Available online at www.sciencedirect.com
SCIENCE DIRECT
Physics Letters A 357 (2006) 420–423
PHYSICS LETTERS A
www.elsevier.com/locate/pla

Application of random matrix theory to biological networks

Feng Luo [a,d], Jianxin Zhong [b,c,*], Yunfeng Yang [c], Richard H. Scheuermann [d], Jizhong Zhou [e,c,*]

[a] *Department of Computer Science, Clemson University, 100 McAdams Hall, Clemson, SC 29634, USA*
[b] *Department of Physics, Xiangtan University, Hunan 411105, China*
[c] *Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[d] *Department of Pathology, U.T. Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, TX 75390-9072, USA*
[e] *Department of Botany and Microbiology, University of Oklahoma, Norman, OK 73019, USA*

Received 9 February 2006; received in revised form 20 April 2006; accepted 21 April 2006

Available online 2 May 2006

Communicated by C.R. Doering

Abstract

We show that spectral fluctuation of interaction matrices of a yeast protein–protein interaction network and a yeast metabolic network follows the description of the Gaussian orthogonal ensemble (GOE) of random matrix theory (RMT). Furthermore, we demonstrate that while the global biological networks evaluated belong to GOE, removal of interactions between constituents transitions the networks to systems of isolated modules described by the Poisson distribution. Our results indicate that although biological networks are very different from other complex systems at the molecular level, they display the same statistical properties at network scale. The transition point provides a new objective approach for the identification of functional modules.

PACS: 05.45.-a; 87.10.+e

Keywords: Random matrix; Biological network; Cell; Protein; Gene

The cell is a complex system that contains numerous functionally diverse elements, including protein, DNA, RNA and small molecules. Understanding the fundamental principles and behavioral properties of the cell as a system has become a key research activity in the post-genomic era. Research on the topological properties of large scale networks of cell constituents has shown that biological networks share some fundamental topological properties, including scale-free, small-world, hierarchical, modular [1] and self-similar [2] properties, with other complex systems, such as the internet and social networks. Inspired by the electrical engineering paradigm, small gene circuit descriptions combined with mathematical modeling have been utilized to understand small subsystems of cellular processes [3]. Unfortunately, the huge number of constituents and their complex relationships in the cell make the mathematical modeling of large-scale biological systems challenging. It is of significant importance to understand the nature of the structure and interactions of biological networks for achieving quantitative description of their functions.

In this Letter, we use RMT to analyze the structure and interactions of biological networks. RMT, initially proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei [4], is a powerful approach for the identification and modeling of phase transitions and dynamics in physical systems. It has been successfully used to study the behaviors of complex systems, such as spectral properties of large atoms [5], metal insulator transitions in disordered systems [6], spectra of quasiperiodic systems [7,8], chaotic systems [9], brain responses [10], and the stock market [11]. One of the essential statistical properties in the RMT is eigenvalue fluctuation. For real and symmetrical random matrices that represent the time-reversal invariant complex systems, the eigenvalue fluctuations follow two universal laws depending on the correlation prop-

* Corresponding authors.
E-mail addresses: zhongjn@ornl.gov (J. Zhong), zhouj@ornl.gov (J. Zhou).

0375-9601/$ – see front matter
doi:10.1016/j.physleta.2006.04.076

erty of eigenvalues. Strong correlation of eigenvalues induced by strong interactions of matrix components leads to eigenvalue fluctuations described by the GOE. On the other hand, eigenvalue fluctuations for a random matrix with various decoupled components follow Poisson distribution due to absence of correlation of large number of eigenvalues of different components.

In this study we have found that the spectral fluctuation of a yeast protein–protein interaction network and a yeast metabolic network is described by the GOE statistics. Furthermore, we demonstrate that while each of these global networks belong to the GOE, removal of interactions between constituents identifies a transitions in which the spectral fluctuation approximates the Poisson statistics of RMT resulting in a decoupled network composed of isolated modules. Such a transition provides a new objective approach for the identification of functional modules within global biological networks.

