Random matrix analysis of localization properties of gene coexpression network

Sarika Jalan,[1] Norbert Solymosi,[2] Gábor Vattay,[2] and Baowen Li[1,3]
[1]*Department of Physics and Centre for Computational Science and Engineering, National University of Singapore, 117456, Republic of Singapore*
[2]*Department of the Physics of Complex Systems, Eötvös University, Pázmány Péter Sétány 1/A, H-1117 Budapest, Hungary*
[3]*NUS Graduate School for Integrative Sciences and Engineering, NUS, Singapore 117546, Republic of Singapore*
(Received 3 November 2009; revised manuscript received 4 March 2010; published 28 April 2010)

We analyze gene coexpression network under the random matrix theory framework. The nearest-neighbor spacing distribution of the adjacency matrix of this network follows Gaussian orthogonal statistics of random matrix theory (RMT). Spectral rigidity test follows random matrix prediction for a certain range and deviates afterwards. Eigenvector analysis of the network using inverse participation ratio suggests that the statistics of bulk of the eigenvalues of network is consistent with those of the real symmetric random matrix, whereas few eigenvalues are localized. Based on these IPR calculations, we can divide eigenvalues in three sets: (a) The nondegenerate part that follows RMT. (b) The nondegenerate part, at both ends and at intermediate eigenvalues, which deviates from RMT and expected to contain information about *important nodes* in the network. (c) The degenerate part with *zero* eigenvalue, which fluctuates around RMT-predicted value. We identify nodes corresponding to the dominant modes of the corresponding eigenvectors and analyze their structural properties.

DOI: 10.1103/PhysRevE.81.046118 PACS number(s): 89.75.Hc, 87.18.−h, 02.10.Yn, 21.60.Ev

I. INTRODUCTION

A. Complex networks

Gene expression information captured in microarrays data for a variety of environmental and genetic perturbations, in conjunction with other sources such as protein-protein or protein-DNA interaction and operon organization data, promises to yield unprecedented insights into the organization and functioning of biological systems [1,2]. It has been increasingly realized that dissecting the genetic and chemical circuitry prevents us from further understanding the biological processes as a whole. In order to understand the complexities involved, all reactions and processes should be analyzed together. To this end, network theory will be used. It has been getting fast recognition to study systems which could be defined in terms of units and interactions among them. These studies revealed that the available data from gene coexpression network share some unexpected features with other complex networks as diverse as the internet routers. In order to understand the behavior of complex systems such as gene coexpression network, several simple models based on the simple principles and capturing essential features of the unerlying system, have been presented [3–5].

In this paper, by using network theory and random matrix theory (RMT), we analyze gene coexpression network. First, we generate network from the gene coexpression data collected form six brain regions that are metabolically relevant to Alzheimer's disease [6] by using appropriate threshold and then study the spectra of this network under the RMT framework. Information about the genes that are preferentially expressed during the course of Alzheimer's disease could improve our understanding of the molecular mechanisms involved in the pathogenesis of this common cause of cognitive impairment in senior persons, provide new opportunities in the diagnosis, early detection, and tracking of this disorder, and provide novel targets for the discovery of interventions to treat and prevent this disorder. Information about the genes that are preferentially expressed in relationship to normal neurological aging could provide new information about the molecular mechanisms that are involved in normal age-related cognitive decline and a host of age-related neurological disorders, and they could provide novel targets for the discovery of interventions to mitigate some of these deleterious effects.

Coexpression networks have also been known as relevance networks. The terminology has been introduced by Butte and Kohane [7]. Since then, properties of the relevance networks have been extensively studied [8].

The paper is organized as follows. After introducing the relevance of network theory and gene coexpression network, we discuss the recent outcome of RMT analysis of complex networks in Sec. I B. The main goals of our eigenvector analysis are written in the Sec. I C. Section II describes the important achievements of RMT and explains its various properties we use in our analysis. Section III sheds light on the data and network construction. Section IV presents various numerical results. Section V concludes the paper with a discussion on the relevance of current analysis as well as suggests future directions.

B. RMT of network spectra

Our previous work [9] showed that various vastly studied model networks follow random matrix predictions of Gaussian orthogonal statistics (GOE) at the level repulsion domain. We demonstrated that nearest-neighbor spacing distribution (NNSD) of protein-protein interaction network of budding yeast follows RMT prediction as well [9]. This is a promising result which suggests that these networks can be modeled as a random matrix chosen from an appropriate ensemble. The universal GOE statistics of eigenvalues fluctuations could be understood as some kind of randomness spreading over the protein-protein interaction network and model networks capturing real-world properties. Recently, covariance

matrix of amino acid displacement has been analyzed under RMT framework [10]. The analysis shows that the bulk of eigenvalues follows universal GOE statistics of RMT. In the present paper, we analyze gene coexpression network [6] under RMT framework. First, we calculate nearest-neighbor spacing distribution of network spectra and then perform eigenvector analysis to detect nodes having specific contribution to network.

