Application of random matrix theory to microarray data for discovering functional gene modules

Feng Luo,[1] Jianxin Zhong,[2,3,*] Yunfeng Yang,[4] and Jizhong Zhou[4,5,†]
[1]*Department of Computer Science, Clemson University, 100 McAdams Hall, Clemson, South Carolina 29634, USA*
[2]*Department of Physics, Xiangtan University, Hunan 411105, China*
[3]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[4]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[5]*Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma 73019, USA*
(Received 9 June 2005; revised manuscript received 3 February 2006; published 29 March 2006)

We show that spectral fluctuation of coexpression correlation matrices of yeast gene microarray profiles follows the description of the Gaussian orthogonal ensemble (GOE) of the random matrix theory (RMT) and removal of small values of the correlation coefficients results in a transition from the GOE statistics to the Poisson statistics of the RMT. This transition is directly related to the structural change of the gene expression network from a global network to a network of isolated modules.

DOI: 10.1103/PhysRevE.73.031924 PACS number(s): 87.17.Aa, 87.80.Vt, 89.75.Fb

Understanding gene expression networks at system level is a key issue in the post genome era [1–4]. The emerging microarray technology [5,6] enables massive parallel measurement of expressions of thousands of genes simultaneously. It has opened up great opportunities to unveil gene expression networks at large scale. Currently, the inference of gene expression networks from microarray profiles is harmed by the dimensionality problem, namely, the number of genes is much larger than available data points. It is essential to develop powerful computational methods to extract as much biological information as possible from microarray data.

The random matrix theory (RMT), initially proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei [7], 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 [8], metal insulator transitions in disordered systems [9], spectra of quasiperiodic systems [10,11], chaotic systems [12], brain responses [13], and the stock market [14]. The RMT focuses on the study of statistical properties of eigenvalue spacing between consecutive eigenvalues. From the RMT, distribution of eigenvalue spacing of real and symmetrical random matrices follows two universal laws depending on the correlativity of eigenvalues. Strong correlation of eigenvalues leads to statistics described by the Gaussian orthogonal ensemble (GOE). On the other hand, eigenvalue spacing distribution follows Poisson statistics if there is no correlation between eigenvalues. A typical example of the GOE distribution is a complete random matrix with a random distribution of all matrix elements. The nonzero off-diagonal elements in this matrix, which represent mutual interactions between diagonal elements, induce strong correlations of eigenvalues and thus the GOE statistics. Differently, eigenvalue spacing distribution of a random matrix with nonzero values only for its diagonal (or block-diagonal) parts follows the Poisson statistics, because eigenvalues of this system are not correlated due to the absence of interaction between diagonal (or block-diagonal) parts. From microarry data, gene coexpression correlation matrices (CCMs) can be constructed. Because of the modularity of gene coexpression networks, after successive removal of lower values of correlation coefficients a CCM begins to have nonzero elements only for its block-diagonal parts corresponding to gene modules. We expect from the RMT that eigenvalue spacing distribution in the CCMs undergoes a transition from the GOE statistics to Poisson statistics.

In this paper, we report our results of application of the RMT to analysis of the CCMs of gene microarray profiles. We have found that eigenvalue spacing distribution of the CCMs of yeast gene microarray profiles is described by the GOE statistics. Furthermore, removal of small values of the correlation coefficients in the CCM results in a transition from the GOE statistics to the Poisson statistics as well as disassociation of the gene coexpression network from a global network to a network of isolated modules. This transition may provide a new objective approach for identification of functional modules from microarray profiles [15].

We used the standard spectral unfolding technique in the RMT to study the eigenvalue spacing statistics of the CCMs. 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. As a result, eigenvalue spacing distribution is a function of E_i and thus system dependent. In order to observe universal (system independent) eigenvalue fluctuations of different types of matrices, the RMT 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 $e_i=N_{\mathrm{av}}(E_i)$ 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, one calculates the nearest neighbor spacing distribution (NNSD) of eigenvalues, $P(s)$, which is defined as the probability density of unfolded eigenvalue spacing $s=e_{i+1}-e_i$. From the RMT, $P(s)$ of the GOE statis-

*Corresponding authors. Electronic address: zhongjn@ornl.gov
†Electronic address: zhouj@ornl.gov

