Analyzing Protein-Protein Interaction Networks by Spectral Clustering-based Graph Segmentation

Ido Heskia[1] and Rahul Singh[2]
[1]Department of Mathematics, [2]Department of Computer-Science
San Francisco State University, San Francisco, CA 94132

Advances in proteomics at the state-of-the-art are generating a staggering amount of experimental protein-protein interaction (PPI) data. Hidden in this data is information essential to developing a systemic understanding of biochemical processes in cells which can lead to revolutionary insights regarding possible treatment for diseases [1]. Developing computational techniques to analyze, interpret, and extract the useful information from this data is therefore a emerging critical problem.

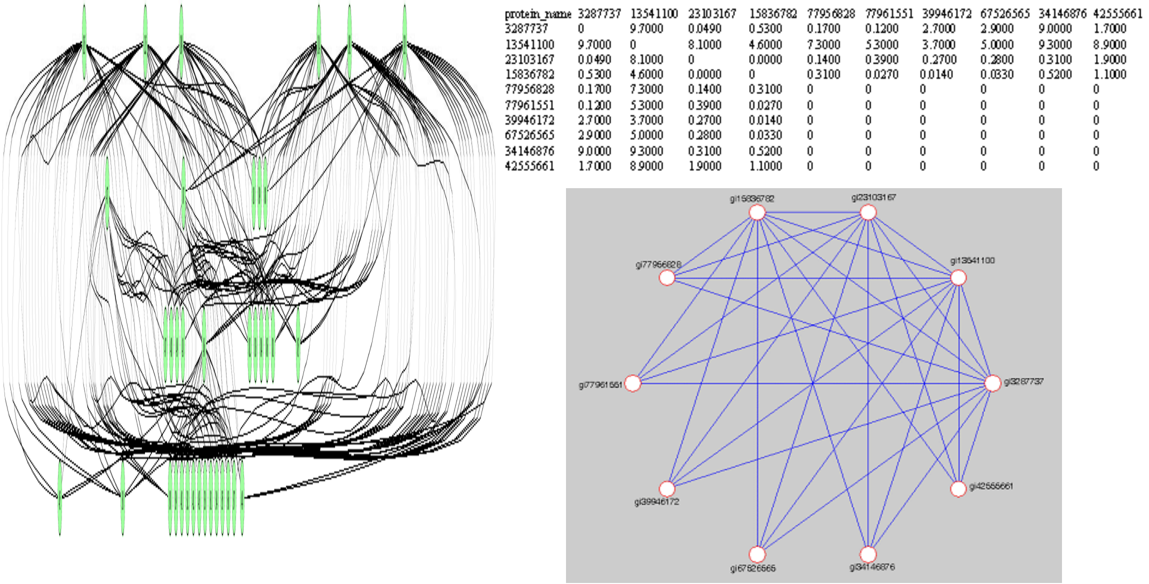
Usually, large PPI networks are represented as an undirected weighted graph *G(V,E)*. Each node $v \in V$ corresponds to a unique protein P_v while each $e_{ij} \in E$, the edge joining node *i* to node *j*, is assigned the weight $w(e_{ij})$ for the particular PPI between P_i and P_j. Therefore, proteins which are closely related, and have many interactions with each other, will have a high weight for the edge joining their corresponding nodes. Because of the large number of interactions, these graphs have dense, non-uniform connectivity (see Figure 1). In order to learn about the most important PPI's in the given network, one needs to partition the graph into disconnected clusters, where each cluster is composed of proteins which have a high PPI (i.e. the edge joining them has a high weight). The basic intuition lies in removing edges from the graph until it is reduced into into smaller disconnected subgraphs. The obvious challenge lies in deciding which edges to remove, and when to terminate the process of removing edges so that we don't lose too much of the available data. Techniques which have been used to analyze such data include, among others, use of random graphs [1] and graph visualization using layout algorithm [2, 3].

