A Network-based Approach for Inferring Thresholds in Co-expression Networks

Nicolás López-Rozo(✉), Miguel Romero, Jorge Finke, and Camilo Rocha

Department of Electronics and Computer Science, Pontificia Universidad Javeriana, Cali, Colombia
nicolaslopez@javerianacali.edu.co

Abstract. Gene co-expression networks (GCNs) specify binary relationships between genes and are of biological interest because significant network relationships suggest that two co-expressed genes rise and fall together across different cellular conditions. GCNs are built by (i) calculating a co-expression measure between each pair of genes and (ii) selecting a significance threshold to remove spurious relationships among genes. This paper introduces a threshold criterion based on the underlying topology of the network. More specifically, the criterion considers both the rate at which isolated nodes are added to the network and the density of its components when the threshold varies. In addition to Pearson's correlation measure, the biweight midcorrelation, the distance correlation, and the maximal information coefficient are used to build different GCNs from the same data and showcase the advantages of the proposed approach. Finally, a case study presents a comparison of the predictive performance of the different networks when trying to predict gene functional annotations using hierarchical multi-label classification.

Keywords: Gene co-expression network · Hierarchical multi-label classification · Gene function prediction · Network density · Correlation metrics.

1 Introduction

A gene co-expression network (GCN) of an organism specifies binary relationships between genes that are likely controlled by the same transcriptional regulatory program, are functionally related, or are members of the same pathway or protein complex. They are of biological interest because significant co-expression relationships show a similar expression pattern across different experimental conditions, meaning that two co-expressed genes rise and fall together across different samples. GCNs are usually built in a two-step approach: first, by calculating a co-expression measure between each pair of genes, and second, by selecting a significance threshold to remove spurious relationships between genes. In practice, a correlation metric between any two genes is unlikely to be zero (even if their expression behavior is completely independent), thus resulting in a GCN

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 265–276, 2023.
https://doi.org/10.1007/978-3-031-21127-0_22

that is a complete weighted graph with some spurious edges. However, dealing with a dense network can render its computational analysis intractable. The challenge is thus to use a correlation metric that identifies significant relationships and makes sense from a biological perspective, while making the resulting GCN accessible and tractable for computational analysis.

This paper proposes a new hard-threshold criterion for discarding relationships from gene correlation networks. The main novelty of the criterion is that it considers the underlying topology of the network, more specifically, the rate at which isolated nodes are connected to the network and the density of the network components as the threshold varies. Unlike other authors that define several threshold ranges based on network density minima and the number of edges and nodes [1,19], the proposed approach selects between two specific values of the selected topological properties of the network.

The Pearson and Spearman correlation measures are commonly used for finding correlations among genes [18]. However, these correlation coefficients are designed to identify linear or monotonic relationships, respectively, and must therefore be interpreted carefully. Moreover, Spearman correlation has been found to achieve a lower accuracy than Pearson when analyzing continuous data [6]. In this paper, different correlation metrics are evaluated to showcase the proposed approach. In particular, the biweight midcorrelation (BICOR) [8], the distance correlation (DCORR), and the maximal information coefficient (MIC) are used to build and compare different GCNs from the same data.

A case study is presented, which analyses the performance of hierarchical multi-label classification (HMC) for predicting gene functions based on different generated GCNs. The analysis focuses on the biological processes of the Gene Ontology (GO) hierarchy [5]. Functional information is imported from the DAVID Bioinformatics Resources [7]. The experimental results suggest that, although performance of the functional prediction for all metrics is similar, an improvement greater than the variance can be observed for large hierarchies using BICOR to construct the network. Ultimately, these results highlight the importance of characterizing the relationships among genes in elucidating biological functionality of organisms and suggest how the proposed thresholding criterion can be extended to other biological contexts.

The remainder of the paper is organized as follows. Section 2 presents preliminary concepts. Section 3 presents the proposed threshold criterion and shows its impact on building a gene co-expression network from a dataset of rice. Section 4 presents a case study in gene function prediction based on a hierarchical multi-label classification model and compares the performance of the resulting co-expression networks. Section 5 summarizes related work and draws some conclusions.

