Global topological study of a biological network – the protein-protein interaction network

Ka-Lok Ng[1]
Department of Bioinformatics
Taichung Healthcare and Management University
No. 500, Lioufeng Road, Wufeng Shiang,Taichung, Taiwan 413
klng@thmu.edu.tw

Chien-Hung Huang
Department of Information Management
Ling Tung College
1, Ling Tung Road, Nantun
Taichung, 408
chhuang@mail.ltc.edu.tw

Abstract

We have employed the combinatory graph theory approach to analyze the protein-protein interacting database, DIP, for five different species (*S. cerevisiae, , H. pylori, H. sapiens, E. coli and M. musculus*). Two global topological parameters (connectivity, interaction path length, i.e. diameter) were used to characterize the protein-protein interaction network. Our study indicates that the protein-protein interaction network is neither a random or scale-free network The maximum degree of connectivity study seems to indicate that the highly developed species, such as mammal, have a lower degree of connectivity. It is interesting to notice that, the average interaction path length is of the order 2.3 to 6.2 for any two proteins. Furthermore, one can concluded that for all the five species, the interaction network is quite robust when subject to random perturbation, that is the average diameter for the perturbed case are slightly differ from the unperturbed cases. This can be interpreted that the protein-protein interaction network is robust against external perturbations or errors.

Keywords: protein-protein interaction, yeast two-hybrid model, combinatory graph theory, random graph

1. Introduction

Up to now, the genomes of more than 30 species and at least 100 microbes have been completely sequenced (http://www.er.doe.gov/production/ober/microbial.html). However, the genome sequences do not shed any light on how genes or proteins interact with each other. Then, it is natural to proceed to study how genes are expressed or switched off, or how proteins are interacting in control of intracellular and intercellular processes. In other words, knowing the functionality and interactions is far more important than just having the genomic or protein sequencing information.

Networks of interactions are fundamental to all biological processes; for example, the cell can be described as a complex network of chemicals connected by chemical reactions. Cellular processes are controlled by various types of biochemical networks [Kanehisa 2000] :

(I) metabolism is the most basic network of biochemical reactions, which generate energy for driving various cell processes, and degrade and synthesize many different bio-molecules;

(II) protein-protein interaction network, such as binding interactions and formation of protein complex, is another important class of biological network, and

(III) in a gene regulatory network, the protein encoded by a gene can regulate the expression of other genes. These genes in turn produce new regulatory proteins that control other genes.

In the last few years, we began to see many progresses in analyzing biological networks using the statistical mechanics of random network approach. The random network approach is becoming a powerful tool for investigating different biological systems, such as the yeast protein interaction network [2], food web [3] and metabolic network [4]. Many studies indicated that there are underlying global structures of those biological networks. Below we highlight the current status of these results.

(I) Metabolic network

Metabolism comprises the network of interaction that provide energy and building blocks for cells and organisms. In many of the chemical reactions in living cells, enzymes

[1] Corresponding author

act as catalysts in the conversion of certain compounds (substrates) into other compounds (products). Comparative analyses of the metabolic pathways formed by such reactions give important information on their evolution and on pharmacological targets [14]. Recently, the large-scale organization of the metabolic networks of 43 organisms are investigated and it is found that they all have the feature of a scale-free small-world network [15], i.e. $P(k) \sim k^{\gamma}$ and the diameter of the metabolic pathway is the same for the 43 organisms.

(II) Protein-protein interaction network

Proteins perform distinct and well-defined functions, but little is known about how interactions among them are structured at the cellular level. Recently, it was reported that [5] in the yeast organism (a total of 3278 proteins by the two-hybrid method [6] measurement), the protein-protein interactions are not random, but well organized. It was found that, most of the neighbors of highly connected proteins have few neighbors, that is highly connected proteins are unlikely to interact with each other.

(III) Gene transcription regulatory network

A genetic regulatory network consists of a set of genes and their mutual regulatory interactions. The interactions arise from the fact that genes code for proteins that may control the expression of other genes, for instance, by activating or inhibiting DNA transcription [16]. Recently, it was reported that [17] in the yeast organism, there is a hierarchical and combinatorial organization of transcriptional activity pattern.

