Gene expression

Using biologically interrelated experiments to identify pathway genes in Arabidopsis

Kyungpil Kim1,†, Keni Jiang2,†, Siew Leng Teng3,†, Lewis J. Feldman2,* and Haiyan Huang1,*

1Division of Biostatistics, University of California, Berkeley, 2Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, 3Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, and 4Department of Statistics, University of California, Berkeley, CA 94720, USA

ABSTRACT

Motivation: Pathway genes are considered as a group of genes that work cooperatively in the same pathway constituting a fundamental functional grouping in a biological process. Identifying pathway genes has been one of the major tasks in understanding biological processes. However, due to the difficulty in characterizing/infering different types of biological gene relationships, as well as several computational issues arising from dealing with high-dimensional biological data, deducing genes in pathways remain challenging.

Results: In this work, we elucidate higher level gene-gene interactions by evaluating the conditional dependencies between genes, i.e. the relationships between genes after removing the influences of a set of previously known pathway genes. These previously known pathway genes serve as seed genes in our model and will guide the detection of other genes involved in the same pathway. The detailed statistical techniques involve the estimation of a precision matrix whose elements are known to be proportional to partial correlations (i.e. conditional dependencies) between genes under appropriate normality assumptions. Likelihood ratio tests on two forms of precision matrices are further performed to see if a candidate pathway gene is conditionally independent of all the previously known pathway genes. When used effectively, this is a promising approach to recover gene relationships that would have otherwise been missed by standard methods. The advantage of the proposed method is demonstrated using both simulation studies and real data sets. We also demonstrated the importance of taking into account experimental dependencies in the simulation and real data studies.

Contact: hhuang@stat.berkeley.edu; ljfeldman@berkeley.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on June 2, 2011; revised on December 21, 2011; accepted on January 17, 2012

1 INTRODUCTION

A biological pathway is a series of chemical reactions that form an integral and critical part of every biological process. Pathway genes, or genes involved in the same biological pathway, constitute a fundamental functional grouping in a biological process. A major task in understanding biological processes is to identify a set of genes in the same biological pathways and elucidating the relationships between them.

Using gene expression data, there have been two popular computational approaches for finding pathway genes: clustering analysis and network models. Clustering analysis uses a co-expression measure to quantify similarities in gene expressions and then assigns similar genes into clusters (Bühlmann et al. 1998). Genes in each cluster are considered to be functionally related, and thus likely to be in the same pathway. This approach works when the pathway genes exhibit strong co-expressions with one another. Network models generally model a pathway as a network, with the genes represented as nodes and the gene relationships represented as edges linking the nodes. The work in Friedman et al. (2004), starting with a full network, a typical pathway can be identified as a connected subgraph and then assigns similar genes into clusters (Eisen et al. 1998).

Despite their appealing features, the approaches described above have limitations. One limitation comes from the high dimensionality of microarray data. The well-known large p, small n problem can result in an unreliable co-expression measure and hence a very high rate of false discoveries in clustering analysis. In the network models, this raises concern about the stability and accuracy of the model inference; it is almost impossible to employ the network models on a genomic scale as the estimation of covariance or its inverse matrices becomes problematic. Although there has been
recent work such as regularized network models to overcome this problem (Schäfer and Strimmer, 2005a), the accuracy of the results remains unclear. As such, in practice, these approaches are usually applied to a rather small number of genes or among a small number of clusters of genes preselected based on some prior knowledge, which as a consequence, makes it difficult for us to explore the whole genomic scale of information.

