2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Prediction of Protein Function from Protein-Protein Interaction Network by Weighted Graph Mining

Taewook Kim [1], Meijing Li [1], Keun Ho Ryu [1,2 +], Jungpil Shin[2]

[1]Database/Bioinformatics Laboratory, Chungbuk National University, Cheongju, Korea

[2]Multimedia Systems Lab, School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu, Fukushima, Japan

Abstract. Protein-protein interaction network plays a key role in protein function prediction. Many previous studies attempted to assign the protein function by neighbor or the connected path way to known protein. However, since interacting proteins do not always correspond, their accuracy is limited. In this paper, we proposed a novel approach to predict the protein function by weighted graph mining. Our proposed approach finds the functional frequent 2-node and 3-node patterns by labeling each protein node as a set of corresponding functions to predict the protein function without functional inconsistent between interacting protein. It also makes a selection from discovered frequent patterns by applying the weight to each edge so that we could compare which pattern is the most reliable for the function prediction. The function prediction is performed by matching the selected 2-node patterns, interacting with unannotated protein, with the frequent 3-node patterns. In the experiment, we used yeast protein interaction network which has functional frequent 3-node patterns and result shows 0.653 of function prediction accuracy. Compared with other methods' performance, our approach is better.

Keywords: protein interaction network, frequent pattern mining, weighted graph mining, protein function prediction.

1. Introduction

Prediction of protein function has become most challenging problem in various areas such as medical and biology. Since it is possible to prevent disease and develop new medicine by predicting protein function and researching interaction between proteins, it is considered to be more important. To assign the protein functions, many experiments are being done. However, because of many existed target proteins, it needs both much time and costs to determine the protein functions. To reduce such waste, many protein function prediction methods have been proposed. Among them, predicting protein function from protein interaction network is one of most challenging method because of complexity of functional relationship between proteins [1]. Recently, many methods which used graph mining and statistical approach in protein interaction network has been attempted for assigning protein functions.

The earlier studies using protein interaction network assigned the protein function based on neighbor proteins that interact with unknown protein or its pathway [2]. Chuan Lin proposed common neighbor method in proteins interaction network [3]. They supposed that if the function of its binding protein is known, function of an annotated protein can be deduced. It means if two proteins have common neighbors, they are likely to have same functions. Alexei Vazquez categorizes protein functions to 12 categories and assigns the proteins to functional classes on their interaction network [4]. They supposed that if proteins that interact with unknown protein have common function, unknown proteins are likely to have common functions. Ozsoyoglu and Kirac proposed a pairwise graph alignment algorithm to measure the similarity between the

+ Corresponding author. Tel.: + 82-43-261-2254; fax: +82-43-275-2254.
E-mail address: khryu@dblab.chungbuk.ac.kr, khryu@u-aizu.ac.jp.

set and an annotation neighborhood of an unknown protein [5]. But in this approach, large amount of noise could be created by splitting the input graph, rendering this approach ineffective. Young Rae Cho presented a frequent association pattern mining method [6]. They used selective joining and apriori pruning algorithm to find frequent subgraph. The function prediction is performed by matching the subgraph including the unannotated protein with the frequent pattern analogous to it. But in such these proposed approaches have presented limited accuracy because predicted functions could be quite different from real functions when the protein has several functions or interacting proteins have completely different functions. Indeed, Lei Shi compared three protein interaction data set such as MIPS, DIP, BIioGRid to confirm the function-relevant interactions [7]. And less than 37 percents of the protein interactions in these databases showed the function-relevant. It means only a few interacting protein has same function and it could make an improper prediction if we use only neighbor of unannotated protein for the function prediction.

For the purpose of solving these problems, in this paper, we explore the efficient discovery of frequent functional patterns in protein interaction network in order to predict protein functions. Interacting Proteins are likely to have same functions and corroborate with one another for common purpose [8]. Therefore, unannotated function can be predicted by its interaction partner. But as previously stated, not all the interacting proteins have same functions. For this reason, we assigned the set of protein functions to each corresponding node and find two neighbours that interact with an unknown protein to find patterns based on each functional label. Also we find 3-node frequent functional patterns in whole protein interaction network to assign the function to unannotated protein. To select the reliable pattern among the discovered patterns, we apply the weight in the network. We calculate weight of each pattern and select pattern which has the biggest weight.

