An Enhanced Markov Clustering Method for Detecting Protein Complexes

Charalampos N. Moschopoulos, Georgios A. Pavlopoulos, Spiridon D. Likothanassis and Sofia Kossida

***Abstract*— With the recent high-throughput methods, large datasets of experimentally detected pairwise protein –protein interactions are generated. However, these data suffer from noise, reducing the quality of the information they bring (identification of protein complexes). This paper introduces a novel methodology for detecting protein complexes in a protein – protein interaction graph. Our method initially uses the Markov clustering algorithm and then filters the derived results in order to obtain the best set of clusters that represent protein complexes. The efficiency of our method is shown in experimental results derived from 7 different yeast protein interaction datasets. Moreover, comparisons with 4 other algorithms are performed proving that our method predicts known protein complexes, recorded in the MIPS database, more accurately.**

I. Introduction

The study of protein interactions has been vital to the understanding of how proteins function within the cell. More specifically, protein interactions are crucial for forming structural complexes, for extra-cellular signalling, intra-cellular signalling, cell communication and several other aspects of cellular function [1].

There are several experimental methods such as pull down assays [2] and tandem affinity purification [3] that are used in order to detect protein interactions in an organism. Today, relatively new high-throughput methods (yeast two hybrid systems [4], mass spectrometry [1], microarrays [5] and phage display [6]) generate enormous datasets of protein – protein interactions. However, despite the wide variety of experimental methods, only a small fraction of protein complexes have been identified due to the weakness of these methods to detect all the proteins composing these complexes [7]. Additionally, high throughput methods are error prone as they miss a fraction of protein interactions and yield several protein interactions that do not exist in nature.

Manuscript received August 4, 2008.

C. N. Moschopoulos is with the Department of Computer Engineering & Informatics at University of Patras and with Bioinformatics & Medical Informatics Team of Biomedical Research Foundation of the Academy of Athens, GR-11527 GREECE (phone: +30-210-6597199; fax: +30-210-6597545; e-mail: mosxopul@ceid.upatras.gr).

G. A. Pavlopoulos is with European Molecular Biology Laboratory, Heidelberg, D-69117 GERMANY (e-mail: pavlopou@embl.de).

S. D. Likothanassis is with the Department of Computer Engineering & Informatics at University of Patras, GR-26500 GREECE (e-mail: likothan@ceid.upatras.gr).

S. Kossida is with Bioinformatics & Medical Informatics Team of Biomedical Research Foundation of the Academy of Athens, GR-11527 GREECE (e-mail: skossida@bioacademy.gr).

Because of the unreliability of the protein interaction data, computational methods of data mining or knowledge discovery are necessary to gain valuable information such as the discovery of protein complexes. Usually, these methods, model the protein interaction datasets as an unweighted and undirected graph defined as $G=(V,E)$ where V represents the set of vertices (proteins) and E represents the set of edges (interactions). In these graphs, a protein complex generally corresponds to a dense subgraph that is an aggregation of vertices that are highly interactive with each other.

Previous approaches use either a local search strategy or a hierarchical one. In the first category, the best known algorithm is the Molecular Complex Detection (Mcode) [8] A year before Mcode was published, another algorithm called TRIBE-MCL, which was based on MCL, was presented for detecting protein families [9]. Besides that, King et al. suggested the RNSC algorithm [10] which uses a cost local search algorithm based loosely on a tabu search meta – heuristic. Another algorithm of the local search approach is the Local Clique Merging Algorithm (LCMA) [11] which locates cliques in a graph and subsequently tries to expand them. On the other hand, most of the hierarchical clustering approaches are based on the concept of dividing the initial graph by removing the minimum set of edges. The Highly Connected Subgraph method (HCS) [12] separates a graph into several subgraphs using minimum cuts and stops when the cut is bigger or equal to the number of graph vertices divided by 2. Koyutürk suggested the SIdeS algorithm [13] which uses the HCS algorithm philosophy with a different stopping criterion which is based on the statistical significance of the derived subgraphs.

