A Proposal for Graph Theoretic Analysis of Protein-Protein Interaction (PPI) Networks

Natarajan Meghanathan
Jackson State University
natarajan.meghanathan@jsums.edu

Abstract

Protein-Protein Interaction (PPI) networks are commonly represented as undirected graphs with nodes corresponding to proteins and unit-weight edges representing the interactions between two proteins. In an on-going research, we intend to improve the knowledge-base of PPI networks by conducting the following graph theoretic analysis studies: (i) We propose to take an intersection of the PPI network graphs for specific commonly-studied organisms extracted from a set of well-known databases and evaluate properties such as degree distribution, diameter and clustering coefficient of the intersection graph, which will retain only interactions reported in all the databases. (ii) We propose to analyze the correlation between the properties observed for the intersection graph and the functionality of the proteins constituting the different PPI components. (iii) We will develop heuristics to search for different graphlets (connected subnetworks with a smaller number of nodes) in the PPI networks and conduct a frequency distribution analysis of these graphlets to capture the structural similarities in the different PPI networks.

1. Introduction

Proteomics is the study of the structure and functions of proteins. Understanding protein-protein interactions (PPI) is an important problem in proteomics [1]. A study of the PPI networks will provide valuable insight about the inner working of cells, leading to more insights about complex biological processes. Various databases report PPI experimental results for a variety of organisms. A list of the PPI databases is available at http://ppi.fli-leibniz.de/jcb_ppi_databases.html. The recent deluge of experimental PPI data (e.g., through high-throughput techniques such as mass spectrometry and yeast 2-hybrid screening) has been largely available in the literature [1]. The PPI data has been mainly studied and reported for the following organisms, which would also be used for our research: (i) *Saccharomyces cerevisiae*, (ii) *Drosophila melanogaster*, (iii) *Escherichia coli* and (iv) *Homo sapiens*. The mostly commonly used PPI databases, which will also be used in this research, are as follows:

- DIP (Database of Interacting Proteins), UCLA [2]
- DroID (DROsophila Interactions Database), Wayne State University [3]
- MIPS (Mammalian Protein-Protein Interaction Database) [4]
- BOND (Bio-molecular Object Network Databank) [5]

The proposed research aims to advance the knowledge-base on PPI through graph-theoretic analysis conducted on the PPI network graphs extracted for the above four organisms from these well-known databases. The specific objectives of this proposal are as follows:

1) Evaluate the graph theoretic global properties, such as degree distribution, diameter and clustering coefficient on an intersection of the PPI network graphs for specific organisms, extracted from the different databases.
2) Analyze the correlation between graph theoretic global properties (such as degree, path lengths, connectivity and clustering coefficient) and functionality of the proteins that would belong to different components of the PPI.
3) Develop heuristics to identify 3-, 4- and 5-node graphlets (connected subnetworks that capture the local properties) on PPI graphs of specific organisms and conduct a frequency distribution analysis of the graphlets.

The significance of this research lies in the following aspects: (i) Objective 1 evaluates the graph theoretic properties on PPI intersection graphs based on interactions that could be accepted with high

confidence as they are reported in all the four well-known databases; (ii) Objective 2 would lead to interesting correlation analysis between specific graph theoretic properties and protein functionality in a specific organism. For example, with our analysis, one would be able to know what is the degree distribution (likewise, the path lengths between any two proteins or the clustering coefficient) of proteins that contribute towards energy production (or any other such functionality) in Drosophila Melanogaster or any other specific organism. (iii) Objective 3 will lead to the development of heuristics that can be used to identify the frequency of occurrence of connected subnetworks (called graphlets) of a smaller number of nodes in a larger PPI network. Exhaustive searching for all instances of a graphlet in a large network would be computationally intensive; thus, we develop heuristics for finding approximate frequencies. Such information on the frequency of occurrence of the graphlets can be then used to calculate the relative graphlet frequency distance between the PPI networks of two different organisms. The larger the value of the relative graphlet frequency distance, the larger would be the difference in the structure and properties of the PPI networks of the two organisms compared.

2. Research Background and Review of Relevant Literature

In this section, we define and briefly review the graph theoretic properties such as degree, diameter and clustering coefficient that have been used to study the global properties of a PPI network as well as the notion of graphlets to capture the local properties of a PPI network.

