Information-incorporated Gene Network Construction with FDR Control

Hao Wang,1 Yumou Qiu,1,* Hongqing Guo,2 Yanhai Yin2 and Peng Liu1,∗

1Department of Statistics, Iowa State University, Ames, 50010, Iowa, USA and 2Department of Genetics, Development and Cell Biology, Iowa State University, Ames, 50010, Iowa, USA

*Corresponding author. yumouqiu@iastate.edu, pliu@iastate.edu

Abstract

Motivation: Large-scale gene expression studies allow gene network construction to uncover associations among genes. To study direct associations among genes, partial correlation-based networks are preferred over marginal correlations. However, FDR control for partial correlation-based network construction is not well-studied. In addition, currently available partial correlation-based methods cannot take existing biological knowledge to help network construction while controlling FDR.

Results: In this paper, we propose a method called Partial Correlation Graph with Information Incorporation (PCGII). PCGII estimates partial correlations between each pair of genes by regularized node-wise regression that can incorporate prior knowledge while controlling the effects of all other genes. It handles high-dimensional data where the number of genes can be much larger than the sample size and controls FDR at the same time. We compare PCGII with several existing approaches through extensive simulation studies and demonstrate that PCGII has better FDR control and higher power. We apply PCGII to a plant gene expression dataset where it recovers confirmed regulatory relationships and a hub node, as well as several direct associations that shed light on potential functional relationships in the system. We also introduce a method to supplement observed data with a pseudogene to apply PCGII when no prior information is available, which also allows checking FDR control and power for real data analysis.

Availability: R package is freely available for download at https://cran.r-project.org/package=PCGII.

Contact: pliu@iastate.edu or yumouqiu@iastate.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Key words: FDR, Gene Network, Gaussian Graphic Model, Information Incorporation

1. Introduction

The advent of microarray and RNA-sequencing technologies allows gene expression studies for the whole transcriptome and offers the opportunity to explore functional relationships among all genes using network analysis (Barabasi and Oltvai, 2004; Boccaletti, 2010; Cowen et al., 2017). The key to building a gene network involves identifying the connection (edges) among genes (nodes) to describe how genes are related. Various methods have been developed to construct gene networks. Some widely applied approaches, such as WGCNA (Zhang and Horvath, 2005) and ARACNE (Margolin et al., 2006), apply marginal correlations that measure whether the expression levels of two genes simultaneously change in the same or opposite directions, regardless of other genes. Some approaches, such as Gaussian graphical models (GGMs), are based on conditional correlations that measure the correlations between genes while controlling for the effects of other genes. Consider a scenario where gene A inhibits gene B and gene C. Even without a direct interaction between B and C, the marginal association between B and C is positive due to the shared inhibitor, gene A. Such marginal correlations can be misleading in network analyses focusing on direct associations (Zuo et al., 2014). In contrast, conditional correlations account for confounding effects and identify connections between genes that others cannot explain. GGMs construct networks based on partial correlation, a type of conditional correlations, and can identify direct associations among genes. In addition, GGM offers advantages in terms of well-established statistical properties and computational efficiency. Hence, this paper focuses on networks built by GGMs.

Building gene networks by GGM faces a challenge arising from the extensive number of nodes. When the number of genes is much larger than the sample size, classical computations of GGM are not applicable. Sparse GGMs have been developed (Schäfer and Strimmer, 2005; Meinshausen et al., 2006; Friedman et al., 2008;
Peng et al. (2009) to identify the network in high-dimensional settings, but none of these methods control false discovery rate (FDR). Given the substantial number of possible connections that need to be evaluated, controlling FDR is essential to generate a reliable network while minimizing false positives. Recently, three approaches (Ren et al., 2015; Bernai et al., 2019; Qiu and Zhou, 2020) have been proposed to recover a large-scale network while controlling FDR, but their performances need more investigation.

For well-studied organisms, some knowledge about direct associations among genes is available before constructing networks. Approaches capable of integrating such existing biological knowledge are particularly appealing. Benedetti et al. (2020) choose network sparsity by comparing marginal correlation-based networks to the available prior knowledge. Yi et al. (2021) incorporates information into data-driven GGM to enhance power. Although these methods allow information incorporation in network analysis, they do not have a procedure to control FDR.

To the best of our knowledge, there are no available methods to build gene networks based on conditional correlations while incorporating biological information and controlling FDR at the same time. To fill the gap, we propose a method called Partial Correlation Graph with Information Incorporation (PCGII), which determines direct associations based on partial correlations, takes prior information while building gene networks, and controls FDR. We estimate partial correlations with estimated coefficients and residuals from regularized node-wise regressions. Given biological information, we leave out penalization on those known direct associations, resulting in more accurate coefficient estimates. We provide a de-biased estimator for partial correlations and propose a method to control FDR to recover nonzero partial correlations, leading to the constructed network. We demonstrate that PCGII controls FDR even when the prior information includes false positives, and it achieves higher power than the methods without utilizing prior biological information.

