Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 472415, 10 pages
doi:10.1155/2009/472415

Research Article

Investigating Topological and Functional Features of Multimodular Proteins

Zelmina Lubovac

Systems Biology Research Centre, School of Life Sciences, University of Skövde, Box 408, 54128 Skövde, Sweden

Correspondence should be addressed to Zelmina Lubovac, zelmina.lubovac@his.se

Received 16 February 2009; Revised 19 July 2009; Accepted 12 September 2009

Recommended by Tatsuya Akutsu

To generate functional modules as functionally and structurally cohesive formations in protein interaction networks (PINs) constitutes an important step towards understanding how modules communicate on a higher level of the PIN organisation that underlies cell functionality. However, we need to understand how individual modules communicate and are organized into the higher-order structure(s) of the PIN organization that underlies cell functionality. In an attempt to contribute to this understanding, we make an assumption that the proteins reappearing in several modules, termed here as multimodular proteins (MMPs), may be useful in building higher-order structure(s) as they may constitute communication points between different modules. In this paper, we investigate common properties shared by these proteins and compare them with the properties of so-called single-modular proteins (SMPs) by analyzing three aspects: functional aspect, that is, annotation of the proteins, topological aspect that is betweenness centrality of the proteins, and lethality. Furthermore, we investigate the interconnectivity role of some proteins that are identified as functionally and topologically important.

Copyright © 2009 Zelmina Lubovac. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

One of the challenges that systems biology is facing consists of explaining biological organisation in the light of the existence of modules in networks [1–4]. A proposal that cellular function is carried out by modules [5] has fired a "modular era" of systems biology in which the focus has been on studying modularity at different levels of cellular organisation. A series of studies attempting to reveal the modules in cellular networks, ranging from metabolic [6] to protein networks [7, 8], support the proposal that modular architecture is one of the principles underlying biological organisation.

To generate functional modules as functionally and structurally cohesive formations in PINs is an important step towards understanding how individual modules communicate and are organised on a higher level of the PIN organisation that underlies cell functionality. We here investigate whether the proteins that appear in several modules, that we term multimodular proteins (MMPs), may be useful in building higher-order structure(s) as they may constitute communication points between different modules.

In this paper, we investigate common properties shared by these proteins and compare them with the properties of single-modular proteins (SMPs), that is, proteins that occur in only one module, by analysing three aspects: functional aspect, that is, annotation of the proteins, using Gene Ontology (GO), topological aspect that is betweenness centrality of the proteins, which is used to find topologically important proteins, and their lethality. Furthermore, we investigate the interconnectivity role of some proteins that are identified as functionally and topologically important.

2. Materials and Methods

2.1. Experimental Data Sets. The data set referred to as CORE data consists of protein-protein interactions that were downloaded from the Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu/). DIP stores and organises experimentally determined interactions between proteins in *Saccharomyces cerevisiae* [9]. The majority of the interactions were identified with high-throughput yeast

two-hybrid (Y2H) screens. [10] We used the subset of DIP-YEAST, denoted as CORE, which has been validated in [11]). After removal of 195 self-interactions, the CORE subset contained 6375 interactions between 2231 proteins.

The second data set, referred to as von Mering data, consists of protein interactions critically evaluated by von Mering et al. (2002) [11], where a quality assessment of large-scale data sets of protein-protein interactions in *Saccharomyces cerevisiae* was performed. In [12], data sets from yeast two-hybrid (Y2H) systems, protein complex purification techniques that rely on mass-spectroscopy (TAP and HMS-PCI), correlated mRNA expression profiles, genetic interactions, and *in silico* interaction predictions were analysed. As stated further in this study, each of these methods can be used to predict protein interactions, even though their goals are slightly different. While the main purpose with yeast two-hybrid and mass spectrometry is to identify physical binding between pairs of proteins, the remaining of the mentioned methods is mainly focused on predicting functional associations, which in many cases also requires physical binding [12]. The authors integrated about 80 000 interactions between proteins in and found that only 2455 were supported by more than one method. This low overlap between sets of protein interactions obtained from different methods may be due to the high fraction of false positives but may also be caused by the difficulties for some methods to capture certain types of interactions. All interactions are classified by the level of confidence (low, medium, high), based on the evidence that supports them. In this paper, we have used the interaction set with high level of confidence, meaning that all interactions are confirmed by several methods. We will refer to this data set as "von Mering." The data set contains 2455 interactions between 988 proteins.

2.2. Algorithm for Module Identification. In previous work by Bader and Hogue (2003), an algorithm for finding complexes in large-scale networks, called MCODE, based on the weighting of nodes with a core-clustering coefficient was proposed. The core-clustering coefficient of a node i is defined as the density of the highest k-core of the closed neighbourhood $N[i]$. The highest k-core of a graph is the central most densely connected subgraph. We have earlier proposed a weighted core-clustering coefficient for identifying topologically and functionally cohesive clusters [13]. The weighting scheme uses the weighted core-clustering coefficient of node i, which is defined as the weighted clustering coefficient of the highest k-core of the closed neighbourhood $N[i]$ multiplied by the highest core number.

We called the algorithm SWEMODE (Semantic Weights for MODule Elucidation). SWEMODE has three options concerning traversal of nodes that are considered for inclusion in a module, as described in [13]. Here, we use depth-first search; that is, the protein graph is searched starting from the seed node, which is the highest weighted node, followed by recursively traversing the graph outwards from the seed node, identifying new module members according to the given *NWP* (Node Weight Percentage) criterion. As in [14], the requirement for inclusion of the neighbours in a module is that their weights are higher than a threshold, which is a given *NWP* of the seed node. At this stage, once a node has been visited and added to the module, it cannot be added to another module [13]. However, in the postprocessing step, overlap is allowed to some extent. Because we here choose to go further by inspecting the interconnectedness, it is valuable to traverse not only the immediate neighbours but also other indirect neighbours.

