*2011 International Conference on Bioscience, Biochemistry and Bioinformatics*
*IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore*

# Self-similarity in Weighted PPI Networks

Dan-Ling Wang[1], Zu-Guo Yu[1,2*] and Vo Anh[1]
[1]Discipline of Mathematical Sciences, Faculty of Science and Technology,
Queensland University of Technology, Brisbane, Q4001, Australia.
[2]School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China.
Email: wang_dan_ling@sina.com (D.-L Wang), yuzg@hotmail.com (Z.-G. Yu), v.anh@qut.edu.au(V. Anh)
[*] Corresponding author

***Abstract*—Protein-protein interactions (PPI) play a critical role in most cellular processes and form the basis of biological mechanisms. With developments of high-throughput methods, vast amounts of PPI data are available which makes it possible to study biology systems at the network level. Recent developments have indicated that network theory is making an important contribution to the topological study of PPI networks. These networks have been shown to have some characteristic properties such as small-world effect, scale-free degree distribution and self-similarity. However, the unweighted PPI networks are far from being optimal because of the varying reliability of the interactions data. In this paper, we adopt the iterative scoring method to generate weighted PPI networks. By using the random sequential box covering algorithm, we calculate the fractal dimensions for both the original unweighted PPI networks and the generated weighted PPI networks. The results show that self-similarity is still present in generated weighted PPI networks. This implies that it is viable to expand the study of properties of complex networks to a wider field including more complex weighted networks and possibly directed complex networks.**

***Keywords-component; Protein-protein interaction networks; self-similarity; fractal scaling;***

## I. INTRODUCTION

Protein-protein interactions (PPI) play a critical role in most cellular processes and form the basis of biological mechanisms. In the post-genome era, the developments of high-throughput methods, such as yeast-two-hybrid and mass spectrometry, have produced vast amounts of PPI data, which makes it possible to study genes and proteins at the network level [1].

The simplest representation of protein-protein interaction networks takes the form of a mathematical graph consisting of nodes and edges. Proteins are represented as nodes and an edge represents a pair of proteins which physically interact. PPI networks are normally represented as unweighted graphs. However, given the varying reliability of interactions, these unweighted graphs are far from being optimal in representing the data [2,3]. More effective analysis would be achieved by considering weighted PPI networks in which each edge is associated with a weight representing the probability of an interaction. For this aim, many computational approaches have been proposed. Deane *et al.* [2] proposed two methods for assessing the overall quality of an interaction dataset. Patil and Nakamura [3] used a combination of genomic features including sequence, structure and gene ontology annotation to assign reliability to protein-protein interactions in Saccharomyces cerevisiae. Besides gene annotation, gene expression and sequence homology, several methods based on the topology of PPI networks have been proposed such as CDdistance [4], FSWeight [5] and the iterative scoring method [6]. These methods have been applied to predict more reliably protein interactions, essential proteins and protein complexes, etc. [2-6].

At the same time, topological properties of complex networks have attracted much attention in diverse areas of science. Many networks such as the World Wide Web, metabolic networks as well as PPI networks have been shown to share similar characteristic properties including the small-world property [7,8], the scale-free degree distribution [9-11] and self-similarity [12].

The **small-world property** means the characteristic path length $L$ and the number of nodes $N$ have the following relationship:

$$L \propto \log(N). \tag{1}$$

The small-world effect means that any two nodes can be connected via a short path of a few links [7]. A famous example is the so-called 'six degrees of separation' in social networks [8]. A large number of real networks are called '**scale-free**' because they show a power-law distribution of the number of links per node, i.e. the probability distribution of the number of links per node $P_k$ (also known as the degree distribution) satisfies a power-law $P_k \sim k^{-\gamma}$ with the degree exponent $\gamma$ varying in the range $2< \gamma <3$ [9]. Originated from the small-world property, it is widely believed that complex networks are not self-similar under a length-scale transformation. After analyzing a variety of real complex networks, Song *et al.* [12] found that they in fact consist of self-repeating patterns on all length scales, i.e., they have self-similar structures. In order to confirm the self-similarity of complex networks, Song *et al.* [12] analyzed the PPI networks of Homosapiens, the fruit fly D. melanogaster, the bacteria E.coli and H. pylori, the baker's yeast S. cerevisiae and the nematode worm C. elegans which are all available via the DIP database [18]. They calculated their fractal dimension, which is a widely used tool to characterize complex fractal sets [12-17]. The results indicated that among these organisms, E.coli and Homosapiens possess