The function prediction of an unknown protein is based on the two functional frequent patterns. So, the prediction is performed by matching these patterns. In other world, we annotate the protein functions by searching the frequent 3-node patterns including frequent 2-node patterns.

2. Methods

We showed our proposed function prediction method in Figure 1. Our function prediction method consists of four steps, generate the linkage, frequent pattern mining, applying the weight, function annotation. We first find two neighbor of given unannotated protein. And find the functional frequent 2-node and 3-node patterns using the set of function label which correspond to each node. And then, we calculate the weight of each frequent 2-node patterns to select the pattern for comparison with frequent 3-node patterns.

2.1. Problem definition

We represented a protein interaction network as an undirected and weighted graph G(V,E) [9]. $V(v_1 \ldots v_k)$is a set of nodes that denotes proteins and $E(e_1 \ldots e_k)$ is a set of edges that denotes interactions between proteins. Since one protein could occupy several functions, we assigned the set of functional categories to each node as a label $F(f_1 \ldots f_k)$.

After inserting unannotated protein into the graph, its interaction partners can be expressed in the graph as well. Our interest is to detect the frequent patterns using the set of functions which are labelled to each node. Thus, we searched the functional patterns that occur frequently using the neighbors that interact with unannotated protein. And calculate the weight of each pattern. Also we find frequent 3-node patterns in the whole network for the comparison. Our objective is to assign the function to all unanoatated proteins based on two and tree node functional frequent patterns.

Followings, we give three definitions for the function prediction method.

Definition1: A two-node functional pattern is a pattern that has two functional labels. Each functional label corresponds to a graph node in the protein interaction network. One of functional label must be linked to unknown protein directly. In other world, one of the two nodes must be neighbour of unannotated proteins.

Definition2: Frequent 2-node pattern is a pattern whose support count is bigger than minimum support. Support count is the number of appearing of the patterns in the network. We select five most frequent 2-node patterns for the prediction and calculate weight of each pattern.

Definition3: For the function prediction, we found the 3-node functional patterns which appear frequently in the whole network. We use apriori algorithm to find 3-node functional frequent patterns. To assign the protein function, 3-node functional pattern must include frequent 2-node pattern that is founded previous step. If there are two or more patterns that have same frequency, we sorted them arbitrarily.

Assigning the protein function for unannotated protein is based on discovery of frequent two-node pattern as well as three-node pattern in graph.

2.2. Pre-processing

Because the data is not suitable for the experiment, we had to go through pre-processing step. In this step, we removed protein nodes which do not interact with any other protein nodes and then we make linkages between interacting proteins. The linkage is to find the interacting protein. We found not only just two interacting protein also three interacting protein so that we could find the functional patterns. So, we have found 1249 nodes and 2985 interactions between them. Also, we make an annotation each node to fit the protein functions that have each protein. After annotating functional label, we categorized them into 16 function category for the experiment.

2.3. Searching the Unannnotated Protein's 2-Node Neighbour

In our approach, to find frequent 2-node patterns, one node must be connected to unannotated protein node directly. In other word, among the two neighbors that we want to know, one neighbour must be adjacent neighbor of unannoatated protein. So, our algorithm searches all adjacent neighbours of unannotated protein and then searches neighbors of previously discovered neighbors. As a result, our algorithm finds two neighbor proteins that are connected to unannotated protein. By giving the functional label to each founded neighbor, we could find the frequent functional patterns. Figure 2 shows the process of searching the neighbors as an example. In Figure 2, Suppose P is an unannotated protein, according to our algorithm, two nodes 1, 2 interacting with P are found. And then, it also finds neighbors of node 1 and node 2. So, we can get four different neighbors which are linked 2-nodes of P.

