2012 International Conference on Environment Science and Engieering
IPCBEE vol.32(2012) © (2012)IACSIT Press, Singapoore

Structure and Centrality of the Largest Fully Connected Cluster in Protein-Protein Interaction Networks

Azadeh Jafarnejad and Mahmood A. Mahdavi [+]

Department of Chemical Engineering, Ferdowsi University of Mashhad, Azadi Square, Pardis Campus, 91779-48944, Mashhad, Iran

Abstract: Biological systems are composed of interacting and functional modules. Identifying these modules is essential to understand structure of biological systems. One of the essential modules in each protein-protein interaction network is ribosomal cluster. Our analysis on 20 protein-protein interaction networks of bacteria showed that this cluster is the largest and mostly connected cluster in all networks. The properties of this cluster were described using centrality parameters of the nodes in this sub network. It was found that centrality parameters are effective information to interpret biological aspects of networks with the support of biochemical information. Ribosomal cluster was analyzed in this study as a typical example.

Keywords: "Centrality", "Cluster", "Protein-Protein interaction".

1. Introduction

Protein-protein interactions (PPI) are crucial for important molecular processes in the cell and PPI networks provide insights into protein functions [1]. These networks are modular and clusters correspond to two types of modules: protein complexes and functional modules. Protein complexes are permanent interacting proteins at same time and places but functional modules are transient interacting proteins that participate in a particular cellular process at different time and places [2]. Proteins with high betweenness and low connectivity act as important links between these modules and are more likely to be essential [3]. Bottlenecks are proteins with a high betweenness centrality [4]. In scale-free protein interaction networks, most proteins have few connections, whereas a small proportion of proteins interact with many partners. High-degree proteins are hubs and are approximately threefold more likely to be essential than non-hubs [5]. It was shown that the most central metabolites in the *E.coli* metabolic network have the largest closeness centrality value [6]. A vertex with a high radiality value can easily reach other vertices and is generally closer to the other nodes [7]. Thus, centrality parameters in networks are meaningful biological terms and are very good indicators of essential proteins in PPI networks.

In this study, PPI networks of twenty microorganisms, all bacteria, were analyzed and centrality parameters were obtained. Using the parameters the mostly populated and mostly connected cluster of all studied networks was identified as "ribosome" cluster. The features of this cluster were studied and some biological facts were confirmed based on the findings.

2. Materials and Methods

2.1. PPI networks

The PPI catalogues of twenty microorganisms were retrieved from STRING (version 8, 2010). This version covers about 2.5 million proteins from 630 organisms, providing the most comprehensive view on protein–protein interactions currently available [8]. There is a ‘combined score’ between any pair of proteins in STRING. In this study, interactions with a combined score of 900 or more were used. The visualized

[+] Corresponding author. *E-mail address*: mahdavi@ferdowsi.um.ac.ir.

network of interactions resulted in a majority of nodes connected to each other as a network and a minority of nodes disconnected from the whole network. Only the connected portions of the PPI datasets were selected for analysis.

2.2. Software

Duplicated interactions were removed using a Perl script and modified data was imported to the Cytoscape 2.8, a bioinformatics package for biological network visualization and data integration [9]. Centrality analysis was performed using CentiScaPe, a Cytoscape plugin for computing network centrality parameters [10]. Clusters were detected using ClusterONE, a plugin to discover densely connected and possibly overlapping regions within the Cytoscape network [11]. Topological parameters were computed using Network Analyzer, a Cytoscape plugin for computing and displaying a comprehensive set of topological parameters of networks [12].

3. Results and Discussion

3.1. Overall structure of the networks

By plotting the degree probability P(k) as a function of degree (k) for 20 datasets, it was observed that most nodes have low degrees and a few nodes have high degrees, as observed in scale free networks. However, small networks were far from scale free distribution more than the larger ones. For example in the largest studied network i.e. *Escherichia coli* with 2063 nodes and 7112 interactions the exponent of the degree (γ) was 1.8 while that was 1.3 for the smallest studied network related to *Pelotomaculum thermopropionicum* whit 334 nodes and 1056 edges as shown in Figure 1. As seen in this figure there is a hump shape that demonstrates a trace of initial conditions of the degree distribution which may be found for any network size. As the size of networks grows, humps become smoother and degree distribution approaches to scale free distribution [13].