In a postprocessing step, modules that contain less than three members may be removed, both before and after applying a so-called "fluffing" step. The degree of "fluffing" is referred to as "fluff" parameter and can vary between 0.0 and 1.0 [14]. For every member in the module, its immediate neighbours are added to the module if they have not been visited and if their neighbourhood weighted cohesiveness is higher than the given fluff threshold f.

To identify topologically and functionally important proteins, we calculated the number of module occurrences for each protein across 200 sets of overlapping modules (the fluff parameter was varied between 0 and 1 in increments of 0.1 and the *NWP* parameter was varied between 0 and 0.95 in increments of 0.05). All three GO aspects were combined into a single weight for each protein. All modules that only contain a single member are removed from further analysis.

For each seed protein, we calculated the number of times each protein appears in different modules in each module set, divided by the number of module sets it appears in. For example, if protein Nup100 is member of 10 modules in one module set and 20 modules in the another module set, the average number of module occurrences of the protein will be $(10 + 20)/2 = 15$.

2.3. Betweenness Centrality. Betweenness centrality has been applied in the context of social networks, to measure the centrality and influence of a person or a group [15]. The betweenness centrality of a node v is originally defined by Freeman (1977) as the number of shortest paths between other nodes that pass through v and it is given by

$$C_B(v) = \sum_{i,j \in V:\, i \neq j, i \neq v, j \neq v} \frac{g_{ivj}}{g_{ij}}, \tag{1}$$

where g_{ivj} is the number of the shortest path linking i and j that contain v, and g_{ij} is the total number of the shortest path between i and j. High-betweenness nodes occur on large number of nonredundant shortest paths between other nodes. If a node with high-betweenness centrality is removed, it may disconnect different parts of the network completely. Thus, such nodes may be thought of as potential bridges between modules in network and have most influence on the information transfer.

2.4. Lethality. We obtained lethality data from the MIPS database [16]. There are 1015 lethal proteins obtained from manually curated MIPS database. The list of MMPs and SMPs observed across modules in both data sets was compared to the list of lethal proteins.

Table 1: Annotation statistics for top ten multimodular proteins.

Proteins	Cdc28	Nap1	Prp43	Pre1	Pwp2	Sed5	Tfp1	Nop4	Utp7	Rpc40
Module frequency	4.2	3.9	2.9	2.7	2.7	2.6	2.6	2.6	2.5	2.5
GO biological process	cell organization and biogenesis									
GO frequency	80%									
P value	$3.8 \cdot 10^{-4}$									

Table 2: Statistics for the most significant GO terms based on GO biological process. Module frequency decreases from left to right, and the last column contains a group of proteins that occur in only one module or are not present in any of the modules.

Module frequency	≥1.9	≥1.7	≥1.4	≥1.2	>1	≤1
[#] proteins	[50]	[100]	[150]	[200]	[250]	[250]
GO biological process	GO term frequency P value					
Ribonucleoprotein complex biogenesis and assembly (5.5%)	42% $9.3 \cdot 10^{-18}$	36% $3.2 \cdot 10^{-18}$	41% $4.9 \cdot 10^{-37}$	40% $1.9 \cdot 10^{-47}$	41% $3.9 \cdot 10^{-64}$	16% $1.1 \cdot 10^{-06}$
Cellular component organization and biogenesis (30%)	70% $1.7 \cdot 10^{-06}$	62% $1.2 \cdot 10^{-08}$	65% $1.0 \cdot 10^{-16}$	65% $2.2 \cdot 10^{-22}$	66% $1.9 \cdot 10^{-29}$	56% $3.9 \cdot 10^{-55}$
Organelle organization and biogenesis (17.8%)	50% $5.8 \cdot 10^{-05}$	45% $1.0 \cdot 10^{-07}$	51% $2.1 \cdot 10^{-18}$	50% $3.1 \cdot 10^{-23}$	53% $2.1 \cdot 10^{-35}$	35% $2.2 \cdot 10^{-08}$
RNA metabolic process (14.2%)	44% $8.9 \cdot 10^{-05}$	48% $1.8 \cdot 10^{-13}$	46% $1.3 \cdot 10^{-18}$	45% $5.9 \cdot 10^{-24}$	46% $1.3 \cdot 10^{-31}$	32% $1.1 \cdot 10^{-10}$
Primary metabolic process (44%)	74% $4.7 \cdot 10^{-03}$	79% $2.1 \cdot 10^{-10}$	79% $1.1 \cdot 10^{-15}$	81% $8.7 \cdot 10^{-25}$	80% $2.5 \cdot 10^{-30}$	78% $3.8 \cdot 10^{-25}$

3. Results

3.1. GO Annotation of Multimodular Proteins

3.1.1. CORE Data Set. We started by analysing annotations with help of SGD GO Term Finder (http://www.yeastgenome.org/help/goTermFinder.html), in order to identify the most significantly shared GO terms among the MMPs with varying number of module occurrence. The subontology "biological process" was chosen. The majority of the most frequent multimodular proteins (top 10) are annotated with the GO biological process term "cell organization and biogenesis," which has the following GO definition: "the processes involved in the assembly and arrangement of cell structures, including the plasma membrane and any external encapsulation structures such as the cell wall and cell envelope," as described in [17]. Table 1 shows the top ten MMPs, where 80% (highlighted proteins) belong to the above mentioned class. GO Frequency in Table 1 shows the percentage of those proteins that are annotated with the given GO term. The most significantly shared term is obtained by examining the group of proteins to find the GO term to which the highest fraction of the proteins is associated, compared to the number of times that the term is associated with other yeast proteins. The significance (P value) of the shared GO term describing the biological process for the ten most frequent proteins is shown in the last row in Table 1.

