Bioinformatics, 37(14), 2021, 1982–1989
doi: 10.1093/bioinformatics/btab041
Advance Access Publication Date: 1 February 2021
Original Paper

Gene expression

Robust gene coexpression networks using signed distance correlation

Javier Pardo-Diaz [1,2,*], Lyuba V. Bozhilova [1], Mariano Beguerisse-Díaz[3], Philip S. Poole[2], Charlotte M. Deane [1] and Gesine Reinert[1,*]

[1]Department of Statistics, University of Oxford, Oxford OX1 3LB, UK, [2]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK and [3]Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on June 24, 2020; revised on November 30, 2020; editorial decision on January 18, 2021; accepted on January 21, 2021

Abstract

Motivation: Even within well-studied organisms, many genes lack useful functional annotations. One way to generate such functional information is to infer biological relationships between genes/proteins, using a network of gene coexpression data that includes functional annotations. However, the lack of trustworthy functional annotations can impede the validation of such networks. Hence, there is a need for a principled method to construct gene coexpression networks that capture biological information and are structurally stable even in the absence of functional information.

Results: We introduce the concept of signed distance correlation as a measure of dependency between two variables, and apply it to generate gene coexpression networks. Distance correlation offers a more intuitive approach to network construction than commonly used methods, such as Pearson correlation and mutual information. We propose a framework to generate self-consistent networks using signed distance correlation purely from gene expression data, with no additional information. We analyse data from three different organisms to illustrate how networks generated with our method are more stable and capture more biological information compared to networks obtained from Pearson correlation or mutual information.

Contact: jdiaz@stats.ox.ac.uk or reinert@stats.ox.ac.uk

Availability and implementation: Code is available online (https://github.com/javier-pardodiaz/sdcorGCN).

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression data, while noisy, contains key information about biological processes (Kothapalli *et al.*, 2002). Such data are often represented as gene coexpression networks, where nodes are genes and edges represent correlations in their expression across multiple samples (Lee *et al.*, 2004). Representing gene coexpression as networks eases the study and visualization of the expression data (Magwene and Kim, 2004; Weirauch, 2011). One motivation behind creating these networks is that genes which are coexpressed across multiple samples are likely to have related functions (Hughes *et al.*, 2000; Makrodimitris *et al.*, 2020; Stuart *et al.*, 2003; van Noort *et al.*, 2003), allowing inference of gene function using *guilt by association* approaches (Wolfe *et al.*, 2005). This procedure is especially useful if the studied organism is poorly annotated. For example, *Rhizobium leguminosarum*, a soil bacteria important in agriculture that can infect plants of the legume family and provide them organic nitrogenous compound, has no functional information available for ~25% of its predicted genes. Most methods to generate and validate gene coexpression networks use exogenous biological information (such as gene ontologies and metabolic information) to select which edges need to or do not have to be present (e.g. Bar-Joseph *et al.*, 2003; Ihmels *et al.*, 2002; Ucar *et al.*, 2007). Therefore, the lack of reliable genomic functional information may hinder the construction of gene coexpression networks and the validation of their accuracy.

The most widely used methods to generate gene coexpression networks in the absence of exogenous information are based on the absolute value of the Pearson correlation of the expression of gene pairs across samples (Weirauch, 2011). After computing the correlation, there are two alternatives: (i) construct a fully connected weighted network, or (ii) impose a threshold to construct unweighted networks with edges connecting genes whose expression correlation is high enough. The former approach is widely used thanks to the R package WGCNA (Langfelder and Horvath, 2008), but results in noisy networks where gene relationships may not be easy to identify. The latter approach (e.g. George *et al.*, 2019) keeps the strongest relationships; however, it is not obvious which threshold value strikes the right balance. A threshold too low, results in overly dense networks that are difficult to analyse; a threshold that is too high risks discarding valuable information.

A natural way of studying gene expression data is to compare the expression of a gene in different samples to assess how it changes. For any two genes, the most intuitive approach should follow the same straightforward procedure: evaluate how the expression of each gene changes across the different samples and then compare the patterns of changes between the two genes. This same idea underpins the concept of *distance correlation* (Székely *et al.*, 2007). While Pearson correlation measures linear relationships, distance correlation measures the dependence, both linear and nonlinear, between two vectors, and provides a non-negative score that is zero if and only if the vectors are statistically independent (Székely *et al.*, 2007). Thus, distance correlations allow us to identify relationships between the expression of genes beyond linearity; this type of correlation has been successfully used in bioinformatics settings to predict miRNA-disease associations (Zhao *et al.*, 2018) and to generate gene regulatory networks from expression data (Ghanbari *et al.*, 2019; Guo *et al.*, 2014). Gene regulatory networks are different from gene coexpression networks because they are directed, much sparser and aim to identify regulatory pathways rather than functional associations. We give a detailed definition of distance correlation in Section 2 below.

Distance correlation is always non-negative; that is, it does not capture whether an association between the expression of two genes is positive or negative. Naturally, this information may be biologically relevant; to overcome this shortcoming, we introduce a *signed* distance correlation. After calculating the distance correlation between each pair of genes, we impose a sign, which corresponds to the sign of the Pearson correlation between the expression of the genes.

Here, we propose a method to construct gene coexpression networks using signed distance correlation as an intuitive alternative to networks from Pearson correlation, Spearman correlations and mutual information. To highlight the strengths of our approach, we construct and compare networks using the four methods. This work is to our knowledge the first work using distance correlation to construct gene coexpression networks.

We construct networks by including only edges between genes for which the signed distance correlation of their expression exceeds a threshold. We select the threshold based on the internal consistency of the networks using the R package COGENT (Bozhilova *et al.*, 2020) instead of using exogenous biological information known (or imposed) *a priori*. We evaluate our method in data from three different organisms. First, we generate an unweighted gene coexpression network for the bacteria *R.leguminosarum* from microarray data. We then analyse RNA-Seq data from the yeast *Saccharomyces cerevisiae*, and single-cell RNA-Seq data from human liver cells. The results of our analysis of yeast and human data can be found in the Supplementary Material.

