

Assessing and Predicting Protein Interactions Using Both Local and Global Network Topological Metrics

Guimei Liu¹ Jinyan Li² Limsoon Wong¹
liugm@comp.nus.edu.sg jyli@ntu.edu.sg wongls@comp.nus.edu.sg

¹*School of Computing, National University of Singapore, Singapore*

²*School of Computer Engineering, Nanyang Technological University, Singapore*

High-throughput protein interaction data, with ever-increasing volume, are becoming the foundation of many biological discoveries. However, high-throughput protein interaction data are often associated with high false positive and false negative rates. It is desirable to develop scalable methods to identify these errors. In this paper, we develop a computational method to identify spurious interactions and missing interactions from high-throughput protein interaction data. Our method uses both local and global topological information of protein pairs, and it assigns a local interacting score and a global interacting score to every protein pair. The local interacting score is calculated based on the common neighbors of the protein pairs. The global interacting score is computed using globally interacting protein group pairs. The two scores are then combined to obtain a final score called *LGTweight* to indicate the interacting possibility of two proteins. We tested our method on the DIP yeast interaction dataset. The experimental results show that the interactions ranked top by our method have higher functional homogeneity and localization coherence than existing methods, and our method also achieves higher sensitivity and precision under 5-fold cross validation than existing methods.

Keywords: protein-protein interaction; network topology

1. Introduction

Protein-protein interactions play a critical role in most cellular processes and form the basis of biological mechanisms. Protein interactions have been traditionally studied on an individual basis, which is accurate but is often slow and laborious. In the past several years, high-throughput experimental techniques—such as yeast two-hybrid assay, mass spectrometry, protein chip and phage display—have been introduced to detect a large number of interactions simultaneously, which enables the study of protein-protein interactions at the proteome scale. However, high-throughput protein interaction data are often associated with high false positive and false negative rates due to the limitations of the associated experimental techniques and the dynamic nature of protein interaction maps. It is therefore desirable to develop computational methods to identify these errors.

Many computational approaches have been proposed to assess the reliability of high-throughput protein interaction data or predict new protein interactions. Various information has been used in these approaches, including protein primary struc-

tures and associated physicochemical properties [1], 3D structures of protein complexes [10], gene fusion [18], protein domains [13, 14], literature [23], co-localization information [8] and co-evolution information [11, 22]. Every method for protein interaction assessment and prediction is limited by the availability and reliability of the information it uses, and methods using different information sources are complementary to one another. Some work integrates multiple information sources to achieve better performance [12, 20].

Recent screening techniques have made large amounts of protein-protein interaction data available, which makes it possible to assess or predict protein interactions using solely the topology of the protein interaction networks [4, 5, 24, 25, 29]. Saito et al. [24, 25] introduced two measures called IG1 and IG2 which use the local topological structure of protein pairs to assess their reliability, and they do not consider topological properties beyond the candidate protein pair and their neighbors. Chen et al. [4] proposed a more global measure called IRAP, which is defined as the collective reliability of the strongest alternative path between two proteins. The authors later improved the IRAP measure by iteratively removing low-confidence interactions from the network and adding high-confidence new interactions to the network [5]. Yu et al. [29] proposed a method to predict new protein interactions by completing defective cliques. Chua et al. [6] proposed a measure called FSweight which exploits indirect neighbors to predict protein functions. The same group of authors later showed that FSweight could also be used to predict protein interactions and it outperformed IG1, IG2 and IRAP on large interaction datasets [3].

FSweight is still a local measure. In this paper, we propose a computational method which uses both local topological information of protein pairs and global topological structures discovered from the whole interaction network to assess and predict protein interactions. The local interacting score of a protein pair is calculated based on the neighbors of the two proteins, and the reliability of the interactions between these two proteins and their neighbors is also taken into consideration. The global interacting score is obtained based on the observation that if one group of proteins interact with another group of proteins, then it is likely that the interaction between these two protein groups is mediated by an underlying complementary binding domain/motif pair. The above observation has been used to discover interacting motif pairs [16, 19, 27]. We call such protein group pairs interacting protein group pairs. If a protein pair participates in an interacting protein group pair, that is, the two proteins belong to different groups of the interacting protein group pair, then the interaction between the two proteins is likely to be true.