C. Important nodes and connections

It is now well known that various real-world systems are scale-free networks [3]. The scale-free nature of networks suggests that there exist few nodes with very high degrees. Motivated by this finding, they suggested that since these nodes are responsible to hold the whole networks, henceforth they are the most important ones. Some other analyses (by Newman and others) of real-world networks show that complex networks have community or module structure [11,12]. Modules are the division of network nodes within which the network connections are dense, but between which they are sparser. The modularity concept assumes that system functionality can be partitioned into a collection of modules and each module performs an identifiable task, separable from the functions of other modules [13]. Analysis of module structure involves *betweenness* measure. Betweenness of an edge is defined as the number of shortest path between pairs of nodes going through the edge. Betweenness studies of real-world networks suggest that the nodes connecting the different communities are the most important ones, which has been verified in the metabolic networks by Guimerá and Amaral [13].

Above description emphasizes on the importance of nodes depending on their position in the network, as these nodes characterize network properties. On the other hand, Erdös-Rényi (ER) and Strogatz-Watts (SW) models emphasize on the importance of random connections in the networks. In the ER model, any two nodes are connected with probability p. One of the most interesting properties of ER model is the sudden emergence of various global properties, such as emergence of a giant cluster. As p increases, while number of nodes in the graph remains constant, the giant cluster emerges through a phase transition [14]. Further, the SW model shows the small world transition with the fine tuning of number of random connections [15]. Our previous RMT analysis of the spectra of SW model networks [9] shows that at the SW transition there is a transition to the *spreading of randomness* in the network characterized by the correlations between nearest eigenvalues. In the current paper, we analyze spectra of the gene coexpression network under RMT framework. Particularly, we study eigenvectors of the adjacency matrix of this network. The spectra have two parts: one part which follows RMT predictions of universal GOE statistics and other part which does not follow RMT prediction. The eigenvectors deviating from the RMT prediction provide information about the *influential or important nodes* in the network.

II. RANDOM MATRIX STATISTICS

RMT deals with the statistical properties of matrices with independent random entries. To be self-consistent, we give a brief introduction of the RMT here and explain various RMT properties of eigenvector components which we will use in our analysis. RMT was initially proposed to explain the statistical properties of nuclear spectra [16]. Later, this theory was successfully applied in the study of the spectra of different complex systems such as disordered systems, quantum chaotic systems, and large complex atoms [17]. Recent studies illustrate the usefulness of RMT in understanding the statistical properties of the empirical cross-correlation matrices appearing in the study of multivariate time series of the following: the price fluctuations in the stock market [18], electro encephalogram (EEG) data of brain [19], variation of various atmospheric parameters [20], etc. Recent analysis of complex networks under RMT framework [9,10,21,22] shows that various network models and real-world network also follow universal GOE statistics. Furthermore, localizations of eigenvectors have also been used to analyze various structural and dynamical properties of real and model networks [23,24].

In the following, we present spacing distribution and Δ_3 statistics of random matrices. We denote the eigenvalues of a network by λ_i, $i=1,\ldots,N$, where N is size of the network and $\lambda_1<\lambda_2<\lambda_3<\cdots<\lambda_N$. In order to get universal properties of the fluctuations of eigenvalues, people usually unfold the eigenvalues by a transformation $\bar{\lambda}_i=\bar{N}(\lambda_i)$, where $\bar{N}$ is averaged integrated eigenvalue density [16]. Since we do not have any analytical form for $\bar{N}$, we numerically unfold the spectrum by polynomial curve fitting (for elaborate discussion on unfolding, see Ref. [16]). After unfolding, average spacing is *unity*, independent of the system. Using the unfolded spectra, we calculate spacings as $s_i=\bar{\lambda}_{i+1}-\bar{\lambda}_i$. NNSD is defined as the probability distribution $[P(s)]$ of these s_i's. In the case of GOE statistics,

$$P(s)=\frac{\pi}{2}s\exp\left(-\frac{\pi s^2}{4}\right). \tag{1}$$

The Δ_3 statistic measures the least-squares deviation of the spectral staircase function representing the averaged integrated eigenvalue density $\bar{N}(\lambda)$ from the best straight line fitting for a finite interval L of the spectrum, i.e.,

$$\Delta_3(L;x)=\frac{1}{L}\min_{a,b}\int_x^{x+L}[N(\bar{\lambda})-a\bar{\lambda}-b]^2d\bar{\lambda}, \tag{2}$$

where a and b are obtained from a least-squares fit. Average over several choices of x gives the spectral rigidity $\Delta_3(L)$. For the GOE case, $\Delta_3(L)$ depends *logarithmically* on L, i.e.,

$$\Delta_3(L)\sim\frac{1}{\pi^2}\ln L. \tag{3}$$

The following section explains the properties of eigenvectors of random matrices.