Another concern is related to the limited biological inference of these approaches. Cluster methods are based on a marginal co-expression measure between two genes independent of other genes. Similarly in a network model using covariance matrix, an edge only connects genes with strong marginal correlations. Such approaches potentially expose us to the risk of missing higher level interactions such as group interactions, i.e. gene A interacts with a group of genes but it does not possess any strong relationship with the individual ones. This group interaction is frequently observed in real biological pathways when a group of genes cooperatively regulate one gene. Using the inverse covariance matrix of genes in a network model has a better hope for detecting such kinds of higher level interactions. The inverse covariance matrix is also known as the precision matrix, whose elements have an interpretation in terms of partial correlations (i.e. the correlation between any two genes conditioned on one or several other genes). However, in the current literature, partial correlation is mostly calculated as the precision matrix, whose elements have an interpretation in terms of partial correlations (i.e. the correlation between any two genes conditioned on one or several other genes). However, in the current literature, partial correlation is mostly calculated on either all the available genes or a more-or-less arbitrary subset of them that likely contain noisy (i.e. non-pathway or biologically unrelated) genes. It is reported that conditioning on all genes simultaneously can introduce spurious dependencies which are not from a direct causal effect or common ancestors (De la Fuente et al., 2004). This problem may be circumvented to some extent by considering lower order partial correlations, e.g. calculating a partial correlation of two genes conditioned on every other individual variables (first-order partial correlation), and on every other two variables (second-order partial correlation) (De la Fuente et al., 2004; Magwene and Kim, 2004; Wille et al., 2004; Wille and Bühlmann, 2006). However, one concern on lower order partial correlation is its insensitivity for inferring higher level gene associations such as group interactions. More importantly, if the conditioned genes are biologically unrelated, the corresponding conditional dependence properties would be difficult to interpret and verifying the biological relevance of the recovered networks becomes challenging. Further discussions on the adverse effects of conditioning on noisy genes are given in Sections 2 and 3.

In this article, we introduce a new pathway gene search algorithm, designed based on evaluating partial correlations between genes, for a particular biological pathway of interest. The motivation of using partial correlation is based on its ability to detect complex gene relationships under appropriate normality assumptions of the data: (i) a strong partial correlation between two genes suggests a direct interaction despite a weak marginal correlation; (ii) a negligible partial correlation suggests no direct relationship after removing influences from other genes and the two genes are conditionally independent. To overcome the concerns and limitations of current methods for using partial correlations, we require a few (e.g. 3–5) preselected biologically related pathway genes, upon which the partial correlation is conditioned on, to guide the search. Specifically, we perform the likelihood ratio tests to see if a candidate gene is conditionally independent of all the preselected known pathway genes. The requirement of pre-known pathway genes seems a limitation of our approach. However, by incorporating this small amount of biological knowledge, huge advantages on biological inference can be gained and false positive discoveries have been reduced dramatically (Section 3). Furthermore, by conditioning on preselected pathway genes, the resulting partial correlation coefficients can be directly interpreted as a similarity measure to the considered pathway. In addition to suggesting satisfying mathematical and biological properties, the proposed approach is also advantageous computationally since we only need to estimate a moderate dimensional precision matrix once for each candidate gene.

Moreover, we also take into account the presence of experiment dependencies in the gene expression data when estimating a precision matrix [elements in a precision matrix are proportional to partial correlations between genes (Schäfer and Strimmer, 2005a)]. In current studies of gene relationships, the presence of expression dependencies attributable to the biologically interrelated experiments has been widely ignored. When unaccounted for these (experiment) dependencies can result in inaccurate inferences of functional gene relationships and hence incorrect biological conclusions (Saw and Huang, 2006). Our simulation and real data study supports this conclusion and confirms that considering those dependencies indeed plays a critical role in correctly inferring pathway gene relationships.

The rest of the article is organized as follows. In the Section 2, we introduce the mathematical and statistical background of our approach. In Section 3, we demonstrate the model validity and evaluate the performance of our approach using extensive computer simulations and real data applications. In real data applications, we apply it to genomic scale Arabidopsis thaliana datasets obtained from four different types of environmental stresses (oxidation, wounding, UV-B light and drought). We examine the effects of these stresses by focusing on the genes associated with the glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways. Finally, we discuss the advantages as well as potential drawbacks of our framework and consider further directions for research.

2 METHODS

As we discussed earlier, in contrast to marginal calculation (e.g. the Pearson correlation), partial correlation can work as a more effective tool for inferring complex gene interactions in pathways when it is properly computed. Below, we first provide a brief review on the concept of partial correlation, followed by a detailed description of a new search strategy, designed based on a likelihood ratio test on partial correlations, for finding pathway genes.

2.1 Partial correlation

When an expression matrix (with genes in rows and experimental conditions in columns) is multivariate normally distributed, standard graphical model theory (Edwards, 1995) shows that the partial correlation between genes can be equivalently represented by the corresponding elements in the precision matrix \( (\Sigma^{-1})^{-1} \), where \( \Sigma \) is the covariance matrix. That is, for a set of genes \( G \), the partial correlation between genes \( i \) and \( j \) can be expressed as

\[
\rho_{ij} = \text{cor}(\{X_i | W \cup \{j\}) = \frac{\sqrt{\omega_i} \sqrt{\omega_j}}{1, i, j} \frac{\Sigma^{-1}(i,j)}{(1)}
\]

where \( \omega_i \) are elements in the inverse gene covariance (or inverse gene correlation) matrix \( \Sigma^{-1} \). With the normality assumption on expression measurements, when \( \rho_{ij} \) vanishes, two genes \( i \) and \( j \) are conditionally independent given the remaining genes.
2.2 Method motivation

