REFINING MARKOV CLUSTERING FOR PROTEIN COMPLEX PREDICTION BY INCORPORATING CORE-ATTACHMENT STRUCTURE

SRIGANESH SRIHARI[1] KANG NING[2] HON WAI LEONG[1]
srigsri@comp.nus.edu.sg kning@umich.edu leonghw@comp.nus.edu.sg

[1] *School of Computing, National University of Singapore, Singapore 117590*
[2] *Department of Pathology, University of Michigan, Ann Arbor, MI*

Protein complexes are responsible for most of vital biological processes within the cell. Understanding the machinery behind these biological processes requires detection and analysis of complexes and their constituent proteins. A wealth of computational approaches towards detection of complexes deal with clustering of protein-protein interaction (PPI) networks. Among these clustering approaches, the Markov Clustering (MCL) algorithm has proved to be reasonably successful, mainly due to its scalability and robustness. However, MCL produces many noisy clusters, which either do not represent any known complexes or have additional proteins (noise) that reduce the accuracies of correctly predicted complexes. Consequently, the accuracies of these clusters when matched with known complexes are quite low. Refinement of these clusters to improve the accuracy requires deeper understanding of the organization of complexes. Recently, experiments on yeast by Gavin *et al.* (2006) revealed that proteins within a complex are organized in two parts: core and attachment. Based on these insights, we propose our method (MCL-CA), which couples core-attachment based refinement steps to refine the clusters produced by MCL. We evaluated the effectiveness of our approach on two different datasets and compared the quality of our predicted complexes with that produced by MCL. The results show that our approach significantly improves the accuracies of predicted complexes when matched with known complexes. A direct result of this is that MCL-CA is able to cover larger number of known complexes than MCL. Further, we also compare our method with two very recently proposed methods CORE and COACH, which also capitalize on the core-attachment structure. We also discuss several instances to show that our predicted complexes clearly adhere to the core-attachment structure as revealed by Gavin *et al.*

Keywords: Protein complex; PPI network; Markov Clustering.

1. Introduction

One of the most interesting challenges of postgenomic biology is to understand how proteins interact and organize themselves to generate vital biological functions. Biological processes such as cell cycle, replication, and signal transduction require precise arrangement of protein molecules for proper assembly and function. However, many principles as to how individual proteins form such molecular structures are still unknown. *Protein complexes* are amongst the fundamental units of this macromolecular organization and are responsible for most of vital biological processes within the cell. For example, the HOPS complex associates with the vacuolar membrane and is involved in homotypic fusion and vacuole protein sorting in yeast [11]. Therefore, understanding the underlying organizational principles of cellular machinery requires detection and analysis of protein complexes.

Most computational methods for detection of protein complexes work on protein interaction data. The interactions among proteins are assembled in the form of a protein-protein interaction (PPI) network: each node represents a unique protein, and an edge between a pair of nodes represents an interaction (verified through reliable biological experiments) between the two corresponding proteins. Transient interactions are not considered. PPI networks are undirected and may be unweighted or weighted, with the weight on an edge representing the confidence score (usually between 0 and 1) for the interaction. Protein complexes within the PPI network form stable subnetworks which are believed to have relatively more number of interactions. Therefore, most computational methods focus on detecting dense subnetworks within the PPI network.

A wealth of methods make use of graph clustering procedures to detect complexes [2, 6, 13]. The K-means and hierarchical clustering methods are the classical ones, along with some more recent ones (see [6]). Clustering is an unsupervised learning method that identifies intrinsic similarities between data elements in order to group them into disjoint or overlapping substructures. It typically involves a metric or similarity measure to achieve this grouping. It requires solving an optimization problem to arrive at the best solution (global minimum of a defined error quantity). However, the application of clustering approaches to biological networks encounters yet another challenge: their ability to deal with high levels of noise.

Among the clustering algorithms, the Markov Clustering Algorithm (MCL) has become very popular since its proposal in 2000 [2]. It is a fast and scalable unsupervised clustering algorithm. It makes use of the concept of random walks: given a graph G, assume a walk taken along the graph starting at an arbitrary node v. The walker visits every neighbor u of v with equal probability. If she enters a dense region of G, she will remain in the dense region with high probability. So, by simulating a large number of random walks (called a *flow*), the underlying cluster structure in the graph can be identified. Unlike other clustering procedures, MCL considers the connectivity properties of the underlying network, and is therefore capable of deriving clusters that are relatively dense, even though they may not have high absolute densities.

