THRESHOLD SELECTION BASED ON RANDOM MATRIX THEORY FOR GENE CO-EXPRESSION NETWORK

Laura BARACALDO [1]
Luis LEAL[2]
Liliana LOPEZ-KLEINE[1]

- ABSTRACT: Random Matrix Theory (RMT) methods for threshold selection had only been applied in a very low number of studies aiming the construction of Gene Co-expression Networks (GCN) and several open questions remained, especially regarding the general applicability regardless the diverse data structure of gene expression data sets. Moreover, no clear methodology to follow at each step was available. Here, we show, that RMT methodology is, in fact, capable to differentiate Gaussian Orthogonal Ensemble (GOE) from Gaussian Diagonal Ensemble (GDE) structure for a great number of simulated data sets and that results are similar to those obtained with the reference method of clustering coefficient.

- KEYWORDS: Similarity matrices; threshold selection; random matrix theory; gene co-expression networks

1 Introduction

The cell is a system of multiple interacting entities with specific functions, whose intrinsic complexity can be studied under mathematical frameworks such as networks. Gene co-expression networks (GCNs) are a common representation of this complex system as they depict those pair of genes having similar expression profiles, and therefore, highlight those genes that might be functionally related to the same pathway or protein complex. In GCNs, nodes represent genes and significant co-expression relationships are represented by edges. They are constructed through two main steps. First, a similarity measure is assessed on the gene expression profiles of each pair of genes conducing to a gene by gene similarity matrix. Second, a similarity threshold is selected, above which the relationship is assumed to be significant. Once this threshold is applied, only genes with significant similarities will be kept for the network and it can be described as an adjacency matrix with zeros for all non-significant similarities.

The step of choosing the threshold should be as objective and efficient as possible in order to reflect biological mechanisms eliminating noise. Concerning the threshold selection, several methodologies have been used and reviewed elsewhere (LÓPEZ-KLEINE and LEAL, 2014). There are methodologies based on either some statistical

[1] Universidad Nacional de Colombia - Department of Statistics, Bogotá, Colombia. E-mail: *lnbaracldol@unal.edu.co*; *llopezk@unal.edu.co*
[2] Imperial College London - Department of Life Sciences, London, UK. E-mail: is: *luis.leal-ayala@imperial.ac.uk*

criteria (FREEMAN *et al.*, 2007; CARTERS *et al.*, 2004) or criteria regarding the nature of the final graph (GUPTA *et al.*, 2006; ELO *et al.*, 2007). Methods based on Random Matrix Theory (RMT) belong to this last group because they aim to detect the transition between a random network and a network with a systemic structure as is expected for a biological network (GUPTA *et al.*, 2006).

RMT was initially proposed to explain statistical properties of the nuclear spectrum. Predictions for different spectral conditions of complex systems, such as unordered systems or chaotic quantic systems, were obtained using this approach (SNAITH *et al.*, 2003). From the beginning of the 20th century, different applications of RMT have been found in other sciences such as physics, engineering, finances, meteorology, etc. (SARIKA *et al.*, 2007).

RMT approaches for the threshold selection in complex networks are based on the characterization of the statistical distribution of the nearest neighbour spacing distribution (NNSD) of the eigenvalues of the adjacency matrix (CVETKOVIT *et al.*, 1980; SARIKA *et al.*, 2007) NNSD represents the probability of finding neighbour eigenvalues with any given spacing. This probability is expected to have certain probabilities, depending on the correlation structure underlying the eigenvalues. Therefore, given the nature of correlations, eigenvalues are able to differentiate random from non-random networks (i.e. systemic networks) (SARIKA *et al.*, 2007). This property of RMT makes it useful for the study of biological networks, such as co-expression networks, which have a systemic structure (BARABASI and ALBERT, 1999; WATTS *et al.*, 1998) and clearly differentiate from random networks (ERDOS and RENYI, 1960). Because the nature of the graph is selected in RMT, makes it a completely objective and effective methodology for threshold selection.

