




Effects of threshold on the topology of gene co-expression networks†

Cynthia Martins Villar Couto, * César Henrique Comin and Luciano da Fontoura Costa

Cite this: *Mol. BioSyst.*, 2017, 13, 2024

Received 21st February 2017,
Accepted 21st July 2017

DOI: 10.1039/c7mb00101k

rsc.li/molecular-biosystems

Several developments regarding the analysis of gene co-expression profiles using complex network theory have been reported recently. Such approaches usually start with the construction of an unweighted gene co-expression network, therefore requiring the selection of a suitable threshold defining which pairs of vertices will be connected. We aimed at addressing such an important problem by suggesting and comparing five different approaches for threshold selection. Each of the methods considers a respective biologically-motivated criterion for electing a potentially suitable threshold. A set of 21 microarray experiments from different biological groups was used to investigate the effect of applying the five proposed criteria to several biological situations. For each experiment, we used the Pearson correlation coefficient to measure the relationship between each gene pair, and the resulting weight matrices were thresholded considering several values, generating respective adjacency matrices (co-expression networks). Each of the five proposed criteria was then applied in order to select the respective threshold value. The effects of these thresholding approaches on the topology of the resulting networks were compared by using several measurements, and we verified that, depending on the database, the impact on the topological properties can be large. However, a group of databases was verified to be similarly affected by most of the considered criteria. Based on such results, it can be suggested that when the generated networks present similar measurements, the thresholding method can be chosen with greater freedom. If the generated networks are markedly different, the thresholding method that better suits the interests of each specific research study represents a reasonable choice.

1 Introduction

The study of complex systems, such as in biology, is a particularly hard task because one cannot totally understand a phenomenon just by reducing it to isolated components.¹ A more accurate approach would be to consider the whole interaction scenario, but this would be a very complicated task, unfeasible currently. Yet, the studies of complex structures will greatly benefit from the consideration of pairwise interactions, examining not only the individual components, but also how they relate to each other. Such studies are inherently part of the new area of complex networks, which is itself an extension of graph theory.² Complex network research aims at characterizing, analyzing, modeling and simulating complex systems.³ The relationship between network theory and biological systems became even more important as a consequence of the awareness that biological networks are not random, and follow several organizing principles of topology and structure.⁴ This set of principles is meaningful to

characterize the network, and reveals important features of the studied system. The complex network principles can be applied in a specific field of biological research, namely genomics.

The genomic era was concerned mostly with the specific analysis of the genetic code. Currently, the challenge has shifted to the interpretation of the function of these genes in the genome, so as to clarify how they interact one another as the means to execute certain biological functions.⁵ The aim of the post-genomic era of systems biology places more emphasis on the modeling of biological interactions (genes, small and interference RNA, proteins, and metabolites, among others), in order to bring these isolated components together in resulting networks.^{3,6}

A particularly interesting kind of genetic network is the gene co-expression network, which can be used to analyze the gene expression data on a global scale. This approach is based on the fact that genes encoding proteins that are part of some complex, or that are engaged in the same pathway, are usually co-regulated, and may exhibit a correlated expression pattern.⁵ The construction of the gene co-expression network starts with a gene-expression database, which contains different expression profiles showing how gene expression is influenced by some specific biological activity (*e.g.* disease, developmental

São Carlos Institute of Physics, University of São Paulo, PO Box 369, 13560-970, São Carlos, SP, Brazil. E-mail: roravillar@gmail.com

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7mb00101k

stage, growth condition, *etc.*). The intensities of gene expression are normalized, an association index is applied between every pair of genes, and their correlations are ranked. Some association indexes that have been used to measure similarity between biological molecules are: the Jaccard index, Simpson index, Geometric index, Cosine index, Pearson correlation coefficient, hypergeometric index and connection specificity index.⁷ Afterwards, a probabilistic method is used to evaluate the probability of observing a particular configuration of ranks across the different organisms by chance, and a cut off (threshold) has to be chosen to indicate if a correlation means association (above the threshold value) or not (under the threshold value). Among the indexes mentioned before, one of the most frequently used is the Pearson correlation coefficient,^{8,9} a straightforward and objective way to measure the correlation between expression profiles.^{1,10–18}

Various studies in the systems biology field have been based on the analysis of weighted gene co-expression network analysis to reveal genes that play an important role in the regulation of some biological dysfunction, as a manner to show potential gene markers.^{19,20} A typical use of a co-expression network is to functionally annotate genes, based on interaction-profile similarities, with the benefit of associating genes with a known function.⁷ The well-known study of Zhang and Horvar's¹⁶ proposed a method of “soft thresholding” co-expression networks, determining the connection strengths between each gene pair and maintaining a biologically motivated criterion, based on the fact that many biological networks approximate to a scale-free topology.^{21–24}

In general, graph theory is a powerful tool to describe relationships among genes based on co-expression over distinct biological scenarios.²² In Stuart *et al.*,⁵ a multiple-species gene co-expression network was built to elucidate gene function on a global scale. Microarray data from humans, flies, worms and yeast were united in groups of orthologous genes to be assigned to a single metagene, and then a multiple-species gene co-expression network was generated. The network showed several co-expression relationships, conserved across evolutionary time, between newly evolved and ancient metagenes, implying that these groups of genes may be functionally related. The biological function of some of the genes were confirmed, which also shed light on those with no previously known function, based on the co-regulation and evolutionary conservation.⁵

As mentioned before, after the calculation of correlation between every pair of genes, it is necessary to set a value – or threshold – that indicates whether a pair of vertices is connected or not. The threshold choice constitutes an important, even critical, step during the network construction, because a value too high would imply few connections among vertices, possibly resulting in unconnected networks. Alternatively, a low threshold value would imply several connections among vertices, making the network nonspecific and less significant.

Thresholding a network is desirable to keep connections that are significant according to some reasonable criterion and reject those that are not. The approach has been systematically used in network science, including climate,²⁵ transportation²⁶

and social²⁷ networks. Also, many studies involving biological networks^{5,28–33} have employed thresholding methodologies, since it presents several advantages, such as lower computational cost and elimination of less relevant relationships. Even more important, there is no established definition of many topological measurements in weighed networks.³⁴ For instance, Saramaki *et al.*³⁴ analyze four alternative definitions of clustering coefficient in such networks.

The question of what would be the ideal threshold value is as important as it is difficult, and ultimately can only be answered by taking into account the full biological mechanisms underlying the studied networks. As this is currently impossible, we need to resort to other approaches capable of shedding some light into this problem. Indeed, even if the full biological mechanisms were known, they would be represented as a weighted graph. This kind of graph is relatively hard to analyze compared to binary graphs, which can be characterized by a more established set of measurements. The thresholding of a weighted graph is desirable as a way to simplify it, and also to emphasize topological properties, which depends on the focus of interest. At the same time, relatively little attention has been paid to the development of mathematical methods to select an appropriate threshold value while studying gene co-expression networks. One of the main purposes of this work is to propose an empirical approach to this problem, accompanied by some statistical analysis and validation. In particular, we aimed at characterizing distinct gene expression networks, and comparing the behavior of network topological measures under different threshold values chosen by different approaches.

We propose and characterize 5 methodologies for setting a suitable threshold value for gene co-expression networks. Each method considers a distinct criterion for setting the threshold, and depending on the type and purpose of the research, the proposed criterion can be more or less desirable. If the aim of the research is to identify the most relevant relationships while simplifying the network (eliminating redundancies that do not necessarily contribute to information gain), one could resort to the entropy maximization of selected network properties.³⁵ If the objective of the study is to analyze gene relationships over several neighborhoods,⁷ such as the connected components of the network,² it would be interesting to set the threshold according to the desired size of the giant component of the network. Since many biological networks are scale free,^{21–24} imposing a threshold where the degree distribution of the resulting network is as close as possible to a power-law may generate a network that has a clearer biological meaning.²¹ Another interesting approach is to constrain the average number of relationships that genes may have in the network, which immediately translates to setting the average degree of the network. We also consider a situation in which the threshold is set at the most stable topological configuration, that is, a threshold value in which small changes to the threshold hardly change any topological property of the network. Our aim is to identify how the resulting topology of the network changes according to the method of choice. Such an analysis can lead to interesting results. For instance, cases in which the thresholds

obtained by most of the aforementioned methods are similar avoid the need to choose a specific method. In contrast, in cases where rather different threshold values are obtained, it is important to carefully consider which aspect of the research should be emphasized, implying the respective choice of thresholding methodology.

