Protein Function Prediction Using Frequent Patterns In Protein-Protein Interaction Networks

Peipei Li[1], Lyong Heo[1], Meijing Li[1], Keun Ho Ryu[1]
[1]Database/Bioinformatics Lab
Chungbuk National University
Chungbuk, Korea
[1]{lipeipei, heoly, mjlee, khryu}@dblab.chungbuk.ac.kr

Gouchol Pok[1,2]
[2]Department of Computer Science
Yanbian University of Science and Technology
Yanji, China
[2]gcpokyust@gmail.com

***Abstract*— Protein function prediction is one of the most challenging problems in the post-genomic era. Previous prediction methods using protein-protein interaction networks relied on the neighborhoods or the connected paths to known proteins. Still new algorithm is required to increase the accuracy. In this paper, we propose a novel protein function prediction approach on the basis of frequent pattern mining in graph data. A protein-protein interaction network is represented as an unweighted, undirected graph with nodes denoting proteins and edges denoting interactions between proteins. Each node is labeled with a set of corresponding protein functions. The function prediction method is processed in three steps, neighbor finding, pattern finding and function annotation. Using our approach we predict protein functions on a core set of protein-protein interaction data from DIP (Database of Interacting Proteins) and function annotation data from FunCat of MIPS (the Munich Information Center for Protein Sequences). The experimental results show better performance in prediction accuracy than existing neighbor counting methods.**

Keywords-protein function prediction; protein-protein interaction networks; frequent pattern; graph mining

I. INTRODUCTION

Protein function prediction is a fundamental problem in computational and experimental biology. Dramatic growth of the number of predicted proteins cause a large fraction of these newly discovered proteins need to have a functional assignment. Still, the experimental determination of the function of a protein with known sequence and structure remains a difficult, time and cost intensive task.

To overcome the difficulties, many protein function prediction methods have been proposed from protein sequential [1], structural [2-3] and protein-protein interaction information [4]. Especially protein-protein interactions are of great interest because interaction proteins are likely to collaborate on a common purpose. Protein-protein interaction network has been proved as a feasible approach for unannotated protein by many studies.

Schwikowski proposed neighbor counting approach to predict the function of an unannotated protein based on the frequencies of the functions among its neighbors [5]. Hishigaki used a chi-square statistics [6] to calculate the significance of the functions of neighbors considering both directly and indirectly connected proteins. These two methods are based on the local graph and limit the annotations of an unknown protein by its interacting neighbors. Chuan Lin developed a common-neighbor-based prediction model and a Bayesian framework to predict protein function based on small-world property of protein-protein network [7]. They supposed that two proteins are likely to have same functions if they share common neighbors, and the more common neighbors they have the more likely they have same functions. Lei Shi used a distance matrix to filter the noise in the network, and designed an artificial neural network based method to predict protein functions [8]. In [9] and [10], the authors applied markov random field model to simulate the protein interaction network and develop prediction models for unknown proteins. Although previous studies using machine learning and statistical methods improve function prediction performances, the link-based approaches are still under critical challenges. In this respect, the annotation pattern-based approach has been recently developed. Experiment results show that pattern-based approach outperforms link-based approach like neighbor search method [12]. Kirac and Ozsoyoglu [11] proposed a pairwise graph alignment algorithm to measure the similarity between the set and an annotation neighbor of an unannotated protein. Young-Rae Cho [12] presented a frequent labeled subgraph mining approach using selective joining and apriori pruning algorithms. The function annotation is performed by matching the subgraph including the unannotated protein with the frequent patterns analogous to it. However, if the subgraph has too many nodes, it is difficult to search the frequent patterns.

In order to solve these problems, in this paper, we proposed a novel protein function prediction approach on the basis of frequent two-node functional pattern in protein-protein interaction networks. Interacting proteins in the network are likely to collaborate on a common purpose. Therefore the function of an unannotated protein can be predicted by its interacting proteins. If we can find the protein that interacts with the interacting proteins of the unannotated protein, the protein might in turn have a potential function as the unannotated protein. The proposed protein function prediction method is based on this theory. We focus on the protein functions in the global graph and not on the protein nodes or protein interacting neighbor protein nodes limited in the local graph in link-based approach. And because the proposed method use two-node pattern, it is easier to search the frequent

978-1-61284-181-6/11/$26.00 ©2011 IEEE

patterns than existing pattern-based method as mentioned above.

