Predicting Protein-Protein Interactions Based on PPI Networks

Hui Li, Chunmei Liu, and Legand Burge

***Abstract*—Protein-protein interactions play a key role in the completion of cellular functions and usually correlate to each other in the form of a protein-protein interaction network. In this paper, we propose an algorithm to study the protein-protein interactions using the properties of protein-protein interaction networks. First, the algorithm constructs a protein-protein network according to the two query proteins. All the neighbors of the two query proteins obtained from the online protein databases are also included in the network. Second, the improved network partition algorithm was used to split the network into sub networks. Finally, a scoring function was proposed based on network clusters to predict the protein-protein interactions. The experimental results show that the scoring function based on the PPI network predicts the protein-protein interactions accurately.**

***Index Terms*—Protein-protein interaction; graph theory; score function.**

I. Introduction

In a biological system, most of the cellular functions are the direct result of the interactions of among proteins. Therefore, protein-protein interactions (PPIs) can help us understand the biological complexity within an organism. For this reason, great efforts have been devoted to the study of protein-protein interactions [1]. In particular, identification of protein interaction plays a key role in computational protein docking, structurally undetermined or protein function annotation.[1]

PPIs have been extensively studied for decades. Various machine learning based methods have been proposed and are applied successfully in this field such as SVM [2], HMM [3], artificial neural networks [4], Bayesian networks [5,6] and so on. Although machine learning has been successfully applied to this field, due to the inherent complexity of the biological system, the computational interpretation of PPIs is biologically questionable. Predictions that only consider individual information can not take advantage of the biology system for protein function. Graph theory based methods are superior in analyzing the protein interactions from the network properties. Current graph approaches that find the functional related clusters in the networks include spectral bisection [7, 8], Girvan and Newman (GN) [8], Kernighan-Lin (KL) [9] algorithm and hierarchical clustering. The GN algorithm is one of the community identification approaches and it has been broadly applied to functional modularity identification and pr[otein function predication. Newman has improved their algorithms afterwards. Among the existing algorithms, the Newman algorithm for community discovery is one of the remarkable approaches for decomposing networks into modules in complex networks, and it has been applied successfully to the biological networks [10-13].

In this paper, we propose a novel method to identify PPIs based on the protein-protein interaction network. The proposed method lies in two perspectives. One is protein-protein interaction network built dynamically from online databases, and the other is the protein interacting pairs extracted from the complicated network based on community discovery from the network. The data is queried and parsed dynamically to build the protein-protein interaction network for query protein pairs. A network partition algorithm is used for splitting protein-protein interaction network. First, we build the protein-protein interaction network through the relationship between the query proteins and their neighbors. In the dynamic data parsing process, we need to specify the number of layers. We dynamically parse a protein and all the neighbors of the current protein are put in one layer. We obtain the second layer by selecting one neighbor in order so that we can continue parsing. The layers will determine the composition of the size of the PPI network. When we select one neighbor to continue parsing, we obtain the second layer, so the layers will determine the composition of the size of the PPI network.

Then we split the protein interaction networks into several sub networks. Next, protein pairs are stemmed from different sub networks using our algorithm. Finally a scoring function based on the properties of the PPI network rather than the individual information of the protein pairs is established. During the validation we assume that protein functions and structures are known for all proteins (fully annotated) and the protein interaction networks are known for all but one protein. Experimental results show that the proposed method can improve the accuracy of the prediction of protein-protein interactions.

II. Data Sources

BioGRID [14] is a publication search based and freely accessible PPI database. It covers raw protein data and genetic interactions to Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster and Homo sapien. BioGRID dataset can be downloaded at http://www.thebiogrid.org. We use BioGRID dataset, and UniProtKB [16] as our data source.

[1] Manuscript received September 25, 2012; revised November 11, 2012. This work was supported by NSF CAREER (CCF-0845888) (Li and Liu), NSF Science & Technology Center grant CCF-0939370 (Li, Liu, and Burge).

The authors are with the Department of Systems and Computer Science, Howard University, Washington, DC 20059, USA.(e-mail: hli@scs.howard.edu; chunmei@scs.howard.edu; blegand@scs.howard.edu).