The article is organized as follows. First, we describe the selection of a representative group of biological public databases to investigate the thresholding methodologies. By using each one of these databases, we constructed 20 different co-expression networks varying the threshold value. Next, different approaches for threshold selection are detailed and applied, and then compared to verify the influence of the threshold selection criteria on the final network. The results obtained by each criterion are presented and it is shown how principal component analysis can be used to provide the basis for comparing the effects of the considered thresholding approaches respectively to a variety of gene co-expression databases.

2 Dataset and methodology

2.1 Gene expression profiling data

The dataset used in our study consists of 21 genetic microarray experiments that were retrieved from the Gene Expression Omnibus³⁶ and the Array Express database (www.ebi.ac.uk/arrayexpress).

The considered datasets are indicated in (Table S1, ESI[†]), which includes for each entry: the biological group to which the data belong, a reference code for the experiment, the experiment access number, a key bibliographical reference, the available number of genes and biological replicas, and the data organization.

2.2 Network creation and analysis

Each experiment was selected from the Microarray database, as illustrated in Fig. 1a, containing a respective gene expression profile, which has information about the gene expression level under distinct experimental conditions (Fig. 1b). For each experiment, we used the Pearson correlation coefficient to

quantify the co-expression between each pair of genes,^{5,19,31} (Fig. 1c). The Pearson correlation coefficient measures the linear relationship between two random variables. This coefficient provides a value between -1 and 1 , where 1 corresponds to a perfect linear overlap, -1 indicates perfect linear anticorrelation, and 0 characterizes the absence of correlation. The calculated Pearson values can be understood as a weighted co-expression network containing $N(N - 1)/2$ relationships between the N considered genes.

The resulting weighted matrix (Fig. 1d) is subjected to different threshold values (Fig. 1e), and the respective adjacency matrices are used to build the corresponding gene co-expression network (Fig. 1f). Finally, the different criteria for threshold determination can be applied on these generated networks (Fig. 1g) in order to choose the most appropriate one (Fig. 1h). Note that the statistics used for threshold selection are calculated over the giant component of the network, *i.e.* the major group of connected vertices after thresholding.

2.3 Threshold selection methods

The criteria to be used and compared are:

- (1) Maximum variation of topological properties (based on PCA) (2.3.1);
- (2) Maximum measurement entropy (2.3.2);
- (3) Size of the giant component (2.3.3);
- (4) Power-law degree distribution (2.3.4);
- (5) Specified average degree (2.3.5).

2.3.1 Maximum variation of topological properties. For this thresholding criterion, we considered the following topological properties characterizing each network:

(i) Degree statistics. The degree of a vertex is the number of edges incident on that vertex.³⁷ We considered the minimum, maximum, average, median, standard deviation and entropy of the degree values.

(ii) Local transitivity. The local transitivity measures the probability that two neighbors of a vertex are connected.³⁸ We considered the minimum, maximum, average, median, standard deviation and entropy of the transitivity values.

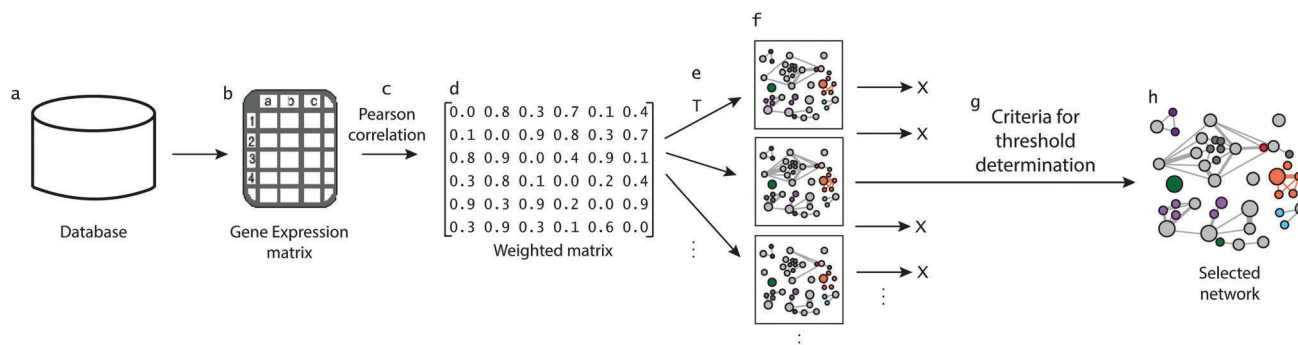


Fig. 1 Steps of the proposed threshold selection methodology. A set of experiments are selected from the Gene Expression Omnibus and Array Express databases (a). Each experiment contains a respective gene expression dataset (b). The Pearson correlation coefficient is then calculated for all gene pairs in each experiment (c), generating a respective weighted matrix (d). Next, a set of thresholds are applied to the correlation values (e), resulting in a respective set of networks (f). The criteria for threshold determination are applied to the networks (g), in order to choose the more appropriate alternative (h). T represents the candidate threshold value.

(iii) Betweenness centrality. In order to calculate this measure for a given vertex, we first count the number of shortest paths between two vertices that pass through the vertex. Then, the result is divided by the total number of shortest paths between the two vertices. Next, this calculation is repeated for all pairs of vertices in the network, the sum of the obtained values defines the betweenness of the vertex.^{37,39,40} We considered the minimum, maximum, average, median, standard deviation and entropy of the betweenness values.

(iv) Size of the giant component. The number of vertices in the giant component tends to increase with the network average degree and provides an interesting indication about the uniformity of the connections along the network.²

(v) Number of edges in the giant component. Number of edges that connect the vertices in the giant component.²

(vi) Clustering coefficient. The clustering coefficient is a measure of the tendency of vertices to cluster together in a network.^{41,42} This measure can be thought of as a global transitivity of the network.

(vii) Degree assortativity. Degree assortativity expresses the preference of vertices in a network to attach to others that have similar degrees.⁴³

(viii) Average shortest path length. Average shortest path length is a concept in network topology that is defined as the average number of steps along the shortest paths for all possible pairs of network vertices.⁴⁴ It is a measurement of the effectiveness of information transport on a network.⁴⁵

(ix) Diameter. The diameter of a network is the largest of the smallest number of vertices which must be traversed in order to travel from one vertex to another. Paths with backtracks, detours, or loops are excluded from consideration.⁴⁶

(x) Rich club coefficient. The rich-club coefficient is a measurement of how much well-connected vertices interconnect one another.⁴⁷ Networks which have a relatively high rich-club coefficient are said to demonstrate the rich-club effect and will have many connections between vertices of high degrees.

(xi) Matching index. The matching index can be assigned to each edge in a network in order to quantify the similarity between the connectivity of the two vertices adjacent to that edge.⁴⁸

For each gene expression experiment, 20 thresholds linearly spaced were applied to the respective weighted Pearson correlation matrix. The topological properties described above were applied to each generated network, and principal component analysis (PCA) was used to project the properties into a low dimensional space defined by two representative features (given by the first two principal axes). PCA is a statistical procedure that uses an orthogonal transformation to convert a set of variables, that are possibly correlated, into a lower dimensional set of uncorrelated variables called principal components.⁴⁹ So, the first component has the major possible variance, and each successive component has the largest possible variance, limited by being orthogonal to the previous component (showing the structure of data in a way that better explains its variance).