Very few studies have applied methods based on RMT for threshold selection and open questions remain about its applicability (LÓPEZ-KLEINE and LEAL, 2014; LUO *et al.*, 2007). Here, we revisit the method and prove its applicability on GCNs. We use simulations based on RMT theory of several datasets, constructing random and systemic graphs in order to proof that the methodology can be applied to a wide range of gene expression data structures. This allowed us to establish that the method is very sensitive to some steps of the RMT methodology, but not to data nature, as suggested before (LÓPEZ-KLEINE and LEAL, 2014). The method is compared against a traditional method based on topological properties of the graph using simulated and real data sets. Scripts have been written in R language (R CORE TEAM, 2017) and are available under request.

2 Materials and methods

2.1 Structure of adjacency matrices in RMT

An adjacency matrix is an $N \times N$ matrix with i rows and j columns in which N represents the number of nodes (genes) of the graph. In the most simple adjacency matrix, the connecting nodes are represented by A_{ij}=1 and non-connecting ones by A_{ij}=0. The connecting elements may also be indicated as $A_{ij}=\rho_{ij}$, where ρ_{ij} is some similarity measure between nodes, for genes generally the Pearson correlation between gene expression profiles.

A non-random graph is represented by an adjacency matrix with few highly connected nodes or hubs. These graphs are known in RMT as Gaussian Diagonal Ensemble (GDE) graphs, which's structure can be assessed by a Poisson distribution of the NNSD of the

eigenvalues. GDEs are real, diagonal matrices with uncorrelated (independent) eigenvalues (SARIKA *et al.*, 2007) These graphs generally represent real systems, such as biological graphs.

On the other hand, a random graph in RTM is represented by a Gaussian Orthogonal Ensemble (GOE) distribution of NNSD of the eigenvalues, which are correlated (SARIKA *et al.*, 2007). These graphs do generally not represent real systems such as a biological network, because every node is randomly connected to all other nodes.

2.2 RMT Methodology for threshold selection

RMT methods aim at differentiating between these two types of structures by analysing NNSD of the eigenvalues after a previous step called unfolding or transformation of eigenvalues.

Unfolding of eigenvalues of the obtained matrices

In order to find out the universal properties of the fluctuations of the eigenvalues $\{\lambda_i\}$ associated to the network, spurious effects are removed using unfolding or transformation $\in_i = \bar{N}(\lambda_i) = \int_{\lambda_{min}}^{\lambda} \bar{\rho}\,(\lambda')d\lambda'$, where $\bar{\rho}(\lambda')$ is the eigenvalues density. The estimation of the function $\bar{\rho}(\lambda')$ was achieved by the non-parametric kernel smoothing based on the empirical distribution of the eigenvalues. (See appendix, script C).

Statistics associated to NNSD

This step will define if the graph described by the adjacency matrix is GOE or GDE. Two main properties are considered: 1) global properties like spectral density or distribution of eigenvalues $\rho(\lambda)$ and 2) Local properties, like fluctuations around $\rho(\lambda)$. Fluctuations are more popular and studied through the NNSD of eigenvalues. They indicate the probability of finding neighbouring values, given a specific spacing and follow two distinct universal distributions depending on the underlying correlation structure (CVETKOVIC *et al.*, 1980, SARIKA et al, 2007). Once eigenvalues are scaled as $\in_i$, spacings to the nearest neighbour are obtained as $s^{(i)} = \in_{i+1} - \in_i$; which, due to the previous scaling have a mean value of 1 in both, GDE and GOE systems. The distribution of the spacings $P(s)$ will then be defined as the probability distribution of $s^{(i)}$ (see appendix, script C).

The goodness of fit to GOE or GDE was evaluated using a Kolmogorov-Smirnov test. The chosen threshold, test statistic and P-value were retained.

Border effect on Statistics associated to NNSD

We evaluated the border effect, eliminating a certain percentage of eigenvalue spacings at both ends: 1%, 5%, 10% and 20% of bordering spacings were eliminated before unfolding in order to test if the adjustment changes.