In addition, we have repeated the same evaluation procedure by adding proteins with decreasing module frequency to analyse how the annotation statistics is affected by adding those proteins. The summary of those results may be found in Table 6. The first column shows the statistics for the top 50 protein, where all proteins are present in approximately 2 modules in average. Still, the majority of the proteins share the GO term "cell organization and biogenesis", which is also the most significant term ($P = 1.3 \cdot 10^{-11}$), and the GO frequency has increased slightly from 80% to 82%. For comparison, 50 random SMPs were evaluated with the same procedure. Here we found that the most significant term that is shared among 96% of those proteins is the GO biological process term "cellular process" ($P = 2.1 \cdot 10^{-5}$), which may not help us to derive any conclusions about the more specific roles of those proteins. Also in this subset of proteins, we found that the GO term "cell organization and biogenesis" is shared among proteins, but the GO frequency for this term is 63%, compared to 82% of most frequent MMPs that are annotated with this term.

GO term frequency for the most significant terms decreases gradually as we add more proteins with decreasing module frequency. Several nonsignificant annotation terms appear as we add proteins with decreasing module frequency, meaning that those proteins have more dispersed annotation, while high-frequent MMPs seem to have more consistent annotation dominated by their participation in cellular organisation.

Cdc28, which appears most frequently in modules, is one of five different cyclin-dependent protein kinases (CDKs) in yeast and has a fundamental role in the control of the main

Table 3: Comparison between top 100 most frequent multimodular proteins and most frequent "bottle neck" proteins, identified by Przulj et al. (2003).

Module freq.	≥2.1		≥1.9		≥1.8		≥1.7	
"bottle necks"	≥25		≥18		≥14		≥11	
[#] Proteins	[25]		[50]		[75]		[100]	
GO biological Process	GO term frequency *P* value							
Cellular process (64.1%)	—	100% $5.1 \cdot 10^{-3}$	—	98% $2.5 \cdot 10^{-5}$	93% $1.0 \cdot 10^{-6}$	95% $2.0 \cdot 10^{-7}$	—	94% $9.7 \cdot 10^{-10}$
Ribonucleoprotein complex biogenesis and assembly (5.5%)	40% $5.6 \cdot 10^{-5}$	—	42% $9.3 \cdot 10^{-18}$	32% $1.9 \cdot 10^{-6}$	39% $1.7 \cdot 10^{-15}$	27% $1.0 \cdot 10^{-6}$	36% $3.2 \cdot 10^{-18}$	25% $4.8 \cdot 10^{-8}$
Cellular component organization and biogenesis (30%)	—	—	70% $1.7 \cdot 10^{-06}$	66% $6.3 \cdot 10^{-5}$	63% $1.5 \cdot 10^{-6}$	61% $9.2 \cdot 10^{-6}$	62% $1.2 \cdot 10^{-8}$	63% $4.2 \cdot 10^{-9}$
Organelle organization and biogenesis (17.8%)	—	—	50% $5.8 \cdot 10^{-5}$	46% $1.6 \cdot 10^{-3}$	48% $5.7 \cdot 10^{-7}$	43% $2.2 \cdot 10^{-4}$	45% $1.0 \cdot 10^{-7}$	43% $2.0 \cdot 10^{-6}$
Cellular metabolic process (46.6%)	—	—	—	76% $8.6 \cdot 10^{-3}$	79% $3.6 \cdot 10^{-6}$	79% $5.9 \cdot 10^{-6}$	81% $4.7 \cdot 10^{-10}$	77% $2.4 \cdot 10^{-7}$
RNA metabolic process (14.2%)	—	—	48% $1.8 \cdot 10^{-13}$	—	52% $2.8 \cdot 10^{-12}$	35% $3.4 \cdot 10^{-3}$	48% $1.8 \cdot 10^{-13}$	32% $2.4 \cdot 10^{-3}$
Primary metabolic process (44%)	—	—	79% $2.1 \cdot 10^{-10}$	—	79% $2.7 \cdot 10^{-7}$	73% $1.2 \cdot 10^{-4}$	79% $2.1 \cdot 10^{-10}$	73% $2.0 \cdot 10^{-6}$

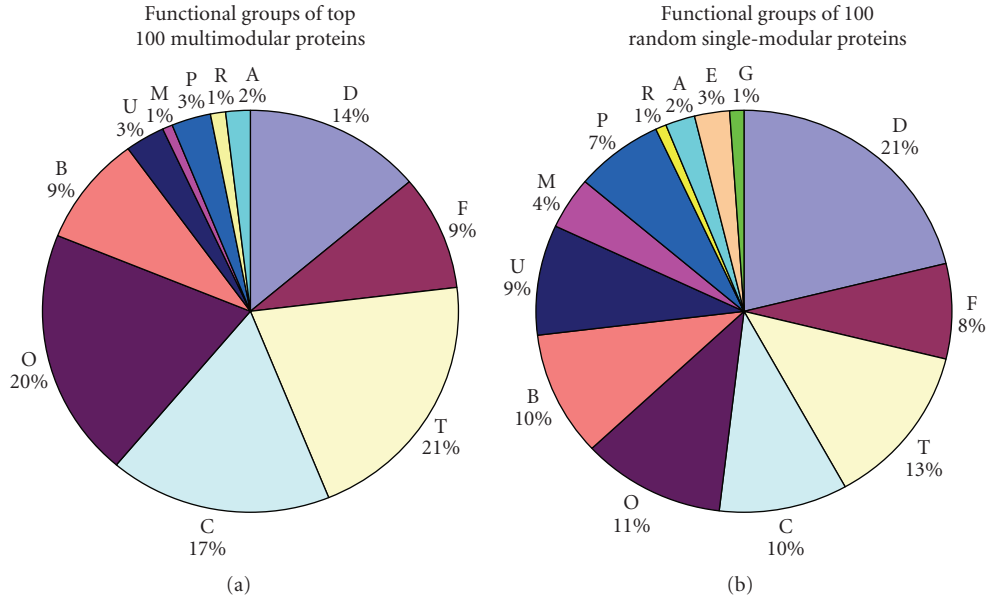

Figure 1: Statistics for MIPS functional categories: D: genome maintenance; T: transcription; F: protein fate; C: cellular fate/organisation; O: cellular organisation; G: amino acid metabolism; M: other metabolism; E: energy production; R: stress and defence; B: transcriptional control; P: translation; A: transport and sensing; U: uncharacterized.