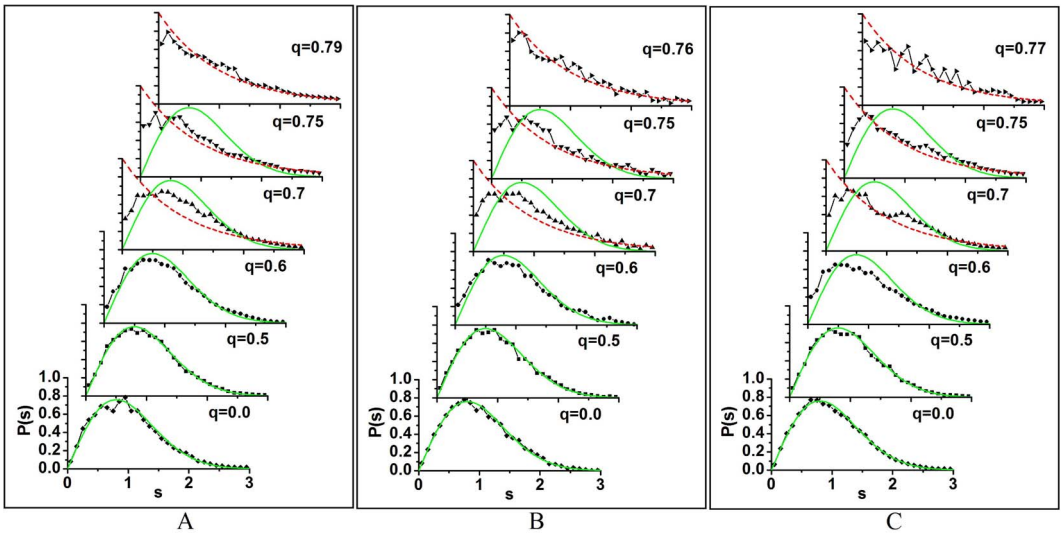

FIG. 1. (Color online) Nearest neighbor spacing distributions (symbols) of gene coexpression correlation matrices constructed from microarray profiles of yeast mutants by different cut methods with cutoff value q. The solid green is the Wigner-Dyson distribution and the dashed line is the Poisson distribution. (a) Method I: $c=0$ if $|c|<q$. (b) Method II: $c=0$ if $c<q$. (c) Method III: $c=0$ if $c>-q$.

tics closely follows the Wigner-Dyson distribution

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

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

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

The difference between the Wigner-Dyson and Poisson distributions manifests in the regime of small s, where, $P_{\mathrm{GOE}}(s\to 0)=0$ and $P_{\mathrm{Poisson}}(s\to 0)=1$.

Matrix elements in the coexpression correlation matrix CCM for our RMT analysis are the standard Pearson correlation coefficients of gene expressions between different genes defined as

$$c(g_i,g_j) = \frac{1}{N}\sum_{k=1,N}\left(\frac{g_{ik}-M_{g_i}}{\sigma_{g_i}}\right)\left(\frac{g_{jk}-M_{g_j}}{\sigma_{g_j}}\right),$$

where M_{g_i}, M_{g_j} are the average gene expression of gene g_i and g_j, respectively, σ_{g_i}, σ_{g_j} are their corresponding standard deviations, and N is the total number of experiments.

As examples, we studied two microarray expression profiles of yeast. The first one is the microarray expression data of 287 yeast mutants [16]. We selected genes that have expressed in most mutant experiments for our study. As a result, there is a total of 6209 genes. The second profile is the gene expression data of yeast cells response to environment changes [17]. We selected 6090 genes that have expressed in most experiments. Figures 1 and 2 show the NNSDs for the yeast mutants and the yeast cells response, respectively (curves with $q=0$). One can see that in both cases, the NNSD is well described by the Wigner-Dyson distribution.

It has been widely believed that a cell system, like many other engineering synthetic systems, is modular. Its cellular functionality is performed by a collection of modules, which are groups of physically or functionally linked genes. Given a specific condition and stage, there are particular gene modules expressed in the cell. The expression patterns of genes in the same functional module often exhibit higher similarity in microarray experiments [2–4]. Therefore, the CCM of a microarray profile consists of a strong correlation part, C_s, which corresponds to the correlation between genes in the same module, and a weak correlation part, C_w, which corresponds to the correlation between genes in different modules or unexpressed isolated genes. To test the modularity of genes, we gradually remove lower values of correlation coefficients. Three cut methods are investigated to keep significant correlations: (1) $c=0$ if $|c|<q$ (method I); (2) $c=0$ if $c<q$ (method II); and (3) $c=0$ if $c>-q$ (method III); where $0<q<1$ denotes the level of cutoff. Evidently, both positive and negative significant correlations are retained in method I. In methods II and III, only positive or negative significant correlations are kept, respectively. The three different cut methods allow us to study a series of CCMs with different cutoff values.