2.3 RMT Methodology for threshold selection

With the aim of testing the RMT approach to establish the difference between random graphs and non-random graphs, 5000 simulations of each GOE and GDE matrices were generated, their eigenvalues unfolded and NNSD tested. In this section we describe, how data were simulated. For these simulations, statistics were applied in order to retrieve the expected NNSD. Nevertheless, no threshold was established.

Simulation of random and non-random adjacency matrices

If A is a Gaussian random matrix $m \times n$, denoted by $G_\beta(m, n)$ then the density function of its elements is given by:

$f(a_{ij}) = \frac{1}{(2\pi)^{\beta mn/2}} exp\left\{-\frac{1}{2}\|A\|_F^2\right\}$, where $\|\|A\|_F^2$is the Frobenius norm of matrix A. $G_\beta$is orthogonally invariant (see appendix, script A).

Adjacency matrices with uncorrelated eigenvalues can be considered as a random Poisson process (BERRY, 1981), which is obtained from an integral system that can be simulated by Gaussian diagonal random matrices (see appendix, script B).

2.4 Actual data

In order to evaluate all steps of the RMT method to select the threshold, real gene expression data available at NCBI (http:/www.ncbi.nlm.nih.gov/geo/), obtained on plants, were analysed. The data used was obtained comparing control conditions to presence of pathogens (platform accession numbers: GPL2025 for rice data set (6 samples and GPL198 for the Arabidopsis data set (18 samples)). Data was processed using the Bioconductor package affy (GAUTIER *et al*., 2004). Additionally, filtering of non-informative genes was accomplished eliminating genes with no differential expression among treatments. Pearson correlation between all pairs of genes was used to obtain the similarity matrices for each data set.

2.5 Reference method for threshold selection: Clustering coefficient

Our results were compared against a method based on network topological properties (ELO *et al*., 2007) on the real data set. The method computes the network's clustering coefficient at different thresholds, and find deviations from a randomized network with the same degree distribution. First, the clustering coefficient is computed at a given threshold (e.g., τ_v= 0.01) by equation #1. Then, the expected value of the clustering coefficient under a random network model is calculated by using equation #2. The threshold is gradually increased to evaluate the difference between the network's clustering coefficient and its randomized counterpart network. Such difference between clustering coefficients is expected to increase monotonically as long as noisy edges are removed from the network. This is a discrete optimization problem formulated by (ELO *et al*., 2007) in equation #3.

$$C(\tau_v) = \frac{1}{K}\sum_{k_i>1}\frac{2D_i}{k_i(k_i-1)} \quad (1)$$

In equation (1), $C(\tau_v)$ is the observed clustering coefficient in the network, k_i denotes the number of neighbours of gene i or node degree; D_i denotes the number of edges between the neighbors of gene i. K is the number of genes with $k_i > 1$.

$$C_r(\tau_v) = \frac{\left(\overline{k_d} - \bar{k}\right)^2}{\bar{k}^3 N} \qquad (2)$$

In equation (2), N denotes the number of connected nodes in the network, $\bar{k} = {}^1/_N \sum_{i=1}^{N} k_i$, and $\overline{k_d} = {}^1/_N \sum_{i=1}^{N} {k_i}^2$.

The optimum similarity threshold $\tau^*$is determined by finding the minimum threshold τ_v for which the difference between the clustering coefficients is maximum. Thus, τ^* is the first local maximum of the curve $C(\tau_v) - C_r(\tau_v)$.

$$\tau^* = \min_{v} \{\tau_v : C(\tau_v) - C_r(\tau_v) > C(\tau_{v+1}) - C_r(\tau_{v+1})\} \qquad (3)$$

In equation (3), τ^* is the selected similarity threshold; $\tau_{v+1} = \tau_v + 0.01$ with $\tau_v \in [0.01,\ 0.99]$

3 Results and discussion

No clear methodology to follow at each previously described step was available. Here, we show, that RMT methodology is in fact capable to differentiate GOE from GDE structure for a great number of simulated gene sets and that results are similar to those obtained with the reference method of clustering coefficient.