Eigenvector analysis

The distribution of eigenvector components is studied to obtain system-dependent information. Let u_l^k be the lth component of kth eigenvector u^k. The eigenvector components of

a GOE random matrix are Gaussian-distributed random variables. For this the distribution of $r=|u_l^k|^2$, in the limit of large matrix dimension, is given by Porter-Thomas distribution [25], i.e.,

$$P(r) = \frac{N}{\sqrt{2\pi r}} \exp\left(\frac{-Nr}{2}\right). \tag{4}$$

Shannon entropy for the state whose components are described by the above distribution would be given by in large N limit as [25]

$$H_s \sim -N\int_0^\infty r\ln(r)P(r)dr \sim \ln\left(\frac{N}{2}\right). \tag{5}$$

Additionally, inverse participation ratio (IPR) is also considered to study the RMT features of the eigenvectors. The IPR of eigenvector is defined as

$$I^k = \sum_{l=1}^{N} [u_l^k]^4, \tag{6}$$

where u_l^k, $l=1,\ldots,N$ are the components of eigenvector u^k. The meaning of I^k is illustrated by two limiting cases: (i) a vector with identical components $u_l^k \equiv 1/\sqrt{N}$ has $I^k=1/N$, whereas (ii) a vector with one component $u_1^k=1$ and the remainder zero has $I^k=1$. Thus, the IPR quantifies the reciprocal of the number of eigenvector components that contribute significantly. For a vector with components following distribution (4), $I^k \sim 3/N$.

III. DATA AND NETWORK CONSTRUCTION

The data set (GSE5281) was obtained from gene expression omnibus [6]. Liang *et al.* [2] studied gene expression profiles from laser capture microdissected neurons in six functionally and anatomically distinct regions from clinically and histopathologically normal-aged human brains. From these data sets, only 74 normal samples were used to construct the coexpression networks. In the original study, the Affymetrix Human Genome U133 Plus 2.0 Array was used. This microarray contains 54675 oligonucteotids (probe sets) representing the expressed human genes for each samples. On the microarray, one gene is represented by one or more probe sets. Each probe set is built up from 25 mer length oligonucleotides, so-called probes [26]. In the present study, probe sets are the units of observation. For the identification of probe sets, the Affymetrix IDs were used. The Pearson's product-moment correlation was calculated for each probe set–pair expression level and those which have value greater than 0.88 are used to construct the gene coexpression network. This network consists of 5000 nodes and 1 201 480 undirected edges. Nodes represent probe set denoting genes and edges denote their coexpression levels.

From this weighted network, we construct a sparse binary network as following. We choose the value of threshold being $r=0.89$. If the coexpression strength is greater than r than the corresponding element in the matrix gets value 1, otherwise it takes value 0. Threshold value of $r=0.89$ leads to a network with much less number of edges and results into many disconnected components. Note that choosing the threshold value is a crucial step and different schemes have been proposed to select it [27,28]. We sort out the nodes and edges forming largest connecting cluster, which is of the size $N=3179$ and 46 033 connections. The average degree of this network is $\langle k\rangle \sim 30$. RMT analysis is done for this biggest component. Figure 1 shows the adjacency matrix of this component and Fig. 2 is the degree distribution.

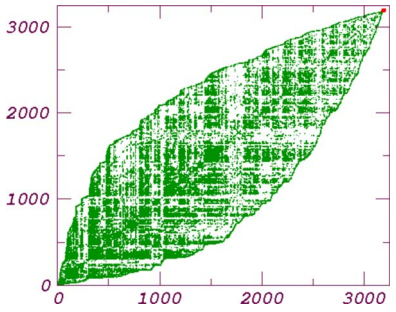

FIG. 1. (Color online) Adjacency matrix of the largest connected component of the gene coexpression network with the threshold value of ~0.89. Nodes forming largest connecting cluster are renumbered in the sequential order for a clear visualization.

IV. RESULTS

In the following, we present the various RMT results for gene coexpression network constructed above. We calculate the eigenvalues and eigenvectors of the adjacency matrix corresponding to the largest connected network. Since this is an undirected network, eigenvalues of adjacency matrix are real and we denote them as λ_i, $i=1\ldots N$. Eigenvectors are denoted as u^k, $k=1\ldots N$.

A. Spacing distribution and Δ_3 analysis

From this spectrum, we calculate NNSD $P(s)$ as described in Sec. II and $\Delta_3(L)$ statistic using Eq. (2). Figure 3(a) shows that NNSD agrees well with the NNSD of GOE matrices (1) with the value of Brody parameter [9,29] $\beta \sim 1$.

