Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 523634, 6 pages
http://dx.doi.org/10.1155/2014/523634

Research Article

Protein-Protein Interactions Prediction Based on Iterative Clique Extension with Gene Ontology Filtering

Lei Yang and Xianglong Tang

School of Computer Science and Technology, Harbin Institute of Technology, Mailbox 352, 92 West Dazhi Street, Nan Gang District, Harbin 150001, China

Correspondence should be addressed to Lei Yang; yangl@hit.edu.cn

Received 24 August 2013; Accepted 5 November 2013; Published 22 January 2014

Academic Editors: C. Angelini and Y. Lai

Cliques (maximal complete subnets) in protein-protein interaction (PPI) network are an important resource used to analyze protein complexes and functional modules. Clique-based methods of predicting PPI complement the data defection from biological experiments. However, clique-based predicting methods only depend on the topology of network. The false-positive and false-negative interactions in a network usually interfere with prediction. Therefore, we propose a method combining clique-based method of prediction and gene ontology (GO) annotations to overcome the shortcoming and improve the accuracy of predictions. According to different GO correcting rules, we generate two predicted interaction sets which guarantee the quality and quantity of predicted protein interactions. The proposed method is applied to the PPI network from the Database of Interacting Proteins (DIP) and most of the predicted interactions are verified by another biological database, BioGRID. The predicted protein interactions are appended to the original protein network, which leads to clique extension and shows the significance of biological meaning.

1. Introduction

Identifying protein-protein interaction (PPI) and constructing biological networks are vital to understand the molecular function and cellular organization [1]. In recent year, protein interactions have been enriched by high-throughput biology experimental methods [2, 3]. Furthermore, there are a large number of computational methods to improve these datasets and overcome the experimental limitations of time-consuming and cost [4–6]. The categories of computational methods for predicting PPI differ among studies. For example, in paper [5], methods of predicting PPI are classified according to the structural, genomic, and biological contexts of proteins. Of them, topology methods based on PPI networks are simple to use and demand few additional information. However, the reliability based on topology methods usually fluctuates because of false positive and negative interactions in PPI network. A clique within a PPI network is an induced complete subgraph, with constituent proteins that completely interact with each other. Cliques in PPI networks are related to protein complexes and functional modules tightly and have a biological significance [6]. And components in protein complexes or functional modules are prone to interacting with each other. Thus, protein interactions can be predicted based on completing defective cliques in a PPI network. Besides, the estimation based on gene ontology (GO) annotations can enhance the accuracy of PPI predictions [7]. This is because PPIs from cliques usually have common terms in GO annotations of cellular component (CC) or molecular function (MF), due to the correlation of cliques with complexes or functional modules. Therefore, we design a predicting PPI method combining cliques and GO terms to overcome the unreliability of predicting PPI based on topology methods with the interference of noise in datasets.

Applying different GO rules, we generate two predicted PPI sets, that is, CORE and ALL which insure the quality and quantity of predictions, respectively. The proposed method is applied on DIP dataset [8]. The two predicted sets are estimated with a statistical method based on gold standard datasets [9], and the results show the effectiveness of our method. Furthermore, we introduce another dataset, BioGRID [10], recording a larger number PPIs from biological experiments to verify the correction of the predictions.

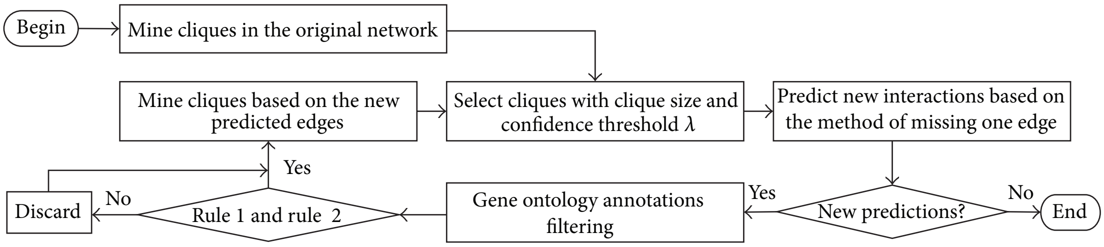

Begin
Mine cliques in the original network
Mine cliques based on the new predicted edges
Select cliques with clique size and confidence threshold λ
Predict new interactions based on the method of missing one edge
Yes
Discard
No
Rule 1 and rule 2
Gene ontology annotations filtering
Yes
New predictions?
No
End

Figure 1: Flowchart of predicting protein interactions in a PPI network.