RMT methods for threshold selection had only been applied in a very low number of studies aiming the construction of GCNs (ELO *et al.*, 2007; LÓPEZ-KLEINE and LEAL, 2014; LUO *et al.*, 2007). Five thousand simulations of random (GOE) and non-random networks (GDE) were obtained. For each simulation, all steps of the RMT were undertaken. On real data, the clustering coefficient methodology was applied in order to compare. Figure 1 shows distribution of one of the simulations of each system in order to illustrate that the simulations conducted to the expected distributions.

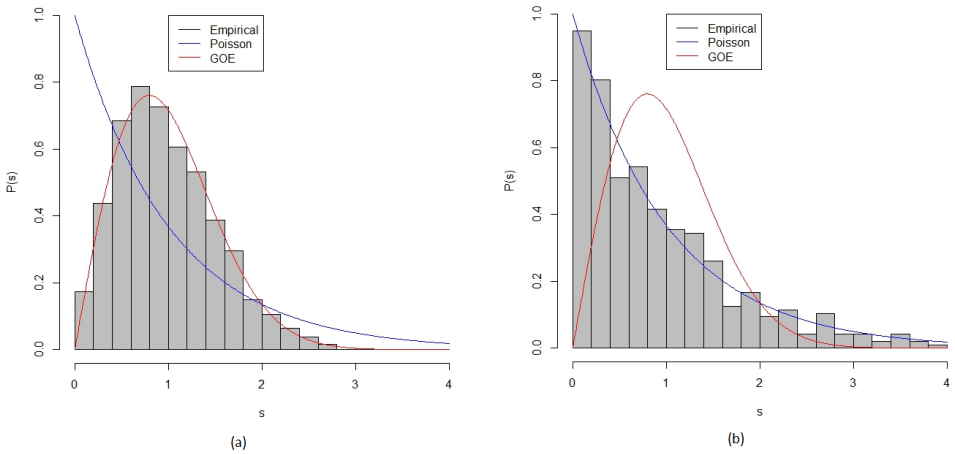

Figure 1 - Histogram of the empirical distribution of NNSD of a simulated GOE matrix (a) and GDE-Poisson matrix (b) to the expected theoretical distribution.

P-values obtained after applying Kolmogorov-Smirnov test to all simulated data sets are shown in Figure 2. A change is observed when more than 10% of the bordering values are eliminated. Elimination of 1% seems to have a positive effect on P-values, increasing adjustment to the theoretical distributions slightly and is therefore encouraged. These test results indicate that, a similarity matrix conducting to a random graph or a non-random graph can be accurately be detected using RMT methodology (after scaling appropriately), and that the final goodness to fit test is reliable.

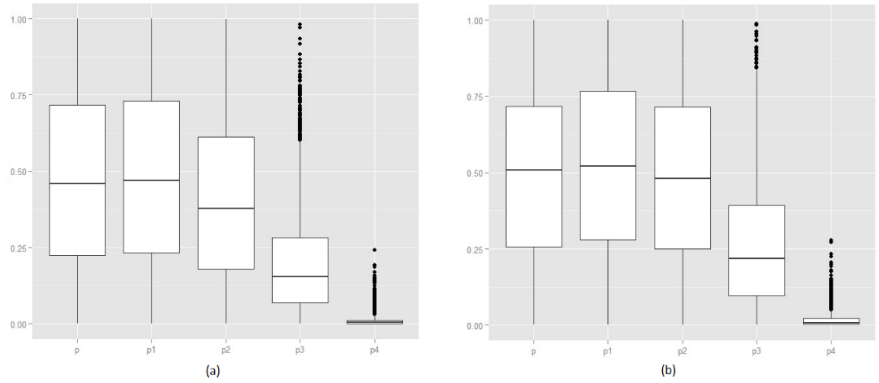

Figure 2 - Boxplot of P-values obtained for the Kolmogorov-Smirnov goodness to fit test to the theoretical distribution of 5000 simulations of (a) GOE matrices and (b) GDE-Poisson matrices. P1,P2,P3,P4 are boxplots of the same P-values after eliminating 1%,5%,10% and 20% of bordering spacings respectively.

