A Unified Representation of Multi-Protein Complex Data

Chris Ding, Xiaofeng He, Richard F. Meraz, Stephen R. Holbrook
Computing Research Division and Physical Biosciences Division
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
{chqding,xhe,rfmeraz,srholbrook}@lbl.gov

Abstract

The protein interaction network presents one perspective for understanding cellular processes. Recent experiments employing high-throughput mass-spectrometric characterizations have resulted in large datasets of physiologically relevant multi-protein complexes. We present a dual representation of such datasets based on an underlying bipartite graph model that is an advance on existing models of the network where the connections between proteins are uniformly weighted. Our dual representation allows for additional weighting of connections between proteins shared in more than one complex as well as addressing the higher level of organization that occurs when the network is viewed as consisting of protein complexes that share components. The dual representation also allows for the application of the rigorous MinMaxCut graph clustering algorithm for the determination of relevant protein modules in the networks. Statistically significant annotations of clusters in the network using terms from the Gene Ontology suggest that this method might be useful for posing hypothesis about uncharacterized components of protein complexes or uncharacterized relationships between protein complexes.

Introduction

Proteins carry out most essential cellular processes in complex multi-protein assemblies. These protein complexes perform activities needed for metabolism, communication, growth and structure. A systematic identification, characterization and understanding of these molecular machines of life will provide an essential knowledge base and link proteome dynamics and architecture to cellular function and phenotype. A variety of experimental and computational approaches have been employed to deduce the constituents of protein macromolecular complexes Experimental approaches such as the yeast two-hybrid genetic screen yield binary interaction data while more recent high throughput methods combine tagged "bait" proteins and protein-complex purification schemes with mass spectrometric measurements to yield physiologically relevant data on intact multi-protein complexes (Schwikowski et al., 2000; Ho et al., 2002; Gavin et al., 2002). Taken together, data from these experiments approximate the network of interactions between proteins and protein complexes that govern most cellular processes.

An important issue is the effective representation of the functional relationships between various parts of the interaction network (Alm and Arkin, 2003). So far most studies have represented protein interaction data as a map of binary interactions with uniformly weighted connections between interacting proteins (Bader and Hogue, 2002). For multi-protein complex data, this binary model assumes a pairwise interaction between all constituents in a complex. This equal weighting, however, is an oversimplification since physical interactions between constituents cannot be unambiguously described for all complexes without rigorous structural analysis. Some efforts have been made to move beyond the binary interaction model. The "spoke" model (Bader and Hogue, 2002) assumes pairwise interactions only between the purification "bait"

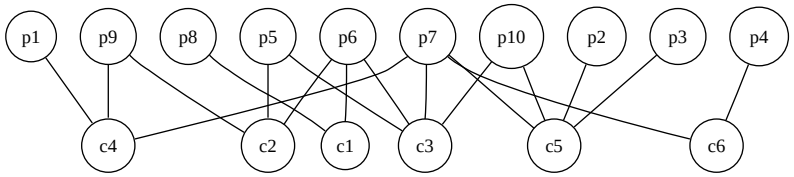

Figure 1: A bipartite graph representation of a hypothetical protein-complex dataset. The p-nodes represent proteins and c-nodes represent experimentally-determined protein complexes. An edge between a p-node and a c-node indicates that the protein is contained in the protein complex.

and proteins that co-purify in the complex. A hypergraph model (Pothen, 2003) allows protein to connect to more than one protein.

The most important limitation of existing models of the protein interaction network is their inability to represent a higher order organization of the proteome that results from the consideration of network relationships between protein complexes. A recent review by Gavin and Superti-Furga discusses the major issues concerning protein complexes and proteome organization and gives several examples of the modularity of protein complexes and their ability to share components and interact in complex cellular processes (Gavin and Superti-Furga, 2003). A model of the protein interaction network that adequately deals with relationships between protein complexes would be an important step toward a framework for a systems-level understanding of cellular processes.

A Bipartite Graph Model of Protein Complex Data

In this paper we propose a novel representation of multi-protein complex data that treats proteins and protein complexes on equal footing.. This representation emphasizes the "duality" of the relationship: a protein complex is characterized by its constituent proteins, while the interaction between two proteins can be gauged by the protein complexes that contain these proteins. This duality is best captured by a bipartite graph (Figure 1) specified by an adjacency matrix B, in which a protein-complex is represented by a column and a protein is represented by a row.

This bipartite representation of multi-protein complex dataset leads to a coherent framework for interaction networks: (1) The protein-protein (p-p) interaction network arises naturally. If we define the interaction strength between two proteins as the number of complexes that contain the two proteins, this interaction strength is given precisely by the adjacency matrix BB^T. (2) More importantly, the protein complex - protein complex (c-c) interaction network arises naturally. If we define the interaction strength between two protein complexes as the number of common proteins shared between them, then this interaction strength is precisely given by the adjacency matrix B^TB. (see the *Methods* Section for more details.) This framework overcomes the shortcomings in previous work: (a) The c-c interaction network yields a higher level organization of cellular processes. (b) The interaction strength of connections in the network is more realistic than simple uniform weighting.

The more realistic interaction strength of network connections from our dual representation allows for the application of a rigorous graph clustering algorithm (MinMaxCut) which has been shown to be successful with difficult datasets (Ding, 2002). The goal of clustering the protein interaction network is to determine its component modules, their functional annotations and some notion of the relationships between them. A module in a biological network is loosely defined as a functional unit separable from the rest of the network. In this context the use of the terms modules and computationally discovered clusters is interchangeable. Our hypothesis is that computationally discovered modules would encompass proteins