As a result, most of predictions are identified with the various records of biological experiments. Appending the predictions into the PPI network, the mined cliques are close to complete complexes.

2. Materials and Methods

A PPI network is used to describe a set of protein pairs, and it can be modeled as a graph. We use an undirected graph $G = (V, E)$ to represent a PPI network, where V is the set of all nodes (proteins) and E is the set of all edges (interactions) in graph G.

We propose an iterative method based on cliques and GO annotations to predict PPI (see Figure 1). First, clique mining is performed in a PPI network. Second, cliques satisfied with two conditions are selected to participate in PPI prediction. Third, the missing-one-edge method is used to predict interactions based on the mined cliques. Then, the predictions are corrected via gene ontology annotations. This process is repeated until no new prediction of PPI generates.

Mining cliques in a PPI network is a NP-complete problem according to graph computing theory [11]. However, the existing methods of mining cliques work well based on the characteristic of scale-free topology [12]. Based on an original network, the method of Gendreau et al. [13] is brought in to obtain all cliques. Then, new interactions are predicted based on these cliques.

2.1. Selection of Cliques Participated in PPI Prediction. Not all cliques participate in the process of predicting PPI. We set a minimum threshold of clique size. Only the cliques whose sizes are more than that threshold have the qualification to participate in PPI prediction. Furthermore, we set a clique confidence score to avoid that the selected cliques include many predicted PPIs. The clique confidence is determined by sum score of all the edges in it. That is,

$$\text{Clique}_{\text{score}} = \frac{\sum_{e \in k\text{-clique}} e_{\text{score}}}{k(k-1)/2}, \qquad \text{where } e_{\text{score}} = \begin{cases} 1 & \text{original} \quad \text{PPI} \\ 0 & \text{predicted} \quad \text{PPI.} \end{cases} \tag{1}$$

We set a threshold λ of the clique score for PPI prediction. Obviously, cliques selected as the participants for predicting PPI must satisfy with the two mentioned thresholds.

2.2. Prediction of PPIs with Missing-One-Edge Method. We count the number of the nodes connected with the members of the known cliques. Based on a known k-clique, the appending of one edge extends it to $(k+1)$-clique(s). Corresponding to the direct connection, the missing edge that links the nodes in and out of the clique to form a larger clique is a predicted interaction. As to a given k-clique, the description of prediction with missing-one-edge method is as below.

2.2.1. Building a Candidate Set C to Deposit All Possible Extending Clique Nodes. $C = N(n_1) \cup N(n_2) \cup \cdots \cup N(n_k) - \{n_1, n_2, \ldots, n_k\}$, where n_i is the node in k-clique and $N(n_i)$ represents the neighbor set of n_i.

2.2.2. Filtering Nodes in the Candidate Set C. $C \leftarrow C - \{n \mid \text{Degree}(n) < k-1, n \in C\}$, where Degree($n$) represents the number of the neighbors of node n.

2.2.3. Predicting

(1) Getting a node w in C.

(2) Judging whether $\exists i$ making $n_i w \notin E_{\text{all}}$ and $\forall j, n_j w \in E_{\text{all}}$, $j = 1, 2, \ldots, k$, $j \neq i$, where $E_{\text{all}} = E \cup \{\text{predictions}\}$.

(3) If condition (2) is satisfied, $n_i w$ is a new predicted PPI.

(4) Repeating (1) until all of the nodes in C have been reached.

2.3. Correction of Predicted PPIs by GO Terms. Proteins in a complex always have same cellular components, and proteins in a function module often perform similar biological roles. Therefore, GO terms contribute to reducing false-positive predictions [14] before the next round of clique mining based on the predictions. We set two GO rules as follows.

Rule 1. Two proteins in the predicted interaction should have a common cellular component at least.

Rule 2. Two proteins in the predicted interaction should have a common molecular function at least.

If only Rule 1 is used to filter the predicted PPIs, a loose PPI prediction set (ALL) is gotten. If both Rules 1 and 2 are applied to filter the predicted PPIs, a tight prediction set (CORE) is obtained.