A negligible element in the precision matrix suggests conditional independence between two genes. This motivates us to use precision matrix as a key component in our method for detecting higher level gene interactions, e.g. group gene interactions in a pathway. However, the successful use of partial correlation highly relies on two issues. One issue is about the selection of a proper set of genes upon which the correlation is conditioned on, i.e. \( W(x) \) in formula (1). When this set of genes contains noisy (i.e. non-pathway) genes, the derived partial correlation would be unreliable for detecting gene relationships. We can see this explicitly through a linear regression interpretation of partial correlation: the partial correlation \( \rho_{ij} \) between gene \( i \) and \( j \) conditioned on a set of genes \( Z \) is simply the correlation \( \text{corr}(x_i, x_j) \) of the residuals \( e_i \) and \( e_j \) resulting from linearly regressing gene \( i \) and gene \( j \) against the genes in \( Z \), respectively. Assume we have pre-known pathway genes \( x_i \) and \( y_i \), and a non-pathway gene \( h \) that is independent of genes \( x \) and \( y \) in \( Z \). Now we consider two candidate genes \( w = x + y \) and \( v = h(x+y) + h \) (note that these two equations only represent the expression relationship between the genes), where \( h \) is small and close to 0. Clearly, \( a \) is more likely to be a pathway gene due to its direct and strong relationship with two pre-known pathway genes \( x \) and \( y \), while \( v \) is more likely to be a non-pathway gene since it is almost a replicate of the non-pathway gene \( h \). However, the partial correlations \( \text{corr}(x_i, x_j|Z) = \text{corr}(x_i, y_j|Z) = \text{corr}(v_i, v_j|Z) = 1 \), showing no advantages of gene \( a \) over gene \( v \) for their partial correlations with the pathways genes \( x \) and \( y \). This undesired performance is due to the inclusion of noisy genes in the gene set \( Z \) upon which the partial correlation was computed. Recognizing this, we decide to build up our approach by conditioning only on a small set of pre-known pathway genes to reduce noise in partial correlation estimation. We call this set of pre-known pathway genes as seed genes. Though the requirement of seed genes seems a limitation, only 3-5 seed genes are really needed for our method to run and generate reliable results. In brief, by incorporating a small amount of prior biological information, we can gain huge advantages in detecting genes involved in a particular pathway (Section 3). Furthermore, in Section 3, by using both simulation and real data, we additionally demonstrate the adverse effects of having noisy genes in the set of seed genes in detecting pathways genes. The other issue critical to the proper use of partial correlation is on the estimation of gene precision matrix (see Formula (1)). Given a gene expression matrix with genes in rows and experiments in columns, an effective estimation of gene precision matrix is challenging especially when there are dependencies (or when the row-wise and column-wise dependencies co-exist) in the original gene expression. Experiment dependencies can be defined as the dependencies in gene expression between experiments due to the similar or related cellular states induced by the experiments (Teng and Huang 2009). Such dependencies cause dependent elements in a gene expression vector. When unaccounted for, they can result in inaccurate inferences of gene relationships, and hence incorrect biological conclusions. To take into account the experiment dependencies in partial correlation estimation, we adapt a model and an estimation procedure, named Known from Teng and Huang (2009), for inferring gene correlation matrix when there are both the gene-wise and experiment-wise dependencies in the gene expression matrix. The main aspect of the framework is the use of a Kronecker product covariance matrix to model the gene–experiment interactions. The Known estimation is mainly achieved by an iterative estimation of the two-covariance matrices: one covariance matrix is estimated through a weighted correlation formula assuming the other covariance matrix is known. In addition, a row subsampling technique (to enable a comparable number of rows and columns in estimation) and a covariance shrinkage technique (to stabilize the estimated covariance matrices) are employed to ensure a robust estimation. Compared with the Pearson coefficient, the Known correlation has a smaller estimation variance when experiment dependencies exist. More details of incorporating Known in our estimation procedure are presented in next section.