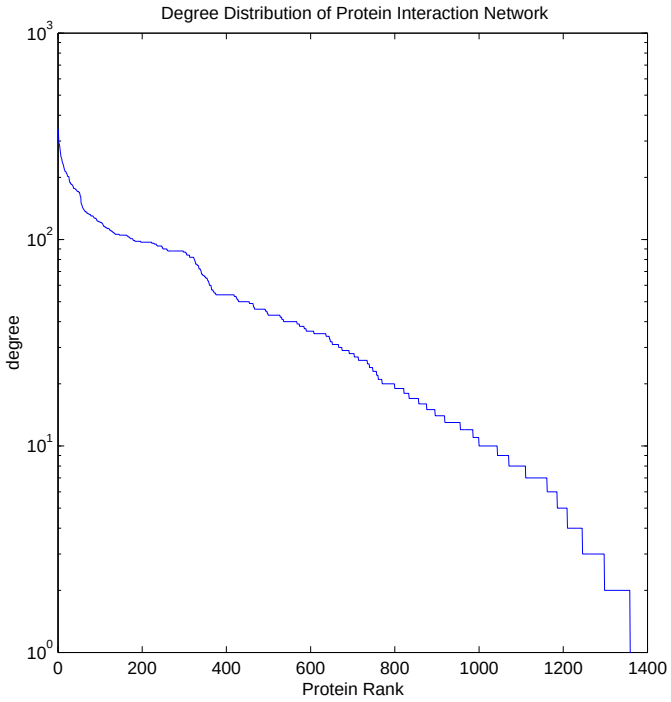

Figure 2: Distribution of the degree (number of proteins a given protein interacts with) in the protein-protein interaction network. This curve approximates a power-law distribution indicating that it is a scale-free network topology.

related through physical, and possibly temporal, associations in functionally coincident macromolecular complexes (p-p network), or define more diverse relationships of cellular process between functionally related protein complexes (c-c network).

Results and Discussion

Multi-Protein Complex Dataset

Two datasets summarizing high-throughput analysis of multi-protein complexes are available for the yeast *Saccharomyces cerevisiae* (Gavin et al., 2002; Ho et al., 2002). Coupling different purification (immuno-precipitation and tandem affinity purification (TAP)) and labeling schemes with mass spectrometry (MS), both studies used bait proteins to identify physiologically intact protein-complexes. A recent analysis used a maximum likelihood model and gene expression correlation coefficients to evaluate the reliability of various high-throughput protein-protein interaction datasets and concluded that the TAP dataset had the highest accuracy for predicting protein function (Ding et al., 2001). Another analysis compared the accuracy and coverage of protein interactions for several high-throughput datasets relative to a trusted reference set of protein complexes annotated manually from the Munich Information Center for Protein Sequences (MIPS) and the Yeast Proteome Database (YPD) (von Mering et al., 2002). This analysis also revealed a superior accuracy to coverage tradeoff for the TAP-MS data relative to other methods. Hence we have chosen this dataset to illustrate our model.

We represented this dataset as a bipartite graph with adjacency matrix B. The symmetric matrix BB^T defines the interaction strength of the protein-protein interaction network from the underlying bipartite

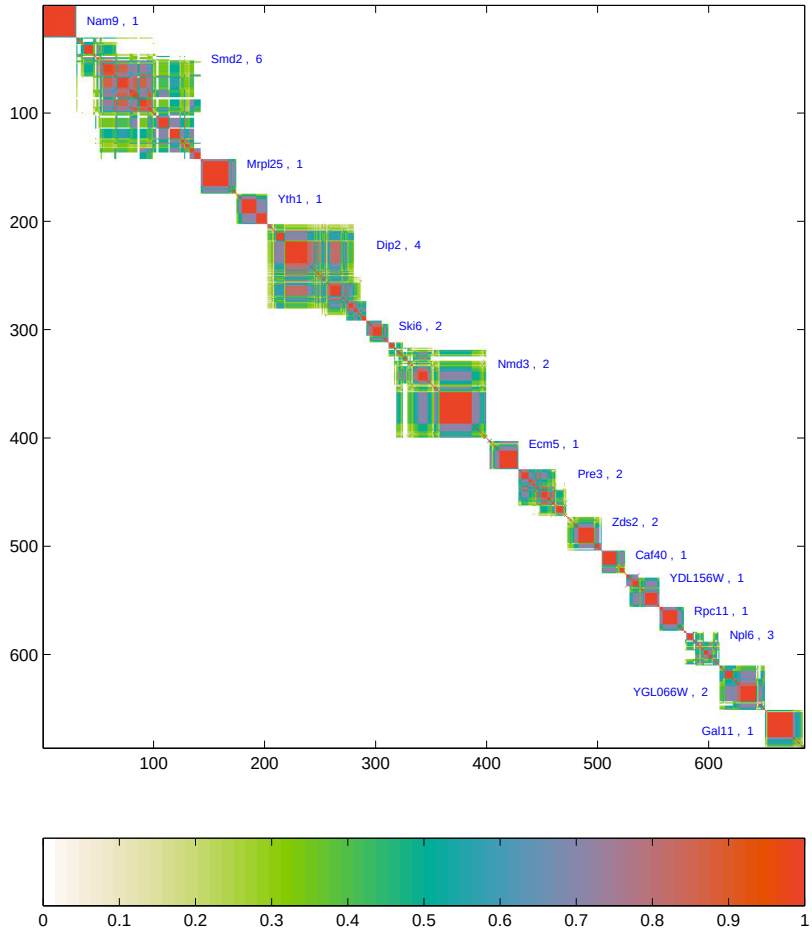

Figure 3: Predicted clusters of the p-p network. The colors show the normalized interaction strength of the p-p network. Clusters with less than 20 proteins are not shown. The most highly connected protein in each cluster is shown by its protein name, and the number of TAP protein complexes this cluster matches to (with $\rho > 0.7$) is shown as the number after the protein name. Axes are protein IDs in the p-c network. They help to show the size of each clusters. For example, cluster P_{28} with protein Smd2 has 112 proteins.

graph model. This p-p network shows a scale free topology indicating that proteins in the network have a wide range of connectivities (Figure 2). Previous work has speculated that connectivity in the network might correlate with observable biological properties such as the rate of protein evolution (Fraser et al., 2002).

