Protein Function Prediction Based on Neighborhood Profiles

Kire Trivodaliev, Ivana Cingovska, Slobodan Kalajdziski, and Danco Davcev

Ss. Cyril and Methodius University, Faculty of Electrical Engineering and Information Technologies, Computer Science Department, Karpos 2 BB,
1000 Skopje, Macedonia
{kiret,ivanac,skalaj,etfdav}@feit.ukim.edu.mk

Abstract. The recent advent of high throughput methods has generated large amounts of protein interaction network (PIN) data. A significant number of proteins in such networks remain uncharacterized and predicting their function remains a major challenge. A number of existing techniques assume that proteins with similar functions are topologically close in the network. Our hypothesis is that the simultaneous activity of sometimes functionally diverse functional agents comprises higher level processes in different regions of the PIN. We propose a two-phase approach. First we extract the neighborhood profile of a protein using Random Walks with Restarts. We then employ a “chi-square method”, which assigns k functions to an uncharacterized protein, with the k largest chi-square scores. We applied our method on protein physical interaction data and protein complex data, which showed the later perform better. We performed leave-one-out validation to measure the accuracy of the predictions, revealing significant improvements over previous techniques.

Keywords: Protein interaction networks, Neighbourhood extraction, Protein function prediction.

1 Introduction

The rapid development of genomics and proteomics has generated an unprecedented amount of data for multiple model organisms. As has been commonly realized, the acquisition of data is but a preliminary step, and a true challenge lies in developing effective means to analyze such data and endow them with physical or functional meaning [1]. The problem of function prediction of newly discovered genes has traditionally been approached using sequence/structure homology coupled with manual verification in the wet lab.

The first step, referred to as computational function prediction, facilitates the functional annotation by directing the experimental design to a narrow set of possible annotations for unstudied proteins.

Significant amount of data used for computational function prediction is produced by high-throughput techniques. Methods like Microarray co-expression analysis and Yeast2Hybrid experiments have allowed the construction of large interaction networks. A protein interaction network (PIN) consists of nodes representing proteins, and edges representing interactions between proteins. Such networks are stochastic as edges are weighted with the probability of interaction. There is more information in a PIN compared to sequence or structure alone. A network provides a global view of the context of each gene/protein. Hence, the next stage of computational function prediction is characterized by the use of a protein's interaction context within the network to predict its functions. A node in a PIN is annotated with one or more functional terms. Multiple and sometimes unrelated annotations can occur due to multiple active binding sites or possibly multiple stable tertiary conformations of a protein. The annotation terms are commonly based on an ontology. A major effort in this direction is the Gene Ontology (GO) project [2]. GO characterizes proteins in three major aspects: molecular function, biological process and cellular localization. Molecular functions describe activities performed by individual gene products and sometimes by a group of gene products. Biological processes organize groups of interactions into "ordered assemblies." They are easier to predict since they localize in the network. In this paper, we seek to predict the GO molecular functions for uncharacterized (target) proteins. The main idea behind our function prediction technique is that function inference using only local network analysis but without the examination of global patterns is not general enough to cover all possible annotation trends that emerge in a PIN.

According to a recent survey [3], most existing network-based function prediction methods can be classified in two groups: module assisted and direct methods. Module assisted methods detect network modules and then perform a module-wide annotation enrichment [4]. The methods in this group differ in the manner they identify modules. Some use graph clustering [5, 6] while others use hierarchical clustering based on network distance [4, 7, 8], common interactors [9] and Markov random fields [10].

Direct methods assume that neighboring proteins in the network have similar functional annotations. The Majority method [11] predicts the three prevailing annotations among the direct interactors of a target protein. This idea has later been generalized to higher levels in the network [12]. Another approach, Indirect Neighbor [13], distinguishes between direct and indirect functional associations, considering level 1 and level 2 associations. The Functional Flow method [14] simulates a network flow of annotations from annotated proteins to target ones. Karaoz et al. [15] propose an annotation technique that maximizes edges between proteins with the same function.

A common drawback of both the direct and module-assisted methods is their hypothesis that proteins with similar functions are always topologically close in the network. The direct methods are further limited to utilize information about neighbors up to a certain level. Thus, they are unable to predict the functions of proteins surrounded by unannotated interaction partners.

We hypothesize that the simultaneous activity of sometimes functionally diverse functional agents comprise higher level processes in different regions of the PIN. Our hypothesis is more general, since a clique of similar function proteins can be equivalently treated as a set of nodes that observe the same functional neighborhood. A justification for our approach is provided by Fig.1 which shows that proteins of similar function may occur at large network distances.

