A Unified Representation of Multi-Protein Complex Data for Modeling Interaction Networks

Chris Ding,[1] Xiaofeng He,[1] Richard F. Meraz,[1] Stephen R. Holbrook[1]

Keywords: protein interactions, supercomplex, bipartite graph, MinMaxCut clustering

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 unified representation of such datasets based on an underlying bipartite graph model that is an advance on existing models of the network. Our unified representation allows for 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. This 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 protein-protein and complex-complex network using terms from the Gene Ontology suggest that this method will be useful for posing hypothesis about uncharacterized components of protein complexes or uncharacterized relationships between protein complexes.

Protein Complex Data Modeled as a Bipartite Graph

A bipartite graph has two types of nodes: p-nodes that denote proteins and c-nodes that denote protein complexes. A protein complex (c-node) connects to each of its constituent proteins (p-nodes). A bipartite graph is specified by its adjacency matrix $B = (b_{ij})$ where

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

Thus, a protein complex is represented by a column in B and a protein is represented by a row in B. We call the relation between proteins and complexes, 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)

Our unified representation goes beyond the conventional uniformly-weighted protein interactions. The interaction strength between two proteins p_i, p_j is

$$(BB^T)_{ij} = \#\text{ of protein complexes containing both } p_i \text{ and } p_j \quad (2)$$

$(BB^T)_{ii} = \sum_j b_{ij}$ = number of protein complexes containing protein p_i, called the weight of p_i.

Protein Complex - Protein Complex Associations (c-c network)

The interaction strength between two protein complexes c_i, c_j is

$$(B^TB)_{ij} = \#\text{ of proteins shared by } c_i \text{ and } c_j \quad (3)$$

$(B^TB)_{jj} = \sum_i b_{ij}$ = number of proteins contained in c_j, called the weight of c_j.

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: {chqding, xhe,rfmeraz,srholbrook}@lbl.gov

The p-c network B, the p-p network BB^T and the c-c network B^TB are the three main components of the unified representation framework.

MinMaxCut Clustering Result Analysis

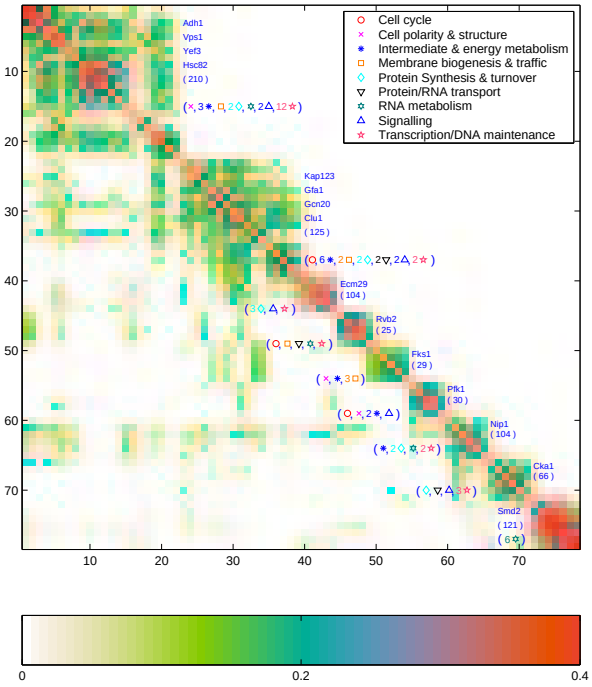

Figure 1: Predicted protein supercomplexes (clusters of the c-c network). Color represents normalized interaction strength.

Figure.1 shows the result of MinMaxCut clustering on c-c network. Clusters (called supercomplexes) are labeled with the most frequently occurring proteins and the number of TAP-MS protein complexes with related biological processes. Gene Ontology annotations show that supercomplexes represent the diversity of interconnected cellular processes. For instance, GO annotations on the largest supercomplex (the first cluster shown in figure) suggest that it encompasses complexes involved in chromatin dynamics, transcriptional regulation and initiation, cell cycle control, DNA replication and repair, and signal transduction.

References

[1] Ho, Y., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415:180-193.

[2] Gavin, A.-C., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141-147.

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

[4] 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)*. pp. 107–114.