events of the yeast cell cycle [18]. Topologically, it acts as a hub; that is, it holds together several functionally related clusters in the interaction network. In previous work, this protein was suggested to be a part of the intramodule path within the yeast filamentation network, because it had the highest intracluster connectivity; that is, it was the protein with the highest number of interactions with other members of the same cluster [1]. It is therefore highly interesting that we have identified this protein as the most frequent in our modules, as described in [17].

We further evaluated the proteins by analysing their MIPS functional categories [16], to determine what functional characteristics may be derived by studying proteins based on their module frequency. We observed that proteins involved in cellular organisation (O) appear more frequently among the top 100 MMPs, compared to the random

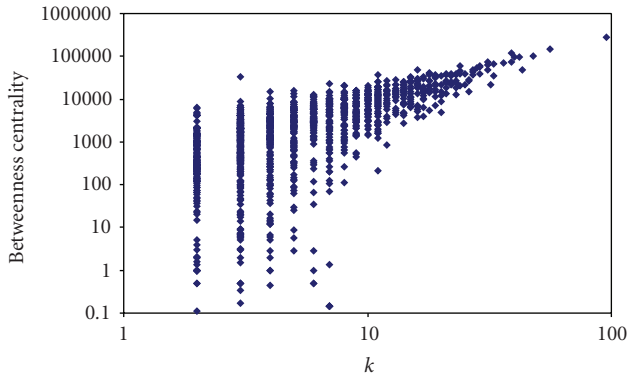

Figure 2: Degree (k) versus betweenness centrality plotted on algorithmic scale.

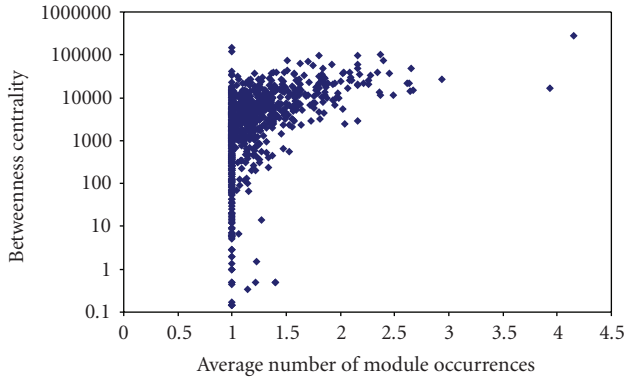

Figure 3: Average number of module occurrences versus betweenness centrality plotted on algorithmic scale.

Table 4: Lethality among multimodular proteins (MMPs) across both data sets.

	No. of MMP	No. of lethal proteins	Percentage
CORE	480	222	46.3%
Von Mering	83	57	68.7%

Table 5: Lethality among single-modular proteins (SMPs) across both data sets.

	No. of SMP	No. of lethal proteins	Percentage
CORE	502	173	34.5%
Von Mering	213	116	54.5%

set of SMPs. Among SMPs, we found that transcription seems to be enriched, as 13% of proteins are annotated with T-transcription and 10% were annotated with B-transcriptional control. This result supports our findings based on studying GO biological process annotation, where "cell organization and biogenesis" were the most significant term among multimodular proteins.

We have also found a lower percentage of uncharacterised proteins in the chart that shows the statistics for the 100 most frequent MMPs (see Figure 1), while none of the proteins

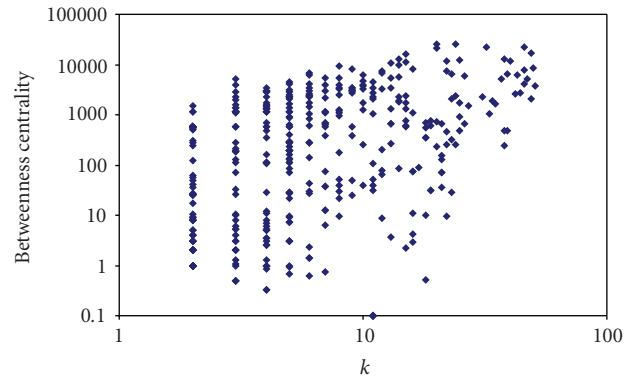

Figure 4: Degree (k) versus betweenness centrality plotted on algorithmic scale.

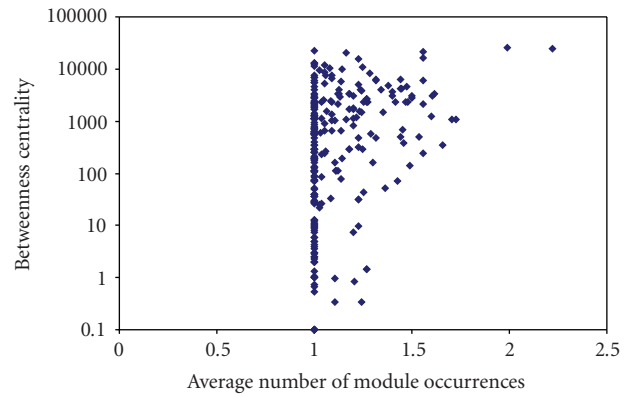

Figure 5: Average number of module occurrences versus betweenness centrality plotted on algorithmic scale.

in the top 50 MMPs is uncharacterised (see Figure 7). This indicates that the more often the protein takes part in the different modules, the higher is the probability that the protein has a defined function. In the same chart (see Figure 1(a)), the proteins that belong to amino acid metabolism and energy production are absent. By studying Figure 7, we can conclude that there is a high fraction of the proteins belonging to the cellular organisation category in each of the module frequency intervals. To make the charts comparable, we have sorted the proteins in decreasing order of module frequency and divided them into the four groups of high-frequent proteins, where each group contains 50 proteins (see pie charts in the first row), and four different groups that contain SMPs (see pie charts in the bottom row). The fraction of proteins that belong to the category "cellular organisation" in multimodular proteins is constantly higher (varies between 18% and 26%) than the fraction of such proteins in the single-modular groups of proteins (varies between 4% and 8%).

