Analyzing Protein Interaction Networks via Random Graph Model

Xiao-Run Wu[1], Yunping Zhu[2] and Yixue Li[3]

[1]Institute of Intelligent Machines, Chinese Academic of Sciences
PO Box 1130 Hefei Anhui, China 230031
[2]Beijing Institute of Radiation Medicine, Taiping Road 27, Beijing 100850, China
[3]Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai, 200031, China

xrwu@iim.ac.cn, zhuyp@hupo.org.cn, yxli@sibs.ac.cn

Abstract

Many complex systems may best be described as networks, which we can use graph theory to analyze their topological properties. In an organism, protein-protein interactions may also be mapped into complex network. Here we use random graph theory to analyze seven different organism protein interaction networks. Three topological properties (degree distribution, clustering coefficient and average shortest path) were used to characterize these networks. The logarithm of the node degree distribution vs. the logarithm of the node degree plot shows that all seven species follow a power-law distribution quite well. In addition, we also obtained the relatively high clustering coefficient of these protein interaction networks. The distance between two nodes of these protein interaction networks indicates that it is quite short comparing with the large network size. The plot of the logarithm of the frequency vs. the shortest path length also indicates that the shortest path length distribution follows a normal distribution quite well.

Keyword: protein-protein interaction, network, random graph theory

I. Introduction

The completion of genome sequencing projects gives us a chance to analyze organisms on a genome level for the first time. The challenge became how to understand the roles of a huge number of gene products and their interaction to create an organism. In parallel to an ever-increasing number of genomes becoming available, some high-throughput protein-protein interaction detection methods have also been introduced in the past couple of years that produce a huge amount of interaction data. Such methods include yeast-hybrid systems [1] [2] [3] [4], protein complex purification method using mass spectrometry [5] [6], correlated messenger RNA (m-RNA) expression profiles [7], genetic interactions [8], and *in silico* interaction predictions derived from gene fusion [9], gene neighborhood [10], and gene co-occurrences or phylogenetics profiles [11]. Traditionally, protein interactions have been studied individually by genetic, biochemical and biophysical techniques. But with the speed that new proteins are being discovered and predicted increases, the yeast two-hybrid (Y2H) method has been proposed for high-throughput interaction-detection [1]. Consequently, these protein interaction detection methods have led to the discovery of thousands of interactions between proteins [2] [3]. These data generated from these methods can be represented graphically as an interaction network in which the nodes represent proteins and pirewise interactions are denoted as

edges. Topological analysis of this network can help us to understand the inner working principle of cells [12] and how the protein interaction networks evolve [13] [14] [15]. Here we used graph theory based analysis to describe the topological structure of this network. Three topological properties (node degree, average shortest path and clustering coefficient) were used to characterize these protein interaction networks.

II. Material and Methods

A. Data Collection

The Database of Interacting Proteins (DIP) [16] [17] [18] [19] is a curated database containing information about experimentally determined protein-protein interactions. We analyze the DIP (Apr. 24 and Apr. 17, 2005) for seven different organisms, i.e., *S. cerevisiae, D. melanogaster, C. elegans, H. pylori, H. sapiens, E. coli* and *M. musculus*. In Table.1 we list the statistics of the numbers of proteins, and the number of interactions in our analysis for the seven organisms.

Table 1. The statistics of DIP database for S. cerevisiae, D. melanogaster, C. elegans, H. pylori, H. sapiens, E. coli and M. musculus

Organism	Proteins	Interactions
S. cerevisiae(CORE)	2640	6600
D. melanogaster	7451	22819
C. elegans	2638	4030
H. pylori	710	1420
H. sapiens	1065	1369
E. coli	553	761
M. musculus	329	286

For the sake of minimizing experimental uncertainty, we used the CORE subset of DIP database, which contains the pairs of interacting proteins identified in the budding yeast, *S. cerevisiae* that were validated according to the criteria described in [16] [17] [18] [19].

