

Extracting information from gene coexpression networks of *Rhizobium leguminosarum*

Javier Pardo-Díaz^{1,3,*}, Mariano Beguerisse-Díaz²,
Philip S. Poole³, Charlotte M. Deane¹ and Gesine Reinert¹

¹ Department of Statistics, University of Oxford, UK

² Mathematical Institute, University of Oxford, UK

³ Department of Plant Sciences, University of Oxford, UK

March 1, 2022

Abstract

Nitrogen uptake in legumes is facilitated by bacteria such as *Rhizobium leguminosarum*. For this bacterium, gene expression data are available, but functional gene annotation is less well developed **than for other model organisms**. More annotations could lead to a better understanding of the pathways for growth, plant colonisation and nitrogen fixation in *Rhizobium leguminosarum*. In this paper we present a pipeline which combines novel scores from gene coexpression network analysis in a principled way in order to identify genes which are associated with certain growth conditions or highly coexpressed with predefined set of genes of interest. This association may lead to putative functional annotation or to a prioritised list of genes for further study.

Key words. Gene co-expression network analysis; functional annotation; *Rhizobium leguminosarum*.

1 Introduction

Rhizobium leguminosarum is a bacterium that fixes atmospheric nitrogen when associated with legumes (e.g. peas, beans, lentils). *R. leguminosarum*

transforms molecular nitrogen in the air into ammonia which can be assimilated by plants. Nitrogen fixation improves the growth of plants as nitrogen is one of the limiting factors during the growth process (Gutiérrez, 2012). Bacteria such as *R. leguminosarum* grow in the soil in the absence of their legume host (Downie, 2010) (*free-living conditions*) and when sensing the plant, they approach its root hairs and attach. The bacteria are eventually released into the developing nodule, where they differentiate to a nitrogen-fixing state called *bacteroid* (Downie, 2010). Thus, better understanding the metabolic and regulatory pathways in *R. leguminosarum* for growth, plant colonisation, and nitrogen fixation may lead to improved farming methods of legumes; see also Waterman (2016).

Currently, functional gene annotation in *R. leguminosarum* is far from complete: we know the precise function for only around 25% of the genes in the genome, and have no functional annotation at all for another 25%; the remaining genes have limited functional annotation. To facilitate the identification of functional groups of genes, based on gene expression data we aim to find *R. leguminosarum* genes (*new candidate genes*) which are highly connected to a pre-selected group of genes (*seed list*), or to genes which have a high expression under a particular experimental condition.

Our objectives are to assign the new candidate genes a putative molecular function and unravel which signals trigger their expression.

As main tool we use gene coexpression network analysis. Gene coexpression can be assessed via correlation measures of the expression across multiple samples, see for example Wang *et al.* (2014) and Pardo-Diaz *et al.* (2021b). Gene coexpression networks are networks where nodes are genes and edges indicate (high) coexpression of these genes (Lee *et al.*, 2004). Representing gene coexpression as networks eases the study and visualisation of the expression data (Weirauch, 2011; Magwene and Kim, 2004) and helps exploit the structure of gene interactions at a whole-system level. One motivation behind creating these networks is that genes which are highly coexpressed across multiple samples are likely to have related functions (Hughes *et al.*, 2000; Stuart *et al.*, 2003; van Noort *et al.*, 2003; Makrodimitris *et al.*, 2020), allowing inference of gene function using *guilt by association* approaches (Wolfe *et al.*, 2005) in particular if the studied organism is poorly annotated.

Gene coexpression networks can be weighted (so that there is a *weight* value associated with each edge that represent the level of correlation of the expression) or unweighted. According to Pardo-Diaz *et al.* (2021a), keeping the correlation values of the gene expression as weights only if they are higher

than a chosen threshold results in thresholded and weighted networks that can capture more biological information than unweighted networks, while reducing noise. In this paper, we only analyse thresholded and weighted networks which are constructed using signed distance correlation, following the pipeline described in Pardo-Diaz *et al.* (2021a), with the threshold value chosen using COGENT (Bozhilova *et al.*, 2021). Networks based on signed distance correlation can be more robust and capture more biological information than those constructed using Pearson correlation, Spearman correlation or Mutual Information (Pardo-Diaz *et al.*, 2021a,b). This paper builds on Pardo-Diaz *et al.* (2021a) by taking the networks from Pardo-Diaz *et al.* (2021a) as starting points to extract biological signals from the gene expression data.

In the literature, there are different approaches to extract information from gene networks in general and gene coexpression networks in particular (Thalamuthu *et al.*, 2006). Here, we identify candidate genes for annotation in two different *R. leguminosarum* coexpression networks. In Cowen *et al.* (2017) an excellent survey is provided about methods for gene function prediction which are based on propagating information about available functional annotation through the network. This paper uses these ideas by introducing a score that quantifies such propagation via PageRank, but also introduces a score based community detection as well as a score based on edge weights of nearest neighbours. We also provide a novel way to combine these scores in order to enhance the signals from the scores. We assess how the results depend on the different techniques and approaches employed, as well as on the input network. Our main findings are that all three scores can successfully retrieve genes which are highly coexpressed with those in the seed list using a set of control genes, and that a combination of these scores adds considerable signal. We also present these scores as a tool to compare different experimental conditions. While the methods in this study are used to explore the gene coexpression networks for *R. leguminosarum*, they could be applied to other (weighted or unweighted) coexpression networks.

The paper is structured as follows. In Section 2 we recall some basic network notions, the problem of community detection in networks, and personalised PageRank. Section 3 details the three scores we use in our analysis; it also includes our methodology to evaluate them, combine them, and compare them across different experimental set-ups. In Section 4 we present the results of the evaluation of the scores, and then we illustrate with two examples how to apply our approach. Lastly, in Section 5 we summarise our

findings, compare them with the information available in the literature, and present directions for future work.

The data, the code and the main outputs are available online at <https://github.com/javier-pardodiaz/NetAnalysis>.

2 Background

In this paper a network is simple and undirected, on a set V of N nodes with adjacency matrix A , whose elements are edge weights; $A_{i,j} = 0$ indicates that there is no edge between nodes i and j . The *degree* of node i is the number j such that $A_{i,j} > 0$ and is denoted by k_i , for $i = 1, \dots, N$. The number of edges is then $m = \frac{1}{2} \sum_{i=1}^N k_i$. The *strength* of node i is $s(i) = \sum_{j:j \neq i} A_{i,j}$; for an unweighted network, the strength equals the degree.

2.1 Community detection

Detecting communities allows non-trivial internal network organisation to be unveiled at a coarse grain level (Yang *et al.*, 2016). Communities are usually understood to be groups of nodes that are connected “more densely” to each other than to the rest of the network. Some community detection methods are based on the maximization of the *modularity*, that is, the difference between the number of edges in communities and the expected number under a configuration model (Newman and Girvan, 2004). For a given partition $P = \{P_1, \dots, P_k\}$ of the set of nodes, for node i we set $\sigma_i = k$ if $i \in P_k$. Reichardt and Bornholdt (2006) add a resolution parameter (γ) to the modularity to yield as quality function $Q(P) = \sum_{i,j} \left(A_{i,j} - \gamma \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$, where $\delta(\sigma_i, \sigma_j) = 1$ if i and j are in the same set in the partition P , and 0 otherwise. High values of the resolution parameter γ result in small communities; low γ values lead to large communities. A popular algorithm for maximising $Q(P)$ is the Leiden algorithm (Traag *et al.*, 2019). The modularity function in this algorithm is piecewise constant in γ ; the algorithm also provides the values of γ at which the modularity function changes. These values are interpreted as “optimal” choices for γ .