We propose a conceptually novel approach to this problem motivated by research in image segmentation [5], where the image is modeled as a weighted graph, with nodes corresponding to pixels and edges representing pixel similarity. The proposed approach is based on the use of isoperimetric partitioning [4]. This algorithm is computationally fast, finds high quality clusters and doesn't require coordinate information.

For a finite graph *G* the isoperimetric number, h_g, is defined as:

$$h_g = \min_{S \subset V} \frac{Vol\,\partial S}{Vol\,S} \quad where \quad Vol\,S = \sum_{e_{ij} \in S} w(e_{ij})v \quad and \quad Vol\,S \le \frac{1}{2} Vol\,V$$

∂S , the boundary of the set *S*, is defined as $\{e_{ij} \mid i \in S, j \notin S\}$. Maximizing *Vol S* ensures that the nodes (proteins) in the cluster would be closely related, and minimizing $Vol\,\partial S$ ensures that the nodes across different clusters would be dissimilar. Thus h_g provides us with the least expensive cut of the graph into disconnected components. This basic strategy is recursively applied to analyze the PPI graph.



protein_name	3287737	13541100	23103167	15836782	77956828	77961551	39946172	67526565	34146876	42555661
3287737	0	9.7000	0.0490	0.5300	0.1700	0.1200	2.7000	2.9000	9.0000	1.7000
13541100	9.7000	0	8.1000	4.6000	7.3000	5.3000	3.7000	5.0000	9.3000	8.9000
23103167	0.0490	8.1000	0	0.0000	0.1400	0.3900	0.2700	0.2800	0.3100	1.9000
15836782	0.5300	4.6000	0.0000	0	0.3100	0.0270	0.0140	0.0330	0.5200	1.1000
77956828	0.1700	7.3000	0.1400	0.3100	0	0	0	0	0	0
77961551	0.1200	5.3000	0.3900	0.0270	0	0	0	0	0	0
39946172	2.7000	3.7000	0.2700	0.0140	0	0	0	0	0	0
67526565	2.9000	5.0000	0.2800	0.0330	0	0	0	0	0	0
34146876	9.0000	9.3000	0.3100	0.5200	0	0	0	0	0	0
42555661	1.7000	8.9000	1.9000	1.1000	0	0	0	0	0	0

Figure 1: Section of the PPI graph (left). The entire graph consists of 1401 proteins/enzymes and has over 750 thousand non-zero weighted edges. One of the clusters obtained from this graph is presented on the left (bottom). At the left (top) is the connectivity matrix for this cluster indicating the edge weights.

An added advantage of this approach is that the value of h_g can be systematically explored (varied) to analyze the graph at different resolutions of putative biochemical relevance. We have applied this algorithm to a PPI network data set consisting of 1401 proteins/enzymes with 788,743 nonzero weighted edges. We found 78 clusters of proteins of varying sizes and biochemical relevance. An example showing one of the clusters is shown in Figure 1.

Our talk will cover the technical details of the approach underscoring its link to image analysis and present results analyzing the biochemical relevance of the protein-protein interactions discovered using this algorithm.

References

[1] N. Przulj, Knowledge Discovery in Proteomics: Graph Theory Analysis of Protein-Protein Interactions.

[2]A.J. Enright and C.A. Ouzounis, BioLayout: an automatic graph layout algorithm for similarity visualization. Bioinformatics. 17;9: 853-854, 2001

[3] T. Frickey and A. Lupas, CLANS: a Java application for visualizing protein families based on pairwise similarity, Bioinformatics. 20;18:3701-3704, 2004

[4] L. Grady and E.L. Schwartz, Isoperimetric Partitioning: a New Algorithm for Graph Partitioning, Society for Industrial and Applied Mathematics SIAM J. SCI. COMPUT. Vol. 27, No. 6 (2006), pp. 1844–-1866.

[5] J. Shi and J. Malik, Normalized cuts and image , IEEE Trans. Pattern Anal. Mach. Intelligence, 22 (2000), pp. 888–-905.