2 Preliminaries

2.1 Correlation Metrics

Correlation metrics are used for measuring the degree of association (i.e., linear relationship) between two random variables. The most common correlation metric is perhaps the Pearson correlation coefficient (PCC). It is computed as the covariance ratio of two variables and the product of their standard deviations. Given two variables X and Y, the PCC is defined as

$$\mathrm{PCC}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where n is the number of points for each variable, x_i refers to the i-th data point, and $\bar{x}$ is the sample mean of X. The definition is analogous for y_i and $\bar{y}$.

The biweight midcorrelation (BICOR), an alternative to PCC, is a measure of similarity between samples that relies on the median rather than the mean for sample centrality [8]. For two random variables X and Y, BICOR is defined as

$$\mathrm{BICOR}(X,Y) = \sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i,$$

where $\tilde{x}_i$ and $\tilde{y}_i$ are normalized versions of the original values x_i and y_i, which in turn, are defined as

$$\tilde{x}_i = \frac{(x_i - \mathrm{med}(x))w_i^{(x)}}{\sqrt{\sum_{j=1}^{n}[(x_j - \mathrm{med}(x))w_j^{(x)}]^2}},$$

where $\mathrm{med}(x)$ corresponds to the median of x and $w_i^{(x)}$ is a weight assigned to each value x_i based on its deviation from the median. The definition of $\tilde{y}_i$ and $w_i^{(y)}$ are analogous (more details on computing these weights can be found in [15]).

Distance correlation (DCORR) is a correlation metric that provides a better performance than PCC regarding complex relationships and outlier effect. In contrast to PCC, DCORR can measure non-linear relationships and is less affected by outliers [6]. Furthermore, the DCORR between a pair of variables is equal to zero if and only if they are independent. The metric does not assume that data is normally distributed and results from the distance covariance between X and Y, DCov^2. In particular, the distance covariance is defined as

$$\mathrm{DCov}^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n}[A_{ij}B_{ij}],$$

where A_{ij} and B_{ij} are defined as $A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$ and $B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}$, given that $a_{ij} = \|x_i - x_j\|$ and $b_{ij} = \|y_i - y_j\|$ are Euclidian norms. The

DCORR between X and Y is defined as

$$\mathrm{DCORR}(X,Y) = \frac{\mathrm{DCov}^2(X,Y)}{\sqrt{\mathrm{DCov}^2(X,X)\mathrm{DCov}^2(Y,Y)}}.$$

Other type of metrics that can be used to assess correlation have their origins in information theory. They are based on the concept of mutual information of two variables to measure mutual dependence between them. The maximal information coefficient (MIC) quantifies the degree in which a discrete variable can give information about another variable. It is based on computing a grid and observing how well the grid encapsulates the relationship between the two variables [12]. MIC has been applied in the context of gene co-expression network analysis. It has been shown to outperform other models (such as WGCNA [8]) on finding non-linear relationships [11].

2.2 Gene Co-expression Networks

Gene co-expression networks are represented as undirected graphs where each node corresponds to a gene and a pair of nodes is connected with an edge if there exists a significant co-expression relationship between them.

Definition 1. Let V be a set of genes, E a set of edges that connect pairs of genes, and $w : E \to \mathbb{R}_{\geq 0}$ a weight function. A *(weighted) gene co-expression network* is a weighted graph $G = (V, E, w)$.

Gene co-expression networks are of biological interest since co-expressed genes usually belong to the same regulatory pathway or protein complex and can elucidate biological functionality for unlabeled genes [8]. High-throughput gene expression profiling technologies such as microarrays or RNA-Seq yield datasets that can be used for generating co-expression networks [13]. If the expression profiles for several genes under different experimental conditions are measured, then the network can be constructed by connecting with an edge those pairs of genes showing similarity in their expression patterns.

For the purpose of this paper, a hierarchical multi-label classification model is used to fairly evaluate the performance of the different correlation metrics and their resulting GCNs in predicting gene functional annotations.