So far, MCL seems to be the most successful clustering algorithm for deriving protein complexes from PPI networks [3, 13]. It was shown to outperform a number of algorithms like MCODE [1] specifically designed for partitioning PPI networks. It has been used for comprehensive analysis of complexes derived from the yeast interactome [7, 12]. Very recently, Vlasblom *et al.* [13] showed that MCL performed far better than other successful algorithms like Affinity Propagation (AP) [4] on

PPI networks. MCL was shown to be highly robust and tolerant to varying levels of noise, typical to biological networks.

Inspite of these advantages, MCL has some limitations when applied to PPI networks. MCL mainly produces disjoint clusters from the underlying network (though rarely it may report overlapping clusters by changing some ‘options’ in the software). However, it is known from previous studies [1] that protein complexes may be overlapping: a protein may take part in more than one complex. Secondly, MCL is a general network clustering algorithm and does not take into account the biological properties or organization within protein complexes. Our experiments show that MCL produces many noisy clusters, which either do not represent any known complexes or have additional proteins (noise) that reduce the accuracies of correctly predicted complexes. Consequently, the accuracies of these clusters when matched with known complexes are quite low. Additionally, MCL generates many clusters without actually ranking them using some biologically meaningful criteria.

In view of the existing interest in applying MCL (and clustering methods in general) to identify complexes from PPI networks, this paper attempts to make use of the advantages of MCL, and also overcome some of the limitations mentioned above. We refine the clusters produced by MCL to improve their accuracies. These refinement steps make use of deeper understanding of protein complex organization, recently revealed by Gavin *et al.* [5].

The experiments by Gavin *et al.* [5] on yeast revealed that proteins within a complex are organized as *cores* and *attachments* (see Figure 1 in supplementary materials). These core proteins show high-level of functional similarity and have relatively more interactions among themselves. Attachment proteins are closely-associated with these core proteins, and they together form the complex. An attachment protein may be present in more than one complex. Among the attachments there may be *modules*, which are sets of two or more proteins always together and present in multiple complexes.

We propose our method to refine the clusters produced by MCL using the core-attachment structure as revealed by Gavin *et al.*’s work. We call our method MCL coupled with Core-Attachment (MCL-CA). Very recently (2009), Henry Leung *et al.* proposed the CORE algorithm [8], and Min Wu *et al.* proposed the COACH algorithm [9], which also make use of the core-attachment structure to predict complexes from PPI networks.

We tested MCL-CA on two high-quality yeast PPI datasets from Gavin *et al.* [5] and Krogan *et al.* [7]. We show that MCL-CA is able to significantly improve the accuracies of complexes predicted by MCL when matched with known complexes. A direct result of this is that MCL-CA is able cover larger number of known complexes than MCL. Further, we also compare our method with CORE and COACH, since these two methods also capitalize on the core-attachment structure.

2. Methods

In this section, we describe the various steps in our computational approach. Before we describe our algorithm, we introduce a few notations. Our PPI graph is undirected and unweighted, and is denoted as $G = (V, E)$, where V is the set of proteins and E is the set of interactions between the proteins. Any subgraph of G is represented as $S = (V_s, E_s)$ for $V_s \subseteq V$ and $E_s \subseteq E$. For a protein $p \in V$, $N(p)$ is the set of neighbors of p. Our algorithm is structured as a sequence of following steps:

(i) Cluster the PPI graph using MCL.
(ii) Determine core proteins.
(iii) Filter out noisy clusters.
(iv) Determine attachment proteins.
(v) Determine module proteins.
(vi) Determine complexes and rank them.

2.1. *Clustering the PPI graph using MCL*

The first step in our algorithm is to cluster the PPI graph using the MCL algorithm. Upon running MCL on $G = (V, E)$, we obtain a set of k disjoint clusters of proteins, given by $\{C_i : C_i = (V_i, E_i), 1 \leq i \leq k\}$, where $V_i \subseteq V$ and $E_i \subseteq E$. Also, $\bigcup_i V_i = V$. And, for any $1 \leq i, j \leq k, i \neq j$, $V_i \cap V_j = \emptyset$. Among these clusters, we discard away all clusters of size 1 and retain the remaining.

