Function prediction of proteins in yeast networks based on the MCL algorithm

Ke Zhan[a,b], YunQuan Zhang[a,b,c],
[a] University of Chinese Academy of Sciences ,Beijing 100190, China
[b] Laboratory of Parallel Software and Computational Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China.
[c] State Key Lab. of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.
Email: zhankecas@gmail.com, yunquan.cas@gmail.com

Abstract—**Large-scale Protein-Protein interaction data sets exist in Saccharomyces cerevisiae due to many interaction detection methods such as yeast two-hybrid assay, mass spectrometry of purified complexes, correlated mRNA expression profile and so on. How to make use of these data sets to understand the protein function is very important. We use the algorithm [17] developed by Stijn van Dongen to describe the functional modules in PPI networks.We analyze four protein-protein networks from Saccharomyces cerevisiae, and our results suggest that the functional modules detected are consistent with the biology knowledge. Protein-Protein interaction network was separated into clusters using MCL algorithm. Based on the clusters resulted from MCL algorithm, we assign the function annotations using P-value and majority methods. The majority method is based on the majority rule [15]. The predicted function of proteins provide clue to biology experiments. Two methods are used to assign function annotations for the known clusters and unknown proteins, we compare the two predicted results, the results show that the two methods are consistent with each other.**

Index Terms—**function prediction , protein-protein interaction, MCL algorithm.**

I. Introduction

IN/the post-genomic era one of the most important tasks is how to mine biology information from those protein-protein interaction networks. Various biological experiments were used to study the function of unknown protein. Small scale experiments [7] were designed to study the individual gene function. Large scale methods include yeast two-hybrid assay [11], [20], mass spectrometry of purified complexes [6], [9], correlated mRNA expression profile [5], [10]. Many interaction data sets of kinds of model organism accumulate based on these experiments.

In a protein-protein interaction network, one node represent one protein, if two proteins interact with each other, there is one edge between them. One protein-protein interaction network is abstracted as one graph in view of the above mentioned condition. Several progress reports have been published by the graph-theoretical methods. For example, the scale-free topological structure of the protein-protein network from S.cerevisiae has been described in [3], [12] . Intuitively, proteins which have similar function interact more likely with each other. So the other method such as clustering of protein-protein interaction network is also significant to analyze the network. The popular clustering algorithms include: Clique Finder [1], Density-periphery based clustering [2], Network Blast [16] and so on. These algorithms are used to cluster the proteinprotein interaction network to uncover the topological structure and predict the function of unknown proteins. Different clustering algorithms are used in different networks.

In this paper, we focus on the MCL algorithm. The subnetwork which are isolated from the largest connected subgraph are removed before the program run. Based on the clusters resulted from the program, we assign function annotations to the unknown proteins and known clusters. The function annotations of proteins were downloaded from MIPS [13]. The data version is 2.1(09.01.2007). As a single node, if it is assigned 99 from MIPS, we call this node an unknown node. The cluster which includes unknown nodes is called an unknown cluster. If a single node is assigned functional annotation which is different from 99, we call the node a known node. The nodes included by the cluster are all known nodes, in this case, we call the cluster as a known cluster. From the compared results based on the two predicted methods(P- value and majority methods), we find that the two methods are consistent with each other.

The rest of the paper is organized as follows. In section II, we list the source of the data sets, give a brief description of the platform and the methods used during the running program. Section III reveals the results. Conclusions are presented in section IV.

II. MATERIALS AND METHODS

A. Interaction Data Sets

We use four data sets which come from Saccharomyces cerevisiae protein-protein interaction networks. The basic information of the four data sets are summarized in Table 1:

This work was supported in part by National Natural Science Foundation of China (No.61133005, No. 61272136), The National High-tech R&D Program of China(No. 2012AA010903).

doi:10.4304/jsw.9.5.1157-1162

In 2002, von Mering,C. et al. count up the numbers of proteins and interactions. There were 5400 yeast proteins and 80000 interactions among them. Proteins are represented by the nodes, interactions between proteins are represented by the edges. Low level of reliability edges and related nodes are removed. Then 2617-11855 data set came into being. The first data set includes 11855 edges among 2617 nodes [4]. The data set is denominated as the numbers of nodes and edges related to the network. The second data set includes 711 nodes and 704 edges. The third network include 2455 high-confidence protein-protein interactions during 988 proteins. The fourth data set involve 2238 interactions during 1827 proteins. The interactions are based on direct interactions identified by biochemical experiments and two-hybrid experiments, but not protein complexes. We derive the function annotations of proteins from Munich Information Center for Protein Sequences (MIPS) [13], [19].