We evaluate the biological information in our networks using the STRING database (Szklarczyk *et al.*, 2019), which is a protein–protein interaction database with scores for pairs of proteins. The higher the STRING score, the more likely they are to have a biologically meaningful functional interaction. Using STRING, we show that networks from signed distance correlation capture more biological information and are structurally more stable than networks based on Pearson or Spearman correlation or mutual information.

While we apply our method to gene expression data, our method to construct networks from signed distance correlations (in combination with COGENT) can be used in applications beyond bioinformatics.

Data and source code are available from https://github.com/javier-pardodiaz/sdcorGCN.

2 Materials and methods

The method we propose generates an unweighted coexpression network from gene expression data that may come from different sources, such as microarrays, RNA-Seq and single-cell RNA-Seq assays. Figure 1 illustrates the main steps of the method, which includes data pre-processing, computing correlations, and thresholding to create networks.

The input to the method is a $m \times n$ gene expression matrix M, where each of the m rows correspond to a gene, each of the n columns is a different sample, and the entries are the expression values of each gene in each sample.

2.1 Data

We analyse gene expression data from three different organisms: *R.leguminosarum*, Yeast (*S.cerevisiae*), and Human (*Homo sapiens*), obtained using different experimental techniques (microarrays, RNA-Seq and single-cell RNA-Seq). Below, we present our results on the *R.leguminosarum* dataset. The description and analysis of yeast and human datasets can be found in Supplementary Sections 3 and 4.

The *R.leguminosarum* bv. *viciae* 3841 data contains gene expression information observed under 18 different growth conditions. These data come from $n = 54$ microarray channels with 3 independent samples per condition (Karunakaran *et al.*, 2009; Pini *et al.*, 2017; Ramachandran *et al.*, 2011). The complete list of the conditions is in Supplementary Section 1. From the total 7263 genes in the current genome annotation (Young *et al.*, 2006), we remove genes that do not appear in all the microarrays or appear as pseudogenes, leaving $m = 7,077$ genes for which we calculate the mean expression within each microarray. These data are encoded in the $7,077 \times 54$ matrix M.

The data for all three organisms in the form of expression matrices (genes in rows and samples in columns) are available at https://github.com/javier-pardodiaz/sdcorGCN and http://opig.stats.ox.ac.uk/resources.

2.2 Pre-processing

Gene expression data is noisy and the raw values are only comparable within the same experiment due to their arbitrary scales. Therefore, gene expression data requires some pre-processing before we can use it to generate networks (Libralon *et al.*, 2009). We apply quantile normalization (Bolstad *et al.*, 2003) to the gene expression matrix M. This normalization step renders the distribution of the expression values in different samples (i.e. the columns of M) identical in their statistical properties, such as maximum value and quantiles. This normalization enables us to compare data from different experiments.

To avoid interference from low expression values in the quantile normalization, we ignore the 20% least expressed genes from each sample before the normalization step. After the quantile normalization, we set the ignored values to the lowest expression value in M to decrease the level of noise. In practice, we have observed that 20% offers a good balance between preserving as much information as possible, and weeding out noisy measurements. We denote the pre-processed expression data by M^*, and its i-th row by $M^*_{i,\cdot}$.

2.3 Computing correlations between the expression of the genes

Distance correlation is as a measure of association between random vectors that addresses some of the limitations of linear measures such as Pearson's (Székely *et al.*, 2007). To compute the distance correlation between the expression of two genes i and j, let the vectors $M^*_{i,\cdot} = (M^*_{i,1}, \dots M^*_{i,n})$ and $M^*_{j,\cdot} = (M^*_{j,1}, \dots, M^*_{j,n})$ contain their expression values across the n samples. For each gene $i \in \{1, \dots, m\}$, we calculate the $n \times n$ expression distance matrix $\tilde{Y}^{(i)}$ whose entries are the absolute differences between the expression values across samples:

$$\tilde{Y}^{(i)}_{h,k} = |M^*_{i,h} - M^*_{i,k}|, \tag{1}$$

where $h, k \in \{1, \dots, n\}$ are all the samples in the data. Then, we compute the double-centred expression distance matrix $Y^{(i)}$:

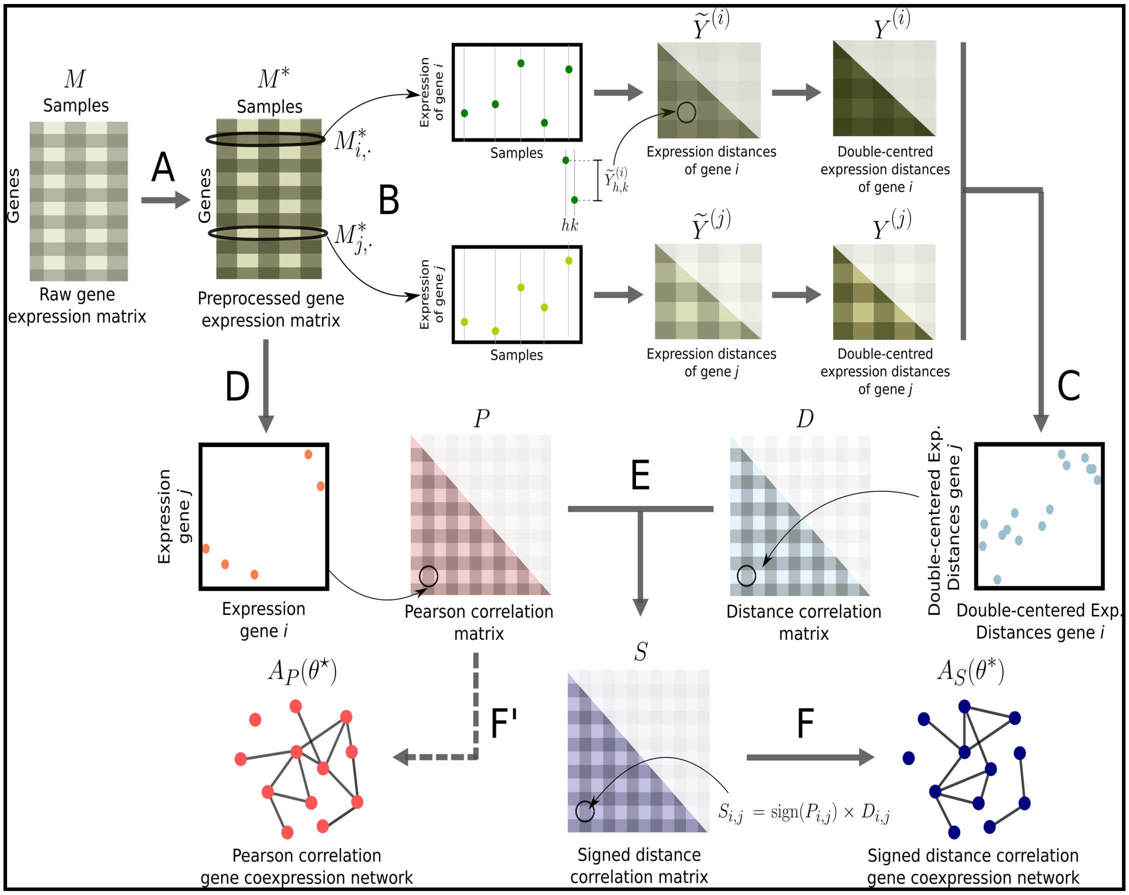
M
Samples
Genes
Raw gene expression matrix
A
M^*
Samples
Genes
$M^*_{i,\cdot}$
$M^*_{j,\cdot}$
Preprocessed gene expression matrix
B
Expression of gene i
Samples
Expression of gene j
Samples
$\tilde{Y}^{(i)}_{h,k}$
hk
$\tilde{Y}^{(i)}$
Expression distances of gene i
$Y^{(i)}$
Double-centred expression distances of gene i
$\tilde{Y}^{(j)}$
Expression distances of gene j
$Y^{(j)}$
Double-centred expression distances of gene j
C
D
Expression gene j
Expression gene i
P
Pearson correlation matrix
E
D
Distance correlation matrix
Double-centered Exp. Distances gene j
Double-centered Exp. Distances gene i
$A_P(\theta^\star)$
Pearson correlation gene coexpression network
F'
S
$S_{i,j} = \mathrm{sign}(P_{i,j}) \times D_{i,j}$
Signed distance correlation matrix
F
$A_S(\theta^*)$
Signed distance correlation gene coexpression network

Fig. 1. Pipeline to construct networks from gene expression data using signed distance correlation. (**A**) We pre-process the input matrix M with raw gene expression data using quantile normalization and setting the lowest 20% values from each sample to the minimum value in M to obtain M^*. (**B**) We compute the expression distance matrices $\tilde{Y}^{(i)}$ and $\tilde{Y}^{(j)}$ for each gene $i, j \in \{1, \ldots, m\}$, and we double center them to obtain $Y^{(i)}$ and $Y^{(j)}$. (**C**) We compute the distance correlation matrix D, whose entries $D_{i,j}$ are the positive root of the Pearson correlation between $Y^{(i)}$ and $Y^{(j)}$, for every pair of genes. (**D**) We compute the Pearson correlation between each pair of rows in the M^* to obtain the Pearson correlation matrix P. (**E**) To construct the signed distance correlation matrix S, we multiply every distance correlation between the expression of two genes $D_{i,j}$ by the sign of their Pearson correlation $\mathrm{sign}(P_{i,j})$. (**F**) Using COGENT (Bozhilova *et al.*, 2020), we find the optimal threshold θ^* that produces the most self-consistent network $A_S(\theta^*)$ from S. (**F′**) Analogously, we find the optimal threshold θ^* to generate the network $A_P(\theta^*)$ from P; this step is not part of the pipeline and only necessary to be able to compare Pearson and signed distance correlation networks.

$$Y^{(i)}_{h,k} = \tilde{Y}^{(i)}_{h,k} - \frac{1}{n}\sum_{k} \tilde{Y}^{(i)}_{h,k} - \frac{1}{n}\sum_{h} \tilde{Y}^{(i)}_{h,k} + \frac{1}{n^2}\sum_{h,k} \tilde{Y}^{(i)}_{h,k}. \qquad (2)$$

where $a, b \in \{1, \ldots, n\}$. The distance covariance of the expression of genes i and j is:

$$\mathrm{dcov}(i,j) = \frac{1}{n}\left(\sum_{h,k} Y^{(i)}_{h,k} Y^{(j)}_{h,k}\right)^{1/2}, \qquad (3)$$

where $h, k \in \{1, \ldots, n\}$. Finally, the distance correlation between the expression of the genes is

$$\mathrm{dcor}(i,j) = \frac{\mathrm{dcov}(i,j)}{\sqrt{\mathrm{dcov}(i,i)\mathrm{dcov}(j,j)}}. \qquad (4)$$

This expression is the non-negative square root of the Pearson correlation between $Y^{(i)}$ and $Y^{(j)}$. We store the pairwise distance correlation between the expression of all genes in the $m \times m$ symmetric matrix D with entries $D_{i,j} = \mathrm{dcor}(i,j)$.

