Multifractal analysis on gene and PPI networks

Danling Wang[1], Yanfei Wang[2*]

[1]*School of Mathematics and Physics, University of Science and Technology, Beijing, China*

[2]*School of Sciences China Agricultural University Tsinghuadonglu 17, 100083, Beijing, China*

Received 1 August 2014, www.tsi.lv

Abstract

Multifractal analysis is a useful way to systematically describe the spatial heterogeneity of both theoretical and experimental fractal patterns. In this paper, we introduce a new box-covering algorithm to compute the generalized fractal dimensions of complex networks. We apply our method on networks built on disease-related gene microarray data and PPI networks. For each microarray data, we compare the difference of multifractal behaviour between gene networks that reconstructed from patients and normal micorarrays. The result suggests that multifractality exists in all the gene networks we generated and the differences in the shape of the D_q curves are obvious for all microarray data sets. Meanwhile, multifractal analysis could provide a potentially useful tool for gene clustering and identification between healthy people and patients. For the analysis of PPI networks, the results support that the algorithm is a suitable and effective tool to perform multifractal analysis of complex networks, and this method can be a useful tool to cluster and classify real PPI networks of organisms.

Keywords: multifractal analysis, self-similarity, gene networks, PPI networks

1 Introduction

Complex networks have been studied extensively due to their relevance to many real-world systems such as the World Wide Web, the internet, energy landscapes, and biological and social systems. After analysing a variety of real complex networks, Song et al. [1] found that they consist of self-repeating patterns on all length scales, i.e., they have *self-similar structures*. In order to unfold the self-similar property of complex networks, Song et al. [1] calculated their fractal dimension, a known useful characteristic of complex fractal sets [2-4], and found that the box counting method is a proper tool for further investigations of network properties. The tools of fractal analysis provide a global description of the heterogeneity of an object. However, this approach is not adequate when the object may exhibit a multifractal behaviour. Multifractal analysis is a useful way to systematically characterize the spatial heterogeneity of both theoretical and experimental fractal patterns. It was initially proposed to treat turbulence data, and has recently been applied successfully in many different fields including time series analysis [5], financial modelling [6], biological systems [7,8] and geophysical systems [9].

In recent years, bioinformatics has become a more and more notable research field since it allows biologists to make full use of the advances in computer science and computational statistics in analysing the data of an organism at the genomic, transcriptomic and proteomic levels [10]. DNA technology, i.e. microarray of large sets of nucleotide sequences, is a modern tool that is used to obtain information about expression levels of thousands of genes simultaneously. The gene networks built based on microarray data become a popular research field.

In this paper, we aim to compare the difference of multifractal behaviours between gene networks that reconstructed from patients and normal people microarrays and some PPI networks. However, work in such high dimensional and large data is extremely difficult. So for gene microarrays, firstly, we apply *Fuzzy Membership test* (FM-test) [11] to get the most important genes that are related with the disease; then we construct networks based on the microarray data of the selected genes by calculating the correlation coefficient. Next we apply the modified fixed-size box-covering method on them to detect their multifractal behaviours. For PPI networks, we firstly use cytoscape to we need to find the largest connected part of each data set, and then we adopt our method to check multifractal characteristics in different organisms. Secondly, we also we randomly chose several sub-networks from different parts of the human PPI network with the same nodes and compare multifractal characteristic between them.