2.3 Hierarchical Classification

Classification problems may be defined using binary, multi-class, or multi-label prediction tasks, where predictions consist of a single class, a single class from a set of mutually exclusive classes, and a subset of classes, respectively. Hierarchical multi-label classification (HMC) addresses the task of structured output prediction where the classes are organized into a hierarchy and an instance may belong to multiple classes. In many problems, such as gene function prediction, classes inherently satisfy these conditions [9]. The problem of predicting gene

functions refers to the task of identifying associations between genes and functions based on biological information, such as gene co-expression networks. This problem is usually addressed using HMC.

Moreover, it is often the case in HMC datasets that individual classes have few positive instances. In gene function prediction, typically only a few genes are associated to specific functions. This implies that for most classes, the number of negative instances by far exceeds the number of positive instances. Hence, the focus is on recognizing the positive instances, i.e., on identifying the associations between genes and functions. For this reason, the area under the average precision-recall (PR) curve introduced by [16] is used for evaluation, denoted as AU($\overline{\text{PRC}}$). This metric transforms the multi-label problem into a binary one by computing the precision and recall for all functions A' together. This corresponds to micro-averaging the precision and recall. Multiple thresholds are used to create a PR curve, where each point represents the precision and recall for a give threshold that can be computed as:

$$\overline{\text{Prec}} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \quad \text{and} \quad \overline{\text{Rec}} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}.$$

3 Co-expression Network Construction

This section introduces the method for constructing co-expression networks based on the expression profiles of genes.

According to [14], two possible approaches are feasible to filter spurious edges, namely, *hard-* and *soft-thresholding*. In the former, a threshold value that 'cuts' the network is defined to discard the edges whose weight falls below. In the latter, a function is designed to map each correlation value to an edge weight, suppressing the influence of weak edges and enhancing the influence of strong ones. Here, a hard-thresholding approach is defined based on the observed co-expression values, which considers the topology of the network.

3.1 Relationship between Threshold and Network Density

Defining the threshold to remove spurious relationships is not a trivial task [4]: setting the threshold too high results in many disconnected small network components; setting it too low introduces noise in the topology of the network. Furthermore, it is impossible to define a universal correlation coefficient threshold because underlying biochemical processes change from one organism to another, among different tissues of the same organism, or even among different experimental designs [18].

Some network properties are relevant for analyzing the network growth and should be introduced for further discussions. The average degree of a network represents the average number of connections of any node in a network. A normalized version of this metric represents a measure of network *density* (which ranges from 0 to 1). Note that a network with no edges has density of 0, while

a fully connected clique has a density of 1. Since a network may be composed of several connected components, consider the measure of *average density* $\bar{\rho}$ as the average density of its connected components (an isolated node is a component with $\bar{\rho} = 1$; similarly, a network consisting of cliques also has $\bar{\rho} = 1$). The work in [1,17,19] provides an important insight on how the network behaves with respect to changes in the threshold. In the proposed method, the number of nodes with at least 1 edge, the number of edges, and the average density of the network are relevant properties.

Consider a threshold that is just a fraction higher than the highest co-expression value across the entire network. At this starting point, the network will consist only of isolated nodes and therefore $\bar{\rho} = 1$. As the threshold starts to decrease, some edges are sparsely added to the network and, up to a certain point, tree-like structures start to emerge. Trees are graphs with low density and therefore the average density of the network decreases. As the threshold continues to decrease, more connections are expected to appear between nodes already connected than between disconnected nodes. Further decrease in the threshold increases the average network density until all edges are covered and the network becomes a clique.