2.2 PageRank and personalised PageRank

A typical objective when studying a network is to identify which nodes are the most *important* ones (those with the highest centrality). The PageRank algorithm, originally developed to rank web-pages (Brin and Page, 1998), tackles this problem. This algorithm produces a probability distribution over the N nodes of a directed network G that represents the fraction of time that a random walker spends in each node in an infinitely long random walk. At each step, the random walker can, with probability α , choose to follow an outgoing edge from the current node (chosen with probability proportional to its weight), or, with probability $1 - \alpha$, “teleport” to any node chosen from a distribution v . This algorithm can also be applied to undirected networks. For simplicity assume that there are no isolated nodes in the undirected network, and that all node strengths are positive, $s_i > 0$ for all $i = 1, \dots, N$. Let $x(t) = (x_1(t), \dots, x_N(t)) \geq 0$ be the **vector with $x_i(t)$ denoting the** probability that **node i** receives a random walker at time t (so that $\sum_{i=1}^N x_i(t) = 1$, for all t), which evolves according to

$$x(t) = \alpha (D^{-1})^T x(t-1) + (1 - \alpha)v. \quad (1)$$

The matrix D is the diagonal matrix of the strength of the nodes, so that $D_{i,i} = s_i > 0$ is the strength of node i and $D_{i,j} = 0$ if $i \neq j$. The PageRank vector π is the solution of equation (1) when $t \rightarrow \infty$, that is $\pi = \alpha (D^{-1})^T \pi + (1 - \alpha)v$; see Xing and Ghorbani (2004).

Typically the teleportation distribution v is chosen to be uniform (i.e., $v_i = \frac{1}{N}$ for all i); however, *personalised PageRank* Gleich (2015) manipulates v to bias teleported walkers to a specific node or group of nodes. For example if $v_i = 0.9$ and $v_j = \frac{0.1}{N-1}$ for $j \neq i$, teleported walkers will go to node i with probability 0.9. This has the effect that nodes near i will have a higher personalised PageRank than with uniform teleportation.

3 Methods

3.1 Data

We analyse two different *R. leguminosarum* networks (*standard* and *complete*), both of them obtained applying the approach described in Pardo-Diaz *et al.* (2021a) but using two different microarray datasets, each of which contain gene expression values. The *standard* dataset includes three replicates

of 18 different conditions, making a total of 54 microarrays; these conditions include free-living bacteria, bacteria from the soil in which legumes are grown (rhizosphere), and bacteroids. What we call the *complete* dataset includes a total of 87 microarrays from 35 different conditions; however not all conditions have replicates. Both networks have the same set of 7,077 nodes, representing genes. These expression data are published in Karunakaran *et al.* (2009); Ramachandran *et al.* (2011); Pini *et al.* (2017); Prell *et al.* (2009); Terpolilli *et al.* (2016); Garcia-Fraile *et al.* (2015) but have never been analysed jointly. The list of microarrays and the summary statistics of both networks are in the Supplementary Information SI A.

For the analysis in Section 4.4, we use a subset of microarrays of the published data. These microarrays include expression values for ten conditions of interest (sample channel) and also “control” conditions (control channel). The expression values from the sample channel of all these microarrays are included (among other data) in our “complete” dataset. These different conditions, as well as the control conditions in the microarray, are given in Table 3.1. We generate an expression matrix in which the rows are all the genes in the network and the columns are the expression values of both channels of each microarray. On the columns of this matrix we perform quantile normalization (Bolstad *et al.*, 2003) so that the values in the different columns are comparable. We note that instead of using microarray data one could use RNA-Seq data. If using TPM (Transcripts Per Million) (Li and Dewey, 2011) as expression units, there is no need to quantile-normalise the expression data since the values for the different samples are already comparable.

3.2 Identification of new candidate genes

To study the networks, we present three complementary scores to identify *new candidate genes* which are highly coexpressed with a given set of genes, or with genes that are highly expressed under a particular condition. For each of the scores, we include two approaches, depending on the input.

The first approach identifies genes which are highly coexpressed with a predefined selection of genes we want to study (*seed genes*). This predefined selection of genes could include genes involved in a particular process, genes with a particular expression pattern, or just a gene which is studied as part of an exploratory analysis. We denote this preselected set of genes by L . The

Abbr.	Condition of interest	Abbr.	Control condition	Type
Pyr	Pyruvate NH ₄	Glu	Glucose NH ₄	OA
Suc	Succinate NH ₄	Glu	Glucose NH ₄	OA
OAC	Acetate NH ₄	Glu	Glucose NH ₄	OA
ACAC	Acetoacetate NH ₄	Glu	Glucose NH ₄	OA
PCA	Protocatechuate NH ₄	Pyr	Pyruvate NH ₄	OA
For	Formate NH ₄	Pyr	Pyruvate NH ₄	OA
HBA	Hydroxybenzoate NH ₄	Pyr	Pyruvate NH ₄	OA
Ino	Inositol NH ₄	Glu	Glucose NH ₄	O
Ara	Arabinose NH ₄	Glu	Glucose NH ₄	S
Gal	Galactose NH ₄	Glu	Glucose NH ₄	S

Table 1: Experimental conditions for both channels in the microarrays (sample and control) used to study the gene expression in different growth media. All of these conditions are free-living. The growth conditions are OA for organic acid, O for other, and S for Sugar.

seed genes may not be straightforward to select. For example, for selecting genes which are highly expressed in a particular experimental condition, a typical method is to choose all genes whose expression exceeds a pre-set threshold, but an objective method for setting such a threshold is lacking. Different threshold choices can lead to different results.

To overcome this problem, the second approach retrieves scores which reflect the level of coexpression of genes with genes which are highly expressed in a given condition, taking the expression values of all the genes in that condition as an input. We denote the input expression vector by E and the expression of gene i by $E(i)$. An advantage of using these scores instead of simply selecting the genes which show the highest expression, or the highest increase in expression between conditions, is that we can extract more information from the data. For example, one gene might increase its expression only slightly between two given conditions but could be highly coexpressed with a gene that shows a high increase in its expression. Using our methods we can capture this gene with relatively low expression which would not have been picked otherwise. This second approach might also be useful when using other types of continuous values instead of expression data.

We use three ways of obtaining scores for every gene in the network. The interpretation for each of the scores is that the higher the score, the more

connected the gene is with the genes in the list L , or with the genes which have high expression values. The three scoring systems are as follows.

