Published in IET Systems Biology
Received on 16th July 2010
Revised on 5th November 2011
doi: 10.1049/iet-syb.2010.0066

ISSN 1751-8849

Identifying similar functional modules by a new hybrid spectral clustering method

S. Madani[1] *K. Faez*[2] *M. Aminghafari*[3]

[1]*Department of Computer Science, Amirkabir University of Technology, Tehran, Iran*
[2]*Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran*
[3]*Department of Statistics, Amirkabir University of Technology, Tehran, Iran*
E-mail: kfaez@aut.ac.ir

Abstract: Recently, a large number of researches have focused on finding cellular modules within protein–protein interaction networks. Until now, most of the works have concentrated on finding small modules and protein complexes. The authors have extended the concept of functional module and have identified larger functional modules which are the most similar to the entire network. To this end, a new hybrid spectral-based method is proposed here. First, the original graph is transformed into a line graph. Next, the nodes of the new graph are represented in the Euclidean space by using spectral methods and finally, a self-organising map is applied to the points in the new feature space. The experimental results show that similar modules, obtained from the proposed method, have own local hubs and lots of significant functional subunits concerning each other. These modules not only detect general biological processes that each protein is involved in, but also due to great similarities to the original network, it can be used as significant subnetworks for predicting protein function as detailed as possible. Some interesting properties of these modules are also investigated in this research.

1 Introduction

Understanding the large-scale organisation of protein interactions is one of the most important goals of systems biology. Since proteins rarely act as single isolated components in cells, cellular functions are probably to be carried out in a modular framework [1]. Nowadays, many studies have been conducted to find functional units in the protein–protein interaction (PPI) networks. Since identifying these modules provides a great view of what happens in cells, it has an advantage of annotating uncharacterised proteins within each module over the direct function assignment methods.

So far, several methods have been introduced to identify these cellular modules in PPI networks [2–8]. Enright *et al.* [3] applied the Markov clustering (MCL) algorithm to find protein families. This algorithm simulates random walks within a graph by the expansion and inflation operations. As [9] shows, it is one of the best approaches among the related works. In another work, MCODE was proposed by Bader and Hogue [10]. This algorithm has three steps: vertex weighting by using the density of its local neighbourhood, prediction of molecular complexes and an optional post-processing step. Spirin and Mirny [4] found dense subgraphs in PPI networks by three approaches and categorised them into two main groups: functional modules and protein complexes. The first algorithm was to identify all the cliques by complete enumeration. The second approach was based on super paramagnetic clustering and the third one was a Monte-Carlo algorithm which maximises the density of each cluster. In King *et al.* [5], proposed the restricted neighbourhood search clustering. This local search algorithm partitions the node set of the network based on cost function. In another work, Dunn *et al.* used Girvan and Newman's edge-betweenness algorithm to discover protein complexes [6, 11, 12]. They iteratively removed the edges with the highest betweenness value until achieving the desired clusters. Later, Chen and Yuan [7] modified the edge-betweenness criterion to make the original algorithm more robust against unbalanced partitioning.

There are also a few spectral-based approaches to cluster the PPI networks. For example, in Bu *et al.* [13], calculated the spectrum of the adjacency matrix corresponding to the PPI network. Then, they selected the top 10% of proteins in an eigenvector after sorting by the absolute value. Finally, by defining some constraints 48 quasi-cliques and 6 quasi-bipartites were found. In another work, Sen *et al.* [14] identified clusters by an eigenmode analysis of the connectivity matrix which is the same as the negation of unnormalised Laplacian matrix. They found proteins corresponding to the most significant components in each eigenvector as members of that eigencluster.

To our knowledge, previous studies have mostly found protein complexes in comparison to functional modules. Perhaps because the protein complexes form at the same time and place in real world and, hence, they can be found by experimental purification methods but functional modules do not exist as a complex at a single time point. But exactly for the same reason, computational methods can

be more worthwhile to detect these modules which are hard to be found in experimental manners. On the other hand, all the studies which detect functional modules concentrated on nearly small ones. Hence, most of these functional modules are also protein complex and there is a very little difference between them.

In this paper, we extend the concept of functional module and identify larger modules, which are the most similar to the whole PPI network and preserve some important network properties such as degree distribution. For this purpose, we developed a new hybrid spectral-based method for detecting these modules in PPI networks. First, we transformed the original PPI network into line graph. After that, we extracted the most significant features for each node of graph by using spectral analysis. Finally, a self-organising map (SOM) was applied to cluster points in the new feature space.

We applied our approach to a high-quality dataset of *Saccharomyces cerevisiae* PPI network produced by Krogan group. The results indicate that our method has found significant overlapping modules which have some interesting properties and for this reason, they are suitable to predict protein function by more complicated methods in the future. In fact, these modules, which are general biological processes where each protein involved is in, provide great knowledge about topology of functional area and indirect partners of each protein and, thus, from now on, they can be used to give an answer to some important questions about proteins such as their functions. This paper is organised as follows: The next section presents our method. In Section 3 the results are derived. The significant functions of modules and their local hubs are discussed in Section 4, and Section 5 concludes the paper.

2 Method

So far, several clustering algorithms have been applied to detect protein complexes and functional modules in PPI networks. Although both complexes and modules have more interactions among their members than with the rest of the network, they are biologically different [4]. As noted in the introduction, most researches concentrated on detecting either protein complexes or tiny functional modules. Here, we use network topology to find larger functional modules which have interesting attributes. In the following, the biological and computational points are clarified separately.

2.1 Biological points

The molecular modules including protein complexes and functional modules were investigated by Spirin and Mirny in 2003. They identified some of these cellular modules and demonstrated some of their properties and differences. The size of their proposed modules ranged from 5 to 25 [4]. Up to now, several articles have tried to detect these small modules and even in some cases, the results have been verified by comparing cluster sizes [7]. As Spirin and Mirny mentioned, functional modules consist of proteins which involve in the same cellular process while binding to each other at a different time and place. But by this definition, the functional modules can be small or even very large because cellular processes are arranged in a hierarchy. Indeed, the functional module concept depends on the corresponding cellular process. If we consider more general cellular processes, we will earn larger modules.

Let us take a proper look at the hierarchy of the modules in a cell where at the lowest level, each protein is a module itself with its own functions and at the highest level, the whole network could be considered as a module whose main duty is the cell survival. The present work attempts to find the functional modules in a more general level. But an important issue is how much the modules could be larger? This research is aimed at probing into this hierarchy from top to bottom to find that whether or not there are modules which have still the structural similarity to the entire network despite more specific functions than the PPI network. The reason why we pursue in this direction is that the assumption of self-similarity has already been verified in many fields of research. In addition, such similar modules can be very valuable because they can code the structural information of the protein network into smaller subnetworks.

Recently, a few interesting researches have investigated the self-similarity and fractality in complex networks by using the box-covering method [15–18]. In this approach, the network is covered by distinct boxes of a given maximum diameter l_β. Then, a renormalisation transformation is applied for finite times in way each box is replaced by a node and the nodes will be connected if there is at least one edge among the corresponding boxes. Lastly, fractality is perused by calculating d_β as the fractal dimension by $N_\beta(l_\beta) \sim l_\beta^{-d_\beta}$ where $N_\beta(l_\beta)$ is the minimum number of needed boxes for l_β. Song *et al.* [15] showed that real complex networks such as PPI networks are self-similar under different length scales and their degree distribution remains invariant during alternative renormalisation processes. As Gallos *et al.* [16] mentioned the inverse of the renormalisation could be seen as the time evolution of the PPI network. Indeed, this kind of similarity depends on the duration of time and it means that the network preserves its degree distribution during the process of growing. Gallos *et al.* at the end of their paper [16] pointed towards the relationship between fractality and modularity, and mentioned that the boxes can be considered as different functional modules. However, this claim still has not been confirmed.

In this paper, we would like to investigate the similarity in a PPI network. To achieve this goal, we apply a hybrid spectral-based method to discover similar overlapping modules which preserve some important properties of the PPI network. The features considered here are the degree distribution of clusters, the density of clusters, the diameter of clusters and the average shortest path within clusters. It should be noted that at first sight, there is no relationship between our similar modules and the boxes provided by box-covering method. In self-similarity context, there is a similarity between the PPI networks at different times; and the boxes produced by the box-covering method are only blocks for network reduction. Moreover, these boxes have only one property; that is, the distance between any pair of their nodes is less than l_β, and there are not any other significant properties. Instead, our desired modules are the most similar subgraphs in four important topological properties to the whole PPI network. In spite of these differences, these two topics could be complementary. In other words, a PPI network preserves not only its degree distribution during the time but also includes the similar functional modules and this demonstrates the complexity of the similarity in these networks. In the next section, we describe the computational method used for finding these interesting modules.

2.2 Computational points

Based on the argument above, we tend to discover the most similar modules to the entire network. To do this, a

doi: 10.1049/iet-syb.2010.0066

spectral-based method was applied to cluster nodes according to the significant features extracted for each node. The most important point which should be noted is that our goal of clustering node set of graph is not finding most dense connected regions, but rather we only use an efficient clustering method to find our desired modules. In this way, the resulting modules would have more interactions among their own proteins and fairly few interactions to the proteins in the rest of the network addition to the properties above. To have this property is useful because it is known that the existence of interaction can be an encouraging sign of functional dependency.

Lately, spectral-based approaches have become one of the most popular clustering algorithms. These approaches have not only a rich mathematical background but also very often work well in practice [19]. They use eigenvectors of different Laplacian matrices to determine clusters and in general, consist of recursive and multiway methods [19–22]. Experimentally, it has been observed that the multiway spectral clustering algorithms are a little better than the recursive one [20, 21]. Thus, here, a multiway spectral-based approach is used to partition graph nodes to k almost large clusters which are similar to the entire graph. But the main problem is how to choose the appropriate number of clusters. Generally, estimating the number of clusters is a challenge in clustering context and this problem is more complicated in our special case because we look for some clusters with specific properties.

To solve this problem, we cluster graph nodes to k groups for k from 2 to a sufficient upper bound. As each cluster is a graph itself, the similarity of each cluster to the entire network is evaluated in terms of some topological properties. The assumed similarity criteria here are degree distribution of clusters, density of clusters, diameter of clusters and average shortest path within clusters. Finally, we select k, the optimal number of clusters, for which the clusters are the most balanced and remarkably similar to the PPI network in these features.