For small threshold values, we expect that the properties of the network will not significantly change when varying the

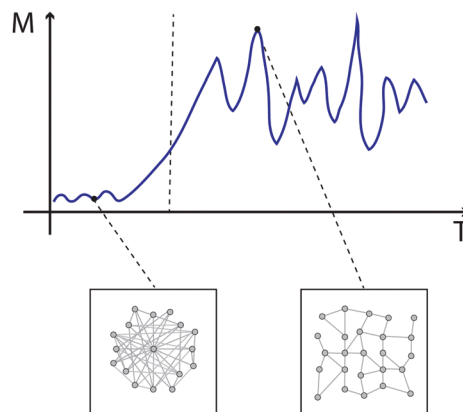


Fig. 2 Illustration of the expected variation of a network measure as the threshold increases. Low threshold values should result in stable measure values, while high threshold values may lead to a more unstable topology.

threshold, since the network tends to be highly connected. As the threshold increases, the network starts to break apart into smaller connected components, and the properties of the network may drastically change for slightly different thresholds. The point where the network goes from a stable state (*i.e.*, small change in properties as the threshold varies) to an unstable one (large property variation with the threshold) should represent a suitable choice for thresholding the correlation values. An example is shown in Fig. 2.

The networks generated from each threshold value can be represented as points in the 2D space defined by the PCA. In order to find the threshold for which the topology goes from the stable to the unstable state, we first measured the euclidean distance between points associated with two consecutive threshold values. Then, the first threshold for which the distance is larger than d_{\max}/f is chosen as the more appropriate threshold for the dataset. d_{\max} is the maximum euclidean distance between PCA points observed for the experiment. f is a constant used to set the amount of variation in the PCA, relative to the maximum, where the network is considered unstable. In our analysis we consider $f = 1.7$, which we found to provide reasonable thresholds for all datasets.

This criterion preserves the characteristics of a network with a relatively robust structure, because the optimum threshold value will be selected before the measurements undergo stronger variations resulting from, for instance, a break of the giant component into other smaller components. It will also tend to result in a relatively sparse network, since larger topological variations are more common in networks with low average degree.^{50,51}

The maximum variation of the topological property criterion will be useful for showing the genes that have highly correlated expression, and also contribute to the integrity of a system of co-expression gene products. This characteristic allows the clustering of genes based on interaction-profile similarities, and can be used to functionally annotate genes, associating unknown ones to genes with known functions.⁷

2.3.2 Maximum measurement entropy. For this criterion, we rely on the assumption that an optimal threshold will create

a network with as much information as possible, so that the structure will be particularly informative about the system it represents. One way to measure the amount of information carried by the network is the Shannon entropy,⁵² which can be expressed as:

$$H(X) = - \sum_{i=1}^N P(x_i) \log_b P(x_i) \quad (1)$$

where b is the basis of the used logarithm, $P(x)$ represents the probability mass function of x , and N is the number of possible values of x_i .

One way to implement such a measure for a network is to calculate the entropy of a set of vertex measurements. This can be carried out as follows. Given a set of M measures, defined for each node in the network, an M -dimensional histogram is calculated. This procedure defines a probability mass function $P(x^{(1)}, x^{(2)}, \dots, x^{(M)})$, where $x^{(l)}$ represents the l -th considered measure. Then, the Shannon entropy of this function is calculated. We applied 20 thresholds and verified which value resulted in vertex measures having the highest entropy.

When the threshold used to generate the network is very small, the structure of the network tends to a complete graph, in which all measures have the same value for all nodes. For very large thresholds, the network will present many isolated nodes, and, again, nodes will have similar properties. Therefore, we expect that, for some intermediate threshold values, nodes will present a maximum measure heterogeneity, which should be reflected in the entropy. This concept is illustrated in Fig. 3. The first and third networks (Fig. 3a and c) have low entropy due to the excess and lack of links, respectively. The second network (Fig. 3b) is more informative, reflecting a more complex structure and topology. For the analysis of this criterion, the local transitivity and betweenness centrality were used as node measures.

The maximum measurement entropy can be chosen if the purpose of the research is to analyze the most relevant gene relationships, in terms of maximizing the amount of information about the network structure, while simplifying the network (eliminating redundancies). This property can be used to analyze the information flow along the network, based on the property that genes with comparable mRNA expression profiles have a high chance to be regulated by the same genetic mechanisms.^{35,53–55}

2.3.3 Size of the giant component. This criterion is based on the fact that the higher the chosen threshold to build a

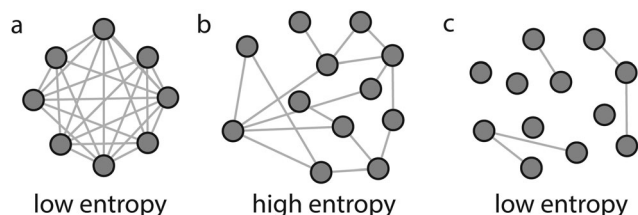


Fig. 3 Networks and their expected entropy. Networks (a and c) have low entropy due to the excess and lack of links, respectively. Network (b) can be understood as holding a more complex structure, reflected in a higher entropy value.

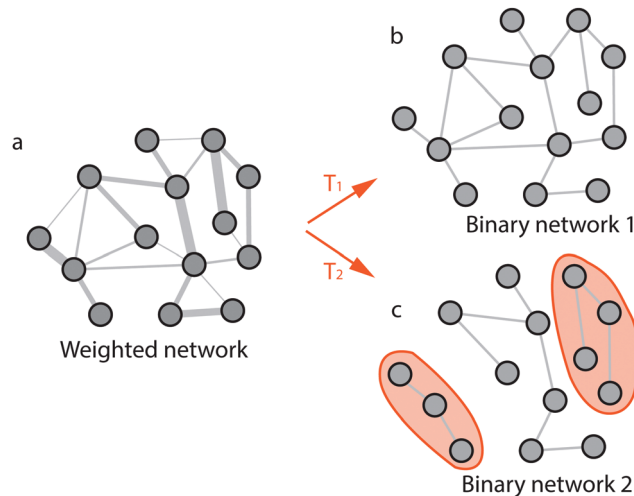


Fig. 4 Weighted network (a) and the result of applying a relatively low threshold (T_1) (b) and a relatively high threshold (T_2) (c). When a sufficiently high threshold is used, the network may break into groups of vertices that are disconnected (highlighted nodes in c).

network, the more sparse the structure will be, since a smaller number of gene relationships will be kept in the final network. As the network becomes more sparse, its nodes will become separated into distinct connected components. The largest of these components (*i.e.*, with the largest number of nodes) is called the giant component of the network. The size of the giant component is strongly influenced by the threshold used for binarizing the connections. Fig. 4 exemplifies this. A weighted network (Fig. 4a) is used for representing the relationship between genes: the edge thickness indicates the strength of the association between each gene pair. After applying a certain threshold (T_1 and T_2 in Fig. 4), some edges will be removed, and the resulting graph becomes unweighted. If the threshold value is low, only the weakest links will be deleted (T_1 resulting in network 1 in Fig. 4b). On the other hand, if the chosen threshold is high, only the strongest edges will be preserved (T_2 resulting in network 2 in Fig. 4c). When the applied threshold is high enough, some edges that connected the groups of vertices may disappear, causing the network to “break up” into groups of vertices that are not connected (highlighted groups in Fig. 4). To analyze this criterion, we have chosen thresholds that generated networks with three sizes of giant component: approximately 40, 60 and 80 percent of the original number of nodes.

The size of the giant component criterion, as well as the maximum variation of topological properties criterion (Section 2.3.1) will suit the purpose of analyzing genes that have highly correlated expression. This criterion can also be used to choose genes to be functionally annotated.⁷ The smaller the chosen size of the giant component is, the higher the average correlation between linked genes will be.