B. Algorithm Used For Clustering Yeast Networks

In this paper, the MCL algorithm is used for the yeast networks. The MCL algorithm is short for the Markov Cluster Algorithm. The algorithm use two simple algebraic operations on matrices to simulate flow. Expansion is the first operation, this step models the spreading out of flow. Expansion squares the matrix which is converted from the adjacency matrix of a network. Inflation is the second operation, this step models the contraction of flow. Inflation is a Hadamard power followed by a diagonal scaling. Expansion and inflation are repeated until there is no change on the matrix. We download the MCL program from http://micans.org/mcl/. Install the software under Fedora 15 operating system. The kernel version is 2.6.42.3 . The inflation factor -i which affect the cluster granularity was chosen as 1.4. Different inflation factor represents different levels of granularity, different clustered results will be derived. Finally we get the clusters resulted from MCL algorithm.

C. Computation Of P-value

Assuming that unknown proteins should have the same function of the clusters. P-value [8], [18], [23] is used to assign every cluster a main function. Then the unknown proteins are assigned function of the cluster which include the node. The equation used for computing P-value is:

TABLE I
BASIC INFORMATION OF THE FOUR NETWORKS

Name	NN	NE	NNLCS	References
2617-11855	2617	11855	2375	von Mering C et al.
711-704	711	704	168	Shiwei Sun et al.
988-2455	988	2455	573	von Mering C et al.
1827-2238	1827	2238	1299	Vazquez A et al.

NN: Numbers of Nodes, NE: Numbers of Edges, NNLCS: Numbers of Nodes from Largest Connected Subgraph

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

Hypergeometric distribution is used for each function categorization to model the probability of at least k proteins from a cluster of size n by chance in a category containing C proteins from a total network size of G proteins. The equation test whether a cluster is enriched with proteins which are from a particular category or which are from a random category. The smaller the P-value, the probability that the function category come from a particular category is higher.

We use C++ language to implement the P-value algorithm. Because the computation of P-value is one high precision operation, we use NTL(A Library for Doing Number Theory) which is download from http://www.shoup.net/ntl/download.html to improve accuracy. NTL provides arbitrary length integer arithmetic and arbitrary precision floating point arithmetic. We choose the unix version of the software.

Every cluster contains different function categories, we choose the function category of the lowest P-value as the main function of the cluster. If multiple function categories have the same lowest P-value, those function categories are assigned to the cluster.

We assign the main function of the cluster to the unknown node. If a node belongs to multiple clusters, we choose the function of the cluster whose P-value is minimum . Then this function will be assigned to the node.

D. Computation Of Majority

Roded Sharan et al. studied conserved subnetworks in multiple species [16], and they predicted protein functions when the proteins in a cluster were significantly enriched and at least half of the proteins had the same annotations. The prediction was based on conserved subnetworks in multiple organism. Function annotations which came from directed partners of known proteins were used to predict the function of unknown proteins [15], [22]. In the set of 554 unknown proteins, the number of proteins which have at least one partner of known function is 364. The numbers mean that there are 190 proteins which can not be assigned function making use of the information of directed partners. And two or more partners of 69 proteins have known function. Two or more partners of only 29 proteins have common function. These figures mean that few fraction of unknown proteins may be predicted exactly. This method is called majority rule.

To better take advantage of the global information, first a protein-protein network is clustered into different clusters, then the most common function annotations in one cluster are assigned to the unknown nodes which are belong to the cluster.

Based on this idea, we choose the most common function of known proteins as the function of unknown cluster. Analogously, we predict the function of known cluster. We call this method majority method. This method will assign function annotation to a unknown protein whenever the unknown node belongs to a cluster. Even if the partners of the unknown protein are unknown proteins.