10.7763/IJMLC.2012.V2.239

III. Gold Standard

Gold standard is normally used for labeling the protein pairs and validating the prediction of PPIs from the database. It includes known PPI pairs as positive cases and non-interacting PPI pairs as negative cases. Selecting gold standard and features definitely affect the validity and reliability of the prediction of PPIs, so the validity of gold standard plays a key role in the quality of PPI predictions. However, there is no universal gold standard available in PPIs. The most challenging problem is gold standard negative sample selection. It is very hard to state the negative interactions because the non-interacting protein pairs are not easy to verify. Generally, random selection of protein pairs is the generic method for negative cases. Particularly, the gold standard of positive cases is the protein pairs known to be in the same complex. In this paper, we use PPIs from BioGRID as gold standard positive cases and protein pairs randomly selected from other databases as negative cases. Positive cases are selected from 1400 interactions (physical interactions) from BioGRID.

On the other hand, we build approximate gold standard negative samples using a random selection method from the Uni-ProtKB/Swiss-Prot database. In order to ensure the identified protein pairs are non-interacting, the protein pairs that have known interactions are removed from the selected sets.

One way of selecting negative samples that are non-interacting protein pairs suggested by researchers is to select protein pairs from different cellular localizations [11]. The protein pairs in different cellular localization may be interacting [7], so we use MIPS [15] to remove the interacted pairs out of the negative samples.

IV. Method

By dynamically parsing diverse online protein-protein interaction databases, a pair of unknown query proteins can form a protein-protein interaction network and the annotated information of each protein. The algorithm can be extended to gather information from partially annotated proteins. The dynamic parsing protein annotation allows more protein pairs and more neighbors to be included in the protein-protein network. Most of annotated proteins provide protein neighborhood information. We can also obtain protein sequences by dynamic parsing, therefore most useful annotated information is available in the protein-protein networks.

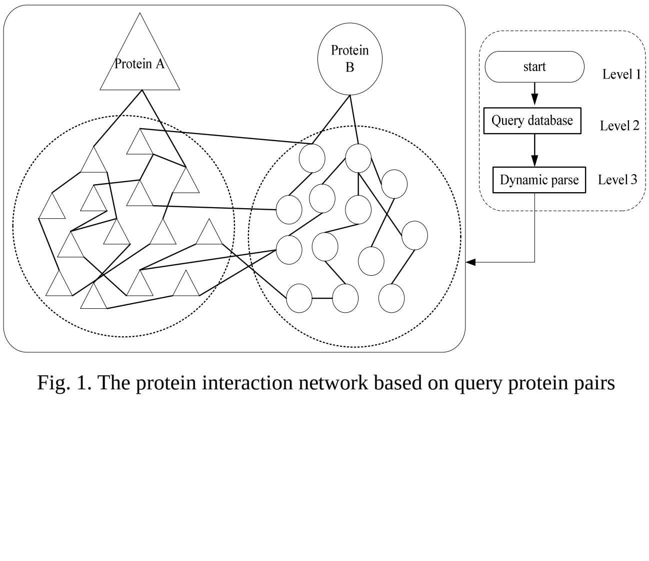

Fig. 1. The protein interaction network based on query protein pairs

After we obtain the protein-protein network, we can use the network partition method to partition the PPI network. In a PPI network, a protein is represented as a node. The edge symbolizes an interaction between corresponding two proteins. Given a PPI network $N = (V, E)$ and query protein pairs P_a and P_b included in the V, we use the partition algorithm [17] to partition N into several sub-networks which satisfy community conditions. The process of the network partitions are as follows:

1. Compute the degree of each node in the PPI-network
2. Remove the nodes whose degree is the maximum in the graph.
3. Ranking the vertex by descending order of the degrees in the graph
4. Select the busiest edge and remove it. The busiest edge is the edge that there are maximal numbers of the shortest paths which cross the edge.
5. Repeat 3 and 4 until protein Pa and protein Pb stem from different sub-networks.
6. Obtain a set of sub-networks *SubNetSet*, where *SubNetSet={Subnet1, Subnet2, Subneti, Subnetn}*

We use the Q value which is computed according to Equation 1 to evaluate the sub-networks divided by the algorithm [6]:

$$Q = \sum_{i} e_{ij} - \sum_{ijk} e_{ij} e_{ki} \qquad (1)$$

where e_{ij} is the edge in the original network that connect vertices in subnet i to those in subnet j.

The biochemical property reflects the bonds between the pair of proteins. Therefore it is helpful to predict the protein-protein interaction by using the biochemical property for the protein. The method for extracting biochemical feature from the protein sequence is as the following.