2.2. *Determining core proteins*

Core proteins, as described by Gavin *et al.* [5], show greatest degree of physical association, high similarity in expression levels, and represent functional units within complexes. In our model, we consider core proteins as the set of proteins that satisfy the following two properties: (a) Every complex has a set of core proteins; (b) The core proteins in a protein complex have relatively more interactions among themselves, and less interactions with proteins outside the complex.

We categorize a protein $p \in V_i$ to be a core protein in cluster C_i, given by $p \in Core(C_i)$, if: (a) The *in-degree* of p with respect to the cluster C_i is greater than the *average in-degree* of C_i, given by: $d_{in}(p, C_i) \geq d_{avg}(C_i)$; and (b) The in-degree of p with respect to C_i is greater than the *out-degree* of p with respect to C_i, which is given by: $d_{in}(p, C_i) > d_{out}(p, C_i)$. The in-degree of $p \in V_i$ with respect to C_i is the number of interactions p has with proteins within the cluster C_i. It is defined as $d_{in}(p, C_i) = |N(p, C_i)| = |\{(p, q) : (p, q) \in E_i, q \in V_i\}|$. Similarly, the out-degree of $p \in V_i$ with respect to C_i is the number of interactions p has with proteins outside the cluster C_i. It is defined as $d_{out}(p, C_i) = |\{(p, r) : (p, r) \in E, r \notin V_i\}|$. Then the average in-degree of cluster C_i is the average of in-degrees of all proteins within C_i, given by $d_{avg}(C_i) = \sum d_{in}(p, C_i)/|C_i|$.

2.3. *Filtering out noisy clusters*

As per Gavin *et al.*'s work [5], core proteins represent functional units within complexes. Therefore, we consider clusters without any core proteins as noisy clusters. Additionally, our experiments showed that MCL produces many isolated two-protein clusters (two proteins linked to each other, and not linked to any other protein) that have very low significance and do not represent any real complexes. We consider all such clusters as noisy and filter them out.

2.4. *Determining attachment proteins*

Attachment proteins, as described by Gavin *et al.* [5], are densely associated with the core proteins, and show greater heterogeneity in expression levels. In our model, we consider attachment proteins as the set of proteins that satisfy the following property: The attachment proteins of any complex are densely connected to the core proteins of that complex. An attachment protein is not unique to a complex, instead may be present in more than one complex.

We categorize the attachment proteins in cluster C_i into *local* and *foreign* attachments, together represented by the set $Attach(C_i)$. We consider a protein p to be a *local* attachment in cluster C_i if: (a) $p \in V_i$; and (b) p is a common neighbor to at least half the core proteins in C_i, that is, $|N(p, C_i) \cap Core(C_i)| \geq |Core(C_i)|/2$. However, our experiments revealed that with (b) we might miss out some true local attachments and therefore, we also include the following heuristic: if $p \in V_i$ is connected to at least two core proteins in C_i but has more neighbors within C_i than outside, that is, $|N(p, C_i) \cap Core(C_i)| \geq 2$ and $d_{in}(p, C_i) > d_{out}(p, C_i)$, then p is a local attachment in C_i. We consider a protein p to be a *foreign* attachment in cluster C_i if: (a) $p \notin V_i$; and (b) p is a common neighbor to more than half core proteins of C_i, that is, $|N(p, C_i) \cap Core(C_i)| > |Core(C_i)|/2$ and $|C_i| > 2$. With these properties, p may be an attachment in two clusters C_i and C_j, that is, $p \in Attach(C_i)$ and $p \in Attach(C_j)$.

2.5. *Determining module proteins*

Modules, as described by Gavin *et al.* [5], are groups of proteins that are most likely to be in direct physical contact with cores, show greatest degree of functional similarity, are least likely to be present partially, and give rise to "cross-talk" between various functional categories. In our model, we consider modules as proteins that satisfy the following two properties: (a) The set of module proteins of a complex form a proper subset of the attachment proteins in that complex; (b) The set of module proteins are present in more than one complex in *entirety.*

Let proteins p and q be attachments in two clusters C_i and C_j. We consider the set $M = \{p, q\}$ to be a *module* in clusters C_i and C_j, if M is present in entirety within the attachment sets of C_i and C_j. This is given by the condition: $M \subseteq Attach(C_i)$ and $M \subseteq Attach(C_j)$. Note that M can have more than two proteins and may be present in more than two clusters as well.