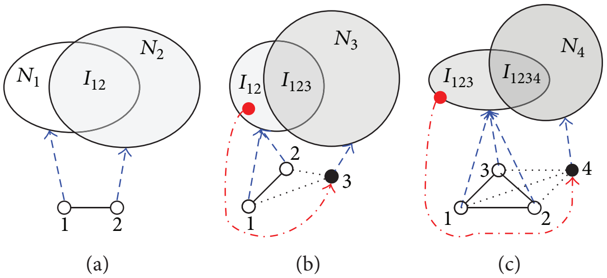


Figure 2: An illustration of mining cliques based on predicted protein pairs. (a) Components of intersection I_{12} between the neighbors of node 1 and node 2 are candidates to form 3 cliques. (b) Node 3 derived from I_{12} with nodes 1 and 2 forms 3 cliques. The neighbors of node 3 intersect I_{12} and construct the candidate set I_{123} of 4 cliques. (c) The node in the I_{123} composes of nodes 1, 2, and 3 to form 4 cliques. The neighbors of node 4 derived from I_{123} intersect I_{123} and construct the candidate set I_{1234} for forming 5 cliques. This procedure is repeated until the maximum clique is found based on node 1 and node 2.

2.4. Clique Generation Based on New Predicted Interactions. After we obtain the predicted PPIs of the first round, the new predictions are merely related with the cliques including interactions just predicted in the previous round. Therefore, we design an appropriate method to generate cliques only based on the predicted interactions to minimize the computational cost. Mining cliques is a recursively bottom-up procedure, as illustrated in Figure 2. First, all 3 cliques are found out based on the new predicted PPIs (see Figure 2(a)). Then, all 4 cliuqes are found out based on the 3 cliques (see Figures 2(b) and 2(c)). In turn, k-cliques are mined based on the $(k-1)$-cliques until the maximum cliques are found.

3. Results and Discussion

3.1. Performance on DIP Dataset. Our method is applied on a dataset of PPIs in *Saccharomyces cerevisiae* downloaded from the Database of Interacting Proteins (DIP, version of 2010/6/14) [8]. DIP is generally acknowledged as an excellent data source containing experimentally determined protein-protein interaction. This version dataset contains 26718 interactions. Getting rid of proteins of self-interacting, we achieve 4997 protein nodes and 23233 protein-protein interaction pairs from DIP database.

The maximum size of mined clique is 10. According to this maximum size, we set the minimum size of clique participating in predicting PPI to be 6. The threshold of clique confidence score is set to be 0.7. Based on the PPI network of the original DIP dataset, 442 PPI predictions are generated. Then, according to the cliques derived from these predictions and different GO rules, CORE and ALL are generated, respectively. CORE includes 352 predictions and ALL contains 874. Adding these predictions into the original dataset, the size of the maximum clique is enlarged to 16 and the number of small cliques is reduced, that is, consolidation of smaller cliques into larger ones.

3.2. Estimation with BioGRID Dataset. We expect predictions in CORE and ALL to be validated by the other records of biological experiments. Database BioGRID collects sufficient and reliable data in *Saccharomyces cerevisiae* from primary literatures. Therefore, we compare CORE and ALL with the dataset of BioGRID (version 3.2.98) and find that the overlaps between them are very high. Of 352 predictions in CORE, 323 are found in BioGRID and the overlapping ratio is near 92%. Of 874 predictions in ALL, 774 are found in BioGRID and the overlapping ratio is nearly 89%. The process of predicting and validating PPI in each round is shown in Tables 1 and 2.

Table 1: Prediction steps of CORE validated with BioGRID.

Round	Max. clique	Predicted PPIs	PPIs after GO	In BioGRID	Confirmed ratio
1	10	442	149	143	0.96
2	13	421	130	114	0.88
3	15	154	64	57	0.89
4	15	29	8	8	1
5	11	6	1	1	1

Table 2: Prediction steps of ALL validated with BioGRID.

Round	Max. clique	Predicted PPIs	PPIs after GO	In BioGRID	Confirmed ratio
1	10	442	368	335	0.91
2	16	539	357	305	0.85
3	15	161	122	109	0.89
4	15	21	19	17	0.89
5	13	4	4	4	1
6	9	4	4	4	1

3.3. Estimation with Statistics. A statistical method based on gold standard (GS) datasets is introduced to estimate the PPI predictions [9], where the GS is essentially a likelihood ratio L, that is, $L = (P_+/G_+)/(P_-/G_-)$. P_+ is the number of repaired edges of cliques contained in the true-positive GS set. P_- is the number in the true-negative GS set. G_+ is the number of the true-positive set of GS and G_- is the number of the true-negative set of GS. Jansen et al. [9] have pointed out that the PPI predictions are acceptable with $L > 600$. G_+ and G_- are constant and equal to 8250 and 2705844, respectively. The estimation of value L in CORE and ALL is listed in Table 3. The L values of two predicted PPI sets are more than

Table 3: L value estimated by the GS sets.