self-similarity with nearly the same fractal dimension of 2.3. The other PPI networks also showed fractal scaling, but the estimates of their fractal dimensions have larger error range, which may be due to their incomplete data. Kim *et al.* [16] introduced another method of random sequential box covering and applied to a variety of complex networks.

However, these studies of self-similarity are mainly based on unweighted networks. So we might wonder if weighted networks still have the property of self-similarity and whether the generated weighted networks still have the same fractal dimension as the original unweighted networks.

In this paper, we consider the PPI networks of Homosapiens, E.coli (a bacterium), Arabidopsis Thaliana (a plant), C. elegans and baker's yeast S. cerevisiae from two databases. Firstly, we adapt the random sequential box-covering algorithm to calculate their fractal dimensions. Then, by using the iterative scoring method, we generate their weighted PPI networks and use the same algorithm to calculate the fractal dimensions of the generated weighted PPI networks. We will investigate the self-similarity in both PPI networks and generated weighted PPI networks.

## II. METHODS

In this section, we introduce the box-covering methods for calculating the fractal dimension of complex networks and an iterative scoring method for generating weighted PPI networks based on the original unweighted PPI networks.

### *A. The box covering method to calculate fractal dimension*

The box covering method is a basic tool to measure the fractal dimension of conventional fractal objects embedded in Euclidean space. However, such a method cannot be applied to real networks such as PPI networks because the Euclidean metric is not well defined for such networks. Song *et al*. [12, 14] studied the fractality and self-similarity in complex networks by using box covering techniques. They proposed several possible box covering algorithms [14] and applied them to a number of models and real-world networks. Meanwhile, Kim *et al* [15-17] introduced another method called the random sequential box-covering which shares a common spirit with the other algorithms introduced by Song *et al.* [14]. The random sequential box covering method contains a random process of selecting the position of the center of each box. In this study, we adapt the random sequential box-covering algorithm [16] to measure the fractal dimension of PPI networks. The details of this algorithm are as follows. For a given network $A$, let $N_B$ the number of boxes with linear size $r_B$ that are needed to cover the entire network. The fractal dimension $d_B$ is then given by

$$N_B \propto r_B^{-d_B} \tag{2}$$

By measuring the distribution of $N_B$ for different box sizes, the fractal dimension $d_B$ can be obtained by fitting the power law distribution through the following steps [16].

(i) Randomly select a node at each step, and the selected node would be the center or seed of a box;

(ii) Search the network at distance $r_B$ from the seed and cover all the nodes that have been found within distance $r_B$ but not covered yet. Assign newly covered nodes to the new box. If no newly covered nodes have been found, then the box is discarded.

(iii) Repeat (i) and (ii) until all the nodes in the network have been assigned to their respective boxes.

By using this algorithm, we calculated the fractal dimension for the same data of the human PPI network as in Kim *et al* [16] and obtained a similar fractal dimension of 2.20± 0.09. We then used this method to estimate the fractal dimension of the other PPI networks and their weighted PPI networks.

### *B. The iterative scoring method*

Many methods have been proposed to assess the reliability of protein interactions. These methods usually assign a score to each protein pair such that the higher the score is, the more likely the proteins interact with each other. Among these methods, CDdistance [4] and FSWeight [5] are measures calculated using the number of common neighbors of two proteins. They are initially proposed to predict protein functions, and have been shown to perform well for assessing the reliability of protein interactions.