Figure 3(b) plots the $\Delta_3(L)$ statistics. It can be seen that $\Delta_3(L)$ statistics agrees well with the GOE statistics up to the value of $L \sim 25$ (which is much less than the same for the corresponding random and scale free model networks [9]). According to the RMT, this implies that besides randomness,

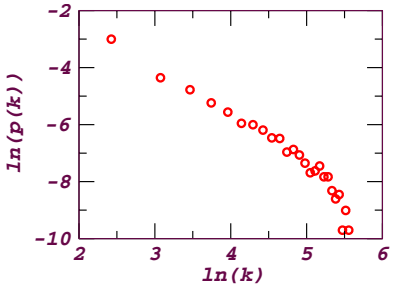

FIG. 2. (Color online) Degree distribution of the largest connected part of the gene coexpression network for threshold 0.89.

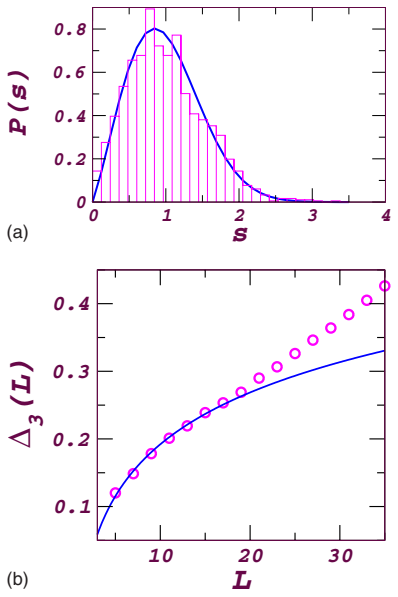

FIG. 3. (Color online) (a) Spacing distribution and (b) $\Delta_3(L)$ statistics for the eigenvalue spectra of the gene coexpression network. The histogram in (a) corresponds to the numerical values and solid line is GOE prediction (1) of RMT. The circles in (b) are numerical results (2) and the solid curve is GOE prediction (3) of Δ_3.

the network has some specific features. Note that the points which deviate from GOE statistics ($L>20$), as shown in the Fig. 3(b), can also be analyzed using deformed GOE statistics as shown in [21].

B. Eigenvector analysis

Having calculated spacing distribution and Δ_3 statistics, now we use eigenvector analysis to study the factors responsible for the deviation from RMT. We calculate IPR and entropy for all the eigenvectors. The eigenvectors, whose IPR and entropy deviate from the random matrix predictions, carry the relevant information. The nodes corresponding to the top contributing components of these vectors may be *important nodes* in terms of functionality of the whole network. In the following, we present the eigenvector analysis results for the gene coexpression network.

Figure 4(a) shows eigenvalues in the increasing order. Apart from distinguishably seen high eigenvalues toward the end of the spectra, there is a flat part around the zero eigenvalue. Real world networks, in general, are very sparse and are reported to have large number of *zero* eigenvalues [30,31]. Though for the network we consider here, out of 3179 eigenvalues, only approximately 73 ($\sim$2.5% of all eigenvalues) are degenerate with the value *zero*. The degeneracy at zero eigenvalue is lesser than many other real-world networks [9]. There are nearly 3106 nondegenerate eigenvalues, which could be taken as the effective dimensionality of the network.

We also calculate Shannon entropy for all the eigenvectors using Eq. (5) and compare them to those of the random vectors. Figure 4(b) shows the entropy as a function of

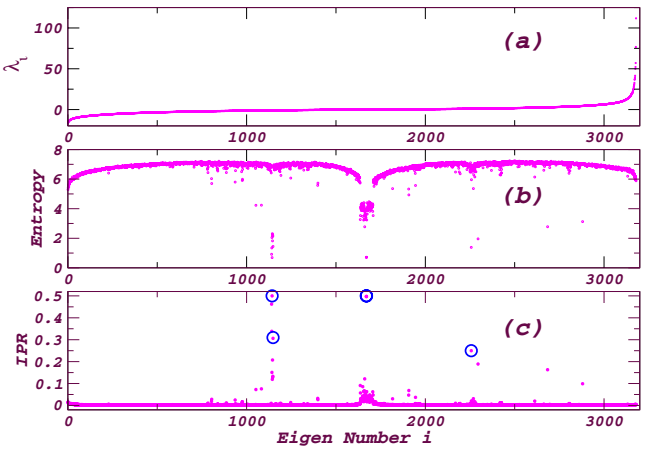

FIG. 4. (Color online) (a) Eigenvalues, (b) entropy, and (c) IPR as a function of eigennumber for the threshold value of 0.89. Open blue circles in (c) correspond to the localized eigenvectors whose top contributing nodes are listed in the Table I