To calculate global interacting scores, we first generate groups of proteins that have common interacting partners from the interaction network using frequent item-set mining techniques, and then for every pair of discovered protein groups, we calculate their interacting scores. The global interacting score of a protein pair is computed based on the interacting score of the interacting group pairs it participates in and the degree of its participation. We studied the performance of our method on the DIP yeast interaction dataset. Our experiment results showed that

our method outperforms FSweight, especially for predicting new interactions.

The rest of the paper is organized as follows. Section 2 describes our method, and the experiment results on the DIP yeast interaction dataset are presented in Section 3. Section 4 discusses and concludes the paper.

2. Method

In this section, we first describe how to calculate local interacting scores and global interacting scores of protein pairs, and then discuss how to combine them together to get the final score. The following notations are used in this section. A protein interaction network can be modeled as an undirected graph $G = (V, E)$ where vertex set V is the set of proteins and edge set E is the set of interactions between proteins. We use u, v, x to denote individual vertices (proteins), V_1, V_2 to denote vertex sets (protein groups), and (u, v) to denote the edge between u and v . The neighbor set of a vertex u in G , denoted as N_u , is defined as $N_u = \{v | (u, v) \in E\}$.

2.1. Local interacting score

The local interacting score is defined based on the observation that if two proteins have many common neighbors, then these two protein are likely to interact with each other. We use a variant of the CD-distance measure to calculate local interacting score of protein pairs. The CD-distance measure was originally proposed by Brun et al. [2] for function prediction, and later was shown to be very effective in assessing the reliability of high-throughput interaction data [3].

It has been estimated that more than half of current high-throughput data are spurious [15, 26, 28], and these spurious interactions usually have a low score. To alleviate the impact of spurious interactions, we iteratively apply the scoring method on the weighted interaction network. The local interacting score of a protein pair in the k -th ($k > 0$) iteration, denoted as $w_L^k(u, v)$, is defined as follows:

$$w_L^k(u, v) = \frac{\sum_{x \in N_u \cap N_v} w_L^{k-1}(x, u) + \sum_{x \in N_u \cap N_v} w_L^{k-1}(x, v)}{\sum_{x \in N_u} w_L^{k-1}(x, u) + \sum_{x \in N_v} w_L^{k-1}(x, v) + \lambda_u^k + \lambda_v^k} \quad (1)$$

where $w_L^{k-1}(x, u)$ is the score of (x, u) in the $(k-1)$ -th iteration, $w_L^0(x, u)=1$ if $(x, u) \in E$ and $w_L^0(x, u)=0$ if $(x, u) \notin E$. The two terms, λ_u^k and λ_v^k , are used to penalize proteins with very few neighbors (as in [6]), and they are defined as follows:

$$\lambda_u^k = \max\{0, \frac{\sum_{x \in V} \sum_{v \in N_x} w_L^{k-1}(v, x)}{|V|} - \sum_{v \in N_u} w_L^{k-1}(v, u)\} \quad (2)$$

When $k=1$, the local interacting score is similar to the CD-distance score except that it uses λ_u^1 and λ_v^1 to penalize proteins with very few neighbors. In our experiments, we have found that the local interacting score reaches the best performance when $k=2$, and the subsequent iterations do not improve the performance further.

2.2. Global interacting score

The global interacting score is based on the observation that if one group of proteins interact with another group of proteins, then it is likely that the interaction between these two protein groups is mediated by an underlying complementary binding domain/motif pair. We call such protein group pairs *interacting protein group pairs*. Given a protein pair (u, v) and an interacting protein group pair (V_1, V_2) , we say (V_1, V_2) *contains* (u, v) if $u \in V_1$ and $v \in V_2$. We also say that (u, v) *participates in* the interacting protein group pair (V_1, V_2) . If a protein pair participates in an interacting protein group pair whose two groups are densely connected, then the interaction between these two proteins is likely to be true.