2.6. *Determining complexes and ranking them*

As per Gavin *et al.*'s work [5], protein complexes are composed of cores and attachments. In our model, we form a unique protein complex C'_i from each cluster C_i: The constituent proteins of the complex C'_i include all the core and attachment proteins of C_i. We discard away all proteins categorized as neither cores nor attachments. Therefore, $C'_i = (V'_i, E'_i)$, where $V'_i = [Core(C_i) \cup Attach(C_i)]$ and $E'_i = \{(p,q) : (p,q) \in E, p \in V'_i, q \in V'_i\}$. The resulting protein complexes may be overlapping with attachment proteins forming part of multiple complexes.

After the protein complexes are constructed, we rank them based on a scoring function $Score$, which is based on the *edge density* γ and *in-to-out ratio* R of the complexes. The scoring function for C'_i is defined as: $Score(C'_i) = \gamma(C'_i) * R(C'_i)$, where $\gamma(C'_i) = |E'_i|/|E'_{max_i}|$ and $R(C'_i) = |E'_{out_i}|/|E'_i|$. Here, E'_{max_i} is the set of all possible interactions between proteins in V'_i, that is, $|E'_{max_i}| = (|V'_i|) * (|V'_i| - 1)/2$, and E'_{out_i} is the set of all external interactions involving proteins in V'_i, that is, $E'_{out_i} = \{(p,q) : (p,q) \in E, p \in V'_i, q \notin V'_i\}$.

3. Results and discussions

We extended the MCL software [2] to implement MCL-CA using C/C++ combined with PL/SQL on a Pentium P4 Dual Core 3GHz 2GB RAM Linux machine. We used two high-quality protein-protein interaction datasets of yeast from Gavin *et al.* [5] and Krogan *et al.* [7][a]. The details of the datasets are shown in Table 1. As of now, biological experiments to show the core-attachment structure have been revealed only on yeast, so we do not know yet whether complexes of other organisms display similar structures.

Table 1. Details of datasets

Dataset	*Number of proteins*	*Number of interactions*	*Average number of interactions per protein*
Gavin, 2006	1430	6531	10.62
Krogan, 2006	2675	7080	6.98

For the evaluation of our predicted complexes, we used the manually curated yeast complexes from Wodaklab [10] as the benchmark. The Wodaklab CYC2008[b] catalogue contains 408 yeast complexes. Each of our predicted complexes was compared to the known complexes from the benchmark. For this, we used the accuracy measure as proposed in [1], given by $Acc = (N_c)^2/(N_p.N_t)$, where N_c is the number of common proteins shared by the predicted and known complexes, and N_p and N_t are the numbers of proteins in the predicted and known complexes, respectively.

[a] Gavin and Krogan download from GRID database [BioGRID version 2.0.33]: `http://www.thebiogrid.org/`

[b] Wodaklab Curated complexes 2008: `http://wodaklab.org/cyc2008/`

If the accuracy is greater than or equal to a threshold t, we assume the known complex has been found in our dataset. Sensitivity is the proportion of known complexes found for a given threshold. Finally, we used the co-annotations from Gene Ontology[c] to validate the functional similarities between proteins within complexes.

3.1. *Improvement over MCL*

We did a three-fold comparison between MCL and MCL-CA in terms of: (a) the total number of predicted complexes; (b) the number of predicted complexes that matched known complexes; (c) the accuracies of predicted complexes. Table 1 in supplementary materials compares the total number of complexes predicted by MCL and MCL-CA (with $inflation = 2.0$) from the two datasets. It shows that MCL produced many insignificant (noisy) clusters, which were discarded by MCL-CA in the filtering step. Table 2 shows the number of known complexes that were correctly predicted by MCL and MCL-CA for thresholds $t = \{0.6, 0.7, 0.8\}$. The number of known complexes covered by MCL-CA was significantly higher than that of MCL for all threshold values. There was no complex correctly predicted by MCL that was missed by MCL-CA. The difference in the numbers of known complexes covered by MCL and MCL-CA was larger for threshold $t = 0.6$ than for $t = \{0.7, 0.8\}$. For example, for the Gavin dataset, MCL-CA covered 14 more known complexes for $t = 0.6$, and 8 more known complexes for $t = 0.8$.

Table 2. No. of known complexes predicted by MCL and MCL-CA

No. of known complexes correctly predicted for $t = \{0.6, 0.7, 0.8\}$.

