The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009


# Maximizing Modularity Density for Exploring Modular Organization of Protein Interaction Networks*

Shihua Zhang[1,†] Xuemei Ning[2] Chris Ding[3]

[1]Academy of Mathematics and Systems Science, CAS, Beijing 100190, China
[2]College of Science, Beijing Forestry University, Beijing 100083, China
[3]Department of Computer Science and Engineering, University of Texas at Arlington Arlington, TX 76019, USA

**Abstract** The purpose of this study is to introduce a new quantitative measure modularity density into the field of biomolecular networks and develop new algorithms for detecting functional modules in protein-protein interaction (PPI) networks. Specifically, we adopt the simulated annealing (SA) to maximize the modularity density and evaluate its efficiency on simulated networks. In order to address the computational complexity of SA procedure, we devise a spectral method for optimizing the index and apply it to a yeast PPI network. Our analysis of resulted modules suggests that most of these modules have well biological significance in context of protein complexes. Comparison with the MCL and the modularity based methods shows the efficiency of our method.

**Keywords** Modular organization; network clustering; modularity density; spectral method; protein interaction network

## 1 Introduction

Modularity has been considered to be one of the main organization principles of biological networks in the past decade years. Biological modules as a critical level of biological hierarchy and relatively independent units play special roles in biological systems [8]. How to uncover modular structures in various biological networks is a basic step for understanding cellular functions and organizational mechanisms of biosystems. For example, by using the network partition, Zhao *et al.* (2006) investigated the functional and evolutionary modularity of human metabolic network from a topological perspective.

One popular class of methods for dissecting modular structure in the field of general complex networks is based on optimizing a global quality function called modularity [2, 11] to partition the network into modules. And it has been comprehensively adopted to analyze biological networks [12, 4, 3, 5]. However, it has recently been shown that the resolution of the modularity based methods is intrinsically limited. It fails to find small

*This work is partially supported by the National Natural Science Foundation of China under grant No.60873205, No.10801131, Innovation Project of Chinese Academy of Sciences, kjcs-yw-s7 and the Ministry of Science and Technology, China, under Grant No.2006CB503905.
†zsh@amss.ac.cn

communities in large networks—instead, groups of small communities turn out merged as larger ones [13]. Li *et al.* (2008) proposed a novel quality function called modularity density ($D$) which aims to conquer the resolution limit problem in modularity. They have tested it on many kinds of small networks for illustration but not on large real networks.

In this study, we aim to introduce the new quantitative measure modularity density into the modular analysis of biomolecular networks and develop new algorithms for detecting functional modules in protein-protein interaction (PPI) networks. We first adopt the simulated annealing (SA) technique to maximize the modularity density and evaluate its advantages on a suit of simulated networks where the modules are known. In order to conquer the computational burden of SA procedure, we adopt a spectral $k$-means method for optimizing the measure and apply it to a yeast PPI network. Our biological analysis of resulted modules suggests that most of these modules carry distinguished biological significance. We also make a comparison of our method with other two methods including the popular MCL and modularity based methods to verify its effectiveness.

# 2 Materials and Methods

## 2.1 Definition of modularity and modularity density

The popular modularity $Q$ is defined by Newman and Girvan (2004). Briefly, when the nodes of a network are divided into modules, one can compute it as follows:

$$Q = \sum_{s=1}^{m} \left[ \frac{l_i}{L} - \left( \frac{d_i}{2L} \right)^2 \right],$$

where $m$ is the number of modules, $L$ is the total number of edges in the network, $l_i$ is the number of edges between nodes in module $i$, and $d_i$ is the total number of degrees of the nodes in module $i$. The highest $Q$ value of all possible module separations is called the network modularity. In the past studies, empirical and simulation studies showed that the network partition method of maximizing modularity $Q$ (MQ) has good performance. However, Fortunato and Barthélemy (2007) recently pointed out the serious resolution limits of this method, and claimed that the size of a detected module depends on the size of the whole network. The main reason is that the modularity $Q$ does not capture the information of the number of nodes in a module, and the choice of partition is highly sensitive to the total number of links in the network.