3.2.1 Community detection co-occurrence score

We perform community detection analysis on the networks using the Leiden algorithm (Traag *et al.*, 2019) to optimise the network partitions. Different values in the resolution parameter γ lead to different partitions, but, as a function of γ , the quality function $Q(P)$ has only finitely many local maxima. These local maxima are estimated using the Optimiser object included in the Leiden module, and they result in a set of partitions. For a given resolution value, we then select the community that contains the highest number of seed genes. This approach is based on the assumption that highly coexpressed genes tend to be at the same community; for an illustration which supports this assumption see Fig. 2. To combine the output of the community detection algorithm at different resolutions, we propose a *co-occurrence score* that evaluates each gene on the frequency of it belonging to communities enriched in seed genes. Let R be the set of resolutions for which partitions are considered, let $C_r^{(i)}$ the set of genes belonging to the same community as gene i at resolution $r \in R$, and let $\tau(i)$ be the indicator function which equals 1 if i is in the list L , and 0 otherwise. Then we define the community detection co-occurrence score $S_{CD}^{(i)}$ for gene i as

$$S_{CD}^{(i)} = \sum_{r \in R} \frac{|C_r^{(i)} \cap L| - \tau(i)}{(|C_r^{(i)}| - \tau(i))|L|}. \quad (2)$$

Thus, genes that tend to belong to small communities with a high number of seed genes across most of the partitions will have a high score. The addition of τ adjusts the co-occurrence score for genes in the seed list L , to help avoid over-estimation. When including gene expression (or any other continuous value) in our analysis, we modify the score function $S_{CD}^{(i)}$ to $Z_{CD}^{(i)}$ so that $Z_{CD}^{(i)}$ gives a high score to those genes that belong to communities which are enriched in highly expressed genes. We define

$$Z_{CD}^{(i)} = \frac{1000}{\sum_{v \in V} E(v)} \sum_{r \in R} \frac{\sum_{j \in C_r^{(i)}} E(j)}{|C_r^{(i)}|}, \quad (3)$$

where $E(j)$ denotes the expression value of gene j .

3.2.2 Personalised PageRank score

In each of the networks we analyse, we compute the PageRank vector π (Brin and Page, 1998). This vector provides a centrality measurement $\pi^{(i)}$ for each node i in the network. These values represent the fraction of time that a random walker spends in each node in an infinitely long random walk across the network (see Section 2.2). At each step, we use $\alpha = 0.85$ so that the teleporting probability is $1 - 0.85 = 0.15$ (the standard parameter value, according to Brin and Page (1998)). **An extensive discussion about choices of teleporting probabilities in different types of networks can be found in Gleich (2015); often the standard parameter provides a good starting point for exploration.**

In addition, for each gene $l \in L$ in the seed list we compute its personalised PageRank vector π_l (Gleich, 2015) (see Section 2.2), with the personalisation that teleportation takes place only to gene l itself. We set the entry $\pi_l^{(l)}$ to zero so that in the final score the genes in the seed list can be compared to the other genes in the network. Then, we compute the average of the personalised PageRank vectors of the genes included in the seed list;

$$\rho_L(i) = \frac{1}{|L| - \tau(i)} \sum_{l \in L} \pi_l^{(i)}.$$

The vector ρ_L offers a score for each of the genes in the network, including those in the seed list. We expect genes which are highly connected to those in the seed to have higher values in ρ_L than in π . For this reason, we define our PageRank score for gene i as the ratio

$$S_{PR}^{(i)} = \rho_L^{(i)} / \pi^{(i)}.$$

Similarly to the co-occurrence score, we may like to take gene expression into account instead. To calculate a PageRank vector $\rho_{L,E}$ that takes gene expression into account, we use a personalised PageRank in which when teleporting, the probability to teleport to a gene k is equals to its gene expression value $E(k)$ divided by the sum of the expression of all the genes in the network, so that genes with a low expression will be visited less often than those with a high expression. With the resulting personalised PageRank vectors $\rho_{L,E}$ we set

$$Z_{PR}^{(i)} = \rho_{L,E}^{(i)} / \pi^{(i)}.$$

3.2.3 Edge weight score

The scores in subsections 3.2.1 and 3.2.2 are relatively computer-intensive. We also introduce a straightforward edge-based score. For each gene $i \notin L$ we calculate its edge weight score $S_{EW}^{(i)}$ as

$$S_{EW}^{(i)} = \frac{1}{|L|} \sum_{l \in L} A_{i,l},$$

the average weight of the edges connecting i to the genes in the seed list. For a seed list gene $i \in L$ we set

$$S_{EW}^{(i)} = \frac{1}{|L| - 1} \sum_{l \in L, l \neq i} A_{i,l}.$$

This score favours genes with high degree or strength. For using gene expression instead of a seed list, we introduce the score

$$Z_{EW}^{(i)} = \frac{1}{N - 1} \sum_{v \in V} A_{i,v} E(v).$$

3.3 Evaluation of the scores using control genes

To evaluate the performance of the different scores for studying seed lists from sections 3.2.1 to 3.2.3, we use some *control genes*. These genes are 49 highly coexpressed genes that code for ribosomal proteins, according to the genome annotation in Young *et al.* (2006). The average signed distance correlation (Pardo-Diaz *et al.*, 2021b) of the expression of the control genes is 0.79 when using the standard dataset and 0.70 when using the complete one. The control genes are listed in Table SI B in the Supplementary Information. According to Wheatley *et al.* (2020), most of these genes are essential **for bacterial growth** in all three studied conditions, namely free-living, rhizosphere, and bacteroid. The control genes show a high expression in free-living conditions and a lower expression in the rhizosphere and bacteroid stages. This difference in their expression might relate to the fact that the growth of bacteria and the synthesis of proteins decreases during the two latter stages (especially in bacteroids, which do not divide) and therefore ribosomes are less needed.

To evaluate the scores, we take a random subset of five control genes and use them as seed list (*control seed list*). Then we score all the genes in the

coexpression network using the three different scores from Subsections 3.2.1 - 3.2.3. The higher the scores, the more related the genes are deemed to be to the selected five genes. As all the control genes are coexpressed and functionally related, we expect the remaining 44 control genes to retrieve high score values. We compare the rank of these 44 genes for the different scores to their ranks when sampling all the genes in the network uniformly at random. If the scores are informative then the rank of the 44 genes should be larger using the control seed list compared to ranking them randomly. We use the scores as a classifier of all genes, classifying whether or not they belong to the list of 49 ribosomal proteins, so that there are exactly 44 true positives. We repeat this simulation process 25 times, at each iteration taking a potentially different random subset of five genes as control seed list. To evaluate the results we use the Area Under the Receiver Operating Characteristics curve (AUROC).

3.4 Combination of the scores

Given the scores from Subsections 3.2.1 - 3.2.3, we select genes based on their individual scores as well as their performance across all three scores. Specifically, we select a set of genes based on the rule “among the top M for at least one of the three scores, and, of those, we select the genes which are among the top T for all three scores, or among the top S for at least one of the scores”. Here we assume that $T \geq M \geq S$.

If we were just to select the top S genes for each of K scores, we would expect to select roughly KS genes. To assess the effect of selecting also genes which are among the top M for any of the K scores, and which are among the top T for all K scores, let W denote the number of such genes selected. Using a Poisson approximation, we can approximate the probability of selecting at least as many genes as W if the ranks were allocated at random for each score. Let

$$\lambda = \lambda(M, T; K, N) = N \left\{ \left(\frac{T}{N} \right)^K - \left(\frac{T-M}{N} \right)^K \right\}. \quad (4)$$

Then, for $m \geq 3\lambda$,

$$\mathbb{P}(W \geq m) \leq 2 \frac{m+1}{m+1-\lambda} \frac{1}{\sqrt{2\pi m}} \exp \left\{ -\frac{(m-\lambda)^2}{2(m+\lambda)} \right\}. \quad (5)$$

More details can be found in Supplementary Information SI D.