2.3 Likelihood ratio tests for pathway gene search

Let \( S = \{x_1, \ldots, x_k\} \) denote the set of seed genes for a pathway of interest and \( G \) denote the set of all genes whose expression measurements in \( T \) experiments (each experiment may have > 1 replicates) are available. Usually \( |G| \gg |S| = k \) and \( T > |S| \). Motivated by the arguments in the above section, we formulate a searching strategy, based on performing likelihood ratio tests, for pathway genes as follows.

(1) We first estimate the experiment correlation matrix \( \Sigma_e^T \) using the Known R package provided by Teng and Huang (2009). The input data are the expression measurements of the \( G \) genes in \( T \) experiments, and there are > 1 replicates sampled for each experiment. To generate expression matrices, we randomly choose one replicate from each experiment to compose a sample matrix \( X_{0,t} \) of dimension \( |G| \times T \) and by repeating this process, we generate \( B \) sample matrices \( X_1, \ldots, X_B \) with \( B \) large enough. By the model in Teng and Huang (2009), \( X_0 \) is considered to be generated from a multivariate normal distribution with mean \( M \) (a matrix of dimension \( |G| \times T \)) and a covariance matrix \( \Sigma_e^o \otimes \Sigma_o^T \), where \( \Sigma_e^o \) represents the gene covariance matrix and \( \Sigma_o^T \) is the experiment correlation matrix. The output of the Known R package is the estimated \( M \) and \( \Sigma_o^T \), denoted as \( \hat{M} \) and \( \hat{\Sigma}_o^T \), by an iterative estimation procedure. More details on the Known estimation procedure can be found in Teng and Huang (2009).

(2) For a candidate gene \( g_j \in G \backslash S \) (the set of \( k \) seed genes), we estimate the gene covariance matrix for genes \( g_j \) in \( S \), e.g. by \( \Sigma_j = \sum_{k=1}^{B} X_j \), where

\[
\Sigma_j = \frac{1}{B} \sum_{k=1}^{B} (X_j - \hat{M}) (X_j - \hat{M})^T
\]

\( \hat{M} \) represents one of the sample matrices of dimension \( (k+1) \times T \) constructed by bootstrapping the replicates of each experiment for the expression measurements of the genes in \( S \cup g_j \), the first \( k \) rows correspond to the \( k \) seed genes and the row \( k+1 \) corresponds to the candidate gene \( g_j \).

(3) Obtain the precision matrix, \( \Sigma_j^{-1} = (\Sigma_j)^{-1} \). Note that we usually require \( T > 1.5M \). When \( \Sigma_j \) is invertible, we use its pseudo inverse. \( \Sigma_j^{-1} \) will be used as an approximate Maximum Likelihood Estimate (MLE) of the precision matrix under the alternative model in Equation (3). We further write

\[
\Sigma_j^{-1} = \begin{bmatrix}
\hat{a}_{11} & \cdots & \hat{a}_{1,k+1} \\
\cdots & \cdots & \cdots \\
\hat{a}_{k+1,1} & \cdots & \hat{a}_{k+1,k+1}
\end{bmatrix}
\]

where \( \hat{a}_{ij} = a_{ij} \) for \( i = 1, \ldots, k+1 \) and \( j = 1, \ldots, k+1 \).

(4) Obtain matrix \( \Sigma_j^{*} \) from \( \Sigma_j^{-1} \) by replacing the offdiagonal elements in the bottom row and rightmost column of \( \Sigma_j^{-1} \) by zeros. That is,

\[
\Sigma_j^{*} = \begin{bmatrix}
\hat{a}_{11} & \cdots & \hat{a}_{1,k+1} & 0 \\
\cdots & \cdots & \cdots & \cdots \\
\hat{a}_{k+1,1} & \cdots & \hat{a}_{k+1,k+1} & 0 \\
0 & \cdots & 0 & 0
\end{bmatrix}
\]

\( \Sigma_j^{*} \) will be used as an approximate MLE of the precision matrix under the null hypothesis in Equation (3).