We use C++ language to implement the majority algorithm.

III. Results

A. Clusters In Networks

In this paper, we choose MCL algorithm to cluster the protein-protein networks. There are several parameters of the MCL algorithm to control the results. The inflation value is the main parameter. We choose the inflation value 1.4 so that the number of predicted unknown proteins are much more. The clusters of the four yeast networks are listed in the supplementary data(Clustering results of the MCL algorithm.I14.pdf). For example, the title out.168.txt.I14 is the cluster result of the network 711-704. The rest files are other results respectively.

B. Function Annotation Of Known Clusters And Unknown Clusters

For every network, first we extract the largest connected subgraph from the original network. Next we make use of the MCL algorithm to cluster the subgraph. Based on the results of the algorithm, we assign function annotation to the known clusters and unknown proteins. During the process of function annotation, we use two different methods which include P-value and majority methods.

For the function prediction of a known cluster with P-value method, the nodes in the cluster are known for they have a single function annotation or more than one function annotation. We compute the P-value for every function annotation. The function annotation which have the lowest P-value was assigned to the cluster as the function of the cluster. If multiple function annotations have the same lowest P-value, the set of function annotations were assigned to the cluster. For the function prediction of a known cluster with majority method, we count the number of every function annotation which is belonged to the nodes of the cluster. The function annotation which has the max number is assigned to the cluster. If multiple function annotations have the same max number, the set of the multiple function annotation are assigned to the cluster.

For a unknown cluster which includes known nodes and unknown nodes whose function annotation are ”99”, we ignore the unknown nodes, compute the P-value of known function annotation, assign the function annotation which have the lowest P-value to the cluster as above. Then the majority method is used for the function prediction of the unknown clusters. We count the number of every function annotation which is belonged to the known nodes of the cluster. The most common function annotation that the function annotation has the max number is assigned to the cluster. If multiple function annotations have the same max number, the set of multiple function annotations are assigned to the unknown cluster.

C. Function Prediction Of Unknown Proteins

Based on the function prediction of the unknown clusters, we predict the function of the unknown proteins. If the unknown protein belongs to a single cluster, then the function of the cluster are assigned to the unknown protein. If the unknown protein belongs to more than one cluster, then the function of the clusters which have the lowest P-value or the maximal majority value are assigned to the unknown protein. Every unknown protein is assigned function due to the above methods.

We compare the function prediction results of unknown proteins which are belonged to 2617-11855 network [4]. Table 7 − 10(supplement data) summarize the predicted function annotations of 194 unknown proteins. For the same 2617-11855 network, based on the clustering results, the number of predicted proteins by two methods is 194, 76 unknown proteins are predicted in the study of [4]. Only 10 proteins are the same proteins between two unknown sets. Our new prediction of unknown proteins provide more clues to biology experiment. The rest unknown proteins of the other three networks are predicted using the same methods. The resulted prediction are listed in table 1(network of 711-704), table 2(network of 988-2455), table 3 − 6(network of 1827-2238)(supplement data). Function prediction of unknown proteins.pdf summary the results.

D. Validation of the prediction

1) High consistency of the two predicted methods: From the prediction results of known clusters, we found that function annotations of at least 85 percentages of known clusters under two different predication methods are the same or similar. For example of the 2617-11855 network: for a cluster, the function predicted by P-value method are ribosome biogenesis(12.01) and ribosomal proteins(12.01.01). The function predicted by the majority method are also ribosome biogenesis(12.01) and ribosomal proteins(12.01.01). We call the prediction results are the same results by two methods. For the unknown protein YPL077c, the function predicted by P-value method are DNA synthesis and replication(10.01.03), and DNA binding(16.03.01). The function predicted by majority method are :DNA processing(10.01), DNA synthesis and replication(10.01.03), nucleic acid binding(16.03) and DNA binding(16.03.01). These predicted results are called similar results by two methods. The third situation, the unknown protein YPL159c is assigned lipid/fatty acid transport(20.01.13) by P-value method and nucleotide/nucleoside/nucleobase metabolism(01.03) by majority method. These results are called different results by two method. For the known clusters of the four networks, the overlapped ratio by two methods are 90%, 95%, 85%, 86% respectively. For the unknown clusters of the four networks, the overlapped ratio by two methods are 100%, 100%, 77%, 79% respectively. The results demonstrate that the methods are validity indeed. The predicted function annotations can be used as the reference information for the biological experiments. The results are summarized in table 2.