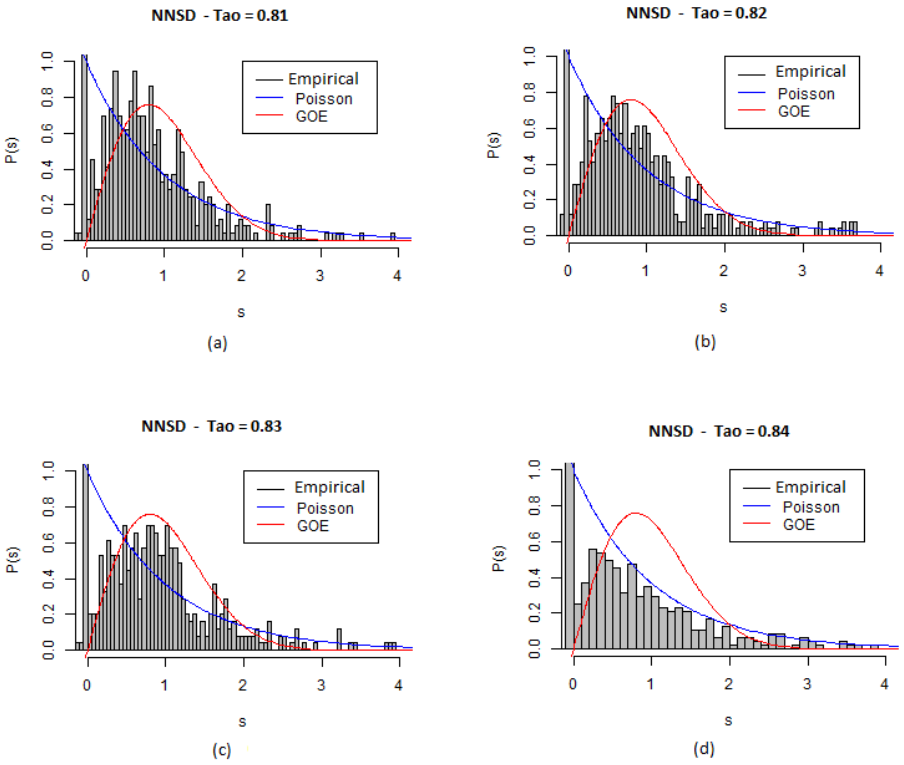

Figure 3 - Histogram of the empirical distribution of NNSD of real data sets of rice (a, b) and Arabidopsis (c, d) showing how the transition between GOE adjustment and GDE-Poisson adjustment allows detecting the threshold.

For the real data sets transitions from GOE to GDE adjustment were detected at $\tau = 0.98$ for rice and $\tau = 0.88$ for Arabidopsis. Results for the rice data set are different and for the Arabidopsis data set are similar. This indicates that the transition between random and non-random systems is detected at a slightly different point with RMT methods, than with clustering coefficient, although thresholds are similar. Nevertheless, for this particular rice data set, only three new edges are added when the threshold is increased from 0.82 to 0.98 and no edge is added when threshold is changed from 0.84 to 0.88 for the Arabidopsis data set.

These results suggest that both methods conduct to similar results and are objective in selecting the correct threshold. Nevertheless, other statistical methods like proposed in (Velez et al, 2014) and elsewhere should be investigated and compared to the clustering coefficient and RMT.

Acknowledgements

The authors would like to thank Gabriel Tellez (Physics Department, Universidad de Los Andes, Bogota Colombia) and Campo Elías Pardo (Statistics Department, Universidad

Nacional de Colombia, Bogotá) for their valuable comments on the master final work that conducted to this manuscript.

BARACALDO, L.; LEAL, L.; LOPEZ-KLEINE, L. Seleção do limite com base na teoria da matriz aleatória para a rede de co-expressão genética. *Rev. Bras. Biom*. Lavras, v.36, n.2, p.376-384, 2018.