The reminder of the paper is further organized as follows. In Methods we present an unweighted undirected graph model for protein-protein interaction network, and then give two definitions. Then the prediction method is explained in three steps and an example is given. In Results and discussion section, data set for experiment is described and analyzed. The function prediction method is evaluated in term of partly prediction accuracy, and compared with neighbor counting method. Finally the paper is concluded in Conclusion section.

II. METHODS

A. Problem Definition

Protein-protein interaction network has been typically represented as a random graph by previous studies [13]. In this paper we represent it as an undirected unweighted graph G(V, E), which V(v_1,…v_m) is a set of nodes denoting proteins, and E(e_1,…e_n) is a set of edges denoting interactions between them. As known each protein falls into a set of functions, so each node in the network is labeled with a set of functional categories F(f_1,…f_k).

For an unannotated protein, after it is inserted into the graph, its interaction proteins can be represented in the graph. Our objective is to predict its functions set using the existing graph model.

Following we give three definitions for the function prediction method.

Definition 1: A two-node functional pattern is a pattern with two function itemsets which each function itemset corresponds to a graph node in the protein-protein network.

Definition 2: The support of a two-node functional pattern is the number of all the patterns with the same function sets in the graph.

Definition 3: A most frequent two-node functional pattern is a two-node functional pattern whose support is the largest among all the two-node functional patterns. When searching the most frequent two-node functional patterns, if the top two or more are with the same support, we sort them arbitrarily.

The unannotated protein function prediction is processed based on finding the most frequent two-node functional patterns in the graph.

B. Frequent Two-node Functional Pattern Based Function Prediction Method

The flowchart of the proposed function prediction method is shown in Figure 1. The function prediction method includes three steps, neighbor finding, pattern finding and function annotation.

Given an unannotated protein, at first, it searches all the interacting neighbor proteins around it. All function sets corresponding to the interacting proteins can be obtained.

Next, it finds all most frequent two-node functional patterns, which one node is fixed as the function of each interacting neighbor protein. The process of finding frequent two-node functional patterns is based on pattern growth method [14]. Pattern growth method is processed by adding a new edge to extend patterns from a single pattern directly. In this work, we first find all the nodes those have the same function with the interacting neighbor protein, and then extend it to a two-node pattern by interacting edges. The supports of all two-node functional patterns are calculated, and the most frequent two-node functional patterns are obtained.

Then each candidate function from all the functional subterms which are left by excluding the function subterms corresponding to the protein neighbors from the most frequent two-node functional patterns, is ordered in a list from the most frequent to the least frequent. The support is calculated by summation of supports of most frequent two-node functional patterns which the function is derived from. Finally the function of the unannotated protein is given by the most frequent one in the order list.

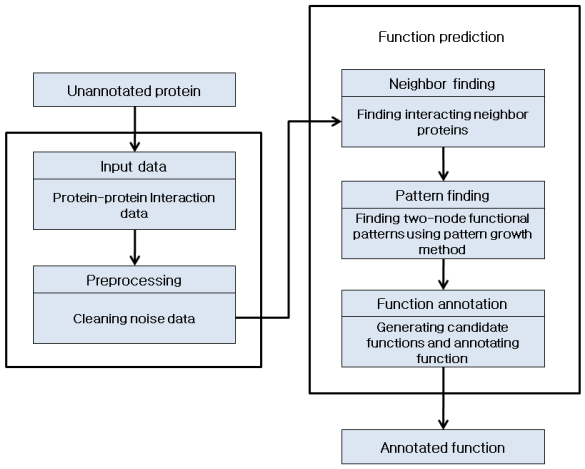

Figure 1. Our approach for protein function prediction

C. A Function Prediction Example

Figure 2 shows a five nodes graph as an example of the function prediction method. In Figure 1, five protein nodes {n_1, n_2, n_3, n_4, n_5} are labeled with function categories in the graph. Eight edges are connected between the five nodes. Node n_6 connected with n_1 and n_5 is the unannotated protein that needs to be annotated with a set of functions. The neighbors n_1 and n_5 are labeled with functions {f_1} and {f_1, f_3} respectively.