We found that the NNSDs of the CCMs constructed from microarray expression profiles of yeast using the tree cut methods all exhibit sharp transitions from a Wiger-Dyson distribution to a Poisson distribution, as the cutoff level q increases. This transition behavior is clearly shown in Figs. 1 and 2. Transition points were obtained by the chi-square test, which is a standard technique for goodness-of-fit test that determines whether a set of sample data have been drawn from a hypothetical solution. In our calculation, q corresponds to the transition point when the chi-square test to Poisson distribution is less than a critical value of confidence level 99.999%. We found that the NNSDs of microarray profiles of yeast mutants are well described by the Poisson distribution as $q\geqslant 0.79$, 0.76, 0.77 obtained by methods I, II, and III, respectively. For microarray profiles of yeast cells response to environmental changes, we found Poisson distribution as $q\geqslant 0.89$ by method I, $q\geqslant 0.88$ by method II, and $q\geqslant 0.86$ by method III.

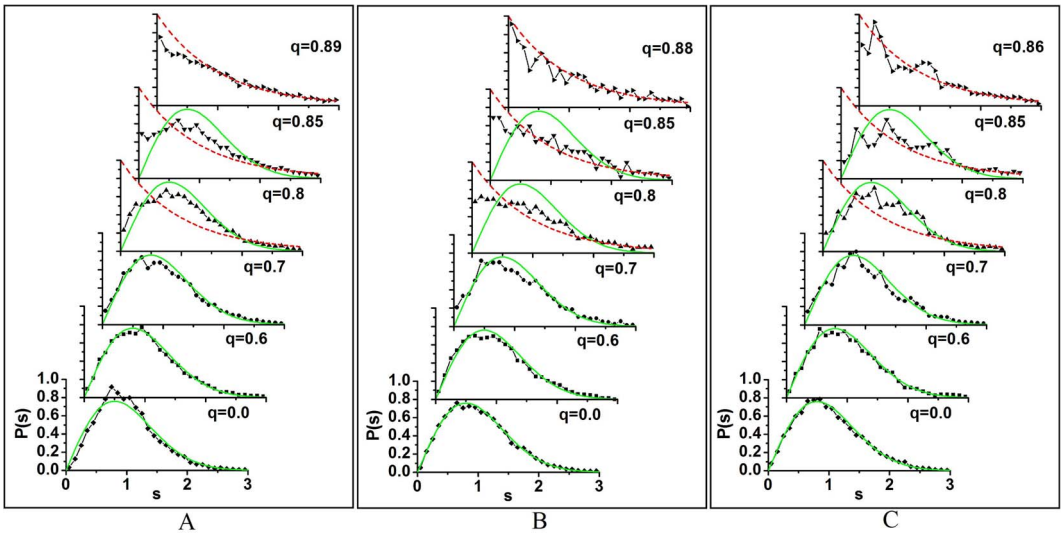

FIG. 2. (Color online) Nearest neighbor spacing distributions (symbols) of gene coexpression correlation matrices constructed from microarray profiles of yeast cells response to environment changes by different cut methods with cutoff value q. The solid line is the Wigner-Dyson distribution and the dashed line is the Poisson distribution. (a) Method I: $c=0$ if $|c|<q$. (b) Method II: $c=0$ if $c<q$. (c) Method III: $c=0$ if $c>-q$.

