2005

The Architecture of a Proteomic Network in the Yeast

Emad Ramadan

Old Dominion University, erama001@gsa.odu.edu

Christopher Osgood

Old Dominion University, cosgood@odu.edu

Alex Pothern

Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/biology_fac_pubs

Part of the [Biology Commons](#), and the [Computer Sciences Commons](#)

Repository Citation

Ramadan, Emad; Osgood, Christopher; and Pothern, Alex, "The Architecture of a Proteomic Network in the Yeast" (2005). *Biological Sciences Faculty Publications*. 380.

https://digitalcommons.odu.edu/biology_fac_pubs/380

Original Publication Citation

Ramadan, E., Osgood, C., & Pothern, A. (2005). *The architecture of a proteomic network in the yeast*. Paper presented at the First International Symposium, CompLife 2005, Konstanz, Germany, September 25-27, 2005.

The Architecture of a Proteomic Network in the Yeast^{*}

Emad Ramadan¹, Christopher Osgood², and Alex Pothen¹

¹ Computer Science Department, Old Dominion University,
Norfolk, VA 23529, USA
{eramadan, pothen}@cs.odu.edu

² Biological Sciences Department, Old Dominion University,
Norfolk, VA 23529, USA
cosgood@odu.edu

Abstract. We describe an approach to clustering the yeast protein-protein interaction network in order to identify functional modules, groups of proteins forming multi-protein complexes accomplishing various functions in the cell. We have developed a clustering method that accounts for the small-world nature of the network. The algorithm makes use of the concept of k -cores in a graph, and employs recursive spectral clustering to compute the functional modules. The computed clusters are annotated using their protein memberships into known multi-protein complexes in the yeast. We also dissect the protein interaction network into a global subnetwork of hub proteins (connected to several clusters), and a local network consisting of cluster proteins.

1 Introduction

Systems biology involves the study of complex biological structures and processes by identifying their molecular components and the interactions among them. Looking across the evolutionary landscape, biological subsystems performing discrete functions are capable of being linked together in different ways without lethality to an organism, and often with positive gains in complexity and adaptation. Among the properties that are now recognized in multiple biological systems are: modularity (sets of semi-autonomous molecules that perform specific functions); robustness (the ability of biological systems to tolerate perturbations and noise); and emergence (new properties that emerge from the interaction of functional modules) [14].

One of the challenges in computational systems biology is to create tools that enable biologists to identify functional modules and the interactions among them from large-scale genomic and proteomic data. We report the results of a study on an organism-scale protein-protein interaction network in the yeast with the goal of identifying proteins that form functional modules, (i.e., multiple proteins involved that have identical or related biological function), by clustering techniques. Furthermore, we propose a hierarchical organization of the proteomic network.

^{*} Research supported by NSF grant CCR0306334, by subcontract B542604 from the Lawrence Livermore National Laboratory, and by a grant from the Office of Research at Old Dominion University.

Methods for clustering proteomic networks have to cope with several features specific to protein interaction data. High-throughput experiments such as the yeast 2-hybrid system and the tagged affinity purification (TAP) [1,15,16], have high error rates, nearing 50% in some instances. Proteomic networks are *modified power-law* networks and *small-world* networks [6]. That is, the distribution of the fraction of vertices with a given degree follows a modified power-law; and the average path length between vertices is of the order of $\ln n$ (or smaller), where n is the number of vertices in the network. Hence there is a large number of low degree proteins, and a significant number of high degree proteins. The latter make it harder to discover clusters in the data, while the former increase the computational requirements. Cluster validation is hampered by the fact that there is often little overlap between different experimental studies due to the limited coverage of the interactome [13]. Finally, the predicted clusters must be biologically significant: e.g., functionally homogeneous.

In spite of these difficulties, we believe that we have successfully clustered a yeast proteomic network, with the predicted clusters overlapping well with multi-protein complexes and organelles. Our approach is based on identifying *hub proteins*, proteins that connect to a large number of clusters, and low-shell proteins (defined in the next section), and clustering the residual network. Low-shell proteins can be added to the cluster network at a later stage. We validate the clusters by comparing the clusters against experimental data on multi-protein complexes.