Since a protein can participate in several cellular processes, it can belong to more than one cluster. Therefore first of all, we transform the original graph into a line graph to create overlap. A primary graph and its line graph are shown in Fig. 1, which extracted from [23]. In graph theory, given an undirected graph $G = (V, E)$, its line graph $L(G)$ is a graph where each node represents an edge of G that is $V(L(G)) = E(G)$ and two nodes are adjacent if and only if their corresponding edges have a common endpoint in G; that is, $E(L(G)) = \{((u, v), (v, w))|(u, v), (v, w) \in E(G)\}$. We weight each edge $((u, v), (v, w))$ in the line graph $L(G)$ by averaging the weight of the edges (u, v) and (v, w) in the original graph. As Pereira–Leal mentioned, the line graph has some advantages over the initial graph for clustering purpose. For example, its clustering coefficient is much higher and the structure of original network can be completely recovered from it [23]. Hence, this preprocessing step is efficient for clustering approaches and because of these good properties, this transformation has been previously applied to PPI networks [23, 24].

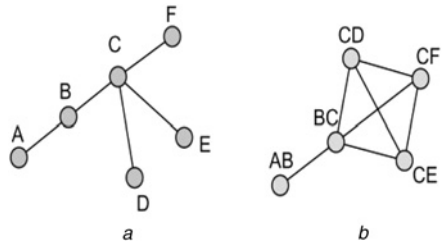

Fig. 1 *Original graph and the line graph*

a Original graph
b Line graph corresponding to the original graph

After transformation, the Ng, Jordan and Weiss (NJW) algorithm with some modifications is applied to cluster nodes of new graph (for details about NJW algorithm see [21]). First, we used the normalised graph Laplacian $\boldsymbol{L}_{\mathrm{rw}} = \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}$ instead of $\boldsymbol{L}_{\mathrm{sym}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$ where $\boldsymbol{D}$ is a diagonal matrix with diagonal entries $d_{ii} = \sum_{j=1}^{n} w_{ij}$. As $\boldsymbol{L}_{\mathrm{rw}}$ is closely related to a random walk, the eigenvectors of normalised Laplacian matrix $\boldsymbol{l}_{\mathrm{rw}}$ has been used to extract the significant features for each node in this research. Secondly, we solved the generalised eigen problem $\boldsymbol{L}x = \lambda \boldsymbol{D}x$ to find the top k eigenvectors of $\boldsymbol{L}_{\mathrm{rw}}$. It should be noted that the weight of each edge is a figure between 0 and 1. Thus, as Verma and Meila [20] mentioned, if some entities of $\boldsymbol{D}$ are very small, it will be more effective not to compute $\boldsymbol{D}^{-1}$ and generally solving the generalised eigen problem is numerically more stable. Therefore our method is more stable than the NJW algorithm. For the same reason, each method which uses $\boldsymbol{D}^{-1}$ can be unstable in similar cases. Thirdly, as we were going to apply our algorithm to the largest connected component of network, the eigenvector corresponding to the smallest eigenvalue, that is, u_1 was not used in the new feature vector. Since first eigenvector for connected graphs is a constant vector, it does not have any information about clusters. But by dividing each feature vector by its norm, the first eigenvector will not be constant anymore and in this way, norm of feature vectors again has effect on clustering procedure.

Moreover, here, a SOM network [25] with k neurons was applied to cluster points in the new feature space in place of k-means algorithm. In spite of similarity between these two clustering approaches, SOM has some advantages over k-means. For example, it preserves the close topological neighbourhoods of the data vectors. This advantage provides a suitable biological view about the functional modules and in this way, we can find out the relations between these modules. For initialising the weights of SOM, the most significant two principal components were provided by using principal components analysis (PCA) and the initial weights were distributed across the space spanned by principal components of the inputs. That way, SOM network starts with a reasonable initial configuration and in conclusion, the learning phase can be conducted faster [26].

Since clustering of the line graph yields the clusters of interactions, the resulting clusters in the line graph are transformed back to the clusters in the original PPI network. Eventually, these final clusters will be used for all the further analyses. It should be noted that our goal is to discover the functional modules which are the most similar to the whole PPI network. Hence, we used four well-known criteria which are the degree distribution of clusters, the density of clusters, the diameter of clusters and the average shortest path within clusters to calculate similarity each cluster to the entire network. Our hybrid method can be described as follows.

2.2.1 Algorithm for detecting the most similar modules to PPI network:

- Transform original graph into its line graph with adjacency matrix $\boldsymbol{W}$ such that $0 \leq w_{ij} \leq 1$.

• For $k = 2$ to $K_{\max}$
1. Let the degree matrix $\boldsymbol{D}$ be a diagonal matrix with the entries

$$d_{ii} = \sum_{j=1}^{n} w_{ij}$$

2. Compute the unnormalised Laplacian $\boldsymbol{L}$ as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.
3. Compute the k smallest generalised eigenvalues of $\boldsymbol{L}_{\text{rw}}$ by solving the generalised eigen problem $\boldsymbol{L}x = \lambda \boldsymbol{D}x$.
4. Form $\boldsymbol{U} = [u_2, \ldots, u_k] \in \mathbb{R}^{n \times k}$ with the vectors $u_2, \ldots, u_k$ as its columns where u_i is the eigenvector corresponding to the ith smallest eigenvalue.
5. Construct the matrix V from U by normalising each row of U to have a unit norm (i.e. $V_{ij} = U_{ij}/(\sum_{j=2}^{k} U_{ij}^2)$).
6. Let each row of V be a point in the feature space $\mathbb{R}^k$. Cluster points with SOM whose weight vectors are initialised by PCA method.
7. Assign node i to cluster j if and only if the ith row of the matrix V was assigned to cluster j.
8. Transform back the obtained clusters of the line graph.
9. Estimate the scaling parameter α for the power law degree distribution $p(x) \sim x^{-\alpha}$ of the PPI network and the modules using a maximum likelihood based method proposed in [27].
10. Calculate the density of each cluster C with vertex set $V_C = \{v_1, \ldots, v_n\}$ by

$$\text{density}(C) = \frac{\sum_{v_i, v_j \in V_c} w_{ij}}{\binom{n}{2}}$$

11. Calculate the diameter of each cluster C.
12. Calculate the average shortest path within each cluster C.
• Analyse the similarity of the kth bunch of clusters by using mean, standard deviation, min and max of each property.
• Select k for which the k clusters are the most balanced and remarkably similar subgraphs to the PPI network.

3 Experimental results

In this section, we report on experiments and discuss the results in detail. Since it is well known that there is much noise among protein interactions, here, we used core dataset of Krogan group published in Nature [28] to reduce the noise by weighting the interactions. This dataset is available in [29]. This *Saccharomyces cerevisiae* PPI network which consists of 7123 interaction pairs involving 2708 proteins is an undirected weighted graph with proteins as nodes and interactions as edges and the weights assigned to the edges, show edge reliability. The downloaded network has 63 connected components such that the largest connected component consists of 2559 nodes and 7031 edges and the other components are tiny subgraphs so that the average of their number of nodes is equal to 2.4032. For the sake of increasing accuracy, our algorithm was applied to the largest connected component. After transforming the original network into its line graph, it became an undirected weighted graph with 7031 nodes and 107 230 edges.

It is common for spectral clustering approaches to estimate the number of clusters by a so-called criterion eigengap heuristic [19]. First of all, we tried to determine the number of clusters by this criterion. Fig. 2 shows the eigenvalue distribution of unnormalised and normalised Laplacian of the transformed graph. It is easy to see there is no clear gap between eigenvalues and, therefore the number of clusters cannot be determined by using this eigengap heuristic. In other words, it is not easily possible to observe the modular structure in our network. In conclusion, we look for our desired modules by considering the steps 9–12 in our algorithm. Notice that the SOMs were trained by a batch train algorithm. The neurons were arranged in a hexagonal grid model. The initial neighbourhood radius and the number of iterations were set to 3 and 4000, respectively. The Euclidean distance was utilised in the training phase of SOMs, because Nadler *et al.* [30] showed that the Euclidean distance between the new feature vectors is the same as diffusion distance in random walks.

As noted in Section 2.2, we considered several criteria such as the degree distribution of clusters, the density of clusters, the diameter of clusters and the average shortest path within clusters to find the most similar clusters to the whole PPI

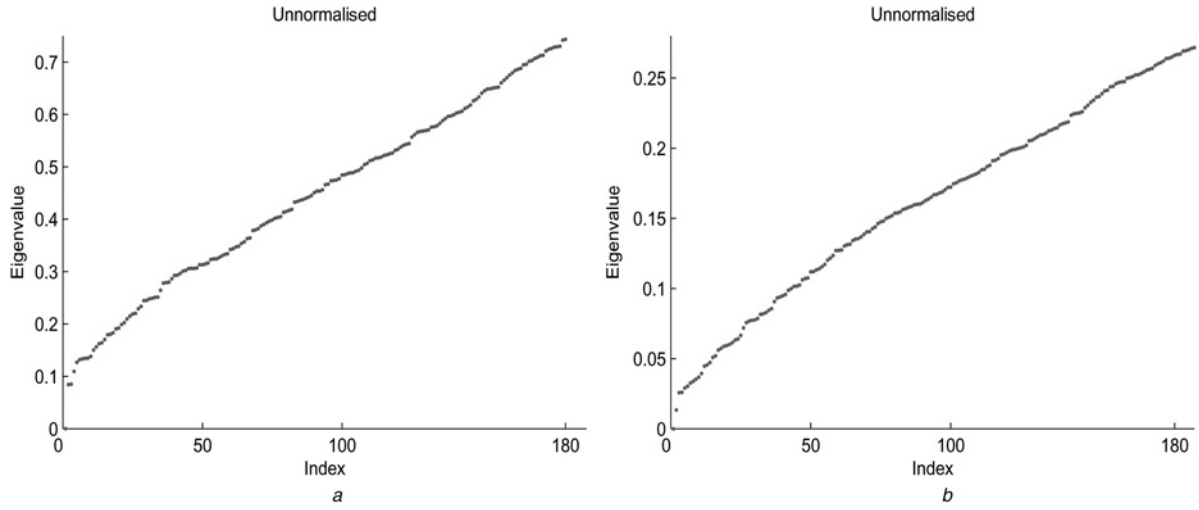

Fig. 2 *Eigenvalue distribution of Laplacian matrix*
a Eigenvalue distribution of unnormalised Laplacian matrix
b Eigenvalue distribution of normalised Laplacian matrix

doi: 10.1049/iet-syb.2010.0066

network. Fig. 3 shows how these criteria vary with k, the number of clusters. Here, the desirable k is one for which the k clusters have the following properties:

- The average values of the four properties should be as close as possible to corresponding values of the entire network.
- The clusters should have a balanced view of our selected properties. In other words, the variance in those properties should be small. Also, the box plot for the selected k should not have any outliers. It is because each outlier indicates a cluster which has a significant difference with the rest of clusters in that specific property.