	Predictions	P_+	P_-	L
CORE	352	88	23	1255
ALL	874	134	45	977

the acceptable threshold. The predicting quality of CORE is obviously better than ALL.

3.4. Comparison with Other Methods. Based on a network topology to predict protein interactions, clique methods are usually more reliable and accurate than clustering methods in the dense region of protein interaction network. Therefore, we select two methods based on cliques to compare with our method. The first is DC method of Yu et al. [15], which predicts interactions in protein interaction network by completing defective cliques. It is similar with our method in the first step to generate predicted PPI candidate set. DC method predicts PPIs only once; yet, our method generates them step by step with GO filtering. The second is the molecular complex detection algorithm (IPC-MCE) [16], which uses node extension approach based on the interaction probability with the known cliques to extend cliques in dense regions. IPC-MCE algorithm can also be viewed as a method of predicting interactions, if we consider all proteins in a predicted complex to interact with each other.

Using DC method, the parameters of nonoverlapping size and overlapping size of a clique are, respectively, sets 3 and 4 [15]. The lower size of non-overlapping parts and the higher size of overlaps can achieve good performance of predictions. Hence, in this paper, we set them 3 and 6, respectively. For IPC-MCE algorithm, we set the probability threshold t to 0.8. These parameters of two compared methods are set to be stricter than applied in the original papers for getting better results. DC method predicts 465 protein-protein interactions and 407 are hit in the BioGRID dataset. Furthermore, value L is estimated in GS datasets. There are 70 in true positives and 26 in true negatives based on gold standard datasets, and the value L is 883. IPC-MCE algorithm generates 496 predicted interactions and 436 are hit in BioGRID. And 79 are in true-positive dataset and 31 are in true-negatives based on gold standard datasets. The value L is 836. The results compared between various methods are shown in Figure 3.

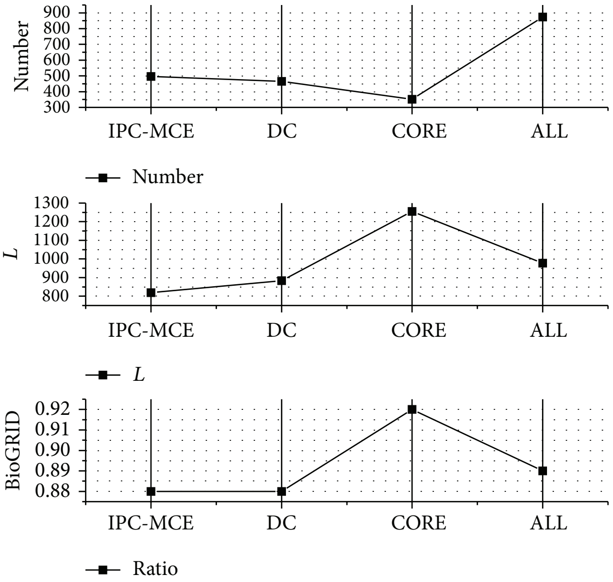


Figure 3: ALL has the largest number of predicted pairs. The estimation results with value L and BioGRID are better than the other compared methods. CORE has the largest accurate but small number. These predictions in CORE pursue high quality by sacrificing the predicted quantity.

3.5. Intersection of CORE and ALL. There are 321 interactions between CORE and ALL, as shown in section (2) of Figure 4. The remaining areas of CORE and ALL are represented with sections (1) and (3), respectively. The predictions in section (2) are validated to be better than PPIs in CORE and ALL. Therefore, the overlapping interactions between two sets have high reliability and are more suitable to complement the PPI datasets from biological experiments.