The hub proteins carry interesting information about the architecture of proteomic network, and are organized into a subnetwork of their own. Thus we propose a two level architecture for the yeast proteomic network, consisting of a global subnetwork of hubs, and a local subnetwork of clusters and low-shell proteins. A schematic of this architecture is shown in Fig. 1, where the top level corresponds to the global hub network, and the lower level corresponds to the local cluster network.

Fig. 1. A schematic representation of the yeast proteomic network as a hub-cluster interaction network. The top level corresponds to a global network of hub proteins, and the bottom level to a local network of cluster proteins.

2 Materials and Methods

2.1 k -Cores and k -Shells in Graphs

We begin by describing the concepts of a k -core and a k -shell in a graph, since our clustering method makes use of these.

Given a natural number k , the k -core of a graph G is the maximal subgraph of G in which every vertex has degree at least k in the subgraph (provided it is not the empty graph). The k -cores in a graph are nested: the $(k + 1)$ -core is contained in the k -core, for $k = 0, 1, \dots, K - 1$, where K is the value of the maximum core in the graph. The k -core of a graph need not be a connected subgraph even if the original graph is connected. Note that if a graph contains a k -vertex connected component or a clique on $k + 1$ vertices, then it is contained in a k -core; however, the k -core need not contain a k -connected subgraph or a clique on $k + 1$ vertices.

The k -shell of a graph is the set of vertices that belong to the k -core, but not to the $(k + 1)$ -core. The k -shell includes vertices with degree k from the k -core, but also other vertices whose degree in the residual graph becomes less than $(k + 1)$ when low degree vertices are removed.

There is a well-known linear-time algorithm, in the number of edges, for computing the k -core (indeed, for finding all k -cores, for $k = 0$ to the maximum core value) of a graph. The idea is to repeatedly remove vertices v of degree less than k from the graph and all edges incident on v , updating the degrees of the neighbors of v in the residual graph as edges are deleted. The algorithm repeats this step until all vertices that remain have degree k or higher in the residual subgraph.

We have extended the concept of a k -core to a hypergraph in earlier work [19]. k -cores have been used earlier for clustering proteomic networks as a way of identifying highly connected subnetworks and for removing proteins belonging to low shell values [4].

We claim that clustering the k -core of a network removes noise in the data, in the same spirit as computing a shared nearest neighbor similarity (SNN) network. In an SNN network, two vertices are joined by an edge with weight equal to the number of their common neighbors at a distance less than or equal to d , where d is a natural number parameter. The SNN network includes only those edges that have weight higher than a threshold, and clustering algorithms have been designed to work with this network [17]. Unfortunately for large networks of small average path lengths, the computation of the SNN network can be prohibitively expensive. We suggest that the k -core is an efficient way to compute a network that approximates an SNN network. Every vertex in a k -core is adjacent to at least k other vertices in the subgraph, each of which is adjacent to k vertices with high core values.

2.2 Clustering Algorithms

Three major clustering approaches have been employed to identify functional modules in proteomic networks. The first approach searches for subgraphs with specified connectivities, called network motifs, and characterizes these as functional modules or parts of them. A complete subgraph (clique) is one such candidate, but other network

motifs on small numbers of vertices have been identified through exhaustive searching or statistical methods [21]. This approach is not scalable for finding larger clusters in large-scale networks. The second approach, recently proposed in this context by Bader and Hogue [4], computes a weight for each vertex (depending on the density of a maximum core in the neighborhood of the vertex); it then grows a cluster around a seed vertex, a vertex with the largest weight in the currently unclustered graph. A vertex in the neighborhood of a cluster is added to it as long as its weight is close (within a threshold) to the weight of the seed vertex. Once a cluster has been identified, the procedure is repeated with a vertex of largest weight that currently does not belong to a cluster as the seed vertex. However, our experience comparing this approach with the spectral algorithms that we describe next shows that this method is less stable than the latter (i.e., the clusters depend on the seed vertices chosen).

We now discuss a spectral algorithm for clustering.

Let $G = (V, E, W)$ denote a weighted graph with vertex set V , edge set E , and weights on the edges W . Consider the problem of partitioning V into two sets $V_1 \cup V_2$. We consider the weights

$$W_{il} \equiv W(V_i, V_l) = \sum_{j \in V_i, k \in V_l, (j,k) \in E} w_{jk},$$

where $i, l = 1, 2$. Minimizing the objective function

$$J(V_1, V_2) = \frac{W_{12}}{W_{11}} + \frac{W_{12}}{W_{22}}$$

minimizes the sum of weights of the edges between distinct clusters, while simultaneously maximizing the sum of the weights of the edges within each cluster. This objective function for clustering has been called the MinMaxCut [10], and it measures a ratio related to the *separability* of a cluster to its *cohesion*. We prefer this function to related objective functions that have been proposed such as Normalized Cut.