B. Theory and Methods

A graph is usually denoted by G, or by $G(V,E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges of G. We often use n to represent $|V|$, and m to represent $|E|$. We also use $V(G)$ to represent the set of nodes of a graph G, and $E(G)$ to represent the set of edges of a graph G. Nodes joined by an edge are said to be adjacent. A neighbor of a node v is a node adjacent to v. We denoted by $N(v)$ the set of neighbors of node v (referred to as the neighborhood of v), and by $N[v]$ the closed neighborhood of v, which is defined as $N[v] = N(v) \bigcup \{v\}$. The degree of a node is the number of edges incident with the node. A graph is complete if it has an edge between every pair of nodes. Such a graph is also called a clique. A complete

graph on n nodes is commonly denoted by K_n. A path in a graph is a sequence of nodes and edges such that a node belongs to the edges before and after it and no nodes are repeated; a path with k nodes is commonly denoted by P_k. The path length is the number of edges in the path. The shortest path length between nodes u and v is commonly denoted by $d(u,v)$. The diameter of a graph is the maximum of $d(u,v)$ over all nodes u and v. If a graph is disconnected, we assumed that its diameter is equal to the maximum of the diameters of its connected components.

B.1. Degree distribution

Generally, degree is the simplest and the most intensively studied one-vertex characteristic. Degree, k, of a vertex is the total number of its connections. From the adjacent matrix, one can obtain a histogram of k interactions for each protein. Dividing each point of the histogram with the total number of proteins then $P(k)$ can be produced. In a random network [20] [21], the edges are randomly connected and most of the nodes have degree close to $< k >$. The degree distribution is generally a Poisson distribution, i.e., $P(k) \sim e^{-k}$, for $k << < k >$ and $k >> < k >$. In many real networks, degree distribution has no well-defined peak but has a power-law distribution [23] [24]. Such networks are referred to as scale-free network. The power-law form of the degree distribution implies that the networks are extremely inhomogeneous. In the scale-free network, there are many nodes with few edges and a few nodes with many edges. In general, the highly connected nodes play a crucial role in the functionality of the network.

B.2. Clustering coefficient

The second topological quantity, which is measurable, is known as the clustering coefficient [22] [24]. The coefficient is a measure of the tendency of the nodes of the network towards clustering. The clustering coefficient is generally defined in the following manner. Assume a specific node i in the network is connected by k_i edges to other nodes. If all these first neighbors are located within a cluster, there would be $\frac{k_i(k_i-1)}{2}$ edges between them. Consequently, the clustering coefficient C_i of node i can be written as:

$$C_i = \frac{2E_i}{k_i(k_i-1)}. \qquad (1)$$

where E_i is the number of actual edges which exist between the k_i nodes. As a result, the clustering coefficient C of the whole network can be obtained by taking an average over all the C_i values.

B.3. Shortest path

One may define a geodesic distance between two nodes u and v of a graph with unit length edges, which is the shortest-path length $d(u,v)$, from the node u to the node v. It is necessary for us to introduce the distribution of the shortest-path lengths between pairs of nodes of a network and the average shortest-path length of a network. The average here is made over all pairs of nodes between which a path exists and over all realizations of network.

The average shortest-path length is often called as the characteristic path length of a network, and it usually determines the effective linear size of a network, the average separation of pairs of nodes.

Here, we also calculated the maximal shortest-path length over all the pairs of nodes between which a path exists. In general, this characteristic will determine the maximal extent of a network.

III. Results

In Figure 1, we plot the logarithm of the node degree distribution $P(k)$ vs. the logarithm of the node degree k for the seven organisms' protein interaction networks, respectively. It is evidently seen from Figure 1 that the number of the proteins decreases with the number of degrees increases. That is, they follow an inverse relation law, which shows that proteins with high degree are rare in practice. From Figure 1, it can be also found that the log-log plot follows a straight line distribution with a negative slop. This result suggests that the protein interaction networks are scale-free networks.

Table 2. Comparison of the clustering coefficients of protein interaction networks and random network for the seven organisms, where *C* is clustering coefficient.

Organism	Proteins	*d*	*C* measured	*C* for random graph (10^{-4})
S. cerevisiae(CORE)	2640	5.0	0.3315	9.5
D. melanogaster	7451	6.2	0.0243	4.1
C. elegans	2638	3.1	0.0634	5.8
H. pylori	710	4.0	0.0755	28.2
H. sapiens	1065	2.6	0.2056	12.1
E. coli	553	2.8	0.6223	24.9
M. musculus	329	1.7	0.1545	26.4

Real world network also behaves strong clustering property. Here we calculated the average clustering coefficient of the seven networks. In Table.2 we gave the comparison of the clustering coefficients of protein interaction networks and random network for the seven organisms, where *C* is clustering coefficient. In addition, we also gained the relatively high clustering coefficient of these protein interaction networks. These results show that the protein-protein interaction networks behave also strong clustering property. Although we are mainly to focus on two-body interactions, the method can be completely extended to multi-body interactions in the protein interaction networks, where the clusters of proteins with many interactions will be also found [25].