We used the standard spectral unfolding technique in our study. In general, the density of eigenvalues of a matrix varies with its eigenvalue E_i $(i = 1, 2, 3, \ldots, N)$, where N is the order of the matrix. In order to observe the universal eigenvalue fluctuations of different matrices, random matrix theory requires spectral unfolding to have a constant density of eigenvalues. To fulfill this, one can replace E_i by the unfolded spectrum e_i, where $e_i = N_{\text{av}}(E_i)$ and N_{av} is the smoothed integrated density of eigenvalues obtained by fitting the original integrated density to a cubic spline or by local density average. With the unfolded eigenvalues, We calculated the nearest neighbor spacing distribution (NNSD) of eigenvalues, $P(s)$, which is defined as the probability density of unfolded eigenvalue spacing $s = e_{i+1} - e_i$. We know from RMT that $P(s)$ for the GOE statistics closely follows the Wigner–Dyson distribution

$$P_{\text{GOE}}(s) \approx \frac{1}{2}\pi s \exp\bigl(-\pi s^2/4\bigr).$$

In the case of Poisson statistics, $P(s)$ is given by the Poisson distribution

$$P_{\text{Poisson}}(s) = \exp(-s).$$

One difference between the Wigner–Dyson and Poisson distributions is their behavior at small values of s, where: $P_{\text{GOE}}(s \to 0) = 0$ and $P_{\text{Poisson}}(s \to 0) = 1$.

We applied the random matrix theory to two biological networks of yeast. The first network is the core protein interaction network of yeast obtained from the DIP [12] database (version ScereCR20041003) generated from the filtering of large high-throughput protein interaction data using two different computational methods [13]. After removal of all self-connecting links, the final protein interaction network includes 2609 yeast proteins and 6355 interactions. The second network is the yeast metabolic network constructed by Jeong et al. [14] from the data in the WIT database [15]. After removal of redundant links, the final metabolic network has 1511 chemical substrate and intermediate states and 3807 interactions. The original metabolic network is a directed network. To make the metabolism network symmetric for RMT study, we changed the directed network to an undirected network by replacing all directed links in the network with undirected links.

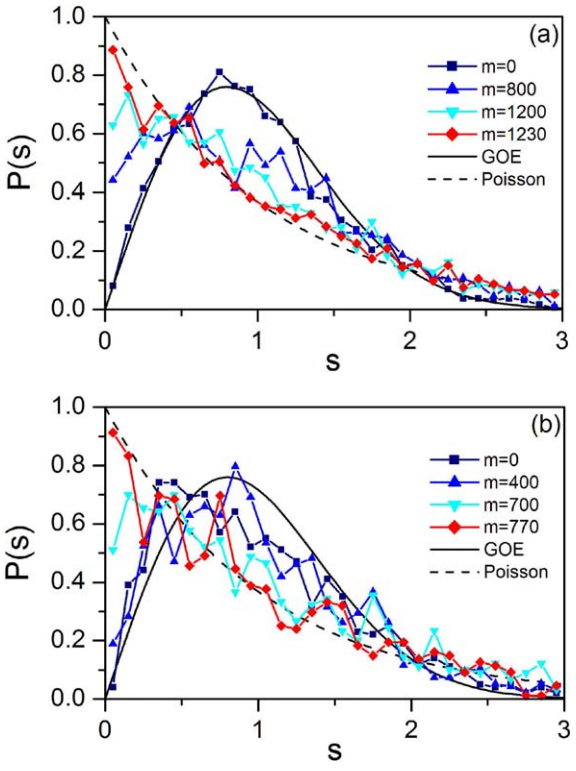

Fig. 1. The NNSDs of yeast biological networks. Smooth and dashed black lines are the GOE distribution and the Poisson distribution, respectively. (a) Yeast core protein–protein interaction networks with different number of removed links (m): 0 (navy), 800 (blue), 1200 (cyan), and 1230 (red). (b) Metabolic networks with different number of removed links (m): 0 (navy), 400 (blue), 700 (cyan), and 770 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In our RMT analysis, the two biological networks are represented by two real symmetric matrices. The dimension of each matrix is the number of constituents in the network. The elements in the matrices are set to 1 if there is a direct interaction between the constituents; otherwise, the elements are set to 0. We calculated the NNSD of these two matrices for RMT analysis by direct diagonalization of the matrix. Fig. 1 shows the NNSDs of these two networks. One can see that the NNSDs of the protein interaction network are well described by the Wigner–Dyson distribution. The NNSDs of the metabolism network are also very close to the Wigner–Dyson distribution, especially in the region representing small values of s. The slight deviation may be due to the incomplete nature of the defined network.