The intuition behind the iterative scoring method is that if the score of an interaction reflects its reliability, then the scored interactions should better represent the actual interaction network than the initial binary ones, and we should be able to further improve score computation by re-computing the score of each protein pair using the scored interactions. Here, we use the AdjustCD distance [6] which is a variant of CDdistance to calculate the score of protein pairs.

A PPI network could be represented as an undirected network $G = (V, E)$, where the node set $V$ is the set of proteins and the edge set $E$ is the set of interactions between proteins. We use $u$, $v$, $x$ to denote individual nodes (proteins) and $(u, v)$ to denote the edge between node $u$ and node $v$. The neighbor set of a node $u$ in $G$, denoted as $N_u$, is defined as $N_u = \{v | (u, v) \in E\}$. For a given pair of proteins $u$ and $v$, the distance AdjustCD [6] of edge $(u, v)$ is defined as

$$AdjustCD(u,v) = \frac{2\left|N_u \cap N_v\right|}{\left|N_u\right| + \lambda_u + \left|N_v\right| + \lambda_v} \tag{3}$$

where $\lambda_u$ and $\lambda_v$ are used to penalize proteins with very few neighbors as in FSWeight [5] defined as

$$\lambda_u = \max\left\{0, \frac{\sum_{x\in V}\left|N_x\right|}{\left|V\right|} - \left|N_u\right|\right\} \tag{4}$$

$$\lambda_v = \max\left\{0, \frac{\sum_{x\in V}\left|N_x\right|}{\left|V\right|} - \left|N_v\right|\right\} \tag{5}$$

Based on this definition, if the degree of a node $u$ is below the average degree, then it is adjusted to the average degree.

The iterative version of AdjustCD is defined as follows:

$$w^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} (w^{k-1}(x,u) + w^{k-1}(x,v))}{\sum_{x \in N_u} w^{k-1}(x,u) + \lambda_u^k + \sum_{x \in Nv} w^{k-1}(x,v) + \lambda_v^k} \quad (6)$$

where $w^{k-1}(x,u)$ and $w^{k-1}(x,v)$ are scores of ($x$, $u$) and ($x$, $v$) respectively in the ($k$-1)-th iteration.

Initially, if there is an edge between $x$ and $u$ in the original PPI network, then $w^0(x, u) = 1$, otherwise, $w^0(x, u) = 0$. The two terms $\lambda_u^k$ and $\lambda_v^k$ are also defined based on weighted degree:

$$\lambda_u^k = \max\left\{0, \frac{\sum_{x \in V}\sum_{y \in N_x} w^{k-1}(x,y)}{|V|} - \sum_{x \in N_u} w^{k-1}(x,u)\right\} \quad (7)$$

$$\lambda_v^k = \max\left\{0, \frac{\sum_{x \in V}\sum_{y \in N_x} w^{k-1}(x,y)}{|V|} - \sum_{x \in N_v} w^{k-1}(x,v)\right\} \quad (8)$$

It is not difficult to see that $w^1(u,v)$ = AdjustCD ($u$, $v$). CD-distance and FSWeight can be iterated in a similar way. Liu *et al.* [6] showed that the iterative scoring method can improve functional homogeneity and localization coherence of top ranked interactions, and the iterative scoring method performs best when $k$ = 2, and the subsequent iterations do not improve the performance further. By using this method, we generate a weighted PPI network based on the original unweighted PPI network, and we take the score of each protein pair as the weight of the edge between them.

## III. RESULTS

The protein-protein interaction data used here are were downloaded from two databases. The PPI networks of **C.elegans** and **Arabidopsis thaliana** were downloaded from BioGRID [19]. The PPI networks of baker's yeast **S.cerevisiae**, **E.coli** and **Homosapiens** were downloaded from DIP [18].

Our fractal scaling analysis is based on connected networks, which mean there are no isolated nodes or all the nodes in the network must be reachable.