(5) Perform the following hypothesis test \( H_0: \Sigma_j \in \Theta_0 \) vs \( H_1: \Sigma_j \in \Theta_1 \), where \( \Theta_0 \) is the collection of precision matrices (for the genes in \( S \cup g_j \)) with zero offdiagonal elements in the bottom row and rightmost column, and \( \Theta_1 \) is the collection of all possible precision matrices. The null hypothesis assumes conditional independence between the candidate gene and each of the seed genes given all other seed genes. Then the test statistic is

\[
-2 \log LR^T = -2 \log \frac{\text{det}(\Sigma_j^{*})}{\text{det}(\Sigma_j)} = -2 \log \left( \prod_{i=1}^{k+1} \prod_{j=i+1}^{k+1} \lambda_i \right) = -2 \left( \sum_{i,j=1}^{k+1} \lambda_i \lambda_j - \sum_{i,j=1}^{k+1} \lambda_i \lambda_j - h(\Sigma_j; X_1, \ldots, X_B, \hat{M}) \right)
\]
We evaluate the performance of the proposed method in identifying repeatedly run genes or on which expression datasets to use, an optional method is to genes using statistical thresholds (see Supplementary Material for the list, we can decide how many of them should be declared as pathway candidates) this way would be robust against the change of data and the choice of seed genes. If there are questions on which known pathway genes to include as seed genes, the corresponding elements in a precision matrix will be close to zeros (i.e. null is true and the test statistic will be small). In contrast, if a seed gene has no relationship with the pathway genes, the test statistic will be large and declared as significant under the test.

(vi) Repeat steps (iii)-(v) for all candidate genes in G/S. Given the test statistic values for all the candidate genes, we rank them in decreasing order.

It is a natural interpretation that the higher a candidate gene is ranked in the list, the more likely that gene is associated with the seed genes. Based on the list, we can decide on which genes should be declared as pathway genes using statistical thresholds (see Supplementary Material for details). The candidate pathway genes identified this way would be robust against the change of data and the choice of seed genes.

2.4 Other methods for comparison

For performance comparison, four additional methods, pwsrc.null, pearson.mean, pearson.max and GLM are considered. The first pwsrc.null is designed by replacing $\Sigma$ in Equation (2) with the identity matrix to represent the case ignoring experiment dependencies. pearson.mean and pearson.max adopt the Pearson correlation as a distance measure. Specifically, they calculate pair-wise correlation coefficients between each candidate gene and the seed genes and take either mean (pearson.mean) or maximum (pearson.max) of them. GLM adopts the regression model of the candidate gene $g$, on the seed genes $S$ as follows:

$$E_g = \alpha + \sum_{j=1}^{m} \alpha_j S_j + \epsilon,$$

where $\epsilon$ is assumed to be normally distributed with zero mean. Since a negligible residue implies a possible interaction between the candidate gene and the seed genes, naturally we can use the residuals as our test statistics (all the genes are scaled to have unit norm before doing regression analysis). For a fair comparison with our method, the experiment dependencies in the gene expression are removed by projecting the data matrix onto the eigenspace of $\hat{\Sigma}$ by $X = (X - \mu) / (\hat{\Sigma}^{-1} \hat{\Sigma}^{1/2})$.

3 RESULTS

We evaluate the performance of the proposed method in identifying pathway genes using simulation data and real microarray datasets. In both studies, we calculated precision = TP/(TP+FP) and recall = TP/(TP+FN) to assess the results from our approach and several other methods mentioned in Section 2. Here TP is the number of true positive findings of pathway genes, FP is the number of false positives and FN is the number of false negatives. Note that precision and recall are popular measures for evaluation of classification performance. In the context of this study, they can be regarded as a measure of exactness and completeness of our pathway gene searching results, respectively. In the real data analysis, we will present the strength and usefulness of our approach as a tool for identifying pathway genes, particularly in glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways. In our study, the pathway genes are defined as composed of structural genes that encode an enzyme, whereas regulator genes are defined as genes controlling the expression of the structural genes.

3.1 Simulation study

We simulate a microarray dataset consisting of 500 genes and 30 experiments, with 5 replicates for each experiment. To make the approach more realistic, we introduce experiment dependencies, multiple distinct pathways and some random noise into the simulated data. The simulation parameters are as follows:

(i) Experiment correlation matrix, $\Sigma^E$. This matrix characterizes the experiment dependencies. For illustrative purposes, we set the experiment correlation matrix to have various dependencies such as 10, 33, 50 and 67%. In the case of a 33% dependency, for example, ~33% of the experiments have high dependencies while the remaining experiments are uncorrelated with one another, i.e. the first $10 \times 10$ elements in $\Sigma^E$ lie between 0.5 and 0.6, with the rest being zeros. Diagonals on $\Sigma^E$ are set to 1. Figure 1 shows the heatmaps for three of the four experiment correlation matrices mentioned above.