3.1.2. Von Mering Data Set. One of the important goals in systems biology is to find relations between the topological properties and functional features of genes and proteins in the networks. In previous network studies, the focus has

Table 6: Annotation statistics for multimodular proteins of different module frequency versus single-modular proteins from Yeast CORE data set. Statistics for the most significant annotation terms of the multimodular proteins with varying occurrences intervals, compared to the corresponding statistics for single-modular proteins (CORE data set).

Module frequency	≥1.9	≥1.6	≥1.4	≥1.3	≥1.3	≥1.2	≥1.2	≥1.1	≥1.1	=1
No. proteins	50	100	150	200	250	300	350	400	450	450
GO biological process	GO term frequency *P* value									
GO: 0016043 (30%) cellular component organization and biogenesis (30%)	82% $1.3 \cdot 10^{-11}$	77% $1.1 \cdot 10^{-19}$	76% $6.2 \cdot 10^{-29}$	78% $5.4 \cdot 10^{-42}$	75% $4.5 \cdot 10^{-47}$	74% $3.2 \cdot 10^{-54}$	73% $1.0 \cdot 10^{-60}$	71% $1.5 \cdot 10^{-66}$	72.7% $4.7 \cdot 10^{-80}$	66% $2.1 \cdot 10^{-57}$
GO: 0006996 (17.8%) organelle organization and biogenesis (17.8%)	60% $1.2 \cdot 10^{-08}$	53% $5.8 \cdot 10^{-13}$	49% $3.0 \cdot 10^{-16}$	51% $7.4 \cdot 10^{-24}$	47% $9.3 \cdot 10^{-25}$	46% $2.9 \cdot 10^{-27}$	46% $4.9 \cdot 10^{-33}$	44% $7.2 \cdot 10^{-34}$	46% $1.3 \cdot 10^{-43}$	43% $4.5 \cdot 10^{-36}$
GO: 0043283 (30.2%) biopolymer metabolic process (30.2%)	66% $7.2 \cdot 10^{-05}$	68% $2.8 \cdot 10^{-12}$	63% $1.8 \cdot 10^{-14}$	62% $1.3 \cdot 10^{-17}$	58% $1.4 \cdot 10^{-17}$	58% $6.0 \cdot 10^{-21}$	59% $3.5 \cdot 10^{-28}$	59% $1.2 \cdot 10^{-31}$	59% $2.4 \cdot 10^{-36}$	57% $2.3 \cdot 10^{-30}$
GO: 0006139 (20.7%) nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (20.7%)	54% $8.3 \cdot 10^{-05}$	52% $1.8 \cdot 10^{-09}$	47% $1.1 \cdot 10^{-10}$	49% $4.3 \cdot 10^{-16}$	47% $3.9 \cdot 10^{-18}$	46% $2.1 \cdot 10^{-21}$	47% $7.5 \cdot 10^{-26}$	47% $2.2 \cdot 10^{-30}$	47% $5.4 \cdot 10^{-35}$	45% $4.9 \cdot 10^{-31}$
GO: 0016070 (14.2%) metabolic process (14.2%)	42% $5.6 \cdot 10^{-04}$	40% $8.0 \cdot 10^{-08}$	37% $2.1 \cdot 10^{-09}$	39% $3.5 \cdot 10^{-15}$	36% $9.2 \cdot 10^{-16}$	36% $6.7 \cdot 10^{-20}$	37% $6.7 \cdot 10^{-26}$	37% $2.0 \cdot 10^{-28}$	37% $3.2 \cdot 10^{-32}$	34% $8.9 \cdot 10^{-25}$
GO: 0044238 (44.0%) primary metabolic process (44.0%)		74% $4.6 \cdot 10^{-07}$	70% $4.6 \cdot 10^{-08}$	69% $8.4 \cdot 10^{-10}$	68% $4.7 \cdot 10^{-12}$	68% $2.3 \cdot 10^{-14}$	69% $1.8 \cdot 10^{-19}$	69% $8.0 \cdot 10^{-22}$	68% $3.1 \cdot 10^{-24}$	67% $5.0 \cdot 10^{-21}$

been on highly connected proteins, so called "hubs", and proteins with high-betweenness centrality, so called "bottle necks" [19–21]. We aim here to show that multimodularity feature of the proteins that is proposed here may also indicate protein essentiality in the network, especially considering the fact that the underlying module-identification method relies on both topological and functional information about proteins.

For this purpose, the method proposed here is compared with another related method. In previous work by Pržulj et al. (2004) [18], topologically important proteins are identified by using the most frequent "bottle neck" nodes [19]. The method starts from a tree of the shortest paths for each node v. Such tree consists of n_v nodes that are directly or indirectly connected to v. All nodes w from the tree, such that more than $n_v/4$ paths from v to other nodes meet at node w, are defined as "bottle necks". Pržulj et al. (2004) presented only the top ten most frequent "bottle neck" proteins, and stated that 70% of those are involved in supporting cellular structure and organisation. We here evaluate the annotations for different groups of proteins based on how often they appear in different modules (see Table 2). After each specific GO term in the first column, the total percentage of all proteins that are annotated with this term is given. It can be noticed that the percentage of proteins that are annotated with the chosen terms drops for the proteins with module frequency ≤1, with the exception of the term in the last row "primary metabolic process", which is the most common of all presented terms.