Assume that the shortest-path length of the pairs of nodes for seven protein interaction networks is calculated. As a result, the shortest-path length distribution *vs*. the shortest path length for H. pylori is plotted in Figure 2. From Figure 2, it can be found that the distribution of the shortest path length follows the normal distribution quite well. Figure 3 shows the distribution of H. sapiens protein-protein networks. In addition, we also calculated the characteristic path length and the longest path length of seven protein interaction networks. The results were shown in Table.3. From Table.3, we can find that the characteristic path length and longest path length are both relatively short comparing their large size of the protein interaction networks.

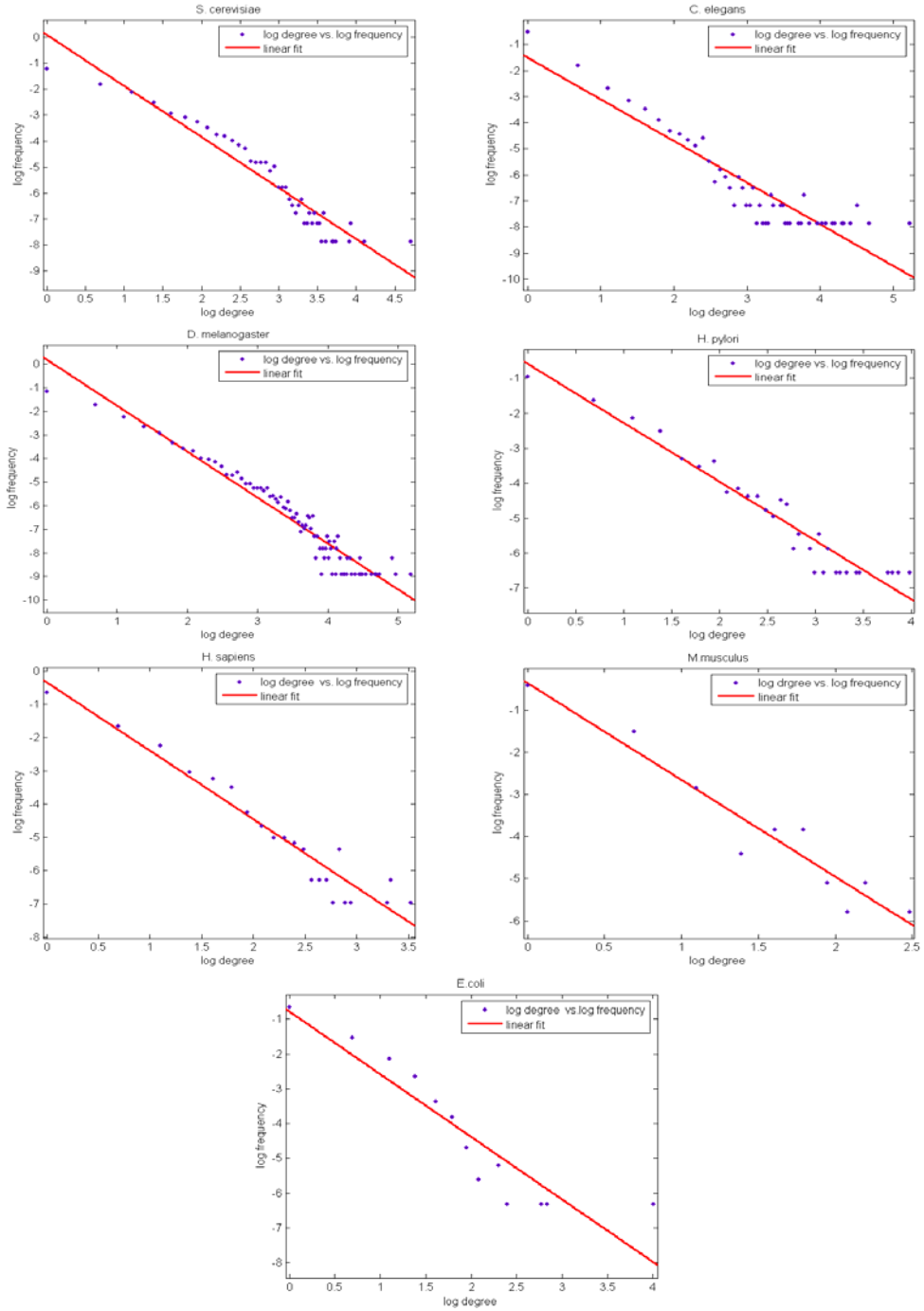

Fig. 1. Degree distribution of different protein interaction networks of the seven organisms. The power-law degree distribution is a robust feature of the protein-protein interaction networks of the seven organisms.