(ii) Gene covariance matrix, $\Sigma^G$. This matrix characterizes the gene dependencies among one another. An illustrative example, we introduce two distinct pathways with 15 genes in each pathway; genes in the same pathway have high correlation while genes not in the same pathway are uncorrelated. Specifically, in each pathway the first four genes designated to be seed genes show high correlation (correlation coefficient changes between 0.5 and 0.6) between each other. The remaining 11 genes are separated into three subgroups and are designed to have high correlation with 1, 2 or 4 of the seed genes, respectively, and low correlation with the others (correlation coefficient changes between 0.1 and 0.2).

The simulated data is generated as follows. First, we generate a $500 \times 30$ gene expression matrix $X$, with vec($X^T$), from a multivariate normal distribution with mean $\mu$ (zero matrix) and a covariance matrix $\Sigma^G \otimes \Sigma^E$. To make the pathway genes more realistic, for each pathway two randomly chosen genes in each subgroup are linearly combined to make a new pathway gene. The same procedure generated all the final 11 pathway genes for each pathway (replacing the original 11 pathway genes generated above). Using the final $500 \times 30$ gene expression matrix, we add random noise with a small SD (e.g. 0.01) to each column (i.e. experiment) to generate the 5 replicates for each experiment. Repeating this process, we generate 1000 simulation datasets.

In this analysis, we compare our approach to that of others and evaluate the performance using precision and recall measures. All the approaches are implemented as follows: given seed genes, run the pathway search algorithms as described in Section 2 and rank the genes by their measured relationships to the seed genes. Calculate precision and recall for the top $p$ (i.e. $n = 1, \ldots, 15$) genes.

As this is a simulation study and we know the true experiment correlation matrix, we add one more method pwsrc.true into the comparison. The only difference between pwsrc.true and pwsrc.knorm is that pwsrc.true uses the true experiment correlation matrix ($\Sigma^E_{true}$) instead of the estimated one in Equation (2). We denote the estimated correlation matrix used by our method as $\Sigma^E_{true}$ for clarity. The results are summarized in Figure 2. When the dependencies among experiments are low, pwsrc.knorm performs worse than pwsrc.null. However, this performance discrepancy becomes smaller as the experiment dependency increases and finally pwsrc.knorm outperforms pwsrc.null when the experiment dependency exceeds 33%. This situation can be easily understood in Figure 1. When the dependencies among experiments are low,
We next test the validity of our approach by applying it to biological groups that have applied transcriptome co-expression to analyze the pathways that are induced by broad environmental stresses, and which are extensively characterized. A considerable number of genes in glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways have been identified in Arabidopsis (Arabidopsis thaliana), thus providing us with a rich source of data for validating our approach.

To determine the importance of the seed gene quality, we added two randomly chosen, non-pathway genes into the seed-gene-set, which is originally composed of four pathway genes. The results are summarized in Supplementary Figure S1. Regardless of the experiment dependencies, the performance of \( \text{pearson.max} \) becomes worse when the seed-gene-set contains noisy genes.

### 3.2 Application to real datasets

We next test the validity of our approach by applying it to biological pathways composed of genes that are known to operate in tandem. For this test set, we selected two secondary metabolic pathways from the model plant Arabidopsis thaliana: the pathway leading to GSLs, sulfur-rich amino acid-containing compounds which become active in response to tissue damage, and believed to offer a protective function (Hammond-Kosack and Jong 2001; Banderby et al. 2010). Verkerk et al. (2004) and Yun and Chud (2004) identified the pathway leading to flavonoids, compounds of diverse biological activities such as anti-oxidants, functioning in UV protection, in defense, in auxin transport inhibition, and in flower colouring (Cachon et al. 2005; Yano and Koyama 2006; Taylor and Grootewold 2008; Woo et al. 2005). In Arabidopsis, the regulators and structural genes in glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways have been extensively characterized. A considerable number of genes in both pathways are induced by broad environmental stresses, and regulated at the transcriptional level. Furthermore, several research groups have applied transcriptome co-expression to analyze the pathway genes.

![Pathway Genes](https://academic.oup.com/bioinformatics/article-abstract/28/6/815/311628/819)

**Fig. 1.** Heatmaps of (Top) true and (Bottom) estimated experiment correlation matrices of the simulation datasets having different experiment dependencies (10, 33 and 67%).

**Fig. 2.** Graphical summary of the simulation study. Simulation datasets are generated with different experiment dependencies (a) 10%, (b) 33%, (c) 50% and (d) 67%. For each plot, precision and recall are calculated from the top \( n \) genes in the list obtained by five different methods.