2 Methods

The most common algorithms of traditional multifractal analysis are the fixed-size box-counting algorithms [12]. For a given measure μ with support E in a metric space, we consider the partition sum:

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q, q \in R,$$

* ***Corresponding author*** e-mail: yfmu@sina.com

where the sum is evaluated over all different nonempty boxes B of a given size ε in a grid covering of the support T. The exponent $\tau(q)$ is defined by:

$$\tau(q) = \lim_{\varepsilon \to 0} \frac{\ln Z_\varepsilon(q)}{\ln \varepsilon}$$

and the generalized fractal dimensions of the measure are defined as:

$D(q) = \tau(q)/(q-1)$, for $q \neq 1$,

and:

$$D(q) = \lim_{\varepsilon \to 0} \frac{\ln Z_{1,\varepsilon}}{\ln \varepsilon}, \text{ for } q = 1,$$

where $Z_{1,\varepsilon} = \sum_{\mu(B)\neq 0} \mu(B) \ln \mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of $(\ln Z_\varepsilon(q))/(q-1)$ against $\ln \varepsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\varepsilon}$ against $\ln \varepsilon$ for $q = 1$. The $D(q)$ corresponding to negative values of q deal with the structure and the properties of the regions where the measure value is small.

Our group proposed a new box-covering algorithm to compute the generalized fractal dimensions of network [13]. For a network, we denote the matrix of shortest path lengths by $B = (b_{ij})_{N\times N}$, where b_{ij} is the length of the shortest path between nodes i and j. Then we use $B = (b_{ij})_{N\times N}$ as input data for multifractal analysis based on our fixed-size box counting algorithm as follows:

a) Initially, all the nodes in the network are marked as uncovered and no node has been chosen as a seed or centre of a box.
b) Set t=1, 2,…, T appropriately. Group t nodes into T different ordered random sequences. More specifically, in each sequence, nodes which will be chosen as seed or centre of a box are randomly arrayed.
   ***Remark*: T** is the number of random sequences and is also the value over which we take the average of the partition sum $\bar{Z}_r(q)$. In this study, we set T=1000 for all the networks in order to compare them.
c) Set the size of the box in the range $r \in [1, d]$, where d is the diameter of the network.
   Remark: When r=1, the nodes covered within the same box must be connected to each other directly. When r=d, the entire networks could be covered in only one box no matter which node was chosen as the centre of the box.
d) For each centre of a box, search all neighbours within distance r and cover all nodes which are found but have not been covered yet.
e) If no newly covered nodes have been found, then this box is discarded.
f) For the nonempty boxes B, we define their measure as $\mu(B) = N_B / N$, where N_B is the number of nodes covered by the box B, and N is the number of nodes of the entire network.
g) Repeat (d) until all nodes are assigned to their respective boxes.
h) Repeat (c) and (d) for all the random sequences, and take the average of the partition sums $\bar{Z}_r(q) = (\sum^t Z_r(q))/T$, and then $\bar{Z}_r(q)$ for linear regression.

Linear regression is an essential step to get the appropriate range of $r \in [r_{\min}, r_{\max}]$ and to get the generalized fractal dimensions D_q. In our approach, we run the linear regression of $[\ln \bar{Z}_r(q)]/(q-1)$ against $\ln(r/d)$ for $q \neq 1$, and similarly the linear regression of $\bar{Z}_{1,r}$ against $\ln(r/d)$ for q=1, where $\bar{Z}_{1,r} = \sum_{\mu(B)\neq 0} \mu(B) \ln \mu(B)$ and d is the diameter of the network.