Proteins on one side of an interacting group pair are expected to have some common domains or motifs, so we expect that they have some common interacting partners. Also it is not desirable to have very few proteins on either side of an interacting group pair, because otherwise, the underlying interacting domain/motif pair may not be significant. Here we use two thresholds *min_sup* and *min_size* to restrict the minimum number of common neighbors and the minimum size of a protein group. We call *min_sup* the minimum support threshold and *min_size* the minimum size threshold. For an interacting protein group pair, each of its two protein groups must have at least *min_sup* common neighbors and contain at least *min_size* proteins.

The calculation of global interacting scores of protein pairs consists of three steps. In the first step, protein groups that have at least *min_sup* common interacting partners and contain at least *min_size* proteins are generated. In the second step, the interacting score of every pair of discovered protein groups is calculated. In the last step, the global interacting score of a protein pair is computed.

2.2.1. Generating protein groups

The protein groups that have at least *min_sup* common interacting partners and contain at least *min_size* proteins are generated using frequent itemset mining techniques. The adjacency matrix of an undirected graph can be regarded as a transaction database where each adjacency list is a transaction and each vertex (protein) is an item. The support of an itemset (protein group) is defined as the number of transactions (adjacent lists) containing it, which is equal to the number of common partners of the corresponding protein group. Finding protein groups that have at least *min_sup* common interacting partners and contain at least *min_size* proteins is equivalent to finding frequent itemsets occurring in at least *min_sup* transactions and containing at least *min_size* items. Frequent itemset mining algorithms use the anti-monotone property of itemsets to prune the search space, that is, if an itemset appears in less than *min_sup* transactions, then all of its supersets also appear in less than *min_sup* transactions, thus the itemset can be pruned. Given that the adjacency matrix of a protein interaction network is usually sparse, frequent itemset mining algorithms can generate the desired protein groups within several minutes.

In this paper, we use the AFOP algorithm [17] to generate the protein groups.

2.2.2. Calculating interacting confidence score of protein group pairs

Let V_1 and V_2 be two protein groups generated in the first step. The interacting confidence score of V_1 and V_2 , denoted as $conf(V_1, V_2)$, is defined as the ratio of the number of interactions between V_1 and V_2 to the total number of distinct protein pairs contained in (V_1, V_2) :

$$conf(V_1, V_2) = \frac{|\{(u, v) | (u, v) \in E, u \in V_1, v \in V_2\}|}{|V_1| \cdot |V_2| - |V_1 \cap V_2| \cdot (|V_1 \cap V_2| + 1)/2} \quad (3)$$

When calculating the total number of distinct protein pairs contained in (V_1, V_2) , we need to consider the situation that V_1 and V_2 may contain some common proteins. In the simple case that the two protein groups contain no common proteins, the total number of distinct protein pairs contained in (V_1, V_2) is simply $|V_1| \cdot |V_2|$. Otherwise, among the $|V_1| \cdot |V_2|$ protein pairs, there are $|V_1 \cap V_2|$ self-interactions and $|V_1 \cap V_2| \cdot (|V_1 \cap V_2| - 1)/2$ duplicated protein pairs, and they should be discarded. Therefore, the total number of distinct protein pairs contained in (V_1, V_2) is $|V_1| \cdot |V_2| - |V_1 \cap V_2| - |V_1 \cap V_2| \cdot (|V_1 \cap V_2| - 1)/2 = |V_1| \cdot |V_2| - |V_1 \cap V_2| \cdot (|V_1 \cap V_2| + 1)/2$.

2.2.3. Calculating global interacting score of protein pairs

The global interacting score of a protein pair is computed based on the interacting confidence score of the interacting group pairs it participates in and the degree of its participation, and it is defined as follows:

$$w_G(u, v) = \max\{conf(V_1, V_2) \cdot \frac{2|N_u \cap V_2|}{|V_2| + |N_u|} \cdot \frac{2|N_v \cap V_1|}{|V_1| + |N_v|} | u \in V_1, v \in V_2\} \quad (4)$$

where $\frac{2|N_u \cap V_2|}{|V_2| + |N_u|}$ and $\frac{2|N_v \cap V_1|}{|V_1| + |N_v|}$ are the participation degree of protein u and v respectively.