As described above, because nodes n_1 and n_5 are the neighbors of n_6, we first discover all the two-node functional patterns including the function of n_1 and n_5, which are {f_1} and {f_1, f_3} respectively, using pattern growth method. Table I lists all three two-node functional patterns including {f_1}, and the supports are 1, 3, and 2. Table II lists all three two-node functional patterns including {f_1, f_3}, and the supports are 2, 1, and 0. The most frequent two-node functional patterns corresponding to the two function categories are {f_1} - {f_1, f_2}

for $\{f_1\}$, which has a largest support of 3, and $\{f_1, f_3\}$ - $\{f_1\}$ for $\{f_1, f_3\}$, which has a largest support of 2 respectively.

Excluding the original function sets $\{f_1\}$ from $\{f_1\}$ - $\{f_1, f_2\}$ and $\{f_1, f_3\}$ from $\{f_1\}$ - $\{f_1, f_3\}$, each candidate function from the left $\{f_1, f_2\}$ and $\{f_1\}$ is ordered in Table III. The support of function f_1 is added by 3 and 2, which are the same supports with the most frequent two-node functional patterns where function f_1 is derived from. Finally function f_1 with the largest support 5, is predicted as the final annotated function to node n_6.

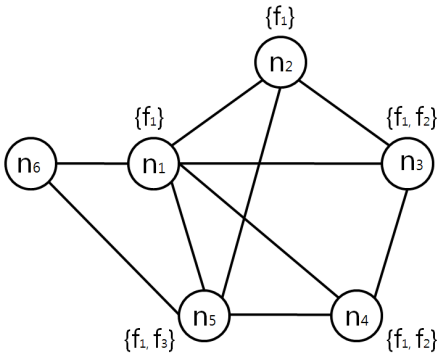

Figure 2. An example for function prediction method

TABLE I. TWO-NODE FUNCTIONAL PATTERNS INCLUDING $\{F_1\}$ AND CORRESPONDING SUPPORTS

Two-node Functional Pattern Including $\{f_1\}$	Support
$\{f_1\}$ - $\{f_1, f_2\}$	3
$\{f_1\}$ - $\{f_1, f_3\}$	2
$\{f_1\}$ - $\{f_1\}$	1

TABLE II. TWO-NODE FUNCTIONAL PATTERNS INCLUDING $\{F_1, F_3\}$ AND CORRESPONDING SUPPORTS

Two-node Functional Pattern Including $\{f_1, f_3\}$	Support
$\{f_1, f_3\}$ - $\{f_1\}$	2
$\{f_1, f_3\}$ - $\{f_1, f_2\}$	1
$\{f_1, f_3\}$ - $\{f_1, f_3\}$	0

TABLE III. CANDIDATE FUNCTIONS AND THEIR CORRESPONDING SUPPORTS

Candidate Function	Support
f_1	5
f_2	3

III. RESULTS AND DISCUSSION

A. Data Set

Database of Interacting Proteins (DIP) [15] provides a collection of experimentally determined protein interactions. It can be considered as a representative of more reliable protein-protein interaction data since all the interactions in this data set have been carefully examined. The core protein-protein interaction data of the baker's yeast Saccharomyces cerevisiae contains 1274 protein nodes and 3222 interactions between them. This data set is downloaded from the DIP website and used in the experiments. The functional annotation was performed using the Functional Catalogue (FunCat) [16], which is a well-established hierarchical classification system enabling the functional description of proteins from any organism. FunCat is one of the database resources of the Munich Information Center for Protein Sequences (MIPS) [17], which combines automatic processing of large amounts of sequences with manual annotation of selected model genomes.

B. Preprocessing

Preprocessing step should be applied to make the experimental data more suitable for using. As a preprocessing step, we delete all the protein nodes that do not have any interactions with others from the data set. Finally 1249 protein nodes and 2985 interactions between them are used in the experiment. After annotating each protein nodes, 16 function categories are used in the experiment.

C. Function Categories

As known, a protein can have several different functions, so one function category can be overlapped by different proteins. The function categories with the count of proteins annotated on them are enumerated in Table IV. From the list, function category Transcription is found to have a most number of proteins, which has 498 proteins annotated on it, and function category Transposable Elements, Viral and Plasmid Proteins has a least number of proteins, which has 2 protein annotated on it.