Biological networks have modular structures with more interactions between elements inside the same module and fewer interactions between different modules [16–18]. Girvan and Newman [19] proposed the concept of edge (or link) betweenness for describing the modular structure of a complex network, which is defined as the number of shortest connection paths between all pairs of network vertices that run through the edge. Edges between different modules tend to have more shortest-paths running through them compared with the edges inside

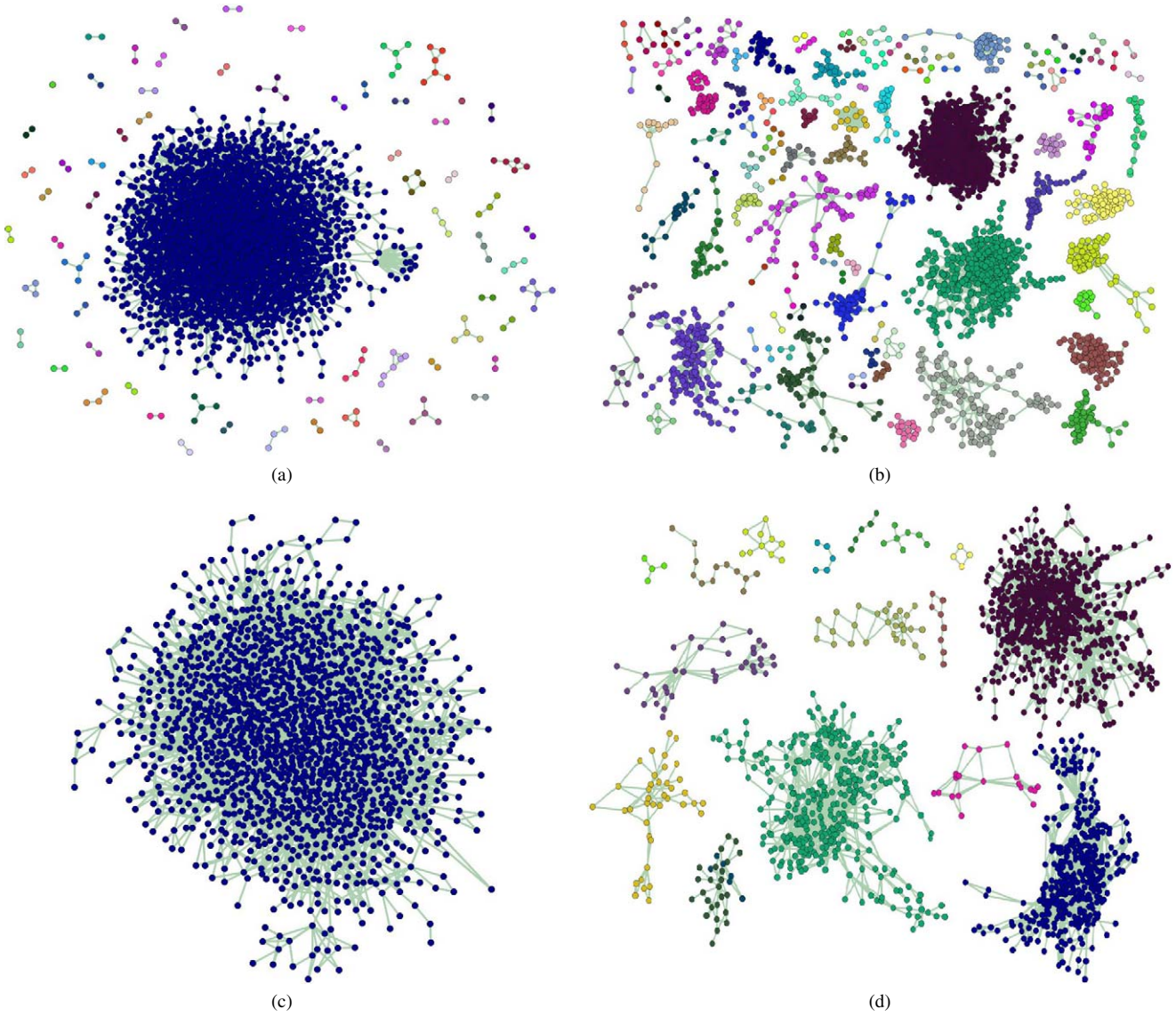

Fig. 2. Graph view of the yeast biological networks. (a) The original yeast core protein interaction network. (b) The yeast core protein interaction network with 1230 links removed. (c) The original yeast metabolic network. (d) The yeast metabolic network with 770 links removed. The graphs were produced using Biolayout [20].