TABLE II
RESULTS OF FUNCTION ANNOTATION

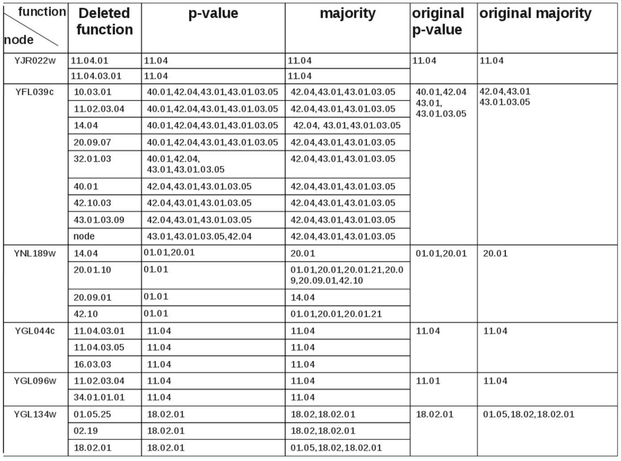
Name	NNLCS	NKC	NUC	NUP	Max	Min	PUP	PKC
711-704	168	10	4	7	39	3	100	90
988-2455	573	42	8	11	102	2	100	95
1827-2238	1299	46	65	142	70	2	77	85
2617-11855	2375	66	64	194	168	2	79	86

NNLCS: Numbers of Nodes from Largest Connected Subgraph, NKC: Numbers of Known Clusters, NUC: Numbers of Unknown Clusters, NUP : Numbers of Unknown Proteins, Max: Max Numbers of Nodes in Cluster, Min: Min Numbers of Nodes in Cluster, PUP: Similarity Percentage of Predication Unknown Proteins, PKC: Similarity Percentage of Predication Known Clusters

2) Function prediction after blank out the function of the nodes: We blank out the function of known nodes one by one. Every network have many known nodes, we choose three nodes whose degree are max and three nodes whose degree are minimum as the representative nodes for every network. The fist column are the nodes whose function are blanked out one by one. For example, in figure 1, The first node is YJR022w whose degree is 20, the degree of YFL039c is 18, the degree of YNL189w is 13. The degree of the last three nodes(YGL044c, YGL096w, YGL134w) are all 1.

The second column is ”deleted function” which represent the function deleted. For example, the node YJR022w has two functions:11.04.01 and 11.04.03.01. The first step, we delete the function of 11.04.01, predict the function of the cluster which YJR022w belongs to. The second step, we continue to delete the function of 11.04.03.01, predict the function of the same cluster. We do the same operation on the rest nodes. For the node of YFL039c, the last function is represent as ”node”, because there are many similar function, if we delete one by one, there will no effect to the final predicted results, the similar function we don’t delete. ”node” represents we delete the all the functions of the node YFL039c which means that we delete the node YFL039c from the cluster at the last step.