eigennumbers. According to RMT, Shannon entropy of a random vector of dimension $N=3106$ is $\ln(3106/2)\simeq 7.35$. Furthermore, RMT-predicted value for Shannon entropy of a random vector of dimension $N=73$ (corresponding to degenerate part) is $\ln(73/2)\simeq 3.6$. Based on these calculations, we can divide eigenvalues into three sets: (a) The nondegenerate part that follows RMT. (b) The nondegenerate part, at both ends and at intermediate eigenvalues, which deviate from RMT and expected to contain information about *important nodes* in the network. (c) The degenerate part with *zero* eigenvalue, 1636–1708, which fluctuates around RMT-predicted value.

Furthermore, we calculate IPR of all the eigenvectors using Eq. (6) and plot in Fig. 4(c). It shows that IPR of several eigenvalues are localized. For example, vectors corresponding to the 1140–1148 eigenvalues have $I^k \geq 0.1$, showing that few components contribute more than the other components. Following, we enlist some localized eigenvectors corresponding to nondegenerate eigenvalues from set (b): u^{1143} (with $I^k \sim 0.5$), u^{1148} (with $I^k \sim 0.31$), and u^{2257} (with $I^k = 0.25$). Some of the localized eigenvectors corresponding to zero eigenvalues are [set (c)] u^{1636} (with $I^k=0.1$), u^{1670}, and u^{1671} (with $I^k \sim 0.5$). We next analyze the significant contributors of eigenvectors deviating from the RMT predictions. The eigenvector u^{1143} contains approximately $1/\mathrm{IPR}^{1143}=20$ significant participants. Table I presents top five significant contributors (nodes) corresponding to the localized eigenvector mentioned above. Note that original gene numbers are written as in the data sets [6]. As shown in the Fig. 2, degree distribution of the connected network analyzed above follows a power law with a fat tail, which means that few nodes are hubs, and carries the whole network. But random matrix analysis of eigenvectors reveals that all the most contributing nodes listed above have rather small degree. They are all almost toward bottom of the power-law distribution.

The degrees of all the top contributing nodes in the localized eigenvectors are either well below the average degree or around the average degree of the network. Gene, assigned

TABLE I. Top five largest contributing nodes in localized eigenvectors for network constructed with the threshold value of 0.89. The nodes are written in the original gene number as given in the data sets [6].

Set B			Set C	
u^{1143}	u^{1148}	u^{2257}	u^{1670}	u^{1671}
202060_at	227636_at	202916s_at	225921_at	21435x_at
217731s_at	205003_at	226832_at	212635_at	203034s_at
201121s_at	211940x_at	209860s_at	208645s_at	200673_at
221775x_at	224616_at	218175_at	221511x_at	221471_at
229630s_at	222203s_at	221810_at	231896s_at	225950_at

with probe set 202060_at (corresponding to the node 2299 in the renumbered network), which is the first top contributing node corresponding to eigenvector u^{1143}, has a degree 15, the second top contributing node has a degree 17, the third node has a degree 20. Fourth and fifth top contributing nodes have degree 9 each. The top five nodes corresponding to u^{1148} have degrees 21, 14, 7, 17, and 24. Those are corresponding to eigenvector u^{2257} have degrees 1, 1, 6, 3, and 1, respectively. The localized eigenvectors corresponding to set (c) are u^{1670}, u^{1671}, and top five contributing nodes have degree, in sequential order from first to the fifth contributing nodes (see Table I), 2, 4, 8, 1, 3 and 10, 9, 23, 14, 2, respectively.

Now we change the threshold value to 0.91. This threshold value leads to 25 000 connections in the whole network. This network has largest connected cluster of size 2439 and number of connections 22 546. The average degree of this network is $\langle k \rangle \sim 20$. Again, we renumber the nodes such that nodes in the connected component take value from 1 to 2439 and calculate the eigenvalues and eigenvectors of the adjacency matrix corresponding to this largest connected network. From the spectrum NNSD and Δ_3 statistics are calculated and these two show similar GOE statistics as shown in Fig. 3 for $r=0.89$.