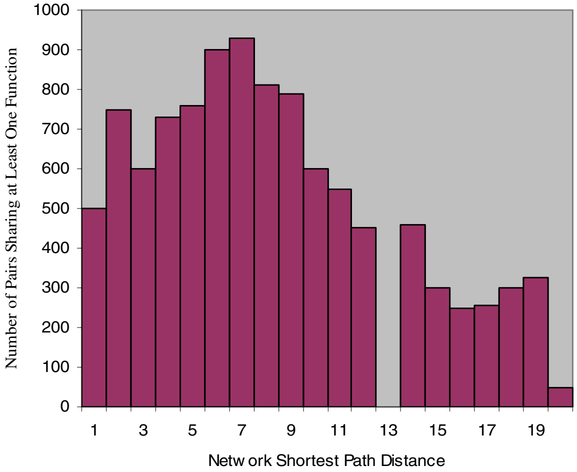

Fig. 1. Proteins sharing annotations do not always interact in the Filtered Yeast Interactome (FYI) [16]

2 Research Methods

Our approach divides function prediction into two steps: extraction of neighbourhood profile, and prediction based on the computed neighbourhood (Fig. 2). According to our hypothesis, we summarize the functional network context of a target protein in the neighbourhood extraction step. We compute the steady state distribution of a *Random Walk with Restarts (RWR)* from the protein. The steady state is then transformed into a functional profile. In the second step, we employ a *chi-sqare* method to predict the function of a target protein based on its neighbourhood profile.

2.1 Extraction of Neighbourhood Profiles

We summarize a protein's neighborhood by computing the steady state distribution of a *Random Walk with Restarts (RWR)*. We simulate the trajectory of a random walker that starts from the target protein and moves to its neighbors with a probability proportional to the weight of each connecting edge. We keep the random walker close to the original node in order to explore its local neighborhood, by allowing transitions to the original node with a probability of *c*, the restart probability.

Let $G = (V;E)$ be the graph representing a protein-protein interaction network, where V is the set of nodes (proteins), and E is the set of weighted undirected edges, where the weight shows the probability of interaction (or functional association) between protein pairs. We define the proximity of a node *v* to a start node *s*, $p_s(v)$, as the steady state probability that a random walk starting at node s will end at node v.

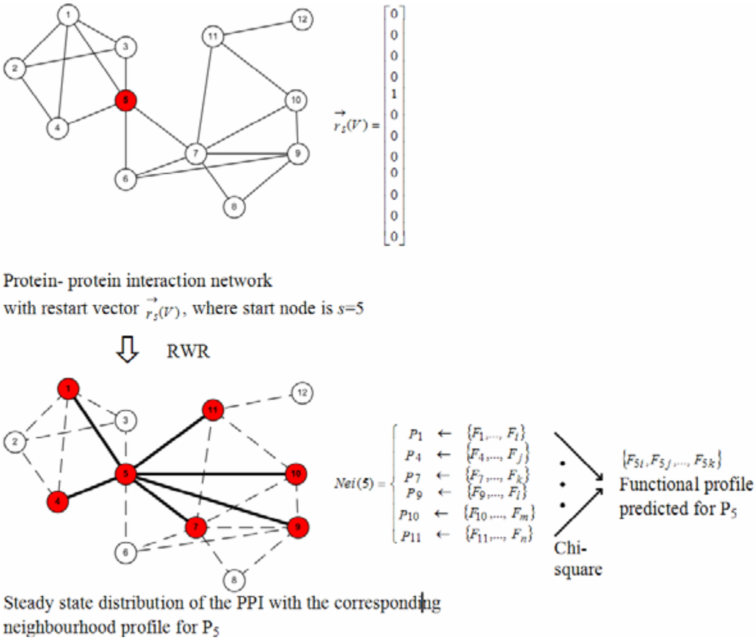
1
2
3
4
5
6
7
8
9
10
11
12
$\vec{r}_s(V) = [0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0]^T$
Protein- protein interaction network
with restart vector $\vec{r}_s(V)$, where start node is s=5
RWR
$Nei(5) = \{ P_1 \leftarrow \{F_1,\dots,F_i\};\ P_4 \leftarrow \{F_4,\dots,F_j\};\ P_7 \leftarrow \{F_7,\dots,F_k\};\ P_9 \leftarrow \{F_9,\dots,F_l\};\ P_{10} \leftarrow \{F_{10},\dots,F_m\};\ P_{11} \leftarrow \{F_{11},\dots,F_n\}$
Chi-
square
$\{F_{5i}, F_{5j},\dots,F_{5k}\}$
Functional profile
predicted for P_5
Steady state distribution of the PPI with the corresponding
neighbourhood profile for P_5