2.3.4 Power-law degree distribution. Another way of choosing a threshold consists of determining in advance a model to which the network structure should adapt. Most biological networks tend to present a scale free structure, *i.e.* their degree distribution is approximately a power-law,^{21,56} $P(k) \propto k^{-\gamma}$, where γ is the

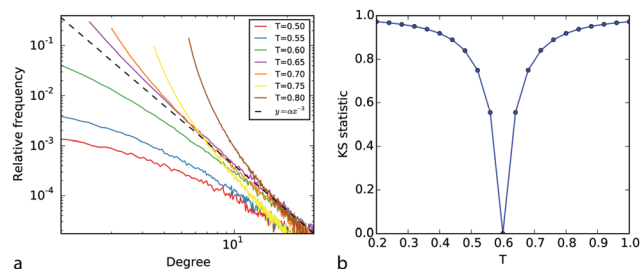


Fig. 5 Plot of the degree distribution of hypothetical networks (a). The black dashed line represents a power-law function. Plot (b) represents the distance that different hypothetical degree distributions have from a power-law distribution. T represents the threshold value.

exponent and the \propto symbol indicates “proportional to”. The γ exponent determines several properties of the system, and the lower the value of γ , the more important the role of hubs becomes.

We adopted a threshold criterion based on the imposition of a power-law degree distribution, that is, the optimal threshold would be the one that would make the distribution of the network degrees the most similar to a power-law.¹⁶ An example of the procedure is shown in Fig. 5. In Fig. 5a we show a plot of the nodes degree distribution for different hypothetical networks generated by different threshold values. The plot on the right illustrates the distance a hypothetical network distribution has from a power-law distribution (Fig. 5b). In order to measure the similarity of the degree distribution to a power-law, we first fit a power-law function to the distribution using the procedure described by Clauset *et al.*⁵⁷ (*i.e.*, a maximum likelihood estimator of the γ exponent). Then, the value of the Kolmogorov–Smirnov test statistic⁵⁸ that the distribution follows the fitted power-law is used as a similarity measure.

The power-law degree distribution criterion is based on co-expression networks modeled as scale-free gene networks,^{21,22} so this method will fit the purpose of achieving a network following, as close as possible, this model.

2.3.5 Specified average degree. This criterion consists of choosing threshold values that generate networks with specified average degrees, as illustrated in Fig. 6. Using this criterion, one can set the average number of interactions a gene will have in the network. We considered two values of average degree, 100 and 1000, and we have chosen the threshold providing the closer average degree to the desired one. This criterion can be used to functionally annotate genes,⁷ as well as assist in the identification of a family of transcripts involved in similar biological processes.^{22,55}

3 Results and discussion

3.1 Application of the threshold criteria

The next sections describe the results obtained for each thresholding criterion. Fig. 7 depicts the distribution of all values of Pearson correlation coefficients (corresponding to each pair of gene expressions, yielding the edge weights) obtained for all original datasets. Also included are the thresholds given by the respective application of the several considered criteria. Observe that in 71.4% of the databases the chosen thresholds

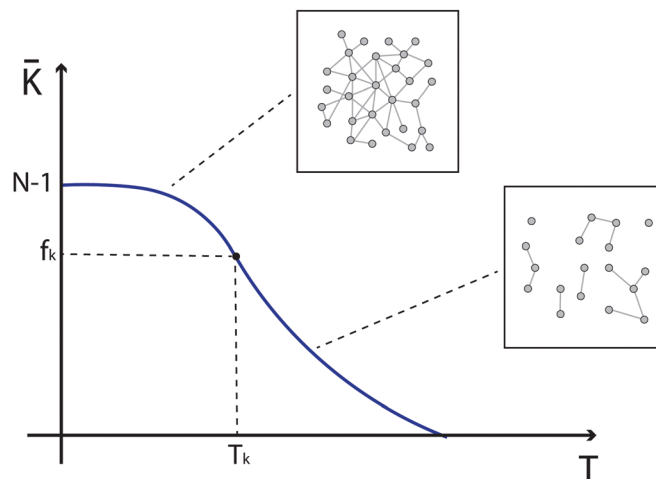


Fig. 6 Average degree for different threshold values. The higher the network's average degree, the more connected the nodes will be. The average degree is represented by \bar{K} . A choice of average degree f_k has a corresponding threshold T_k .

lie between 0.70 and 1.0. Interestingly, most weight distributions in this figure show a similar profile, resembling the right portion of a Gaussian, indicating a predominance of relatively small values, which are being filtered by the proposed criteria as originally desired.

Table S2 (ESI[†]) shows a complementary analysis using the Pearson correlation coefficient, corresponding to the absolute values obtained with respect to strength (sum of the edge weights connected to each node) *versus* degree. The results indicate a significant relationship between strength and degree, which justifies that the thresholding criterion removed the weakest links among genes in most cases. Considering all criteria, it can be verified that the Pearson correlations that are greater than 0.5 (green cells of the table) correspond to 66.5% of the total. Excluding the maximum variation of topological property criteria, the percentage of green cells goes to 71.42%. These percentages are substantially larger than the null hypothesis of 50%, corroborating the fact that weighted relationships tend to be incorporated into the thresholded connections.

3.1.1 Maximum variation of topological properties. Fig. S2 (ESI[†]) shows the variation between consecutive points in the PCA projection of each experiment (the PCA projections are shown in Fig. S1, ESI[†]). There is a stabilizing tendency at the beginning of each trajectory, and a fluctuating and unstable region at the end of the trajectories. The networks at the beginning of these trajectories were generated by lower thresholds (near 0.85), and therefore have a higher concentration of edges, since a lower threshold allows a greater number of levels of gene expression to be classified as correlated.

The detected thresholds defining the transition between stable and unstable regions are indicated by red dots in Fig. S2 (ESI[†]). This point corresponds to the transition of the network from a dense connectivity to a more sparse, broken, one.

3.1.2 Maximum measurement entropy. We applied the Shannon entropy to the local transitivity and betweenness

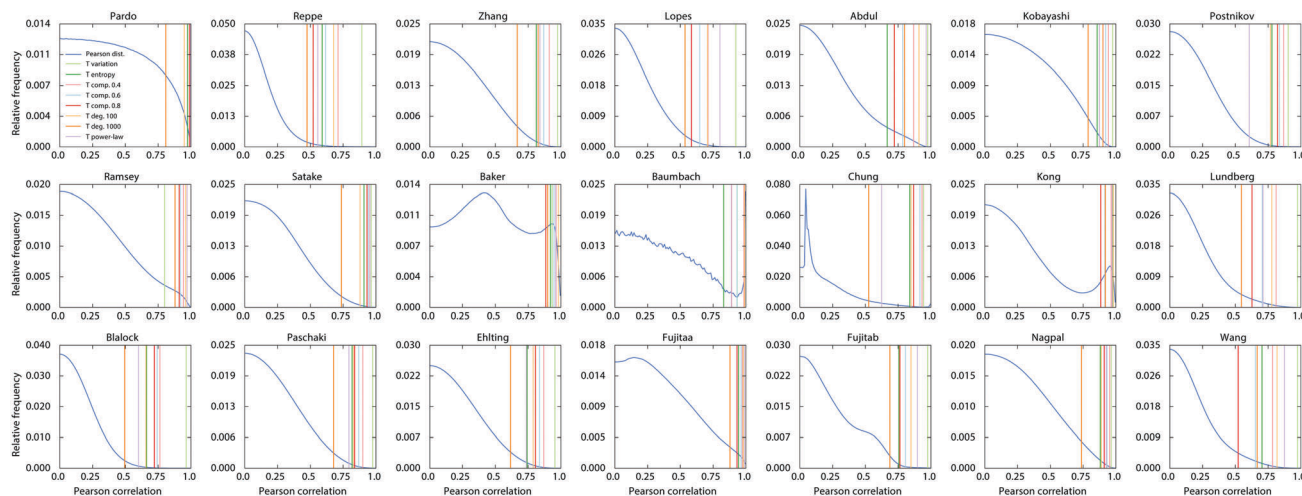


Fig. 7 Distributions of the Pearson correlation coefficients for each dataset (blue line). The vertical lines correspond to the thresholds selected according to each adopted criterion. "Pearson dist." means Pearson correlation distribution.

centrality of the nodes. The result is shown in Fig. S3 (ESI[†]). Red dots indicate the maximum entropy reached by each experiment.