TABLE IV. FUNCTION CATEGORIES WITH THE COUNT OF ANNOTATED PROTEINS ON THE CATEGOPIES

Id	Function Category	Count
01	Metabolism	333
02	Energy	81
10	Cell Cycle and DNA Processing	364
11	Transcription	498
12	Protein Synthesis	192
14	Protein Fate	370
16	Protein with Binding Function or Cofactor Requirement	407
18	Regulation of Metabolism and Protein Function	70
20	Cellular Transport	241
32	Cell Rescue, Defense and Virulence	126
34	Interaction with the Environment	112
38	Transposable Elements, Viral and Plasmid Proteins	2
40	Cell Fate	83
41	Development	17
42	Biogenesis of Cellular Components	278
43	Cell Type Differentiation	132

D. Frequent Two-node Functional Pattern Detection

Table V lists the top 10 frequent two-node functional patterns with their corresponding supports found by the function prediction method. We use the function category id defined by FunCat as the function label in the experiment. It is obvious that two-node functional pattern {11, 16}-{11, 16} has a largest support of 165, which is much more than other two-node patterns.

TABLE V. TOP 10 FREQUENT TWO-NODE FUNCTIONAL PATTERNS WITH CORRESPONDING SUPPORTS

Two-node Functional Patterns	Support
{11,16}-{11,16}	165
{11,16} – {11}	69
{12,42} – {12,42}	66
{11} – {11}	63
{14} – {14}	43
{12} – {12}	34
{11} – {10,11}	24
{20} – {20}	22
{11,16} – {10,11,16,43}	19
{12,42} – {12,16,42}	18

E. Comparision of Function Prediction Accuracy with Neighbor Counting Method

In this paper, we use partly prediction accuracy used to evaluate our proposed function prediction method. The correct prediction is defined as the predicted function set of the unannotated function is included in the original function set. In other words, the predicted function set is a subset of the original function set.

In order to minimize the bias associated with the random sampling of the training and hold out data samples, we tend to use 10-fold cross-validation method. 10-fold cross-validation method is a special case of k-fold cross-validation method [18]. In 10-fold cross-validation method, it segments the data into 10 equal-sized partitions. During each run, one of the partitions is chosen as testing data, while the rest of them are used for training. This procedure is repeated 10 times so that each partition is used for testing exactly once. And the total accuracy is calculated as the average of all 10 runs.

For comparison, we implemented neighbor counting approach using the same data set as described above. It is processed as ordering the annotated functions of all neighbors of unannotated protein from the most frequent to the least frequent, and the most frequent one is declared as predictions for unannotated protein. The method is also evaluated by 10-fold cross-validation method.

The accuracy of each run and average of our function prediction method comparing with neighbor counting approach are listed in the following Table VI. From the table, for all of the ten test runs, the accuracies of the predicted method are all higher than neighbor counting method. And on average, the predicted method gets a partly accuracy of 0.600, and neighbor counting method gets a partly accuracy of 0.397. It is obvious that the proposed method has a much better performance than neighbor counting method.

TABLE VI. PARTLY ACCURACY FOR PREDICTED METHOD AND NEIGHBOR COUNTING APPROACH BY 10-FOLD CROSS-VALIDATION

Run	Partly Accuracy of Predicted Method	Partly Accuracy of Neighbor Counting Method
1	0.685	0.452
2	0.540	0.403
3	0.574	0.484
4	0.533	0.355
5	0.518	0.339
6	0.711	0.355
7	0.631	0.379
8	0.615	0.387
9	0.514	0.363
10	0.640	0.452
Average	0.600	0.397

IV. CONCLUSION

With the growth of newly discovered proteins, function prediction method is needed to be developed to give each of them a function annotation. In this paper, we propose a novel approach to predict the function of unannotated proteins based on frequent pattern mining in graph structure. We represent a protein-protein interaction network as an unweighted undirected graph with nodes denoting proteins and edges denoting interactions between proteins. Especially each node in the graph has a set of corresponding protein functions label. The function prediction method is processed in three steps, neighbor finding, pattern finding and function annotation. Pattern growth method is used to find protein frequent functional patterns including the neighbors' function. The unannotated protein function is predicted as the most frequent function calculated from the function sets which are obtained excluding the function term corresponding to each neighbors in frequent two-node functional pattern of each neighbors. The experiments are based on a core set of protein-protein interaction data from DIP and function annotation data from FunCat of MIPS. By 10-fold cross-validation method, the partly prediction accuracy of the proposed method has an average partly accuracy of 0.600, and has a much better performance than existing neighbor counting method of an average partly accuracy of 0.397. The proposed method has improved the prediction accuracy, and can be useful in protein function prediction area. In the future work we will continue to focus on pattern-based protein prediction method in protein-protein interaction networks and improve the method to increase prediction accuracy.