3.6. Assessment of Cliques Extended by Predicted PPIs. The predicted PPIs using our methods can well complement missing interactions to extend potential cliques (see Figure 5). We construct a component set of a clique mined from original DIP database within the Origin area. The Core area is extended by the predicted PPIs from CORE based on the Origin area. The All area is extended by the predicted PPIs from ALL based on the Origin area. We find that many extended cliques exhibit better quality than original cliques under GO annotations. In Figure 5(a), a 7-clique mined in the original network is extended to a 12-clique, then extended to an 18-clique. The number of U4/U6 × U5 tri-snRNP complex included in clique is increased from 6 to 11, then to 16. The components of original cliques are improved in the protein complexes. In Figure 5(b), a 9-clique mined in the original network is extended to a 14-clique, then extended to an 18-clique. The number of complex components included in the clique is increased from 9 to 14, then to 18. In Figure 5(c), a 7-clique mined in the original network is extended to an 11-clique, then extended to a 16-clique. The number participating in the mRNA processing included in the clique is increased from 6 to 10, then to 15. The components of original cliques increase in the same biological process. In Figure 5(d), a 7-clique mined in the original network is extended to a 9-clique, then extended to a 21-clique. The number participating in the ribosome biogenesis process included in the clique is increased from 7 to 9, then to 19. In Figure 5(e), a 7-clique mined in the original network is extended to a 16-clique, then extended to a 19-clique. The number participating in RNA binding included in clique is increased from 6 to 15, then to 16. The components of the original clique are extended in the molecular function.

4. Conclusions

Cliques in the PPI network are good resource and worth to study properly. Predicting PPIs by recursively extending

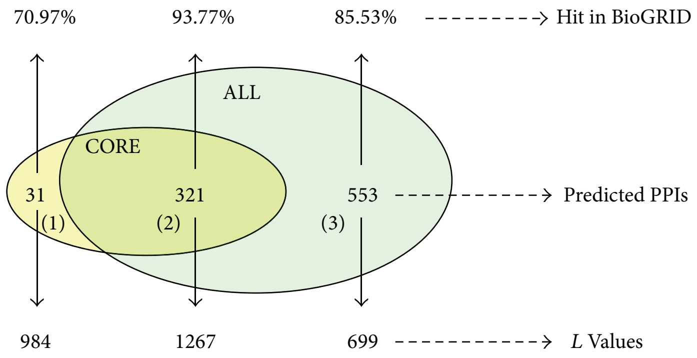
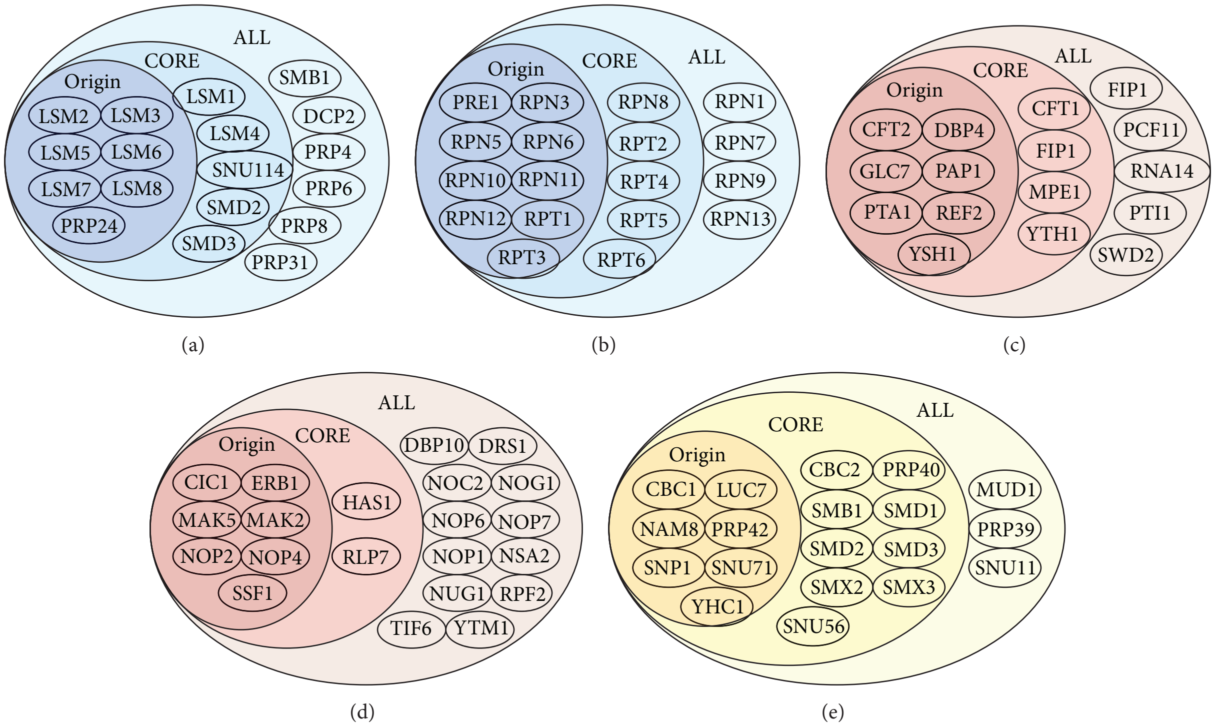


FIGURE 4: The overlap and difference between CORE and ALL are estimated with the dataset BioGRID and statistical likelihood benchmark *L*.