Dataset	$t = 0.6$		$t = 0.7$		$t = 0.8$	
	MCL	*MCL-CA*	*MCL*	*MCL-CA*	*MCL*	*MCL-CA*
Gavin	53	67	38	45	27	35
Krogan	81	100	48	65	36	50

In order to analyse the variation in increase in the number of known complexes covered by MCL-CA over MCL, we considered two sets of complexes predicted each from Gavin and Krogan datasets: (a) set $A = \text{MCL} \cap \text{MCL-CA}$ consisted of all complexes correctly predicted by both methods, but with different accuracies; (b) set $B = \text{MCL-CA} \setminus \text{MCL}$ consisted of all complexes correctly predicted by MCL-CA and not by MCL. The accuracy threshold was set to $t = 0.6$. For the Gavin dataset, $|A| = 16$ and $|B| = 14$, while for the Krogan dataset, $|A| = 23$ and $|B| = 19$. Table 3 shows the minimum, maximum and average increase in accuracies for MCL-CA over MCL for the complexes in sets A and B. The increase in accuracies for predicted complexes in A was noticably high, with the average being 16.01% and 15.29% for Gavin and Krogan datasets, respectively. The increase in accuracies for predicted complexes in B was significant, with the average being 53.02% and 51.34% for Gavin

[c] Gene Ontology: `http://geneontology.org`

and Krogan datasets, respectively. Tables 2 and 3 show that MCL-CA significantly improved the accuracies of many low quality complexes present in the datasets. As a result, these low quality complexes, which were difficult to be covered by only MCL, matched known complexes with better accuracies and therefore covered by MCL-CA.

Table 3. Overall increase in accuracy for MCL-CA over MCL for $t = 0.6$

$A = MCL \cap MCL\text{-}CA$; $B = MCL\text{-}CA \setminus MCL$.

	Increase in accuracy(%)					
	A			*B*		
Dataset	*Min*	*Max*	*Avg*	*Min*	*Max*	*Avg*
Gavin	5.0%	40.84%	16.01%	13.21%	103.03%	53.02%
Krogan	4.16%	63.93%	15.28%	11.11%	91.0%	51.34%

Table 2 in supplementary materials displays the improvement in accuracies for a sample of 10 complexes predicted from the Gavin dataset with threshold $t = 0.6$. The complexes in the upper half belonged to set A, while those in lower half belonged to set B. The improvement was noticably high for complexes in set A. For example, the Exocyst complex. The improvement was significant for complexes in set B. These were the low quality complexes that did not match any known complex with accuracy $\geq t = 0.6$ using only MCL. MCL had induced many additional (noise) proteins in these complexes. With the refinement due to MCL-CA, these matched known complexes with better accuracies. For example, the COPI complex.

Table 4. Comparisons between all four methods

No. of known complexes correctly predicted for $t = \{0.6, 0.7, 0.8\}$

Threshold t	*Dataset*	*MCL*	*MCL-CA*	*COACH*	*CORE*
0.6	Gavin	53	67	59	73
	Krogan	81	100	96	106
0.7	Gavin	38	45	43	48
	Krogan	48	65	59	89
0.8	Gavin	27	35	39	31
	Krogan	36	50	53	65

3.2. *Comparisons with CORE and COACH*

CORE [8] and COACH [9] are two very recently (2009) proposed methods that also make use of the core-attachment structure to detect complexes from yeast PPI networks. However, all the three methods (CORE, COACH and MCL-CA) are different from one another in the computational essense. While CORE calculates p-values between proteins within the entire network and builds cores from them, COACH finds dense subgraphs (preliminary cores) and adds attachments to them. COACH and MCL-CA are significantly faster computationally compared to CORE.

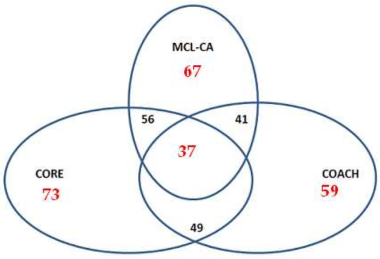

Fig. 1. Overlaps in known complexes covered by CORE, COACH and MCL-CA for $t = 0.6$ on the Gavin dataset.