Fig. 2. Function prediction process: extraction of neighbourhood profile, and prediction based on the computed neighbourhood

Random walk method simulates a random walker that starts on a source node, *s* (or a set of source nodes simultaneously). At every time tick, the walker chooses randomly among the available edges (based on edge weights), or goes back to node s with probability *c*. The restart probability *c* enforces a restriction on how far we want the random walker to get away from the start node *s*. In other words, if *c* is close to 1, the affinity vector reflects the local structure around *s*, and as *c* gets close to 0, a more global view is observed.

The probability $p_s(v)^{(t)}$, describes the probability of finding the random walker at node *v* at time *t*. The steady state probability $p_s(v)$ gives a measure of proximity to node *s*, and can be computed efficiently using iterative matrix operations. Fig. 3 shows the iterative algorithm, which provably converges. The number of iterations to converge is closely related to the restart probability *c*. As *c* gets smaller the diameter of the observed neighborhood increases, thus the number of iterations to converge gets larger. The convergence check requires the L_1-norm between consecutive $\vec{p}_s(V)$ s to be less than a small threshold, e.g., 10^{-12}. In our experiments, for *c*=0,30 the average number of iterations to converge is around 55.

A possible interpretation of the neighborhood profile is an affinity vector of the target node to all other nodes based solely on the network structure.

```
Input: the interaction network $G = (V;E)$;
       a start node $s$;
       restart probability $c$;
Output: the proximity vector $\vec{p}_s(V)$;

Let $\vec{r}_s(V)$ be the restart vector with 0 for all its entries except a 1 for the entry
denoted by node $s$;
Let A be the column normalized adjacency matrix defined by $E$;
Initialize $\vec{p}_s(V) := \vec{r}_s(V)$;
while ( $\vec{p}_s(V)$ has not converged):
    $\vec{p}_s(V) := (1 - c)A\vec{p}_s(V) + c\vec{r}_s(V)$;
```

Fig. 3. The iterative algorithm to compute the proximity of all the nodes in the graph to a given start node *s*

2.2 Chi-Square Method for Protein Function Prediction

The second step in our approach is predicting the annotations of a given protein P_i based on its *neighborhood profile* Nei(i). We use a method to infer protein functions based on $\chi2$-statistics. For a protein P_i, let $n_i(j)$ be the number of proteins interacting with P_i and having function F_j. Let $e_i(j)$ = #Nei(i) × π_j be the expected number of proteins in Nei(i) having function F_j , where #Nei(i) is the number of proteins in Nei(i). Define

$$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)} \quad (1)$$

For a fixed *k*, they assign an unannotated protein with *k* functions having the top *k* $\chi2$-statistics.

The two steps of our approach are completely independent, different approaches can be adopted for neighboorhood extraction and classification.

3 Results

Our study assessed the reliability of two different groups of protein-protein interaction data: the protein physical interaction data and the protein complex data. The protein physical interaction data included two yeast two-hybrid data sets, by Uetz et al. [17]

and DIP [18], a collection of protein interactions from the literature and the yeast two-hybrid assays. The protein complex data included experimentally determined MIPS physical interaction data set [19], obtained by systematic purification of protein complexes and protein identification via mass spectrometry, and a set of experimentally determined protein complexes called "MIPS Complex" [19].

We separated protein complex data from protein physical interaction data because of their obvious difference: not all protein pairs in a complex interact with one another, and not all physically interacting protein pairs are in the same complex. Many protein complexes such as ribosomes and RNA Polymerases are essential for a cell, and the interactions within a complex are generally more stable and stronger and have a longer life span than most other physical interactions, while other physical interactions include other important interactions such as signal transductions.

We compare the accuracy of the techniques by performing leave one-out validation experiments. We use leave-one-out validation because many annotations in the actual network are of relatively low frequency, and thus limiting the training set. Our method is working with actual networks, containing significant number of uncharacterized proteins and hence this is a realistic measure of the accuracy. In this setup, a target protein is held out (i.e. its annotations are considered unknown) and a prediction is computed using the rest of the annotation information in the network.