The distance correlation in Equation (4) is always non-negative; however, the association between the expression of two genes can be either positive (i.e. both genes are expressed at the same time) or negative (i.e. one gene is expressed when the other is not). Thus, the sign of the association may contain crucial biological information that is lost if we only use distance correlation. We introduce sign into distance correlation by using the correlation matrix P that contains the Pearson correlation coefficients between the expression of each pair of genes:

$$P_{i,j} = \frac{\mathrm{cov}(M^*_{i,\cdot}, M^*_{j,\cdot})}{\sigma_{M^*_{i,\cdot}} \sigma_{M^*_{j,\cdot}}}. \qquad (5)$$

The matrix P also has size $m \times m$ but, unlike D, it may have negative values: values close to 1 mean strong positive correlation, values close to -1 mean strong negative correlation, and values around 0 mean no correlation. We generate a signed distance correlation matrix S whose values are the entries in D multiplied by the sign of the corresponding entries in P:

$$S_{i,j} = \mathrm{sign}(P_{i,j})D_{i,j}. \qquad (6)$$

For comparison purposes, we create the matrix I which contains the mutual information values for each pair of gene expression vectors from the preprocessed dataset M^*. To compute I, we use the R package minet (Meyer *et al.*, 2008). We select the empirical probability distribution as the entropy estimator, and use $\sqrt{n}$ bins with equal frequencies for the discretization.

2.4 Thresholding correlation matrices

Once we have a signed distance correlation S, we generate an unweighted network with adjacency matrix $A_S(\theta)$ by applying a threshold θ to S:

$$A_S(\theta)_{i,j} = \begin{cases} 1 & \text{if } S_{i,j} \geq \theta, \\ 0 & \text{otherwise}. \end{cases} \tag{7}$$

The matrix $A_S(\theta)$ encodes an unweighted undirected network in which pairs of genes are connected if there is a strong and positive signed distance correlation in their expression. Naturally, different values of θ result in networks with different properties. We want to find the θ that minimizes the number of edges between genes that are not coexpressed, and maximizes the number of edges between genes that are coexpressed. For this purpose, we use the R package COGENT (Consistency of Gene Expression NeTworks) (Bozhilova *et al.*, 2020).

The main COGENT functions evaluate the internal consistency of a method to generate networks from a specific dataset. First, COGENT splits the gene expression data in two possibly overlapping groups of samples, and then constructs a network with the same node set from each group, G_1 and G_2. Then, COGENT measures the similarity between G_1 and G_2; the more similar the networks, the higher the internal consistency of the method. COGENT helps to find a threshold θ^* that results in the most internally consistent networks. The COGENT function getEdgeSimilarityCorrected provides a score of edge similarity between G_1 and G_2 and adjusts it so that results obtained for different values of θ are comparable. We use the semi-random density adjustment implemented in COGENT and select the function parameter that allows to keep the isolated nodes during the analysis.

The edge similarity between G_1 and G_2 is the Jaccard index of the set of edges. This index is adjusted using the randomized networks G_1^* and G_2^* from a a configuration-type model from the degree sequences d_1^* and d_2^*. The degree sequences d_1^* and d_2^* are random permutations of the degree sequences of G_1 and G_2. More precisely, the similarity between G_1 and G_2 is:

$$\mathrm{sim}(G_1, G_2) = \frac{|E_1 \cap E_2| - \beta}{|E_1 \cup E_2| + \beta}, \tag{8}$$

where

$$\beta = \frac{|E_1| \sum_{(u,v)\in E_1} d_2^*(u) d_2^*(v)}{\sum_{u\in V_2} \sum_{\substack{v\in V_2 \\ v\neq u}} d_2^*(u) d_2^*(v)} + \frac{|E_2| \sum_{(u,v)\in E_2} d_1^*(u) d_1^*(v)}{\sum_{u\in V_1} \sum_{\substack{v\in V_1 \\ v\neq u}} d_1^*(u) d_1^*(v)},$$

and E_i, V_i are the set of edges and vertices for network G_i, $i = 1, 2$.

The value of β is the expected edge overlap between G_1 and G_2 if they were random networks. In general, β is higher for denser networks. The similarity $\mathrm{sim}(G_1, G_2)$ in Eq. 8 is a value between $-1/3$ and 1; this value is high if G_1 is more similar to G_2 than to a randomization of G_2 and vice versa (Bozhilova *et al.*, 2020).

In our analysis, we create two random subsets of columns from M to obtain two $m \times \lfloor \frac{3n}{4} \rfloor$ matrices M_1 and M_2, such that half of the total number of samples n (i.e. the columns of M) are shared between both subsets, and 1/3 of their columns are different. Here, $\lfloor \frac{3n}{4} \rfloor$ denotes the greatest integer less or equal to $\frac{3n}{4}$. We choose this amount of overlap to evaluate how the relatively small change of about 1/3 of the data affects the final result. The choice of amount of overlap is user defined and may also depend on the research question. When the interest lies in clustering of genes, then one may like to choose an amount of overlap which leads to a moderately sparse network. Our previous *in silico* experiments show that if the proportion of samples from M shared between M_1 and M_2 is much smaller than 50%, the similarity between the obtained networks is low, and proportions much larger than 50% produce almost identical networks. We pre-process M_1 and M_2 and compute the signed distance correlation matrices S_1 and S_2 as outlined in Section 2.3. We test different values of θ, and for each of them, we obtain two unweighted networks and their similarity with COGENT. We repeat this whole process 25 times, every time with different subsets of columns of M. Finally, we compute the similarity score $s(\theta)$, which is the average of the similarity of the networks (in Eq. 8) over the 25 samples.

To favour signal over noise, we create a score that balances the similarity of the networks in $s(\theta)$ with the density of $A_S(\theta)$:

$$\mathrm{Score}(\theta) = s(\theta) - \mathrm{density}(A_S(\theta)). \tag{9}$$

We select the threshold θ^* that retrieves the highest Score(θ) and use this value to generate the unweighted gene coexpression network $A_S(\theta^*)$ from the signed distance correlation matrix S. We also calculate in the same manner the optimal thresholds $\theta^\star$ and $\theta^\diamond$ to construct the unweighted networks $A_P(\theta^\star)$ and $A_I(\theta^\diamond)$ from the Pearson correlation matrix P and the mutual information matrix I.