3.1.3 Size of the giant component. Fig. S5 (ESI[†]) shows the size of the giant component as a function of the considered thresholds. Note that the sizes are normalized by the total number of genes of each experiment. The red, green and purple dots indicate networks having giant components comprising around, respectively, 40%, 60% and 80% of the total number of genes. The distance among these highlighted dots may also indicate the resilience of the network structure in the face of threshold variations. Some of them (e.g.: Postnikov,⁵⁹ Ramsey⁶⁰ and Blalock⁶¹) have the giant component broken into smaller components very quickly, while others (Reppe,⁶² Abdul⁶³ and Wang⁶⁴), show a more gradual transition between the highlighted markers.

3.1.4 Power-law degree distribution. Fig. S5 (ESI[†]) shows the Kolmogorov–Smirnov statistics (K–S) of the hypothesis that the degree distribution of the networks follows a power-law function. Red dots indicate the minimum value obtained for each dataset. Note that, in many cases, the K–S statistic shows large fluctuations as the threshold is varied.

3.1.5 Specified average degree. Fig. S6 (ESI[†]) shows the networks' average degrees as a function of threshold. The red and green dots indicate the thresholds for which the average degree is, respectively, 100 and 1000.

3.2 Comparison among methods

Table S3 (ESI[†]) shows 9 relevant properties obtained for the networks generated by each thresholding method. It is clear from Table S3 (ESI[†]) that the N giant of a given dataset varies substantially across the criteria. For some datasets, the thresholding implied keeping most genes in the giant component: the Postnikov dataset⁵⁹ maintained all the genes when thresholded by the specified average degree (1000) and the power-law degree distribution criteria; the Ramsey dataset⁶⁰ also maintained all the genes considering the maximum variation of topological properties criterion; and the Satake dataset⁶⁵ behaved similarly under the specified average degree (100) criterion.

Each criterion is used to select appropriate thresholds for the 21 datasets. Therefore, a useful approach for comparing the values selected by each criterion is to calculate the Pearson correlation coefficient between the obtained thresholds. The result can be represented as a weighted graph, where each node is a criterion and edges indicate the Pearson correlation among pairs of nodes. This graph is shown in Fig. 8a. The size of the nodes is associated with the average Pearson correlation between the node and all other nodes, that is, it indicates how strongly the respective criterion is related to the others. The scatter plots used for calculating these correlation values are shown in Fig. S8, the calculated values are shown in Tables S4–S6 (ESI[†]).

The results indicate that, besides the maximum variation of topological properties, the correlation among all methods is similar. The referred criterion showed low correlation with the other methods, meaning that the networks generated by the maximum variation of topological properties method have topological properties that are highly distinct from the networks generated by the other methods. An example of this divergence can be illustrated by the huge difference in the N giant values for this criterion. One of such cases is the Lundberg dataset,⁶⁶ which goes from 85 (maximum variation of topological properties criteria) to 18 000 nodes (specified average degree (1000) criteria). Also, we noticed that the maximum variation of topological properties criterion has the tendency to yield networks with particularly low N giant values (such as for the Blalock,⁶¹ Paschaki,⁶⁷ Ehling,⁶⁸ and Wang⁶⁴ datasets).

Nevertheless, it is important to notice that the correlation result (Fig. 8) is not straightforward to interpret. In order to explain why, Fig. S9 (ESI[†]) shows two different cases of threshold comparison among the obtained results: entropy *versus* giant component 60% (Fig. S9a, ESI[†]), and power-law *versus* average degree 1000 (Fig. S9b, ESI[†]). The red line indicates the linear least squares fit of the points. The fact that both cases have a high correlation (ρ) does not mean that the chosen threshold values will be the same. For instance, the result

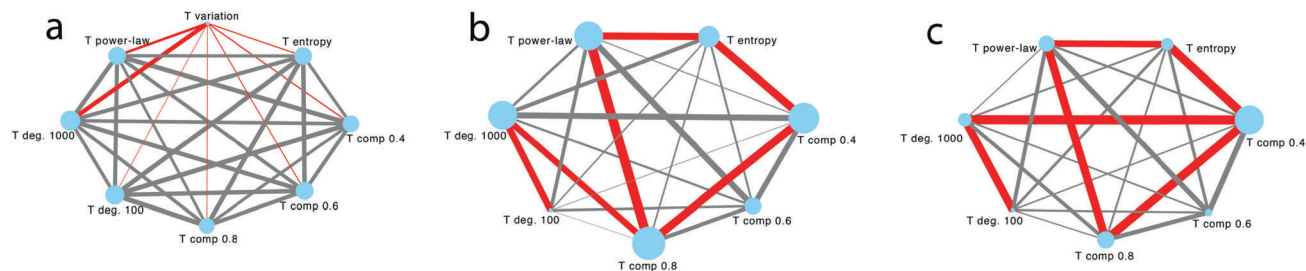


Fig. 8 (a) Weighted graph representing the Pearson correlation among the criterion results. The node size is given by the average of the logarithm of the Pearson correlation between the criterion and all other criteria. Thicker edges correspond to higher correlation values. Weighted graphs representing the obtained (b) $\tilde{\alpha}$ and (c) $\tilde{\beta}$ are also shown. Please refer to the caption of Fig. 9 for the meaning of each node name. Red edges indicate criteria that diverged the most between each other, therefore having either low Pearson correlation or high $\tilde{\alpha}$ or $\tilde{\beta}$. The T variation node was excluded in (b and c) in order to not mask the differences among the other methods.

shown in Fig. S9b (ESI[†]) indicates that thresholds selected by the power-law criterion tend to be higher than those selected by the deg. 1000 criterion, even though they are strongly correlated.

In order to properly compare the selected thresholds, the angular and linear coefficients of the relationships between the methods also need to be taken into account. These coefficients are obtained by applying a linear least squares regression to the data⁶⁹ (see Fig. S7, ESI[†] for the definition). Regarding the angular coefficient, α , when $\alpha \approx 1$ the selected thresholds will tend to vary in a similar manner. Therefore, we define the angular relationship between two criteria as $\tilde{\alpha} = 1 - \alpha$. Small values of $\tilde{\alpha}$ mean that the methods might be related (the actual relationship still depends on the correlation coefficient), while large values indicate the absence of a relationship. Conversely, the absolute value of the linear coefficient, $\tilde{\beta} = |\beta|$, is also related to the similarity between methods, since $\tilde{\beta} \approx 0$ means that the methods might be related. The values of $\tilde{\alpha}$ and $\tilde{\beta}$ can be different depending on the choice of explanatory variable. In other words, applying the linear least squares method to a plot of y as a function of x may give different $\tilde{\alpha}$ and $\tilde{\beta}$ than when applying the same method to a plot of x as a function of y . The former considers x as an explanatory variable, while in the latter the explanatory variable is y . Therefore, for each pair of criteria being compared, we choose the minimum $\tilde{\alpha}$ and $\tilde{\beta}$ obtained when considering the thresholds from each criterion as the explanatory variable.

Fig. 8b and c show a graphical representation of the relationships among the criteria considering, respectively, $\tilde{\alpha}$ and $\tilde{\beta}$. Edges highlighted in red indicate the pairs of methods that diverged the most. In order to better interpret the results, we calculated the difference between the threshold values (using Table S7, ESI[†]) obtained by these highlighted pairs of methods, and the result is shown in Fig. S10 (ESI[†]). We notice that these criteria have selected different threshold values for many datasets, although the difference can widely vary between cases. For instance, the comparison between the maximum measurement entropy criterion and power-law degree distribution criterion resulted in eleven out of twenty-one (11/21) threshold values with a difference of less than 0.05 (Fig. S10b, ESI[†]) (bases from Lopez,⁷⁰ Abdul,⁶³ Kong,⁷¹ Blalock,⁶¹ Wang⁶⁴), and the comparison between the specified average degree (100) criterion

and specified average degree (1000) criterion resulted in all thresholds (21/21) with a difference higher than 0.05 (Fig. S10f, ESI[†]).