Fig. 1. Function prediction of 711-704 network



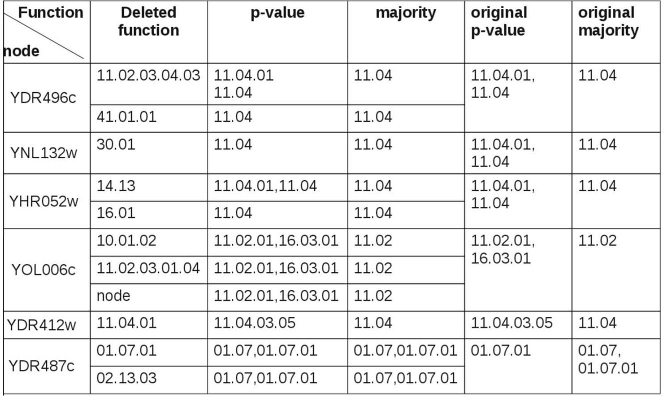
function / node	Deleted function	p-value	majority	original p-value	original majority
YJR022w	11.04.01	11.04	11.04	11.04	11.04
	11.04.03.01	11.04	11.04		
YFL039c	10.03.01	40.01,42.04,43.01,43.01.03.05	42.04,43.01,43.01.03.05	40.01,42.04 43.01, 43.01.03.05	42.04,43.01 43.01.03.05
	11.02.03.04	40.01,42.04,43.01,43.01.03.05	42.04,43.01,43.01.03.05		
	14.04	40.01,42.04,43.01,43.01.03.05	42.04, 43.01,43.01.03.05		
	20.09.07	40.01,42.04,43.01,43.01.03.05	42.04,43.01,43.01.03.05		
	32.01.03	40.01,42.04, 43.01,43.01.03.05	42.04,43.01,43.01.03.05		
	40.01	42.04,43.01,43.01.03.05	42.04,43.01,43.01.03.05		
	42.10.03	42.04,43.01,43.01.03.05	42.04,43.01,43.01.03.05		
	43.01.03.09	42.04,43.01,43.01.03.05	42.04,43.01,43.01.03.05		
	node	43.01,43.01.03.05,42.04	42.04,43.01,43.01.03.05		
YNL189w	14.04	01.01,20.01	20.01	01.01,20.01	20.01
	20.01.10	01.01	01.01,20.01,20.01.21,20.0 9,20.09.01,42.10		
	20.09.01	01.01	14.04		
	42.10	01.01	01.01,20.01,20.01.21		
YGL044c	11.04.03.01	11.04	11.04	11.04	11.04
	11.04.03.05	11.04	11.04		
	16.03.03	11.04	11.04		
YGL096w	11.02.03.04	11.04	11.04	11.01	11.04
	34.01.01.01	11.04	11.04		
YGL134w	01.05.25	18.02.01	18.02,18.02.01	18.02.01	01.05,18.02,18.02.01
	02.19	18.02.01	18.02,18.02.01		
	18.02.01	18.02.01	01.05,18.02,18.02.01		

The third column is ”p-value” which represents the predicted function of the unknown cluster using P-value method after blank out the function of one node. The fourth column ”majority” has the similar meaning. The difference is that we use majority method as the predicted method. The fifth column ”original p-value” represents the predicted function of the same unknown cluster , in which the function of the node is not deleted. The sixth column ”original majority” represents the similar meaning as the fifth column. The difference is that function predicted method is majority method.

For example of the node YFL039c, we first delete the function of 10.03.01, the function of unknown cluster is predicted as 40.01, 42.04, 43.01, 43.01.03.05 using P-value method and 42.04, 43.01, 43.01.03.05 using majority method. Before blank out the function, the predicted function of the same cluster is 40.01, 42.04, 43.01, 43.01,03.05 using P-value method and 42.04, 43.01, 43.01.03.05.

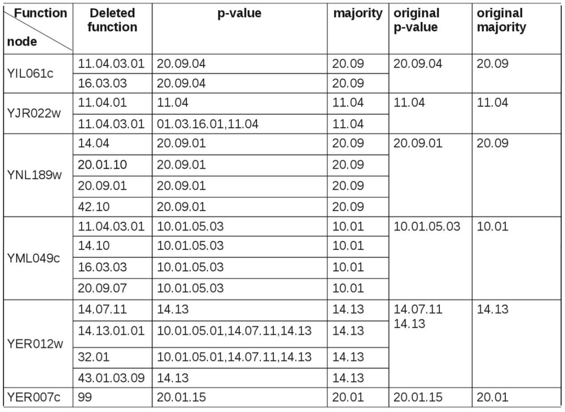
We use the same steps to the other three networks. The result are illustrated by fig 2 ,fig 3 and fig 4 respectively.