3 Results and discussion

3.1 MULTIFRACTAL ANALYSIS OF GENE NEWORKS

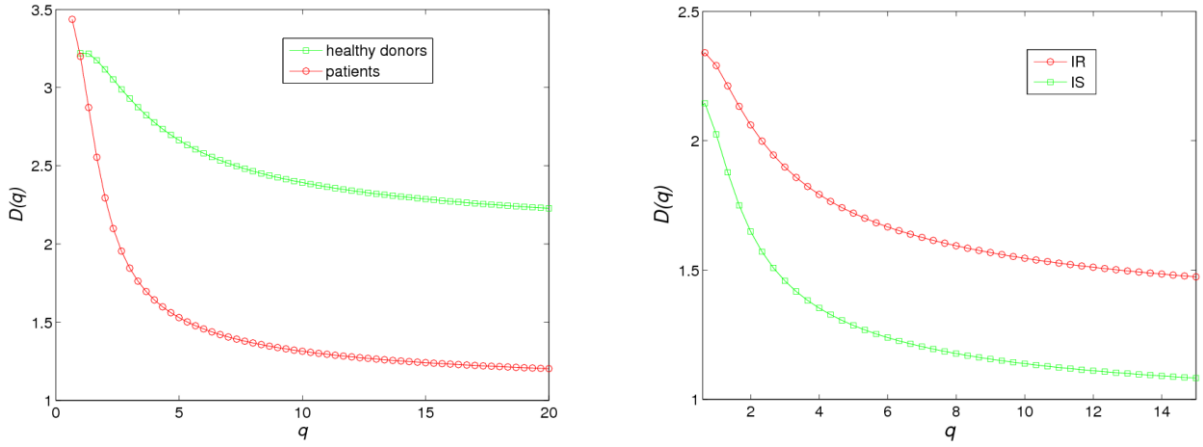
Four different gene microarray data sets are used in our work: Colorectal cancer gene microarray data [16]; Type II diabetes gene microarray data [17]; Type I diabetes gene microarray data [18] and Lung cancer gene microarray data [11]. There are two parts in each data, the first part consists of genes expression values from patients or drugs sensitive people; the second part consists of genes expression values from healthy donators or patients after medication or treatment. For each original data, we firstly use FM-test [11] to select around 2000 genes which are most possibly related with disease, then build patients gene networks and normal people gene networks respectively. For colorectal data, CP is the patient gene networks, HP is healthy people gene network; for type II diabetes data, IR is insulin resistant people network and IS is insulin sensitive people network; for type I diabetes data, TP is patients network and TM is patients after medication network; for lung cancer data, LP is patients network and LN is normal people network. We analysed multifractal behaviour of two networks for each microarray data. All the networks are full connected.

We calculated the D_q spectra for these gene networks of different datasets and then summarize the numerical results in Table 1 including the number of nodes (N), number of the threshold t, maximum value of D_q, limit of D_q, and ΔD_q. Figures 1 and 2 show that the generalized fractal dimension D_q of these gene networks are decreasing functions of q and multifractality exists in these networks. From the table and figures we see, multifractal characteristic exists in all the gene networks

we analysed. Meanwhile, the D_q curves of gene networks from normal people are mostly higher than the ones from patients, especially for the first three microarray data.

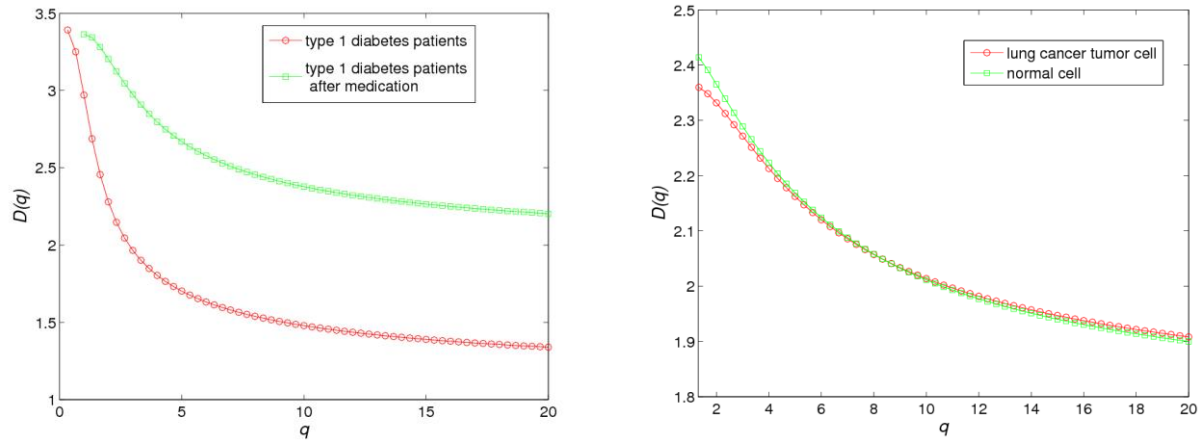
TABLE 1 Numerical results on gene networks