Let Q denote the Laplacian matrix of a graph with weights w_{ij} on its edges (i, j) ; thus $q_{ij} = -w_{ij}$ for $i \neq j$, and each diagonal element q_{ii} is the sum of the weights of the edges incident on the vertex i . Let D be a diagonal matrix with its i -th component $d_{ii} = \sum_{(i,j) \in E} w_{ij}$; $d_1 = \sum_{i \in V_1} d_{ii}$, and $d_2 = \sum_{i \in V_2} d_{ii}$. Let p be a ‘generalized partition vector’ with $p_i = \sqrt{d_2/d_1}$ for $i \in V_1$; and $p_i = -\sqrt{d_1/d_2}$ for $i \in V_2$; let e be the n -vector of all ones. Then we have $p^T D e = 0$, and $p^T D p = d_1 + d_2$. Ding et al. [9] have shown that

$$\min_{V_1, V_2} J(V_1, V_2)$$

is equivalent to

$$\min_p p^T Q p / p^T D p, \quad \text{subject to} \quad p^T D e = 0.$$

This minimization problem is NP-hard since the generalized partition vector p is restricted to have elements from one of two values. However, we can relax this constraint and let p take values from $[-1, +1]$ to obtain an approximate solution. This problem is solved by the eigenvector x corresponding to the smallest positive eigenvalue of the generalized eigenproblem $Qx = \lambda Dx$.

The partition is obtained by choosing the vertices in one part to consist of vertices with eigenvector components smaller than a threshold value, while the other part has the remaining vertices. The threshold value could be chosen so as to locally minimize the MinMaxCut objective function. For details, see [8,10].

A clustering method is obtained by recursively applying the spectral partitioning method, by splitting each current cluster into two subclusters. The MinMaxCut objective function can be used to determine if a given cluster should be split further.

2.3 Algorithm

The yeast protein interaction network under study has 2610 proteins and 6236 interactions; we work with its largest connected component, which has 2406 proteins and 6117 interactions.

In the first step, we separate the high degree proteins, which are candidates for *hub proteins*. A hub is a protein that connects several different clusters in the network together, and these form a subset of the high degree proteins. After some experimentation, we chose candidate hub proteins to be those with degree 15 or higher in the network we study. The residual network has 2241 proteins and 3057 interactions, and consists of 397 connected components. The largest connected component of the residual graph has 1773 proteins and 2974 interactions (and hence most of the other components have few or no edges). We chose the largest component for further analysis.

In the second step, we compute the 3-core of the residual graph in order to remove the low- shell proteins (the 0-, 1-, and 2-shells) from the network. As discussed earlier, we believe that this step removes some of the noise from the experimental protein interaction data. We have found that this step has two advantages. First, the clustering algorithms generate better clusters of the residual network; the low shell proteins can be assigned to a cluster after it has been identified. Second, this step reduces the graph size substantially since this is a modified power law network with a large number of low degree proteins.

In the third step, we have applied the spectral clustering recursively to cluster the subgraph and identify the clusters, employing the MinMaxCut objective function. Once the clusters are identified, then the high-degree proteins which were removed as candidate hub proteins can be confirmed as hub proteins if they connect multiple clusters, or can be included among the cluster proteins.

Our spectral clustering code is currently written in Matlab for quick prototyping. The current code takes 65 seconds on a PC with a 1.3 MHz Intel processor and 768 MB memory. The hub and k -core computations are faster. Here we have greatly reduced the run times needed by removing the low-shell and hub proteins before clustering.

We have been concerned in this paper with identifying a methodology that can successfully deliver biologically significant clusters in proteomic networks. Distributed computations will be needed when we consider larger proteomic networks such as the human, and networks consisting of heterogeneous data.

We are also concerned with scalable clustering algorithms. The proposed approach requires $O(|E| \log |V|)$ time, where $|E|$ is number of edges in the network, and $|V|$ is the number of vertices. The k -core computation and the eigenvector computation at each clustering step can be performed in time $O(|E|)$; and there are $\log |V|$ partitioning

steps needed to cluster. The spectral clustering could be replaced with a multi-level clustering approach that can also be implemented in time $O(|E|)$.