The degree of a node is the number of neighbors connected to the node. Let P(k) denote the probability that a randomly selected node of a network has degree k. Past studies have illustrated that the PPI networks from specific databases have a power-tail degree distribution, with $P(k) \approx k^{-\gamma}$, where $2 < \gamma \leq 3$. Such networks are referred to as scale-free networks [6] wherein most nodes have relatively low degree, however, a significant numbers of nodes will have an unusually high degree. The diameter is the maximum longest of all the shortest path lengths between any two nodes in the network. Despite their large sizes, most PPI networks have been observed to have small diameters. This property is often referred to as the small-world property [6]. The Clustering Coefficient of a network is defined as the average probability that two neighbors of a given node are adjacent [6]. Formally, if a node v in a network has d_v neighbors, the clustering coefficient of node v is defined as $c_v = \frac{2E_v}{d_v(d_v - 1)}$, where E_v is the actual number of edges between the d_v neighbors of v and $d_v(d_v\text{-}1)/2$ is the largest possible number of edges between the d_v neighbors of v. The clustering coefficient of the whole network is the average of the C_v for all nodes v. For PPI networks (that have a small-world model), the clustering coefficient only slowly increases with the number of nodes.

The term graphlet denotes a connected network with a small number of nodes. Graphlet counts quantify the local structural properties of a network. In this research, we will specifically use a series of 29 different 3-, 4- and 5-node graphlets (Figure 1). These graphlets have been widely observed [1] in the PPI networks of the four candidate organisms studied in this research. The 29 graphlets in Figure 1 are ordered within groups from the least to most dense with respect to the number of edges when compared to the maximum possible number of edges in the graphlet. A related study on detecting locally over-represented gene ontology terms in PPI networks has been recently conducted in [7].

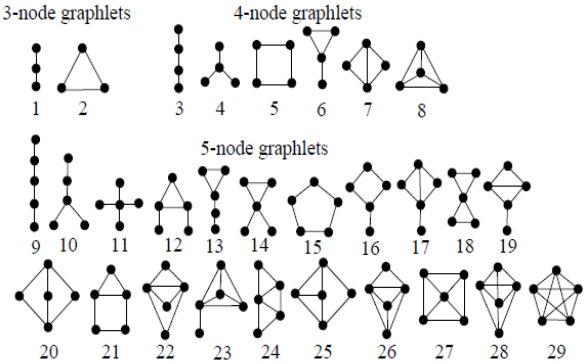

Figure 1: Set of Graphlets [1] used in this Research

3. Research Methodologies

3.1 Research Methodology for Objective 1

For each of the candidate organism s, we extract the PPI network graphs (G_1^s, G_2^s, G_3^s and G_4^s) from the four databases chosen for this study. We will maintain the individual graph information in the form of an Adjacency List. The intersection PPI network graph ($G_{1\text{-}4}^s = G_1^s \cap G_2^s \cap G_3^s \cap G_4^s$), is simply obtained as follows: We first select the individual PPI network graph G_i^s, ($1 \leq i \leq 4$), which has the lowest number of

nodes. We denote such a graph as G_{min}^{s}. A node $v \in G_{min}^{s}$ will be part of $G_{1\text{-}4}^{s}$ only if the node v is also part of the other three individual PPI graphs. Once all the nodes that are to be part of $G_{1\text{-}4}^{s}$ are decided, we will decide on the edges to be part of the intersection graph. An edge (u, v) between two nodes u and $v \in G_{min}^{s}$ will be part of $G_{1\text{-}4}^{s}$ only if $u \in G_{1\text{-}4}^{s}$, $v \in G_{1\text{-}4}^{s}$ and the (u, v) is also part of the other three individual PPI graphs. The run-time complexity to construct such an intersection graph would be $4*(|V| + |E|) = O(|V| + |E|)$ where $|V|$ and $|E|$ are the number of nodes and edges in the graph G_{min}^{s}. We evaluate the graph theoretic properties for the four individual PPI network graphs as well as the PPI intersection graphs as follows:

- The degree distribution is obtained by counting the number of neighbor nodes for each node.
- The diameter of the graph is obtained by running the Breadth First Search (BFS) tree algorithm [8] rooted at each node in the graph. Since the PPI graphs satisfy the small-world property (we anticipate the PPI intersection graphs also satisfy the small-world property), the BFS algorithm is expected to visit a majority of the vertices in the graph with a smaller number of iterations. The distance (number of hops) between the root of the BFS tree and a node in the tree is a measure of the path length. The diameter of a PPI network graph is the maximum of such path lengths across all the BFS trees, one tree per node.
- We obtain the number of edges, E_v, between the neighbors of a node v as follows: If $adj(v)$ is the set of nodes that are neighbors of node v, then for every node $u \in adj(v)$, we traverse $adj(u)$ and increment E_v by 1 if any vertex $w \in adj(u)$ [where $w \neq v$] also exists in $adj(v)$.

