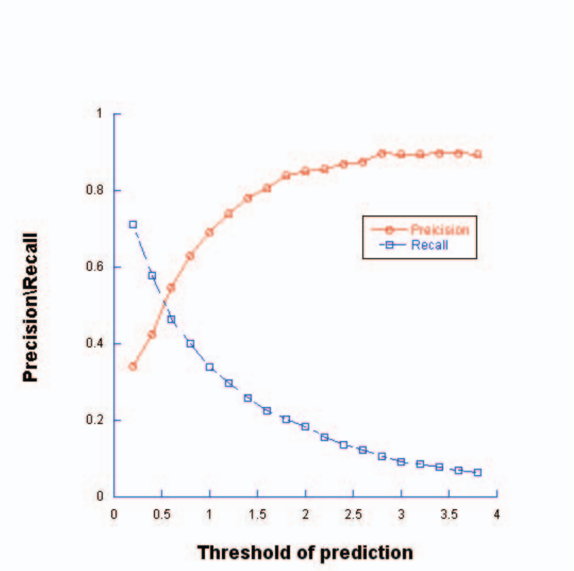

Figure 4: Precision and recall plots by cross-validation for protein function prediction. The performance of our function prediction algorithm was assessed by the leave-one-out cross-validation using the proteins that appear in the interaction data from DIP and are annotated on the functional categories in MIPS. As a higher threshold of prediction confidence is used, precison increases whereas recall decreases.

each protein in annotations, we assumed it is un-annotated and predicted its function using its interaction information and the annotations of the other proteins. Then we compared the predicted functions with the true annotations. Let n_i be the number of annotated functions for protein P_i, m_i be the number of predicted functions for P_i, k_i be the size of common functions of m_i and n_i, and n is the total number of distinct proteins with annotations. Precision and recall are then calculated as :

$$Precision = \frac{\sum_1^n k_i}{\sum_1^n n_i}, \tag{10}$$

and

$$Recall = \frac{\sum_1^n k_i}{\sum_1^n m_i}. \tag{11}$$

When we implement our proposed method, first we integrated those three different protein interaction data sets using Formula 3, and the reliability of each data set was estimated by EPR (Expression Profile Reliability) index [3]: 0.85 for DIP, 0.73 for MIPS, 0.81 for BIOGRID. Then we rebuilt the weighted graph with weight threshold 0.2, which means we only keep interactions whose weight is above 0.2. At this point, we achieved an interaction graph. Then we used the prediction algorithm we proposed above to predict the functions of each protein in the data set. Figure 4 shows the precision and recall plots with respect to the threshold of prediction confidence, which is a user-dependent parameter in our algorithm. When we use 3.8 as the threshold of prediction confidence, our algorithm predicts fewer functions for each protein, but most of the functions are correctly predicted comparing to the actual annotations, and the precision for this threshold is close to 0.9. As a lower threshold is used, recall increases while precision decreases monotonically. Approximately, when the recall is 0.2 and 0.4, we had the precision of 0.8 and 0.6, respectively.

4.2. Comparison with other approaches

We evaluated the performance of our ANN method with two previous approaches: the neighbor-counting method [22], and the indirect-neighbor method [1].

Indirect-neighbor method [1] computes the likelihood that an unknown protein p has a function using the functional similarity weights between p and direct and indirect neighbors. The functional similarity weight of two proteins is calculated by the commonality of their neighbors in the protein interaction network. We used a threshold of the likelihood to generate the output set of predicted functions for each protein. We then obtained different output sets by various thresholds. Neighbor-Counting method computes the frequency of each function among the direct neighbors of protein p and then sorts it to get the top k functions.

Figure 5 shows the precision and recall of the three approaches on the filtered data set. Our ANN based method remarkably outperforms the neighbor-counting method, since neighbor-counting method only considered the direct neighbors and missed lots of functional information from other proteins in the protein interaction network. Our approach is slightly better than indirect-neighbor method when recall is between 0 and 0.2, but when recall is bigger than 0.2, our method has the precison of more than 0.1 higher than the indirect-neighbor method. This result indicates three things: 1) Fully understanding the small world property of the protein interaction network is very important to predict the functions of proteins. 2) The more functional information you use to predict unknown proteins, the better result you may get. 3) A weighted graph is more suitable to represent protein interaction networks than unweighted graph to predict the functions of proteins, and the more reliable the graph is, the more accurate the result will be.

It is worth mentioning that since building a weighted graph and function prediction are completely independent, different approaches can be adopted for these two steps, such as using IRAP [2] or IG2 [20] to build the weighted graph and using KNN or other machine learning methods to predict the functions of proteins.

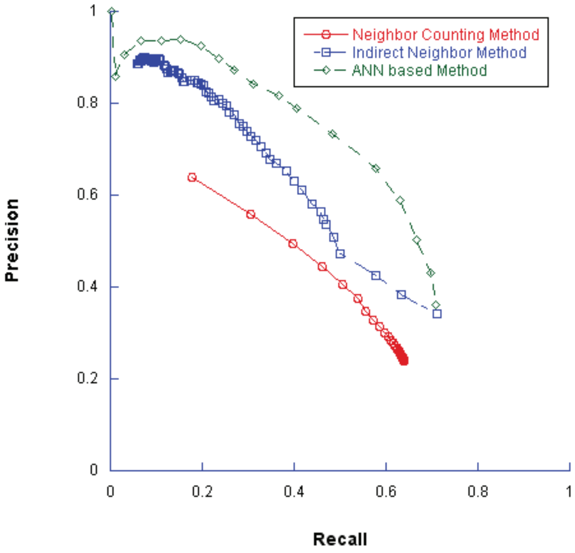

Figure 5: The precision-recall relationships of our ANN based method are compared with two competing methods: Indirect Neighbors methods and Neighbor Counting Methods. For any recall value, our approach substantially outperformed the other two methods