2.5 Network comparison

We compare gene coexpression networks obtained from thresholding the signed distance correlation matrix S, the Pearson correlation matrix P, and the MI matrix I. The density of a network may influence the amount of biological information it captures and hence we cannot compare networks with different densities. Instead, we generate additional networks from each matrix which match the different optimal edge densities. Letting the edge density of $A_S(\theta^*)$, $A_P(\theta^\star)$, and $A_I(\theta^\diamond)$ be d_S, d_P, and d_I, we compare the following nine networks:

- $NS(d_S)$: Network from S with edge density d_S (i.e. $A_S(\theta^*)$).
- $NP(d_P)$: Network from P with edge density d_P (i.e. $A_P(\theta^\star)$).
- $NI(d_I)$: Network from I with edge density d_I (i.e. $A_I(\theta^\diamond)$).
- $NP(d_S)$: Network from P with edge density d_S.
- $NI(d_S)$: Network from I with edge density d_S.
- $NS(d_P)$: Network from S with edge density d_P.
- $NI(d_P)$: Network from I with edge density d_P.
- $NS(d_I)$: Network from S with edge density d_I.
- $NP(d_I)$: Network from P with edge density d_I.

To construct the last six networks, we simply find a threshold manually that produces networks with density d_S, d_P, and d_I. Since for the *R.leguminosarum* dataset d_P is roughly the same as d_I (see Section 3), we use the network $NS(d_P)$ as a proxy for the network $NS(d_I)$ and omit networks $NP(d_I)$ and $NI(d_P)$.

We first evaluate the internal consistency of each network with COGENT. We also evaluate the biological information contained in the networks using STRING, a database of known and predicted protein–protein interactions (Szklarczyk *et al.*, 2019). STRING collects information from numerous sources, including experimental data, computational predictions and textmining. The association evidence in STRING is categorized into independent channels, weighted, and integrated to produce a confidence score C for all recorded protein interactions. Interactions with high C score are more likely to be true than those with a low score.

We work with three different sets of confidence scores:

- C: Total scores provided by STRING.
- C†: Scores that *only* consider coexpression information (coexpression channel combined with coexpression transferred channel).
- C‡: Scores that *exclude* coexpression information.

We provide details of how to retrieve C† and C‡ in Supplementary Section 5.

For each network, we add the confidence score associated with each pair of connected genes and divide by the number of edges. We perform this operation independently for the three sets of confidence scores to obtain three aggregate confidence scores per network. These scores represent the amount of biological information that the networks capture. Then we compare these scores to the expected amount of biological information captured by chance. To do so, we generate three sets of 30 random networks with edge densities d_S, d_P, d_I, and evaluate their biological content following the approach detailed above. We compare the distribution of the aggregate C, C† and C‡ scores from the random networks with density d_S with the scores from networks $NS(d_S)$, $NP(d_S)$, and $NI(d_S)$, the scores from the random networks with density d_P with $NS(d_P)$ and $NP(d_P)$, and the scores from the random networks with density d_P with $NS(d_I)$

and $NI(d_I)$. For the *R.leguminosarum* dataset, we omit the random networks with density d_I and compare $NI(d_I)$ with the network with density d_P due to the similarity in their densities. In Supplementary Section 2, we also perform the biological evaluation of the networks $NR(d_S)$ and $NR(d_P)$ with edge density d_S and d_P, respectively, obtained from a Spearman correlation matrix R.

3 Results

Here, we present our analysis of the *R.leguminosarum* dataset; the results for the yeast and human data are in Supplementary Sections 3 and 4.

3.1 Correlation matrices

The correlation matrices P (Pearson), D (distance), and S (signed distance) are all symmetric with $m = 7,077$ rows and columns. As we show in Table 1 and Figure 2, the distribution of the absolute values in P are different to those in D. In particular, the distribution of the values of P and S are different. Supplementary Figures S2 and S5 and Tables S4 and S7 contain the same analysis for the yeast and human data.

3.2 Gene coexpression networks

We first estimate, using COGENT (Bozhilova *et al.*, 2020), the optimal thresholds θ^*, $\theta^\star$, and $\theta^\diamond$ to construct the networks $NS(d_S)$, $NP(d_P)$, and $NI(d_I)$ from the matrices S, P, and I, respectively. To analyse and compare the networks, it is perhaps more intuitive to compare the networks using their edge density than with the threshold used to produce them. For this, we construct networks for a range of values of θ. For each density (and the θ that produced it), we evaluate its consistency score Score(θ) using Eq. 9. This score is related to the self-consistency of the network. Figure 3 shows the value of Score(θ) as a function of the density of the networks.

Table 1. Statistical summary of the correlation matrices from the *R.leguminosarum* data

Correlation	Min	1st q	Median	3rd q	Max	Mean
D	0.02	0.22	0.28	0.36	0.99	0.30
$\|P\|$	0.00	0.08	0.17	0.29	1.00	0.20
P	−0.86	−0.17	−0.01	0.17	1.00	0.00
S	−0.92	−0.27	−0.14	0.28	0.99	0.00

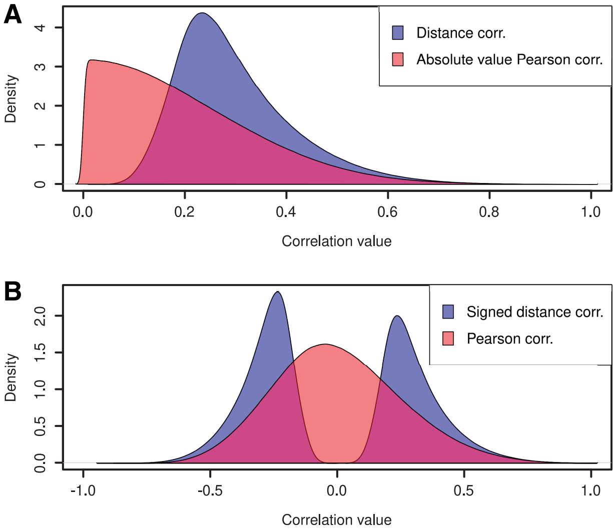

Fig. 2. Distribution of the entries of the correlation matrices from the *R.leguminosarum dataset*. (**A**) Distribution of distance correlations in D (blue), and the absolute of the Pearson correlations $|P|$ (red). (**B**) Distribution of signed distance correlation S (blue), and Pearson correlation P (red).