We also present a more systematic comparison between our protein groups, chosen based on their average occurrence in the modules, and the bottle neck proteins (see Table 3). The top 25 proteins obtained by our approach significantly share the term "ribonucleoproteins complex biogenesis and assembly", which is a child term of "cellular component organization and biogenesis". No significantly shared ontology terms appear in the corresponding set of bottle-neck proteins.

3.2. Topological Features of Multimodular Proteins

3.2.1. CORE Data Set. We started by investigating general properties of the data set by studying the relation between degree and betweenness centrality. Figure 2 shows degree k versus betweenness centrality plotted on algorithmic scale. The few highly connected nodes (hubs) in the PIN must have high betweenness values because there are many nodes directly and exclusively connected to these hubs and the shortest path between these nodes goes through these hubs. However, the low-connectivity nodes also exhibited a wide range of betweenness values in the yeast PIN.

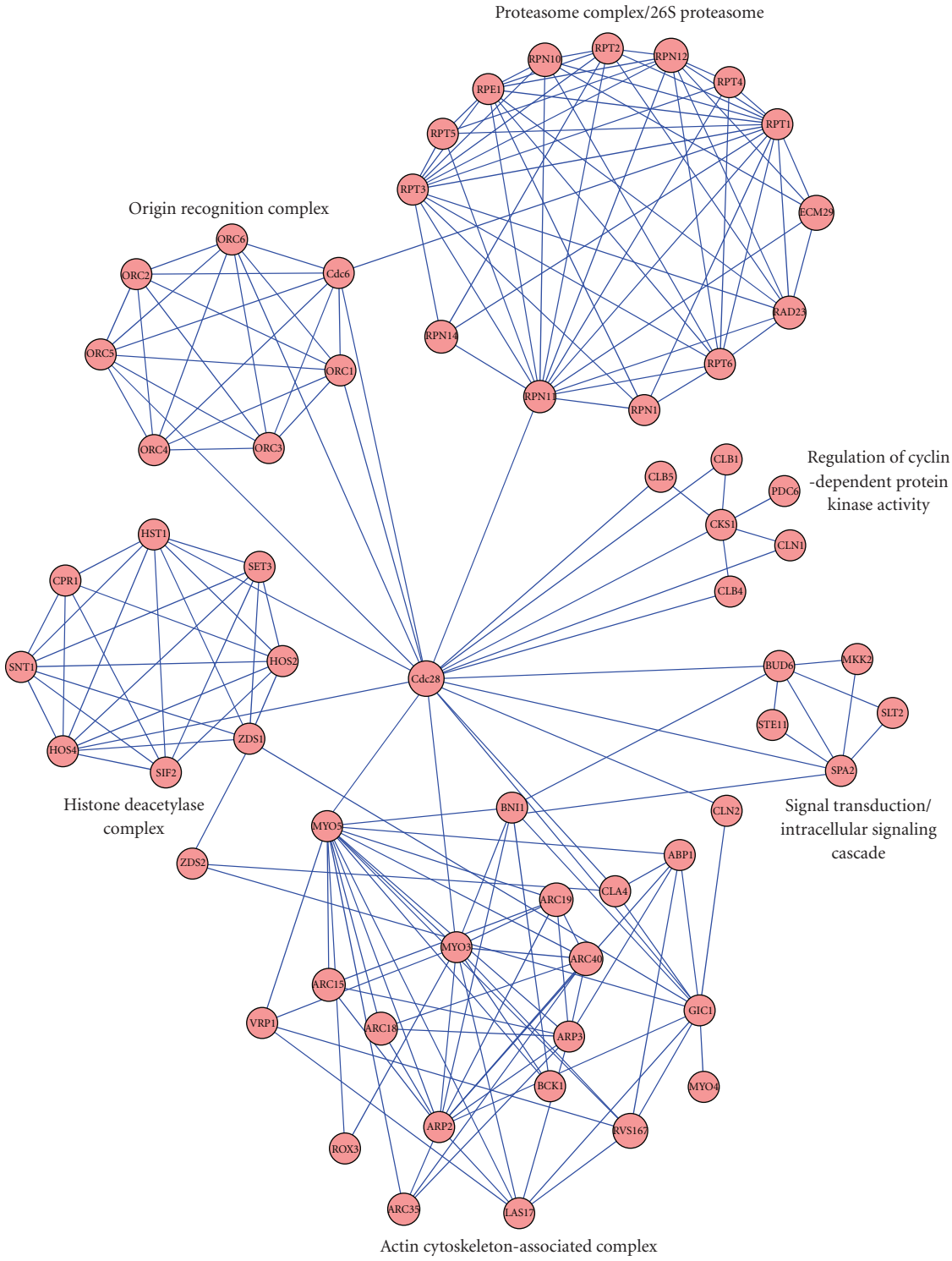

Figure 6: Modular network involving modules in which Cdc28.

In Figure 3, node betweenness centrality is plotted as a function of average number of module occurrences. We can notice that all proteins with average module frequency ≥ 2 have considerably high betweenness values. However, the single-modular nodes also exhibited a wide range of betweenness values in the yeast PIN.

3.2.2. Von Mering Data Set. We repeated the same experiment for the von Mering data set. In Figure 4, betweenness is plotted as a function of degree k. Here, we could not use any characteristic degree k or any interval of k values to denote the importance of nodes (based on the betweenness).