Networks	*N*	*t*	Max D_q	Lim D_q	ΔD_q
CP	*2000*	*0.72*	*3.44*	*1.20*	*2.24*
CH	2000	0.85	3.22	2.23	0.99
IR	2304	0.95	2.34	1.47	0.87
IS	2299	0.95	2.14	1.08	1.06
TP	2000	0.72	3.39	1.3	2.05
TH	2000	0.81	3.36	2.20	1.16
LC	2000	0.97	2.36	1.91	0.45
LN	2000	0.96	2.43	1.90	0.53



a) b)

Figure 1 a) Colorectal cancer microarray data, b) Type II diabetes microarray data



a) b)

Figure 2 a) Type I diabetes microarray data; () lung cancer microarray data

3.2 MULTIFRACTAL ANALYSIS OF PPI NEWORKS

The protein-protein interaction data we used here are mainly downloaded from two databases: the PPI networks of Drosophila melanogaster (fruit fly), C.elegans, Arabidopsis thaliana (a type of plant) are downloaded from BioGRID. The PPI networks of S. cerevisiae (baker's yeast), E.coli and H.pylori are downloaded from DIP [19]. We also use the same human PPI network data as in Lee and Jung [20].

Our fractal and multifractal analyses are based on connected networks which do not have separated parts or isolated nodes. In order to apply them to protein-protein interaction networks, some preparation is needed in advance. Firstly, we need to find the largest connected part of each data set. For this purpose many tools and methods could be used. In our study, we adopt Cytoscape [21] which is an open bioinformatics software platform for visualizing molecular interaction networks and analysing network graphs of any kind involving nodes and edges. In using Cytoscape, we could get the largest

connected part of each interacting PPI data set and this connected part is the network on which fractal and multifractal analyses are performed.

We calculated the D_q spectra for seven PPI networks of different organisms and summarize the corresponding numerical results in Table 2 including the number or nodes (N), number of edges (E), diameter of the network (d), maximum value of D_q, limit of D_q, and ΔD_q. These results show multifractality exists in PPI networks.

TABLE 2 Numerical results of Protein-protein interaction networks

Networks	N	E	d	Max D_q	Lim D_q	ΔD_q
Human	8934	41341	14	4.89	2.65	2.24
D.melanogaster	7476	26534	11	4.84	2.87	1.97
S.cerevisiae	4976	21875	10	4.62	2.48	2.14
E.coli	2516	11465	12	4.15	2.10	2.05
H.pylori	686	1351	9	3.47	1.91	1.56
Arabidopsis thaliana	1298	2767	25	2.51	1.62	0.88
C.elegans	3343	6437	13	4.47	1.49	2.98

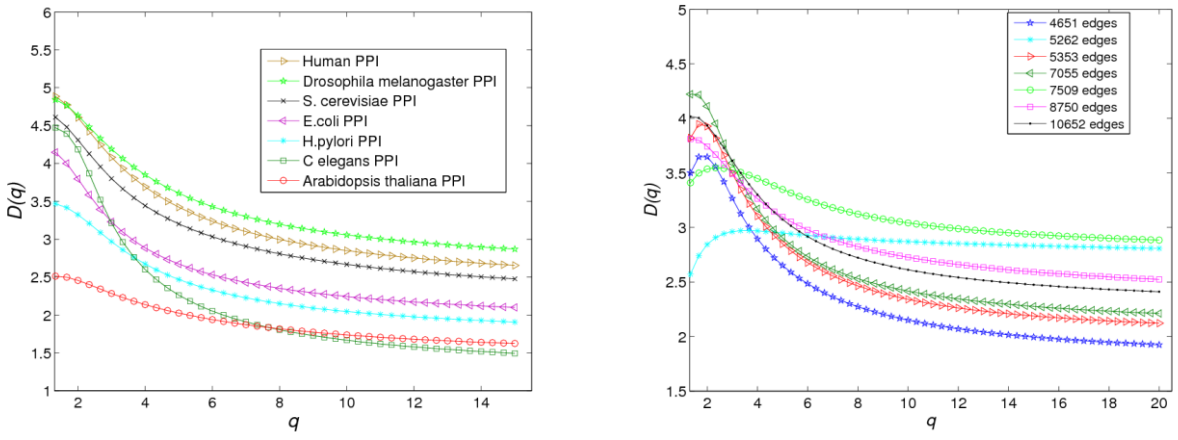
TABLE 3 Numerical results of sub-networks of Human PPI