3 Results

3.1 Data Source and Analysis

Among the protein interactions produced by high-throughput methods such as the yeast 2-hybrid experiment or tagged affinity purification (TAP) [1,15,16], there are many false positives due to experimental limitations as well as biological factors (proteins that are not expressed at the same time or in the same cellular locale) [13]. In order to reduce the interference by false positives, we focused on the protein interaction network from the Database of Interacting Proteins (DIP), circa. April 2004 (URL: dip.doe-mbi.ucla.edu/dip/), consisting of the reliable dataset, which includes only data determined by a small-scale experiment, confirmed by independent high-throughput experiments, or scored highly by a probabilistic method that estimates the reliability of an interaction. This dataset has 2610 proteins that involve 6236 interactions considered to be reliable with high confidence.

3.2 The Cluster and Hub Networks

The local network computed by the clustering algorithm on the yeast protein interaction network, from which high degree proteins (hubs) and low-shell proteins have been removed, is shown in Fig. 2. Colors are used to distinguish the proteins belonging to a cluster, although some colors are reused to color proteins belonging to clusters that are drawn sufficiently far from each other. Thirty-eight clusters are displayed; for clearer presentation, we have omitted the edges joining two clusters when fewer than three edges join a cluster to another. All edges joining proteins within each cluster are shown.

The sum of the numbers of within-cluster edges is 984, while the sum of the between-cluster edges is 239, and the largest number of edges joining one cluster to another is 9. These measures are related to the concepts of *cohesion* and *separation* of the clustering [22], and thus we believe that our method has been able to cluster the residual network well. Each of the clusters is assigned to multi-protein complexes using the Munich Information Center for Protein Sequences (MIPS) database (URL: mips.gsf.de), as described in the next subsection. Each low-shell protein can now be easily assigned to a cluster with whose proteins it has the most number of interactions.

From a topological point of view, our approach to clustering helps to uncover the hidden topological structure of a proteomic network. We found that there are two major subnetworks within the protein-protein interaction network. In addition to the *cluster network*, we also construct a *hub network*, the subnetwork formed by the hub proteins in the protein interaction network; a subnetwork formed by the 5-core of the hub network is shown in Fig. 3. Four ‘super-clusters’ are clearly evident in the hub interaction network: from top to bottom, these correspond to the spliceosome, proteins involved in mRNA export and the nuclear pore complex, the regulatory subunit of the proteasome, and proteins that are transcription factors.

We now consider various subnetworks of the yeast protein interaction network to illustrate the differences between the ‘global’ hub network, and the local ‘cluster’ network. Table 1 lists the sizes of these networks, the average path lengths, the diameters, and the cluster coefficients. (The cluster coefficient measures how likely two neighbors of a vertex are to be adjacent to each other in the network, on the average.) The row ‘C + S’ denotes the ‘cluster and shell’ subnetwork obtained by removing the hub proteins from the whole network. Note that this subnetwork has the highest diameter and average path length, due to the presence of the large number of low-shell proteins. Once they are removed, the cluster network exhibits the highest clustering coefficient, supporting our premise that this is a local network. The hub network has the lowest diameter and average path length due to the edges joining the hub proteins to each other (cf. Fig. 3). The tight clustering seen in the hub network was surprising to us, but it is clear that hub proteins preferentially interact with cluster proteins and with each other, rather than the low shell proteins. We discuss the hub subnetwork and clusters in it in more detail in the next subsection.

The average path lengths in these networks are compared against $\ln n$, where n is the number of vertices in each subnetwork. Random power-law networks with exponent

Fig. 2. Clusters in the yeast proteomic network from which hub and low-shell proteins have been removed. When fewer than three edges join a pair of clusters, such edges have not been drawn in this figure, for clarity in presentation.

Table 1. Properties of various subnetworks of the yeast protein interaction network

Subnetwork	No. of vertices edges		Average Path Length ($\ln n$)	Diameter	Cluster coefficient
Hub	165	507	3.5 (5.1)	7	0.37
Cluster	495	1223	6.5 (6.2)	16	0.43
C +S	1773	2974	7.6 (7.5)	19	0.15
Whole	2406	6117	5.1 (7.8)	13	0.21

β satisfying $2 < \beta < 3$ have expected average path lengths of order $\ln \ln n$, while if the exponent $\beta > 3$, it is $\ln n$ [7]. We see that $\ln n$ is a good approximation for the average path length of the cluster and cluster-shell networks; but any network that includes the hub proteins has an even lower average path length.