Fig. 2. Function prediction of 988-2455 network



Function / node	Deleted function	p-value	majority	original p-value	original majority
YDR496c	11.02.03.04.03	11.04.01 11.04	11.04	11.04.01, 11.04	11.04
	41.01.01	11.04	11.04		
YNL132w	30.01	11.04	11.04	11.04.01, 11.04	11.04
YHR052w	14.13	11.04.01,11.04	11.04	11.04.01, 11.04	11.04
	16.01	11.04	11.04		
YOL006c	10.01.02	11.02.01,16.03.01	11.02	11.02.01, 16.03.01	11.02
	11.02.03.01.04	11.02.01,16.03.01	11.02		
	node	11.02.01,16.03.01	11.02		
YDR412w	11.04.01	11.04.03.05	11.04	11.04.03.05	11.04
YDR487c	01.07.01	01.07,01.07.01	01.07,01.07.01	01.07.01	01.07, 01.07.01
	02.13.03	01.07,01.07.01	01.07,01.07.01		

Fig. 3. Function prediction of 1827-2238 network



Function / node	Deleted function	p-value	majority	original p-value	original majority
YIL061c	11.04.03.01	20.09.04	20.09	20.09.04	20.09
	16.03.03	20.09.04	20.09		
YJR022w	11.04.01	11.04	11.04	11.04	11.04
	11.04.03.01	01.03.16.01,11.04	11.04		
YNL189w	14.04	20.09.01	20.09	20.09.01	20.09
	20.01.10	20.09.01	20.09		
	20.09.01	20.09.01	20.09		
	42.10	20.09.01	20.09		
YML049c	11.04.03.01	10.01.05.03	10.01	10.01.05.03	10.01
	14.10	10.01.05.03	10.01		
	16.03.03	10.01.05.03	10.01		
	20.09.07	10.01.05.03	10.01		
YER012w	14.07.11	14.13	14.13	14.07.11 14.13	14.13
	14.13.01.01	10.01.05.01,14.07.11,14.13	14.13		
	32.01	10.01.05.01,14.07.11,14.13	14.13		
	43.01.03.09	14.13	14.13		
YER007c	99	20.01.15	20.01	20.01.15	20.01

From the result, we find that the predicted function using two methods are high consistency after blank out the function of the nodes. Even though the results are high consistency with the original results.

3) Complement to BLASTP method: We divide every network into two subsets, one subset include known proteins and the other subset include unknown proteins. We use blast-2.2.25+(downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/)to predict the function of unknown proteins based on sequence similarity. The BLAST database are required to run BLAST locally, database downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. The main control parameter evalue which represents expectation value threshold for saving hits. We set the evalue 0.1 to compare the sequence similarity. The result summarize in table 3.

TABLE III
RESULTS FROM BLASTP PROGRAM

Network	711-704	988-2455	1827-2238	2617-11855
NPP	0	4	28	45

NPP: Numbers of Predicted Proteins

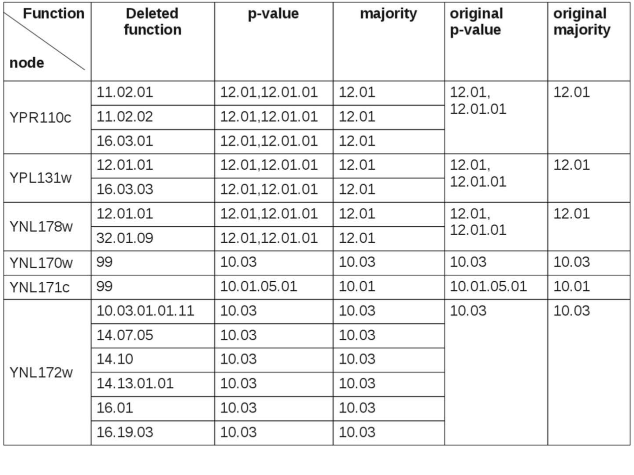
We list the number of predicted proteins in every network. The results show that limited number of proteins are predicted based on BLASTP. Our methods predict much more proteins. The predicted results generate complement effect to BLASTP method. There are one text file of BLASTP results(result of blastp.pdf) in supplementary data. We only list the proteins which have the lowest evalue.

E. Comparison of conductance

Several researches used conductance to measure the goodness of clusters. Roughly speaking, the conductance is the ratio the number of edges on the boundary of a set with the number of edges in the cluster. Let the volume $\mathrm{vol}(S)$ of a set S be the total degree of nodes in it,i.e., $\mathrm{vol}(S) = \sum_{v \in S} \deg(v)$. The conductance $\phi(S)$ of a set S is defined to be the ratio of the number of edges $e(S, \bar{S})$ coming out of S with the minimum of the volume of itself and the volume of its complement $\bar{S}$, i.e. $\phi(S) = e(S, \bar{S}) / \min\{\mathrm{vol}(S), \mathrm{vol}(\bar{S})\}$.