In this paper, a new methodology, called Enhanced Markov Clustering (EMC), is presented. EMC detects protein complexes from protein – protein interaction graph in two steps: In the first step, the protein - protein interaction network is clustered by Markov clustering algorithm (MCL) [9] and in the second step the results are filtered based either on individual or on a combination of 4 different methods (density, haircut operation, best neighbour and cutting edge). Extensive experiments were performed on 7 different datasets which were either derived from individual experiments (Ito [4], Tong [14], Krogan [15] and Gavin [1], [16]) or from online databases (DIP [17] and MIPS [18]). These datasets vary on the number of proteins as well as the number of interactions, composing either sparse (Ito and

Tong datasets) or relatively dense (MIPS and DIP datasets) graphs. Moreover, by using the yeast proteome, the most well studied organism concerning protein – protein interactions, EMC was compared with 4 other algorithms: the Mcode algorithm, the HCS algorithm, the SideS algorithm and the RNSC algorithm and examined the derived results based on 5 different metrics. As it can be seen in the Results section, EMC outperforms all the other algorithms and generates remarkable results.

II. OUR METHOD

To identify accurate protein complexes given a protein-protein interaction network, we built a workflow consisting of a two step procedure. Initially, a protein - protein interaction network is clustered by Markov clustering algorithm (MCL) and in the second step the results are filtered based either on individual or on a combination of 4 different methods. These are density, haircut operation, best neighbour and cutting edge. This two step approach preserves only those clusters that have high probability to be real biological complexes. A brief description of the MCL algorithm and the criteria used for the filtering procedure is given below.

A. Description of the MCL algorithm

The MCL algorithm [9] is a fast and scalable unsupervised clustering algorithm based on simulation of stochastic flow in graphs. The MCL algorithm can detect cluster structures in graphs by a mathematical bootstrapping procedure. The process deterministically computes the probabilities of random walks through a graph, and uses two operators transforming one set of probabilities into another. It does so by using the language of stochastic matrices (also called Markov matrices) which capture the mathematical concept of random walks on a graph.

B. Cluster Density

Protein complexes correspond to dense subgraphs or even cliques in protein interaction graphs [19]. Therefore, clusters of high density are more likely to correspond to known protein complexes. The density of a subgraph is calculated by the formula below:

$$\frac{2|E|}{|V|(|V|-1)},$$

where $|E|$ is the number of edges and $|V|$ the number of vertices of the subgraph.

C. Haircut operation

Haircut operation is a method that detects and excludes vertices with low degree of connectivity from the potential cluster that these nodes belong to. Proportionally, the lower the connectivity of a node is, the lower the probability for this node to belong to a protein complex is. In such a way, the deletion of such nodes that add noise to the cluster leads to protein complexes that are more likely to be present in nature.

D. Best neighbor method

In contrast with haircut operation method, best neighbor method tends to detect and enrich the clusters with candidate vertices that are considered as good "neighbors". Such a node is the one where the proportion of its edges adjacent to the cluster divided by the total degree of the vertex is above a threshold defined by the user:

$$\frac{|\textbf{adjacent edges}|}{|\textbf{total edges}|} > threshold$$

The best neighbor method is mostly suitable to detect larger protein complexes that offer extra information about protein complexes included in a protein interaction dataset.

E. Cutting edge metric

Analyzing the structure of a protein–protein interaction network, molecular modules are densely connected within themselves but are sparsely connected to the rest of the network [20]. To address these cases, a filtering criterion was applied, called cutting edge and is defined as:

$$\frac{|\textbf{inside edges}|}{|\textbf{total edges}|},$$

where $|\textbf{inside edges}|$ is the number of edges inside a cluster and $|\textbf{total edges}|$ is the number of edges that are adjacent to at least one vertex of the cluster. The clusters in which the cutting edge metric is below a user defined threshold are discarded from the filter of our method.

III. DATASETS

To demonstrate the use of our methodology, we used seven datasets derived from various small scale and high-throughput methods. The multifaceted nature of the datasets enables us to perform a more “objective” comparison of the algorithms tested. In this section, we give a short description of the datasets that were used.