It is trivial that as the number of clusters increases, the clusters get smaller and denser and their similarity to the entire network is gradually reduced. In other words, the lesser the number of clusters, the more would be similarity in clusters. Consequently, we look for k that satisfies the constraints above and additionally to be large enough.

As it can be seen from Fig. 3, by increasing k, the mean of each property becomes far from the network corresponding value and the variance of that property increases. In addition, the outliers are presented for larger values of k. Based on these box plots for $k = 12$, there is no outlier in the plots, the mean values are still remarkably close to the PPI network values and the spread of values are acceptable. Hence, these 12 clusters are well balanced in our desirable criteria, in addition to great similarity to the value of the PPI network, and as you can see there does not exist any k greater than 12 so that it satisfies this constraints properly. Consequently, $k = 12$ was picked as the most appropriate number of clusters. In Tables 1 and 2, the values of selected criteria for discovered clusters and the PPI network are summarised, respectively. By comparing these tables to each other, it can be concluded that the means of all four properties are remarkably close to the network corresponding values.

The degree distributions of all 12 clusters are shown in Fig. 4. As we want to compare the degree distribution of modules with each other and with that of the whole network fairly, the minimum node degree in each cluster was considered as $x_{\min}$ and, then, the power-law model

Table 1 Clusters characteristics

Feature	Min	Average	Max
clusters size	186	349.8333	619
clusters degree exponent	1.64	1.7008	1.81
clusters density	0.0045	0.0101	0.0172
clusters diameter	4.873	6.7013	8.356
average shortest paths within clusters	2.0045	2.3718	2.904

Table 2 PPI network characteristics

Feature	Value
degree distribution	$P(x) \sim x^{-1.61}$
density	0.0015
diameter	7.172
average shortest paths	2.4537

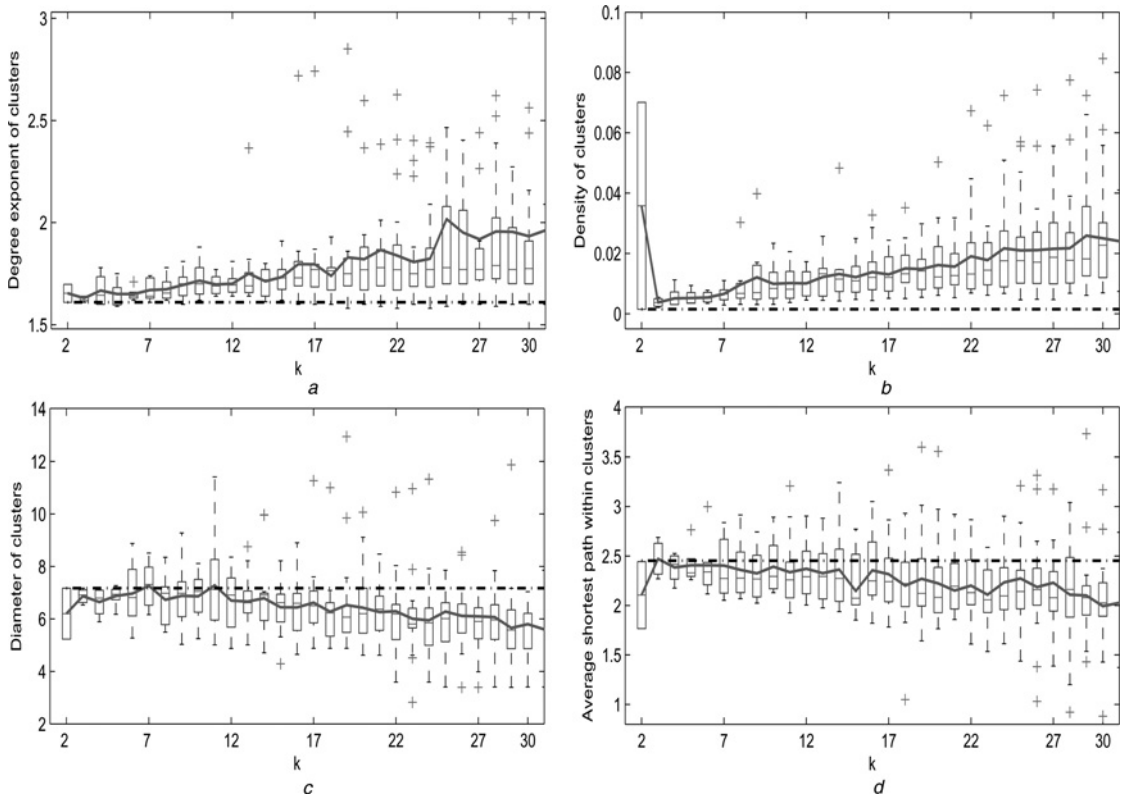

Fig. 3 *In each box plot, the green line shows the average value of the clusters and dash line shows the value of network*

a Defined score for clusters against k
b Density of clusters against k
c Diameter of clusters against k
d Average shortest path within clusters against k

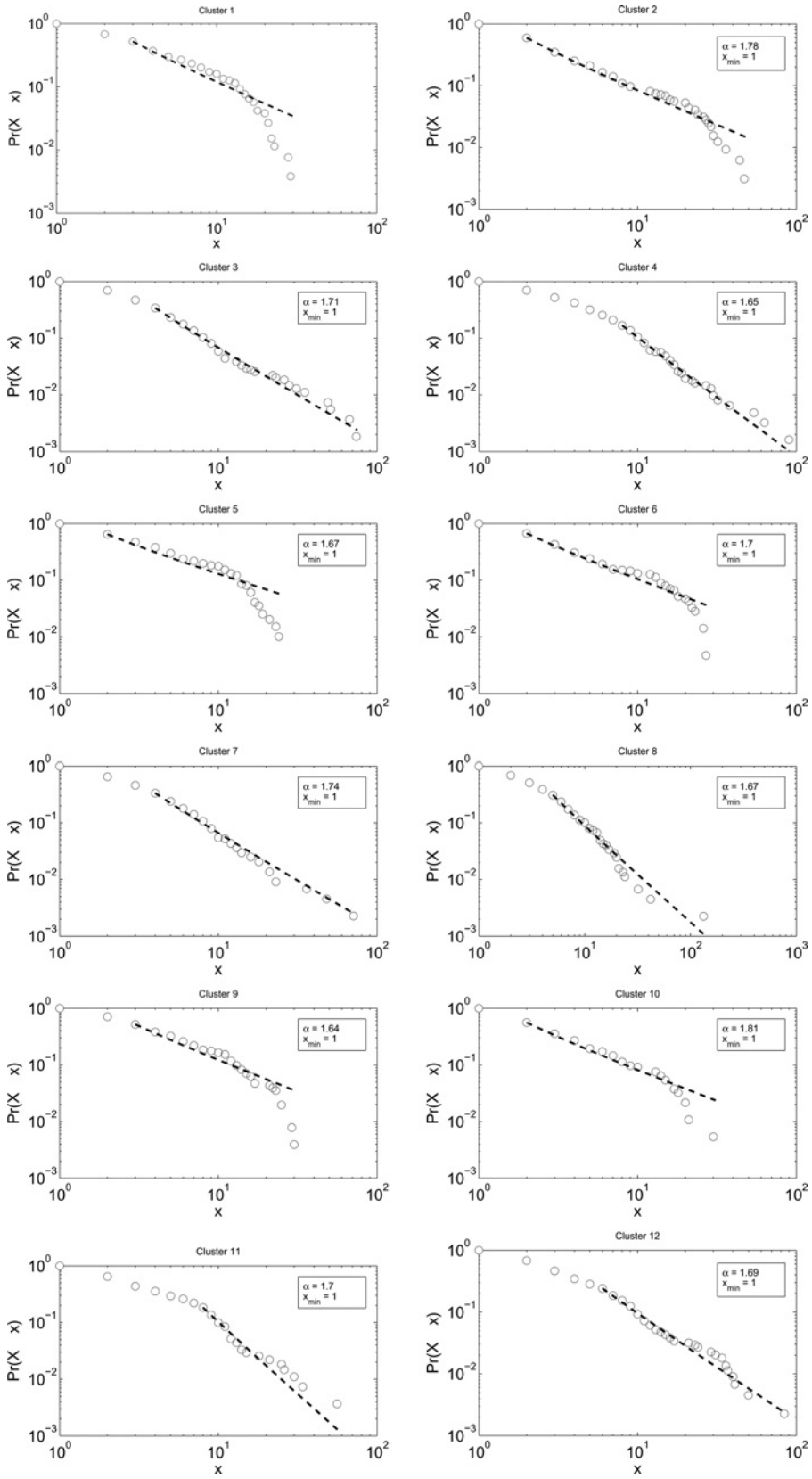

Fig. 4 *Degree distribution of obtained clusters*

$p(x) \sim x^{-\alpha}$ was fitted to the degree distribution of the networks to estimate α using Clauset *et al.* method [27].