2. Input data – Database of Interacting Protein (DIP)

There are thousands of different proteins active in a cell at any time. Many proteins act as enzymes, catalyzing the chemical reactions of metabolism. For our analysis we will use the database DIP [7] (http://dip.doe-mbi.ucla.edu) as the input data. DIP is a database that documents experimentally determined protein-protein interactions (a binary relation). In our computation, we analyze the latest version, July 6, 2003, of DIP, for five different species, *S. cerevisiae, , H. pylori, H. sapiens, E. coli and M. musculus*.

For better accuracy, we employed the CORE subset of DIP which contains the pairs of interacting proteins identified in the budding yeast, *S. cerevisiae* that were validated according to the criteria described in Ref. 8. By analyzing the DIP database one can obtain a matrix representation of protein-protein interactions, M_{int} .

3. Methodology

The biological networks discussed above have complex topology. A complex network can be characterized by certain topological measurements [9]. Erdos and Renyi were the first to propose a model of a complex network known as a random graph [10]. A complex network can be characterized by certain topological measurements [9]. The distance between two nodes is given by the number of links along the shortest path. The number of links by which a node is connected to the other nodes varies from node to node. The diameter of the network also known as the average path length, is the average of the distances between all pairs of nodes.

Connectivity distribution P(k)

Proteins can have direct or indirect interaction among themselves [1]. Direct interactions such as binding interactions, including formation of protein complexes, covalent modifications of phosphorylation, glycosylation and others, and proteolytic processing of polypeptide chains. Indirect interaction refers to be a member of the same functional module (e.g. transcription initiation complex, ribosome) but without directly binding to one another. In this case two proteins (enzymes) are interacted indirectly via successive chemical reactions. Another class of indirect protein-protein interactions is gene expression, where the message of one protein is transmitted to another protein via the process of protein synthesis from the gene. The first topological feature of a complex network is its degree of connectivity distribution. From the interaction matrix M_{int}, one can obtain a histogram of k interactions for each protein. Dividing each point of the histogram with the number of total number of proteins provide $P(k)$. In a random network, the links are randomly connected and most of the nodes have degrees close to $<k>$. The degree distribution $P(k)$ vs. k is a Poission distribution, i.e. $P(k) \sim e^{-k}$, for $k << <k>$ and $k >> <k>$. In many real life networks, the degree distribution has no well-defined peak but has a power-law distribution, $P(k) \sim k^{-\gamma}$,where γ is a constant. Such networks are known 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 links and a few nodes with many links. The highly connected nodes play a key role in the functionality of the net-

work.

Interaction path length

The second topological measurement is the distance between two nodes, which is given by the number of links along the shortest path. The number of links by which a node is connected to the other nodes varies from node to node. The diameter of the network, also known as the average path length, is the average of the distances between all pairs of nodes.

For all pairs of proteins, the shortest interaction path length, *L(j)* (i.e. the smallest number of reactions by which one can reach protein 2 from protein 1) will be determined by using the Floyd's algorithm [11]. Floyd's algorithm is an algorithm to find the shortest paths for each vertex in a graph. It does this by operating on a matrix representing the costs of edges between vertices. The diameter D is given by

$$D = \frac{\sum_j jL(j)}{\sum_j L(j)} \qquad (1)$$

where *j* is the shortest path length and *L(j)* is the frequency of nodes have path length *j*.

Robustness of network

Finally, in order to test whether the interaction network is robust against errors, we slightly perturbed the network randomly. First, we randomly select a pair of edges A-B and C-D. The two edges are then rewired in such a way that A connected to D, while C connected to B. Noticed that this process will not change the degree of connectivity of each node. A repeated application of the above step leads to a randomized version of the original network. Multiple sampling of the randomized networks allowed us to calculate the average diameter of the perturbed network and compare the perturbed results with the unperturbed network.

4. Results

The protein-protein interaction network we obtained has 2164 nodes and 4573 interactions, which is equivalent to a probability of 0.00195 for a random network model. In Table 1, we present the maximum number of connectivity and the average interacting path lengths for both of the unperturbed and perturbed cases for all the five species. The maximum degree of connectivity study seems to indicate that the highly developed species, such as mammal, have a lower degree of connectivity.