Firstly, we need to find the largest connected part of each PPI data. For this step, many tools and methods could be used. In our study, we adopt the **Cytoscape** [20] which is an open bioinformatics software platform for visualizing molecular interaction networks and analyze network graphs of any kind involving nodes and edges. With Cytoscape, we get the largest connected part of each interacting PPI network used in our fractal analysis.

Among the original five PPI networks (human, Arabidopsis thaliana, E.coli, yeast and C.elegans), self-similarity is apparent. By using the iterative scoring method, we then transform the PPI networks into weighted networks and calculate their fractal dimension. The fractal scaling of each PPI network and its weighted PPI network is showed in Figures. 1 -.5, where we use triangle (Δ) for the original network and circle (○) for the weighted network, together with their fitted lines. The fractal dimension is the absolute value of the slope of each fitted line. The fractal dimensions of weighted PPI networks are slightly smaller than those of the original PPI networks. The numerical results of fractal scaling for the original PPI networks and their weighted PPI networks are summarized in Tables I and II respectively. For each PPI network, $N$ is the number of nodes of the largest connected part, $d_B$ is the fractal dimension with error range.

TABLE I. NUMERICAL RESULTS OF FRACTAL SCALING FOR THE ORIGINAL PPI NETWORKS

| PPI | database | $N$ | $d_B$ | error |
|---|---|---|---|---|
| Human | DIP | 503 | 2.20 | ± 0.09 |
| E.coli | DIP | 642 | 2.37 | ± 0.11 |
| Yeast | DIP | 1922 | 2.90 | ± 0.20 |
| C.elegans | BioGRID | 3343 | 3.48 | ± 0.24 |
| Arabidopsis Thaliana | BioGRID | 1298 | 2.26 | ± 0.06 |

TABLE II. NUMERICAL RESULTS OF FRACTAL SCALING FOR THE WEIGHTED PPI NETWORKS

| Weighted PPI | database | $N$ | $d_B$ | error |
|---|---|---|---|---|
| Human | DIP | 417 | 1.85 | ± 0.04 |
| E.coli | DIP | 451 | 1.98 | ± 0.07 |
| Yeast | DIP | 1713 | 2.06 | ± 0.04 |
| C.elegans | BioGRID | 2444 | 2.04 | ± 0.05 |
| Arabidopsis Thaliana | BioGRID | 800 | 2.25 | ± 0.10 |