To further confirm the results, we found local hub of each module, which was considered as a node with the highest degree inside subgraph corresponding to that module. Let us assume that for each node v, $d_{\text{inner}}(v)$, namely the inner degree, is degree of node v in subgraph corresponding to the module it belongs to, and $d_{\text{outer}}(v)$, namely the outer degree, is degree of node v in the whole network. In this research, we calculated the inner and the outer degree of each local hub. Varying minimum and average outer degrees for local hubs against k is shown in Fig. 5.

doi: 10.1049/iet-syb.2010.0066

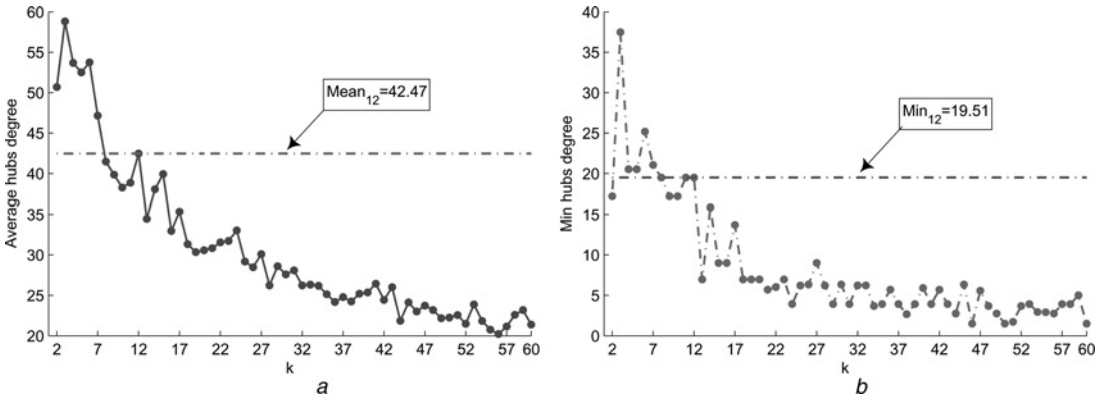

Fig. 5 *Average and minimum outer degree for local hubs against k*

a Average outer degree for local hubs against *k*
b Minimum outer degree for local hubs against *k*

As it can be inferred from Fig. 5, $k = 12$ has the largest average and minimum degree for $k \geq 8$. In other words, for $k = 12$ the vertices which have higher degree in the PPI network were chosen as local hubs. We found that although vertices with very high outer degree, for example 84, have been placed in more than one cluster, they play local hub role only in one cluster and just make a connection between clusters they belong to. This matter can be seen from Fig. 6. It is also interesting that for each local hub except one, the inner and outer degrees are very close to each other. It means that each local hub performs main part of its interactions in its own module.

We evaluated the significance of biological descriptions inside each module by a known statistical test. Assuming hypergeometric distribution for a given module, the probability of observing m or more proteins which have the same annotation out of n proteins is

$$p\text{- value} = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where n is the size of the cluster, N is total number of proteins in the database and M is the number of proteins in a specific functional category. In this paper, protein functions have been extracted from the hierarchical scheme of MIPS [31]. We chose MIPS classification scheme instead of Gene Ontology (GO) as suggested in [31, 32]. As Mewes *et al.* mentioned, FunCat protein function classification has the highest predictive power of all 16 features which have been tested (for details see [32, 33]). In addition, FunCat can have stronger view about functions of each module because it arranged the functions hierarchically. For each cluster some of the most significant functions with corresponding p-values are listed in Table 3 (see supplementary tables). It should be noted that we did not limit ourselves to the functions of a specific level of MIPS functional hierarchy (e.g. second-level or third-level functions) and reported that the functions were remarkably significant. This limitation on the level of function has already been placed in many previous works.

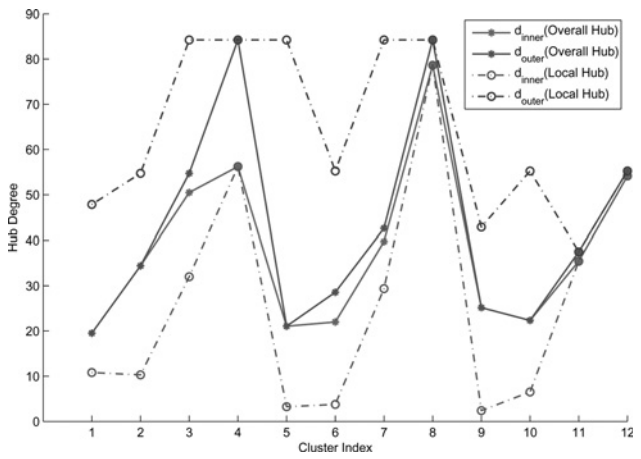

Fig. 6 *Inner and outer degree for local and overall hubs*

However, it seems that some biological experiments could be done to deeply understand what these functional modules do in cells, here, this matter is investigated by analysing function of proteins which have the highest degree in each cluster because in our idea, they can clarify attitude of their clusters and detailed information about cellular function of obtained modules can be inferred from their analysis. To do this, the inner degree of vertices of each cluster was plotted (see Fig. 7) and vertices, whose degree was remarkably high in comparison to others, were selected to be analysed more. We used swissprot database and SGD to extract some information about these selected proteins. Some of the important points about the modules and their selected proteins will be discussed in the next section.

4 Biological discussion

Some biological assessments based on the selected proteins and significant annotations of our modules are discussed in this section. First, the selected proteins of higher inner degree and their functions will be introduced per each module and, then, the interaction among these proteins will be shown in Fig. 8. For more details see supplementary materials as the selected proteins of the modules along with their biological processes are listed in Table 4 and the final biological results are summarised in Table 3.

4.1 Module number 1

For this module, six proteins out of nine selected proteins have GO term nuclear mRNA splicing via spliceosome and, consequently, more general terms RNA splicing and mRNA processing. The proteins YLR275W, YLR147C, YPR182W

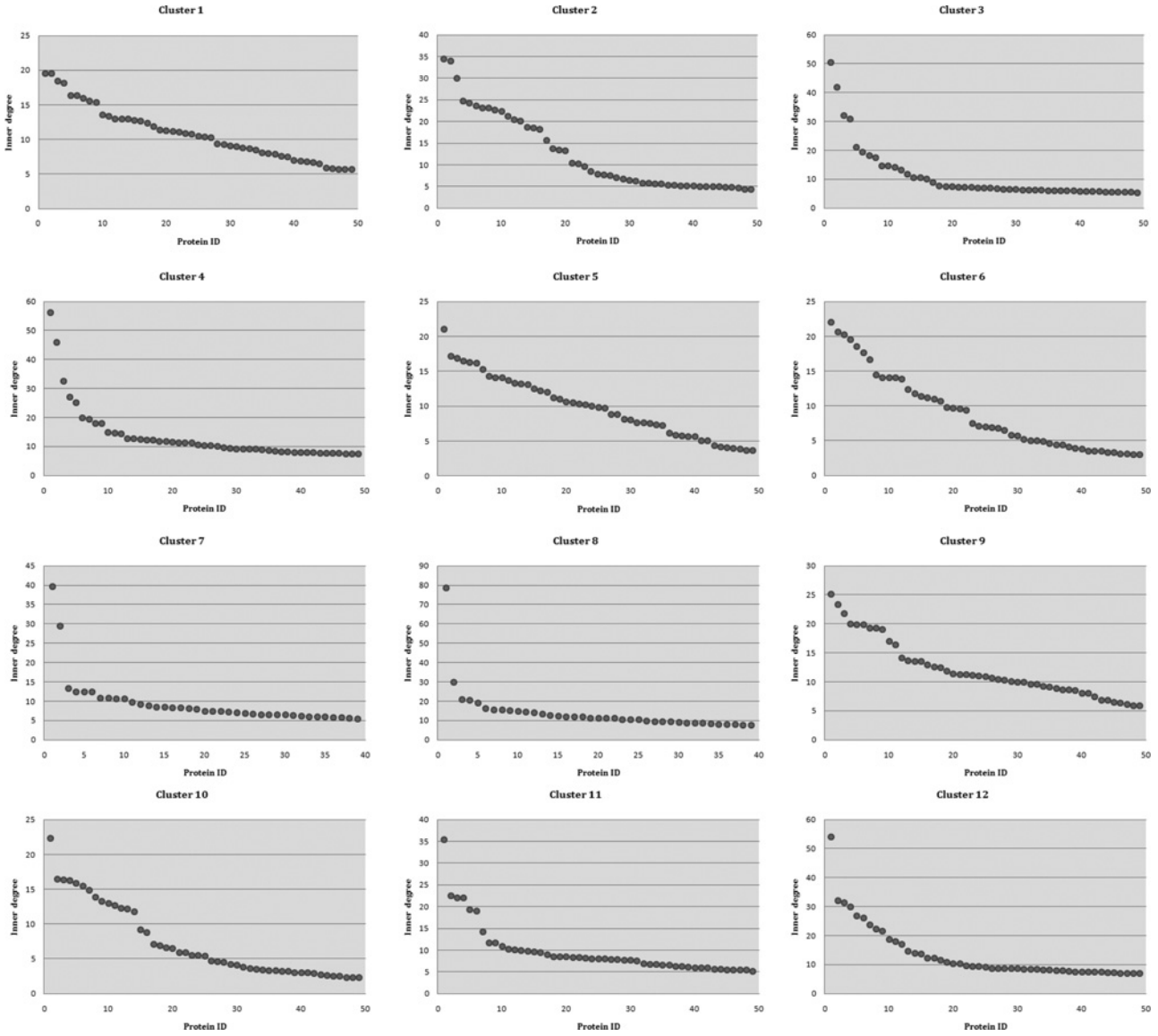

Fig. 7 *Higher degrees of each cluster*

and YGR074W are components of the Sm core complex and present in spliceosomal snRNP U1, U2, U4/U6 and U5. They involve in pre-mRNA splicing. The proteins YHR165C and YKL012W are required for pre-spliceosome formation, which is the first step of pre-mRNA splicing. The proteins YHR165C and YKL012W are associated with snRNP U5 and snRNP U1, respectively. Furthermore, the protein YGL173C which forms a complex with importin alpha subunit (YNL189W) is a multifunctional protein that exhibits several independent functions at different levels of the cellular processes. It is interesting that this protein places in several clusters obtained from our method (the clusters number 1, 4, 7 and 9). In conclusion, the proteins which belong to this module participate in RNA splicing (1.22e-67) and mRNA processing (1.17e-59).

4.2 Module number 2

All the selected proteins except YFR010W are placed in the proteasome complex. The proteins RPN3 acts as a regulatory subunit of the 26S proteasome which is involved in the ATP-dependent degradation of ubiquitinated proteins. The protein YDR363W-A is a subunit of the 26S proteasome which plays a role in ubiquitin-dependent proteolysis. The protein YEL037C plays a central role both in proteasomal degradation of misfolded proteins and DNA repair and it is the central component of a complex required to couple deglycosylation and proteasome-mediated degradation of misfolded proteins in the endoplasmic reticulun that are retrotranslocated in the cytosol. The protein YGL048C is involved in the ATP-dependent degradation of ubiquitinated proteins. The protein YHR200W is a multiubiquitin binding protein. As a result, the general annotations of this module are proteolysis (9.24e-22), cellular protein catabolic process (2.93e-21) and protein complex subunit organisation (1.10e-08).