Table 1. Maximum number of connection and average diameter for both of the unperturbed and perturbed cases.

Species	Maximum connectivity	Average diameter	Average diameter (random network)
S. cerevisiae	48	5.72	5.71
H. pylori	54	4.14	4.14
E. coli	54	3.23	3.37
H. sapiens	26	6.23	6.36
M. musculus	12	2.29	2.44

Furthermore, one can concluded from Table 1 that, for all the five species, the interaction network is quite robust when subject to random perturbation, that is the average diameter for the perturbed case are slightly differ from the unperturbed cases. This can be interpreted that the protein-protein interaction network is robust against external perturbation. This result suggests that the network only have a few paths having very long interaction length.

In Fig. 1 we plot the logarithm of normalized frequency of connections for each species as a function of the logarithm of degree of connectivity.

It is evident from the figure that the number of proteins decrease with increasing number of connections, that is it has an inverse relation. In other words, protein has multiple connections are rare. The log-log plot result seems quite interesting, because it does not have a peak or follow a straight line with negative slope, it indicates that the protein-protein interaction network is neither a random or scale-free network [5].

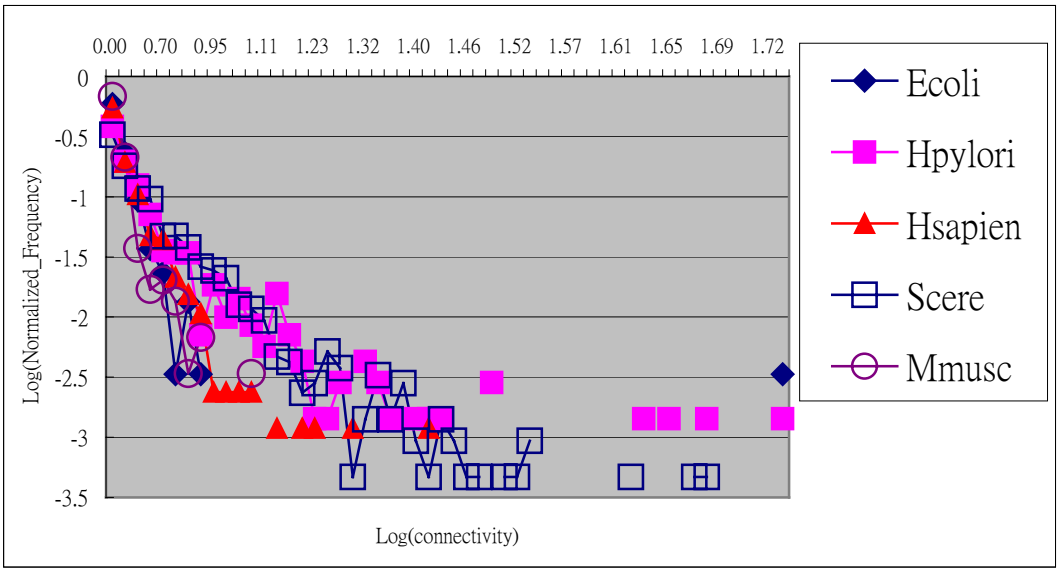

Fig. 1 The logarithm of normalized frequency of connections for each species vs the logarithm of degree of connectivity.

Table 2. The top five yeast proteins that have the highest degree of connections.

Protein name	Swiss-Prot ID	PIR ID	Genebank ID
Calmodulin	P06787	MCBY	71694
Starvation protein 167	P39743	S40887	542352
Actin	P02579	ATBY	14318479
Nuclear pore protein	Q02630	S28925	320799
Serine-rich RNA polymerase I suppressor protein	Q02821	S30884	320855

. In Table 2, we listed the top five proteins that have the highest degree of connections for yeast.

In Fig. 3, we plot the logarithm of the frequency of the shortest path vs the length of the shortest path.