3.5 Joint analysis for different experimental conditions

In this paper, we analyse *R. leguminosarum* gene expression across ten different experimental conditions, as detailed in Table 3.1. We are not interested in the most highly expressed genes under the different conditions as most of them will be the same constitutive genes highly expressed under all conditions. Instead we focus on those genes which are upregulated when compared to a reference condition; in this paper we use growth in glucose as reference condition. To obtain the genes with high scores for the different conditions compared to their scores in the glucose reference condition, we use the quantile-normalised expression matrix presented in Subsection 3.1. We first calculate the Z_{CD} , Z_{PR} , and Z_{EW} scores for all the genes in the network, as detailed in Subsections 3.2.1 - 3.2.3, for each of the ten conditions of interest and for the ten control conditions, by using the expression values from the sample and control channels in the microarrays, respectively. This procedure yields 20 different sets of three scores for each gene (three scores, ten microarrays, two channels per microarray). If for a microarray the control condition is the same condition as the one which we compare against (here, glucose), we just divide the scores for the condition of interest by the scores for the control condition. Otherwise, we use the data from a second microarray that has as condition of interest the control condition of the first microarray: we divide the product of the scores for both conditions of interest by the product of the scores for both control conditions. We thus obtain a set of three scores for each gene for each of the ten conditions. We compare the genes with high scores across the different conditions.

Applying the procedure from Subsection 3.4 to the just obtained scores results in a set of potentially relevant genes for each individual condition; these sets may differ in sizes for different experimental conditions. Now we study the intersection of the sets of genes obtained for different conditions. For this task for simplicity we leave out the pre-filtering using M and instead, for each condition we take the genes which are either in the top S for at least one of the scores, or in the top T for all of the scores. We denote the union of the genes which are selected for the different conditions by Q . We use $S = 5$ and $T = 20$, which makes Q likely to be relatively small. To understand the behaviour of genes in Q across conditions, we now introduce a possibly more lenient threshold, U . For each gene in Q , for each condition we record whether or not it is among the top U genes for at least one of the scores. Thus, every gene will have at least one entry, namely the condition which

qualified the gene to be part of the union Q . It may however have relatively high scores for some other conditions as well, which is what this procedure addresses. The pipeline for this procedure is shown in Fig. 1. This analysis results in a matrix of genes and conditions, which is the main output of the pipeline in this paper. The resulting pattern of genes and conditions could then be examined in further detail, for example using in-vitro experiments.

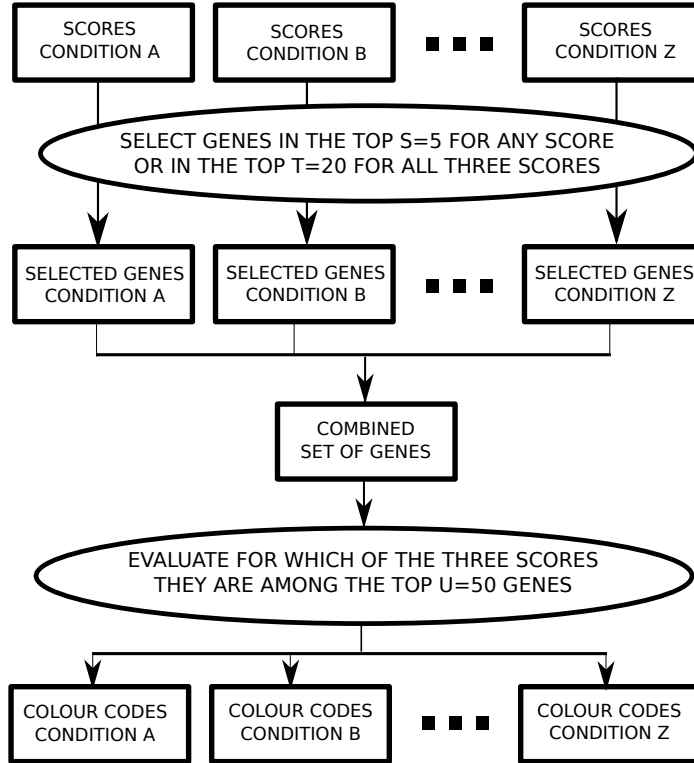


Figure 1: The pipeline for combining the scores for different conditions for retrieving genes for further investigation. **The first step is to select for each condition those genes that are in the top $S = 5$ for any of the three scores or in the top $T = 20$ for all three of them. The second step consists in evaluating at each condition for which of the three scores the selected genes are among the top $U = 20$ scores. The output is a colour coded matrix in which each of the genes in the combined set of genes has a value for each condition (see Figure 6 for reference).**

Summarising, in this paper we present two methods to identify new can-

didate genes: one based in the use of seed gene scores (S_{XX}) and another based on the use of gene expression scores (Z_{XX}). Next, in Section 4.2, we evaluate the scores. Then we illustrate their use on *R. leguminosarum* in two ways. The first way is the use of all the ribosomal genes as seed list (Section 4.3). The second way is the use of the expression across different growth conditions (Section 4.4). These two approaches will turn out to identify different sets of genes for further study and they thus complement each other.

4 Results

4.1 Community detection results

Community detection as described in Subsection 2.1 yields 226 partitions for the standard dataset and 241 partitions for the complete dataset. Fig. 3 presents the number of communities that we retrieve at each partition.

As an illustration that communities include functionally related genes and that the outcomes may depend on the chosen partition, Fig. 2 shows that when using ribosomal genes as seed genes, a community of 48 genes contains a large number of 19 ribosomal genes, which is much higher than the number of genes expected by chance.

Moreover, Fig. 2 shows that communities with different sizes can result in a different subset of ribosomal genes. Hence a principled way of combining the results of community detection outputs is required, further motivating our score from Subsection 3.2.1.

4.2 Performance of the seed list scores

As detailed in Subsection 3.3, we use different subsets of the control genes as control seed lists to evaluate the performance of the scores S_{CD} , S_{PR} , and S_{EW} . We score all the genes in the networks using the scores and rank the genes; then, we compare the rank of the remaining control genes (those not included in the control seed list) to their ranks when sampling all the genes in the network uniformly at random. If the scores were unrelated to the functionality of the genes, then all orderings would be equally likely, and control genes would not have higher scores on average than other genes. Using the score as a classifier of whether or not a gene is a control gene (a ribosomal gene in this case), we assess its performance through the AUROC.