2.3. The final interacting score of protein pairs

The final interacting score of a protein pair is simply defined as the sum of its local interacting score and its global interacting score. For local interacting scores, we set $k = 2$.

$$LGTweight(u, v) = w_L^2(u, v) + w_G(u, v). \quad (5)$$

The higher the interacting score is, the more likely the two proteins interact with each other. After the interacting scores of the protein pairs are calculated, we rank the protein pairs in descending order of their score.

3. Results

In this section, we study the performance of our method and compare it with FSweight [6] and the original CD-distance [2]. We used the DIP (<http://dip.doe-mbi.ucla.edu/>) yeast interaction dataset dated 10/07/2007 in our experiments, which contains 17491 interactions. After removing duplicate interactions and self-interactions, the dataset contains 4932 distinct proteins and 17201 interactions. The DIP yeast core dataset contains 6459 interactions that were validated according to the criteria described in [9], and it is used as golden standard in our experiments.

3.1. *Functional homogeneity and localization coherence*

By the “guilt-by-association” principle [21], true interacting proteins usually share some common functional role or are in the same cellular components. Hence we use the degree of functional homogeneity and localization coherence of protein pairs as one of the measures to evaluate our method. The interacting score of a protein pair indicates the interacting possibility of the protein pair. The higher the score is, the more likely the two proteins interact with each other. If we use a cut-off value *min_score* to select the protein pairs with score no less than *min_score* as interacting protein pairs, we expect that the proportion of the protein pairs sharing common functions or localizations in the selected protein pairs increases with the increase of *min_score*.

We use the annotations in Gene Ontology (GO) (<http://www.geneontology.org/>) to calculate functional homogeneity and localization coherence. The Gene Ontology comprises three orthogonal taxonomies or aspects that hold terms describing molecular functions, biological processes and cellular components of a gene product. We use the terms in the first two taxonomies for functional homogeneity calculation, and the terms in the last taxonomy for localization coherence calculation. The GO terms are organized hierarchically. Two different GO terms may share a common parent or a common child in the hierarchy. GO terms at high levels may occur in many proteins, and they are too common to be useful. GO terms appearing in very few proteins are also not very useful. In our experiments, we select only those informative GO terms. A GO term is informative if itself occurs in at least 30 proteins, but none of its children appears in at least 30 proteins. Using the proteins in the DIP yeast dataset, 50 molecular function terms, 110 biological process terms and 42 cellular component terms are selected.

Among the 4932 proteins in the DIP yeast dataset, 3251 proteins have functional annotations. There are 11229 interactions whose two proteins both have functional annotations, and among them 3660 interactions have common function annotations between its two proteins. We consider only those protein pairs whose two proteins both have functional/localization annotations when calculating the degree of functional homogeneity and localization coherence of a set of protein pairs. Thus the degree of functional homogeneity of the DIP yeast interaction dataset is 32.6%. The

overall functional homogeneity of all the possible protein pairs is 3.4%. There are 1615 proteins with cellular component annotations and 4246 interactions whose two proteins both have localization annotations. Among them, 2321 interactions have common localization annotations between its two proteins, so the degree of localization coherence of the DIP yeast dataset is 54.7%. The overall localization coherence over all protein pairs is 4.9%.

3.1.1. The effect of the number of iterations on local interacting scores

Our first experiment is to study the effect of the number of iterations on the performance of local interacting scores. Figure 1(a) shows the degree of functional homogeneity of the interactions in the DIP dataset ranked using local interacting scores under different k values. It shows that the local interacting score reaches the best performance when $k=2$. The subsequent iterations do not improve the performance much. We use local interacting scores to rank the protein pairs that are not in the DIP dataset and select the top ranked protein pairs as predicted new interactions. Figure 1(b) shows the degree of functional homogeneity of these new interactions ranked under different k values. Again, the performance of the local interacting score reaches the best when $k=2$. We also observed the same trend using localization coherence. In the following experiments, we set $k=2$.