Figure 5 plots eigenvalues (a), entropy (b), and IPR (c) as a function of eigennumber. Entropy and IPR are calculated using Eq. (5) and (6), respectively. Out of 2439 eigenvalues, approximately 96 are degenerate with the value *zero*. It means that there are nearly 2343 nondegenerate eigenvalues, which could be taken as the effective dimensionality of the network. According to RMT, Shannon entropy of a random vector of dimension $N=2343$ is $\ln(2343/2) \simeq 7.0$. On the other hand, RMT-predicted value for Shannon entropy for

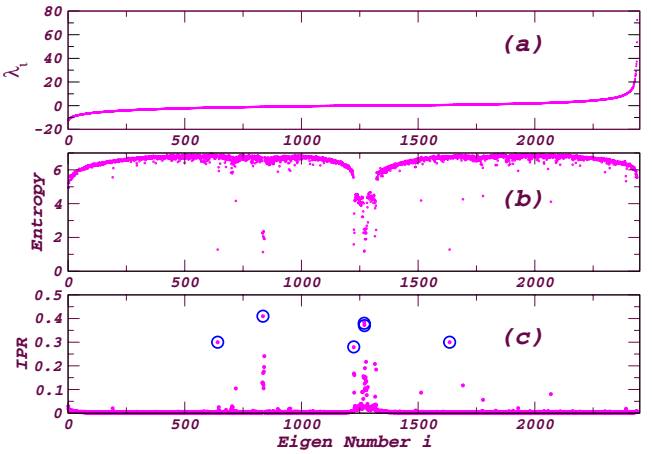

FIG. 5. (Color online) Same as Fig. 4 but for threshold value of 0.91. Open blue circles correspond to localized eigenvectors whose top five contributing nodes are presented in the Table II.

degenerate eigenvectors is $\ln(96/2) \simeq 3.9$. Based on these calculations, again we can divide eigenvalues in three sets (a)–(c). localized eigenvectors corresponding to nondegenerate parts are u^{835}(IPR=0.41), u^{1635} (IPR=0.3), u^{641}(IPR=0.3), u^{840}, and u^{841} (with $\lambda=1$, IPR=0.195 and 0.24). Localized eigenstates corresponding to zero eigenvalues [set (c)] are u^{1269} (IPR=0.38), u^{1270} (IPR=0.37), and u^{1224} (IPR=0.28). Significant contributors in localized eigenvectors are written in Table II.

The degree distribution of the largest component at this threshold follows a power law as well, revealing the scale-free nature of this component. Increasing threshold preserves scale-free property of the network. Some nodes are hubs which carry the whole network and enjoy the structural importance. Again we find that the top contributing nodes are not the ones with very high degree. For two different threshold values, Tables I and II show the largest contributing coexpressing genes in the corresponding localized eigenvectors. We find that choosing threshold is very important for the analysis of gene coexpression networks, as we can see that top five largest contributing nodes differ entirely (except one) as threshold value is changed. This suggests that, though the gross structure of whole network (Fig. 1) and scale-free property remains unchanged, value of threshold has a strong effect on the network leading to entirely different sets (except few) of largest contributing nodes for two different threshold values. The Appendix enlists the gene names corresponding to the probe set identifiers as given in Tables I and II.

TABLE II. Top contributing nodes (genes) in the localized eigenvectors for the threshold value 0.91.

Set B			Set C		
u^{835}	u^{1635}	u^{641}	u^{1269}	u^{1270}	u^{1224}
210338s_at	208666s_at	201121s_at	211733x_at	201494_at	230416_at
210418s_at	224819_at	208667s_at	230869_at	223209s_at	228283_at
202178_at	209460_at	223716_s_at	228045_at	225284_at	238494_at
38398_at	226395_at	224644_at	211733x_at	201494_at	230416_at
213347x_at	201525_at	200626s_at	242317_at	212788x_at	212474_at

V. CONCLUSIONS AND DISCUSSIONS

Using RMT, we have analyzed gene coexpression network constructed by applying two different threshold values to the data obtained from six brain regions that are metabolically relevant to Alzheimer's disease [6]. The NNSD of adjacency matrix of the largest connecting component of the network follows universal GOE statistics (with $\beta \sim 1$). This universality adds one more feature, based on the spectral correlations, to the gene coexpression network which is common with different model networks [9] proposed to capture various structural properties of real-world networks.

The NNSD gives information about the short-range correlations among the eigenvalues. To probe the long-range correlations, we have studied spectral rigidity via $\Delta_3(L)$ statistics. This analysis shows that the gene coexpression network considered here follows RMT prediction of GOE for very long range of L. Beyond this value of L, deviation in the spectral rigidity is seen, indicating a possible breakdown of universality. This means the network under consideration has *sufficient* randomness, which may be important for *robustness* of the systems, with regularity, which may be to *perform functional tasks*. Mixtures of random connections and regular structure have been emphasized at various places. For instance, information processing in the brain is considered to be random connections among different modular structures [32].