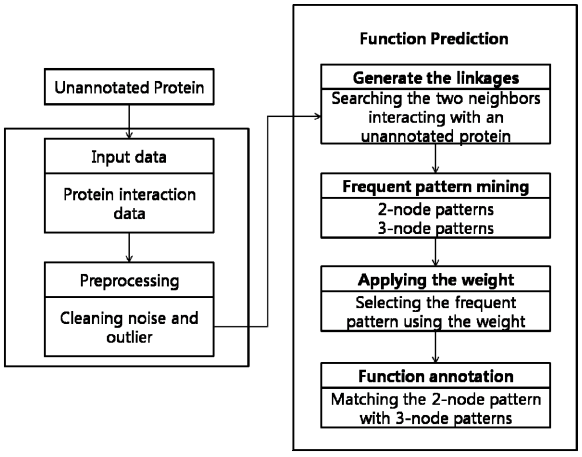

Fig. 1: The process of our proposed method

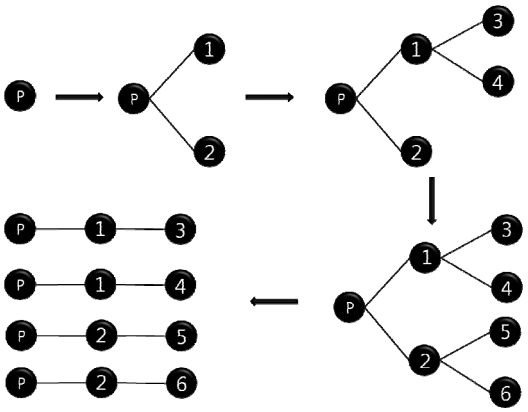

Fig. 2: An example of searching linked neighbours

2.4. Selecting the Frequent 2-node Pattern using Weighted Graph Mining

In this paper, we build a weighted graph model in protein interaction network. Chuntao Jiang showed structure base weighting approach in graph mining method [10]. The approach is based on frequency counts of individual nodes and edges. They adopt Pearson's Correlation Coefficient, PCC, using these frequency counts to measure the weight of edge. We do not select most frequent 2-node pattern for the comparison with 3-node patterns. To find most reliable patterns, we apply the weight to each frequent pattern and select the pattern whose weight is biggest. We showed the formula which they proposed. In this formula A and B mean occurrence number of nodes that belong to frequent patterns. So, we can calculate the weight of frequent patterns by number of occurrence and co-occurrence of two nodes.

$$w_{pcc} = \frac{\sup(A,B) - \sup(A)\sup(B)}{\sqrt{\sup(A)\sup(B)(1-\sup(A))(1-\sup(B))}}$$

2.5. Process of Function Annotation

To detect the functional patterns in a protein interaction network, our frequent functional pattern mining approach is based on apriori algorithm [11]. To predict exact function of protein, we found two and three-node patterns in protein interaction network. First, we searched 2-node frequent patterns in the whole network using apriori algorithm. To obtain the frequent patterns, we confirmed the proteins that interact with unknown protein directly. And then expand the node for searching the 2-node functional pattern. In other word, two-node pattern is a pattern consisting three nodes and one node must be unknown protein node.

We find two-node functional patterns using each discovered functional label in the whole protein interaction network and then count its frequency so that we can remove the patterns that appear not frequently. We select five most frequent functional patterns. And calculate the weight of each pattern to select the patterns which have the biggest weight among them. And then, we compare the 3-node frequent functional pattern with 2-node pattern to selected patterns which use for assigning the function of unannoatated protein.

3. Experiments and Results

3.1. Data set

We used protein-protein interaction data of Saccharomyces cerevisiae obtained by Database of Interacting Protein (DIP) [12] for our experiment. They provide a set of data that is experimentally determined protein interaction. Also it can be considered as a representative of more reliable protein-protein interaction data since all the interactions in data set have been examined carefully. There are some species of protein interactions and as mentioned earlier, we used Saccharomyces cerevisiae protein interaction data among them in experiment. It contains 1274 protein nodes and 3222 interactions between proteins. We also used Functional Catalogue (Funcat) for functional annotation [13]. It is provided by Munich Information Center for Protein Sequence (MIPS) which shows hierarchical classification system [14]. Each functional category has a description of protein function to make understanding easier.

3.2. Function categories

Because some Proteins have not just one function, one protein can be designated various function categories. 16 functional categories derived from MIPS. We counted the number of protein which belongs to each category. The largest number of protein belongs to Transcription which has 498 proteins annotated on it, also Transposable Elements, Viral and Plasmid Proteins has a least number of proteins.

3.3. Comparison of Function Prediction Accuracy with other methods

To evaluate our method, we used function prediction accuracy. We compared the predicted set of functions to actual set of functions to confirm the accuracy. During the comparison, we regard the prediction to be correct prediction if the predicted set of functions is part of the actual set of functions. We compare our method with the neighbor counting method using same data set. The accuracy of our method is 0.653 and accuracy of neighbor counting method is 0.440. This result shows that our proposed method has better performance than previous neighbor counting method.