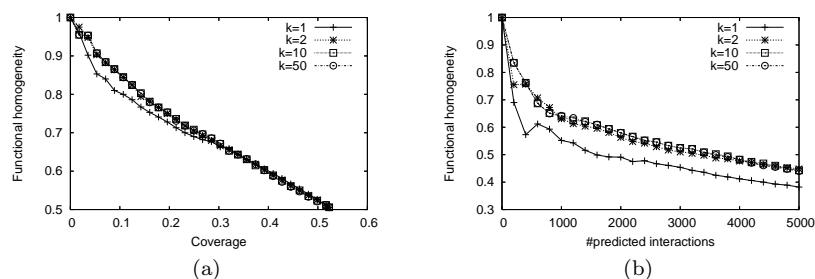


Fig. 1. The effect of the value of k (a) interactions in the DIP yeast dataset (b) New interactions predicted.

3.1.2. Assessing and predicting interactions

Our second experiment is to compare the performance of our method with that of FSweight and CD-distance in terms of functional homogeneity and localization coherence. When calculating global interacting scores, we set $min_sup=1$ and $min_size=5$. More specifically, the generated protein groups have at least one common neighbors and contains at least five proteins. Frequent itemset mining algorithms use the minimum support threshold min_sup to prune the search space. Here the value of min_size is larger than that of min_sup , so we swapped the values of the two thresholds and used min_size as the minimum support threshold to first

find the partner groups of the desired groups, and then generated the desired protein groups in a post-processing step. The time used for generating the protein groups is less than one minute on a PC with 2.33GHz CPU.

Every protein group pair has an interacting confidence score. In our experiments, we retained only those protein group pairs with a confidence score no less than 0.1. We assessed the significance of these retained protein group pairs as follows. For a protein group pair (V_1, V_2) , we randomly generate 1000 protein group pairs (V'_1, V'_2) such that $|V'_1| = |V_1|$, $|V'_2| = |V_2|$ and $|V'_1 \cap V'_2| = |V_1 \cap V_2|$. We then calculate the interacting confidence score of those random protein group pairs, and use the percentage of those random group pairs whose confidence score is no less than $conf(V_1, V_2)$ to approximate the p-value of (V_1, V_2) . We have found that the p-value of all of the retained protein group pairs is no larger than 0.005.

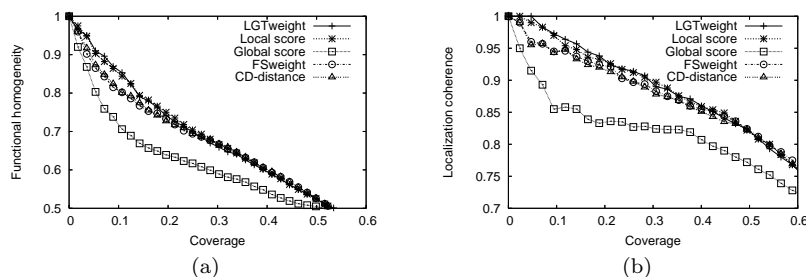


Fig. 2. Performance of the four methods in assessing reliability of interactions (a) functional homogeneity (b) localization coherence

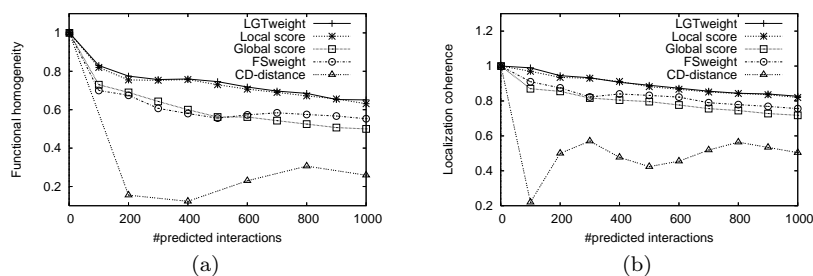


Fig. 3. Performance of the four methods in predicting new interactions (a) functional homogeneity (b) localization coherence

Figure 2(a) shows the functional homogeneity and localization coherence of the interactions in the DIP datasets ranked using five methods: LGTweight, local interacting score, global interacting score, FSweight and CD-distance. Protein pairs

ranked by global interacting score show lower functional homogeneity and localization coherence than those ranked by other methods. The reason being that local interacting score, FSweight and CD-distance rank protein pairs based on their level-1 neighbors, and proteins are more likely to share common functions or localizations with their level-1 neighbors than with other proteins. The global interacting score ranks protein pairs based on interacting protein group pairs. A protein pair contained in an interacting group pair may have no common neighbors at all. The local interacting score performs better than FSweight and CD-distance, and its performance can be improved when combined with global interacting score.