A. ITO dataset

Based on a system that examines every possible two-hybrid pair of protein interaction of the budding yeast *Saccharomyces cerevisiae,* this dataset consists of 4038 two-hybrid interactions among 3279 proteins [4]. Initially, a single huge network linking the vast majority of the proteins was produced. This network was reduced by selecting biologically relevant interactions highlighting various intriguing subnetworks. Our method locates successfully these subnetworks and allows us to expand and improve the protein interaction map for the exploration of genome functions by finding the complexes that are biologically more relevant.

B. Tong dataset

This network consists of 7430 edges and 2262 vertices [14]. A genetic interaction network was mapped by crossing mutations in several genes into a set of viable gene yeast deletion mutants scoring the double mutant progeny for fitness defects. The interactions of this network were produced by predicting the functions of the interactive elements often produced by bringing together functionally related genes or components or elements that belong to the same pathway. The genetic network exhibited dense local neighborhoods; our method aims to go one step further by predicting these neighborhoods but also by spliting them in smaller groups that are functionally more significant.

C. Krogan dataset

This dataset consists of 7088 edges and 2675 vertices and contains different tagged proteins of the yeast *Saccharomyces cerevisiae*. In a previous analysis [15], the MCL algorithm was used to cluster and organize the proteins into several groups that about half of them were absent from the MIPS database. We observed that a small amount of noise was added to these data and therefore we applied our method to detect and filter the groups detected by MCL.

D. Gavin_2002-2006 datasets

In this case, we used two networks, the first consisting of 3210 edges and 1352 vertices and the second consisting of 6531 edges and 1430 vertices [1], [16]. In the first dataset, large scale tandem affinity purification and mass spectrometry were used to characterize multiprotein complexes in *Saccharomyces cerevisiae*. Extending this information to human genome, this dataset provides an outline of the eukaryotic proteome as a network of protein complexes. Using the whole network, we try to see how successfully our method isolates the network complexes. The second dataset comes with the first genome-wide screen for complexes in yeast.

E. DIP dataset

The Database of Interacting Proteins (*DIP*) is a database that documents experimentally determined protein-protein interactions [17]. We used this database to isolate a network consisting of 17491 edges and 4934 vertices. One of the reasons why we included this source for our experiments is because beyond cataloging details of protein-protein interactions, the DIP database helps us not only understand protein functions but the value of protein-protein relationships as well.

F. MIPS dataset

The Munich Information Center for Protein Sequences provides resources mainly related to genome information [18]. Most of the databases that contain information about a variety of genomes of different organisms are manually curated. Furthermore 400 genomes that were automatically annotated are also included. One of the aims of this database is to provide information related to interactions such as protein–protein interactions. In this study case we isolated a network consisting of 12526 edges and 4554 vertices given by the MIPS database.

IV. EXPERIMENTAL RESULTS

A. Evaluation procedure

The benchmark that we used to evaluate the algorithms tested consists of known yeast protein complexes retrieved from the MIPS database Furthermore, MIPS protein complexes composed from smaller ones, also recorded in MIPS database, were removed. The final evaluation dataset comprises 220 complexes.

In addition to the collection of MIPS protein complexes, we also used the same evaluation metric adopted in [8], called geometric similarity index. This method considers a predicted complex as valid if $\frac{I^2}{A*B} > 0.2$ where **I** is the number of common proteins**, A** the number of proteins in the predicted complex and **B** the number of proteins in the recorded complex. We used this measurement to evaluate our results.

Moreover, 4 different matching statistic metrics, that were presented in [21], were used in the evaluation process of the algorithms tested. These are *sensitivity* (Sn), *Positive Predictive Value* (PPV) and *Geometrical Accuracy* (Acc_g). These metrics are typically used to measure the correspondence between the result of a classification and a reference.

B. Results

All the results of the experiments performed are demonstrated in Table 1. It is clear that EMC methodology outperforms all the other algorithms in the percentage of successful predictions in all the cases. In Figure 1, we show the percentage of successful predictions, where EMC overpasses all the other algorithms. Moreover, EMC achieves better approximations of real protein complexes than the other algorithms in almost all of the cases, something that is proven by the high values of the mean score of the valid predictions.

Additionally, the performance of EMC following the classical statistics is exceptional, comparing to the other algorithms. Figure 2 shows the results of the algorithms concerning the Acc_g metric which indicates the trade off between the metrics of Sn and PPV ($Acc\_g = \sqrt{Sn * PPV}$).