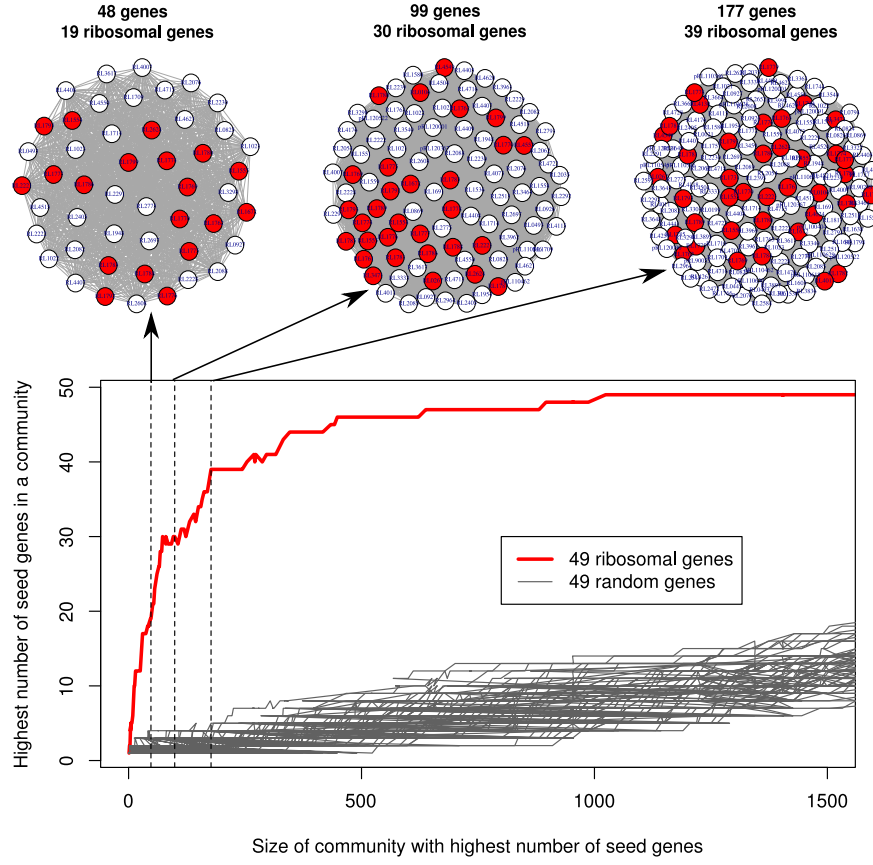


Figure 2: **Top:** Communities with the highest number of ribosomal genes (in red) for different partitions. **Bottom:** Plot representing for different partitions the size of the community with the highest number of ribosomal genes and the number of ribosomal genes in those communities.

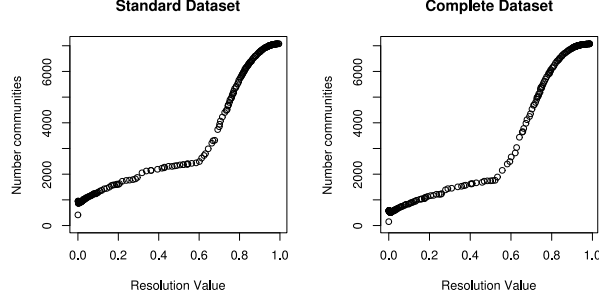


Figure 3: Number of communities in partitions depending on the value of resolution value. The behaviour in the standard dataset and in the complete dataset show a similar nonlinear dependence on the resolution value.

We expect to observe several “not ribosomal genes” with high scores since genes coexpressed with the ribosomal genes do not necessarily need to be ribosomal genes themselves; nevertheless, we expect all ribosomal genes to have a high score. Only the 44 ribosomal genes which are not in the control seed list are counted as true positives.

The upper panels in Fig. 4 show the relationship between the true positive rate and the false positive rate for the studied networks (standard and complete) when using control seed lists of size five. The boxplots reflect the distribution of the AUROC. We observe that the AUROC for the three scores is higher than the one obtained when ordering the genes randomly, which indicates that the remaining control genes have higher scores than expected. We obtain the same results when we include a different number of genes in the control seed list (Fig. SI 1 in the Supplementary Information SI C). These results suggest that all the three scores capture some signal and are appropriate for retrieving genes functionally related to those included in the seed list.

Fig. 4 indicates that, for both networks, when including five control genes in the control seed list, the AUROC for the edge weight scores is significantly higher than for the other two scores (ANOVA tests with p -values < 0.05). However, we do not observe this trend when using other control seed list sizes instead of five, see Fig. SI 1 in the Supplementary Information SI C; we note that for Fig. SI 1 we only perform 10 instead of 25 iterations. The only significant difference we observe is that the edge weight score performs

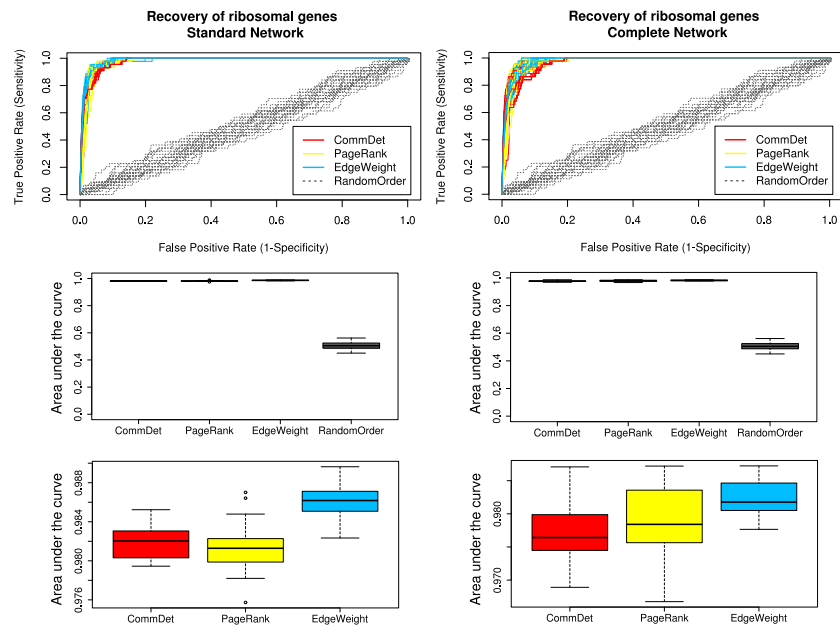


Figure 4: Scores obtained using the control seed list. All scores perform significantly better than random. There is no clear ranking between the scores.

better than the PageRank score when the size of the control seed list is 25 and using the standard network. Therefore, we do not have strong evidence to favour one score above the others.

We also observe that for most cases this case that the AUROC for the standard network is higher than for the complete network. We discuss this point further in Section 5.

4.3 Scoring genes based on the ribosomal genes

We study which genes in the network are highly coexpressed with the ribosomal genes listed in Table SI 4 in the Supplementary Information. For this purpose, we use all the ribosomal genes as a seed list (in contrast to Subsection 4.2 where we use subsets of the ribosomal genes as control seed lists), and retrieve the scores S_{CD} , S_{PR} , and S_{EW} . Fig. SI 2 in the Supplementary Information SI E shows that there is clearly some, but not a perfect, correlation between the scores; we observe Spearman’s rank correlation coefficients between 0.65 and 0.88 when comparing the different scores against each other and Spearman’s rank correlation coefficients between 0.79 and 0.91 when comparing the scores for the different datasets. A Friedman test of the null hypothesis that there is no association between the different rankings is rejected with p-value 2.2×10^{-16} , providing very strong evidence that there is an association between the rankings. For the standard network and the community detection co-occurrence score, we observe that out of the top 48 genes, 33 are ribosomal genes, whereas, if considering the different partitions separately, according to Fig. 2, in a community of 48 genes there are only 19 ribosomal genes. This shows the potential of combining the different resolutions.