Figure 3 shows the functional homogeneity and localization coherence of the new interactions predicted by the five methods. CD-distance performs the worst among the five methods. The global interacting score still performs worse than FSweight, but the gap between it and FSweight becomes smaller. Local interacting score and LGTweight perform significantly better than FSweight. LGTweight performs better than local interacting score due to the use of global interacting score.

3.2. Five-fold cross validation

Our last experiment is to study the performance of our method using five-fold cross validation. Here we use the DIP yeast core dataset as the golden standard. We divide the proteins into five disjoint groups. For each group, we remove the interactions between proteins in that group, and use the remaining interactions as the training dataset. The testing dataset contains all the possible pairs of proteins in the group. The removed interactions that are in the DIP yeast core dataset are regarded as the correct answers. The number of proteins in each of the five groups is 986, the average number of interactions in the five training datasets is 16723, the number of testing interactions in each of the five testing datasets is 486591 and the average number of correct-answer interactions is 307.

Sensitivity and specificity are two commonly used measures to assess prediction algorithms, and they are defined as follows.

$$sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$specificity = \frac{TN}{TN + FP} \quad (7)$$

where TP (True Positive) is the number of true interacting protein pairs that are also predicted to be interacting, FN (False Negative) is the number of true interacting protein pairs that are predicted to be non-interacting, TN (True Negative) is the number of non-interacting protein pairs that are predicted to be non-interacting, and FP (False Positive) is the number of non-interacting protein pairs that are predicted to be interacting.

In our testing data, the number of non-interacting protein pairs is orders of magnitude larger than the number of interacting protein pairs. Only 0.063% testing

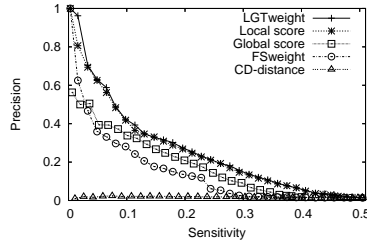


Fig. 4. Sensitivity vs. precision.

protein pairs are considered as truly interacting. In this case, the specificity of an algorithm can be always very high. In our experiments, the specificity of all the algorithms is no less than 97.8% when they reach their maximal sensitivity. To have a clearer comparison of the algorithms, here we use another measure called precision to assess the algorithms, and it is defined as the percentage of true interactions among all the predictions made by the algorithms.

$$precision = \frac{TP}{TP + FP} \quad (8)$$

Figure 4 shows the precision of the five methods with respect to their sensitivity. CD-distance shows very poor performance. FSweight performs worse than the other three methods. Under the same sensitivity, the precision of FSweight is lower than that of the other methods. It indicates that FSweight makes more false positive predictions than other methods. However, the maximal sensitivity achieved by FSweight is 50.5%, which is higher than the other methods. The maximal sensitivity achieved by LGTweight is 49.9%, which is higher than that of local interacting score (46.3%) and global interacting score (40.0%). Under the same sensitivity, the precision of LGTweight is also higher than that of local interacting score and global interacting score, which shows that by combining local interacting score and global interacting score, we can obtain both higher sensitivity and higher precision than using them alone.

Note that here we regard only those interactions in the DIP core dataset as true interactions. However, some interactions not in the core dataset may be true interactions, so using the core dataset as the golden standard may underrate the performance of the methods. The actual performance of the methods tested should be better than what reported here.

4. Discussion and Conclusion

In this paper, we have proposed a computational approach to assessing and predicting protein interactions. The proposed method uses both local topological information of protein pairs and global topological structures discovered from the whole network to calculate interacting scores of protein pairs, and it outperforms existing

methods, especially for predicting new interactions. We used an iterative approach to calculate local interacting scores. We have tried to apply this iterative approach to FSweight, and we also observed a significant improvement on the performance of FSweight. Here we use a simple method to combine the local interacting score and global interacting score of a protein pair. It is possible to use a more sophisticated method to achieve better results.