**Table 1.** A summary of the experimental sets used in our study. Two secondary metabolic pathways (GSL and FB) are compared. The known pathway genes are selected from the list generated by our approach, and demonstrated that our approach is much more effective than any previous approaches for detecting the known pathway genes. Finally, having investigated the validity of this approach, and demonstrating that our approach is much more effective than any previous approaches for detecting the known pathway genes, we asked whether we could identify other possible candidate pathway (new) genes. Initially, we investigated the two pathways genes set in shoot tissue only, but then later expanded the study to include root tissue.

**3.2.1 Studies on GSL pathway** Based on an extensive literature search, we determined 64 genes that can be associated with the GSL pathway (Supplementary Table S2). These 64 genes include, in addition to core genes involved in GSL biosynthesis, regulators of this biosynthesis, early steps of side chain elongation/modification and late steps of catabolism (Supplementary Fig. S2 in detail). For our study, two seed-gene-sets are proposed: (i) seed-gene-set I: AT5G08090 (ATR1), AT4G39950 (CYP79B2), AT2G20610 (SUR1), AT4G31500 (CYP83B1) and (ii) seed-gene-set II: AT5G08090 (ATR1), AT5G07690 (MYB29), AT5G61420 (MYB28), AT4G39940 (AKN2). In seed-gene-set I, only ATR1 encodes a transcription factor
(TF), whereas three other genes encode enzymes. In contrast to seed-gene-set I (composed of the core pathway genes), seed-gene-set II is composed of four regulatory genes (Supplementary Fig. S2).

Using two seed-gene-sets, we first analyzed only the shoot tissue data from tissues subjected to oxidative stress. This dataset is composed of 22,810 genes and 13 experiments with two biological replicates for each experiment. The number of identified GSL pathway genes is summarized in Table 2(a)–(b) for the top 10, 20, 30 and 50 genes from the list obtained by pwsrc.knorm, pwsrc.null, pearson.mean, pearson.max and GLM. With seed-gene-set I in Table 2(a), pwsrc.knorm works best, finding 4, 6, 7 and 8 pathway genes out of the top 10, 20, 30 and 50 genes, respectively. With seed-gene-set II in Table 2(b), a significant increase is observed in the number of identified pathway genes, especially for pwsrc.knorm. For example, among the top 30 genes in the list, pwsrc.knorm finds 7 more pathway genes, while pwsrc.null, pearson.mean, pearson.max and GLM find 2, 2, 1 and 4 more genes compared to Table 2(a), respectively. This increase demonstrates that seed-gene-set II indeed carries more influential information than seed-gene-set I, which enables us to examine the GSL pathway more thoroughly. Furthermore, our method pushes the pathway genes to rank higher positions in the list so that the final precision becomes 60, 55 and 47%, respectively, for the top 10, 20 and 30 genes.

Next, the dataset is expanded to include the root tissue as well, so now the dataset consists of 26 experiments with two replicates each. Again, the combined dataset is analyzed with the two seed-gene-sets I and II. As expected, the performance of GLM and GLM finds no pathway genes in this data, which demonstrates the importance of designing the seed-gene-set. When the seed set is exclusive to each other, and thus there is no overlap in the genes of the datasets from plants subjected to UV-B light and drought stresses are about the same, compared to the previous results with oxidative and wounding stresses (summarized in Supplementary Tables S3 and S4, respectively, and the corresponding graphical summaries are also provided in Supplementary Figs S5 and S6).

3.2.2 Studies of the FB pathway

The flavonoid pathway is derived from the upstream phenylpropanoid pathway, beginning at coumaroyl-CoA (Supplementary Fig. S7). Based on an extensive literature search, we found that at least 26 genes can be associated with the FB pathway (Supplementary Table S5). Genes encoding enzymes in this pathway are regulated by at least 12 TFs belonging to different families, including bZIP WD40, WRKY, MADS-box, R2R3-MYB, and the basic helix–loop–helix (bHLH) family (Yonekura-Sakakibara et al. 2008). It is also worth noting that the genes we considered for the two pathways (GSL and flavonoid) are exclusive to each other, and thus there is no overlap in the genes of the pathways we consider.