ACKNOWLEDGMENT

This work was supported by a grant from the Korean Ministry of Education, Science and Technology (The Regional Core Research Program / Chungbuk BIT Research-Oriented University Consortium), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2011-0001044), and Korea Biobank project (4851-307) of the Korea Centers for Disease Control and Prevention.

REFERENCES

[1] M. Wang, X. Shang and Z. Li, "Sequential Pattern Mining for Protein Function Prediction", Lecture Notes in Computer Science, , Vol. 5139, pp. 652-658, 2008

[2] O. Mason, M. Verwoerd and P. Clifford, "Inference of Protein Function from the Structure of Interaction Networks", Structural Analysis of Complex Networks, pp. 439-461, 2011

[3] K.M. Borgwardt, C. Ong, S. Schönauer, S.V.N. Vishwanathan, A.J. Smola and H. Kriegel, "Protein function prediction via graph kernels", Bioinformatics, Vol. 21, pp. i47–i56, March 2005

[4] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks", Nature biotechnology, Vol. 21, No. 6, pp. 679-700, 2003

[5] B. Schwikowski, P. Uetz and S. Field, "A network of protein protein interactions in yeast", Nature Biotechnology, Vol. 18, pp. 1257-1261, 2000

[6] H. Hishigaki, K. Nakai, T.Ono, A. Tanigami and T. Takagi, "Assessment of prediction accuracy of protein function from protein-protein interaction data", Yeast, Vol. 18, pp. 523-531, 2001

[7] C. Lin, D. Jiang and A. Zhang, "Prediction of protein function using common-neighbors in protein-protein interaction networks", Six IEEE Symposium on BioInformatics and BioEngineering, pp. 251-260, 2006

[8] L. Shi, Y. Cho, and A. Zhang, "ANN based protein function prediction using integrated protein-protein interaction data", 2009 Interactional Joint Conference on Bioinformatics, Systems Biology, and Intelligent Computing, pp. 271-277, 2009

[9] M. Deng, K. Zhang, S. Mehta, T. Chen and F. Sun, "Prediction of protein function using protein-protein interaction data", IEEE Computer Society Bioinformatics Conference, pp. 197-206, 2002

[10] S. Letovsky and S. Kaisif, "Predicting protein function from protein/protein interaction data: a probabilistic approach", Bioinformatics, Vol.19, pp. i197-i209, 2003

[11] M. Kirac and G. Ozsoyoglu, "Protein function prediction based on patterns in biological networks", 12th Interaction Conference on Research in Computational Molecular Biology, pp. 197-213, 2008

[12] Y. Cho and A. Zhang, "Predicting protein funtion by frequent functional association pattern mining in protein interaction networks", Vol. 14, No. 1, pp. 30-36, January 2010

[13] P. Lee, C. Huang, J. Fang, J.J.P. Tsai and K. Ng, "Study of the protein-protein interaction networks via random graph approach", Fourth IEEE International Conference on Cognitive Informatics, pp.110-119, July 2005

[14] X. Yan and J. Han, "Discovery of frequent substructures", Mining Graph Data, pp. 99-115, 2006

[15] 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

[16] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhanupt, M. Munsterkotter and H.W. Mewes, "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes", Nucleic Acids Research, Vol. 32, No. 18, pp. 5539-5545, 2004

[17] H.W. Mewes, S. Dietmann, D. Frishman, R.Gregory, G.Mannhaupt, K.F.X. Mayer, M. Munsterkotter, A. Ruepp, M. Spannagl, V. Stumpflen and T. Rattei, "MIPS: analysis and annotation of genome information in 2007", Nucleic Acids Research, Vol. 36, pp. D196-D201, 2008

[18] P.N. Tan, M. Steinbach, and V. Kumar, "Introduction to date mining", pp: 186-188, 2006