To measure the goodness of a cluster in a network, we compute the conductance of the cluster. For the $2617 - 11855$ network, we get sixty-six known clusters, sixty-four unknown clusters. The conductances of most of these clusters are larger than 0.5. The same network was studied by [4] and they got 48 quasi-clique, during the 48 conductance values, there were 23 values larger than 0.5. Our result is different from that. The different result may be due to the used different clustering methods for the same network.

Fig. 4. Function prediction of 2617-11855 network



Function / node	Deleted function	p-value	majority	original p-value	original majority
YPR110c	11.02.01	12.01,12.01.01	12.01	12.01, 12.01.01	12.01
	11.02.02	12.01,12.01.01	12.01		
	16.03.01	12.01,12.01.01	12.01		
YPL131w	12.01.01	12.01,12.01.01	12.01	12.01, 12.01.01	12.01
	16.03.03	12.01,12.01.01	12.01		
YNL178w	12.01.01	12.01,12.01.01	12.01	12.01, 12.01.01	12.01
	32.01.09	12.01,12.01.01	12.01		
YNL170w	99	10.03	10.03	10.03	10.03
YNL171c	99	10.01.05.01	10.01	10.01.05.01	10.01
YNL172w	10.03.01.01.11	10.03	10.03	10.03	10.03
	14.07.05	10.03	10.03		
	14.10	10.03	10.03		
	14.13.01.01	10.03	10.03		
	16.01	10.03	10.03		
	16.19.03	10.03	10.03		

IV. CONCLUSIONS

Many interaction detection methods resulted in large-scale protein-protein interaction data sets. How to extract useful knowledge to give indication for the biology experiments in the future is important. Cluster analysis is one of the most important methods. MCL algorithm which is widely used in bioinformatics is unsupervised cluster algorithm for networks. We choose MCL algorithm as the cluster analysis algorithm. The inflation value is set to be 1.4.

In previous study, the researchers [15] use majority rule to predict the unknown proteins, to better make use of the global information, we use majority methods to predict the function of unknown proteins. We also make use of P-value method. We assign function to the known clusters. In the most cases, the function annotation results are consistent. By using the same two methods, we predict the function of the unknown proteins. The least percentage of similarity reach to 77. These results suggest that the clusters found by MCL algorithm are consistent with the biological knowledge.

For the same 2617-11855 network, we compare the predicted results of unknown proteins with the study of Bu,D. et al., 2003. We predict 194 proteins. Only 10 proteins are the same as the unknown 76 proteins predicted by Bu,D. et al., 2003. This result shows that the information of 184 new predicted unknown proteins can be used in biology experiments. But the accuracy of predicted results need to be verified by the biology experiments.

In this paper, we predict the function of unknown proteins which are from previous data sets, the method also can be used for new data sets and based on the predicted results, experiment cost will drop.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for their valuable comments and suggestions to improve the presentation of this paper.

REFERENCES

[1] Adamcsek, B. et al, ”Cfinder: locating cliques and overlapping modules in biological networks,” *Bioinformatics*, vol.22, no.8, pp.1021, 2006.

[2] Altaf-Ul-Amin, M. et al. ”Development and implementation of an algorithm for detection of protein complexes in large interaction networks,” *BMC Bioinformatics*, vol.7, pp.207, 2006.

[3] Barabasi AL, Albert R. ”Emergence of scaling in random networks,”. *Science*, vol.286, pp.509-512, 1999.

[4] Bu,D., Zhao,Y., Cai,L., Xue,H., Zhu,X., Lu,H., Zhang,J., Sun,S.,Ling,L., Zhang,N. et al. ”Topological structure analysis of the protein-protein interaction network in budding yeast,” *Nucleic Acids Res.*, vol.31, pp.2443-2450, 2003.