It is reported that structural genes (encoding enzymes) in the FB pathway are regulated at the transcriptional level, suggesting that the regulation genes would be good candidates as seed genes, as indicated by the result of GSL pathway study in Section 3.2.1.
Then we selected two different seed-gene-sets from four different types of TFs [AT4G09820 (TT8), AT5G23260 (TT16), AT5G24520 (TTG1), AT5G08640 (FLS)] and one structural gene [AT5G07990 (TTG2), AT5G24520 (TTG2)], and two of GSL pathway genes. As summarized in Table 3, the number of identified GSL pathway genes in the A. thaliana microarray dataset from tissues subjected to oxidative stress using (a) shoot tissue only, seed-gene-set I; (b) shoot tissue only, seed-gene-set II; (c) shoot and root tissues, seed-gene-set I; (d) shoot and root tissues, seed-gene-set II:

<table>
<thead>
<tr>
<th>Type</th>
<th>pwsrc.knorm</th>
<th>pwsrc.null</th>
<th>pearson.mean</th>
<th>pearson.max</th>
<th>GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>10 4 0 2 1 0</td>
<td>20 6 1 3 4 0</td>
<td>30 7 1 3 5 0</td>
<td>50 8 1 4 7 0</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>10 6 0 4 2 2</td>
<td>20 11 3 4 4 3</td>
<td>30 14 3 5 6 4</td>
<td>50 14 5 8 8 5</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>10 9 0 3 0 3</td>
<td>20 11 0 3 0 3</td>
<td>30 12 0 3 1 7</td>
<td>50 12 0 3 1 7</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td>10 6 0 5 1 4</td>
<td>20 10 0 7 1 5</td>
<td>30 13 0 9 1 9</td>
<td>50 19 0 9 1 12</td>
<td></td>
</tr>
</tbody>
</table>

Then we selected two different seed-gene-sets from four different types of TFs [AT4G09820 (TT8), AT5G23260 (TT16), AT5G24520 (TTG1), AT5G08640 (FLS)] and one structural gene [AT5G07990 (TTG2), AT5G24520 (TTG2)], and one structural gene [AT5G23260 (TT16), AT5G24520 (TTG1), AT5G08640 (FLS)] and (ii) seed-gene-set IV in the A. thaliana microarray dataset from shoot and root tissues subjected to drought stress:

<table>
<thead>
<tr>
<th>Top pwsrc.knorm pwsrc.null pearson.mean pearson.max GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) 10 6 (2) 0 0 2 4 (1)</td>
</tr>
<tr>
<td>(b) 10 6 (2) 2 1 4 5 (1)</td>
</tr>
<tr>
<td>(c) 10 9 (3) 2 2 4 6 (2)</td>
</tr>
<tr>
<td>(d) 10 9 (3) 2 1 4 6 (2)</td>
</tr>
</tbody>
</table>

For complete results, see Supplementary Table S6. The number of identified genes from phenylpropanoid pathways is designated in the parenthesis adjacent to the total number of identified genes.
Supplementary Table S8, the number of pathway genes identified decreased which implies that the number of pathway gene search becomes less efficient when the seed-gene-set contains noisy genes not biologically related to the pathway of our interest.

Finally, we compared our results with other literatures on the discovery of GSL and FB pathway genes. See the Supplementary Material for details.

4 DISCUSSION

We have proposed a novel approach to identify genes associated with a pathway specified by a set of seed genes. This approach considers the space of pathway genes as a span generated by its pathway genes and uses partial correlation as a distance measure to determine genes interacting with the previously identified pathway genes. This approach differs from many existing approaches in the following aspects: (i) it uses the partial correlation conditioned upon identified pathway genes, not on all genes; (ii) it enables us to identify genes having higher level interaction (i.e. group interaction) although their pair-wise marginal correlations are weak; (iii) it considers experiment dependencies when inferring gene relationships; (iv) its computational workload is less demanding.

The first aspect above implies our method is pathway specific. It focuses its search only on a particular pathway among all existing multiple (unknown) pathways in a dataset. This is a limitation of our approach. But we note that this specific search has shown huge advantages in reducing false positive discoveries in both our simulation and real data studies, and has also led to a deeper and insightful biological interpretation of the results. This approach can potentially be extended to the situation that the seed genes or target pathways are not available, if the seed genes for different pathways can be originated from analysis of other sources of biological data.

Although our approach has yielded encouraging biological results in a real dataset application, there is still room for further improvement, including exploration of properties of this approach to answer questions like, ‘What are the biological properties of the identified genes?’ and, ‘How reliable is the set of identified genes?’.

Further biological understanding of the identified pathway genes would give us deeper insights into the biological process under consideration.

Funding: National Institutes of Health, National Eye Institute (R21EY019094 to H.H.) in part.

Conflict of Interest: none declared.

REFERENCES