On the same Gavin and Krogan datasets, CORE produced 295 and 819 complexes, while COACH ($filter = 0.225$) produced 326 and 345 complexes, respectively. We matched these complexes against the same benchmark. Since each method produced different number of complexes and with or without ranking, we considered all the produced clusters for the comparison. Table 4 summarizes this comparison. Though COACH and CORE performed better than MCL-CA at higher thresholds, MCL-CA was successful in substantially increasing the accuracies of mainly low quality complexes beyond the threshold $t = 0.6$. The ratios of correctly predicted complexes to total clusters produced were also higher for MCL-CA for $t = \{0.6, 0.7\}$. Figure 1 shows the overlaps in known complexes covered by the three methods for $t = 0.6$ on the Gavin dataset: CORE $\cap$ MCL-CA = 56, COACH $\cap$ MCL-CA = 41, CORE $\cap$ COACH = 49, and CORE $\cap$ COACH $\cap$ MCL-CA = 37. This shows among the three methods, there is none which totally covers all complexes predicted by another.

3.3. *Analysis of complexes predicted by MCL-CA*

We visualized the predicted complexes of MCL-CA using *Cytoscape* environment[d]. The supplementary materials contain visualizations and explanations for a sample of complexes that include the Golgi Transport complex (PubMed id: 11703943), Arp 2/3 Protein complex (PubMed id: 10377407) and Kornberg's Mediator complex (PubMed id: 15477388). Additionally, we found a set of module proteins that formed the constituents of RNA polymerase complexes I, II, III. We also discovered potential novel complexes that matched the novel complexes found by Gavin *et al.* Further we also did analysis of our predicted complexes to validate the core-attachment structures (all details in supplementary materials).

4. Conclusions and future work

Considering the immense popularity of the Markov Clustering (MCL) algorithm in clustering PPI networks, we have developed the MCL-CA approach to predict yeast

[d] *Cytoscape:* `http://www.cytoscape.org/`

protein complexes with better accuracies.

There is a lot of scope for further improvements and research in this direction. The proposed ranking method tends to favor larger complexes. Better statistical measures to rank complexes can be developed that reflect vital biological measures. Further, we have shown results only on unweighted PPI datasets. It will be worthwhile to see the results on weighted PPI datasets. Together with these weights, it will be interesting to develop new ranking measures for the complexes. We also intend to analyse the performance of MCL-CA using a larger benchmark set of manually curated complexes. Finally, these techniques are based on core-attachment structures found in yeast complexes. It will be interesting to check by computational means if complexes from other organisms display these structures even before wet lab experiments are performed.

Acknowledgements

This work was supported in part by the National University of Singapore under ARF grant R252-000-361-112. We would like to thank Henry Leung and Qian Xiang for providing the list of predicted clusters from CORE [8], Wu Min for providing the COACH software [9] and the reviewers for their valuable comments.

References

[1] Bader G.D., and Hogue C.W.V., An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*, 4, 2003.

[2] Dongen S., Graph clustering by flow simulation, *PhD thesis*, CIW, University of Utrecht, 2000.

[3] Enright A.J., Dongen S.,Ouzounis C.A., An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Research*, 30, 1575-1584, 2002.

[4] Frey B.J., Dueck D., Clustering by passing messages between data points. *Science*, 315(5814), 972-976, 2007.

[5] Gavin A.C., *et al.*, Proteome survey reveals modularity of the yeast cell machinery, *Nature*, 440, 631-636, 2006.

[6] Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer, 2001.

[7] Krogan N.J., *et al.*, Global landscape of protein complexes in yeast *Saccharomyces cerevisiae*, *Nature*, 440, 637-643, 2006.

[8] Leung H., Xiang Q., Yiu S.M., Chin F., Predicting protein complexes from PPI data: A core-attachment approach, *Journal of Computational Biology*, 16, 133-144, 2009.

[9] Min W., Li X., Kwoh C.K., Ng S.K., A core-attachment based method to detect protein complexes in PPI networks, *BMC Bioinformatics*, 10:169, 2009.

[10] Pu S., *et al.*, Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research.*, 18, 2008.

[11] Seals D.F., *et al.*, A Ypt/Rab effector complex containing the Sec1 homolog Vps33p is required for homotypic vacuole fusion, *Proc Natl Acad Sci*, 97(17), 9402-9407, 2000.

[12] Vlasblom J., *et al.*, Identifying functional modules in the physical interactome of Saccharomyces cerevisiae, *Proteomics*, 7, 944-960, 2007.

[13] Vlasblom J., Wodak S.J., Markov clustering versus affinity propagation for the partitioning of protein interaction graphs, *BMC Bioinformatics*, 10, 2009.