The accuracy of the predictions is measured as follows. The method randomly selects an annotated protein and assumes it as unannotated. Then we predict its functions by using our method. We then compare the predictions with the annotations of the protein. We repeat the leave-one-out experiment for K proteins, P_i, … , P_K. Let n_i be the number of known functions for protein Pi, m_i be the number of *predicted* functions for protein P_i, and k_i be the overlap between the set of observed functions and the set of predicted functions. The specificity (SP) and the sensitivity (SN) can be defined as:

$$SP = \frac{\sum_i^K k_i}{\sum_i^K m_i} \quad . \tag{2}$$

$$SN = \frac{\sum_i^K k_i}{\sum_i^K n_i} \quad . \tag{3}$$

For comparison, we implement the neighbourhood counting method [11] and the $\chi 2$ method [12] for functional annotation. We choose the top 1, 2, 3, 4 and 5 functions, respectively, and assign these functions to each unannotated protein. The results of the comparison with the Uetz, DIP, MIPS physical and MIPS complex data sets are shown in Figures 4,5,6 and 7 respectively. As can be seen our method outperforms by a margin on every single data set. Figure 8 presents the assessment of the data sets in terms of their reliability when used in the process of protein function prediction. Our results confirm that the components of a protein complex can be assigned to functions that the complex carries out within a cell. The complex data sets generally perform better in function predictions than do the physical interaction data sets.

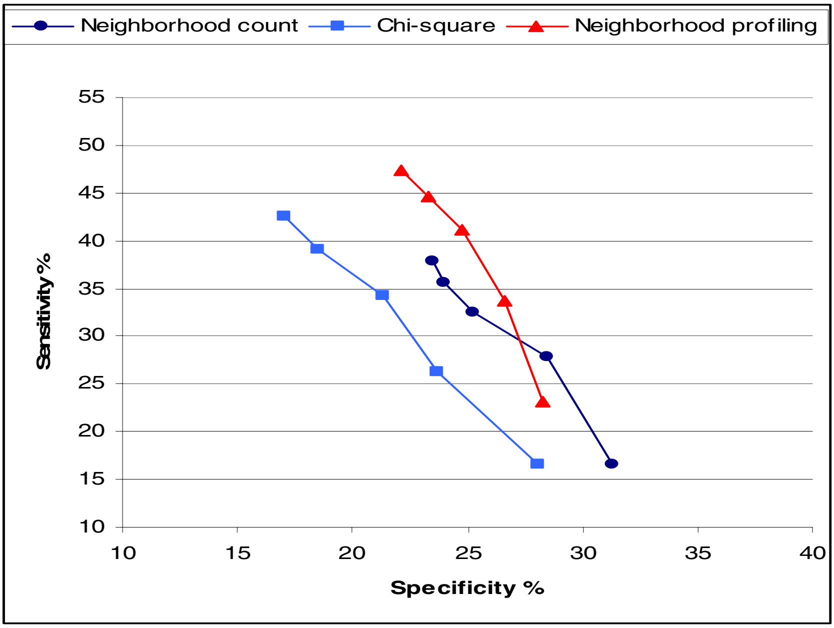

Fig. 4. Sensitivity and specificity of functional prediction using the Uetz data set

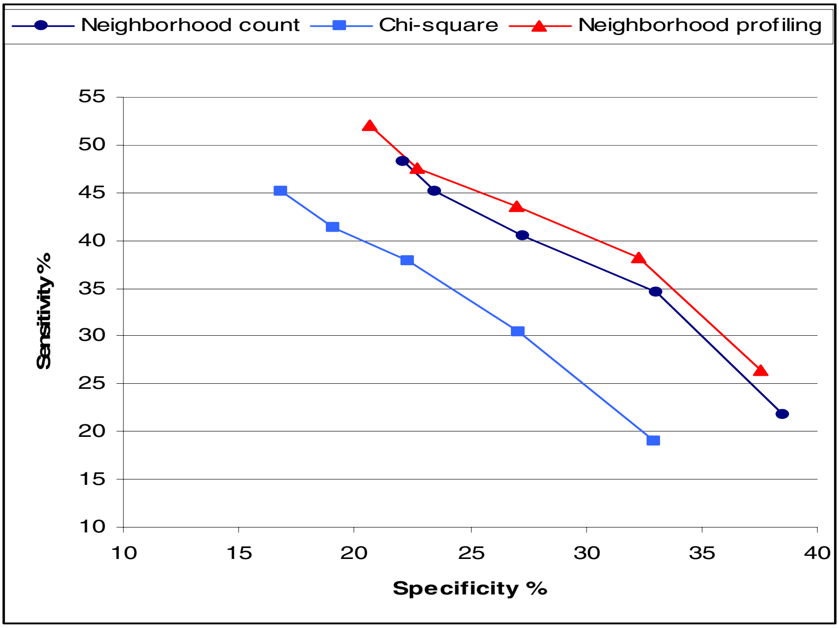

Fig. 5. Sensitivity and specificity of functional prediction using the DIP data set