The next step in our analysis is selecting those genes with a high ranking for one score or, with a possibly slightly lower ranking across all of the scores. We exclude the ribosomal genes from these ranks. First, we select those genes in the top $S = 10$ for any of the three scores, resulting in 26 and 25 genes for the complete and standard networks, respectively. This count is very much in line with what we would expect from independent counts, indicating that different scores might provide different information when attending to the highest scoring genes. Then, we concentrate on those genes which are in the top $M = 50$ for at least one of the rankings and in the top $T = 100$ for all of them and assess whether the number of genes which are selected this way is unusually large, using the Poisson approximation from (5). Their

expected count is 0.01771514, which is calculated using (4). For this category we observe $W = 17$ in the standard network and $W = 26$ in the complete network. Proposition 1, which applies because $17 > 3 \times 0.01771514$, gives that the probability of observing a count of 17 or higher if the rankings were random would be less than 4.05×10^{-5} , which is highly significant. Thus there is strong evidence that the choice of combining the rankings by picking not only genes which are highly ranked in one of them, but also genes which are somewhat less highly ranked, but across all rankings, adds considerable signal.

Overall, for our analysis of the standard as well as the complete network we select the genes which are in the top $M = 50$ for at least one of the rankings and in the top $T = 100$ for all of them, and those in the top $S = 10$ for any of the scores. The final count we get is 35 genes for the standard network and 39 genes for the complete network; 23 genes are shared between the two networks. The selected genes are detailed in Table SI 5 in the Supplementary Information SI F. The average weight of the edges connecting the selected genes with the ribosomal genes is 0.79 (0.73 if including the 131 missing edges) for the standard network and 0.72 (0.68 if including the 121 missing edges) for the complete one. The upper panels in Fig. 5 show the selected genes for the two networks and their ranks. Genes are coloured for a given score if their rank is below 100 and white otherwise; the more intense the colour, the lower the rank. We observe that the rank with the smallest overlap with the other two is the one derived from the PageRank score, suggesting that this score provides some different information. The bottom images represent induced subgraphs of the networks containing the selected genes. The nodes are coloured based on the scores for which the genes are in the top $T = 100$: only S_{CD} (red), only S_{PR} (yellow), only S_{EW} (blue), S_{CD} and S_{PR} (orange), S_{CD} and S_{EW} (purple), S_{PR} and S_{EW} (green) or all of them (grey). These network plots show that the genes are highly connected.

Most of the selected genes are essential or defective for free-living, rhizosphere, and bacteroid conditions, according to Wheatley *et al.* (2017). Out of the selected genes we find genes that code for DNA-binding proteins (RL1551 and RL1691), translation elongation factors (RL1757, RL1772, and RL2222) and subunits of the RNA polymerase (RL1766, RL1767, and RL1798). All these proteins are involved in cell growth and protein synthesis and therefore expected to be coexpressed with ribosomal genes. In the selection we also find several genes without functional annotation (RL1763, RL2291, and

Genes selected using the ribosomal genes as seed lists

Standard Network				Complete Network			
	CommDet	PageRank	EdgeWeight		CommDet	PageRank	EdgeWeight
RL0883	207	6	122	pRL100260	403	10	302
RL1022	15	96	5	pRL120235	407	5	356
RL1023	37	200	9	pRL120521	89	44	83
RL1260	383	2	235	pRL120522	48	37	17
RL1261	476	1	332	RL0883	158	3	53
RL1499	294	3	157	RL0926	31	73	6
RL1551	8	203	11	RL0927	32	62	8
RL1559	6	26	24	RL1022	37	57	13
RL1691	42	51	17	RL1023	12	89	12
RL1719	5	547	411	RL1260	852	2	441
RL1757	47	38	52	RL1551	15	124	10
RL1763	10	30	4	RL1559	93	49	73
RL1766	4	22	13	RL1691	22	36	3
RL1767	1	125	150	RL1709	38	83	24
RL1772	11	98	36	RL1710	52	25	66
RL1948	3	179	1	RL1719	8	833	634
RL2062	84	49	49	RL1757	18	14	14
RL2222	2	59	2	RL1763	9	32	7
RL2228	21	64	10	RL1766	3	19	31
RL2239	33	93	37	RL1767	1	589	474
RL2291	14	86	15	RL1772	4	29	42
RL2573	163	7	79	RL1798	69	26	51
RL2606	849	10	480	RL1948	33	88	4
RL2697	13	154	7	RL2222	2	40	2
RL3333	81	46	64	RL2239	39	87	37
RL3540	51	75	43	RL2526	383	4	383
RL3617	19	136	8	RL2686	725	6	301
RL4007	29	43	18	RL2829	622	7	344
RL4065	25	9	230	RL2964	75	51	46
RL4261	9	40	31	RL3333	20	34	20
RL4295	442	8	354	RL3540	53	15	32
RL4407	23	143	6	RL3617	10	81	18
RL4617	460	4	493	RL4407	5	91	1
RL4634	384	5	231	RL4408	6	33	9
RL4715	7	73	3	RL4409	19	72	19
				RL4617	645	8	392
				RL4634	916	1	417
				RL4635	640	9	358
				RL4715	7	60	5

- RNA polymerase
- Translation elongation factor
- ◆ DNA-binding protein
- ★ Without precise function

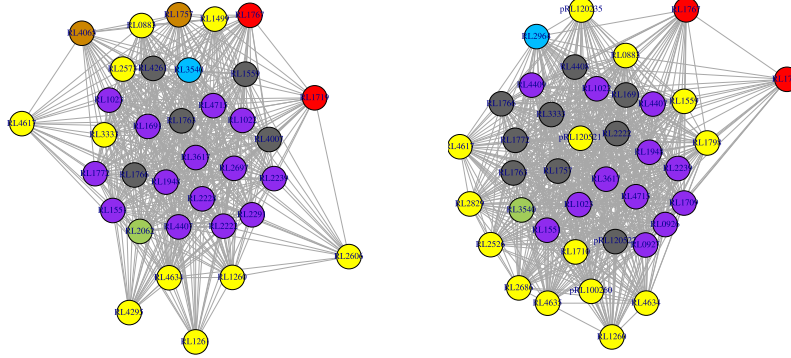


Figure 5: **Top:** the genes which are selected using $M = 50, S = 10$, and $T = 100$, for the standard network and the complete network. The colours refer to the score used, and the shade reflects the rank. The symbols next to the gene names indicate functional annotation. **Bottom:** the induced subgraphs using only these sets of genes. The colour shows for which scores the genes are in the top $T = 100$.

RL4065) and genes (RL2526, RL2697, RL4295, RL4634, and RL4635) for which only a general functional annotation is available; for example a gene may be known to be an oxidoreductase but we do not know which specific process they take part in. From their association with ribosomal genes, we conjecture that these genes might be related to protein synthesis, too.