The comparison between the maximum measurement entropy and power-law degree distribution criterion (Fig. S10b, ESI[†]) resulted in more similar results for the considered measures, because the threshold values are closer to each other, and the resulting networks will have a more comparable topology as well. This similarity is more conspicuous on measures related to the number of nodes and degree (number of nodes and average degree). However, for the measures related to the connectivity among nodes (clustering coefficient, diameter, modularity, average community size) (Table S3, ESI[†]), we do not notice this similarity, which suggests that these networks, though with a comparable size, have different topology.

Comparing criterion specified average degree 100 with criterion specified average degree 1000, we verify large differences between the resulting networks (Table S3, ESI[†]), even for measures that are not directly related to the average degree, such as the modularity and the clustering coefficient of the networks.

In Appendix A, we apply PCA to analyze properties of the graphs generated by each criterion. The results indicate that some datasets behave similarly, that is, their position in the PCA, relative to other datasets, is similar for distinct criteria. These datasets constitute what we call the conserved group.

3.3 General discussion

Considering the conserved group from Fig. 9, we noticed that for each criteria, some measures had a considerable variation, and others followed a trend. For the T comp 0.6 criterion, the modularity and diameter were the measures that varied less, while the number of nodes from the giant component and community average size had a high variation. For the T deg. 100 criterion, the modularity had a small fluctuation, and number of nodes from the giant component had a high variation, as well as clustering coefficient and diameter. Although the criterion takes into account the average degree of the network, the measure average degree had a significant fluctuation for the databases from Reppe⁶² and Fujita b.⁷² This happens because the threshold was specified according to the expected average degree of the entire database, but the measurements in Table S3

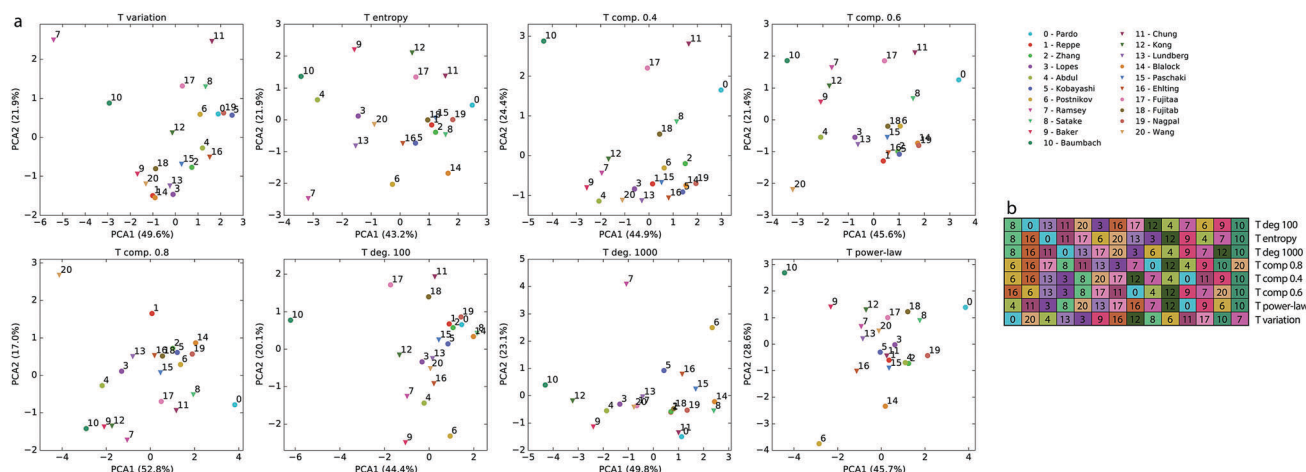


Fig. 9 (a) PCAs of the measures obtained for the considered threshold criteria. Note that each dot corresponds to a respective dataset. *T* variation represents the maximum variation of topological measures; *T* entropy is the maximum measurement entropy criterion; *T* comp. 0.4 means fixing the size of the giant component in 40% of the total number of nodes, with a similar meaning for *T* comp. 0.6 and *T* comp. 0.8; *T* deg. 100 represents specified average degree (100), and *T* deg. 1000 specified average degree (1000); *T* power-law represents the power-law degree distribution criterion. (b) Matrix containing the relative positions of the experiments to the conserved group (closer to farther, from the left to the right); each row corresponds to one criterion.

(ESI[†]) were calculated only for the giant component of the resulting networks. For the *T* deg. 1000 criterion, the average degree had, as expected, smaller variation. However, the standard variation and other measures (number of nodes, clustering coefficient, diameter, modularity, community average size) varied considerably. For the *T* entropy criterion, the clustering coefficient and the modularity were the measures that varied less.

It is interesting to observe that all the biological groups are represented in the conserved group, except the *Drosophila* databases. This may imply that the genes have a similar behavior in these experiments, with analogous interactions among the expressed genes.

The database from Baumbach⁷³ generated the network with measures that diverged the most from the conserved group. That work investigated a biological scenario in which the chromosomes of *Drosophila* flies have a centrosome loss or amplification. Although it would be expected that centrosome anomalies would induce some changes, including transcriptional ones, they found that the centrosome defects caused very few changes in the global transcriptome.⁷³ In the present article, however, the co-expression network generated by the transcriptomic profile of *Drosophila* was found to have different behavior when compared to other co-expression networks (Fig. 9a), which could suggest some type of biological effect.

4 Conclusions

The present article addresses the challenging task of threshold selection for constructing co-expression networks. Although it constitutes a very important step in systems biology, there are few related mathematical studies addressing this problem in a more systematic fashion and proposing possible methodologies. The main purpose of this article was, therefore, to develop a

relatively systematic experimental approach to the problem of threshold selection in co-expression networks, supported by numerical evaluations. We also proposed five different methods for threshold selection, given a dataset. Each of these methods has a respective criterion for selecting a suitable threshold for each dataset in question. The impact on the resulting networks was analyzed by comparing the topology of the obtained networks through several measures with respect to the different threshold selection approaches.

The method of maximum variation of topological properties (Section 2.3.1) generated the networks with the most distinct topological variables. This method resulted in large variations of the number of nodes in the giant component, average degree and its standard deviation, and the community average size. Differently, the method of average degree 1000 (Section 2.3.5) produced networks with a similar structure, naturally showing an average degree around 1000, followed by a clustering coefficient with low variance. The other methods resulted in networks with considerable variation of measurements, especially regarding the average and the standard deviation of the degree.

The database from Baumbach⁷³ seemed not to depend on the method of threshold selection. This dataset showed low variance of topological measures for the different criteria studied in this paper. The databases from the “conserved group”^{61,62,67,72,74–76} also showed relative independence of the method of thresholding. On the other hand, the databases from Lopez,⁷⁰ Zhang,⁷⁷ Abdul,⁶³ Postnikov,⁵⁹ Ramsey,⁶⁰ Satake,⁶⁵ Chung,⁷⁸ Lundberg,⁶⁶ Fujita⁷² and Wang⁶⁴ were more dependant on the applied method.

By taking the above mentioned findings into account, we propose some guidelines that can help researchers to select thresholds for specific networks of interest:

(1) Our results show that, for some types of databases, the chosen threshold can have a great impact on the topological

properties of the resulting networks. So, great attention has to be given to this aspect in each specific application. Contrariwise, good agreement between results obtained by alternative criteria provide some intrinsic support to the obtained network in a similar manner as the boosting concept from artificial intelligence.^{79,80}

(2) Given a specific network of interest, the methodology reported here can be applied while also considering other related databases (e.g. same species, same type of experiment, same tissue/organ, *etc.*). This would allow some comparative context from which to identify how the specific data stand with respect to other similar situations.

(3) In cases where the application of interest involves a critically important aspect or requirement not included in the reported framework, it would be possible to devise a new variation of the methodology aimed at emphasizing such a characteristic in the original data, in a way similar to the regularization approaches adopted in areas such as machine learning and also biology.^{81,82}

It should be observed that more specific guidelines to thresholding gene co-expression networks would require access to complete biological databases, which could be used as gold standards in validation and comparative approaches. Unfortunately, most biological databases, especially those with molecular nature, are incomplete. It is expected that future availability of new, more complete data can lead to complementations and refinements of the above suggested three criteria.