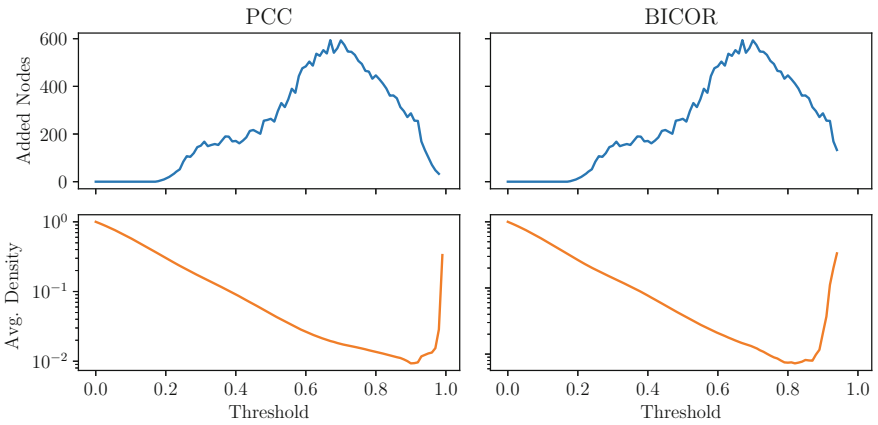

Fig. 1. Example of the behavior of the number of added nodes and average density for PCC and BICOR.

Figure 1 depicts the aforementioned behavior by setting different thresholds to build a co-expression network for rice. The co-expression values are computed using PCC and BICOR (only nodes with a least one edge are considered). In this case, the behavior does not depend on the correlation metric; nevertheless, the dominance of tree-like structures or the densification process depends completely on the underlying correlation values and the intrinsic biological processes involved.

3.2 Threshold Inference

Note that there exists a threshold for which the average network density is minimal (this is, mainly, because with a high threshold the average network density is maximal and equal to 1, and with a low one the resulting network is a clique). This observation is key because it means a transition from a behavior dominated by tree-like structures to one dominated by densification of the existing components. From a biological perspective, at this point, the most intrinsically relevant clusters should have been already established.

Additionally, an important aspect on the network growth is how fast isolated nodes become connected to other components, which contributes to the emergence of the tree-like structures. If the rate at which nodes are being added is growing, network growth is yet unstable and clusters are still forming. In contrast, if the rate of node aggregation is decreasing, then weaker connections are appearing among the existing clusters. In this regard, the turning point of node aggregation is also an important clue to select the threshold, since the most relevant clusters are already formed.

The process of inferring the threshold assumes as input an expression matrix $\mathbf{E}$ containing e expression values for n genes and a correlation metric m. The expression profile for each gene consists of e experimental levels of expression under certain conditions (i.e., $|\mathbf{E}| = n \times e$). The correlation metric m is used to compute the co-expression matrix $\mathbf{C}$ containing the pairwise correlation values between n genes (i.e., $|\mathbf{C}| = n \times n$). The proposed definition of the threshold $t_{\mathbf{C}}$ is given by

$$t_{\mathbf{C}} = \min\left(\text{argmax}(|\Delta V_{I,\mathbf{C}}|), \text{argmin}(\bar{\rho}_{\mathbf{C}})\right),$$

where $\Delta V_{I,\mathbf{C}}$ is the change in the number of isolated nodes as the threshold on the correlation values in $\mathbf{C}$ moves down and $\bar{\rho}_{\mathbf{C}}$ is the average network density.

3.3 Co-expression in Rice

This sections evaluates the selected metrics for computing the co-expression between genes and describes the resulting networks after applying the proposed approach. In each experiment, a discretization of the range of each metric in steps of 0.01 is applied, as in [1].

Data. Expression values for 23,374 genes of rice (*Oryza sativa Japonica*) across 2,678 accessions are retrieved from the NCBI Gene Expression Omnibus (GEO) Datasets [2]. Expression values range from 0.009 up to 280,584, with an average expression value of 1,252. Additionally, the original gene identifiers (AffyID) are converted to Entrez-IDs using the DAVID bioinformatics resources [7].

Analysis. Table 1 presents the main properties of the resulting networks after computing the correlation of rice data using each metric and applying the proposed method for finding the threshold. The column labeled *Nodes* refers to the number of nodes with at least 1 edge.

Table 1. Properties of the resulting networks after applying the proposed threshold method.