4.4 Study of different growth conditions

In this subsection we analyse the genes with high Z_{CD} and Z_{PR} , and Z_{WE} scores for the conditions of interest, detailed in Table 3.1, compared to growth in glucose, following the approach detailed in Subsection 3.5. To select the genes included in Q we use $T = 20$ and $S = 5$. We retrieve 94 and 80 genes for the standard and the complete network, respectively. These genes are detailed in Table SI 6 in the Supplementary Information SI G and illustrated in Fig. 6. In this figure, they are coloured based on the set of scores for which they are among the top $U = 50$ scores: only Z_{CD} (red), only Z_{PR} (yellow), only Z_{EW} (blue), Z_{CD} and Z_{PR} (orange), Z_{CD} and Z_{EW} (purple), Z_{PR} and Z_{EW} (green) or all of them (black). We observe that the score that overlaps the least with the others is the Edge Weight score (blue).

We observe that the sets of genes selected for the different organic acids (Pyr, Suc, OAC, ACAC, PCA, For, and HBA) are more similar to each other than to those for the other conditions. This is unsurprising since these seven conditions are relatively similar to each other.

Moreover, RL0037 (pckA; phosphoenolpyruvate carboxykinase) has high scores (among the top $U = 50$ genes for at least one of the scores) in all the studied conditions except inositol for the complete dataset; for the standard dataset RL0037 appears for all the conditions except for inositol, galactose and formate. Similarly, RL4012 (fbaA; fructose-biphosphate aldolase) appears for all the conditions for the complete dataset. Both enzymes, pckA and fbaA are involved in the synthesis of glucose through the gluconeogenesis pathway. As the conditions are compared against glucose, it is not surprising to retrieve RL0037 and RL4012, since gluconeogenesis will not be active in bacteria grown in glucose. We do not see the same effect for other gluconeogenesis enzymes since they also catalyse the glycolysis reactions.

A group of three genes involved in alanine catabolism: pRL120416 (alanine racemase; dadX), pRL120417 (D-amino acid dehydrogenase; dadA), and RL1966 (alanine dehydrogenase; aldA) are among the top $U = 50$ genes for the three scores for succinate, acetate and formate; **RL1966 is only se-**

Genes selected using the expression in the different growth conditions

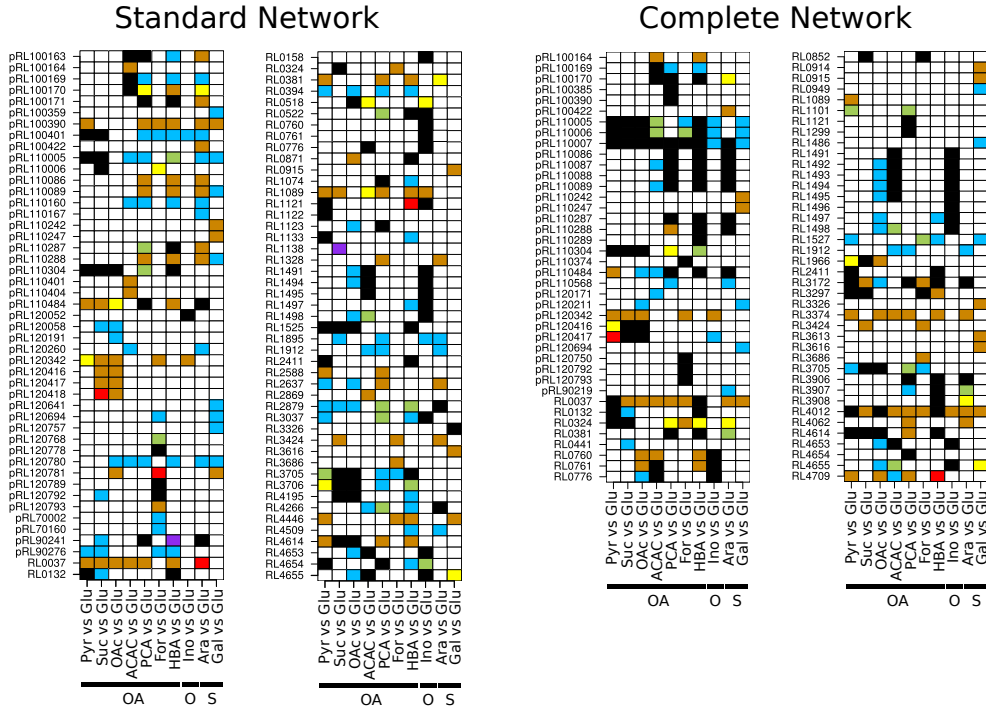


Figure 6: Genes included in Q when $T = 20$ and $S = 5$. The colours indicate for which scores they are in the top $U = 50$ for the different conditions. OA, O, and S denote Organic Acid, Other, and Sugar respectively.

lected for the complete network. These three acids, when metabolised by the cell, are transformed into acetyl-CoA. In order to avoid the accumulation of acetyl-CoA, acetyl-CoA is used to synthesise alanine. We do not observe these genes for pyruvate because growth in pyruvate does not result in an excess of acetyl-CoA.

Furthermore, the genes involved in the catabolism of inositol (RL1491, RL1494 - RL1498) are among the top $U = 50$ for all three scores and both datasets. Also, the genes related to the metabolism of protocatechuate (pRL110086 - pRL110089) are identified for protocatechuate, as well as for hydroxybenzoate and arabinose, for the complete dataset. These findings suggest some similarities between these three conditions.

There is an operon (pRL110005-pRL110007) that for the complete dataset is selected by at least two scores for all the organic acids except protocatechuate. This operon includes a transmembrane protein, an anti-sigma factor and an RNA polymerase ECF factor, suggesting that it is involved in a regulatory process which might be related to growth in this conditions.

Lastly, there are some genes such as RL0132, RL0852, and RL3172 which for which no functional annotation is currently available. We conjecture that these genes might be related to the growth in the conditions for which they are selected.

5 Discussion

Summarising, this paper introduces a principled pipeline to extract biological information from gene coexpression networks. This method is an exploratory tool to analyse and visualise gene coexpression networks which can help to identify new putative functional annotation for some genes.

We introduce three scores to retrieve genes which are highly coexpressed with a pre-selected set of genes (given by a seed list), or with genes which are highly expressed in a particular experimental condition. In both cases, the output is a set of three scores for each gene in the network. These scores are used to rank the genes for each score separately, as well as for a combination of scores. We also present a method to compare these scores across different conditions. The flexibility of choice of the parameters (S , M , T , U) allows the user to tune the size and requirements of the final set of genes.

We illustrate two ways of how to apply our approach to real gene expression data, focusing on two networks for the nitrogen-fixing bacteria *R.*

leguminosarum. These two networks are constructed from different (and overlapping) sets of microarrays following the pipeline described in Pardo-Diaz *et al.* (2021a).

In the first application, we use all 49 ribosomal genes as seed list to identify genes which are highly coexpressed with this set. Applying our approach, the final selection includes a high number of genes which are involved in protein synthesis and which are associated with cell growth. In the second application, we study and compare the genes which are selected under different growth media, illustrating that our method does not only helps to identify genes that are strongly connected with those which are highly expressed, but it also allows to compare the selected genes across conditions. Comparing across conditions may yield insights into which biological processes are shared across different experimental conditions; in this example, these are mainly metabolic processes.