Deviation from the universal RMT predictions identifying system-specific, nonrandom properties of system under consideration might provide clues about important interactions. To extract these system-dependent information, we have performed eigenvector analysis. This analysis reveals that there are some eigenvectors which are highly localized. The component l of a given eigenvector relates to the contribution of node (corresponding gene) l to that eigenvector. Hence, the distribution of the components contains information about the number of genes contributing to a specific eigenvector. Inverse participation ratio IPR, as defined in Eq. (6), distinguishes between one eigenvector with approximately equal components and another with a small number of large components. According to the RMT predictions, the largest contributing nodes (genes) in the localized eigenvectors may have important function or important functional relations among them.

TABLE III. Gene names corresponding to the probe sets for the threshold value 0.89.

Probe set	Gene name
202060_at	Ctr9, Paf1/RNA polymerase II
227636_at	
202916s_at	Family with sequence similarity 20, member B
225921_at	Ninein (GSK3B interacting protein)
214351x_at	Ribosomal protein L13
217731s_at	Integral membrane protein 2B
205003_at	Dedicator of cytokinesis 4
226832_at	
212635_at	Transportin 1
203034s_at	Ribosomal protein L27a
201121s_at	Progesterone receptor membrane component 1
211940x_at	
209860s_at	Annexin A7
208645s_at	Ribosomal protein S14
200673_at	Lysosomal protein transmembrane 4 alpha
221775x_at	Ribosomal protein L22
224616_at	Dynein, cytoplasmic 1
218175_at	Coiled-coil domain containing 92
221511x_at	Cell cycle progression 1
221471_at	Serine incorporator 3
229630s_at	Wilms tumor 1 associated protein
222203s_at	Retinol dehydrogenase 14
221810_at	RAB15, member RAS onocogene family
231896s_at	Density-regulated protein
225950_at	

TABLE IV. Gene names corresponding to the probe sets for the threshold value 0.91

Probe set	Gene name
210338s_at	Heat shock 70kDa protein 8
208666s_at	Suppression of tumorigenicity 13
201121s_at	Progesterone receptor membrane component 1
211733x_at	Sterol carrier protein 2
201494_at	Prolylcarboxypeptidase
230416_at	
210418s_at	Isocitrate dehydrogenase 3 (NAD+)
224819_at	Transcription elongation factor A (SII)
208667s_at	Suppression of tumorigenicity 13
230869_at	Family with sequence similarity 155
223209s_at	Selenoprotein S
228283_at	COX assembly mitochondrial protein homolog
202178_at	Protein kinase C, zeta
209460_at	4-aminobutyrate aminotransferase
223716s_at	Zinc finger, RAN-binding domain
228045_at	
225284_at	DnaJ (Hsp40) homolog, subfamily C
238494_at	TNF receptor-associated factor 3
38398_at	MAP-kinase activating death domain
226395_at	Hook homolog 3 (Drosophila)
224644_at	
211733x_at	Sterol carrier protein 2
201494_at	Prolylcarboxypeptidase
230416_at	
213347x_at	Ribosomal protein S4, X-linked
201535_at	Ubiquitin-like 3
200626s_at	Martin 3
242317_at	HIG1 hypoxia inducible domain family
212788x_at	Ferritin, light polypeptide
212474_at	AVL9 homolog (S. cerevisiase)

The largest connected component is scale-free, indicating the structural importance of few nodes (hubs). Eigenvector analysis shows that top contributing nodes in the localized eigenvectors have relatively low degrees. Note that genes which are hubs or those which connect different communities are also important, as shown by several earlier studies in the network framework [5,13], but the aim of the present work is look for the important genes beyond these structural measures. Changing the value of threshold, while keeping the scale-free structure of network the same, has drastic impact on the localization property of eigenvectors. Almost all the top contributing nodes differ for two different threshold values, indicating impact on the global properties of the underlying network.

Lastly, we discuss here the importance of the analysis and future implications of the results presented in the paper. Several studies have shown that the development of multitarget drugs might give better results than the traditional methods targeting a single protein. Single target design might not always give satisfactory results, as there might be a backup system, which replaces the function of the inhibited target protein. By using multitarget drugs, one can decrease the functionality of entire protein cascades producing more effective results. For example, studies have shown that aging is strongly linked with age-related diseases and they share a common signaling network. Signaling hubs of the age-related protein-protein interaction subnetwork may be good candidates for age-related drug targets. Multitarget drugs attacking hubs of the protein-protein interaction network, "hub-links" (links connecting hubs), bridges (intermodular links having high "betweenness centrality"), or nodes in the overlap of numerous network modules, might give better results [33,34]. Similarly, targeting genes corresponding to the largest contributing nodes in localized eigenvectors may lead to important effect as well. Future investigations are sought in order to know the functionality of these genes corresponding to the top contributing nodes in the localized eigenvectors, which could be then used for such multitarget drug designs.