The amino acids are clustered into seven classes according to their dipoles and volumes of the side chains. Their classification is based on the hypothesis that amino acids within the same cluster have similar characteristics. Table 1 shows the seven classes clustered by the dipole[11].

TABLE I: Classification of Amino Acids by the Dipole

1	2	3	4	5	6	7
Ala, Gly, Val	Ile, Leu, Phe, Pro	Tyr, Met, Thr, Ser	His, Asn, Gln, Tpr	Arg Lys	Asp Glu	Cys

We apply three descriptors composition (C), transition (T) and distribution (D) [18]. C represents the frequency over a certain attribution among seven clusters in a protein sequence. T measures the frequency of one property over another property. D is the distribution of each property. We dynamically obtain the sequence information for each protein and calculate the property for each protein according to the method above. For instance, from the protein pairs P_x and P_y in a protein-protein interaction network:

$$p_x = \{a_1, a_2, a_i, \cdots a_n\}, where P_x \in subneta \qquad (2)$$

$$p_x = \{b_1, b_2, b_i, \cdots b_n\}, where P_x \in subnetb \qquad (3)$$

where a_i and b_i are the values of the feature vector in the sequence for protein P_x and protein P_y, respectively. Let r_{xy}

be the correlation coefficient between P and P_y which come from the subnet*a* and subnet*b*, respectively.

$$r_{xy} = \frac{\sum (\mathrm{X} - \overline{\mathrm{X}})(\mathrm{Y} - \overline{\mathrm{Y}})}{(\sqrt{\sum_{\mathrm{i=1}}^{n} (\mathrm{X} - \overline{\mathrm{X}})^2} \sqrt{\sum_{\mathrm{i=1}}^{n} (Y - \overline{\mathrm{Y}})^2})} \tag{4}$$

Let c^k be the weight value for *subnetk* which is calculated by Equation 1, and let *Sab* be the total value of the correlation coefficient for all neighbors around the query protein P_a and P_b.

$$S_{ab} = \sum_{i=0}^{n} r^{ij} c^{k} \tag{5}$$

Since our scoring function is calculated based on the protein-protein interaction network which is formed by query protein pairs, this scoring function thus includes information not only from the sequence of the query protein pair but also from the property of the protein-protein interaction networks. Therefore, our scoring function considers the protein-protein interaction network based on the annotated information of individual proteins. The higher the score is, the higher the chance that an interaction between the query protein pair is true.

We set a threshold for the scores in order to predict the protein-protein interaction; the scores above that threshold will be classified as interacting, while the scores below that threshold will be classified as non-interacting. We use the sensitivity which is the ratio of the true positives over the amount of true positives and the false negatives on the golden standard dataset to choose the threshold according to the performance.

V. Results

For the comparison between method [19] and our method on the gold standard set, we apply a Receiver Operating Characteristic (ROC) curve to assess our method, which is a useful technique for estimating the performance of a classifier. The ROC curves show that our score function based on network is effective.

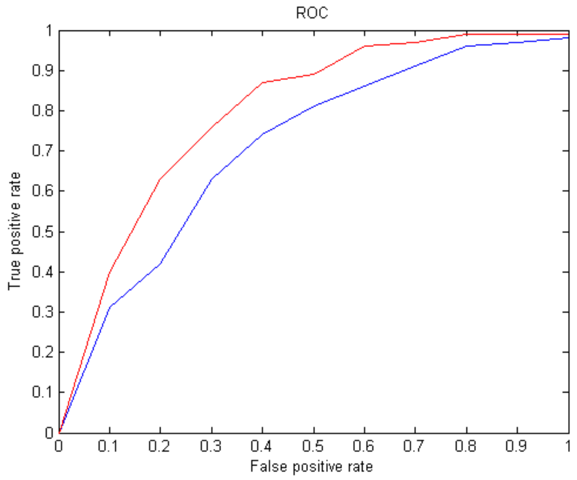

Fig. 2. The comparison for ROC curve of the network and other method, the red one is our method and blue one is method[18]

VI. Conclusion

We provide a novel method with the network score for the prediction of protein-protein interactions. Our method uses a dynamic parsing to form the protein-protein interaction networks online for the query protein pairs. We apply a network partition method to divide the PPI network into several sub-networks. In the corresponding sub-networks, we assign sequence property to each protein, and find all the neighbors around each query protein. Finally, we use the correlate coefficient to calculate the score based on the biochemical property for each protein and set the threshold for the prediction of the protein-protein interactions. The experiment shows that our method achieves a good performance in the prediction with protein-protein network properties.