In the following, we introduce the so-called modularity density $D$ which was proposed as an alternative measure for describing the modular organization [14]. The characteristic of this measure is that it is related to the density of subgraphs. We first define the average modularity degree of subgraph $G_i(V_i, E_i)$ as follows:

$$ad(G_i) = aid(G_i) - aod(G_i) = \frac{2l_i - \overline{l_i}}{n_i},$$

where $aid(G_i)$ is the average inner degree of the subgraph $G_i$, which equals to twice the number of edges in subgraph $G_i$ divided by the number $n_i$ of nodes in this subgraph. $aod(G_i)$ is the average outer degree of the subgraph $G_i$, which equals to the number of edges with one node in the subgraph and the other node outside it divided by the number

$n_i$ of nodes in the subgraph. The intuitive idea is that $ad(G_i)$ should be as large as possible for a valid 'module'. Then the modularity density $D$ of a partition $G_1, \cdots, G_m$ is defined as the sum of all average modularity degree of $G_i$ for $i = 1, \cdots, m$. In contrast to $Q$, $D$ can be calculated as follows:

$$D = \sum_{i=1}^{m} ad(G_i) = \sum_{i=1}^{m} \frac{2l_i - \overline{l_i}}{n_i}. \tag{1}$$

This measure provides a way to determine if a certain mesoscopic description of the graph is accurate in terms of modules. The larger the value of $D$, the more accurate a partition is. So the community detection problem can be viewed as a problem of finding a partition of a network such that its modularity density $D$ is maximized. The search for optimal modularity density $D$ is a NP-hard problem due to the fact that the space of possible partitions grows faster than any power of system size.

Moreover, the phenomenon of multiple resolutions or/and hierarchy of modular structures have been observed in biological networks [16]. The modularity density $D$ can be extended for this more general case using a tuning parameter $\lambda$ as follows [14]:

$$D_\lambda = \sum_{i=1}^{m} \frac{2\lambda(2l_i) - 2(1-\lambda)\overline{l_i}}{n_i} \tag{2}$$

where $\lambda$ is a value ranging from 0 to 1, and when $\lambda = 0.5$, the $D_{0.5}$ corresponds to modularity density $D$. By varying $\lambda$, we can detect detailed and hierarchical organization of biological systems. In other words, we can divide the network into large modules and small modules using a small $\lambda$ and a large $\lambda$ respectively.

## 2.2 Simulated annealing for maximizing $D$ (MD)

In principle, the goal of a module detection is to find the 'optimal' partition with largest modularity $Q$ or modularity density $D$. Several methods have been proposed for optimizing $Q$. Most of them rely on heuristic procedures or approximate strategies. Here, we employ the simulated annealing (SA) technique to maximize $Q$ and $D$ to obtain the 'best' determination of the modules of a network for evaluating.

Simulated annealing is a kind of stochastic search technique for optimization problems. It enables one to find 'low cost' configurations without getting trapped in 'high cost' local minima and has many applications in combinatorial optimization problems. In the searching process, a global parameter $T$ representing temperature is introduced. When $T$ is high, the system can explore configurations of high cost while at low $T$ the system only explores low cost regions. Along with the decrease of $T$, 'low cost' configurations can be reached step by step by overcoming small cost barriers. When identifying modules, the objective is to maximize the quantitative indexes (i.e. $Q$ or $D$), thus, the cost is $C = -Q$ or $-D$. At each temperature, we perform a number of random updates and accept them with probability:

$$p = \begin{cases} 1, & \text{if } C_f \leq C_i \\ \exp\left(-\frac{C_f - C_i}{T}\right), & \text{if } C_f > C_i \end{cases} \tag{3}$$

where $C_i(C_f)$ is the cost before(after) the update.

Specific implementation detail can be seen in [12]. Note that we add a decision cause to ensure that each potential 'module' be connected. The one that performs best consists in isolating the module from the rest of the network, and performing a 'nested' SA, entirely independent of the 'global' one.

In using $Q$ and $D$ as 'fitness functions', the method is more direct than those relying on heuristic procedures. Moreover, SA enables us to carry out an exhaustive search and to minimize the problem of finding sub-optimal partitions. We should note that the SA method can't scale to very large networks, but it is an efficient evaluation method for its exhaustive characteristic. Several efficient methods for optimizing $Q$ have been proposed, but designing efficient algorithms for optimizing the new measure ($D$) is still an essential and challenging problem.