4.3 Module number 3

Although all the selected proteins of the third cluster have a cellular localisation annotation, they are multifunctional and belong to the several clusters. So it is difficult to assign a few definite functions to this cluster but it seems that this cluster contains the structural proteins. As this sort of proteins often regulate another processes, it is completely logical to participate in different processes, especially

doi: 10.1049/iet-syb.2010.0066

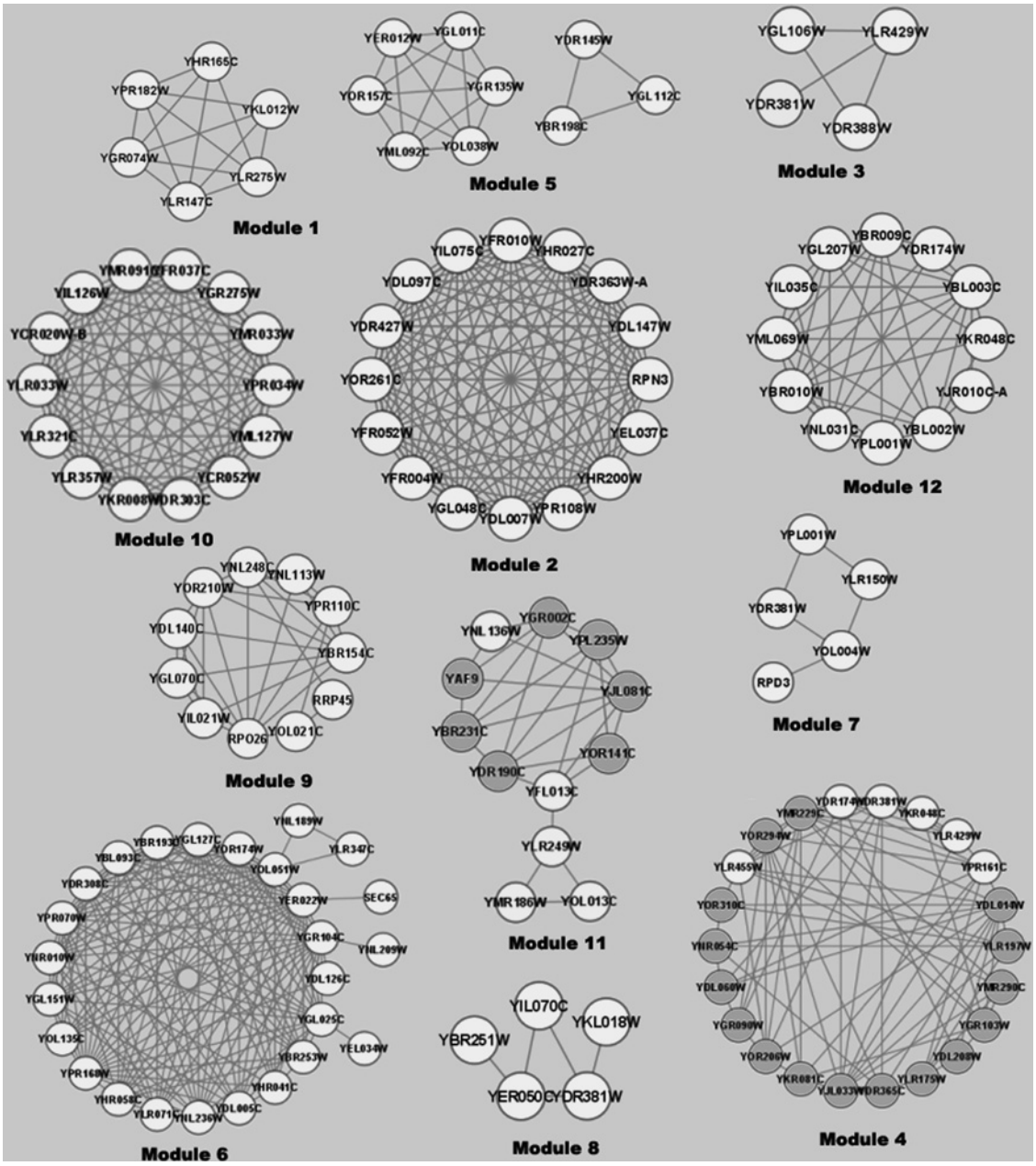

Fig. 8 *The interaction among the selected proteins of higher inner degree for each module*

cellular component organisation or biogenesis (4.09e-20) and regulation of biological processes (3.91e-19) like cellular or metabolic processes.

4.4 Module number 4

All the proteins displayed in Fig. 8 for module 4 except YLR455W, YDR174W, YDR381W, YKR048C, YLR429W and YPR161C have ribosome biogenesis annotation. In general, this module is responsible for the concerning processes like ncRNA processing (6.30e-33), ribosome biogenesis (1.26e-32), rRNA metabolic process (7.18e-30) and rRNA processing (1.14e-27).

4.5 Module number 5

The proteins YGL112C, YER012W and YDR145W are components of the DNA-binding general transcription factor complex TFIID and the transcription regulatory histone acetylation (HAT) complexes SAGA, SALSA and SLIK. The proteins YGR135W, YGL011C, YER012W, YOR157C, YML092C and YOL038W form the proteasome core complex. The local hub of this cluster that is YGL112C is an important protein which exists in 13 protein complexes. This module can be annotated by the terms proteasomal ubiquitin-independent protein catabolic process (4.41e-20), histone acetylation (2.60e-13), protein acetylation (9.88e-12), transcriptional preinitiation complex assembly (6.68e-12), RNA polymerase II transcriptional preinitiation complex assembly (1.57e-11) and protein–DNA complex assembly (8.84e-08).

4.6 Module number 6

All the 26 first proteins except the proteins YNL189W, YDL126C and YLR347C, form the RNA polymerase II mediator complex. The local hub protein, that is, YNL189W mainly is localised at the periphery of the nucleus and controls protein import into nucleus and protein targeting to membrane. All the selected proteins are

annotated to the cellular component disassembly. The transcription (4.91e-08) especially the transcription from RNA polymerase II promoter (2.79e-07) and nucleic acid metabolic process (9.21e-10) can be assigned to this module.

4.7 Module number 7

The functions chromatin silencing at rDNA, chromatin silencing at silent mating-type cassette, chromatin silencing at telomere and chromatin organisation during transcription are the common functions of the selected proteins in this module. Many annotations of this module are related to the regulation such as regulation of biological process (5.27e-16), biological regulation (1.22e-15), regulation of cellular process (1.31e-15), regulation of cellular biosynthetic process (1.32e-13) and regulation of metabolic process (8.98e-13). This module participates in the cell cycle, DNA processing and protein biosynthesis too.

4.8 Module number 8

In this module, the proteins YIL070C, YER050C and YBR251W are placed in the mitochondrion. This cluster has the significant annotations such as mitochondrial translation (4.48e-31), mitochondrion organisation (1.35e-15), gene expression (2.67e-20), protein synthesis (1.29e-33), ribosome biogenesis (4.63e-25), translation (8.79e-13) and translational initiation (3.40e-13).

4.9 Module number 9

All the selected proteins except YOL021C are components of the RNA polymerase I (Pol I), RNA polymerase II (Pol II) or RNA polymerase III (Pol III) complexes. This module have the annotations of rRNA synthesis (1.05e-17), tRNA synthesis (2.49e-21), protein with binding function or cofactor requirement (4.52e-14) and transcription from RNA polymerase I, II and III promoters.

4.10 Module number 10

All the selected proteins are a component of the chromatin structure remodelling complex (RSC), which is involved in transcription regulation and nucleosome positioning. As you know, RSC complex plays different kinds of roles in many processes. So we could assign the functions like chromosome organisation (7.24e-14), primary metabolic process (2.47e-11) and response to stress (3.21e-11) to this category. The interesting point that should be noted is that the proteins which exist in RSC were classified in this module and the most significant functions here are chromosome organisation (6.04e-20), chromatin remodelling (3.98e-09), chromatin assembly or disassembly, and so on.

4.11 Module number 11

The proteins YAF9, YBR231C, YDR190C, YOR141C, YJL081C, YPL235W and YGR002C have a chromatin remodeling annotation. The local hub protein of this module, that is, YMR186W is a molecular chaperone that promotes the maturation, structural maintenance and proper regulation of specific target proteins involved in cell cycle control and signal transduction such as CNA2. This module involves DNA conformation modification (4.94e-10), nuclear and chromosomal cycle (2.63e-05), chromosome segregation/division (4.49e-05), organisation of chromosome structure (1.61e-06), chromatin remodelling (3.98e-09) and protein folding and stabilisation (5.18e-05).

4.12 Module number 12

In this cluster, the proteins YBL003C, YBL002W, YBR009C and YBR010W are a histone and placed in the nuclear nucleosome and other proteins interact with these histones. In our view, instead of assigning a function to this cluster, it is better to say it contains the chromatin-associated proteins. The proteins YBL003C, YGL207W, YBL002W, YBR009C, YML069W and YBR010W are placed in the replication fork protection complex and the proteins YGL207W and YML069W are also placed in the FACT complex.

Finally, an analysis of the biological processes that selected proteins involved in (Table 4) and significant functions of each module (Table 3) showed a strong relationship among them and that way, we could predict general functions of each module that are summarised in the fourth column of Table 3. In this research, first we found 12 functional modules which have wonderful topological properties and then found proteins or protein complexes that each functional module runs under their leadership. The results show that our modules often have one or more protein complexes as an intra-module hubs and it seems that the protein with the highest inner degree in each module that is the local hub defined above, is likely to play more important role in the protein complex it places in.

Lately, Cho and Zhang [34] also paid attention to the hierarchical structures hidden in the PPI networks. They identified structural hubs and functional modules by converting the original PPI network into a hub-oriented tree structure, but their research is markedly different from ours. On the one hand, they first found the structural hubs and finally generated functional modules from the tree structure whereas we first found functional modules and then detected the inner hubs. On the other hand, they concentrated on detecting the structural hubs which are the main proteins to support the hierarchical structure of a PPI network and mainly bridge different functional modules whereas we used inner hubs which are the same as intra-module hubs to predict general functions of our modules. Furthermore, after biological analysis of the highest degree proteins of each module we extended inner hubs until protein complexes in some modules if necessary. It means a protein complex can start or control a large process in a cell. The most important issue which was not addressed in [34] is how to guess proper threshold to extract structural hubs and the corresponding modules. In fact, an important question that should be answered is what are the significant functional modules in the hierarchy they mentioned?