The highest score from the signed distance correlation networks $NS(d_S)$ is Score(θ^*) = 0.456, where $\theta^* = 0.62$, and density $d_S = 1.25\%$. The highest score from the Pearson correlation networks $NP(d_P)$ is Score($\theta^\star$) = 0.426, where $\theta^\star = 0.58$, and density $d_P = 1.63\%$. The highest score from the mutual information networks $NI(d_I)$ is Score($\theta^\diamond$) = 0.140, where $\theta^\diamond = 0.65$, and density $d_I = 1.65\%$. We also create the networks $NP(d_S)$ from P and $NI(d_I)$ from I to match the edge density of $NS(d_S)$, and the network $NS(d_P)$ from S to match the edge density of $NP(d_P)$ and $NI(d_I)$. Table 2 contains a statistical summary of the networks. The discrepancy in the number of edges in $NS(d_S)$ and $NI(d_S)$ is due to several pairs of genes having the same mutual information value. For both edge densities, the networks retrieved using our signed distance correlation matrix S ($NS(d_S)$ and $NS(d_P)$) have a smaller, denser, and with higher global clustering coefficient largest connected component (LCC) than the networks obtained using Pearson correlation ($NP(d_S)$ and $NP(d_P)$). The LCC of the networks obtained using mutual information ($NI(d_S)$ and $NI(d_I)$) are smaller and denser than those from density-matching networks but they show the lowest global clustering coefficient. See Supplementary Tables S5 and S8 for the gene coexpression networks for the yeast and human data.

3.3 Network evaluation

We perform two evaluations of the networks from signed distance correlation, Pearson correlation, and mutual information: self-consistency and biological content. We are interested in the self-consistency of the networks because it reflects their ability to cope with changes in the data used to generate them. If our network is more self-consistent, then we can have greater confidence in the biological conclusions we draw from it, even if the data is imperfect or noisy. The Jaccard similarity in Eq. 8 measures the similarity between networks generated from overlapping, non-identical subsets of a dataset using the same network construction method. The higher the similarity, the higher the internal consistency of the method and the more self-consistent the network obtained from applying it is. We measure the self-consistency Score(θ), obtained from computing the average Jaccard similarity of networks from different randomized subsets of data and subtracting the density of $A_S(\theta)$ ($A_P(\theta)$ and $A_I(\theta)$ for Pearson correlation and mutual information, respectively). As we show in Figure 3, the self-consistency scores of the networks from signed distance correlations are consistently higher than in the Pearson and mutual information networks over an interval of densities that includes the maxima for all three methods. The scores obtained for the mutual information networks are the lowest. From this analysis, we conclude that signed distance correlation networks are more self-consistent than networks based on Pearson correlation or mutual information. Even the optimal threshold for the Pearson matrix P produces a less self-consistent network $NP(d_P)$ than a signed distance correlation network with a matching edge density ($NS(d_P)$); the same applies for mutual information. See Supplementary Figures S3 and S6 for the same analysis on the yeast and human data, with similar results.

To evaluate the biological content of a network, we add the STRING scores of the edges using: all the information in STRING (C), only coexpression information ($C^\dagger$), and everything except coexpression information ($C^\ddagger$), and then divide by the number of edges in the network, as described in Section 2.5. Table 3 and Figure 4 (and Supplementary Fig. S1) show the results for all the networks, and the mean scores from random networks. In every case, the signed distance correlation networks contain more biological information than networks from Pearson correlation, those obtained using mutual information, and the randomized networks. See Supplementary Tables S6 and S9 and Figures S4 and S7 for a similar analysis on the human and yeast data, with similar results. The highest difference of the networks with the randomized networks occurs when we only use coexpression information. The $C^\dagger$ scores of $NS(d_S)$, $NP(d_S)$, and $NI(d_S)$ are 10, 8.5, and 8.7 times higher than the expected ones for random networks. This is not surprising

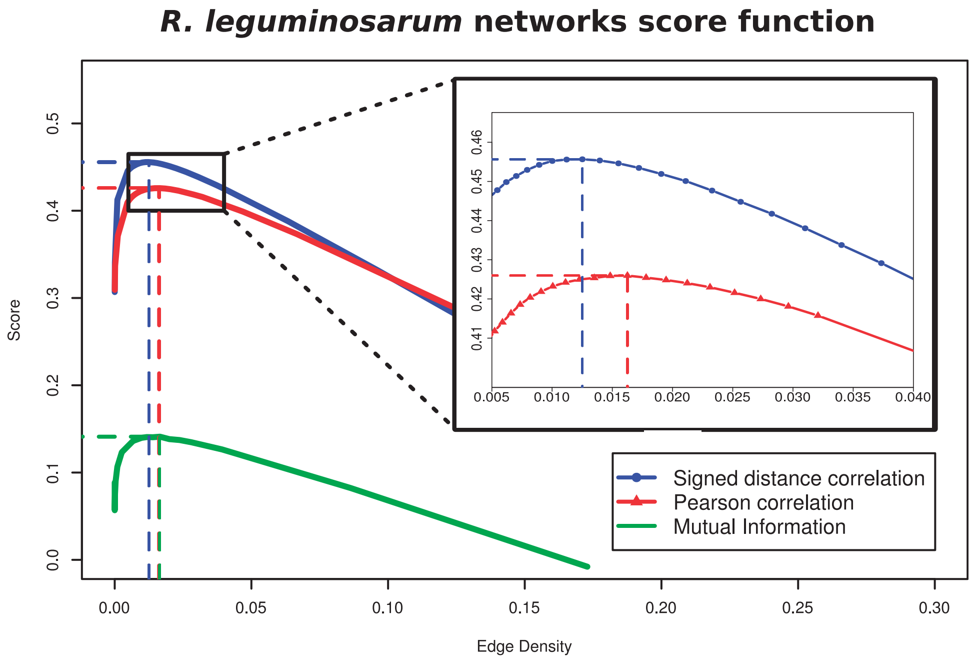

Fig. 3. Self-consistency scores of *R.leguminosarum* networks with edge densities between 0 and 0.3. The blue line with circles shows the scores of networks obtained using signed distance correlations; the red line with triangles shows the score of networks using Pearson correlations; the green line shows the scores of the networks obtained using MI. The dashed vertical lines indicate the density of the most self-consistent network for each type of correlation.