Clusters in the P-P Interaction Network Define Modules of Protein Complexes

Given a network of protein interactions, one can computationally predict modules and annotate these modules with a biological context. A computationally predicted protein module is defined as a highly connected region or structure in the network. Previous work has employed "k-cores" and other density-based methods to partition the protein interaction network (Bader and Hogue, 2002; Bader and Hogue, 2003). In this paper we identify protein clusters using MinMaxCut, a graph clustering algorithm which was shown to be effective for class discovery in gene expression microarray data for lymphoma (Ding, 2002) (see Methods Section). We apply MinMaxCut to the protein interaction network specified by the adjacency

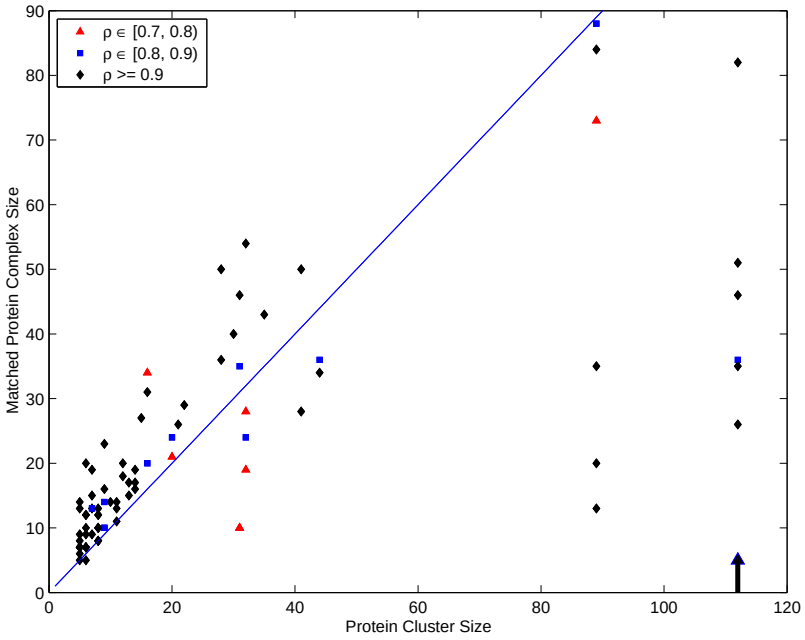

Figure 4: A summary of the overlap between the constituents of predicted p-p clusters and TAP-MS protein complexes. Match coefficients are indicated by the symbols. The solid line indicates where protein complexes and p-p clusters are the same size.

matrix BB^T. The non-uniform interaction strength between proteins gives a more realistic characterization of the network. We now analyze the p-p interaction network highlighting the main results. A comprehensive analysis of these results is deferred to a later paper.

Figure 3 shows the interaction strength (the adjacency matrix BB^T) of the p-p network sorted after clustering. Several clusters exhibit high overall interaction strength and most encompass biologically meaningful complexes. To support our supposition that clusters in the p-p network encompass physiologically relevant protein complexes we compared the discovered p-p clusters to the TAP-MS protein complexes that are the basis of the bipartite graph model. To quantify this correspondence we define the match coefficient

$$\rho = n(P_k, c_j)/\min(|P_k|, |c_j|)$$

where $|P_k|$ =number of proteins in p-p cluster P_k, $|c_j|$=number of proteins in TAP-MS protein complex c_j, and $n(P_k, c_j)$ = number of shared proteins between P_k and c_j. A protein cluster P_k may be entirely contained in a experimental protein complex c_i; or conversely, c_i could be entirely contained in P_k; both cases result in a perfect match with $\rho = 1$. Using this match coefficient and a threshold of 0.8 we found that 65 of 66 predicted p-p clusters match to at least one experimental protein complex (Figure 4). This is strong evidence that clusters in the p-p network define modules of physiologically intact protein complexes and furthermore that any clustered assemblies with uncharacterized constituents might correspond to novel interactions or functional relationships. We note that those protein clusters which match two or more TAP protein complexes are most interesting. For example, Figure 5 details how the largest cluster in the p-p network denoted P_{28} matches to 6 TAP protein complexes. These matching complexes are also shown as 6 points in Figure 4 as indicated by the arrow.

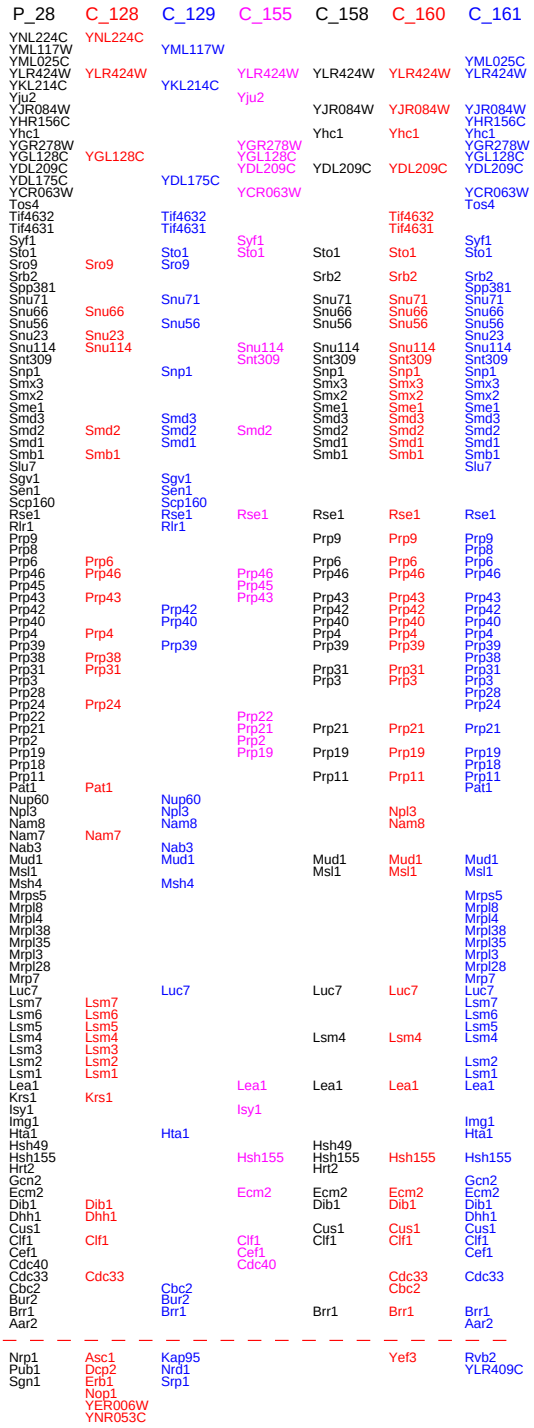

Figure 5: Protein cluster P_{28} matches 6 experimental protein complexes (labeled as published (Gavin et al., 2002)). All proteins in the cluster and protein complexes are listed. Proteins shared by the protein cluster and at least one experimental protein complexes are listed above the dividing line. Below the line are proteins not shared. The matching coefficients are $\rho(P_{28}, c_{128}) = 0.83, \rho(P_{28}, c_{129}) = 0.91, \rho(P_{28}, c_{155}) = 1, \rho(P_{28}, c_{158}) = 1, \rho(P_{28}, c_{160}) = 0.98, \rho(P_{28}, c_{161}) = 0.98$.