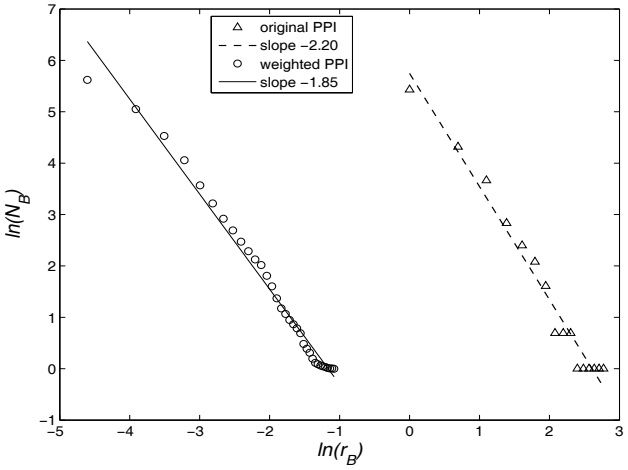


Figure 1. fractal scaling of the Homosapien PPI

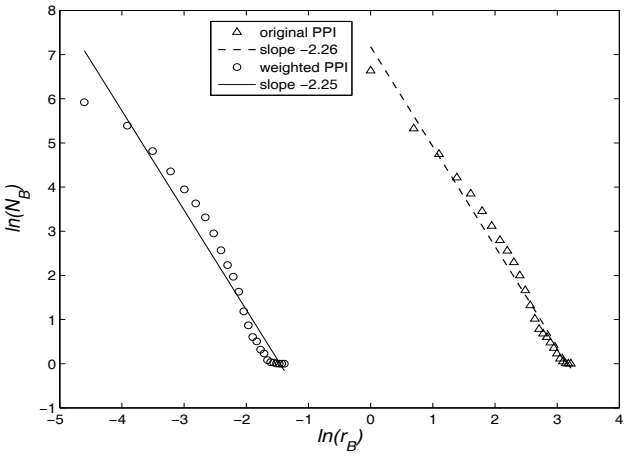


Figure 2. Fractal scaling of the Arabidopsis thaliana PPI

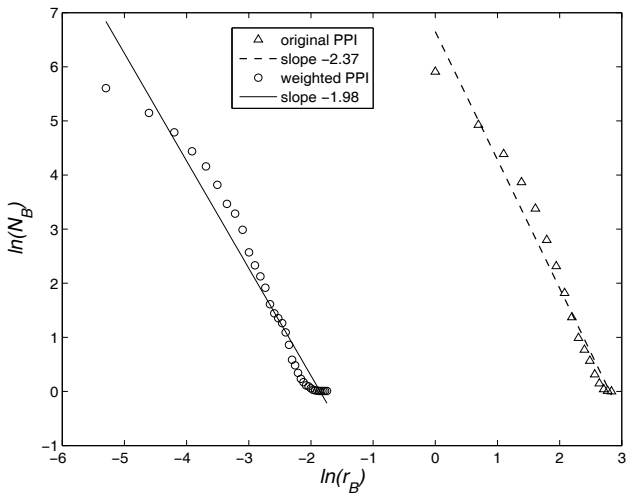


Figure 3. Fractal scaling of the E.coli PPI

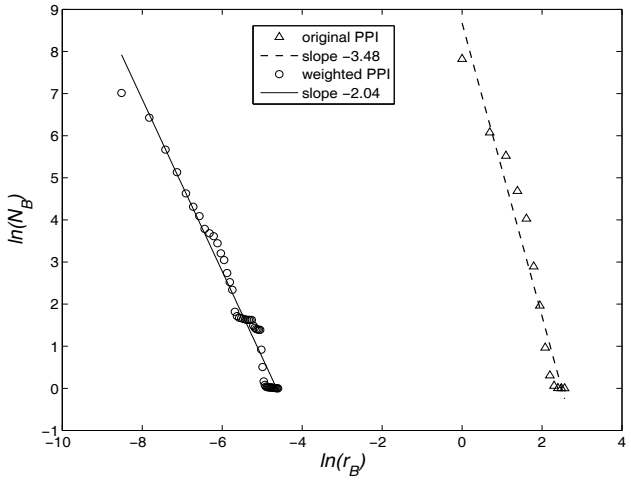


Figure 4. Fractal scaling of the C.elegans PPI

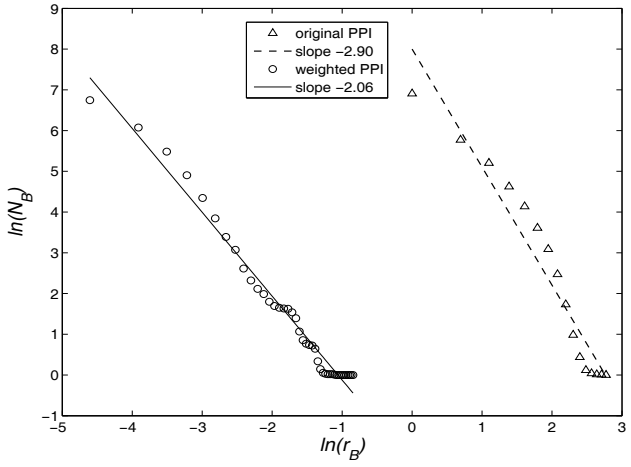


Figure 5. Fractal scaling of the yeast PPI

## IV. CONCLUSION

In conclusion, self-similarity has been found in weighted PPI networks. In this paper, we first use the iterative scoring method to generate weighted PPI networks based on the original PPI networks and then calculate their fractal dimensions. For five PPI networks of Homosapiens, E. coli, Arabidopsis Thaliana, C. elegans and baker's yeast S. cerevisiae, we demonstrate that self-similarity exists in both the PPI networks and their weighted networks. The fractal dimensions of weighted PPI networks are slightly smaller than those of the original PPI networks. We have successfully applied the box-covering algorithm to perform fractal analysis on weighted PPI networks. This suggests that the study of self-similarity of complex networks can be expanded to a wider field including weighted complex networks and possibly directed complex networks.