TABLE 1.

Summary of experimental results. The percentage of successful predictions is shown in the first column; the absolute number of valid predicted complexes is shown in the second column as well as the total number of predicted complexes. The mean score of the valid predicted complexes is shown in the third column. The last three columns present the Sensitivity (Sn), the Positive Predictive Value (PPV) and the geometric Accuracy (Acc_g) respectively.

Algorithms	Percentage of successful prediction	Absolute number of predictions	Mean Score of valid predicted complexes	Sn	PPV	Acc_g
IT0 dataset						
SideS	14.29%	2/14	0.292	85.19%	48.19%	64.04%
Mcode	9.09%	1/11	0.32	100.00%	64.71%	80.44%
HCS	7.69%	1/13	0.333	80.00%	40.00%	56.57%
RNSC	14.28%	1/7	0.563	100.00%	40.00%	63.25%
EMC	18.75%	3/16	0.498	100.00%	92,86%	96.36%
Tong dataset						
SideS	16.67%	4/24	0.317	89.08%	42.02%	61.18%
Mcode	10.81%	4/37	0.517	78.37%	41.89%	57.30%
HCS	13.64%	3/22	0.311	89.68%	40.48%	60.25%
RNSC	17.65%	3/17	0.432	90.00%	50.00%	67.08%
EMC	20.00%	4/20	0.620	98.30%	45.76%	67.07%
Krogan dataset						
SideS	46.15%	36/78	0.519	84.29%	56.94%	70.61%
Mcode	31.94%	23/72	0.614	87.00%	70.00%	78.04%
HCS	44.44%	32/72	0.578	88.40%	56.46%	72.43%
RNSC	42.86%	33/77	0.590	90.28%	58.33%	74.31%
EMC	54.84%	17/31	0.665	98.40%	63.30%	78.92%
Gavin_2002 dataset						
SideS	51.16%	22/43	0.440	78.57%	57.14%	67.01%
Mcode	35.00%	7/20	0.478	95.08%	50.00%	68.95%
HCS	55.56%	20/36	0.423	89.30%	57.93%	71.93%
RNSC	61.70%	29/47	0.512	85.71%	59.52%	71.43%
EMC	62.96%	17/27	0.539	95.24%	55.78%	72.89%
Gavin_2006 dataset						
SideS	37.76%	37/98	0.545	74.86%	57.73%	65.74%
Mcode	50.00%	31/62	0.543	73.44%	51.65%	61.59%
HCS	46.84%	37/79	0.528	81.42%	54.41%	66.56%
RNSC	50.62%	41/81	0.566	74.26%	81.19%	77.65%
EMC	55.17%	16/29	0.678	95.27%	63.51%	77.79%
DIP dataset						
SideS	46.07%	41/89	0.478	75.84%	53.56%	63.73%
Mcode	31.82%	21/66	0.515	71.71%	41.47%	54.53%
HCS	47.95%	35/73	0.504	80.34%	50.57%	63.74%
RNSC	47.54%	29/61	0.553	69.23%	92.31%	79.94%
EMC	58.33%	7/12	0.747	100.00%	72.60%	85.21%
MIPS dataset						
SideS	42.65%	29/68	0.597	80.11%	49.05%	62.68%
Mcode	43.10%	25/58	0.506	82.47%	44.66%	60.69%
HCS	45.00%	27/60	0.666	83.99%	47.75%	63.33%
RNSC	42.62%	26/61	0.661	86.36%	54.55%	68.63%
EMC	62.50%	20/32	0.657	100.00%	71.77%	84.72%