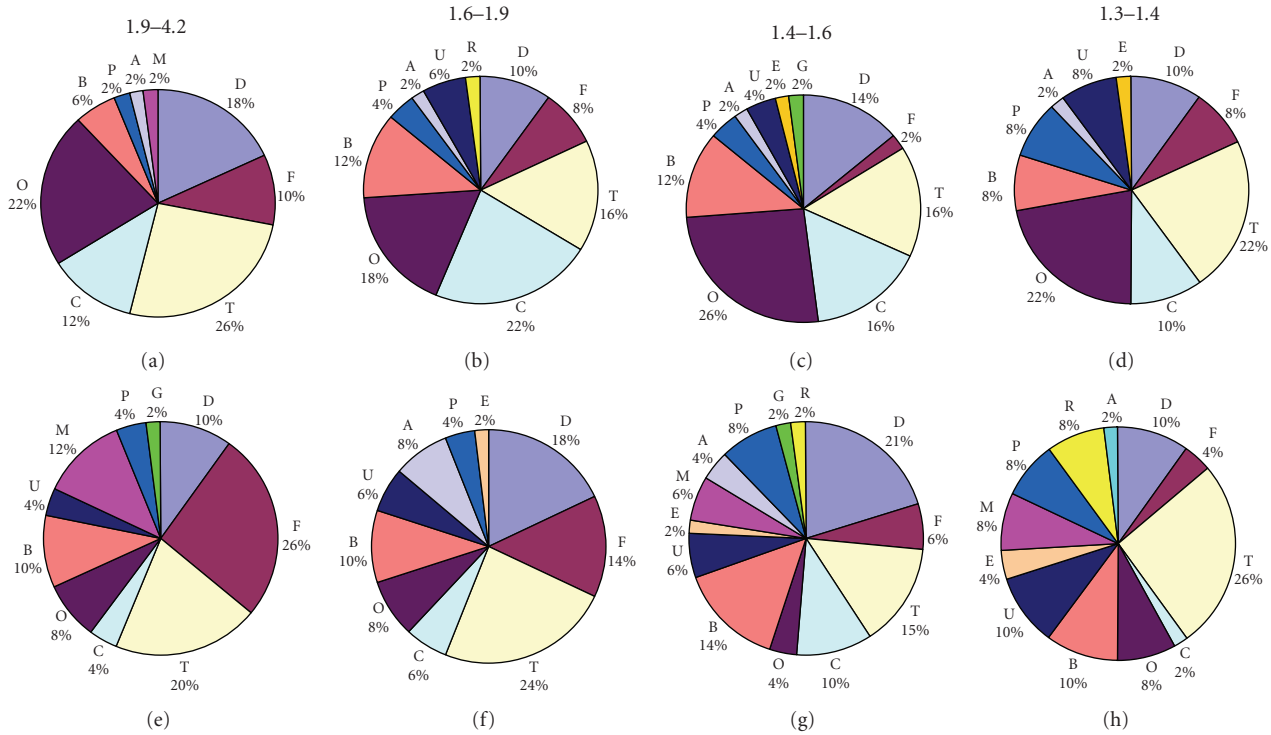

Figure 7: Functional groups statistics for proteins in von Mering data set. The first row shows charts with statistics for multimodular proteins (MMP) in varying intervals of module frequency (in decreasing order of frequency). There are 50 proteins in each interval. For comparison, the second row shows the corresponding statistics for the same number of single-modular proteins.

Also in Figure 5, besides the most frequent multimodular proteins (MMPs) that have high betweenness values, there is a wide range of betweenness centrality values for single-modular proteins (SMPs) as well. However, modular frequency seems to be a better indicator of node importance, in terms of betweenness centrality.

3.3. Lethality. There are 1015 lethal proteins obtained from manually curated MIPS database. The list of MMPs and SMPs observed across modules in both data sets was compared to the list of lethal proteins. The results from this comparison are presented in Tables 4 and 5.

In the CORE data set, we found 222 lethal proteins among the multimodular proteins (MMPs). This corresponds to 46.3%, as there are 480 frequently occurring proteins in total. The corresponding percentage for MMPs derived from modules in the von Mering data set is 68.7, as there are 57 lethal proteins among the 83 MMPs (see Table 4).

We made the same comparison for single-modular proteins (SMPs) across the modules based on both data sets. In the CORE data set, we found 173 lethal proteins among the SMPs, which correspond to 34.5%, as there are 502 SMPs in total (see Table 5). The corresponding percentage for the fraction of lethality in SMPs derived from modules in the von Mering data set is 54.5, as there are 116 lethal proteins among the 213 SMPs, as shown in Table 5.

In both cases, the difference is statistically significant at a 95% confidence level, meaning that there is a significantly larger proportion of lethal proteins, also referred to as important proteins, among multimodular proteins. These results are obtained by performing a z-test for the differences between the two proportions ($z = 3.8$ in the CORE data set, and $z = 2.2$ in the von Mering data set).

3.4. Modular Interconnectivity. Figure 6 shows the result from an example run from module-identifying method, where Cdc28 was predicted as taking part in six modules matching MIPS complexes. In addition, this protein occurs 830 times in 200 module sets and hereby has the highest average number of module occurrences (4.2). Cdc28 is a cyclin-dependent kinase and it is believed to be a key regulator of the cell-division cycle. In this example, it is connected to several proteins from Origin Recognition Complex (ORC), which is involved in DNA replication. Cdc28 is also connected to actin cytoskeleton-associated complex, which is reorganised in accordance with cell-cycle progression. This process is according to previous study believed to be controlled, directly or indirectly, by Cdc28 [22]. Furthermore, there is an important connection between Cdc28 and proteasome complex. The central role of this complex is to direct a cell to proceed with the decision to replicate itself. In yeast cells a critical trigger for cell replication is degradation of Sic1, which is a protein that inhibits the chemical activity of Cdc28. After eliminating

the biochemical Sic1 "brake" due to the action of SCF and the proteasome, the kinase is then free to trigger the progress toward DNA replication and associated events of cell replication.