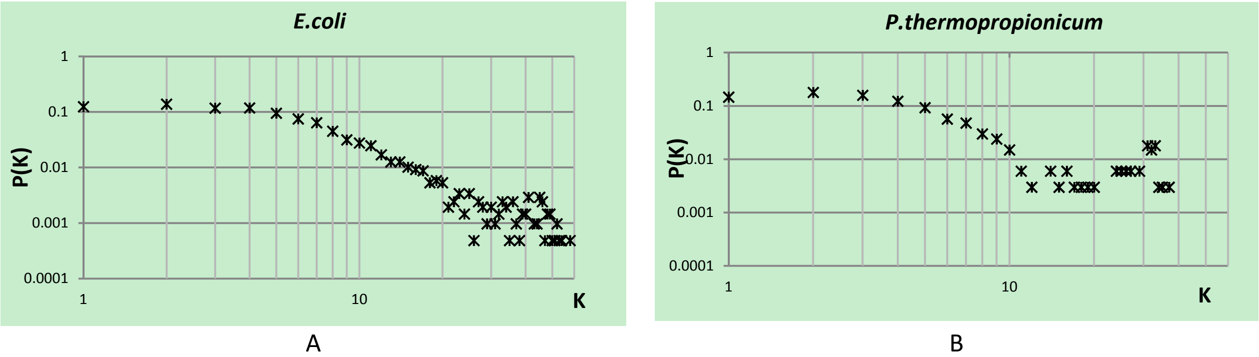

Fig. 1: Degree probability distribution P(k) as a function of degree k. A. *E.coli* B. *P.thermopropionicum*

In disassortative networks high-degree vertices connect preferentially to vertices with low degrees. Average neighbor degree of vertices in the network is a measure to evaluate degree correlations [14]. Technological and biological networks tend to be disassortative [15]. By plotting the neighbor degree of vertices versus degree in interaction networks of 20 microorganisms no evidence of disassortativity mixing in protein level was observed. An important point about these networks is that big clusters in networks are not connected to each other directly but there are some low degree vertices whit high betweenness value acting as linkers between these modules. In other words, in modular level, networks are disassortative; but, inside the modules degree and neighbor degree are correlated. Neighbor degrees in *E.coli* and *Shewanella oneidensis* networks are shown in Figure 2. The trend indicates that in lower degrees the average value increases whereas in higher degrees the average value remains constants in a certain interval.

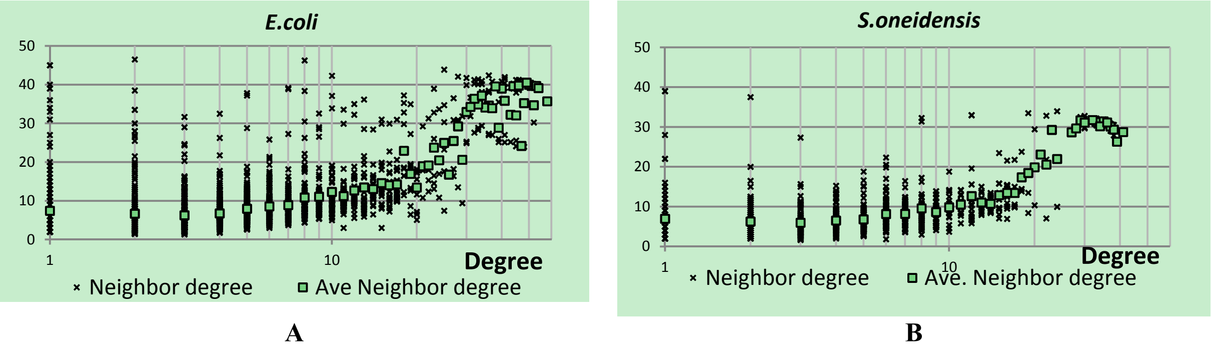

Fig. 2: Neighbor degree vs. degree in *E.coli* and in *S.oneidensis*. Dots in blue represent neighbor degrees and dots in red represent average neighbor degrees in each degree.

Clustering coefficient is an important measure that shows internal structure of a network that is related to the local cohesiveness of a network and measures the probability that two vertices with a common neighbor are connected [14]. Clustering coefficients of vertices versus degree in PPI networks of 20 microorganisms were computed. It was observed that vertices with low degrees had a wide range of clustering coefficients and the maximum value of clustering coefficient that is 1 was found to be abundant in low degrees. High-degree nodes had average clustering coefficients comparable to the average of the whole network. Average clustering coefficients in *E.coli* and *S.oneidensis* networks are shown in Figure 3.