4.3. Discussion

Through recent advances of high-throughput techniques, a significant amount of protein interaction data has been accumulated. Protein functions have been predicted from the interaction data because the evidence of interactions can be interpreted as functional links. However, we observe that only a small fraction of current interaction data from major interaction databases are related to functional linkage. The results indicate that more than 60% of interacting protein pairs are not linked by similar functions. In other words, at most 40% of protein pairs have been motivated by similar functions. This observation has been also demonstrated by the limited accuracy of previous function prediction methods.

Our method uses the small world property of protein interaction networks and derives functional information from both direct neighbors and up to level-3 neighbors, which is more comprehensive than just using direct neighbors and neighborhood information. Also using a weighted interaction network is more suitable than using a unweighted network since different neighbors have different contributions to the functions of unknown proteins.

In our experiments, function prediction has been conducted with yeast protein-protein interaction data. However, our ANN based framework can be well-applicable to higher-level organisims because of its efficiency.

5. Conclusion

Functional characterization of proteome is a central goal in the field of Bioinformatics. The experimentally determined protein interactions are crucial data sources to uncover the functional knowledge of uncharacterized proteins. However, a pre-process to access the functional linkage of interacting proteins from current interaction data is required for predicting protein function successfully.

In this paper, we presented an ANN based method to integrate direct neighbors, level-2 neighbors and level-3 neighbors based on a weighted protein interaction network to predict the functions of proteins. Our results imply that function prediction from protein interaction networks using a weighted network is a promising way, and integrating more data sets and more protein function related information may achieve better results. This is also our future research for functional knowledge discovery.

Acknowledgments

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

References

[1] Chua, HN, Sung, W-K and Wong,L : Exploiting indirect neighbors and topological weight to predict protein function from protein protein interactions. *Bioinformatics* 22(13):1623-1630, 2006.

[2] Chen J., Hsu, W., Lee, M. and Ng, S. :Systematic assessment of high-throughput experimental data for reliable protein interactions using network topology. 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04) pages 368-372, 2004.

[3] Deane, M.C. et al. : Protein interactions : two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1:349-356, 2002.

[4] Deng, M, Zhang, K, Mehta, S,Chen, T and Sun, F: Prediction of protein function using protein protein interaction data. *Journal of Computational biology* 10(6):947-960,2003.

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

[6] Goldberg D.S. and Roth F.P. : Assessing experimentally derived interactions in a small world. *Proc. natl. Acad. Sci. USA,* 100:4372-4376s, 2003.

[7] Hishigaki, H, Nakai, K, Ono, T, Tanigami, A and Takagi, T: Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18:523-531, 2001.

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

[9] 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

[10] Karaoz,U, Murali, TM, Letovsky,S, Zheng, Y, Ding, C, Cantor, CR and kasif, S: Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA* 101(9):2888-2893, 2004.

[11] Lee, H, Tu, Z, Deng, M, Sun, F and Chen, T: Diffusion Kernel-based logistic regression models for protein function prediction. *OMICS A journal of Integrative Biology* 10(1): 40-55, 2006.

[12] Lockhart, D.J., Dong, H., Byne, M.C.,Follettie, M.T., Gallo,M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*, 14:1675-1680, 1996.

[13] Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D Detecting protein function and protein protein interaction from genome sequences. *Science* 285: 751:753, 1999.

[14] Nabieva, E, Jim, K, Agarwal, A, Chazelle, B and Singh, M: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21:i302-i310, 2005.

[15] Mewes, H.W. et al., MIPS : analysis and annotation of proteins from whole genomes in 2005. *Nucl Acids Res*, 34:D169-D172, 2006.

[16] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444-2448, 1988.

[17] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg d, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96:4285:4288, 1999.

[18] Pei, P. and Zhang, A. : A toplolgical measurement for weighted protein interaction network. *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, 268-278,2005

[19] Saito, R.,Suzuki, H. and Hayashizaki, Y. :Interaction generality, a measurement to assess the reliability of a protein protein interaction. *Nucleic Acids Res*, 30:1163-1168, 2002.

[20] Saito, R.,Suzuki, H. and Hayashizaki, Y.: Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19:756-763,2003.

[21] Salwinski, L., et al : The Database of Interacting Proteins: 2004 updata. NAR 32 Database issue, pages D449-51, 2004.

[22] Schwikowski, B., Uetz, P. and Fields, S.: A network of protein protein interactions in yeast. *Nature Biotechnology*,18:1257-1261,2000

[23] Sprinzak, E. , Sattath, S. and Margalit, H. How reliable are experimental protein protein interaction data? *J. Mol. Biol.*, 327(5):919-923, Apr.2003

[24] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: A General Repository for Interaction Datasets. Nucleic Acids Res. 34:D535-9, 2006 Jan 1

[25] Global mapping of the yeast genetic interaction network. Science 303:808-813,2004.

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

[27] Vazquez, A, Flammini, A, Maritan, A, and Vespignani, A: Global protein function prediction from protein-protein interaction networks. *Nature biotechnology* 21(6):697-700, 2003.

[28] Watts, D. and Strogatz, S. Collective dynamics of ”small-world” networks. *Nature*, 393(440-442),1998.