## 2.3 Spectral method for maximizing $D$ (SpeMD)

Given a network $G=(V,E)$, and denote its vertex set as $V$, edge set as $E$ and adjacent matrix as $A$. Given a $m$-partition $P_m$, define a corresponding $n\times m$ assignment matrix $X=[h_1,h_2,\cdots,h_m]$ with $h_{ic}=1$ if $v_i\in V_c$, and $h_{ic}=0$ otherwise, for $1\le c\le m$. Observe that since each vertex can only be in one cluster, $X1_m=1_n$. We can reformulate $D$ in terms of the assignment matrix $X$ as follows:

$$\begin{aligned} D &= \sum_{i=1}^{m}\frac{2l_i-\overline{l_i}}{n_i} \\ &= \sum_{i=1}^{m}\frac{h_i^TAh_i-(h_i^TBh_i-h_i^TAh_i)}{h_i^Th_i} \\ &= \sum_{i=1}^{m}\frac{2h_i^TAh_i-h_i^TBh_i}{h_i^Th_i} \\ &= \sum_{i=1}^{m}\frac{h_i^T(2A-B)h_i}{h_i^Th_i} \end{aligned} \tag{4}$$

where $B$ is the degree matrix. Let $\tilde{h}_i=\frac{h_i}{\|h_i\|}$, $\tilde{H}=[\tilde{h}_1,\tilde{h}_2,\cdots,\tilde{h}_m]$, note that $\tilde{h}_i^T\tilde{h}_j=\delta_{kl}$ or $\tilde{H}^T\tilde{H}=I$, then, we can obtain

$$\begin{aligned} D &= \sum_{i=1}^{m}\tilde{h}_i^T(2A-B)\tilde{h}_i \\ &= Tr\tilde{H}^T(2A-B)\tilde{H}. \end{aligned} \tag{5}$$

So the problem of maximizing $D$ can then be expressed as:

$$\begin{aligned} \max\ & D = Tr\tilde{H}^T(2A-B)\tilde{H} \\ \text{s.t.}\ & \tilde{H}^T\tilde{H}=I \end{aligned} \tag{6}$$

From the standard result in linear algebra, the optimal $\tilde{H}$ of the above trace maximization has close relationship with the leading $k$ eigenvectors of $2A-B$ by relaxing $\tilde{H}$ as an arbitrary orthonormal matrix. We can adopt the corresponding spectral algorithms and use the leading $k$ eigenvectors of $2A-B$ to optimize the modularity density $D$. To obtain the final network partition, we apply the $k$-means clustering method to cluster eigenvectors. Importantly, the same principle can be derived for $D_\lambda$.

### 2.4 The procedure of the algorithm

Given an upper bound $K$ of the number of modules and the adjacent matrix $A = (a_{ij})_{n\times n}$ of a network. The procedure of the algorithm is stated straightforward as follows:

- Spectral mapping:
  1. Compute the diagonal matrix $B = (d_{ii})$, where $d_{ii} = \sum_k a_{ik}$.
  2. Form the eigenvector matrix $U_K = [\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_K}]$, corresponding to the $K$ largest eigenvalues of $2A - B$.
- $k$-means: for each value of $k$, $2 \leq k \leq K$
  1. Form the matrix $U_k = [\mathbf{u_2}, \mathbf{u_3}, \cdots, \mathbf{u_k}]$ from the matrix $U_K$.
  2. Normalize the rows of $U_k$ to unit length using Euclidean distance norm: $\frac{u_{ij}}{\sqrt{\sum_j u_{ij}^2}}$.
  3. Treat the rows of $U_k$ as points in $R^k$ and cluster them into $k$ clusters using $k$-means or even other clustering methods.
- Maximizing modularity density $D$ or $D_\lambda$ with given $\lambda$: Pick the $k$ and the corresponding partition $P_k$ that maximizes $D$ or $D_\lambda$.

We should note that this type of spectral clustering technique has been successfully applied to general clustering problems as well as graph clustering problems [10, 17]. Here, we explore the characteristic of modularity density $D$, and derive a new spectral clustering based method for maximize $D$ ($D_\lambda$) (SpeMD). And the SpeMD procedure described here can be seen as a particular manner of employing the standard $k$-means algorithm on the elements of the leading $k$ eigenvectors to extract $k$ clusters simultaneously.

Convergence and computational complexity of the SpeMD procedure are key problems when this method is applied to large complex networks. Fortunately, several strategies can be employed to improve these problems. First, we can initialize the $k$-means such that the starting centroids are chosen to be as orthogonal as possible [18]. This strategy does not change the time complexity, but can improve the quality of convergence, thus at the same time reduce the need for restarting the random initialization process. Second, several fast techniques for solving eigen system have been developed and several methods of $k$-means acceleration can also be found in the literature. Based on this type of techniques, for large sparse networks with $m \sim n$, and $k \ll n$, the SpeMD procedure will scale roughly linearly as a function of the number of nodes $n$ [17]. Here we didn't consider these ameliorative techniques and only focus on the validity of the SpeMD method.