3.3 Functional Annotation of Clusters

One way to validate the clusters we discovered is to check how homogeneous the proteins in each cluster are with respect to function or the biological process that they are involved in. Each cluster should consist of one or more multi-protein complexes, which are molecular machines responsible for various cellular functions. We compared 38 clusters that we found with multi-protein complexes listed in the MIPS database. We found that in thirteen of the MIPS protein complexes, every protein in the complex was also identified in a cluster corresponding to it; for nine more complexes, we found more than half the proteins involved in the complex in a corresponding cluster. These results are despite the facts that hub and low-shell proteins are not included in this comparison,

Fig. 3. The 5-core of the global hub network. The four clusters evident in this figure correspond to the spliceosome, mRNA export, the proteasome, and various transcription factors.

and that many proteins in the MIPS database are not included in the DIP protein network under study here. When the hubs and low-shell proteins are included, the coverage will increase further. A table containing the number of each cluster, a corresponding MIPS complex name and its MIPS ID, the number of proteins the cluster and the complex have in common, and the names of these proteins, is included in file `table1.xls`, see the Supplementary Materials at www.cs.odu.edu/~pothen/Papers/Cluster/DIP/.

We should note that, in general, the clusters that we have discovered contain more proteins than those reported to belong to a corresponding MIPS complex. This suggests possible biological roles and functional assignments for such proteins, many of which are not currently functionally annotated.

The protein interaction graph of each cluster and a biological process annotation for it, using a directed acyclic subgraph (DAG) derived from the Gene Ontology (URL: www.geneontology.org), are also included in the Supplementary Materials. The cluster subgraphs are included in the `jpg` files, while GO DAGs are listed in the `png` files. While some of the cluster graphs are near-cliques or subgraphs with high edge connectivity, many of them are not. We believe that this validates our approach of finding complexes by a general clustering approach rather than searching for specific subgraph motifs.

3.4 Interactions Between the Hub and the Local Networks

We now consider the hub protein subnetwork and its interaction with the local network in more detail.

One of the complexes in cluster 8, the U4/U6 x U5 tri-snRNP complex, (listed in the file `table1.xls` in the Supplementary Materials), is comprised of a group of proteins involved in spliceosome processing of mRNA. This is the top-most cluster represented in Fig. 3. The spliceosome is required for the ordered and accurate removal of intronic sequences from pre-mRNA and thus plays a key role in alternative splicing, a process of great importance in higher eukaryotes whereby a single gene can generate multiple transcripts (alternatively spliced mRNAs) and thus multiple proteins [12]. The PRP (pre-mRNA processing) and Sm family proteins make up most of the proteins found in cluster 8. Some of the key proteins involved in mRNA processing, including those belonging to the LSM family, are not found in that cluster, but among the hub proteins that interact with multiple clusters.

One of the complexes in cluster 24, the first mRNA cleavage factor complex (represented in the file `c124.jpg` in the Supplementary Materials), includes five proteins involved in mRNA cleavage in preparation for the addition of the eukaryotic signature poly-A tail. Thus, proteins including CLP1 (involved in cleavage of the 3' end of the mRNA prior to tailing), and RNA 14 and 15 (two proteins that participate with CLP1 in formation of the 3' end of mRNA), are collectively implicated in alternative selection of the poly-A addition site [18].

Now we focus our attention on the single protein-protein edge which joins the top-most hub cluster in the figure, corresponding to *mRNA splicing*, to a second hub cluster involving mRNA export and nuclear pore formation proteins, corresponding to *mRNA export*, in Fig. 3. The two hub proteins that form the bridge between these clusters are

PRP6, a component of the mRNA splicing machinery, and PAB1, the poly-A binding protein involved in the final step in mRNA processing. We note that PRP6 is involved in the later stages of mRNA splicing and is in that sense the penultimate step prior to poly-A tailing. Thus, the overall logic of joining these two complexes by these particular hub proteins is compelling.