Mathematical and computational techniques have progressively been proposed and used to analyze biological data, especially given the growing complexity of the systems investigated. The kind of approach reported in the current work may shine some light in the direction of achieving better representations of specific biological scenarios. We applied the principles of Complex Network theory in the study of transcriptomic data, in order to better understand the effects of threshold selection in co-expression networks, hopefully encouraging the development of new criteria as well as methods for respective comparison and validation. More immediate advancements of the reported results include, but are not limited to, incorporating other databases and additional topological measurements. The proposed thresholding criteria could also be integrated into the methodology described by Zhang and Horvath.¹⁶

Appendix A – PCA analysis of the datasets

The space defined by the 7 properties in Table S3 (ESI[†]), for each thresholding criterion, was projected into a respective 2-dimensional space by using PCA. The number of node (N giant) measurements and the average degree were not included as they are directly influenced by the threshold in a trivial manner. Fig. 9a shows the obtained PCAs. PCA is frequently used because it allows the visualization of the data distribution, which would be otherwise impossible in higher dimensions. However, PCA is also particularly useful to identify, given several measurements, which of them account for most of the data variation. The

analysis of the weight of each measurement on the PCA shows that the features that mostly contributed to the variation of PCA 1 is the standard deviation of the average degree, followed by modularity and R_1 ; similarly, the measurements with greatest weights in PCA 2 are the diameter, community average size and clustering coefficient.

The clustering coefficient measurements in the Baumbach database⁷³ were very high (>0.98) for all criteria. The narrow variation for this measurement for some datasets (e.g. Lundberg⁶⁶ and Fujita⁷²), is likely due to the fact that the data are strongly clustered, suggesting a co-expression network where the genes highly influence the expression of each other. In contrast, the clustering coefficient varies widely for other databases (e.g. Blalock⁶¹ and Wang⁶⁴), implying a contrary effect. Low resulting modularities were obtained for the Kong⁷¹ (0.08 to 0.32) and Ramsey⁶⁰ (0.20 to 0.34) databases.

Going back to Fig. 9a, the results indicate that some datasets behave similarly, that is, their position in the PCA, relative to other datasets, is similar for distinct criteria. In order to find which datasets have this property, we selected, by visual inspection, the region in each PCA containing the highest concentration of points. Then, we calculated how many times each dataset appeared in the selected regions. The datasets appearing more than four times were: Pardo,⁷⁴ Reppe,⁶² Kobayashi,⁷⁵ Blalock,⁶¹ Paschaki,⁶⁷ Fujita b⁷² and Nagpal.⁷⁶ These datasets belong to what we call the conserved group, since they behave similarly for the different criteria. Next, we found, for each criterion, the distance between each dataset and the conserved group. This was calculated to be the minimum euclidean distance, in the PCA space, between datasets in the conserved group and the dataset of interest. We ranked the distances for datasets not included in the conserved group, and build a matrix containing the results, which is shown in Fig. 9b. In this matrix, each row corresponds to a criterion. For a given row, the first column contains the index of the dataset that is closest to the conserved group, the second row contains the index for the second closest dataset, and so on.

The obtained matrix shows that the database that was generally closer to the conserved group was Satake,⁶⁵ followed by Ehling.⁶⁸ The Postnikov database⁵⁹ was also usually close to the conserved group, and the Baumbach database⁷³ often resulted apart from this group.

Acknowledgements

C. M. V. Couto thanks FAPESP (Grant No. 2013/19082-7) for financial support. C. H. Comin thanks FAPESP (Grant No. 15/18942-8) for financial support. L. da F. Costa thanks CNPq (Grant no. 307333/2013-2) for support. This work was supported also by FAPESP grant 11/50761-2.

References

- 1 G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider and P. G. Bagos, *BioData Min.*, 2011, **4**, 1–27.

- 2 M. Newman, *Networks: an introduction*, Oxford University Press, Inc., New York, 1st edn, 2010.
- 3 L. D. F. Costa, F. A. Rodrigues and A. S. Cristino, *Genet. Mol. Biol.*, 2008, **31**, 591–601.
- 4 A. L. Barabási, N. Gulbahce and J. Loscalzo, *Nat. Rev. Genet.*, 2011, **12**, 56–68.
- 5 J. M. Stuart, E. Segal, D. Koller and S. K. Kim, *Science*, 2003, **302**, 249–255.
- 6 L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antigueira, M. P. Viana and L. E. Correa Rocha, *Adv. Phys.*, 2011, **60**, 329–412.
- 7 J. I. F. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers and A. J. Walhout, *Nat. Methods*, 2013, **10**, 1169–1176.
- 8 K. Pearson, *Proc. R. Soc. London*, 1895, **58**, 240–242.
- 9 R. A. Fisher, *Biometrika*, 1915, **10**, 507–521.
- 10 S. Sundarajan and M. Arumugam, *Gene*, 2016, **593**, 225–234.
- 11 H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. Sousa, M. Pletikos, K. A. Meyer and G. Sedmak, *et al.*, *Nature*, 2011, **478**, 483–489.
- 12 A. Torkamani, B. Dean, N. J. Schork and E. A. Thomas, *Genome Res.*, 2010, **20**, 403–412.
- 13 R. R. Nayak, M. Kearns, R. S. Spielman and V. G. Cheung, *Genome Res.*, 2009, **19**, 1953–1962.
- 14 L. L. Elo, H. Järvenpää, M. Orešič, R. Lahesmaa and T. Aittokallio, *Bioinformatics*, 2007, **23**, 2096–2103.
- 15 T. F. Fuller, A. Ghazalpour, J. E. Aten, T. A. Drake, A. J. Lusis and S. Horvath, *Mamm. Genome*, 2007, **18**, 463–472.
- 16 B. Zhang and S. Horvath, *et al.*, *Stat. Appl. Genet. Mol. Biol.*, 2005, **4**, 1128.
- 17 I. K. Jordan, L. Mariño-Ramrez, Y. I. Wolf and E. V. Koonin, *Mol. Biol. Evol.*, 2004, **21**, 2058–2070.
- 18 M. M. Babu, *Comput. Genomics*, 2004, 225–249.
- 19 R. Liu, Y. Cheng, J. Yu, Q.-L. Lv and H.-H. Zhou, *Gene*, 2015, **563**, 56–62.
- 20 X. Zheng, C. Xue, G. Luo, Y. Hu, W. Luo and X. Sun, *Cancer Gene Ther.*, 2015, **22**, 238–245.
- 21 A.-L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
- 22 K. K. Harrall, K. J. Kechris, B. Tabakoff, P. L. Hoffman, L. M. Hines, H. Tsukamoto, M. Pravenec, M. Printz and L. M. Saba, *Mamm. Genome*, 2016, **27**, 469–484.
- 23 S. Bergmann, J. Ihmels and N. Barkai, *PLoS Biol.*, 2003, **2**, e9.
- 24 H. Jeong, S. P. Mason, A.-L. Barabási and Z. N. Oltvai, *Nature*, 2001, **411**, 41–42.
- 25 J. F. Donges, Y. Zou, N. Marwan and J. Kurths, *Eur. Phys. J.: Spec. Top.*, 2009, **174**, 157–179.
- 26 M. Barthélemy, *Phys. Rep.*, 2011, **499**, 1–101.
- 27 S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Cambridge University Press, 1994, vol. 8.
- 28 F. Iorio, M. Bernardo-Faura, A. Gobbi, T. Cokelaer, G. Jurman and J. Saez-Rodriguez, *BMC Bioinf.*, 2016, **17**, 542.
- 29 M. Sarkar and A. Majumder, *Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015*, 2016, pp. 79–93.
- 30 F. Y. Yu, Z. H. Yang, N. Tang, H. F. Lin, J. Wang and Z. W. Yang, *BMC Syst. Biol.*, 2014, **8**, S4.
- 31 Z. Shi, C. K. Derow and B. Zhang, *BMC Syst. Biol.*, 2010, **4**, 74.
- 32 A. M. Yip and S. Horvath, *BMC Bioinf.*, 2007, **8**, 22.
- 33 B. H. Voy, J. A. Scharff, A. D. Perkins, A. M. Saxton, B. Borate, E. J. Chesler, L. K. Branstetter and M. A. Langston, *PLoS Comput. Biol.*, 2006, **2**, e89.
- 34 J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski and J. Kertesz, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2007, **75**, 027105.
- 35 P. Michalak, *Genomics*, 2008, **91**, 243–248.
- 36 R. Edgar, M. Domrachev and A. E. Lash, *Nucleic Acids Res.*, 2002, **30**, 207–210.
- 37 L. d. F. Costa, F. A. Rodrigues, G. Travieso and P. R. Villas Boas, *Adv. Phys.*, 2007, **56**, 167–242.
- 38 N. Biggs, *Algebraic graph theory*, Cambridge University Press, 1993.
- 39 L. C. Freeman, *Sociometry*, 1977, 35–41.
- 40 A. M. M. González, B. Dalsgaard and J. M. Olesen, *Ecol. Complex*, 2010, **7**, 36–43.
- 41 P. W. Holland and S. Leinhardt, *Comparative Group Studies*, 1971, 107–124.
- 42 D. J. Watts and S. H. Strogatz, *Nature*, 1998, **393**, 440–442.
- 43 M. E. Newman, *Phys. Rev. Lett.*, 2002, **89**, 208701.
- 44 G. Mao and N. Zhang, *J. Appl. Math.*, 2013, 1–5.
- 45 R. Albert and A.-L. Barabási, *Rev. Mod. Phys.*, 2002, **74**, 47.
- 46 D. B. West, *et al.*, *Introduction to graph theory*, Prentice hall Upper Saddle River, 2001, vol. 2.
- 47 S. Zhou and R. J. Mondragón, *Communications Letters, IEEE*, 2004, **8**, 180–182.
- 48 L. Lovász and M. D. Plummer, *Matching theory*, American Mathematical Soc., 2009, vol. 367.
- 49 I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.
- 50 S. N. Dorogovtsev, A. V. Goltsev and J. F. Mendes, *Rev. Mod. Phys.*, 2008, **80**, 1275.
- 51 R. Guimera, M. Sales-Pardo and L. A. N. Amaral, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2004, **70**, 025101.
- 52 R. Balian, *Poincaré Seminar 2003*, 2004, pp. 119–144.
- 53 R. B. Altman and S. Raychaudhuri, *Curr. Opin. Struct. Biol.*, 2001, **11**, 340–347.
- 54 A. Schulze and J. Downward, *Nat. Cell Biol.*, 2001, **3**, E190–E195.
- 55 D. J. Alocco, I. S. Kohane and A. J. Butte, *BMC Bioinf.*, 2004, **5**, 1.
- 56 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási, *Nature*, 2000, **407**, 651–654.
- 57 A. Clauset, C. R. Shalizi and M. E. Newman, *SIAM Rev.*, 2009, **51**, 661–703.
- 58 F. James, *Statistical methods in experimental physics*, World Scientific, 2006.
- 59 Y. V. Postnikov, T. Furusawa, D. C. Haines, V. M. Factor and M. Bustin, *Mol. Cancer Res.*, 2014, **12**, 82–90.
- 60 J. E. Ramsey and J. D. Fontes, *Mol. Immunol.*, 2013, **56**, 768–780.
- 61 E. M. Blalock, H. M. Buechel, J. Popovic, J. W. Geddes and P. W. Landfield, *J. Chem. Neuroanat.*, 2011, **42**, 118–126.
- 62 S. Reppe, D. Sachse, O. K. Olstad, V. T. Gautvik, H. K. Sanderson, P. Datta, J. P. Berg and K. M. Gautvik, *Bone*, 2013, **53**, 69–78.