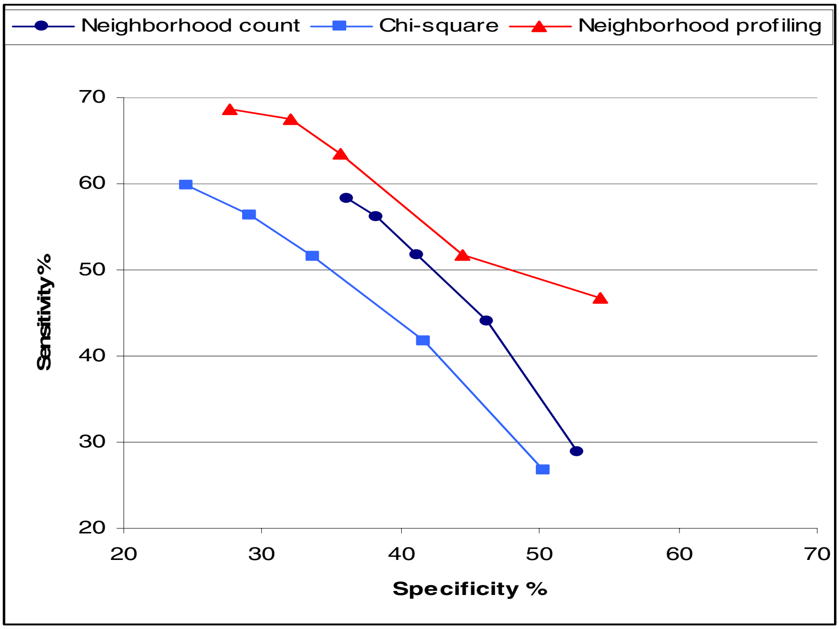

Fig. 6. Sensitivity and specificity of functional prediction using the MIPS physical data set

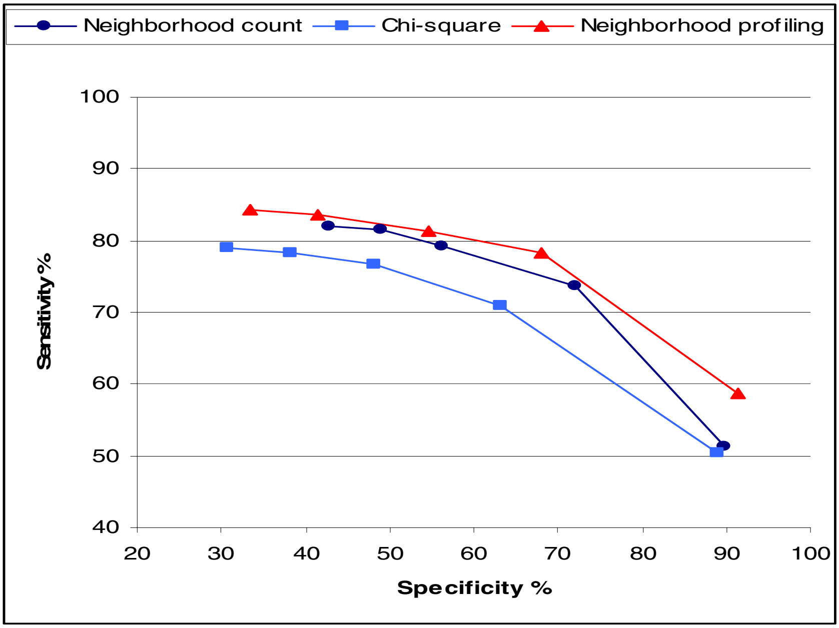

Fig. 7. Sensitivity and specificity of functional prediction using the MIPS complex data set