3.2 Research Methodology for Objective 2

Once we obtain the degree and clustering coefficient values for each protein (as a result of Objective 1) in the high-confidence PPI intersection network graph for each of the candidate organisms, we could determine the degree distribution and average clustering coefficient of proteins that exhibit a particular functionality. The different functionalities that we plan to examine are: translation, transport and sensing, amino-acid metabolism, energy production, stress and defense and transcriptional control. A PPI network is composed of several components (connected subgraphs) that are connected within, but disconnected from each other. A run of the BFS algorithm on the PPI intersection graph would reveal the different connected components that constitute the overall network. We will examine the constituent nodes of each component and determine whether they are composed of protein nodes of the same or different functionality. The BFS iterations on each such component would also give us information about the diameter. With the aid of the degree distribution of the proteins, we will then investigate the additional number of components that get generated because of the removal of selected high-density proteins (i.e., proteins surrounded by large number of neighbors) of particular functionality.

3.3 Research Methodology for Objective 3

The heuristics to determine the presence of a particular graphlet in a PPI network graph will be developed along the following lines: We will pick a node as the reference node in the graphlet and frame rules that will cover the traversal of every other node and all the edges in the graphlet.

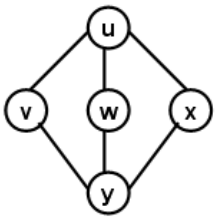

Figure 2: Sample Graphlet

For example if the graphlet of interest is the one shown in Figure 2, the heuristic would be as follows: We will have to identify two vertices in the PPI network (corresponding to vertices u and y in the graphlet) that have the same three neighboring vertices (corresponding to vertices v, w and x). Once we identify a subnetwork, corresponding to a graphlet, in the PPI network, we remove all the constituent edges of the subnetwork identified and repeat this process until we could not find any subnetwork that corresponds to the chosen graphlet. The number of such subnetworks identified will correspond to the frequency of occurrence of the chosen graphlet. The heuristics for the other graphlets will be developed on similar lines and the frequency of occurrence of each of the 29 graphlets in each of the four PPI networks for a specific organism would be determined.

Our conjecture is that the "similarity" between two PPI graphs should be independent of the total number of nodes or edges, and should depend only upon the differences between the relative frequencies of graphlets. The relative graphlet frequency distance of the PPI network graphs of two different organisms

R and *S* would then be determined as follows: If $N_i(G)$ is the number of graphlets of type *i* ($i \in \{1,\ldots, 29\}$) in a PPI network *G* and $T(G) = \sum_{i=1}^{29} N_i(G)$ is the total number of graphlets of *G*, the relative graphlet frequency distance between two graphs *G* and *H* is calculated as [1]: $D(G,H) = \sum_{i=1}^{29} | F_i(G) - F_i(H) |$; where $F_i(G) = -\log(N_i(G)/T(G))$.

4. Conclusions: Expected Results

The expected results of our research are as follows: We ideally would expect the PPI intersection graph to be scale-free. But, given the inconsistencies among the individual PPI databases, we would first have to really conduct the research and see whether the PPI intersection graphs are scale-free or not with respect to the degree distribution. Since the PPI intersection graph would likely have fewer edges than the individual PPI database based graphs, we conjecture that the path lengths would relatively increase in the intersection graph and hence the diameter also would increase relative to the individual graphs. Similarly, with reduction in the number of edges, the clustering coefficient could also be lower for the intersection graph. We expect some of the smaller connected components not to have the scale-free property with respect to degree distribution. Given the heavy-tailed distribution of the entire PPI database, we expect some of the components to be composed of proteins of different functionalities. A relative frequency analysis of the graphlets across two PPI networks would reveal the measure of similarity between the two networks.

Overall, the proposed research will thus enhance our understanding of PPI networks with the aid of graph theoretic analysis and will form the basis for further funded research in this area.

5. Acknowledgments

This research has been funded through the National Science Foundation (NSF) – Mississippi EPSCoR grant (EPS-0556308). The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agency. The author also would like to acknowledge Dr. Raphael Isokpehi (Jackson State University), Dr. Robert Doerksen (University of Mississippi) and Dr. Jinghe Mao (Tougaloo College) for their feedback to the proposed research.

6. References

[1] I. Jurisica and D. Wigle, "Knowledge Discovery in Proteomics," Chapman and Hall/ CRC Press, ISBN: 1584884398, July 2004.

[2] DIP: dip.doe-mbi.ucla.edu/

[3]DroID: http://proteome.wayne.edu/PIMproject1.html

[4] MIPS: http://mips.helmholtz-muenchen.de/proj/ppi/

[5] BOND: http://bond.unleashedinformatics.com/

[6] O. Mason and M. Verwoerd, "Graph Theory and Networks in Biology," *IET Systems Biology*, vol. 1, pp. 89-119, 2007.

[7] M. Lavallee-Adam, B. Coulombe and M. Blanchette, "Detection of Locally Over-Represented GO Terms in Protein-Protein Interaction Networks," *Journal of Computational Biology*, vol. 17, pp. 443-457, 2010.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, "Introduction to Algorithms," MIT Press, 2nd Edition, ISBN: 0262032937, September 2001.