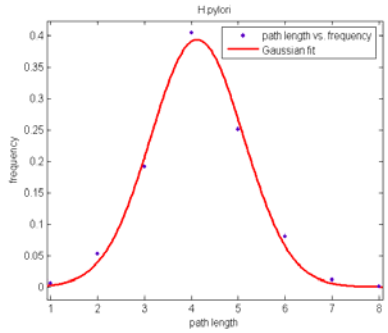

Fig.2. The shortest-path length distribution of *H. pylori* protein-protein networks.

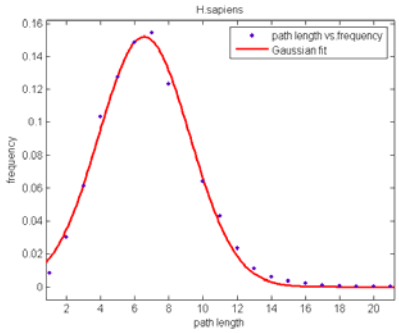

Fig. 3. The shortest-path length distribution of *H. sapiens* protein-protein networks.

Table 3. The characteristic path length and longest path length of the seven organisms' protein-protein interaction networks.

Organism	Characteristic path length	Longest path length
S. cerevisiae(CORE)	5.0	13
D. melanogaster	4.4	11
C. elegans	4.8	14
H. pylori	4.1	9
H. sapiens	6.8	21
E. coli	5.5	16
M. musculus	3.6	9

Ⅳ Conclusions and Discussions

In this paper, we applied graph model theory to analyze the protein-protein interaction networks of seven organisms. Three topological properties were utilized to characterize the process of these protein-protein interaction networks.

The experimental results show that degree distributions of the seven protein interaction networks calculated here all follow the power-law distribution quite well, which means that these protein interaction networks are scale-free network with a few nodes having high degree and the rest having low degree. Usually, real networks often show high clustering property. Clustering coefficient obtained here also indicate high clustering behavior for the seven protein interaction networks. In addition, it can be also found that the shortest-path length and the average shortest-path length calculated here is relatively small comparing with their large network size. This property is usually referred to as a small-world effect.

Although our work is to focus on two-hybrid interactions, it can be conjectured that our analyses on these data can help ones further understand the inner working principle of cells [12] and how the protein interaction networks evolve [13] [14] [15]. In addition, an interesting area tightly related to our work is that if two proteins share significantly large number of common partners, they could have close functional associations [26]. So we can utilize the topology of protein networks to predict protein function of unknown functional protein.

References

[1] S. Fields, O.K. Song, "A novel genetic system to detect protein-protein interactions." Nature 1989,340, 245-246

[2] P. Uetz , L. Giot , G. Cagney , et al, "A comprehensive analysis of protein-protein interactions in *S. cerevisiae*." Nature 2000,403, 623-627

[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc. Natl Acad. Sci. USA, 2001,98, 4569-4574

[4] T. Ito, K. Tashiro, S. Muta, R. Ozawa, et al, "Toward a protein-protein interaction map of the buddin yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins." Proc. Natl. Acad. Sci. USA,2000, 97, 1143-1147

[5] A.C. Gavin ,M. Bosche ,R. Krause , et al, "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature 2002, 415:141-147.

[6] Y. Ho, A. Gruhler, A. Heilbut, et al, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature. 2002 Jan 10; 415(6868):123-4.

[7] T.R. Hughes, M.J. Marton, A.R. Jones, et al, "Functional discovery via a compendium of expression profiles." Cell, 2000, 102:109-126.

[8] H.W. Mewes, D. Frishman, C. Gruber, et al, "MIPS: a database for genomes and protein sequences." Nucleic Acids Res. 2000 Jan 1; 28(1):37-40.

[9] A.J. Enright, I. Iliopoulos, N.C. Kyrpides, C.A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events." Nature. 1999 Nov 4; 402(6757):23, 25-6.