As discussed in this paper, our overlapping functional modules are complex networks which are very similar to the whole PPI network and preserve the degree distribution of network. Thus, from a statistical viewpoint, they can be used for predicting protein functions due to their great similarities to the original network. Currently, there exists two general ways for protein function prediction via analysis of PPI networks. In global methods, the entire topology of the network is used for assigning function to each protein [35–38]. In local methods, first some small neighbourhoods such as 1-neighbourhood, 2-neighbourhood or protein complexes are detected and then a significant annotation has to be assigned to each module by simple ways such as finding the most common annotation [3, 5,

doi: 10.1049/iet-syb.2010.0066

10]. But these approaches have some disadvantages. In the former, function of each protein can affect others and, in this way, some errors occur in the computations because in an organism all the proteins do not affect each other. In the latter, as King *et al.* [5] mentioned, many known complexes show low functional homogeneity. Apart from this, there is also lots of noise in these networks and, hence, in practice, we can hardly discover significant groups in detailed levels. On the other hand, most module detection methods identify a noticeable number of cluster with 1–5 members and due to the lack of enough function, annotating them is an additional problem too.

The obtained clusters here are significant subgraphs which can be used for assigning function to their members instead of the whole network or the previous neighbourhoods. From now on, the function of a protein can be predicted more carefully by using the topology of the cluster protein belongs to, and other biological information such as gene expression profile, 3D structure similarity, phylogenetic trees, and so on. Additionally, as our modules are the protein networks themselves, they can be used for different kinds of analyses like finding local hubs, the protein complexes, the motifs, and so on. It should be noted that the scope of this paper is limited to not using other biological data in addition to the PPI network during the process of clustering. Thus in this paper, larger significant neighbourhoods were identified by clustering the PPI network without using any other biological information and the main focus of future work will be in investigating which kind of biological data can improve the clustering results.

5 Conclusion

In this paper, we gave a new idea of extracting similar meaningful patterns in a PPI network. To do this, a new hybrid spectral-based method was developed. The method combined the line graph transformation, spectral mapping and SOM to identify the most balanced and similar clusters to the PPI network. First, the original graph is transformed into its line graph to create overlap. Next, the nodes of new graph were represented in the Euclidean space and finally, SOM was applied to the points in the new feature space. In this research, the similarity of the clusters to the entire network was evaluated in terms of some important topological features like degree distribution and density to choose the number of clusters appropriately. We have observed that the similar functional modules obtained not only have lots of significant biological subunits concerning each other but also have proteins or protein complexes as local hubs which control their cellular process. In fact, these modules are general biological processes where the proteins can involve in, and provide great knowledge about topology of functional area and indirect partners of each protein. Thus, these modules which are large enough neighbourhoods with significant biological functions could be used for further detailed analyses about their members. It should also be noted that because of feature reduction step, this method is so fast and can be applied to large complex networks. We hope these clusters will be used widely for predicting protein function later.

6 Acknowledgment

The authors are very thankful to Zeinab Bagheri for her worthwhile biological comments and Dr Masoud Pourmahdian for his great assistance. We would also like to thank Dr Maria Costanzo for her valuable guides.

7 References

1 Barabasi, A.L., Oltvai, Z.N.: 'Network biology: understanding the cell's functional organization', *Nat. Rev. Genet.*, 2004, **5**, (2), pp. 101–113
2 Sharan, R., Ulitsky, I., Shamir, R.: 'Network-based prediction of protein function', *Mol. Syst. Biol.*, 2007, **3**, (88), p. 113
3 Enright, A.J., Van Dongen, S.V., Ouzounis, C.A.: 'An efficient algorithm for large-scale detection of protein families', *Nucl. Acids Res.*, 2002, **30**, (7), pp. 1575–1584
4 Spirin, V., Mirny, L.A.: 'Protein complexes and functional modules in molecular networks', *Proc. Natl. Acad. Sci. USA*, 2003, **100**, (21), pp. 12123–12128
5 King, A.D., Przulj, N., Jurisica, I.: 'Protein complex prediction via cost-based clustering', *Bioinformatics*, 2004, **20**, (17), pp. 3013–3020
6 Dunn, R., Dudbridge, F., Sanderson, C.M.: 'The use of edge-betweenness clustering to investigate biological function in protein interaction networks', *BMC Bioinf.*, 2005, **6**, (39)
7 Chen, J., Yuan, B.: 'Detecting functional modules in the yeast protein–protein interaction network', *Bioinformatics*, 2006, **22**, (18), pp. 2283–2290
8 Asur, S., Ucar, D., Parthasarathy, S.: 'An ensemble framework for clustering protein–protein interaction networks', *Bioinformatics*, 2007, **23**, (13), pp. i29–i40
9 Brohee, S., van Helden, J.: 'Evaluation of clustering algorithms for protein–protein interaction networks', *BMC Bioinf.*, 2006, **7**, (488), pp. 1–19
10 Bader, G.D., Hogue, C.W.: 'An automated method for finding molecular complexes in large protein interaction networks', *BMC Bioinf.*, 2003, **4**, (2), pp. 1–27
11 Girvan, M., Newman, M.E.J.: 'Community structure in social and biological networks', *Proc. Natl. Acad. Sci. USA*, 2002, **99**, (12), pp. 7821–7826
12 Newman, M.E.J., Girvan, M.: 'Finding and evaluating community structure in networks', *Phys. Rev. E.*, 2004, **69**, (2), pp. 56–68
13 Bu, D., Zhao, Y., Cai, L., *et al.*: 'Topological structure analysis of the protein-protein interaction network in budding yeast', *Nucl. Acids Res.*, 2003, **31**, (9), pp. 2443–2450
14 Sen, T.Z., Kloczkowski, A., Jernigan, R.L.: 'Functional clustering of yeast proteins from the protein-protein interaction network', *BMC Bioinf.*, 2006, **7**, pp. 355–367
15 Song, C., Havlin, S., Makse, H.A.: 'Self-similarity of complex networks', *Nature*, 2005, **433**, (7024), pp. 392–395
16 Gallos, L.K., Song, C., Makse, H.A.: 'A review of fractality and self-similarity in complex networks', *Phys. A: Stat. Mech. Appl.*, 2007, **386**, (2), pp. 686–691
17 Kim, J.S., Goh, K.-I., Kahng, B., Kim, D.: 'Fractality and self-similarity in scale-free networks', *New J. Phys.*, 2007, **9**, (6), p. 177
18 Kim, J.S., Goh, K.-I., Kahng, B.: 'A box-covering algorithm for fractal scaling in scale-free network', *Chaos*, 2007, **17**, (2)
19 Luxburg, U.V.: 'A tutorial on spectral clustering', *Stat. Comput.*, 2007, **17**, (4), pp. 395–416
20 Verma, D., Meila, M.: 'A comparison of spectral clustering algorithms'. Technical report, University of Washington, 2003
21 Ng, A., Jordan, M., Weiss, Y.: 'On spectral clustering: analysis and an algorithm', *Adv. Neural Inf. Process. Syst.*, 2002, **14**, pp. 849–856
22 Shi, J., Malik, J.: 'Normalized cuts and image segmentation', *IEEE Trans. PAMI*, 2000, **22**, (8), pp. 888–905
23 Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: 'Detection of functional modules from protein interaction networks', *Proteins*, 2004, **54**, (1), pp. 49–57
24 Zhang, S., Liu, H.-W., Ning, X.-M., Zhang, X.-S.: 'A hybrid graph–theoretic method for mining overlapping functional modules in large sparse protein interaction networks', *Int. J. Data Mining Bioinf.*, 2009, **3**, (1), pp. 68–84
25 Kohonen, T.: 'Self-organized formation of topologically correct feature maps', *Biol. Cybern.*, 1982, **43**, (1), pp. 59–69
26 Hastie, T., Tibshirani, R., Friedman, J.: 'Unsupervised learning: the elements of statistical learning' (Springer, 2001, 1st edn.), pp. 485–493
27 Cluaset, A., Shalizi, C.R., Newman, M.E.J.: 'Power-law distributions in empirical data', *SIAM Rev.*, 2009, **51**, (4), pp. 661–703
28 Krogan, N.J., *et al.*: 'Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*', *Nature*, 2006, **440**, pp. 637–643
29 http://tap.med.utoronto.ca
30 Nadler, B., Lafon, S., Coifman, R., Kevrekidis, I.: 'Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators', *Adv. Neural Inf. Process. Syst.*, 2006, **18**, pp. 955–962

31 Ruepp, A., Zollner, A., Maier, D., *et al.*: 'The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes', *Nucl. Acids Res.*, 2004, **32**, (18), pp. 5539–5545

32 Ruepp, A., Mewes, H.W.: 'Prediction and classification of protein functions', *Drug Discov. Today: Technol.*, 2006, **3**, (2), pp. 145–151

33 Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., Gerstein, M.: 'Assessing the limits of genomic data integration for predicting protein networks', *Genome Res.*, 2005, **15**, (7), pp. 945–953

34 Cho, Y.-R., Zhang, A.: 'Restructuring protein interaction networks to reveal structural hubs and functional organizations'. IEEE Int. Conf. on Bioinformatics and Biomedicine, Washington, DC, USA, November 2009, pp. 105–110

35 Chua, H., Sung, W.-K., Wong, L.: 'Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions', *Bioinformatics*, 2006, **22**, (13), pp. 1623–1630

36 Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: 'Whole proteome prediction of protein function via graph–theoretic analysis of interaction maps', *Bioinformatics*, 2005, **21**, (1), pp. i302–i310

37 Karaoz, U., Murali, T.M., Levotsky, S., *et al.*: 'Whole-genome annotation by using evidence integration in functional-linkage networks', *Proc. Natl. Acad. Sci. USA*, 2004, **101**, (9), pp. 2888–2893

38 Shin, H., Tsuda, K., Schlkopf, B.: 'Protein functional class prediction with a combined graph', *Expert Syst. Appl.*, 2008, **36**, (2), pp. 3284–3292

doi: 10.1049/iet-syb.2010.0066

Table 3: Some of the most significant MIPS annotations for clusters along with general functions.