We now examine the connections formed by these two hub proteins with the local clusters that we picture as lying below them in the hierarchy of global (hub) and local (cluster) networks (see Fig. 1). PRP6 interacts with a single cluster (cluster 8) through the protein SMD1. SMD1 further interacts with splicing proteins PRP3 and SMD3 in the hub complex that includes PRP6. SMD1 is involved in the early stages of mRNA splicing and is highly conserved, showing greater than 40% amino acid identity between yeast and human [20]. PRP6 interacts with PAB1 in the second hub complex. PAB1 in turn interacts with three clusters, 22, 24 and 34. As noted above, cluster 24 includes RNA14 and RNA15, both involved in mRNA cleavage, and it is these proteins that interact with PAB1. PAB1 also forms connections with cluster 22 (via its interaction with TIF4632 = eIF4F, file c122.jpg), and with cluster 34 (via PKC1, file c134.jpg). These latter interactions (eIF4F and PKC1) are at first glance puzzling, but in fact they are entirely consistent with emerging evidence of interactions and regulatory loops that exist between distinct components in the gene expression machinery. The poly-A terminus of mRNA, and the associated PAB1, not only interacts with the 5' end of the mRNA, ensuring structural integrity of the transcript prior to its participation in protein translation [5], but the PAB1 terminus also interacts with the translation machinery itself, and specifically with eIF4 initiation factors. Finally, PKC1, a protein kinase crucial to cell signaling pathways, is also implicated in functional interactions with PAB1 and eIFs [3], suggesting global level regulation of protein synthesis from metabolites through mRNA processing.

At the global level of our network model, we find key proteins involved in rate-limiting steps of gene expression, linked in logical order; these are connected to the local network consisting of clusters of proteins involved in execution level functions. Whether this overall pattern is typical of the proteome organizational structure we have identified here, remains to be further investigated.

3.5 Incorporation of Protein Domain Data

Proteins interact with each other through regions that have a specific sequence and fold, called domains. Here we further validate the protein complexes predicted from our clustering approach using information on the domain structure of proteins.

The study of proteins involved in processing eukaryotic mRNAs indicate that virtually all steps involved in gene expression are coordinated and integrated via protein-protein interactions. The LSM proteins provide an informative example of the integration of cellular protein machinery to couple synthesis and quality control in gene expression [2,11]. LSM proteins form heptameric complexes that bind to RNA molecules; one such complex is found primarily in the nucleus where it coordinates splicing of mRNAs, while a second, related, complex of LSM proteins assembles in the cytoplasm to monitor mRNA quality control. LSM proteins have been extensively characterized and include two highly conserved protein interaction domains, SM1 and SM2.

It is proposed that these conserved domains permit each LSM protein to interact with two other LSM proteins in forming the heptameric, doughnut shaped ring structure that is implicated in mRNA splicing. LSM proteins comprise a gene family in which successive rounds of gene duplication have increased LSM copy numbers. LSM proteins have also been shown to form stable interactions with other protein types, including the PRP proteins discussed below.

The PRP proteins similarly carry a protein-protein interaction motif, the tetratricopeptide repeat (TPR) [23]. PRP proteins typically contain multiple copies of the 34 amino acid repeat; Prp1 for example contains 19 repeats of the TPR (ibid). Some PRP proteins contain a second conserved site at the C-terminus of the protein that facilitates interactions between them and LSM proteins, thus coupling two complexes with key roles in mRNA splicing. Our hub network predicts the formation of a complex containing LSM proteins 1-5, 7 and 8, as well as PRP proteins 4, 6, 8, and 31. The specific interactions implicated in our sub-network are, to our knowledge, the first explicit assignments of interactions between these two families of proteins.

We were surprised to find that the hub proteins form a highly interconnected sub-network. Biological evidence indicates that LSM proteins do indeed form multi-protein complexes in the course of performing their key cellular functions. The fact that each LSM protein has at least two protein-protein interaction domains helps us understand how the complexes are formed. Whether similar binding interactions can account for other closely knit hub networks is under investigation.

4 Conclusions

We have proposed a two-level architecture for a yeast protein-protein interaction network. We place a small set of hub proteins, each with at least fifteen interaction partners and involved in gene expression, mRNA export, the proteasome, and transcription factors, into a global subnetwork. A local subnetwork of proteins is organized into clusters that correspond well with multi-protein complexes in the MIPS database. We used the computed clustering to examine the biological significance of some of the interactions observed between the hub and local subnetworks. If the proposed two-level architecture exists in other proteomic networks, then it would be interesting to discover properties that distinguish hub proteins from the proteins in the local network.