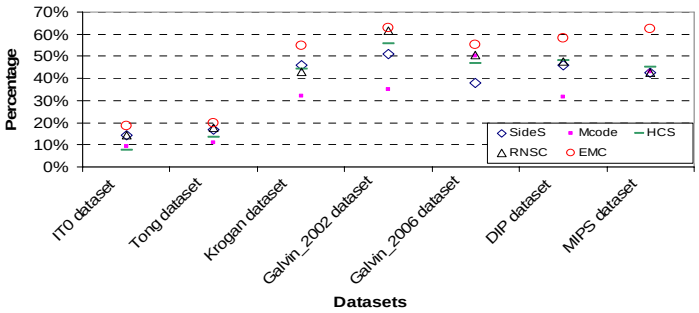

Fig. 1. The percentage of successful predictions of the algorithms tested

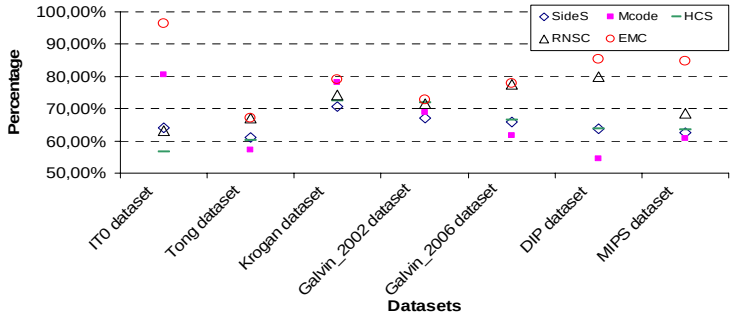

Fig. 2. The performance of the algorithms concerning Acc_g metric.

Going one step further, we tested the parameters of the filter created for the EMC methodology. We tried to achieve good prediction rate without minimizing the number of the final MCL clusters that will pass the filtering process. Specifically, our experiments showed that a density cutoff between 0.6 and 0.75 and a haircut of vertices with less than 2 degree allows a good prediction rate and high values on the metrics used. The methods of best neighbor and cutting edge helped in the improvement of predictions but their values vary depending the dataset tested. Nevertheless, in sparse graphs, better results were obtained without the need of these methods. The best neighbor method was helpful only in two of the denser datasets. Table 2 describes methods and parameters used in the filtering process.

TABLE 2
The methods used in the filtering process

Dataset	Filter
ITO	Density=0.75, Haircut=2
Tong	Density=0.75, Haircut=2
Krogan	Cutting_Edge=0.55, Density=0.7, Haircut=3
Gavin_2002	Cutting_Edge=0.5, Density=0.6, Haircut=2
Gavin_2006	Cutting_Edge=0.75, Density=0.6, Haircut=2, Best_neighbor =0,6
DIP	Cutting_Edge=0.5, Density=0.6, Haircut=3
MIPS	Cutting_Edge=0.5, Density=0.7, Haircut=2, Best_neighbor =0,75

Table 3 shows the results obtained from the MCL algorithm. These results highlight the significance of the filtering process that EMC uses.

TABLE 3.
The results of the MCL algorithm

Algorithms	Percentage of successful prediction	Absolute number of predictions	Mean Score of valid predicted complexes	Sn	PPV	Acc_g
ITO dataset						
MCL	5.56%	35/630	0.372	34.9%	42.66%	38.58%
Tong dataset						
MCL	4.62%	16/346	0.346	44.66%	40.79%	42.40%
Krogan dataset						
MCL	22.57%	58/257	0.484	69.67%	54.45%	61.59%
Galvin_2002 dataset						
MCL	33.49%	71/212	0.574	74.11%	57.01%	65.00%
Galvin_2006 dataset						
MCL	31.22%	59/189	0.527	75.75%	54.26%	64.11%
DIP dataset						
MCl	10.13%	98/967	0.457	47.58%	53.25%	50.34%
MIPS dataset						
MCL	9.70%	87/897	0.478	44.77%	52.40%	48.43%

C. Implementation

The EMC methodology is implemented in C language same as the SideS, RNSC and HCS algorithm. The Mcode algorithm is implemented as a java plugin for the Cytoscape Tool. All the experiments were performed using an Intel Double Core 2.13GHz processor, with 1GB of RAM and Suse Linux 10.1(x86_64) operating system. Loop edges were not taken into account and predicted protein complexes containing less than 3 proteins were discarded during our experimental sets.

The filter we used for the results of the RNSC algorithm was composed by two out of three parameters as they are presented in [10] (size and density). We did not use the third (functional homogeneity) as this kind of information was not available for all datasets so that the comparison with the other algorithms, which did not use this kind of information, would not be biased.