Cluster index	Cluster size	Significant annotation (p-value)	General Functions
1	262	splicing (71/1.76e-61), nucleic acid binding(46/1.85e-12), RNA binding (43/3.38e-20)	mRNA processing mRNA splicing
2	322	cell cycle (2.09e-06), mitotic cell cycle and cell cycle control (1.53e-06), mitotic cell cycle (2.58e-07), M phase (2.48e-09), protein fate (folding, modification, destination) (3.58e-09), protein modification by ubiquitination, deubiquitination (1.10e-08), modification by ubiquitin-related proteins (5.11e-04) , assembly of protein complexes (1.49e-09), protein/peptide degradation (2.55e-16), cytoplasmic and nuclear protein degradation (1.32e-16), proteasomal degradation (ubiquitin/proteasomal pathway) (8.19e-18), ATP binding (4.94e-10), cytoskeleton/structural proteins (1.76e-07), actin cytoskeleton (6.14e-08).	proteasome assembly protein catabolic process protein/peptide degradation [proteasomal degradation (ubiquitin/proteasomal pathway)] mitotic cell cycle and cell cycle control protein modification by ubiquitination, deubiquitination ATP binding
3	544	phosphate metabolism (9.89e-09), cell cycle and DNA processing (2.01e-16), DNA processing (1.17e-07), cell cycle (3.64e-13), protein modification (4.35e-07), protein modification by phosphorylation, dephosphorylation, autophosphorylation (1.62e-10), modification by ubiquitination, deubiquitination (3.33e-08), posttranslational modification of amino acids (e.g. hydroxylation, methylation) (3.81e-05), protein binding (4.72e-08), enzymatic activity regulation / enzyme regulator (8.36e-05), RNA transport (4.60e-08), cytoskeleton-dependent transport (3.88e-06), cellular import (2.41e-04), vesicular cellular import(6.13e-08), endocytosis (6.13e-08), cellular signalling (1.14e-04), protein kinase(3.19e-04), stress response (2.30e-12), pheromone response, mating-type determination, sex-specific proteins (9.35e-05), cell fate(2.80e-10), cell growth / morphogenesis (1.02e-10), biogenesis of cellular component (2.84e-08), cytoskeleton/structural proteins (4.76e-15), actin cytoskeleton (4.97e-14), fungal and other eukaryotic cell type differentiation (5.35e-11), budding, cell polarity and filament formation (5.21e-11).	cytoskeleton/structural proteins cytoskeleton-dependent transport
4	619	DNA processing(6.73e-06), DNA synthesis and replication (3.03e-05), DNA restriction or modification(6.11e-05), DNA conformation modification (3.44e-05), transcription(8.35e-18), RNA processing(4.99e-18), rRNA processing(8.01e-28), RNA modification(9.19e-06), rRNA modification(2.06e-05), protein synthesis (7.89e-11), ribosome biogenesis (6.10e-07), protein modification by phosphorylation, dephosphorylation, autophosphorylation (2.63e-07), posttranslational modification of amino acids (e.g. hydroxylation, methylation)(2.41e-06), RNA binding (8.75e-14), RNA transport (2.46e-04).	ribosome biogenesis rRNA processing rRNA modification RNA binding
5	197	DNA restriction or modification (1.16e-06), DNA conformation modification (e.g. chromatin)(4.10e-07), mitotic cell cycle (2.94e-10), G1 phase of mitotic cell cycle(1.40e-11), nuclear and chromosomal cycle (1.99e-06), mRNA synthesis (9.71e-06), general transcription activities (4.69e-09), transcription initiation (7.17e-11), transcriptional control (1.33e-05), protein modification by acetylation, deacetylation (2.31e-12), protein processing (proteolytic)(1.04e-07), DNA binding(4.73e-03), organization of chromosome structure (3.21e-11)	mRNA synthesis general transcription activities transcriptional control protein modification by acetylation, deacetylation protein processing (proteolytic) organization of chromosome structure DNA binding
6	212	Regulation of C-compound and carbohydrate metabolism (2.63e-05), transcription(5.75e-06), mRNA synthesis (3.23e-09), transcriptional control (2.05e-10), cytoplasmic and nuclear protein degradation (2.13e-	Transcription Transcription regulation transcriptional control

		05), protein binding (2.48e-05), protein transport (9.94e-04).	protein transport protein binding
7	440	cell cycle and DNA processing (1.73e-10), DNA processing (2.27e-11), DNA synthesis and replication (3.64e-06), ori recognition and priming complex formation(1.41e-03), extension/ polymerization activity(9.16e-04), DNA repair (8.37e-04), DNA conformation modification (1.53e-10), cell cycle (2.45e-04), mitotic cell cycle (9.50e-05), G2/M transition of mitotic cell cycle(3.26e-05), nuclear and chromosomal cycle (2.49e-06), spindle pole body/centrosome and microtubule cycle (5.76e-04), transcriptional control (8.92e-04), rRNA processing (4.63e-05), protein synthesis (4.22e-04), translation(9.14e-03), translational control(1.48e-03), modification by acetylation, deacetylation(1.96e-05), nucleic acid binding(3.63e-03), ATP binding (4.51e-03).	Cell cycle and DNA processing Protein biosynthesis
8	446	transcription initiation(6.91e-05), transcription termination(1.40e-06), RNA processing (4.86e-06), 3'-end processing(5.68e-09), protein synthesis (1.29e-33), ribosome biogenesis (4.63e-25), ribosomal proteins (1.52e-21), translation(8.79e-13), translation initiation (3.40e-13), protein modification by phosphorylation, dephosphorylation, autophosphorylation(3.59e-03), posttranslational modification of amino acids (e.g. hydroxylation, methylation)(5.01e-05), RNA binding (1.62e-07), biogenesis of cellular components(1.49e-12), mitochondrion (3.14e-26).	protein synthesis ribosome biogenesis mitochondrion biogenesis of cellular components
9	255	nucleotide/nucleoside/nucleobase metabolism (6.07e-10), RNA degradation (1.95e-16), glyoxylate cycle (1.37e-05), transcription (1.01e-16), RNA synthesis (4.60e-15), rRNA synthesis (1.05e-17), tRNA synthesis (2.49e-21), mRNA synthesis (8.11e-07), rRNA processing (1.21e-04), protein with binding function or cofactor requirement (structural or catalytic) (4.52e-14), protein binding (1.84e-04), nucleic acid binding(8.46e-12), DNA binding (3.00e-20), structural protein binding (2.12e-03), RNA transport (2.19e-04), nuclear transport(4.52e-05), cytoskeleton/structural proteins (1.25e-04), nucleus (3.71e-03), nuclear membrane(1.86e-04).	rRNA synthesis tRNA synthesis protein with binding function or cofactor requirement (structural or catalytic) protein binding nucleic acid binding DNA binding structural protein binding
10	186	metabolism (3.63e-03), phosphate metabolism (1.99e-03), C-compound and carbohydrate metabolism (6.58e-04), glycogen metabolism(4.18e-06), regulation of glycolysis and gluconeogenesis(2.31e-03), metabolism of energy reserves (e.g. glycogen, trehalose)(1.48e-03), DNA processing (2.65e-13), DNA topology (2.09e-04), DNA repair (2.52e-05), DNA conformation modification (3.11e-10), mitotic cell cycle and cell cycle control(2.00e-04), mRNA synthesis (4.72e-07), transcriptional control (6.61e-07), cyclic nucleotide binding (cAMP, cGMP, etc.)(1.11e-04), G-protein mediated signal transduction(4.94e-03).	chromosome organization primary metabolic process response to stress
11	272	amino acid metabolism (2.44e-03), DNA conformation modification (e.g. chromatin) (4.94e-10), cell cycle (4.72e-04), nuclear and chromosomal cycle (2.63e-05), centromere/kinetochore complex maturation (1.97e-03), chromosome segregation/division (4.49e-05), transcriptional control (3.59e-03), rRNA processing (4.15e-03), tRNA processing (4.49e-06), translation (4.59e-04), translation elongation(2.23e-04), protein folding and stabilization (5.18e-05), protein binding (1.93e-05), ATP binding(1.04e-02), stress response (2.64e-03), unfolded protein response (3.19e-03), organization of chromosome structure (1.61e-06).	DNA conformation modification nuclear and chromosomal cycle chromosome segregation/division organization of chromosome structure chromatin remodeling protein folding and stabilization ATP binding
12	443	phosphate metabolism (2.50e-03), regulation of phosphate metabolism(1.24e-03), regulation of C-compound and carbohydrate metabolism (9.95e-05), regulation of lipid, fatty acid and isoprenoid	chromatin associated proteins

		metabolism(1.98e-03), glycolysis and gluconeogenesis(2.40e-04), regulation of glycolysis and gluconeogenesis (2.48e-06), DNA processing(1.80e-22), DNA synthesis and replication(2.03e-09), ori recognition and priming complex formation(4.45e-06), extension/ polymerization activity(9.27e-07), DNA recombination and DNA repair(1.52e-06), DNA conformation modification (e.g. chromatin)(9.38e-24), mitotic cell cycle and cell cycle control(2.13e-08), mitotic cell cycle(4.68e-05), G2/M transition of mitotic cell cycle(7.61e-06), cell cycle checkpoints(2.53e-04), mRNA synthesis (1.55e-11), transcriptional control(5.14e-10), nucleic acid binding (8.44e-07), DNA binding (2.20e-08), regulation of metabolism and protein function(1.29e-03), regulation of protein activity (1.16e-03), enzymatic activity regulation / enzyme regulator (1.16e-03), ER to Golgi transport(1.13e-04), stress response (3.19e-05), DNA damage response (9.95e-05), cellular sensing and response to external stimulus(1.45e-04), chemoperception and response (2.39e-04), cell fate (2.34e-03), cell aging(3.14e-03).	

Table 4: The selected proteins of the modules along with the biological processes they are involved in.