RESUMO: Métodos baseados na Teoria da Matriz Aleatória (RMT) para seleção do limiar tem sido aplicados apenas em um pequeno número de estudos com o objetivo de construir redes de co-expressão de genes (GCNs) e várias questões permaneceram abertas, especialmente no que se refere à aplicabilidade geral, independentemente da heterogeneidade da estrutura dos conjuntos de dados da expressão gênica. Além disso, não estava disponível nenhuma metodologia clara para ser seguida em cada etapa do processo. Neste artigo, mostramos que a metodologia RMT pode, de fato, diferenciar o GOE da estrutura GDE para um grande número de conjuntos de dados simulados, e os resultados são semelhantes aos obtidos com o método de referência do coeficiente de agrupamento.

PALAVRAS-CHAVE: Matrizes de similaridade; seleção do limiar; teoria da matriz aleatória; redes de co-expressão de genes.

References

BARABASI, R.; ALBERT, A. L. Emergence of Scaling in Random Networks. *Science*, v.286, p.509, 1999

BERRY, M. V. Quantizing a classically ergodic system: Sinai's billiard and KKR method, *Annals of Physics*, v.131, p.163-216, 1981

CARTERS, L.; BRECHBÜHLER, C. M.; GRIFFIN, M.; BOND, A. T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, v.20, p.2242–50, 2004

CVETKOVIC, D. M.; DOOB, M.; SACHS, H. *Spectra of graphs: theory and application*. London: Academic Press, 1980

ELO, L. L.; JÄRVENPÄÄ, H.; ORESIC, M., LAHESMAA, R.; AITTOLALLIO, T. Systematic construction of gene co-expression networks with applications to human t helper cell differentiation process, *Bioinformatics* v.23, p.2096–2103, 2007.

ERDOS, P.; RENYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*. v.5, p.17-61, 1960

FREEMAN, T. C.; GOLDOVSKY, L.; BROSCH, M.; VAN DONGEN, S.; MAZIE, P.; GROCOCK, R. J.; FREILICH, S.; THORNTON, J.; ENREIGHT, A. J. Construction, visualisation, and clustering of transcription networks from microarray expression data, *PLoS Computational Biology*, v.3, n.10, p.2032–2042, 2007.

GAUTIER, L.; COPE, L.; BOLSTAD, B. M.; IRIZARRY, R. A. affy - analysis of affymetrix genechip data at the probe level. *Bioinformatics*, v.30, n.3, p.307–315, 2004.

GUPTA, A.; MARANAS, C. D.; ALBERT, R. Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites. *Bioinformatics*, v.22, n.2, p.7, 2006

LÓPEZ-KLEINE, L.; LEAL, L. Similarity threshold selection tools for gene co-expression networks. In: ROGERS, J. V. (Ed.) *Microarrays: Principles, applications and technologies.* New York: Nova Science Publishers, 2014.

LUO, F.; YANG Y.; ZHONG, J.; GAO, H; KHAN, L.; THOMPSON D. K.; ZHOU, J. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, v.8, n.1, p.299, 2007.

NAYAK, R. R.; KEARNS, M.; SPIELMANN, R. S. ; CHEUNG, V. G. Co-expression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Research*, v.19, n.11, p.1953–1962, 2009.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2017.

SARIKA, J.; BANDYOPADHYAY J. N. Random matrix analysis of complex networks, *Phys. Rev*. v.76, 2007.

SCOTT, D. W. *Multivariate density estimation: Theory, practice, and visualization*, New York: John Wiley, 1992.

SHERIF, M. A.; ABUL-MAGD, A. Y. *Effect of unfolding on the spectral statistics of adjacency matrices of complex networks*, 2012.

SNAITH, N. C.; FORRESTER, P. J.; VERBAARSCHOT, J. J. M. Developments in random matrix theory, *J. Phys*., A36, 2003.

VÉLEZ, J. I.; CORREA, J. C.; ARCOS-BURGOS, M. New Method for Detecting Significant p-values with Applications to Genetic Data. *Revista Colombiana de Estadística*, v.37, n.1, p.69-78, 2014.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, v.393, p.440-442, 1998.

Received on 04.10.2016

Approved after revised on 17.10.2017