Network	Nodes	Edges	Components	Avg. degree	Density
PCC	11,241	1,195,905	473	212.78	0.0189
BICOR	7,344	382,202	586	104.09	0.0142
DCORR	13,069	2,184,554	12	334.31	0.0256
MIC	14,233	5,790,949	255	813.74	0.0572

It can be observed that the correlation metric applied to the expression profiles highly influences the resulting network topology, hereby represented with the number of non-isolated nodes, number of edges, number of connected components, average degree, and density. Also note that the amount of nodes and edges greatly changes from metric to metric. Although DCORR and MIC differ on the number of nodes by around one thousand, the number of edges and the network density of MIC increases by more than twice. In contrast, BICOR has around half of the nodes as DCORR, but less than a fifth of the edges. Finally, note that there exists a great variation in the number of connected components: DCORR is not as dense as MIC, but has less components, connected in a sparse manner.

4 Case Study: Gene Function Prediction

This section presents a comparison of the predictive performance of the generated networks on the gene function prediction problem using HMC. This case study is focused on the biological processes of the Gene Ontology (GO) hierarchy [5]. The functional information is imported from DAVID Bioinformatics Resources [7]. It comprises 3,531 biological processes and 6,367 hierarchical relations, which are part of the GO hierarchy, and a total of 289,407 associations between genes and functions.

To fairly compare the generated networks, it is necessary to use the same functions and genes for all co-expression networks. For that purpose, (i) the greatest connected component (GCC) of each network is computed, (ii) the common genes between the GCCs of all networks are selected, and (iii) a subset of GO sub-hierarchies is selected such that at least 10 genes are associated to each biological process. As a result, 10 sub-hierarchies of biological processes are used, represented by their root. Table 2 depicts the sub-hierarchies sorted by their number of biological processes.

The intersection of all GCCs is computed and extracted, so that the same set of genes is used to train the HMC model for each dataset. The HMC models are built based on random forests of decision trees, where all functions of the hierarchy are considered at once. Additionally, k-fold cross validation is used to avoid overfitting in training (the number of folds is $k = 5$). The parameter values used for random forest classifiers are: 200 estimators (*n_estimators*) and

Table 2. Sub-hierarchies of biological processes, represented by their root. Each sub-hierarchy is considered an independent dataset.

Root	Description	Functions
GO:0002376	Immune system process	9
GO:0044419	Biological process involved in interspecies interaction between organisms	13
GO:0032501	Multicellular organismal process	18
GO:0022414	Reproductive process	24
GO:0032502	Developmental process	53
GO:0051179	Localization	86
GO:0050896	Response to stimulus	102
GO:0065007	Biological regulation	184
GO:0008152	Metabolic process	399
GO:0009987	Cellular process	517

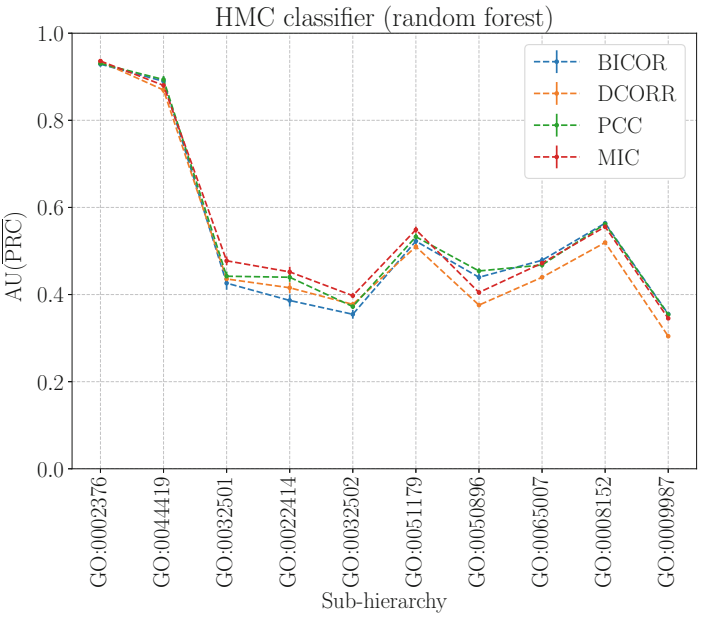

Fig. 2. Prediction performance for the gene function prediction problem on rice, measured with $\overline{\text{AU(PRC)}}$, for the networks generated with all correlation metrics, namely, PCC, BICOR, DCORR and MIC.

minimum number of samples of 5 (*min_samples_split*). Note that one model is build for each network and dataset.