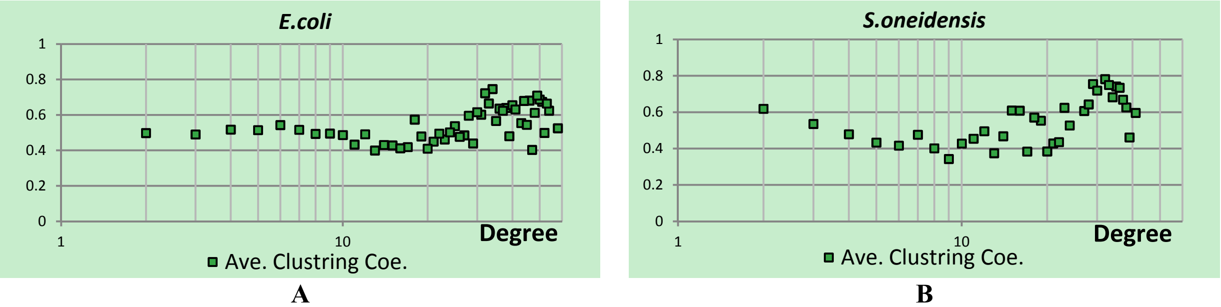

Fig. 3: Average clustering coefficients versus degree in *E.coli* and *S.oneidensis*.

Another important centrality index is betweenness. High betweenness centrality means that the node, for certain paths is crucial to maintain node connections. Betweenness in *E.coli* and *S.oneidensis* networks are shown in Figure 4. The highest values belong to the nodes with higher degrees.

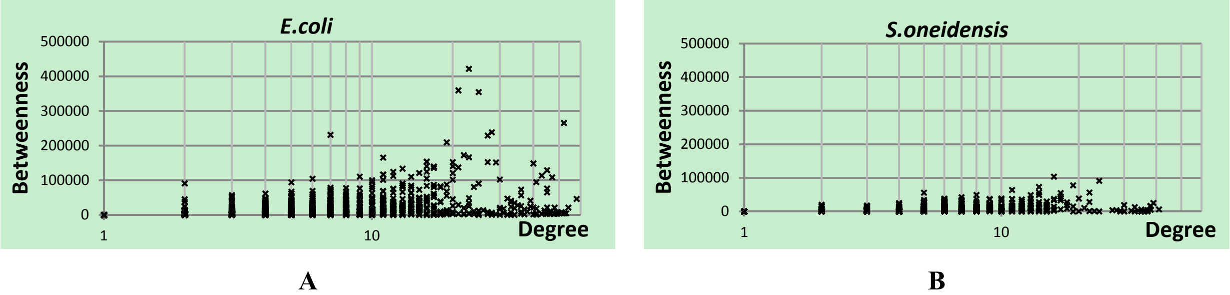

Fig. 4: Betweenness distribution with degree in *E.coli* and *S.oneidensis*

In these networks two important features i.e. clustering coefficient and average neighbor degree, were observed to be high in high-degree nodes. This is an indication that high-degree nodes have a high tendency to be the core of clusters and construct fully-connected components. Although some low-degree nodes have close to maximum clustering coefficients, they can not construct major clusters because they have a few neighbors. There are high-degree proteins with low betweenness centrality in the network. These nodes are the ones that aren't in the paths with high betweenness. These nodes are mostly found inside clusters and have no major role in the whole network connectivity. On the other hand, there are some low-degree nodes with high betweenness in clusters that link one cluster to the other. These nodes are crucial for the connectivity of the network.

3.2. Ribosomal cluster

In all the studied networks, the first and the most important identified cluster was found to be ribosomal cluster, which is the most populated cluster of networks with ribosomes as member protein. Ribosomal functional group plays an important role in existence of the cell and acts as the synthesizer of proteins in the cell. In all networks the highest degrees are related to ribosomal proteins. These proteins are highly connected to each other and their internal connections are much more than their external connections. Most members, despite their high degrees, are only connected to the internal members of the cluster and almost all their proteins in the twenty microorganisms have a very low betweenness. There are only some proteins observed in the cluster with high betweenness playing the role of connecting bridge of this fully connected high populated group with the rest of the network. Although the connections of this sub network are not limited to these proteins but other paths are much less important in comparison with them and the paths leading to them have a very low betweenness. For example in *P.thermopropionicum* network, eliminating five member of ribosomal subgroup causes the disconnection of more than 65 percent of all of the paths in this cluster with the network.