any modules, and thus have higher betweenness values. Gradual deletion of edges with higher betweenness can separate the network while keeping its modules intact. The Girvan and Newman algorithm [19] for identifying modules in a network can be simply stated as follows: (1) calculation of the betweenness for all edges in the network; (2) removal of the edge with the highest betweenness; (3) recalculation of betweennesses for all edges affected by the removal; and (4) repetition of removal until no edges remain. To test the modularity of these two networks, we gradually removed the interaction links between the constituents in the two yeast networks using the Girvan–Newman algorithm and calculated the NNSDs of the remaining networks. A transition of NNSD from a Wiger–Dyson distribution to a Poisson distribution was clearly observed in both cases (Fig. 1). We used the chi-squared test to determine the transition point. For the DIP yeast core protein interaction network, chi-squared testing showed that NNSD follows a Poisson distribution after removal of 1230 links. The remaining protein–protein interaction network contains 107 modules with sizes ranging from 2 to 778 proteins. For the yeast metabolic network, NNSD follows a Poisson distribution after removal of 770 links and the remaining network has 17 modules with sizes ranging from 4 to 602 chemical substrates and proteins.

These biological networks can be easily transformed to graphs by representing each element in the network as a vertex and each link as an edge in the graph. Figs. 2(a) and (c) show graph views of the original DIP yeast core protein interaction network and the yeast metabolic network, respectively. Figs. 2(b) and (d) illustrate the corresponding networks after removal of links at the transition point. One can see from Fig. 2 that the networks described by the Poisson distribution are very different from the original networks described by the GOE statistics. Isolated modules can be easily identified in Figs. 2(b) and (d).

To summarize, we have provided evidence that global biological networks, as represented by the yeast protein–protein

interaction and metabolic networks studied here, belong to the GOE. However, by successive removal of interactions between constituents of the network, a global biological network transitions into a system of isolated modules described the Poisson statistic. The transition from a GOE statistic to a Poisson statistic may open a new avenue for objective identification of functional modules inside global networks.

Acknowledgements

This research was supported by the United States Department of Energy under the Genomics: GTL, Microbial Genome Program and Natural and Accelerated Bioremediation Research Programs of the Office of Biological and Environmental Research, Office of Science, and by the National Institute of Health under N01-AI40076 and N01-AI40041. Jianxin Zhong was supported by the National Natural Science Foundation of China under Grant No. 30570432 and partially by the Materials Sciences and Engineering Division Program of the DOE Office of Science. Oak Ridge National Laboratory is managed by University of Tennessee-Battelle LLC for the Department of Energy under contract DE-AC05-00OR22725. Feng Luo is supported by NSF EPSCoT grant EPS-0447660. We thank Dr. Albert-Laszlo Barabasi, Dr. Hawoong Jeong and Dr. Natalia Maltsev for their help on construction of the metabolic networks.

References

[1] A.-L. Barabasi, Z.N. Oltvai, Nature Rev. 5 (2004) 101.
[2] C.M. Song, S. Havlin, H. A Makse, Nature 433 (2005) 392.
[3] J. Hasty, D. McMillen, J.J. Collins, Nature 420 (2002) 224.
[4] E.P. Wigner, SIAM Rev. 9 (1967) 1.
[5] E.P. Wigner, Proc. Cambridge Philos. Soc. 299 (1951) 189.
[6] E. Hofstetter, M. Schreiber, Phys. Rev. B 48 (1993) 16979.
[7] J.X. Zhong, U. Grimm, R.A. Romer, M. Schreiber, Phys. Rev. Lett. 80 (1998) 3996.
[8] J.X. Zhong, T. Geisel, Phys. Rev. E 59 (1999) 4071.
[9] O. Bohigas, M.J. Giannoni, C. Schmit, Phys. Rev. Lett. 52 (1984) 1.
[10] P. Seba, Phys. Rev. Lett. 91 (2003) 198104.
[11] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A. Nunes Amaral, H.E. Stanley, Phys. Rev. Lett. 83 (1999) 1471.
[12] I. Xenarios, L. Salwínski, X.Q. Duan, P. Higney, S.M. Kim, D. Eisenberg, Nucleic Acids Res. 30 (1) (2002) 303.
[13] C.M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg, Mol. Cell. Proteomics 1 (2002) 349.
[14] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabasi, Nature 407 (2000) 651.
[15] R. Overbeek, et al., Nucleic Acids Res. 28 (2000) 123.
[16] L.H. Hartwell, J.J. Hopfiled, S. Leibler, A.W. Murray, Nature 402 (1999) C47.
[17] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabasi, Science 297 (2002) 1551.
[18] A.W. Rives, T. Galitski, Proc. Natl. Acad. Sci. 100 (2003) 1128.
[19] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. 99 (2002) 7821.
[20] A.J. Enright, C.A. Ouzounis, Bioinformatics 17 (2001) 853.