### 2.5 Performance measures

All performance measures can be seen in the extended version.

## 3 Results

In this section, we apply the present method to a suit of simulated networks and a yeast PPI network to test its efficiency. We first present detailed numerical results to show the difference of network partition determined by maximizing the modularity density $D$ and modularity $Q$ with simulated annealing (SA) technique. In general, maximizing $D$ (MD) can give more detailed and valid results, while maximizing $Q$ (MQ) encounters serious resolution limit in simulated networks.

Then we apply the new spectral method for maximizing the generalized $D_\lambda$ (SpeMD) to a yeast PPI network to identify functional modules which show significant biological relevance. Comparison with MQ and MCL, we show that the SpeMD can obtain competitive performance with the well-known MCL method and resolve much finer modular structure than MQ method. To extract appropriate modules, the SpeMD and MCL both rely on one parameter. Here, we perform the SpeMD and MCL with adjusted parameters to obtain the 'best' geometric accuracy and separation. For SpeMD, we tune $\lambda$ from 0.4 to 0.7 in step of 0.05, and for MCL, we sample inflation parameter values from 1.5 to 2 in steps of 0.1.

## 3.1 Simulated networks

First we do the comprehensive tests on a group of simulated networks which take on significant modular characteristics. In the work of [14], $D$-based method has been showed to be able to obtain competitive performance with $Q$-based method. However, the size of artificial networks generated by using Newman's popular procedure as well as its variant are too small to show the serious resolution limit problem of $Q$. Therefore, we devise a new type of artificial networks. The network is composed of $2m$ cliques ($m$ $n_1$-clique and $m$ $n_2$-clique), and external edges are placed randomly with a fixed expectation values so as to keep the average edge connections $k_{out}$ of each node to nodes of other cliques. So each network has $m(n_1+n_2)$ nodes and about $m(n_1(n_1-1)/2+n_2(n_2-1)/2)+m(n_1+n_2)k_{out}/2$ edges. In the following test, we let $n_1 = 10$ and $n_2 = 15$. Note that we can also relax cliques as dense modules for testing, but here we just show the clique case for convenience.

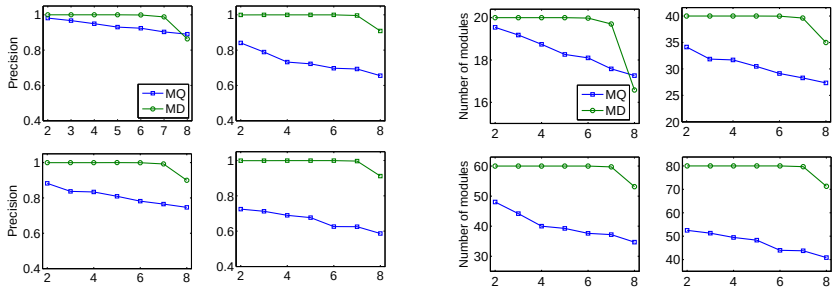


Figure 1: Left figure: Comparative test of MD and MQ on simulated networks with known community structures. It is a plot of the fraction of nodes correctly classified with respect to $k_{out}$. Each point is an average over 50 realizations of the networks. Right figure: Number of modules detected by MD and MQ with the real number of cliques (NC), averaged over 50 network realizations.

The computational results for this experiment are summarized in Fig. 1, where NC is the number of cliques, i.e., $NC = 2m$. The left plot of Fig. 1 shows the fraction of nodes that are correctly classified into the communities (Precision) with respect to $k_{out}$ by MD and MQ respectively. We can see that MD method based on $D$-value performs

much better than MQ method under all the different NC. For instance, for 50 random networks with $NC = 60$ and $k_{out} = 5$, on an average 99.97% nodes are classified correctly by MD, while only about 72.23% nodes by the MQ. When $k_{out} = 8$ which indicates the corresponding networks are difficult to be partitioned, MD still has very high accuracy (>86%).

The most interesting observation is that performance of MD is almost the same, while that of MQ is greatly decreasing with the increase of NC (also the size of networks). For example, for 50 random networks with $k_{out} = 6$, always on an average >99.9% nodes are classified correctly by MD on four different sizes of networks with $NC = 20$, 40, 60, 80, while about 92.40%, 78.18%, 69.75%, 62.59% nodes by the MQ respectively. This fact shows the serious resolution limit problem of modularity $Q$, while that can not be observed on the small networks such as the simulated networks using Newman's method.