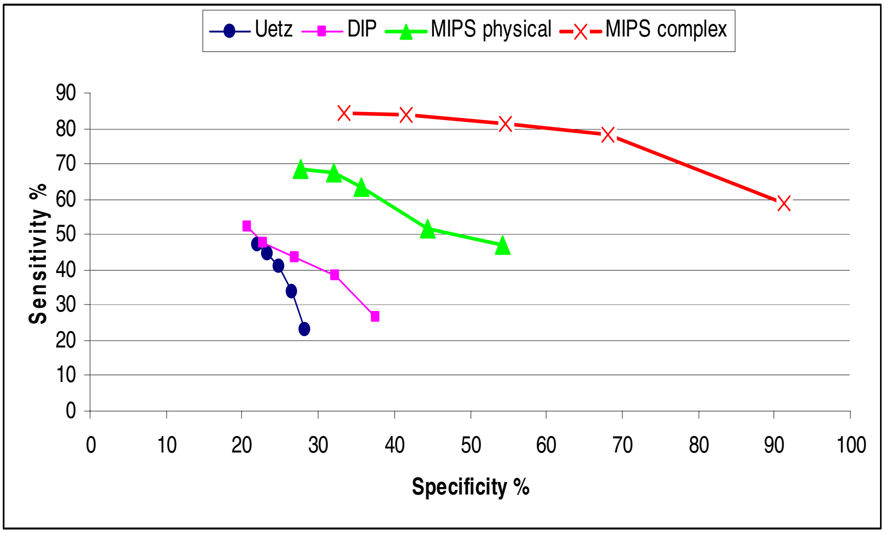

Fig. 8. Sensitivity and specificity of functional prediction for different PIN data sets

4 Conclusion

A new method for protein function prediction using protein interaction networks was presented. It is based on a general hypothesis, since a clique of similar function proteins can be equivalently treated as a set of nodes that observe the same functional neighborhood. We exploit this hypothesis by employing Random Walk with Restarts on the PIN, and extracting the neighborhood profile of an unannotated protein from which we later make the decision of assigning functions to the target by using a "chi-square" method. We validated this two-phase approach by applying it to two different groups of protein interaction data: protein physical interaction data and protein complex data. Experiments revealed that the prediction accuracy of our method outperforms existing techniques by a margin regardless of the data set used. These results are one more proof of the hypothesis that we based our method on. We also assessed the different data sets regarding their reliability on protein function prediction which showed that complex data sets generally perform better than do the physical interaction data sets.

References

1. Yu, G.X., Glass, E.M., Karonis, N.T., Maltsev, N.: Knowledge-based voting algorithm for automated protein functional annotation. PROTEINS: Structure, Function, and Bioinformatics 61, 907–917 (2005)
2. The gene ontology consortium: Gene ontology: Tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)

3. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular Systems Biology 3, 88 (2007)
4. Maciag, K., Altschuler, S., Slack, M., Krogan, N., Emili, A., Greenblatt, J., Maniatis, T., Wu, L.: Systems-level analyses identify extensive coupling among gene expression machines. Molecular Systems Biology 2, 2006.0003 (2006)
5. Spirin, V., Mirny, L.: Protein complexes and functional modules in molecular networks. PNAS 101, 12123–12128 (2003)
6. Dunn, R., Dudbridge, F., Sanderson, C.: The use of edge-betweenness clustering to investigate the biological function in protein interaction networks. BMC Bioinformatics 6, 39 (2005)
7. Arnau, V., Mars, S., Marin, I.: Iterative clustering analysis of protein interaction data. Bioinformatics 21(3), 364–378 (2005)
8. Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., Jacq, B.: Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biology 5, R6 (2003)
9. Samanta, M.P., Liang, S.: Predicting protein functions from redundancies in large-scale protein interaction networks. PNAS 100, 12579–12583 (2003)
10. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19, i197–i204 (2003)
11. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nature Biotechnology 18, 1257–1261 (2000)
12. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18, 523–531 (2001)
13. Chua, H., Sung, W., Wong, L.: Exploiting indirect neighbors and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22(13), 1623–1630 (2006)
14. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21, i302–i310 (2005)
15. Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R., Kasif, S.: Whole-genome annotation by using evidence integration in functional-linkage networks. PNAS 101, 2888–2893 (2004)
16. Han, J., Bertin, N., Hao, T., et al.: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430, 88–93 (2004)
17. Uetz, P., et al.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627 (2000)
18. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M., Eisenberg, D.: DIP: The Database of Interacting Proteins: 2001 update. Nucleic Acids Res. 29, 239–241 (2001)
19. Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., Weil, B.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30, 31–34 (2002)