References

[1] V. Spirin and L. A. Mirny, “Protein complexes and functional modules in molecular networks, ” in *Proc. Natl Acad Sci*, U S A, 2003,100: 12123–8.

[2] J. Y. Shi, S. W. Zhang, Q, Pan, Y. M. Cheng, and J. Xie. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids, vol. 33, no.1, pp. 69~74, 2007.

[3] P . Chen, C. M. Deane, and G. Reinert, A statistical approach using network structure in the prediction of protein characteristics. Bioinformatics , vol. 23, pp. 2314~2321, 2007.

[4] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D . Eisenberg. A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–6, 1999.

[5] R. Jansen, N. Lan, J. Qian, and M. Gerstei, “Integration of genomic datasets to predict protein complexes in yeast,” *J Struct Funct Genomics* ,vol. 2, pp. 71–81,2002.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan , et al, “A Bayesian networks approach for predicting protein-protein interactions from genomic data.,” *Science*, vol. 302, pp. 449–453, 2003.

[7] H. Lee, M., H. Deng, F. Z. Sun, and T. Chen, An integrated approach to the prediction of domain-domain interactions. BMC Bioinformatics 7: 269,2006.

[8] M. Girvan, and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821-7826, 11 June 2002.

[9] C. L. Myers, D. Robson, A. Wible, M. A. Hibbs, C. Chiriac, and et al. Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114,2005.

[10] B. Adamcsek, G. Palla, I. J. Farkas, I. Derenyi, and T. Vicsek Cfinder, “locating cliques and overlapping modules in biological networks,” *Bioinformatics*, vol. 22, pp. 1021–1023, 2006.

[11] A. L. Barabasi, and Z. N Oltvai, “Network biology: understanding the cell’s functional organization,” *Nat Rev Genet*, no. 5, pp. 101–113,2004.

[12] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, and et al. Probabilistic model of the human protein-protein interaction network. NatBiotechnol 23: 951–9,2005.

[13] M. Deng, S. Mehta, F. Sun, and T. ChenInferring, “domain-domain interactions from protein-protein interactions,” *Genome Res*, no. 12, pp. 1540–1548, 2002.

[14] B. J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bähler, V. Wood, K. Dolinski, and M. Tyers, “The BioGRID Interaction Database,” 2008 update. *Nucleic Acids Res*, no. 36(Database issue): pp. D637-640. 2008 Jan.

[15] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, et al, “MIPS: a database for genomes and protein sequences,” *Nucleic Acids Res* ,no. 30, pp. 31–34, 2002.

[16] UniProt Consortium, “Reorganizing the protein space at the Universal Protein Resource,” (UniProt). *Nucleic Acids Res.* 2012 Jan;40(Database issue):D71-75, Epub Nov 18, 2011

[17] P. F. Jonsson, T. Cavanna, D. Zicha, and P. A. Bates (2006) Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7: 2.

[18] H. H. Lin, Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity, Lipid Res, vol. 47, no. 4, pp. 824–831,2006.

[19] Pao-Yang M. Chen, Charlotte. Deane, and Gesine Reinert Predicting, “Validating Protein Interactions Using Network Structure,” *PLoS Comput Bi*ol, vol. 4, no. 7, 25 July 2008.

Hui Li is currently a Postdoctoral Fellow at the Department of Systems and Computer Science in Howard University. He received his PhD in Computer Science from Beijing School of Computer Science, University of Technology, Beijing in 2009. His research interests include computational biology, bioinformatics, pattern recognition and algorithm.

Chunmei Liu is currently an Associate Professor of the Department of Systems and Comp uter Science at Howard University. She got her Ph.D.. in Computer Science from the University of Georgia; Athens, GA in 2006. Her research interests include Graph Theory, Computational Biology, Algorithms, and Complexity.

Legand Burge is Professor and Chair of the Department of Systems and Computer Science at Howard University. Dr. Burge got his Ph.D. in Computer Science from Oklahoma State University in 1998.

He is interested in distributed computing, bioinformatics and machine learning.