To test the performance of MD and MQ in selecting the number of communities, we calculate the number of modules. The right plot of Fig. 1 shows the averaged number of modules on four different sizes of networks ($NC = 20$, 40, 60, 80) with respect to $k_{out}$ by MD and MQ respectively. We can see that MD performs much better than MQ. The MD can almost always identify the right number of modules in four different sizes of networks with $k_{out} \leq 7$. While MQ can not do that. For example, for 50 random networks with $NC = 60$ and $k_{out} = 7$, on an average 59.7 modules are identified by MD, while only about 37.20 modules by the MQ. For the harder case ($k_{out} = 8$), MD can sill do much better than MQ. Actually, even for the easiest case $k_{out} = 2$, MQ can not identify the right modules with 52.50 modules for NC=80. This uncovers the underlying resolution limit just as pointed in [13]. In summary, the MD can recover the underlying community structure more often than the MQ by a sizable margin in the simulated modular networks. The modularity density $D$ more relies on local connectivity of a network and can uncover finer modular structure. While modularity $Q$ more relies on size and total links of a network and can lead to serious resolution limit. Moreover, the limit is more serious as size of networks increasing.

## 3.2 Results on a PPI network

The budding yeast *S. cerevisiae* PPI network was obtained from the DIP database (http://dip.doe-mbi.ucla.edu/dip/), which contains human-curated high-throughput and small-scale binary interactions directly observed in experiments, as well as binary interactions inferred from high-confidence protein complex data. We only considered non-self physical interactions and built the PPI network. The giant component of the PPI network is composed of 2559 proteins linked by 7031 nonredundant interactions. In order to test the ability of SpeMD to extract complexes from the interaction network and compare it with other two methods, we compared the detected modules to known complexes in yeast as annotated by the Munich Information Center for Protein Sequences (MIPS) [15] using the $P_{ol}$ formula.

We apply the SpeMD method to the yeast PPI network to detect functional modules. Totally, we obtain 279 protein modules of sizes from 4 to 38 with $\lambda = 0.6$ (To extract statistically and biologically significant modules, we remove 48 modules with size $\leq$ 3). A complete list of complexes and modules with functional annotation is provided in Supplementary Files. Figure 2 presents three such modules. For example, the second one

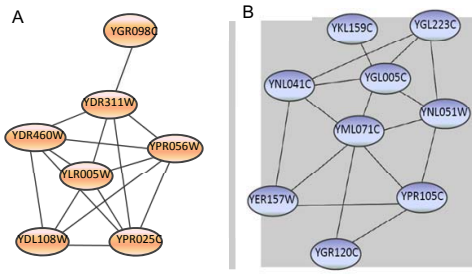


Figure 2: Examples of modules which match the MIPS complexes with great significance. (A) A seven-member module matches with the SSL2-core TFIIH complex when it is part of the nucleotide-excision repair factor 3 (NEF3) with ($P_{ol} = 10^{-14.81}$). (B) A nine-member module matches with Golgi transport complex which stimulates intra-Golgi transport and is composed of eight proteins ($P_{ol} = 10^{-21.7}$).

is a nine-member module which matches with Golgi transport complex for stimulating intra-Golgi transport with $P_{ol} = 10^{-21.7}$.

## 3.3 Comparison with MCL and MQ

There has been many methods for detecting network modules. The comparison of all the methods are not an easy task. Here, we attempt to compare the MD (SpeMD) with two types of classical methods: MQ and MCL. Just as we have mentioned, the modularity ($Q$) maximization based module-detection method has been comprehensively applied in many fields including analysis of biological networks. Another method is the Markov Cluster algorithm (MCL) which was developed by van Dongen [6]. The method simulates a flow on the network by calculating successive powers of the network adjacency matrix. In each iteration, an inflation step is applied to enhance the contrast between regions of strong or weak flow in the network. The process converges towards a partition of the network, with a set of high-flow regions separated by boundaries with no flow. The value of the *inflation parameter* strongly influences the the size and number of the resulted modules.