Lys	100	Asn	56	Val	30	Ile	24
Asp	89	Gln	50	Tyr	29	Ser	23
Arg	73	Cys	39	Met	29	Leu	22
Pro	70	His	33	Trp	28	Gly	21
Glu	66	Ala	31	Thr	28	Phe	21
pI	169	Basic	149	Acidic	97	MW	60
Aromatic	30	Helix	37	Beta-Sheet	33	Coil	27

Table 1: F-statistics of amino acid composition (top) and physical properties (bottom) across all clusters in p-p interaction network.

Modules in the P-P Network Have Characteristic Physical and Chemical Properties

The assembly, thermodynamic stability, and functionality of protein complexes are controlled by various environmental conditions in the cell. Surface accessible amino acid residues can be covalently modified to regulate the functional state of protein-complexes. Non-covalent ligand binding can also modulate the functional state of protein complexes. Hence we would expect that the proteins of discovered clusters in the p-p network would be distinguishable by intrinsic physical and chemical characteristics. We calculated an F-statistic for protein physical-chemical properties and amino acid composition to see if protein clusters exhibit any significant trends that might suggest distinguishing features of their interactions. Given a particular property f across n proteins and K clusters containing these proteins the F-statistic is defined as

$$F = \frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{f}_k - \bar{f})^2 \Big/ \frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1)\ \sigma_k^2$$

where $\bar{f}$ is the average across all proteins, $\bar{f}_k$ and σ_k are the average and variance within p-p cluster P_k, and n_k is the size of cluster P_k. The magnitude of the F-statistic is a measure of how well the given property distinguishes between clusters. The various properties and their F-statistics are listed in Table 1. To assess the statistical significance, we compute the F-statistic for the same dataset when proteins are randomly assigned to classes. The F-statistic for randomly shuffled data are approximately 16 ± 8 across these quantities. Thus quantities above 30 are significant.

Protein complexes can be characterized as non-obligate (temporary) or permanent where the native state is oligomeric. The surfaces that mediate the interactions in these two types of complexes necessarily differ in structural and physical properties (Jones and Thornton, 1996). Since using different values for the cluster cohesion parameter (See *Methods* section) of the MinMaxCut clustering algorithm is likely to result in discovered protein clusters that encompass differing ratios of these two types of complexes we would expect that the calculated physical properties would be somewhere intermediate between those expected for the two types of complexes. Indeed, this seems to be the case if we consider the F-statistics for amino acid composition. Interactions in temporary protein complexes which function dynamically in cellular processes are often tuned by the effects of polar groups (Lys, Arg, Gln, Asn, Asp) which define a complementary electrostatic surface, hydrogen bonding (Arg) and stabilizing hydrophobic interactions (proline). Methylation of Arg and Lys, and acetylation of Lys are well known covalent modifications of surface amino acids that could influence complex formation. Cys participates in the formation of disulfide bridges that can stabilize more permanent complexes as well as more dynamic interactions (Veselovsky et al., 2002; Jones et al., 2000). Finally studies have shown that secondary structural features are often uniformly distributed at protein interaction interfaces consistent with their relative unimportance in the

above calculations (Jones and Thornton, 1996).

Supercomplexes Encompass Modules from the P-P Network

In all previous analysis of protein-complex data only the resulting pairwise interaction network has been examined (von Mering et al., 2002; Bader and Hogue, 2002; Schwikowski et al., 2000). The pairwise interaction network, however, yields an incomplete and noisy version of proteomic organization. As evidenced by recent high-throughput experiments for determining protein complexes and a few other well studied examples: protein complexes are apt to share components and hence define a network of interconnected cellular processes (Gavin and Superti-Furga, 2003). No study to date has adequately represented the higher-order organization of this network. In our dual representation of the data the adjacency matrix $B^T B$ defines the connectivity between protein complexes where the connection is weighted by the number of shared components. Figure 6 shows the result of a MinMaxCut clustering of this network. Clusters are labeled with the most frequently occurring protein as well as the number of protein complexes corresponding to a particular biological process. We introduce the terminology supercomplex to denote a cluster in the complex-complex interaction network.

Since we expect supercomplexes to represent a diversity of interconnected cellular processes it would be consistent if each supercomplex showed high match coefficients with various modules from the p-p interaction network. Figure 7 summarizes the overlap between predicted supercomplexes and predicted protein complexes. Most supercomplexes show overlap with several predicted protein complexes, and in some instances the same predicted protein complex occurs in multiple supercomplexes. In one instance a module in the p-p network and and a supercomplex are in one-one correspondence (the p-p cluster listed in Figure 5).