The SideS and HCS algorithms do not take any parameters, whereas for the use of Mcode and MCL algorithms we used the optimal parameters for accuracy as they are defined in [21].

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a new methodology for detecting protein complexes through clustering protein – protein interaction graphs. This methodology, called EMC, uses MCL algorithm and a filter composing 4 different methods. These methods can be uniquely chosen or combined depending on the study case. We tested our method with 7 different protein interaction datasets and compared it with 4 other algorithms in order to prove its efficiency. For the evaluation process, we used 5 different metrics to prove the quality of our results. The future prospect of our work is to use machine learning techniques in order to optimize the parameters used in the filtering process. This way, it will be reassured that the EMC methodology will obtain satisfactory results whatever the input graph instance happens to be.

APPENDIX

The results of our experiments are available in:
http://www.bioacademy.gr/bioinformatics/projects/EMC

REFERENCES

[1] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature,* vol. 415, pp. 141-7, Jan 10 2002.

[2] H. G. Vikis and K.-L. Guan, "Glutathione-S-Transferase-Fusion Based Assays for Studying Protein-Protein Interactions," in *Protein-Protein Interactions, Methods and Applications,*

Methods in Molecular Biology. vol. 261, H. Fu, Ed. New Jersey: Humana Press, 2004, pp. 175-186.

[3] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Seraphin, "The tandem affinity purification (TAP) method: a general procedure of protein complex purification," *Methods,* vol. 24, pp. 218-29, Jul 2001.

[4] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome " *Proceedings of the National Academy of Science,* vol. 98, pp. 4569-4574, 04/2001 2001.

[5] D. Stoll, M. F. Templin, J. Bachmann, and T. O. Joos, "Protein microarrays: applications and future challenges," *Curr Opin Drug Discov Devel,* vol. 8, pp. 239-52, Mar 2005.

[6] W. G. Willats, "Phage display: practicalities and prospects," *Plant Mol Biol,* vol. 50, pp. 837-54, Dec 2002.

[7] R. P. Sear, "Specific protein-protein binding in many-component mixtures of proteins," *Phys Biol,* vol. 1, pp. 53-60, Jun 2004.

[8] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics,* vol. 4, p. 2, Jan 13 2003.

[9] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Res,* vol. 30, pp. 1575-84, Apr 1 2002.

[10] A. D. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics,* vol. 20, pp. 3013-20, Nov 22 2004.

[11] X.-L. LI, S.-H. TAN, C.-S. FOO, and S.-K. NG, "Interaction Graph Mining for Protein Complexes Using Local Clique Merging," *Genome Informatics,* vol. 16, pp. 260-269, December 2005 2005.

[12] E. Hartuv and R. Shamir, "A clustering algorithm based on graph connectivity " *Information Processing Letters,* vol. 76, pp. 175-181, 2000.

[13] M. Koyuturk, W. Szpankowski, and A. Grama, "Assessing significance of connectivity and conservation in protein interaction networks," *J Comput Biol,* vol. 14, pp. 747-64, Jul-Aug 2007.

[14] A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Menard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A. M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, "Global mapping of the yeast genetic interaction network," *Science,* vol. 303, pp. 808-13, Feb 6 2004.

[15] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature,* vol. 440, pp. 637-43, Mar 30 2006.

[16] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga, "Proteome survey reveals modularity of the yeast cell machinery," *Nature,* vol. 440, pp. 631-6, Mar 30 2006.

[17] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Res,* vol. 28, pp. 289-91, Jan 1 2000.

[18] H. W. Mewes, D. Frishman, K. F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stumpflen, "MIPS: analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Res,* vol. 34, pp. D169-72, Jan 1 2006.

[19] S. H. Yook, Z. N. Oltvai, and A. L. Barabasi, "Functional and topological characterization of protein interaction networks," *Proteomics,* vol. 4, pp. 928-42, Apr 2004.

[20] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proc Natl Acad Sci U S A,* vol. 100, pp. 12123-8, Oct 14 2003.

[21] S. Brohee and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics,* vol. 7, p. 488, 2006.