Networks	N	E	d	Max D_q	Lim D_q	ΔD_q
Subnetwork of Human PPI	3505	4651	24	3.65	1.99	1.66
Subnetwork of Human PPI	3505	5262	27	2.97	2.83	0.14
Subnetwork of Human PPI	3505	5353	22	3.95	2.19	1.76
Subnetwork of Human PPI	3505	7055	15	4.22	2.28	1.94
Subnetwork of Human PPI	3505	7509	15	3.55	2.94	0.61
Subnetwork of Human PPI	3505	8750	16	3.81	2.59	1.22
Subnetwork of Human PPI	3505	10652	10	4.02	2.47	1.55

From Figure 3a we could see that all PPI networks are multifractal and there are two clear groupings of organisms based on the peak values of their D_q curves. The first group includes human, Drosophila melanogaster, S.cerevisiae, and C.elegans. The second group just includes two bacteria E.coli and H. pylori. We could also see that the PPI networks of the seven organisms have similar shape for the D_q curves. They reach their peak values around q = 2, then decrease sharply as q > 2 and finally reach their limit value when q > 10. So we can take limDq = D(20) and use ΔD_q = maxD_q – limD_q to verify how the D_q function changes along each curve.

Then we randomly chose several sub-networks from different parts of the human PPI network. These sub-networks all contain 3505 nodes and different numbers of edges. Since these sub-networks are chosen randomly, overlapping between them is allowed. Then we calculated the D_q spectra for sub-networks of human protein-protein interaction network [20] and summarize the corresponding numerical results in Table 3 including the number or nodes (N), number of edges (E), diameter of the network (d), maximum value of D_q, limit of D_q, and ΔD_q. These results show multifractality exists in PPI networks.

From Figure 3b we could see that not all parts of a PPI network have the same multifractal behaviour. More specifically, among these sub-networks, the ΔD_q values vary from one to another which means that the edge distribution of some parts of a network is symmetric while that of the other parts may not be. This may help to understand the diversity and complexity of protein-protein interactions.



a) b)

Figure 3 (a) The D_q curves for PPI networks; (b) The D_q curves for sub-networks of human PPI networks

4 Conclusions

A modified algorithm for analysing the multifractal behaviours of complex networks is introduced in this paper. We apply this modified fixed-size box-covering method on gene networks reconstructed from patients and normal gene microarrays. Firstly, we use the fuzzy membership test to get the most important genes that related with the disease; then we construct networks based on the microarray data of the selected genes by calculated the correlation coefficient. From the results we see, multifractality exists in all the gene networks we generated and the difference in the shape of the D_q curves are obvious for these microarray datasets. Thus multifractal analysis could provide a potentially useful tool for gene clustering and identification between healthy people and patients. We also apply our method on some PPI networks, these results support that multifractal analysis can be a useful tool to cluster and classify real networks such as the PPI networks of organisms.

Acknowledgment

This work is supported by Chinese Universities Scientific Fund No. 2013XJ010 and the Fundamental Research Funds for the Central Universities No. FRF-TP-13-020A.