Computationally Discovered Modules are Biologically Consistent

We provide anecdotal evidence that computationally discovered modules in the dual representation are biologically consistent. To determine a biological context we used a set of controlled vocabularies defined by the Gene Ontology for which most of the proteins in our dataset have been annotated with at least one term (Dwight et al., 2002). The Gene Ontology consists of three orthogonal ontologies: biological process, molecular function and cellular component (Ashburner et al., 2000). Given that p-p clusters are defined by the proteins sharing maximal membership within the same experimentally determined protein complexes and c-c clusters capture relationships between protein complexes, we would expect the biological process and cellular component ontologies to give the most coherent annotations. We map each protein in a p-p cluster to the most specific ontological term assigned to it. For c-c clusters we determine a non-redundant union of all protein constituents and map these to their most specific annotated terms. The GO ontologies are organized as directed-acyclic graphs. This data-structure allows us to ascend the graph from more specific terms to determine the set of common "parent" terms that describe a predicted cluster's functional categories. We approximate the significance of that annotation by calculating the probability that n or more proteins would be assigned to that term if we selected randomly from the cluster. This probability is calculated as

$$P = \sum_{n \leq j \leq N} \binom{N}{j} p^j (1-p)^{N-j}$$

where p is ratio of proteins in the genome annotated to the given term, and N is the number of proteins in the cluster. This p-value allows us to rank annotations according to significance and to reason about

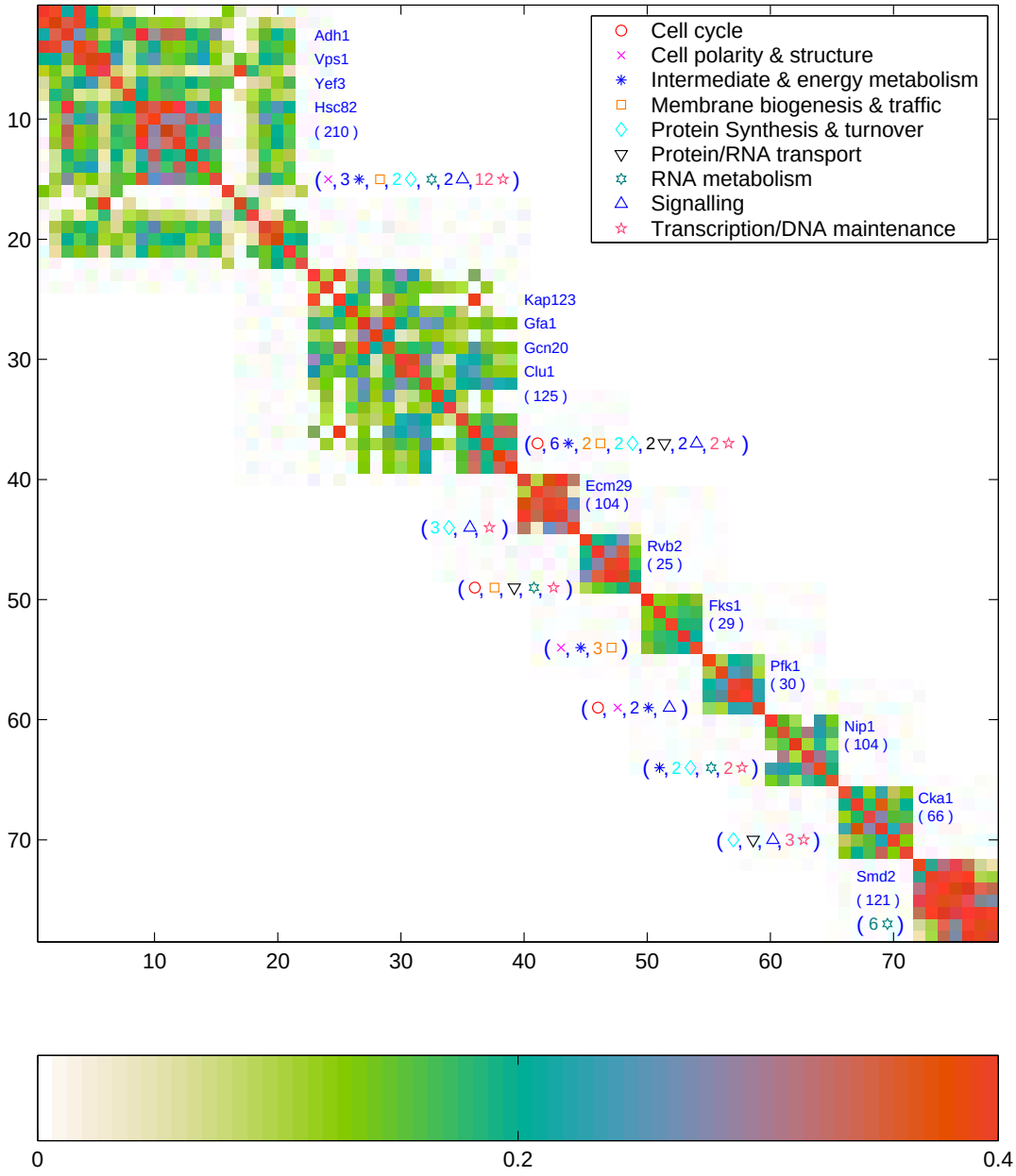

Figure 6: Predicted protein supercomplexes (clusters of the c-c network). Several large supercomplexes are shown. Each supercomplex is labeled with frequently occurring proteins, the number of total non-redundant constituent proteinss, and the relevant biological processes inferred from the participating TAP experimental protein complexes. Axes correspond to arbitrary experimental complex ids.

the cellular roles for a given cluster. If a subgraph composed from the significant terms is biologically consistent, then we may state the validity of the computationally determined module.

We briefly present two examples: the largest cluster in the p-p network denoted P_{28} and the largest cluster in the c-c network denoted C_{47}. P_{28} contains 112 proteins as depicted in Figure 5. Figure 8 shows the most significant ontological terms from the GO - cellular component ontology corresponding to the proteins in this cluster. Annotations to the general terms nucleus (76 proteins) and ribonucleoprotein