The module size distribution of detected modules for each method on the PPI network have been shown in the left plot of Figure 3. The SpeMD and MCL both identify about (279 and 242) modules without extremely large clusters. The major trend generated by MD and MCL are both similar to that of the complexes in MIPS database, which suggest the definition of modularity density is reasonable (Note that the MIPS complex is a combination of hand-curated and experimental complexes. They have some overlap, so complex curve is higher. But the trend is similar). Unfortunately, the module size distribution of MQ is very different from the previous ones. The MQ only detect 21 modules with relative large size raging from 39 to 263. As tested on the simulated networks, the MQ method is highly limited by the resolution problem.

As to biological significance, the accuracy and separation are used for evaluating the correspondence between complexes and modules from each methods [1]. From the right plot of Figure 3, we can easily see that the SpeMD and MCL have consistently better performance than MQ. This means the modularity density based partition method can

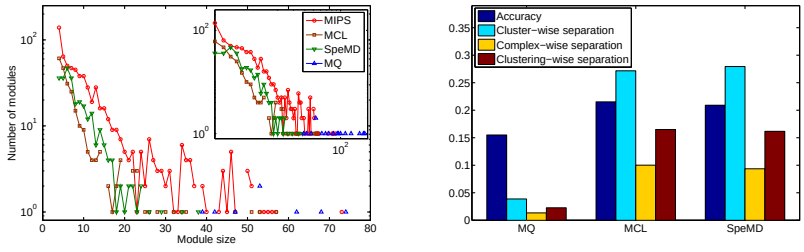


Figure 3: Left figure: Module size distribution of different methods and MIPS protein complex with size $> 3$. Right figure: Performance of different methods on the PPI network. Four different measures including 'Accuracy', 'Cluster-wise separation', 'Complex-wise separation, and 'Clustering-wise separation' have been used.

produce more biologically significant modules than the modularity based method. And the new quality function may become an evaluation index of modularity organization of networks. While MCL has no such evaluation function.

# 4 Discussion and Conclusion

In summary, our method is very effective for uncovering modular organization in biomolecular networks. It provides an objective approach to explore the organization and interactions of biological processes. With the increasing amount of biological 'interaction' data available, MD (SpeMD) can facilitate the construction of a more complete view of the composition and interconnection of functional modules and the understanding of the organization of the whole cellular at system level. We plan to automate this algorithm to compute functional modules for a large number of biological networks. We hope that related studies will benefit from the present method coupled with the modularity density $D$.

# References

[1] Broheé, S. and van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7, 488.

[2] Newman, M.E. and Girvan, M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 026113.

[3] Wang, Z. and Zhang, J. (2007) In search of the biological significance of modular structures in protein networks. *PLoS Comput. Biol.* 3, e107.

[4] Caretta-Cartozo, C., De Los Rios, P., Piazza, F., *et al.* (2007) Bottleneck Genes and Community Structure in the Cell Cycle Network of S. pombe. *PLoS Comput. Biol.* 3(6), e103.

[5] Zhao, J., Yu, H., Luo, J.H., Cao, Z.W., Li, Y.X., (2006) Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics* 7, 386.

[6] van Dongen, S. Graph clustering by flow simulation. In PhD thesis Centers for mathematics and computer science (CWI), University of Utrecht; 2000.

[7] Friedel, C.C., Krumsiek, J., Zimmer, R. (2008) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *RECOMB 2008*, 4955, 3-16.

[8] Barabasi, A., Oltvai, Z. (2004) Network biology: understanding the cell's functional organization. *Nature Rev. Gen.* 2004, **5**, 101-113.

[9] J.F. Kelen and T. Hutcheson. Reducing the time complexity of the fuzzy *c*-means algorithm. *IEEE Transactions on Fuzzy Systems*, **10(2)**, 263-267 (2002).

[10] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. *SIAM International Conference on Data Mining*, (2005).

[11] Newman, M.E.J. (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci., USA* **103**, 8577-582.

[12] Guimer, R., Amaral, L.A.N., (2005) Functional cartography of complex metabolic networks. *Nature*, **438**, 895-900.

[13] Fortunato, S., Barthélemy, M. (2007) Resolution limit in community detection. *Proc. Natl. Acad. Sci., USA* **104**, 36-41.

[14] Li, Z., Zhang, S., Wang, R.S., Zhang, X.S., and Chen, L. (2008) *Physical Review E* 77, 036109.

[15] Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31-34.

[16] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555.

[17] Zhang S., Wang R.S. and Zhang X.S. (2007) Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A*, 374, 483-490

[18] Ng, A., Jordan, M. and Weiss, Y. (2002) On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Systems* 14, 849-856.