References

[1] Song C, Havlin S, Makse H A 2005 Self-similarity of complex networks *Nature London* **433** 392-5
[2] Mandelbrot B B 1983 The Fractal Geometry of Nature *Academic Press* New York
[3] Feder J 1988 Fractals *PlenumPress* New York
[4] Falconer K 1997 Techniques in Fractal Geometry *Wiley* New York
[5] Cansessa E 2000 Multifractality in time series *J Phys A: Math Gen* **33** 3637-51
[6] Anh V V, Tieng Q M, Tse Y K 2000 Cointegration of stochastic multifractals with application to foreign exchange rates *Int Trans Opera Res* **7** 349-63
[7] Yu Z G, Anh V, Lau K S 2001 Multifracal characterisation of length sequences of coding and noncoding segments in a complete genome *Physica A* **301** 351-61
[8] Yu Z G, Anh V, Lau K S 2001 Measure representation and multifractal analysis of complete genome *Phys Rev E* **64** 31903
[9] Kantelhardt J W, Koscielny-Bunde E, Rybski D, Braun P, Bunde A, Havlin S 2006 Long-term persistence and multifractality of precipitation and river runoff records *J Gephys Res* **111** D01106
[10] Keedwell E, Naravanan 2005 Introduction to Artificial Intelligence and Computer Science, in Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems *John Wiley & Sons Ltd* Chichester UK
[11] Liang LR, Wang S X, Lu Y, Mandal V, Patacsil D, Kumar D (2006) FM-test: a fuzzy-set-theory-based approach to differential expression data analysis *BMC Bioinformatics* **7** (S4)
[12] Halsey T C, Jensen MH, Kadanoff LP, Procaccia I, Shraiman B I 1986 Fractal measures and their singularities: the characterization of strange sets *Phys Rev A* **33** 1141-51
[13] Wang D L, Yu Z G, Anh V 2012 Multifractal analysis of complex networks *Chin Phys B* **21** 080504
[14] Werhli A V, Grzegorczyk M, Husmeier D 2006 Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian models *Bioinformatics* **22** 2523-31
[15] Borate B R, Chesler E J, Langston M A, Saxton A M, Voy B H 2009 Comparison of threshold selection methods for micorarray gene co-expression matrices *BMC Research Notes* **2** 240
[16] Collado M, Garcia V, Garcia J M, Alonso O, Lombardia L, Diaz-Uriarte R, Fernandez L A, Zaballos A, Bonilla F, Serrano M 2007 Genomic profiling of circulating plasma RNA for the analysis of cancer *Clin. Chem.* **53**(10) 1860-3
[17] Yang X, Pratley R E, Tokraks S, Bogardu C, Permana P A 2007 Microarray profiling of skeletal muscle tissues from equally obese, non-diabet insulin-sensitive and insulin-resistant Pima Indians *Diabetolotgia* **45** 1584-93
[18] van Oostrom O, de Kleijn D P V, Fledderus J O, Pescatori M, Stubbs A, Tuinenburg A, Lim S K, Verhaar M C 2009 Folic acid supplementation normalizes the endothelial progenitor cell transcriptome of patients with type I diabetes: a case-control pilot study *Cardiovascular Diabetology* **8** 47
[19] DIP: http://dip.doe-mbi.ucla.edu/
[20] Lee C Y, Jung S 2006 Statistical self-similar properties of complex networks *Phys Rev E* **73**(6) 066102
[21] Cytoscape software: http://cytoscapeweb.cytoscape.org/

Authors

Danling Wang, born in 1982, Hebei Province, China

Current position, grades: lecturer at the University of Science and Technology Beijing.
University studies: PhD at Queensland University of Technology (2008-2011).
Scientific interest: fractal and multifractal analysis, chaos systems and dynamics.
Publications: 8

Weiwei Zhu,born in September, 1988, Qingdao, China

Current position, grades: lecturer at China Agricultural University.
University studies: PhD at Queensland University of Technology (2008-2011).
Scientific interest: fuzzy sets theory, data mining, machine learning.
Publications: 9