[5] Cho, R. J. et al. ”A genome-wide transcriptional analysis of the mitotic cell cycle,” *Mol. Cell*, vol.2, pp.65-73, 1998.

[6] Gavin, A. C. et al. ”Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol.415, pp.141-147, 2002.

[7] Golemis E. ProteinCProtein Interactions. A Molecular Cloning Manual. Cold Spring Harbor, *NY: Cold Spring Harbor Laboratory Press*, 2002.

[8] Gyeong-Mi Park, Sung-Hwan Kim, Hwan-Gue Cho. ”Structural Analysis on Social Network Constructed from Characters in Literature Texts,” *JOURNAL OF COMPUTERS*, VOL. 8, NO. 9, 2013.

[9] Ho, Y. et al. ”Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry,” *Nature*, vol.415, pp.180-183, 2002.

[10] Hughes, T. R. et al. ”Functional discovery via a compendium of expression profiles,” *Cell*, vol.102, pp.109-126, 2000.

[11] Ito, T. et al. ”A comprehensive two-hybrid analysis to explore the yeast protein interactome,” *Proc. Natl Acad. Sci. USA*, vol.98, pp.4569-4574, 2001.

[12] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. ”Lethality and centrality in protein networks,” *Nature*, vol. 411, pp.41-42, 2001.

[13] Mewes,H.W., Frishman,D., Gildener,U., Mannhaupt,G., Mayer,K.,Mokrejs,M., Morgenstern,B., Munsterkotter,M. et al. ”MIPS:a database for genomes and protein sequences,” *Nucleic Acids Res.*, vol.30, pp.3134, 2002.

[14] R.D.Luce, A.D. Perry. ”A method of matrix analysis of group structure,” *Psychometrika*, vol.14, no.2, pp.95-116, 1949.

[15] Schwikowski, B., Uetz, P. and Fields, S. ”A network of protein-protein interactions in yeast,” *Nat. Biotechnol.*, vol.18, pp.1257-1261, 2000.

[16] Sharan,R. et al. ”Conserved patterns of protein interaction in multiple species,” *Proc. Natl Acad. Sci. USA* , vol.102, pp.1974-1979, 2005.

[17] Stijn van Dongen. Graph Clustering by Flow Simulation.*PhD thesis*, University of Utrecht, May 2000.

[18] Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. ”Systematic determination of genetic network architecture,” *Nature Genet*, vol.22, pp.281-285, 1999.

[19] Tyler C. McCandless, Sue Ellen Haupt, George S. Young. ”The Effects of Imputing Missing Data on Ensemble Temperature Forecasts,” *JOURNAL OF COMPUTERS*, Vol 6, No 2, 2011

[20] Uetz, P. et al. ”A comprehensive analysis of proteinCprotein interactions in Saccharomyces cerevisiae,” *Nature*, vol.403, pp.623-627, 2000.

[21] von Mering,C., Krause, R., Snel, B., Cornell, M., Oliver, S.G.,Fields, S. and Bork, P. ”Comparative assessment of largescale data sets of proteinCprotein interactions,” *Nature*, vol.417, pp.399-403, 2002.

[22] Wangren Qiu and Xuan Xiao Lidong WangDianxuan Gong. ”A Novel Pseudo Amino Acid Composition for Predicting Subcellular Location of Proteins, ” *JOURNAL OF COMPUTERS*,VOL. 8, NO. 3, 2013

[23] Wu,L.F., Hughes,T.R., Davierwala,A.P., Robinson,M.D., Stoughton,R.and Altschuler,S.J. ”Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters,” *Nature Genet.*, vol.31, pp.255-265, 2002.

YunQuan Zhang received the PhD degree in computer science from Institute of Software,Chinese Academy of Sciences in 2000. He is interested in high performance computing, performance evaluation, parallel numerical software design, parallel computational model, parallel data mining and bioinformatics. He is a research professor, Supervisor of PhD Candidates in S-tate Key Lab. of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences.

Ke Zhan received master degree in biochemistry and molecular biology from HuaZhong Agriculture University, WuHan,China, in 2010. He is currently working toward the PhD degree in computer science at Institute of Software,Chinese Academy of Sciences. He is interested in bioinformatics and parallel computing.