APPENDIX

Tables III and IV correspond to probe set identifiers from Tables I and II, respectively. First columns of these tables are probe set identifiers (Affymetric ID) and second columns dictate the corresponding gene names. However, the functions of some transcripts are not known yet and some of them have no gene name. The value "-" in the gene name column indicates that information is not available. Note that there are many reasons for probe sets without detailed annotation. We know the sequence on microarray for each probe sets. On the chip, we get all expressed genes, but we do not have secure info for all the gene functions. As the knowledge is growing with the latest available technologies, this gap is decreasing with time. One sure information for the probe set is the Affymetric ID as given in the Tables I and II [26].

[1] S. Maslov and K. Sneppen, Science **296**, 910 (2002); H. Hishigaki *et al.*, Yeast **18**, 523 (2001).

[2] W. S. Liang *et al.*, Physiol. Genomics **28**, 311 (2007); W. S. Liang *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 4441 (2008).

[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[4] S. H. Strogatz, Nature (London) **410**, 268 (2001).

[5] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); S. Boccaletti *et al.*, Phys. Rep. **424**, 175 (2006).

[6] http://www.ncbi.nlm.nih.gov/geo/

[7] A. J. Butte and I. S. Kohane, Proc. AMIA Annu. Fall Symp. **25**, 711 (1999).

[8] For a recent review, see D. F. T. Veiga, B. Dutta, and G. Balázsi, Mol. Biosyst. **6**, 469 (2010).

[9] J. N. Bandyopadhyay and S. Jalan, Phys. Rev. E **76**, 026109 (2007); S. Jalan and J. N. Bandyopadhyay, *ibid.* **76**, 046107 (2007).

[10] R. Potestio, F. Caccioli, and P. Vivo, Phys. Rev. Lett. **103**, 268101 (2009).

[11] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002); M. E. J. Newman, Soc. Networks **27**, 39 (2005); Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).

[12] M. J. Krawczyk, Phys. Rev. E **77**, 065701(R) (2008); G. Palla *et al.*, Nature (London) **435**, 814 (2005); M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004); A. Arenas, A. Fernandez, and S. Gomez, New J. Phys. **10**, 053039 (2008).

[13] R. Guimerá and L. A. N. Amaral, Nature (London) **433**, 895 (2005).

[14] Béla Bollobás, *Random Graphs*, 2nd ed. (Cambridge University Press, London, 2001).

[15] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[16] M. L. Mehta, *Random Matrices*, 2nd ed. (Academic Press, New York, 1991).

[17] T. Guhr *et al.*, Phys. Rep. **299**, 189 (1998).

[18] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[19] P. Seba, Phys. Rev. Lett. **91**, 198104 (2003).

[20] M. S. Santhanam and P. K. Patra, Phys. Rev. E **64**, 016102 (2001).

[21] J. X. de Carvalho, S. Jalan, and M. S. Hussein, Phys. Rev. E **79**, 056222 (2009).

[22] S. Jalan, Phys. Rev. E **80**, 046101 (2009).

[23] P. N. McGraw and M. Menzinger, Phys. Rev. E **77**, 031102 (2008).

[24] G. Zhu, H. Yang, C. Yin, and B. Li, Phys. Rev. E **77**, 066113 (2008).

[25] K. Zyczkowski, *Quantum Chaos*, edited by H. A. Cerdeira, R. Ramaswami, M. C. Gutzwiller, and G. Casati (World Scientific, Singapore, 1991).

[26] H. Göhlmann and W. Talloen, *Gene Expression Studies Using Affymetrix Microarrays* (Chapman and Hall, London, 2009).

[27] X. Zhou, M. C. Kao, and W. H. Wong, Proc. Natl. Acad. Sci. U.S.A. **99**, 12783 (2002); D. Smet *et al.*, Bioinformatics **18**, 735 (2002).

[28] F. Luo *et al.*, Bioinformatics **8**, 299 (2007).
[29] T. A. Brody, Lett. Nuovo Cimento **7**, 482 (1973).
[30] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **68**, 046109 (2003).
[31] M. A. M. de Aguiar and Y. Bar-Yam, Phys. Rev. E **71**, 016106 (2005).
[32] J. D. Cohen and F. Tong, Science **293**, 2405 (2001).
[33] P. Csermely, V. Ágoston, and S. Pongor, Trends Pharmacol. Sci. **26**, 178 (2005); T. Korcsmáros *et al.*, Exp Op Drug Discovery **2**, 1 (2007); G. R. Zimmermann, J. Lehár, and C. T. Keith, Drug Discovery Today **12**, 34 (2007); M. Antal, C. Böde, and P. Csemely, Curr. Protein Pept. Sci. **10**, 161 (2009); P. Csermely, Trends Biochem. Sci. **33**, 569 (2008).
[34] G. I. Simkó *et al.*, Genome Medicine **1**, 90 (2009).