In future work, we will consider the computation of an overlapping clustering rather than the exclusive clustering approach considered in this paper, so that a protein could be included in more than one cluster in the local network. We will also investigate additional clustering approaches and biological networks involving heterogeneous data.

References

1. A-C. Gavine, M. Bosche, and R. Krause et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.
2. T. Achsel et al. The Sm domain is an ancient RNA-binding motif with oligo(U) specificity. *Procs. Natl. Acad. Sci.*, 98:3685–3689, 2001.
3. F. Angenstein et al. A receptor for activated C kinase is part of messenger ribonucleoprotein complexes associated with polyA-mRNA in neurons. *J. Neurosci.*, 22:8827–37, 2002.

4. G. D. Bader and C. W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(2):27 pp., 2003.
5. X. Bi and D. Goss. Wheat germ poly (A)-binding protein increases ATPase and the RNA helicase activity of translation initiation factors eIF4A, eIF4B and eIF-iso-4F. *J. Biol. Chem.*, 275:17740–6, 2000.
6. S. Bornholdt and H. G. Schuster, editors. *Handbook of Graphs and Networks*. Wiley VCH, 2003.
7. F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Procs. Natl. Acad. Sci.*, 99(25):15879–15882, 2002.
8. I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Procs. ACM Internatl. Conf. Knowledge Discovery in Data Mining (KDD)*, 2001.
9. C. Ding, X. He, R. F. Meraz, and S. R. Holbrook. A unified representation of multi-protein complex data for modeling interaction networks. *Proteins: Structure, Function, and Genetics*, 2004. To appear.
10. C. Ding, X. He, H. Zha, M. Gu, and H. Simon. A MinMaxCut spectral method for data clustering and graph partitioning. In *Procs. IEEE Internatl. Conf. Data Mining (ICDM)*, pages 107–114, 2001.
11. M. Fromont-Racine et al. Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast*, 17:95–110, 2000.
12. F. Galisson and P. Legrain. The biochemical defects of PRP4-1 and PRP6-1 yeast splicing mutants reveal that the PRP6 protein is required for the accumulation of the [U4/U6.U5] tri-snRNP. *Nucl. Acids Res.*, 21:1555–62, 1993.
13. J. D. J. Han, D. Dupuy, N. Bertin, et al. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23:839–844, 2005.
14. L. H. Hartwell, J. J. Hopfeld, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.
15. Y. Ho, A. Gruhler, and A. Heilbut et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415:180–193, 2002.
16. T. Ito, T. Chiba, and R. Ozawa et al. A comprehensive two hybrid analysis to explore the yeast protein interactome. *Procs. Natl. Acad. Sci.*, 98:4569–4574, 2001.
17. A. A. Jarvis and E. A. Patrick. Clustering based on a similarity measure based on shared nearest neighbors. *IEEE Trans. Computers*, C-22:1025–1034, 1973.
18. C. Noble et al. Rna14-Rna15 assembly mediates the RNA-binding capability of *Saccharomyces cerevisiae* cleavage factor IA. *Nucl. Acids Res.*, 32:3364–75, 2004.
19. E. Ramadan, A. Tarafdar, and A. Pothén. A hypergraph model for the yeast protein complex network. In *Procs. Workshop High Performance Computational Biology (HICOMB)*. IEEE / ACM, 2004. 8 pp., (CDROM).
20. B. C. Rymond. Convergent transcripts of the yeast PRP38-SMD1 locus encode two essential splicing factors, including the D1 core polypeptide of small nuclear ribonucleoprotein particles. *Procs. Natl. Acad. Sci.*, 90:848–52, 1993.
21. V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Procs. Natl. Acad. Sci.*, 100:12123–12128, 2003.
22. P. Tan, M. Steinbach, and V. Kumar. *Introduction to Datamining*. Addison Wesley, 2005.
23. S. Urushiyama et al. The prp1+ gene required for pre-mRNA splicing in *Schizosaccharomyces pombe* encodes a protein that contains TPR motifs and is similar to Prp6p of budding yeast. *Genetics*, 147:101–115, 1997.