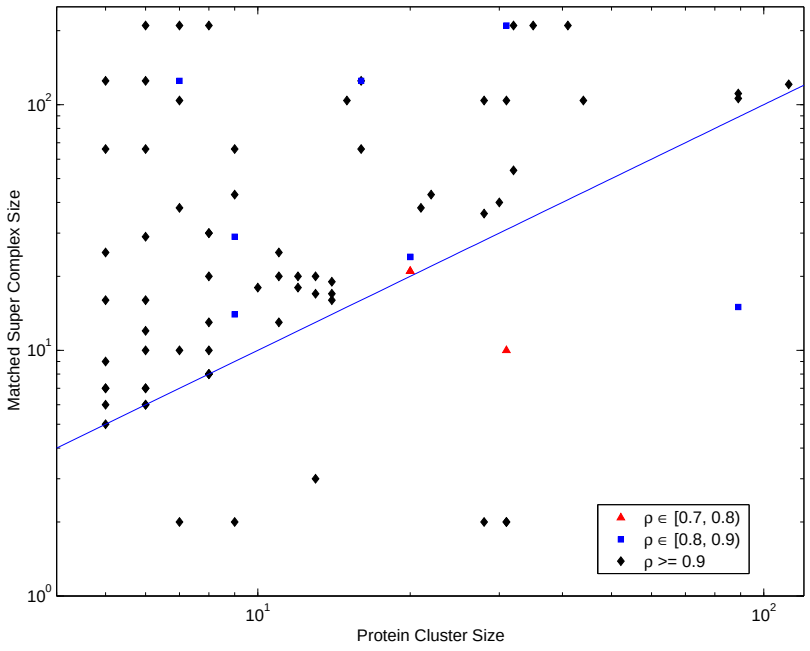

Figure 7: Overlap between computed supercomplexes (clusters of c-c network) and predicted protein complexes (clusters of p-p network). The match coefficient defined through shared protein constituents are indicated.

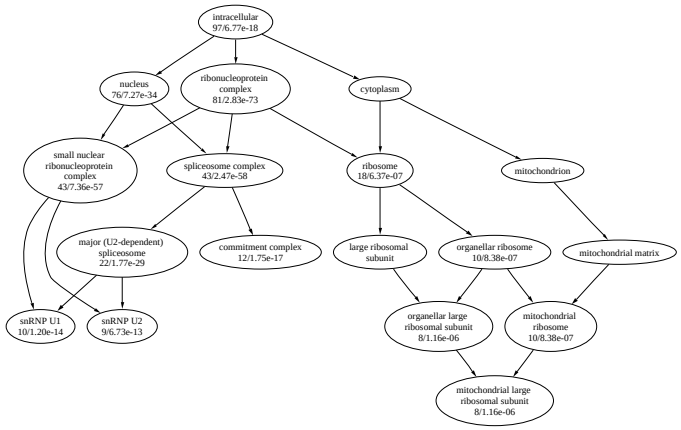

Figure 8: Subgraph of the gene ontology (Component) corresponding to a subset of the most prevalent annotations of proteins in p-p cluster P_{28}.Significant nodes are labeled with the number of proteins annotated directly or indirectly to that term and the p-value for the term.

(RNP) complex (81 proteins) as well as more specific terms such as spliceosome complex (48 proteins), major (U2 dependent) spliceosome (22 proteins) and commitment complex (12 proteins) clearly indicate these proteins are components of the pre-mRNA splicing machinery. It is well known that the transcriptional machinery consists of several coupled multi-protein machines that carry out separate steps in gene expression coordinated via interactions with the carboxy terminal domain of the RNA polymerase II large subunit (Maniatis and Reed, 2002).

The predicted protein complex P_{28} is also the only p-p cluster that corresponds exactly with a supercomplex. Thus, while most of the proteins in the cluster have been accounted for in stable complexes,

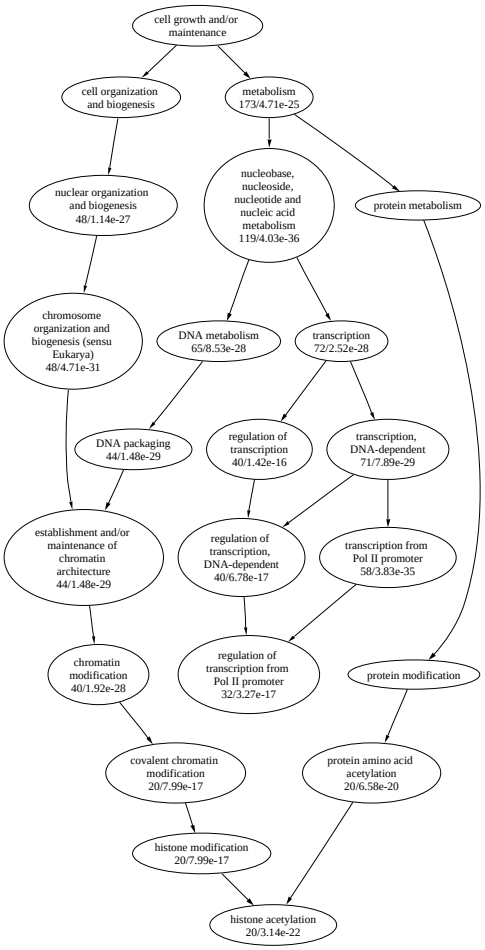

Figure 9: Subgraphs of the gene ontology (Process) corresponding to a subset of the most prevalent annotations of proteins in supercomplex C_{47}. Significant nodes are labeled with the number of proteins annotated directly or indirectly to that term and the p-value for the term.

there are also some more hypothetical relationships suggested by the GO annotations. Ten proteins are predicted to be associated with the mitochondrial ribosome. Constituents of the mitochondrial ribosome are encoded in both the nuclear and mitochondrial genomes. A mechanism that coordinates the expression of these constituents has been hypothesized, given that the stoichiometric synthesis of all mt ribosomal components is likely to be regulated to avoid wasting metabolic energy (Graack and Wittmann-Liebold, 1998). Hence the clustering of these proteins suggests a possible coupling between gene expression in the nucleus and mitochondria. Additionally, there is evidence that splicing can enhance export of mRNA from the nucleus (Reed and Hurt, 2002) and that combinatorial binding of heterogeneous ribonucleoproteins (hnRNPs) to mRNA may regulate post-transcriptional events such as nuclear export, mRNA stability, and nonsense mediated decay (Keene, 2001). That many of our proteins are annotated to these terms (commitment complex, mRNA-nucleus export, translation initiation, polysome, cytoplasmic transport, mRNA splicing, see supplemental data) at least suggests these relationships and their interdependence.