To view the structural change of gene coexpression networks after the removal of weaker correlations, we constructed a series of gene coexpression networks from the CCMs of both microarray profiles. The networks were visualized using Biolayout [18], where each gene in a network is a node and there is a link between two genes if the Pearson correlation between their expressions is not 0 after removal of small values of coexpression correlation coefficients. Figures 3 and 4 show yeast gene coexpression networks at different cutoff values. One can see from Figs 3 and 4 that the networks described by the Poisson distribution are very different from the networks described by the GOE statistics. Furthermore, the transition of the NNSD from the GOE distribution to the Poisson distribution is accompanied by the disassociation of coexpression networks. Isolated modules can be easily identified in Figs. 3 and 4 at the transition points. Our detailed analysis showed that, for the microarray profile of yeast mutants, the coexpression network at $q=0.79$ obtained by method I has 39 modules with number of genes ranging from 3 to 597; the coexpression network at $q=0.76$ by method II has 41 modules with number of genes ranging from 3 to 863; and the coexpression network at $q=0.77$ by method III has 11 modules with number of genes ranging from 3 to 526. For microarray profiles of the yeast cells response, method I at $q=0.89$ generates 13 modules with sizes ranging from 3 to 636; method II at $q=0.88$ gives 17 modules with sizes ranging from 3 to 510 genes; and method III at $q=0.86$ generates 2 modules with 4 and 395 genes. We have also found that the modules obtained by different cut methods have different structures, which is understandable because different cut methods emphasize different types of correlations between genes. We found that structures of gene modules found in Figs. 3 and 4 are very different from the remaining structures constructed from a random matrix after the removal of small values of matrix elements using the same cut methods. We believe that the gene modules obtained by using the RMT are functional modules and we are now experimentally evaluating their biological functionalities [15].

To clearly show that the gene coexpression networks revealed by the RMT from the microarray coexpression data indeed are of a biological origin different from structures

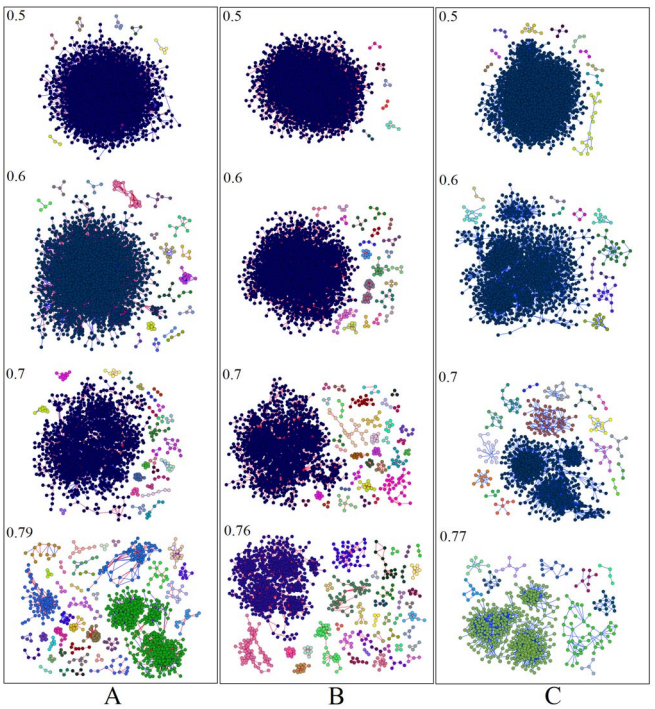

FIG. 3. (Color online) Coexpression networks constructed from microarray profiles of yeast mutants by different cut methods with cutoff value q. Lines represent correlations between genes. (a) Method I: $c=0$ if $|c|<q$. (b) Method II: $c=0$ if $c<q$. (c) Method III: $c=0$ if $c>-q$.

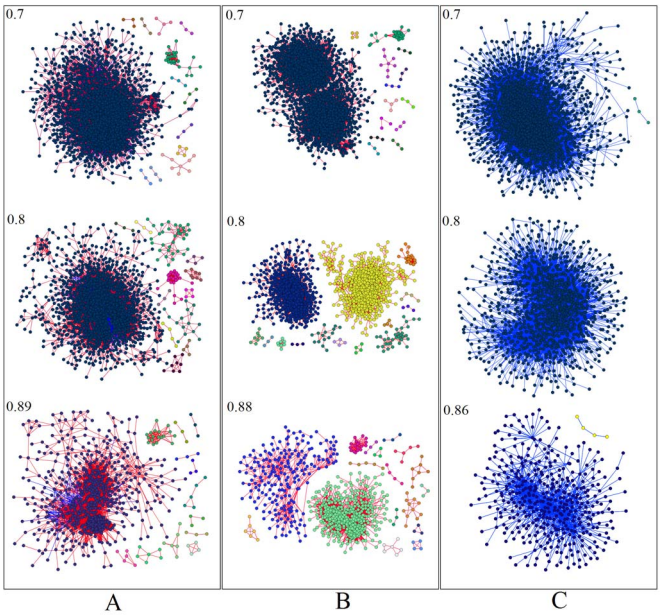

FIG. 4. (Color online) Coexpression networks constructed from microarray profiles of yeast cells response to environment change by different cut methods with cutoff value q. Lines represent correlations between genes. (a) Method I: $c=0$ if $|c|<q$. (b) Method II: $c=0$ if $c<q$. (c) Method III: $c=0$ if $c>-q$.