(a)

(b)

(c)

(d)

(e)

FIGURE 5: Cliques extended by the predicted PPIs from CORE and ALL have high functional significance and physical meaning.

cliques results in a quantity increase of predicted PPIs. And the quality of predictions is significantly improved via combining GO annotations with clique-based method. Obviously, iteratively finding cliques is a time-consuming task. Especially, the PPI network is frequently extended by new PPIs. Therefore, we design a special and novel method based only on the new predictions to mine cliques. This method is feasible and efficient in practice, which decreases the solution searching space and obtains high performance. Our work also partly solves the problem of missing cliques due to the false negative interactions. The clique size increases with continuously appending the predicted PPIs into the original network. Most enlarged cliques show good biological meaning with GO annotations. Besides, the predicted PPIs can complement data defection associated with protein complexes which are too large to detect all of its components with biological experiments. In the future, our methods on predicting PPIs and extending cliques will be applied to other PPI networks, such as human protein network, to help us associate disease relationship in the cliques.

Conflict of Interests

The authors declare that they do not have a conflict of interests for any of the organizations or authors that might have influenced the performance or presentation of the work described in this paper.

Acknowledgment

The authors thank the editors for valuable comments and suggestions. This paper is supported by the National Science

Foundation of China (NSFC) under Grants no. 61173087 and no. 61073128.

References

[1] J. Garcia-Garcia, J. Bonet, E. Guney, O. Fornes, J. Planas, and B. Oliva, "Networks of protein-protein interactions: from uncertainty to molecular details," *Molecular Informatics*, vol. 31, no. 5, pp. 342–362, 2012.

[2] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein-protein interactions. Part I. Experimental techniques and databases," *PLoS Computational Biology*, vol. 3, no. 3, pp. 0337–0344, 2007.

[3] N. Chepelev, L. Chepelev, M. D. Alamgir, and A. Golshani, "Large-scale protein-protein interaction detection approaches: past, present and future," *Biotechnology & Biotechnological Equipment*, vol. 22, no. 1, pp. 513–529, 2008.

[4] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners," *PLoS Computational Biology*, vol. 3, no. 4, pp. 595–601, 2007.

[5] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright, "Computational prediction of protein-protein interactions," *Molecular Biotechnology*, vol. 38, no. 1, pp. 1–17, 2008.

[6] K. A. Theofilatos, C. M. Dimitrakopoulos, A. K. Tsakalidis, S. D. Likothanassis, S. T. Papadimitriou, and S. P. Mavroudi, "Computational approaches for the prediction of protein-protein interactions: a survey," *Current Bioinformatics*, vol. 6, no. 4, pp. 398–414, 2011.

[7] J. A. Blake, M. Dolan, H. Drabkin et al. et al., "Gene ontology annotations and resources," *Nucleic Acids Research*, vol. 41, no. 1, pp. D530–D535, 2013.

[8] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.

[9] R. Jansen, H. Yu, D. Greenbaum et al., "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[10] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, no. 1, pp. D698–D704, 2011.

[11] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, 1979.

[12] G. Bianconi and M. Marsili, "Emergence of large cliques in random scale-free networks," *Europhysics Letters*, vol. 74, no. 4, pp. 740–746, 2006.

[13] M. Gendreau, P. Soriano, and L. Salvail, "Solving the maximum clique problem using a tabu search approach," *Annals of Operations Research*, vol. 41, no. 4, pp. 385–403, 1993.

[14] M. A. Mahdavi and Y.-H. Lin, "False positive reduction in protein-protein interaction predictions using gene ontology annotations," *BMC Bioinformatics*, vol. 8, article 262, 2007.

[15] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein, "Predicting interactions in protein networks by completing defective cliques," *Bioinformatics*, vol. 22, no. 7, pp. 823–829, 2006.

[16] M. Li, J.-X. Wang, B.-B. Liu, and J.-E. Chen, "An algorithm for identifying protein complexes based on maximal clique extension," *Journal of Central South University*, vol. 41, no. 2, pp. 560–565, 2010.