The largest supercomplex C_{47} illustrates how diverse cellular process can be coupled via a nexus of interconnected protein complexes. Figure 9 shows the most significant GO-process annotations for this supercomplex (210 proteins). The GO-process annotations suggest that this supercomplex encompasses

MIPS Listing	# orfs	# orfs in C_{47}
RNA Pol II holoenzyme	35	23
Kornberg's mediator	21	21
Other transcription	73	17
HAT A	15	14
TFIID	13	13
SAGA	14	13
Ada-Spt	14	13
TAFIIs	12	12
DNA repair	33	9
RSC	10	6
ADA	6	6
Replication fork	30	6
DNA mismatch repair	5	5
Cytoplasmic translation initiation	27	4
SAGA-like	5	4
Nucleotide excision repairosome	16	3
RNA Polymerase III	13	3
Replication factor A	3	3
Actin-associated motorproteins	7	3
MSH2/MSH3	3	3
Srb10p	4	3
NEF4	2	2
eIF4A	2	2
NuA4	2	2
Nuclear pore	24	2
Sir	2	2

Table 2: A sample of known protein complexes from the curated MIPS catalog which have many constituents in supercomplex C_{47}. Listed are the name of the complex, the number of known orfs in the complex, and the number of orfs from the complex present in C_{47}.

complexes involved in chromatin dynamics and transcriptional regulation and initiation as well as cell cycle control, cell wall organization and biogenesis, DNA replication initiation and repair, signal transduction, and general transcriptional regulation (for clarity only a subset of the significant annotations are shown). See supplemental materials for the complete annotation). We determined a list of curated protein-complexes from the MIPS Catalog that are highly represented in the supercomplex. A subset of this list is shown in Table 2. Several of these complexes correspond to known participants in chromatin modifications such as histone acetylation and deacetylation which are prerequisite for such processes as transcriptional initiation, certain types of DNA repair, and cell cycle progression (Roth et al., 2001; Green and Almouzni, 2002; Peterson, 2002).

Conclusion

In this paper we propose a dual representation that unifies three interaction networks, the protein - protein complex (p-c) network, the protein - protein interaction (p-p) network and the protein complex - protein

complex (c-c) network under a single framework. The resulting protein - protein and complex - complex interaction networks have more realistic interaction strengths than the conventional binary interaction networks with equal weighting. This results in a coherent framework for computational detection of modules in the dual representations which occur as clusters or densely connected regions. We apply a rigorous graph clustering algorithm to find these modules. Basic statistical analysis revealed that differences between modules in the protein interaction network are reflected by characteristic physical and chemical properties of the protein interactions. We emphasize the protein complex - protein complex (B^TB) network as reflecting a higher-order organization of the proteome. The largest supercomplex has 210 non-redundant constituent proteins and was involved in a number cellular processes. Use of the Gene Ontology revealed that the biological annotations of computationally discovered modules are statistically significant and that this method can facilitate the functional annotation of uncharacterized constituents in future multi-protein complex datasets as well as the discernment of novel functional relationships between protein complexes. As more high quality protein complex data becomes available, we expect this unified representation of interaction networks and associated clustering methodology will evolve into a useful framework for studying systems biology.

Methods

Protein Complex Data Can Be Modeled as a Bipartite Graph

The representation of a multi-protein complex dataset as a bipartite graph allows us to immediately infer a number of important quantities and to apply a large body of existing graph techniques.

A bipartite graph has two type of nodes: p-type nodes that denote proteins (or p-nodes) and c-type nodes that denote protein complexes (c-nodes). This graph structure only allows connections between p-nodes and c-nodes. Thus a protein complex (c-node) has edges connecting to each of its constituent proteins (p-nodes) (Figure 1). A bipartite graph is uniquely determined by its adjacency matrix $B = (b_{ij})$. Let $c_1, c_2, \cdots, c_n$ denote protein complexes and $p_1, p_2, \cdots, p_m$ denote constituent proteins. Define

$$b_{ij} = \begin{cases} 1 & \text{if protein } p_i \text{ is in protein complex } c_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

i.e., a protein complex is represented by a column in B where each entry is either 1 or 0 where a 1 indicates that the complex contains the protein of the corresponding row. Similar, a protein can be viewed as represented by a row in B. For consistency, we call the relations between proteins and complexes, as represented by the bipartite graph, as the p-c network. Starting from the p-c network, we can naturally obtain the following two networks.

Protein-Protein Interactions (P-P) Network)

The interaction strength of between two proteins p_i, p_j is

$$(BB^T)_{ij} = \begin{pmatrix} \text{\# of protein complexes} \\ \text{containing both proteins } p_i, p_j \end{pmatrix} \tag{2}$$

Note $(BB^T)_{ii} = \sum_j b_{ij}$ = the number of protein complexes that protein p_i is involved. We call this the weight of protein p_i.

Complex - Complex Associations (c-c network)

The interaction strength of between two protein complexes c_i, c_j is

$$(B^T B)_{ij} = \left(\begin{array}{c} \text{\# of proteins shared by} \\ \text{protein complexes } c_i, c_j \end{array} \right) \tag{3}$$

Note that $(B^T B)_{jj} = \sum_i b_{ij} =$ the number of proteins contained in the protein complex c_j. We call this the weight of protein complex c_j.

MinMaxCut Clustering

The MinMaxCut graph clustering algorithm (Ding et al., 2001) can be applied equally well to the p-p or c-c networks. Let the weight matrix $W = (w_{ij})$ denote the pairwise connection strength between proteins, or between protein complexes. We wish to partition the connection network G into two subnetworks G_1, G_2, based on a min-max clustering principle. The total connection strength between G_1, G_2 is

$$s(G_p, G_q) = \sum_{i \in G_p} \sum_{j \in G_q} w_{ij}, \tag{4}$$

The total connection strength within a cluster G_1 or G_2 is similarly defined. The clustering principle requires minimizing $s(G_1, G_2)$ (weak connections been clusters) while simultaneously maximizing $s(G_1, G_1)$ and $s(G_2, G_2)$ (strong connections within cluster). These requirements are satisfied by the objective function,

$$J(G_1, G_2) = \frac{s(G_1, G_2)}{s(G_1, G_1)} + \frac{s(G_1, G_2)}{s(G_2, G_2)}. \tag{5}$$