constructed from a random matrix, we studied network properties of two types of random correlation matrices using the RMT. The first one is a completely random correlation matrix constructed by assigning its elements with random values distributed within an interval $[-1,1]$. The second one is the column shuffled microarray correlation matrix, where the expression values of a each gene of a yeast cell response to environment changes are randomly shuffled many times (100 times in this paper) before constructing its corresponding

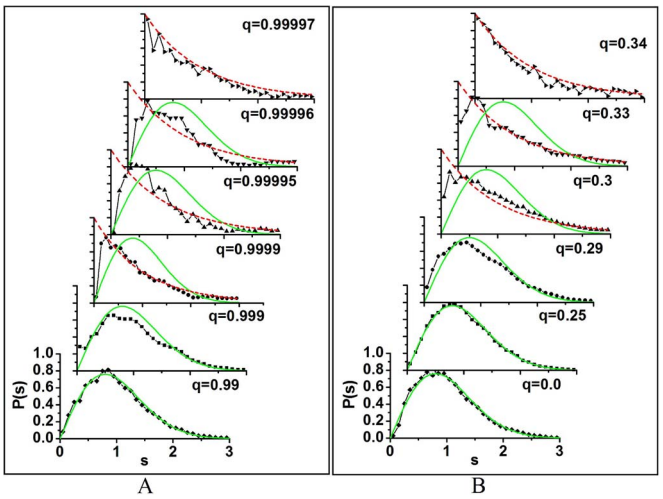

FIG. 5. (Color online) Nearest neighbor spacing distributions (symbols) of random correlation matrices at different cutoff level q constructed from (a) a 2000×2000 completely random matrix and (b) shuffled microarray profiles of yeast cells response to environment changes. The solid line is the Wigner-Dyson distribution and the dashed line is the Poisson distribution.

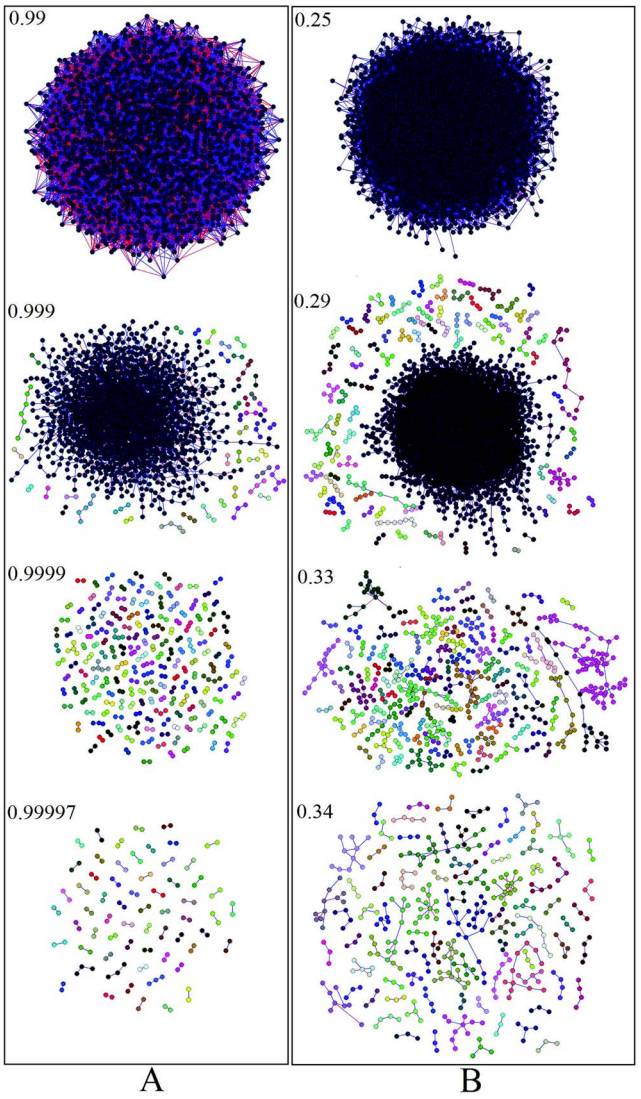

FIG. 6. (Color online) Networks at different cutoff level q constructed from (a) a completely random matrix and (b) shuffled microarray profiles of yeast cells response to environment changes. Lines represent correlations between genes.