Table 2. Statistical summary of *R.leguminosarum* coexpression networks. LCC denotes largest connected component

Network	Number of edges	Number of vertices in LCC	Edge density LCC×100	Global clustering coeff. LCC
$NS(d_S)$	313 348	6431	1.52	0.570
$NS(d_P)$	406 977	6688	1.82	0.570
$NP(d_S)$	313 348	6697	1.40	0.544
$NP(d_P)$	406 977	6880	1.72	0.543
$NI(d_S)$	317 014	5993	1.77	0.279
$NI(d_I)$	414 140	6084	2.24	0.284

Table 3. Evaluation of the biological content of the networks with STRING

Network	All of STRING information (C)	Only coexpression information ($C^{\dagger}$)	All information except coexpression ($C^{\ddagger}$)
$NS(d_S)$	**29.61**	**9.02**	**26.55**
$NP(d_S)$	25.23	7.63	22.72
$NI(d_S)$	24.68	7.81	22.01
RE d_S	7.49 ± 0.08	0.90 ± 0.02	7.09 ± 0.08
$NS(d_P)$	**26.79**	**7.64**	**24.12**
$NP(d_P)$	23.32	6.61	21.08
$NI(d_I)$	22.39	6.60	20.07
RE d_P	7.54 ± 0.08	0.91 ± 0.03	7.13 ± 0.08

Note: RE indicates the expected (mean) result based on random networks with the indicated edge density and its SD. The values in bold correspond to the highest scores for each set of networks and confidence scores.

because we have built the networks using gene expression information. However, the coexpression data used to construct these networks is different to the data in STRING. The scores excluding the coexpression information $C^{\ddagger}$ are 3.7, 3.2, and 3.1 times higher than in the random networks. This is remarkable because this assessment is performed on data that is completely different than the data used to construct the networks. Hence, by applying our pipeline to gene expression data, we can identify new types of relationships between proteins (and genes). These results demonstrate the power of gene coexpression networks to predict functional interactions, especially when constructed using signed distance correlation.

The analysis for networks with d_P and d_I tells a similar story; the scores obtained by the signed distance correlation network $NS(d_P)$ are higher than the scores from the Pearson correlation network $NP(d_P)$ and the mutual information network $NS(d_I)$, despite d_P and d_I being the optimal edge density for Pearson correlation and mutual information. We highlight the fact that in absolute terms (i.e. not dividing the scores by the number of edges in the network) when we use only

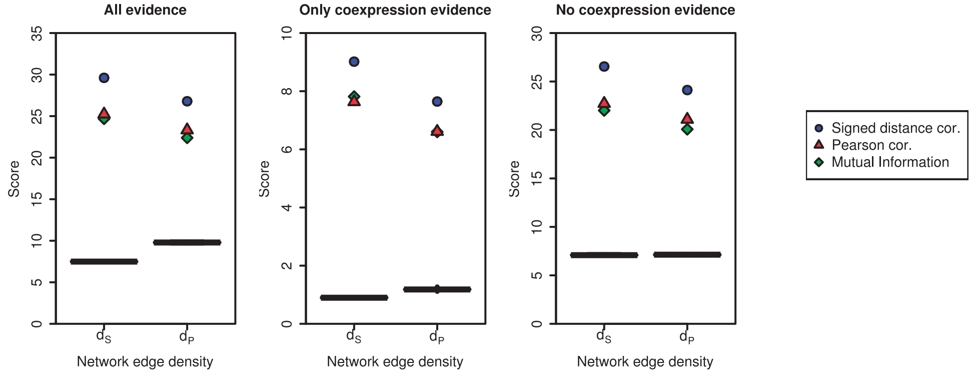

Fig. 4. Scores obtained for the *R.leguminosarum* gene coexpression networks using STRING. All panels show the score for the different networks in the *y* axis, and the network density on the *x* axis. The scores are the result of adding up the confidence scores with all evidence (C), only coexpression evidence ($C^{\dagger}$) and everything excluding coexpression ($C^{\ddagger}$) from STRING associated with the edges in the networks, each computed using different of information. The black box plots correspond to the scores obtained by 30 random networks. Blue circles, red triangles, and green diamonds represent signed distance correlation, Pearson correlation, and mutual information respectively.

coexpression information ($C^{\dagger}$), the score of $NS(d_S)$ is higher than the score of $NP(d_P)$, even though the former has fewer edges than the latter, and therefore its score is the result of adding fewer terms.

For the three datasets we analyse, signed distance correlation networks perform better than Pearson correlation and mutual information networks with matching densities, according to both evaluations and for all tested edge densities.

4 Discussion

In this work, we have introduced signed distance correlation, and presented a method to construct networks in a self-consistent way exclusively from gene expression data. This method has three main steps: data pre-processing, computing correlations, and thresholding. These steps combine well-established methods such as quantile normalization, and the use of COGENT (Bozhilova *et al.*, 2020) to identify the optimal threshold.

Distance correlation is an intuitive approach to study gene expression because it relies on the differences in the expression between samples. By incorporating signs into distance correlation, we can also differentiate between positive and negative relationships, and maintain the advantages of distance correlation.