	Selected proteins	Biological process
1	1.Small nuclear ribonucleoprotein Sm D2(YLR275W) 2.Small nuclear ribonucleoprotein Sm D3(YLR147C) 3. Small nuclear ribonucleoprotein F(YPR182W) 4. Importin subunit alpha(YNL189W) 5. 5'-3' exoribonuclease 1(YGL173C) 6. Importin subunit beta-1(YLR347C) 7. Small nuclear ribonucleoprotein Sm D1(YGR074W) 8. Pre-mRNA-splicing factor 8(YHR165C) 9. Pre-mRNA-processing protein PRP40(YKL012W)	nuclear mRNA splicing, via spliceosome protein import into nucleus protein targeting to membrane mRNA processing assembly of spliceosomal tri-snRNP nuclear mRNA 3'-splice site recognition
2	1. 26S proteasome regulatory subunit RPN3(RPN3) 2. 26S proteasome regulatory subunit RPN5(YDL147W) 3. 26S proteasome regulatory subunit SEM1(YDR363W-A) 4. 26S proteasome regulatory subunit RPN1(YHR027C) 5. Ubiquitin carboxyl-terminal hydrolase 6(YFR010W) 6. UV excision repair protein RAD23(YEL037C) 7. 26S proteasome regulatory subunit RPN2(YIL075C) 8. 26S proteasome regulatory subunit RPN6(YDL097C) 9. 26S proteasome regulatory subunit RPN9(YDR427W) 10. 26S proteasome regulatory subunit RPN8(YOR261C) 11.26S proteasome regulatory subunit RPN12(YFR052W) 12.26S proteasome regulatory subunit RPN11(YFR004W) 13. 26S protease regulatory subunit 8 homolog(YGL048C) 14. 26S protease regulatory subunit 4 homolog(YDL007W) 15. 26S proteasome regulatory subunit RPN 10(YHR200W) 16. 26S proteasome regulatory subunit RPN 7(YPR108W)	regulation of protein catabolic process ubiquitin-dependent protein catabolic process proteasomal ubiquitin-dependent protein catabolic process protein deubiquitination proteasome assembly regulation of cell cycle ER-associated protein catabolic process negative regulation of protein catabolic process nucleotide-excision repair, DNA damage recognition
3	1. Coronin-like protein(YLR429W) 2. Reduced viability upon starvation protein 167(YDR388W) 3. RNA annealing protein YRA1(YDR381W) 4. Myosin light chain 1(YGL106W)	actin cortical patch localization actin filament organization microtubule-based process negative regulation of Arp2/3 complex-mediated actin nucleation endocytosis mRNA export from nucleus regulation of transcription transcription transcription-coupled nucleotide-excision repair cytokinesis, contractile ring formation protein localization vesicle targeting
4	1. RNA annealing protein YRA1(YDR381W) 2. rRNA 2'-O-methyltransferase fibrillarin(YDL014W) 3. Serine/threonine-protein kinase BUR1(YPR161C) 4. rRNA biogenesis protein RRP5(YMR229C) 5. Coronin-like protein(YLR429W)	mRNA export from nucleus regulation of transcription transcription transcription-coupled nucleotide-excision repair Ribosome biogenesis rRNA processing box C/D snoRNA 3'-end processing rRNA methylation tRNA processing
5	1.Transcription initiation factor TFIID subunit 6(YGL112C)	RNA polymerase II transcriptional preinitiation complex

	2.Proteasome component Y13(YGR135W) 3.Proteasome component C7-alpha(YGL011C) 4.Proteasome component C1(YER012W) 5.Proteasome component PUP1(YOR157C) 6.Transcription initiation factor TFIID subunit 12(YDR145W) 7.Transcription initiation factor TFIID subunit 5(YBR198C)	assembly general transcription from RNA polymerase II promoter histone acetylation regulation of transcription factor activity filamentous growth ubiquitin-dependent protein catabolic process ascospore formation response to stress
6	1.Importin subunit alpha(YNL189W) 2.Cell division control protein 48(YDL126C) 3.Mediator of RNA polymerase II transcription subunit 18(YGR104C) 4.Mediator of RNA polymerase II transcription subunit 17(YER022W) 5.Mediator of RNA polymerase II transcription subunit 15(YOL051W) 6.Mediator of RNA polymerase II transcription subunit 4(YOR174W) 7.Mediator of RNA polymerase II transcription subunit 31(YGL127C)	protein import into nucleus protein targeting to membrane ER-associated protein catabolic process apoptosis mitotic cell cycle protein transport vesicle fusion negative regulation of transposition, RNA-mediated regulation of transcription transcription from RNA polymerase II promoter transcription from RNA polymerase II promoter negative regulation of transposition, RNA-mediatedDNA repair meiotic gene conversion meiotic sister chromatid segregation
7	1.Suppressor protein STM1(YLR150W) 2.RNA annealing protein YRA1(YDR381W) 3.Protein LSM12(YHR121W) 4.Transcriptional regulatory protein SIN3(YOL004W) 5.ATP-dependent RNA helicase eIF4A(YJL138C) 6.Histone acetyltransferase type B catalytic subunit(YPL001W) 7.GU4 nucleic-binding protein 1(YGL105W) 8.Histone deacetylase RPD3(RPD3) 9.Elongation factor 2(YDR385W) 10.Glutamyl-tRNA synthetase, cytoplasmic(YGL245W)	Protein biosynthesis chromatin silencing at rDNA chromatin silencing at silent mating-type cassette chromatin silencing at telomere chromatin organization during transcription loss of chromatin silencing during replicative cell aging cell cycle cell division mRNA export from nucleus Transcription Transcription regulation translational initiation RNA metabolic process double-strand break repair via nonhomologous end joining histone deacetylation negative regulation of transposition, RNA-mediated DNA repair histone acetylation histone deacetylation positive regulation of ligase activity tRNA aminoacylation for protein translation double-strand break repair via nonhomologous end joining mitotic recombination phase of mitotic cell cycle positive regulation of translational elongation glutamyl-tRNA aminoacylation
8	1.RNA annealing protein YRA1(YDR381W) 2.Mitochondrial acidic protein MAM33(YIL070C) 3.37S ribosomal protein RSM18, mitochondrial(YER050C) 4.COMPASS component SWD2(YKL018W) 5.37S ribosomal protein S5, mitochondrial(YBR251W)	mRNA export from nucleus regulation of transcription transcription transcription-coupled nucleotide-excision repair aerobic respiration response to drug mitochondrial translation

9	1.DNA-directed RNA polymerases I, II, and III sbunit RPABC1(YBR154C) 2.DNA-directed RNA polymerases I, II, and III subunit RPABC5(YOR210W) 3.DNA-directed RNA polymerases I and III subunit RPAC1(YPR110C) 4.DNA-directed RNA polymerase II subunit RPB1(YDL140C) 5.DNA-directed RNA polymerases I and III subunit RPAC2(YNL113W) 6.DNA-directed RNA polymerase I subunit RPA49(YNL248C) 7.DNA-directed RNA polymerase II subunit RPB9(YGL070C) 8.DNA-directed RNA polymerase II subunit RPB3(YIL021W) 9.Exosome complex exonuclease DIS3(YOL021C) 10.DNA-directed RNA polymerases I, II, and III subunit RPABC2(RPO26) 11.Exosome complex component RRP45(RRP45)	transcription from RNA polymerase II promoter regulation of transcription transcription regulation of cell size DNA damage DNA repair response to drug transcription-coupled nucleotide-excision repair rRNA processing
10	1.Actin-related protein 7(YPR034W) 2.Actin-like protein ARP9(YMR033W) 3.Regulator of Ty1 transposition protein 102(YGR275W) 4.Chromatin structure-remodeling complex protein RSC8(YFR037C) 5.Chromatin structure-remodeling complex subunit RSC7(YMR091C) 6.Nuclear protein STH1/NPS1(YIL126W) 7.High temperature lethal protein 1(YCR020W-B) 8.Chromatin structure-remodeling complex protein RSC58(YLR033W) 9.Chromatin structure-remodeling complex subunit SFH1(YLR321C) 10.Chromatin structure-remodeling complex subunit RSC2(YLR357W) 11.Chromatin structure-remodeling complex subunit RSC4(YKR008W) 12.Chromatin structure-remodeling complex protein RSC3(YDR303C) 13.Chromtin structure-remodeling complex protein RSC6(YCR052W) 14.Chromatin structure-remodeling complex subunit RSC9(YML127W)	ATP-dependent chromatin remodeling RNA elongation from RNA polymerase II promoter nucleosome disassembly regulation of transcription chromatin remodeling chromosome segregation G1/S transition of mitotic cell cycle double-strand break repair via nonhomologous end joining protein import into nucleus response to DNA damage stimulus G2/M transition of mitotic cell cycle chromatin remodeling at centromere chromosome segregation cytoskeleton organization double-strand break repair meiosis chromatin remodeling regulation of cell cycle response to DNA damage stimulus regulation of DNA replication during S phase
11	1.ATP-dependent molecular chaperone HSC82(YMR186W) 2.Actin-related protein 4(YJL081C) 3.ERAD-associated E3 ubiquitin-protein ligase HRD1(YOL013C) 4.RuvB-like protein 2(YPL235W) 5.Elongation factor 3A(YLR249W) 6.RuvB-like protein 1(YDR190C)	proteasome assembly protein folding response to stress telomere maintenance DNA repair chromatin remodeling histone acetylation kinetochore assembly regulation of transcription from RNA polymerase II promoter transcription ER-associated protein catabolic process fungal-type cell wall organization

		rRNA processing snoRNA metabolic process translational elongation
12	1.Histone H2A.2(YBL003C) 2.FACT complex subunit SPT16(YGL207W) 3.High mobility group protein 1(YDR174W) 4.Nucleosome assembly protein(YKR048C) 5.Histone H2B.2(YBL002W) 6.Histone H4(YBR009C) 7.Casein kinase II subunit alpha(YIL035C) 8.FACT complex subunit POB3(YML069W) 9.Histone H3(YBR010W/YNL031C) 10.Histone acetyltransferase type B catalytic subunit(YPL001W) 11.Signal peptidase complex subunit SPC1(YJR010C-A) 12.Serine/threonine-protein kinase BUR1(YPR161C)	nucleosome assembly nucleosome disassembly DNA repair transcription initiation from RNA polymerase II promoter regulation of transcription from RNA polymerase I and II and III promoter RNA elongation from RNA polymerase II promoter regulation of transcription negative regulation of transposition, RNA-mediated chromatin silencing at telomere DNA replication plasmid maintenance M phase of mitotic cell cycle budding cell bud growth protein import into nucleus histone H3-K79 methylation G1/S transition of mitotic cell cycle G2/M transition of mitotic cell cycle cellular ion homeostasis establishment of cell polarity flocculation via cell wall protein-carbohydrate interaction protein amino acid phosphorylation response to DNA damage stimulus DNA-dependent DNA replication histone acetylation protein targeting to ER signal peptide processing phosphorylation of RNA polymerase II C-terminal domain positive regulation of histone H3-K4 methylation transcription