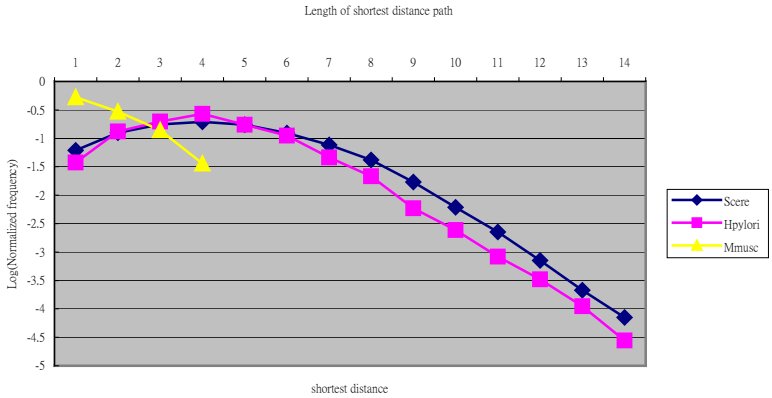

Fig. 3 The frequency distribution of the length of the shortest distance path.

5. Discussion

The degree of connectivity results presented in this paper indicated that the protein-protein interaction neither form a random or scale-free network, where the same conclusion is also reported in Ref. [12]. Here we focused on two-body interactions and it will be interesting

to consider multi-body interactions in the protein network and find clusters of proteins that have many interactions among themselves. Such clusters correspond to protein complexes. Also, Another interesting area of work is to show that if two proteins share significantly large number of common partners than random, they could have close functional associations [13].

6. Conclusions

We have employed the combinatory graph theory approach to analyze the protein-protein interacting database, DIP, for five different species (*S. cerevisiae, H.pylori H. sapiens, E.coli and M. musculus*). Two global topological parameters (connectivity, interaction path length, i.e. diameter) were used to characterize the protein-protein interaction network. The results presented in this paper indicated that the protein-protein interaction neither form a random or scale-free network. It was found that the protein, calmodulin (SWP:P06787, PIR:MCBY, GI:71694), has the maximum connectivity for the *S. cerevisiae*. It is interesting to notice that, the average interaction path length is of the order 2.3 to 6.2 for any two proteins. Furthermore, the robustness of the network can be interpreted that the protein-protein interaction network is robust against external perturbation or errors.

7. References

[1] Kanehisa M., *Post-genome Informatics*. Oxford University Press, Oxford, 2000.

[2] Wagner A., *Mol. Biol. Evol.* 18(7), pp. 1283, 2001.

[3] Melian C. and Bascompte J., *Ecology Letters* 5, pp. 705, 2002.

[4] Ma Hongwu and Zeng An-Ping, *Bioinformatics*, 19, pp. 270, 2003.

[5] Maslov S, and Sneppen K., Science, 296, pp. 910-913, 2002.

[6] Lodish H., Berk A., Zipursky S., Matsudaira P., Baltimore D. and Darnell J. , *Molecular Cell Biology*, 4th ed., W.H. Freeman, N.Y., 2001.

[7] Xenarios I., Frenandez E., Salwinski L., Duan X., Thompson M., Marcotte E., and Eisenberg D., *Nucl. Acid Res*. 29, pp. 239, 2001.

[8] Deane CM, Salwinski L, Xenarios I, Eisenberg D., *Mol Cell Prot* 1: pp. 349-356, 2002.

[9] Bose I., arxiv:cond-mat/0202192, 2002.

[10] Erdos P and Renyi A., *Publ. Math. Inst. Hung. Acad. Sci*. 5, pp. 17, 1960.

[11] Floyd R. W., *Communication of the ACM* 5, no.6, pp. 345, 1962.

[12] Thomas A, Cannings R, Monk N.A.M and Cannings C., LANL e-print archive:q-bio.MN/0309012.

[13] Samanta M.P. and Liang S., LANL e-archive physics/0303027.

[14] Tohsato Y., Matsuda H., and Hashimoto, A. in International Conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, 2000.

[15] Jeong H., Tombor B., Albert R., Oltvai Z.N., and Barabasi A.L., 2002. Science 295, 1662.

[16] Lewin B. 2000. Genes VII. Oxford University Press, Oxford

[17] Farkas I., Jeong H., Viscek T., Barabasi A.L and Oltvai Z.N., 2003. Physica A, 318 601. (arxiv:cond-mat/0205181)