- 63 A. M. Abdul-Nabi, E. R. Yassin, N. Varghese, H. Deshmukh and N. R. Yaseen, *PLoS One*, 2010, **5**, e12464.
- 64 H. Wang, L. Waller, S. Tripathy, S. K. St Martin, L. Zhou, K. Krampis, D. M. Tucker, Y. Mao, I. Hoeschele and S. Maroof, *et al.*, *Plant Genome*, 2010, **3**, 23–40.
- 65 H. Satake, K. Tamura, M. Furihata, T. Anchi, H. Sakoda, C. Kawada, T. Iiyama, S. Ashida and T. Shuin, *Oncol. Rep.*, 2010, **23**, 11–16.
- 66 L. E. Lundberg, M. L. Figueiredo, P. Stenberg and J. Larsson, *Nucleic Acids Res.*, 2012, **40**, 5926–5937.
- 67 M. Paschaki, C. Schneider, M. Rhinn, C. Thibault-Carpentier, D. Dembélé, K. Niederreither and P. Dollé, *PLoS One*, 2013, **8**, e62274.
- 68 J. Ehlting, N. Mattheus, D. S. Aeschliman, E. Li, B. Hamberger, I. F. Cullis, J. Zhuang, M. Kaneda, S. D. Mansfield and L. Samuels, *et al.*, *Plant J.*, 2005, **42**, 618–640.
- 69 D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to linear regression analysis*, John Wiley & Sons, 2015.
- 70 L. López-Corral, L. A. Corchete, M. E. Sarasquete, M. V. Mateos, R. Garca-Sanz, E. Ferminán, J. J. Lahuerta, J. Bladé, A. Oriol and A. I. Teruel, *et al.*, *Haematologica*, 2014, 1–21.
- 71 E. C. Kong, L. Allouche, P. A. Chapot, K. Vranizan, M. S. Moore, U. Heberlein and F. W. Wolf, *Alcohol.: Clin. Exp. Res.*, 2010, **34**, 302–316.
- 72 M. Fujita, Y. Horiuchi, Y. Ueda, Y. Mizuta, T. Kubo, K. Yano, S. Yamaki, K. Tsuda, T. Nagata and M. Niihama, *et al.*, *Plant Cell Physiol.*, 2010, **51**, 2060–2081.
- 73 J. Baumbach, M. P. Levesque and J. W. Raff, *Biol. Open*, 2012, BIO20122238.
- 74 S. J. Pardo, M. J. Patel, M. C. Sykes, M. O. Platt, N. L. Boyd, G. P. Sorescu, M. Xu, J. J. W. A. van Loon, M. D. Wang and H. Jo, *Am. J. Physiol.: Cell Physiol.*, 2005, **288**, C1211–C1221.
- 75 G. S. Kobayashi, L. Alvizi, D. Y. Sunaga, P. Francis-West, A. Kuta, B. V. P. Almada, S. G. Ferreira, L. C. de Andrade-Lima, D. F. Bueno and C. E. Raposo-Amaral, *et al.*, *PLoS One*, 2013, **8**, e65677.
- 76 P. Nagpal, C. M. Ellis, H. Weber, S. E. Ploense, L. S. Barkawi, T. J. Guilfoyle, G. Hagen, J. M. Alonso, J. D. Cohen and E. E. Farmer, *et al.*, *Development*, 2005, **132**, 4107–4118.
- 77 R. Zhang, H. Fang, Y. Chen, J. Shen, H. Lu, C. Zeng, J. Ren, H. Zeng, Z. Li, D. Cai and Q. Zhao, *PLoS One*, 2012, **7**, e32356.
- 78 Y.-S. A. Chung and C. Kocks, *Fly*, 2012, **6**, 21–25.
- 79 Z.-H. Zhou, *Ensemble methods: foundations and algorithms*, CRC Press, 2012.
- 80 L. Breiman, *et al.*, *Annals of Statistics*, 1998, **26**, 801–849.
- 81 M. Veltman, *et al.*, *Nucl. Phys. B*, 1972, **44**, 189–213.
- 82 A. Gábor and J. R. Banga, *BMC Syst. Biol.*, 2015, **9**, 74.