## ACKNOWLEDGEMENT

This project was supported by the Australian Research Council (Grant No. DP0559807), the Natural Science Foundation of China (Grant No. 11071282), the Chinese Program for New Century Excellent Talents in University (Grant No. NCET-08-06867), the Program for Furong Scholars of Hunan province of China, the Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province of China, and a China Scholarship Council-Queensland University of Technology Joint Scholarship.

## REFERENCES

[1] P. Uetz, et al, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," Nature, 2000, 403, pp. 623-627

[2] C. M. Deane, L. Salwinski, L. Xenarios and D. Eisenberg, "Protein interactions: Two methods for assessment of the reliability of high throughput observations," Molecular & Cellular Proteomics, 2002, 1. pp. 349-356

[3] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features," BMC Bioinformation, 2005, 6, 100

[4] C. Brun, F.Chevenet, D. Martin, J. Wojcik, A.Guenoche and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network," Genome Biol., 2003, 5, R6

[5] H.N.Chua, K. Ning, W.K.Sung, H.W.Leong and L. Wong, "Using indirect protein-protein interactions for protein complex predication," J. Bioinform. Comput. Biol., 2008, 6, pp. 435-466

[6] G.Liu, L.Wong and H.N.Chua, "Complex discovery from weighted PPI networks,"Systems biology, 2009, 25, pp. 1891-1897

[7] P.Erdos and A.Renyi, "On the evolution of random graphs," Publ. Math. Inst. Hung. Acad. Sci., 1960, 5, pp. 17-61

[8] S. Milgram, "The small-world problem," Psychol. Today. 1967, 2, pp. 60-67

[9] R.Albert, H.Jeong and A.L.Barabasi, "Diameter of the World Wide Web," Nature, 1999, 401, pp. 130-131

[10] R.Albert and A.L.Barabasi, "Statistical mechanics of complex networks," Rev.Mod. Phys., 2002, 74, pp.47-97

[11] M.Faloutsos, P.Faloutsos and C.Floutsos, "On power-law relationships of the internet topology," Comput. Commun. Rev., 1999, 29, pp. 251-262

[12] C. Song, S Havlin and H. A. Makse, “Self-similarity of complex networks,” Nature, 2005. 433. pp. 392-395

[13] C. Song, L. K. Galos, S. Havlin and H.A Makes, “Origins of fractality in the growth of complex networks,” Nature Physics, 2006, 2, pp. 275-281

[14] C. Song, K. G. Lazaros, S. Havlin and H. A. Makes, “How to calculate the fractal dimension of a complex network: the box covering algorithm,” Journal of Statistical Mechanics: Theory and Experiment, 2007, 3, 03006

[15] J. Kim, B. Kahng and D. Kim, “A box-covering algorithm for fractal scaling in scale-free network,” Chaos, 2007, 17, 026116

[16] J. Kim, K-I Goh, G. Salvi, E. Oh, B. Kahng and D. Kim, “Fractality in complex networks: Critical and supercritical skeletons,” Phys. Rev. E. 2007, 75, 016110

[17] J. Kim, B. Kahng and D. Kim, “A box-covering algorithm for fractal

[18] Database of Interacting Proteins (DIP). http://dip.doe-mbi.ucla.ed

[19] BioGRID:http://thebiogrid.org/download.php

[20] Cytoscape software:http://cytocscapeweb.cytoscape.org/