This is a clear example of the network involving hub that interconnects several functional modules. This example is supported by several topological and functional features, such as average number of occurrences in modules, betweenness centrality, and node degree. However, there are several examples where those features are conflicting, which will be interesting to evaluate in future.

4. Conclusions

We have here presented approaches for identifying topologically and functionally important proteins by calculating the frequency of each protein across 200 sets of overlapping modules. Initial results show that the majority of frequently appearing proteins that connect several modules are involved in the assembly and arrangement of cell structures, such as the cell wall and cell envelope, which indicates that they are involved in supporting the cell structure rather than signal transduction, for example. We also observed by studying MIPS functional classes of the MMPs and SMPs that proteins involved in cellular organisation (O) appear more frequently among the top 100 MMPs, compared to the random sets of SMPs. The results from studying lethality show the significantly higher fraction of lethal proteins among multimodular proteins (MMP), when compared to single modular proteins (SMP) reflecting the tendency of MMP to be more lethal, and hereby indicating their essentiality.

The investigation of different features of so-called multimodular proteins, that is, proteins that take part in multiple modules within the PIN, shows that these may be involved in the assembly and arrangement of cell structures (according to GO annotation) to a greater extent than single-modular proteins or proteins with lower numbers of occurrences across the generated module sets. Also, the analysis of MIPS functional categories, along with the analysis of GO annotation, shows that the fraction of the proteins that belong to the category "cellular organisation" in multimodular proteins is higher than the fraction of such proteins in the single-modular groups of proteins. Another frequently occurring GO term that is assigned to multimodular proteins is "ribonucleoproteins complex biogenesis and assembly" which is a child term of "cellular component organisation and biogenesis". Hence, we find evidence supporting the hypothesis that this GO term reveals the role of modules in building and supporting higher-order structure(s) of the PIN organisation. Other features that we have analysed to characterise possible differences between multimodular and single-modular proteins are betweenness centrality and lethality. In both data sets, it is shown that there is significantly higher fraction of lethal proteins among multimodular proteins, also pointing at their significance. From the analysis of betweenness centrality, it is also notable that proteins with high average module frequency have considerably high betweenness values, while the single-modular nodes exhibit a wide range of betweenness values in the yeast PIN. This also points to the greater importance of the multimodular proteins, as those nodes may be potential bridges between modules in the network and have most influence on the information transfer between communicating modules. If a node with high betweenness centrality is removed, it may disconnect a different part of the network completely.

Possible limitation of this approach should finally be discussed. The method for assigning the weights to proteins, which are used for the purpose of module identification, that, in turn, consists the basis for identifying multimodular feature of the proteins, relies to a great extend on GO terms. Proteins may be annotated at different levels in the hierarchy, that is, some of more specifically described than the others. Another limitation that also should be discussed is that quality of GO annotation in terms of experimental evidence may vary. Currently, all evidence types are used, but some types of evidence such as "traceable author statement" are considered more reliable than others. As we used the protein-protein interactions that are validated by different method, and are generally well annotated it should not affect the performance of module identifying method to a great extent, but the method may benefit from future more fine grained versions of GO.

In future, it would be very interesting to make a systematic comparison with other module-identifying methods and other topological features used to identify essential proteins in protein interactions networks.

References

[1] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 3, pp. 1128–1133, 2003.

[2] J.-D. Han, N. Berlin, T. Hao, et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

[3] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks," *Proteins*, vol. 54, no. 1, pp. 49–57, 2004.

[4] A. A. Petti and G. M. Church, "A network of transcriptionally coordinated functional modules in *Saccharomyces cerevisiae*," *Genome Research*, vol. 15, no. 9, pp. 1298–1306, 2005.

[5] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761, supplement 1, pp. C47–C52, 1999.

[6] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[7] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12123–12128, 2003.

[8] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási, "Functional and topological characterization of protein interaction networks," *Proteomics*, vol. 4, no. 4, pp. 928–942, 2004.

[9] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting

proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, 2000.

[10] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[11] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[12] C. von Mering, R. Krause, B. Snel, et al., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[13] Z. Lubovac, J. Gamalielsson, and B. Olsson, "Combining functional and topological properties to identify core modules in protein interaction networks," *Proteins*, vol. 64, no. 4, pp. 948–959, 2006.

[14] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, article 2, 2003.

[15] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[16] H. W. Mewes, D. Frishman, and U. Güldener, "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 31–34, 2002.

[17] Z. Lubovac, D. Corne, J. Gamalielsson, and B. Olsson, "Weighted cohesiveness for identification of functional modules and their interconnectivity," in *Proceedings of the 1st International Conference on Bioinformatics Research and Development (BIRD '07)*, vol. 4414 of *Lecture Notes in Computer Science*, pp. 185–198, Berlin, Germany, March 2007.

[18] M. D. Mendenhall and A. E. Hodge, "Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*," *Microbiology and Molecular Biology Reviews*, vol. 62, no. 4, pp. 1191–1243, 1998.

[19] N. Pržulj, D. A. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, 2004.

[20] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics," *PLoS Computational Biology*, vol. 3, no. 4, article e59, 2007.

[21] C.-W. Hsu, H.-F. Juan, and H.-C. Huang, "Characterization of microRNA-regulated protein-protein interaction network," *Proteomics*, vol. 8, no. 10, pp. 1975–1979, 2008.

[22] H.-Y. Tang and M. Cai, "The EH-domain-containing protein Pan1 is required for normal organization of the actin cytoskeleton in *Saccharomyces cerevisiae*," *Molecular and Cellular Biology*, vol. 16, no. 9, pp. 4897–4914, 1996.