In this paper, we use only the network topology to assess and predict interactions. It is complementary to those methods using other information for assessing and predicting protein interactions. The performance of our method, and other methods using solely interaction network topology, is limited by the availability and quality of existing interaction data. Chua et al. [7] proposed a framework for integrating multiple information sources. We can use their method or other methods to integrate other information sources into our approach, or integrate our method with other methods to obtain better results.

Acknowledgments

This research was supported in part by Singapore MOE Tier 1 grant R-252-000-274-112 (Liu, Wong) and in part by NTU Tier 1 grant RG66/07 (Li).

References

- [1] Bock JR, and Gough DA, Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.
- [2] Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, and Jacq B, Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6, 2003.
- [3] Chen J, Chua HN, Hsu W, Lee ML, Ng SK, Saito R, Sung WK, and Wong L, Increasing confidence of protein-protein interactomes. In *Proc. of 17th International Conference on Genome Informatics*, pp. 284–297, 2006.
- [4] Chen J, Hsu W, Lee ML, and Ng SK, Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in Medicine*, 35(1-2):37–47, 2005.
- [5] Chen J, Hsu W, Lee ML, and Ng SK, Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, 22(16):1998–2004, 2006.
- [6] Chua HN., Sung WK., and Wong L., Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–30, 2006.
- [7] Chua, HN., Sung, WK., and Wong L., An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 3(24): 3364–3373, 2007.
- [8] Dandekar, T., Snel, B., Huynen, M., and Bork, P., Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–8, 1998.
- [9] Deane CM., Salwinski L., Xenarios I., and Eisenberg D., Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349–56, 2002.

12 G. Liu, J. Li & L. Wong

- [10] Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, and Gerstein M, Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics*, 18(10):529–536, 2002.
- [11] Goh CS, Bogan AA, Joachimiak M, Walther D, and Cohen FE., Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2):283-93, 2000.
- [12] Gomez SM, and Rzhetsky A, Towards the prediction of complete protein-protein interaction networks. In *Pacific Symposium on Biocomputing*, pp. 413–424, 2002.
- [13] Han D, Kim HS, Seo J, and Jang W, A domain combination based probabilistic framework for protein-protein interaction prediction. *Genome Informatics Series: Workshop on Genome Informatics*, 14:250–259, 2003.
- [14] Kim WK, Park J, and Suh JK, Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. *Genome Informatics Series: Workshop on Genome Informatics*, 13:42–50, 2002.
- [15] Legrain P, Wojcik J, and Gauthier JM, Protein-protein interaction maps: a lead towards cellular functions. *Trends in genetics*, 17(6):346–352, 2001.
- [16] Li H, Li J, and Wong L, Discovery motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996, 2006.
- [17] Liu G., Lu H., Lou W., Xu Y., Yu X. J., Efficient Mining of Frequent Patterns Using Ascending Frequency Ordered Prefix-Tree. *Data Mining and Knowledge Discovery*, 9(3): 249-274, 2004.
- [18] Marcotte EM., Pellegrini M., Ng HL., Rice DW., Yeates TO., and Eisenberg D., Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751-3m, 1999.
- [19] Morrison JL, Breitling R, Higham DJ, and Gilbert DR, A lock-and-key model for protein-protein interactions. *Bioinformatics*, 22(16):2012–2019, 2006.
- [20] Ng SK, Zhang Z, and Tan SH, Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–929, 2003.
- [21] Oliver S, Proteomics: guilt-by-association goes global. *Nature*, 403:601–603, 2000.
- [22] Pazos F, and Valencia A., Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, 14(9):609-14, 2001.
- [23] Ramani AK., Bunescu RC., Mooney RJ., and Marcotte EM., Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):R40, 2005.
- [24] Saito R, Suzuki H, and Hayashizaki Y, Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research*, 30(5):1163–1168, 2002.
- [25] Saito R, Suzuki H, and Hayashizaki Y, Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19(6):756–763, 2002.
- [26] Sprinzak E, Sattath S, and Margalit H, How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003.
- [27] Tan SH., Hugo W., Sung WK., and Ng SK., A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, 7:502, 2006.
- [28] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, and Bork P, Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.
- [29] Yu H, Paccanaro A, Trifonov V, and Gerstein M, Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, 2006.