We apply our method to data from *R.leguminosarum*, yeast, and human. In all cases, our method produces networks that are more self-consistent than using Pearson correlation and mutual information. The reason why self-consistency is so important is that it ensures that our results are robust to changes or noise in the data. Therefore, when we cannot assess the biological significance of a network directly, we can use its self-consistency as an indication of biological significance. Networks from signed distance correlation also capture more biological information than networks from Pearson correlation and mutual information, as shown by our evaluation using STRING. In the case of *R.leguminosarum*, the signed distance correlation network (using an optimal threshold θ^* found with COGENT) captures almost four times more biological information than random networks. The amount of captured biological information is less if we use Pearson or Spearman correlations (Supplementary Fig. S1) or mutual information to build the networks.

These results give a weak indication that self-consistent networks derived from biological data might capture more biological information than those with less stability. Therefore, if we cannot assess the biological significance of a network directly, measuring its self-consistency may serve as a proxy. The use of COGENT to select the threshold values requires a sufficiently strong signal in the association between the expression values of different genes. For example, using Euclidean distance, scaled by the square root of the sample size, did not yield sufficient signal, so that networks created by the subsampling method had no more overlap than expected at random. These networks also did not capture much biological information (data not shown).

We applied our method to construct, to our knowledge, the first gene coexpression network for *R.leguminosarum*. This network promises to reveal rich biological information that will illuminate our understanding of plant–bacteria interactions and nitrogen fixation, and it is therefore the starting point for further investigations of the biological mechanisms of this organism. In particular, we plan to identify groups of genes in *R.leguminosarum* which are highly connected in the network and associate them with specific biological processes. To do so, we will make use of community detection techniques and new experimental data. For the human liver dataset one could explore predicting disease-related biological information; see for example Song *et al.* (2019); Chen *et al.* (2018); Li *et al.* (2019).

Finally, we have showcased our method on gene expression datasets from different organisms obtained using different techniques: microarrays (*R.leguminosarum*), RNA-Seq (yeast), and single-cell RNA-Seq (human). However, the methods that we have developed are general, and can also be used to construct networks in a vast range of domains, such as, for example, economics (Wang *et al.*, 2018), neuroscience (Bernhardt *et al.*, 2011), climatology (Donges *et al.*, 2009), or any discipline where networks can be constructed from correlation data.

Acknowledgements

We thank Alison K. East and Florian Klimm for help and advice on how to analyse the datasets used in this manuscript. We thank Katharine Turner (ANU), Florian Klimm, and Malte D. Luecken for fruitful discussions. In addition, we acknowledge support from COST Action CA15109, Keble College Oxford and Keble Association. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558), and Carlos Outerial for assistance in the computing.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (EP/R512333/1 to J.P.-D., M.B.-D., P.S.P., C.M.D. and G.R., EP/L016044/1 to L.V.B., C.M.D. and G.R.), the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/T001801/1 to P.S.P. and G.R.), the COSTNET COST Action (CA15109 to G.R.) and e-Therapeutics plc. M.B.-D. acknowledges support from the Oxford-Emirates Data Science Lab.

Conflict of Interest: none declared.

References

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Bernhardt,B.C. *et al.* (2011) Graph-theoretical analysis reveals disrupted small-world organization of cortical thickness correlation networks in temporal lobe epilepsy. *Cereb. Cortex*, **21**, 2147–2157.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bozhilova,L.V. *et al.* (2020) COGENT: evaluating the consistency of gene co-expression networks. *Bioinformatics*, btaa787.

Chen,X. *et al.* (2018) MDHGI: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS Comput. Biol.*, **14**, e1006418.

Donges,J.F. *et al.* (2009) Complex networks in climate dynamics. *Eur. Phys. J. Spec. Top.*, **174**, 157–179.

George,G. *et al.* (2019) Gene co-expression network analysis for identifying genetic markers in Parkinson's disease-a three-way comparative approach. *Genomics*, **111**, 819–830.

Ghanbari,M. *et al.* (2019) The distance precision matrix: computing networks from non-linear relationships. *Bioinformatics*, **35**, 1009–1017.

Guo,X. *et al.* (2014) Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PLoS One*, **9**, e87446.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ihmels,J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Karunakaran,R. *et al.* (2009) Transcriptomic analysis of *Rhizobium leguminosarum biovar viciae* in symbiosis with host plants *Pisum sativum* and *Vicia cracca*. *J. Bacteriol.*, **191**, 4002–4014.

Kothapalli,R. *et al.* (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Li,H. *et al.* (2019) A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front. Microbiol.*, **10**, 676.

Libralon,G.L. *et al.* (2009) Pre-processing for noise detection in gene expression classification data. *J. Braz. Comput. Soc.*, **15**, 3–11.

Magwene,P.M. and Kim,J. (2004) Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.*, **5**, R100.

Makrodimitris,S. *et al.* (2020) Metric learning on expression data for gene function prediction. *Bioinformatics*, **36**, 1182–1190.

Meyer,P.E. *et al.* (2008) minet: ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Pini,F. *et al.* (2017) Lux bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiol.*, **174**, 1289–1306.

Ramachandran,V.K. *et al.* (2011) Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.*, **12**, R106.

Song,F. *et al.* (2019) mies: predicting the essentiality of mirnas with machine learning and sequence features. *Bioinformatics*, **35**, 1053–1054.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Székely,G.J. *et al.* (2007) Measuring and testing dependence by correlation of distances. *Ann. Stat.*, **35**, 2769–2794.

Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.

Ucar,D. *et al.* (2007) Construction of a reference gene association network from multiple profiling data: application to data analysis. *Bioinformatics*, **23**, 2716–2724.

van Noort,V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Wang,G.-J. *et al.* (2018) Correlation structure and evolution of world stock markets: evidence from Pearson and partial correlation-based networks. *Comput. Econ.*, **51**, 607–635.

Weirauch,M.T. (2011) Gene coexpression networks for the analysis of DNA microarray data. *Appl. Stat. Netw. Biol.*, **1**, 215–250.

Wolfe,C.J. *et al.* (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Young,J.P.W. *et al.* (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.*, **7**, R34.

Zhao,H. *et al.* (2018) Prediction of microRNA-disease associations based on distance correlation set. *BMC Bioinformatics*, **19**, 141.