Figure 2 shows the prediction performance measured with $\overline{\text{AU(PRC)}}$ for the networks generated with all correlation metrics, namely, PCC, BICOR, DCORR, and MIC. It illustrates the variance of the performance for 50 runs of the models. Although the performance for all metrics is similar, MIC outperforms the

other metrics (considering the variance) in the smaller sub-hierarchies, with the exception of GO:0044419. In a similar way, BICOR outperforms the other metrics for the larger hierarchies, i.e, GO:0065007, GO:0008152 and GO:0009987. The remaining hierarchies have inconclusive results.

5 Related Work and Concluding Remarks

A significant effort has been put in comparing different correlation metrics on biological datasets, considering that PCC is used in the majority of co-expression network analysis [18]. BICOR itself appears as part of the WGCNA package of R for gene co-expression analysis and has proved to be robust to outliers. The work in [15] compared BICOR and some mutual information estimators like MIC for Gene Ontology (GO) enrichment analysis. The results suggest that MIC tends to overfit the data from brain cancer, blood lymphocyte, yeast, mouse adipose tissue, and mouse muscle. In fact, the Topological Overlap Matrix method, based on BICOR, dominates 7 of 8 datasets in terms of GO enrichment [10]. Similar to the results in [15], our own findings suggest that the benefits of applying BICOR instead of PCC or Spearman coefficient are marginal.

The recent work in [6] compared the efficacy of using DCORR, PCC, MIC, and Spearman as part of the WGCNA framework for gene co-expression network analysis. Four datasets of expression samples including macrophage, liver, cervical cancer, and pancreatic cancer were used to compared the performance of these metrics on GO enrichment of biological processes. The results show that DCORR is stable for highly correlated pairs of genes as the size of the datasets grows, while PCC only presents a comparable, yet decaying stability in the macrophage dataset. Furthermore, when analyzing the module preservation of GO clusters, DCORR has higher one-to-one correspondence.

Hard- and soft-thresholding procedures have been proposed in the literature [14,18]. A notorious soft-thresholding approach corresponds to WGCNA, which raises the computed PCC to a power $\beta \geq 1$ to ensure a scale-free behavior [8]. This procedure enhances high correlations at the expense of low correlations. Hard-thresholding approaches define a significance value and keep the edges with a correlation value above the threshold. The work in [3] proposed an approach that uses permutation-based significance tests: by independently permuting the components of a gene expression profile, they estimate a p-value for each observed correlation. The authors of [1,19] selected the threshold based on the behavior of the number of edges, number of nodes, and network density, resulting in a range in which the threshold can be selected. The work in [17] extended [1] by adding another consideration: once a range is defined upon the previous criteria, a clustering analysis is carried out to verify the cluster stability and choose the most appropriate end of the range to select as threshold. Unlike [17], the proposed method is based on the growing mechanisms and the topology of the underlying network. Understanding that the appearance of biologically significant modules is closely related to how tree-like structures appear and how the existing components densify, gives an insight on an emergent behavior among genes.

The results in this paper suggest that it is possible (and in some cases convenient) to use robust correlation metrics other than Pearson for obtaining gene co-expression networks. Also, a characterization of the growing mechanisms of co-expression networks in terms of network density and rate of aggregation of isolated nodes has been presented. Although the differences in the performance among some metrics are marginal, the use of BICOR for gene function prediction yields, in terms of three evaluation metrics, promising results for larger GO sub-hierarchies (consisting of more than 180 functions). Similarly, MIC can be used for enhancing gene function prediction performance when the size of the GO sub-hierarchy is small (consisting of less than 90 functions). The hard-thresholding identification was used to build co-expression networks and was used to showcase a solution to the gene function prediction problem for *Oryza sativa Japonica*, a variety of rice.