3.3. Centrality indexes and biological facts

Six centrality indexes for each node including degree, betweenness, stress, closeness, eccentricity and radiality were computed for proteins of the networks. Examination of centrality indexes of ribosomal cluster in the networks shows that in most of the networks the RNA polymerase enzyme and its subunits have the highest betweenness. This observation is relied on the fact that in all of the organisms, especially in bacteria, this enzyme is the main factor in transcription. Transcription will not continue without this enzyme and the next process, which is the translation and synthesizing the protein with the help of ribosome, will not occur. So the main connections of ribosomes with the network are through this enzyme. Energy is supplied for the activity of ribosome and protein synthesis through an enzyme called Adenylate kinase, which is one of the nodes with the most Centrality. It provides the essential energy by reverse transfer of the last phosphate from ATP to AMP. This small enzyme is involved in both energy metabolism and synthesizing the nucleotides and is one of the essential enzymes for growth and reconstruction of cell. Hence, all ribosomal proteins have interactions with this enzyme to gain energy. This enzyme is one of the hubs and bottlenecks of the network. RNA polymerase and Adenylate kinase are two typical enzymes which are as nodes with the highest betweennesses in all studied networks.

In ribosomal clusters other centrality parameters are of interest. Eccentricity values of the member nodes are high as these enzymes need to be in contact with the furthest nodes of the network due to their importance. These nodes also have high closeness and Radiality values because these nodes are not only bottleneck, but also hub and other nodes are in the proximity of these nodes. There are some members in ribosome cluster which have high betweenness parameter in spite of their low degrees. A typical member is the sigma subunit of polymerize RNA enzyme. This subunit is responsible for connecting RNA polymerase to DNA and start transcription. The Eccentricity of this enzyme is high as the function of all ribosomal proteins starts with this enzyme. But the Closeness of this node, in comparison with other members with high betweenness, is low. This is due to the observation that this member is connected to ribosome through other subunits of RNA polymerase. This parameter also shows that the sigma subunit is a bottleneck but not a hub.

The least betweenness in the ribosome cluster is related to ribosomes and their connecting proteins to tRNA. These proteins are only connected with each other and don't have much connections with outside. They rather have a high Closeness parameter because of their high numbers in the cluster, but due to the lack of connection with other members of the network, their eccentricity is low. Some ribosomal proteins with unexpected high betweenness exist in the cluster. These nodes have low eccentricities despite their high betweennesses. This shows that they are not central nodes in the network. The reason of their high betweennesses is that they connect members of the cluster and don't have much connection with outside of the cluster independently.

4. Conclusion

The protein-protein interaction networks of twenty bacteria were analyzed in terms of centrality parameters and it was found that ribosomal cluster is the largest and mostly connected cluster in all studied networks. Centrality parameters also assist in recognizing hubs, bottlenecks, and effective members of the cluster. Combination of these parameters would have biological meanings related to the key proteins of a network. Centrality parameters are effective information for interpretation of PPI networks that requires the extensive support of biochemistry of proteins and nucleotides.

5. References

[1] U. Stelzl, et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell.*2005, 122: 957–968.

[2] C. Lin, et al. Clustering Methods in Protein-Protein Interaction Network. In: *Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications.* John Wiley & Sons, Inc. 2006, pp. 319-356.

[3] M. P. Joy, et al. High-Betweenness Proteins in the Yeast Protein Interaction Network. *Journal of Biomedicine and Biotechnology* . 2005, **2**: 96–103.

[4] Yu, H, et al. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Comput Biol*, 2007, **3** : 713-720.

[5] J. J. Han, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *NATURE*, 2004, **430**: 88-93.

[6] H. W. Ma, A. P. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics,* 2002, **19**: 1423–1430.

[7] D. Koschützki, F. Schreiber. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regulation and Systems Biology*, 2008, **2**: 193–201.

[8] C. V. Mering, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 2005, **33**: D433–D437.

[9] M. E. Smoot, et al. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 2011, **27**: 431-432.

[10] G. Scardoni, M. Petterlini, C. Lauda. Analyzing biological network parameters with CentiScaPe. *Bioinformatics*, 2009, **25**: 2857-2859.

[11] T. Nepusz, H. Yu, A. Paccanaro, Detecting overlapping protein complexes in protein-protein interaction networks. In preparation.

[12] Y. Assenov, et al. Coputing topological parameters of biological networks. *Bioinformatics*, 2008, **24:** 282-284.

[13] S. N. Dorogovtsev, J. F. Mendes, A. N. Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E* , 2001, **63**: 062101- 062104.

[14] B. H. JUNKER, F. SCHREIBER, *Analysis Of Biological Networks.* WileyInterscience, 2008.

[15] M. E. J. Newman, Assortative mixing in networks. *Physical Review Letters*, 2002, **89**: 208701-208705