[10] T.Dandekar, B. Snel, M.A. Huynen, & P. Bork, "Conservation of gene order: a fingerprint of proteins that physically interact." TIBS,1998, 23, 324-328

[11] M.A. Huynen, & P. Bork, "Measuring genome evolution." PNAS, 1998,95, 5849-5856

[12] N. Przulj, D.A. Wigle, I. Jurisica, **"**Functional topology in a network of protein interactions." Bioinformatics. 2004 Feb 12; 20(3):340-8.

[13] J. Berg, M. Lassig, A. Wagner, "Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications." BMC Evol Biol. 2004 Nov 27; 4(1):51.

[14] A. Wagner: "How the global structure of protein interaction networks evolves." Proc Biol Sci. 2003 Mar 7; 270(1514):457-66.

[15] A. Wagner: **"**The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes." Mol Biol Evol. 2001 Jul; 18(7):1283-92.

[16] I. Xenarios, D.W. Rice, L. Salwinski, et al, "DIP: The Database of Interacting Proteins." *NAR*,2000,**28**:289-91

[17] L. Salwinski, C.S. Miller, A.J. Smith, et al, "The Database of Interacting Proteins: 2004 update." *NAR*, 2004,**32 Database issue**:D449-51

[18] I. Xenarios, L. Salwinski, X.J. Duan, et al, "DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions." *NAR*, 2002,**30**:303-5

[19] I. Xenarios, E. Fernandez, L. Salwinski, et al, "DIP: The Database of Interacting Proteins: 2001 update." *NAR*, 2001,**29**:239-41

[20] B. Bollobas, *Random Graphs.* Academic Press, 1985, London

[21] P.Erdos, A.Renyi, "On random graphs." Publicationes Mathematicae, 1959,6, 290-297

[22] D.J.Watts, S.H.Strogatz, “Collective dynamics of ‘small-world’ networks.” Nature 1998,393, 440-442
[23] A.l. Barabasi, R.Albert, “Emergency of scaling in random networks.” Science,1999, 286(5439), 509-512
[24] R.Albert, A.l.Barabasi, **“**Statistical Mechanics of Complex Networks.” Reviews of Modern Physics,2002, 74, 47
[25] V. Spirin, L.A. Mirny, “Protein complexes and functional modules in molecular networks.” PNAS, 2003, 100:12123.
[26] M.P. Samanta and S. Liang, “Predicting protein functions from redundancies in large-scale protein interaction networks.” Proc Natl Acad Sci USA, 2003, 100(22):12579-83

Xiao-Run Wu is a postgraduate student at the Institute of Intelligent Machines, Chinese Academy of Sciences. He graduated with a BSc. in Inorganic Non-metal Materials from the School of Material Science and Engineering of Chongqing University in 2001. His current research interests are bioinformatics, neural networks and intelligent computing.

Yunping Zhu, Director of Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine. Principle Investigator of Department of Bioinformatics, Beijing (National) Proteome Research Center. Associate Professor. Zhu is Council Member of Chinese Human Proteome Organization (CNHUPO), member of the academy committee of Beijing Institute of Radiation Medicine. The research interests are bioinformatics for human proteome, and systems biology. His research projects include two from the High-Tech Research and Development Program of China (863), one from the National Basic Research Program of China (973), and one from the Beijing Municipal Key Research Program. Zhu has developed the biggest human proteome database in the world and also the data management platform. He has systematically studied the human liver proteome, including the protein identification, modification, localization, protein- protein interaction network, and metabolic pathways. Now he is also working on the biomarker discovery for tumor. The research papers are published on ***EMBO J***, ***Cancer Research***, and ***Proteomics***, etc.

Yixue LI, Director, Professor, PI Shanghai Center for Bioinformation Technology, and Bioinformation Center of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. Committee leader, Bioinformation Technology, Chinese National High-Tech Program (863). 1997-2000, Post-doctor, Bioinformatics and software development, EMBL, Heidelberg, Germany. 1996-1997, Post-doctor, Computational Mathematics, Applied Computational Institute, Stuttgart University, Stuttgart, Germany. 1992 – 1996, Ph.D., Theoretical Physics, Theoretical Physics Institute of Heidelberg University, Heidelberg, Germany. 1984 – 1987, M.Sc., Theoretical Physics, Department of Physics, Xinjiang University, Urumuqi, Xingjiang, China. 1978 – 1982, B.Sc., Physics, Department of Physics, Xinjiang University, Urumuqi, Xingjiang, China. Research interests include Computational algorithms for biological data mining, Genome annotation, proteomics data analysis, comparative genomics, epitope prediction, biological database construction.