The solution of the clustering problem is represented by an indicator vector $\mathbf{q}$, where the i^{th} entry of $\mathbf{q}$ is

$$q(i) = \left\{ \begin{array}{ll} a & \text{if} \quad i \in G_1 \\ -b & \text{if} \quad i \in G_2 \end{array} \right. \tag{6}$$

where a and b $(0 < a, b < 1)$ are constants. One can prove that

$$\min_{\mathbf{q}} J(G_1, G_2) \Rightarrow \min_{\mathbf{q}} \frac{\mathbf{q}^T (D - W)\mathbf{q}}{\mathbf{q}^T D \mathbf{q}}, \tag{7}$$

where $D = (d_i)$ is a diagonal matrix, $d_i = \sum_j w_{ij}$. Now, relaxing $q(i)$ from discrete indicator in Eq.6 to a continuous values in $[-1, 1]$, the solution $\mathbf{q}$ of the minimization problem satisfies

$$(D - W)\mathbf{q} = \lambda D \mathbf{q}. \tag{8}$$

The desired solution is the eigenvector $\mathbf{q}_2$ corresponding to the second smallest eigenvalue. From Eq.6, we can recover clusters by the sign of $\mathbf{q}_2$, i.e., $G_1 = \{i \mid q_2(i) \leq 0\}, \quad G_2 = \{i \mid q_2(i) > 0\}$. In general, the optimal dividing point could shift away from 0; we search the dividing point $q(i_{cut})$

$$G_1 = \{i \mid q(i) \leq q(i_{cut})\}, \quad G_2 = \{i \mid q(i) > q(i_{cut})\}.$$

$(i_{cut} = 2, \cdots, n-1)$ such that $J(G_1, G_2)$ is minimized. This gives the final clusters G_1 and G_2.

Hierarchical Divisive Clustering

Divisive clustering starts from the top by treating the whole dataset as a single initial cluster. It recursively splits the current cluster (a leaf node in a binary clustering tree) into two sub-clusters. Two important issues are: (1) how to select the next candidate cluster to split and (2) when to terminate the recursive process.

Given a current cluster G_k, we wish to decide whether to further split it into two sub-clusters. We apply MinMaxCut to G_k. If J^{opt} is large, the overlap between two resulting sub-clusters is large in comparison to the within-sub-cluster similarity and hence cluster G_k should not be further split. Thus the optimal value J^{opt} is a measure of "cluster cohesion".

At every cluster splitting in the divisive process, we compute the cluster cohesion for each of the sub-clusters. To choose the next cluster to split, we choose among all current clusters the one with the smallest cohesion. As the cluster splitting process continues, clusters with small cohesion are split and the cohesion of the resulting clusters increases. To terminate the divisive process, we set a threshold for cohesion $h = 0.6$, i.e., clusters with cohesion greater than h will not be further split. A greater cohesion threshold will lead to "tighter" clusters. h is the only parameter in the MinMaxCut algorithm.

Bioinformatics

The February 2003 release of the Gene Ontology (GO) (http://www.geneontology.org) was used to obtain the annotated terms for yeast proteins from the TAP-MS dataset (Gavin et al., 2002). A freely distributed perl library interface to the Gene Ontology database (Ashburner et al., 2001) was employed for all calculations relating to GO annotations and a perl library interface to the GraphViz package (http://www.research.att.com/sw/tools/graphviz/) was used to create the graph representations. The primary sequences for all proteins analyzed were obtained from the Saccharomyces Genome Database (Dwight et al., 2002). The EMBOSS toolkit (Rice et al., 2000) was used for calculations of sequence properties and the PsiPred program (Jones, 1999) was used for secondary structure determination.

References

Alm, E. and Arkin, A. P. (2003). Biological networks. *Current Opinion in Structural Biology*.

Ashburner, M., Ball, C. A., Blake, J. A., Bostein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*.

Ashburner, M., Ball, C. A., Blake, J. A., Butler, H., Cherry, J. M., Corradi, J., Dolinski, K., Eppig, J. T., Harris, M. A., Hill, D. P., et al. (2001). Creating the gene ontology resource: Design and implementation. *Genome Research*.

Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*.

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

Ding, C. (2002). Analysis of gene expression profiles: class discovery and leaf node ordering. *Proc. 6th Int'l Conf. Comp. Mol. Bio.(RECOMB)*, pages 127–136.

Ding, C., He, X., Zha, H., Gu, M., and Simon, H. (2001). A min-max cut algorithm for graph partitioning and data clustering. *Proc. IEEE Int'l Conf. Data Mining (ICDM)*, pages 107–114.

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., et al. (2002). Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). *Nucleic Acids Research*.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Curciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*.

Gavin, A.-C. and Superti-Furga, G. (2003). Protein complexes and proteome organization from yeast to man. *Current Opinion in Chemical Biology*.

Graack, H.-R. and Wittmann-Liebold, B. (1998). Mitochondrial ribosomal proteins (mrps) or yeast. *Biochem J.*

Green, C. M. and Almouzni, G. (2002). When repair meets chromatin. *EMBO Reports*.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutillier, K., et al. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*.

Jones, D. T. (1999). Protein secondary structure prediction based on position specific scoring matrices. *Journal of Molecular Biology*.

Jones, S., Marin, A., and Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Engineering*.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*.

Keene, J. D. (2001). Ribonucloprotein infrastructure regulation the flow of genetic information between the genome and proteome. *Proceedings of the National Academy of Sciences*.

Maniatis, T. and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*.

Peterson, C. L. (2002). Chromatin remodeling enzymes: taming the machines. *EMBO Reports*.

Pothen, A. (February, 2003). Graph and hypergraph models of protein interaction networks alex pothen. *SIAM Conf. Comput. Science and Eng. Conference presentation*.

Reed, R. and Hurt, E. (2002). A conserved mrna export machinery coupled to pre-mrna splicing. *Cell*.

Rice, P., Longden, I., and Bleasby, J. (2000). Emboss: The european molecular biology open software suite. *Trends in Genetics*.

Roth, S. Y., Denu, J. M., and Allis, C. D. (2001). Histone acetyltransferases. *Annual Review of Biochemistry*.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*.

Veselovsky, A. V., Ivanov, Y. D., Ivanov, A. S., Archakov, A. I., Lewi, P., and Janssen, P. (2002). Protein-protein interactions: mechanisms and modification by drugs. *Journal of Molecular Recognition*.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. P., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*.