correlation matrix. Small values of matrix elements in these two matrices are then removed by using cut method I to construct CCMs and networks at a different cutoff levels. We found that NNSDs of these randomized CCMs also exhibit transitions from a Wiger-Dyson distribution to a Poisson distribution, as shown in Fig. 5. The critical cutoff value for the completely random correlation matrix is $q=0.999\,97$. Our detailed analysis showed that the critical cutoff value increases as the matrix dimension increases. For the shuffled microarray profiles of the yeast cells response to environmental changes, we found a Poisson distribution as $q\geqslant0.34$. Figure 6 illustrates the gene networks constructed from these two random correlation matrices at a different cutoff value. Notably, the networks obtained from the random correlation matrices are very different from the coexpression networks constructed from the original microarray profiles, even though they are all described by the Poisson distribution. Comparing with large highly connected clusters

(hundreds of genes) in the coexpression network, the network of the completely random matrix at $q=0.999\,97$ only has isolated small clusters of a few nodes (2–3); the network of the shuffled microarray profiles at $q=0.34$ only has small chainlike or treelike small clusters with size from 3 to 15 nodes. We searched the gene ontology (GO) concurrence of these clusters using the GO term finder [19] of the Saccharomyces Genome Database. No significant GO term exists for all top 20 large clusters, indicating that the clusters obtained from the shuffled microarray profiles are not biologically meaningful.

In summary, we have provided evidence that genome-wide gene coexpression, as represented by the CCMs of microarray profiles studied here, is described by the GOE statistics of the RMT. However, the successive removal of small correlations in the CCM leads to a transition from the GOE statistics to the Poisson statistics. Our results indicate that although biological systems are very different from complex physical systems [20], they follow the same universal Wigner distribution and the Poisson distribution. The transition we found may open a new avenue for the identification of functional modules from microarray profiles. Different from existing clustering methods, cutoffs or thresholds used for separating genes in our approach is determined self-consistently by the transition given by the random matrix theory.

This work was supported by The United States Department of Energy under the Genomics: GTL through Shewanella Federation, Microbial Genome Program and Natural and Accelerated Bioremediation Research Programs of the Office of Biological and Environmental Research, Office of Science. F.L. was also supported by the NSF EPSCot Grant No. EPS-0447660. J.Z. 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 No. DE-AC05-00OR22725.

[1] A.-L. Barabasi and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[2] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, Science **302**, 249 (2002).
[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Bostein, Proc. Natl. Acad. Sci. U.S.A. **95**, 14863 (1998).
[4] X. Zhou, M. C. Kao, and W. H. Wong, Proc. Natl. Acad. Sci. U.S.A. **99**, 2783 (2002).
[5] D. J. Lockhart *et al.*, Nat. Biotechnol. **14**, 1675 (1996).
[6] M. Schena, D. Shalon, R. Davis, and P. Brown, Science **270**, 467 (1995).
[7] E. P. Wigner, SIAM Rev. **9**, 1 (1967).
[8] E. P. Wigner, Proc. Cambridge Philos. Soc. **299**, 189 (1951).
[9] E. Hofstetter and M. Schreiber, Phys. Rev. B **48**, 16979 (1993).
[10] J. X. Zhong, U. Grimm, R. A. Romer, and M. Schreiber, Phys. Rev. Lett. **80**, 3996 (1998).
[11] J. X. Zhong and T. Geisel, Phys. Rev. E **59**, 4071 (1999).
[12] O. Bohigas, M. J. Giannoni, and C. Schmit, Phys. Rev. Lett. **52**, 1 (1984).
[13] P. Seba, Phys. Rev. Lett. **91**, 198104 (2003).
[14] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).
[15] F. Luo, Y. F. Yang, J. X. Zhong, H. C. Gao, L. Khan, D. K. Thompson, and J. Z. Zhou (unpublished).
[16] T. R. Hughes *et al.*, Cell **102**, 109 (2000).
[17] A. P. Gasch *et al.*, Mol. Biol. Cell **11**, 4241 (2000).
[18] A. J. Enright and C. A. Ouzounis, Bioinformatics **17**, 853 (2001).
[19] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.
[20] L. H. Hartwell, J. J. Hopfiled, S. Leibler, and A. W. Murray, Nature (London) **402**, C47 (1999).