As future work, the characterization of other weighted networks needs to be developed and the elucidation of similar network growing mechanisms in other contexts, as well as the exploration of other correlation metrics (e.g., information-based) for generating co-expression networks, needs to be pursued. Furthermore, this study applies a fixed range discretization with steps of 0.01 to compute the evolution of network density and isolated node aggregation, but alternative discretization approaches need to be assessed.

Acknowledgments. This work was funded by the OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y Validación en Arroz y Caña de Azúcar), anchored at the Pontificia Universidad Javeriana in Cali and funded within the Colombian Scientific Ecosystem by The World Bank, the Colombian Ministry of Science, Technology and Innovation, the Colombian Ministry of Education, and the Colombian Ministry of Industry and Tourism, and ICETEX, under GRANT ID: FP44842-217-2018.

References

1. Aoki, K., Ogata, Y., Shibata, D.: Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. **48**, 381–90 (2007)
2. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A.: NCBI GEO: archive for functional genomics data sets-update. Nucl. Acids Res. **41**(D1), D991–D995 (2012). https://doi.org/10.1093/nar/gks1193
3. Carter, S.L., Brechbühler, C.M., Griffin, M., Bond, A.T.: Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics **20**(14), 2242–2250 (2004). https://doi.org/10.1093/bioinformatics/bth234
4. Couto, C.M.V., Comin, C.H., Costa, L.D.F.: Effects of threshold on the topology of gene co-expression networks. Molecular Biosyst. **13**, 2024–2035 (2017)
5. Gene Ontology Consortium: The gene ontology resource: 20 years and still going strong. Nucl. Acids Res. **47**(D1), D330–D338 (2019)

6. Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., Wei, Y.: Distance correlation application to gene co-expression network analysis. BMC Bioinform. **23**, 81 (2022)
7. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature Prot. **4**, 44–57 (2009)
8. Langfelder, P., Horvath, S.: WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. **9**, 559 (2008)
9. Levatić, J., Kocev, D., Džeroski, S.: The importance of the label hierarchy in hierarchical multi-label classification. J. Intell. Inform. Syst. **45**(2), 247–271 (2015). http://link.springer.com/10.1007/s10844-014-0347-y
10. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform. **7**(Suppl 1), S7 (2006)
11. Rau, C., Wisniewski, N., Orozco, L., Bennett, B., Weiss, J., Lusis, A.: Maximal information component analysis: a novel non-linear network analysis method. Front. Genet. **4** (2013). https://www.frontiersin.org/article/10.3389/fgene.2013.00028
12. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science (New York, N.Y.) **334**, 1518–1524 (2011)
13. Schaefer, R.J., Michno, J.M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., Myers, C.L.: Integrating co-expression networks with gwas to prioritize causal genes in maize. The Plant Cell **30**, 2922–2942 (2018)
14. Silverman, E.K., Schmidt, H.H.H.W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., Balligand, J.L., Benincasa, G., Capasso, G., Conte, F., Di Costanzo, A., Farina, L., Fiscon, G., Gatto, L., Gentili, M., Loscalzo, J., Marchese, C., Napoli, C., Paci, P., Petti, M., Quackenbush, J., Tieri, P., Viggiano, D., Vilahur, G., Glass, K., Baumbach, J.: Molecular networks in network medicine: development and applications. Wiley Interdisciplinary Reviews. Syst. Biol. Med. **12**, e1489 (2020)
15. Song, L., Langfelder, P., Horvath, S.: Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinform. **13**, 328 (2012)
16. Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Mach. Learn. **73**(2), 185–214 (2008)
17. van Verk, M.C., Bol, J.F., Linthorst, H.J.M.: Prospecting for genes involved in transcriptional regulation of plant defenses, a bioinformatics approach. BMC Plant Biol. **11**, 88 (2011)
18. Wang, Y.X.R., Li, L., Li, J.J., Huang, H.: Network modeling in biology: statistical methods for gene and brain networks. Stati. Sci. Rev. J. Inst. Math. Stat. **36**, 89–108 (2021)
19. Zhang, L., Yu, S., Zuo, K., Luo, L., Tang, K.: Identification of gene modules associated with drought response in rice by network-based analysis. PLOS ONE **7**, e33748 (2012)