4. Conclusion

In this paper, we proposed novel approach to assign the protein function from protein-protein interaction network. To avoid inconsistent function prediction, we find 2-node frequent functional patterns and 3-node frequent functional patterns for the comparison between them. Also, by applying the weight to each frequent 2-node patterns, we could select most reliable pattern for the prediction. So, unlike most of previous methods which based on neighborhood or connected pathway of unannotated protein for function prediction from a protein interaction network, our method could be independent from assumption that two interacting proteins are likely to have the same functions. For the experiment, we used Saccharomyces cerevisiae protein interaction data. We predicted 1249 protein nodes and 65 percent of proteins are predicted correctly among them. The weakness of our method is selecting the 3-node frequent pattern for the comparison with frequent

2-node pattern. We just select most frequent 3-node pattern for the function prediction. So, we could expect to increase function prediction accuracy by applying the weight or selecting the pattern among the frequent 3-node functional patterns.

5. Acknowledgements

This study was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2011-0001044) and Korean Ministry of Education, Science and Technology (The Regional Core Research Program/Chungbuk BIT Research-Oriented University Consortium)

6. Reference

[1] S. Jeon, T. Kim, H.S. Shon, M. Li, K.H. Ryu, Frequent Pattern Mining for Protein Function Prediction in Protein-Protein Interaction Network. *Proc.of international conference on convergence technology*. 2012, pp.35-37.

[2] Y. Cho, A. Zhang. Predicting protein function by frequent functional association pattern mining in protein interaction networks. *Information Technology in Biomedicine*, 2010, **14**: 30-36.

[3] C. Lin, D. Jiang, A. Zhang. Prediction of protein function using common-neighbors in protein-protein interaction networks. *Proc. Of IEEE Symposium on BionInformatics and BioEngineering (BIBE)*. 2006, pp. 251-260.

[4] Vazquez. A, Flammini. A, Maritan. A, Vespignani. A. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol*. 2003, **21**: 697-700.

[5] Kirac. M, Ozsoyoglu. G. Protein Function Prediction Based on Patterns in Biological Networks. In: Vingron, M., Wong, L. (eds.).*Proc. of RECOMB 2008. LNCS (LNBI).* 2008, pp. 197-213.

[6] Y. Cho, A. Zhang. Predicting Protein Function by Frequent Functional Association Pattern Mining in Protein Interaction Networks [J]. *IEEE Transactions on information technology in biomedicine*, 2010,**14**(1): 30-36.

[7] L. Shi, Y. Cho, A. Zhang. ANN based protein function predictionusing integrated protein-protein interaction data. *Proc. of Interactional Joint Conference on Bioinformatics*. Systems Biology and Intelligent Computing, 2009, pp. 271-277.

[8] P. Li, G.C. POK, S.J .Kwang, H.S. Shon, K.H. Ryu. QSE: A new 3-D solvent exposure measure for the analysis of protein structure. *Proteomics*, 2011,**11**(19): 3793-3801.

[9] P. Li, L. Heo, M. Li, K.H. Ryu. Frequent Pattern Based Protein Function Prediction Using Protein-Protein Interaction Networks. *Proc. Of Eighth International Conference on Fuzzy Systems and Knowledge Discovery*, 2011, pp. 1644-1688.

[10] C. Jiang, F. Coenen, R. Sanderson and M. Zito, Text classification using graph mining-based feature extraction. *Knowledge-Based Systems*, 2010, **23**: 302-308.

[11] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. *Proc. of the 20th International Conference on Very Large Databases*. 1994, pp. 487-499.

[12] 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*. 2004, **32**: D449-D451.

[13] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhanupt, M. Munsterkotter, H.W. Mewes. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*. 2004, **32**(18): 5539-5545.

[14] H. W. Mewes, S. Dietmann, D. Frishman, R. Gregory, G. Mannhaupt, K. F. X. Mayer, M. Münsterkötter, A. Ruepp, M. Spannagl, V. Stümpflen and T. Rattei. MIPS: Analysis and annotation of genome information in 2007. *Nucleic Acid Research*. 2008, **36**(1): D196-D201.