Most of our results are in line with those included in Karunakaran *et al.* (2009), namely the findings related to gluconeogenesis, and alanine and inositol catabolism. Nevertheless, to our knowledge, in the literature there is no reference to the operon pRL110005-pRL110007 we mention in the results section, nor to the genes RL0132, RL0852, and RL3172. Thus, our pipeline suggests potential functional annotations for some genes which currently do not have detailed functional annotation and that have not been related to cell growth in the conditions we study. These results evince that our principled pipeline can help to increase our biological insights while agreeing with previously published data and illustrate that using gene coexpression networks from transcriptomic data can offer an advantage compared to only looking at expression levels, as was the approach taken for example in Karunakaran *et al.* (2009).

Finally there are two issues to flag. Firstly, we have shown that the three scores can successfully retrieve genes which are highly coexpressed with those in the seed list using a set of control genes. Here the control genes are ribosomal genes which are involved in protein synthesis; they have a high strength (418.57 and 490.54 for the standard and complete network, respectively) compared to the average node strength in the network (74.51 and 117.47). This difference in strength may be advantageous for our scoring methods. Ideally a list of control genes would be relatively large but comparable in strength to the full list of genes. Unfortunately, for *R. leguminosarum* we could not identify such an alternative list of control genes.

Secondly, throughout the paper we use two datasets. For both of them,

the results are similar but there are some noticeable differences. When analysing the AUROC for the different control seed lists, the standard network retrieves a higher AUROC. This behaviour might relate to the fact that the standard dataset contains a higher proportion of bacteroid and rhizosphere samples, where ribosomal genes are downregulated, which may result in a higher correlation between them. The differences when analysing the growth conditions might be related to the fact that the standard dataset contains fewer samples; in particular some specific conditions such as growth in protocatechuate and hydroxybenzoate are not included. Therefore, different networks might offer slightly different information.

In future work, we will use the approach in this paper to better understand the metabolic and regulatory changes in *R. leguminosarum* when infecting the plant. For this purpose we will use new RNA-Seq datasets from rhizosphere and bacteroid samples.

Acknowledgements. The authors would like to thank the anonymous referees for their helpful suggestions. This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/R512333/1 to JPD, MBD, PSP, CMD and GR; EP/T018445/1 and EP/R018472/1 to GR], as well as by the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/T001801/1 to PSP and GR]. MBD acknowledges support from the Oxford-Emirates Data Science Lab.

References

- Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson Approximation*. Oxford University Press.
- Bolstad, B. M., Irizarry, R. A., Åstrand, M., *et al.* (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.
- Bozhilova, L. V., Pardo-Diaz, J., Reinert, G., *et al.* (2021). COGENT: evaluating the consistency of gene co-expression networks. *Bioinformatics*, **37**(13), 1928–1929.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine.
- Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551–562.
- Downie, J. A. (2010). The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. *FEMS Microbiology Reviews*, **34**(2), 150–170.
- Garcia-Fraile, P., Seaman, J. C., Karunakaran, R., *et al.* (2015). Arabinose and protocatechuate catabolism genes are important for growth of *Rhizobium leguminosarum* biovar *viciae* in the pea rhizosphere. *Plant and Soil*, **390**(1), 251–264.

- Gleich, D. F. (2015). PageRank beyond the web. *Siam Review*, **57**(3), 321–363.
- Gutiérrez, R. A. (2012). Systems biology for enhanced plant nitrogen nutrition. *Science*, **336**(6089), 1673–1675.
- Hughes, T. R., Marton, M. J., Jones, A. R., *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**(1), 109–126.
- Karunakaran, R., Ramachandran, V., Seaman, J., *et al.* (2009). Transcriptomic analysis of *Rhizobium leguminosarum biovar viciae* in symbiosis with host plants *Pisum sativum* and *Vicia cracca*. *Journal of Bacteriology*, **191**(12), 4002–4014.
- Lee, H. K., Hsu, A. K., Sajdak, J., *et al.* (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, **14**(6), 1085–1094.
- Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**(1), 1–16.
- Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, **5**(12), R100.
- Makrodimitris, S., Reinders, M. J., and van Ham, R. C. (2020). Metric learning on expression data for gene function prediction. *Bioinformatics*, **36**(4), 1182–1190.
- Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**(2), 026113.
- Pardo-Díaz, J., Poole, P., Beguerisse-Díaz, M., *et al.* (2021a). Generating weighted and thresholded gene coexpression networks using signed distance correlation. *bioRxiv*, <https://doi.org/10.1101/2021.11.15.468627>.
- Pardo-Díaz, J., Bozhilova, L. V., Beguerisse-Díaz, M., *et al.* (2021b). Robust gene coexpression networks using signed distance correlation. *Bioinformatics*. btab041.
- Pini, F., East, A. K., Appia-Ayme, C., *et al.* (2017). Lux bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiology*, **174**(3), 1289–1306.
- Prell, J., Bourdès, A., Karunakaran, R., *et al.* (2009). Pathway of γ -aminobutyrate metabolism in *Rhizobium leguminosarum 3841* and its role in symbiosis. *Journal of Bacteriology*, **191**(7), 2177–2186.
- Ramachandran, V. K., East, A. K., Karunakaran, R., *et al.* (2011). Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biology*, **12**(10), R106.
- Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, **74**(1), 016110.
- Stuart, J. M., Segal, E., Koller, D., *et al.* (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**(5643), 249–255.
- Terpolilli, J. J., Masakapalli, S. K., Karunakaran, R., *et al.* (2016). Lipogenesis and redox balance in nitrogen-fixing pea bacteroids. *Journal of bacteriology*, **198**(20), 2864–2875.
- Thalamuthu, A., Mukhopadhyay, I., Zheng, X., *et al.* (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**(19), 2405–2412.
- Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, **9**(1), 1–12.
- van Noort, V., Snel, B., and Huynen, M. A. (2003). Predicting gene function by conserved co-expression. *TRENDS in Genetics*, **19**(5), 238–242.

- Wang, Y. R., Waterman, M. S., and Huang, H. (2014). Gene coexpression measures in large heterogeneous samples using count statistics. *Proceedings of the National Academy of Sciences*, **111**(46), 16371–16376.
- Waterman, M. (2016). *Getting Outside : A Far-western Childhood*. CreateSpace Independent Publishing Platform.
- Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Senomes*. Taylor & Francis/CRC.
- Weirauch, M. T. (2011). Gene coexpression networks for the analysis of DNA microarray data. *Applied Statistics for Network Biology: Methods in Systems Biology*, **1**, 215–250.
- Wheatley, R. M., Ramachandran, V. K., Geddes, B. A., *et al.* (2017). Role of O₂ in the growth of *Rhizobium leguminosarum* bv. *viciae* 3841 on glucose and succinate. *Journal of Bacteriology*, **199**(1), e00572–16.
- Wheatley, R. M., Ford, B. L., Li, L., *et al.* (2020). Lifestyle adaptations of rhizobium from rhizosphere to symbiosis. *Proceedings of the National Academy of Sciences*, **117**(38), 23823–23834.
- Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics*, **6**(1), 227.
- Xing, W. and Ghorbani, A. (2004). Weighted PageRank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research*, pages 305–314. IEEE.
- Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, **6**, 30750.
- Young, J. P. W., Crossman, L. C., Johnston, A. W., *et al.* (2006). The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biology*, **7**(4), R34.