We apply PCGII to a plant study of the endoplasmic reticulum (ER) body formation-associated genes (Wang et al., 2017) and identify the confirmed regulatory relationship and a hub gene that functionally interacts with others. PCGII also discovers connections among some important genes, indicating more direct gene-gene associations related to ER body formation than currently known. We also develop and apply an analysis technique on an E. coli dataset, illustrating network construction and evaluation when prior knowledge is not available.

The paper is organized as follows. We describe our proposed PCGII procedure in section 2. In section 3, we present simulation studies under different network settings to compare the performance of PCGII with available methods that build partial correlation-based networks and control FDR. Results of applying PCGII to two real datasets are presented in section 4 followed by a discussion in section 5.

2. Methods

Suppose a gene expression dataset has $p$ genes and $n$ samples. Let $Y_{i,j}$ denote the observed expression level of the $j$th gene from the $i$th sample, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Assume $Y_i = (Y_{i,1}, \ldots, Y_{i,p})'$ to be independent and identically distributed (i.i.d.) random vectors with mean $\mu$ and covariance matrix $\Sigma = (\sigma_{j,j} \sigma_{j,j}^T)_{p \times p}$. Let $\Omega = \Sigma^{-1} = (\omega_{j,j} \omega_{j,j}^T)_{p \times p}$ be the corresponding precision matrix and $\rho_{j,j} = -\omega_{j,j} / \sqrt{\omega_{j,j} \omega_{j,j}^T}$ be the partial correlation between the $j$th and $j$th nodes (genes).

2.1. Node-wise Regression

Lemma 1 in Peng et al. (2009) established the connection between a partial correlation matrix and the variance-covariance matrix of the errors from $p$ node-wide regressions. For each of the $p$ variables, the node-wise regression of the $j$th variable takes the form

$$Y_{i,j} = \alpha_{j,0} + \sum_{j \neq j'} \alpha_{j,j'} Y_{i,j'} + \xi_{j,j'},$$

where $\xi_{j,j'}$ is un-correlated with $Y_{i,j'}$ for $j \neq j'$. Let $V = (v_{j,j'})$ be the variance-covariance matrix of $\xi$, $\xi = (\xi_{1,1}, \ldots, \xi_{p,p})'$. The partial correlation between $Y_{i,j}$ and $Y_{i,j'}$ can be written as

$$\rho_{j,j'} = \frac{v_{j,j'}}{\sqrt{v_{j,j} \times v_{j,j'}}}.$$  

As the dimension of genes is typically high, lasso estimation cannot be used to fit the node-wise regressions in Equation (1) by the minimization program:

$$\argmin_{\alpha_{j,0}, \alpha_{j,j'}, \ldots, \alpha_{j,j'}} \sum_{i=1}^{n} \left( Y_{i,j} - \alpha_{j,0} - \sum_{j \neq j'} \alpha_{j,j'} Y_{i,j'} \right)^2 + \lambda_j |\alpha_{j,0}|,$$

where $\alpha_{j,j'} = (\alpha_{j,1}, \ldots, \alpha_{j,p})'$ with $\alpha_{j,j'}$ set as $-1$, and $\lambda_j$ is the penalty parameter (Liu, 2013). Let $\hat{\alpha}_{j,j'} = (\hat{\alpha}_{j,1}, \ldots, \hat{\alpha}_{j,p})'$ be the lasso estimate of $\alpha_{j,j'}$ and $\hat{\xi}_i = (\hat{\xi}_{i,1}, \ldots, \hat{\xi}_{i,p})'$ be the residuals from the lasso fit. Liu (2013) and Qiu and Zhou (2020) showed the sample covariance of residuals $\{(\hat{\xi}_i)_i \}_{i=1}^n$ is biased for estimating $v_{j,j'}$, and proposed the bias corrected estimator

$$\hat{v}_{j,j'} = -\frac{1}{n} \sum_{i=1}^n (\hat{\xi}_{i,j} \hat{\xi}_{j,j'} + \hat{\alpha}_{j,j'} \hat{\alpha}_{j,j'}^2 + \hat{\alpha}_{j,j'}^2 e_{j,j'}^2).$$

Note that, since $\hat{\alpha}_{j,j'} = -1$ by our definition, $\hat{v}_{j,j} = \sum_{i=1}^n \hat{\xi}_{i,j}^2 / n$ is the sample variance of the $j$th residual. The bias correction of estimating $v_{j,j'}$ is needed for $j_1 \neq j$. From Equation (2), the node-wise regression estimate for partial correlation is $\hat{\rho}_{j,j'} = -\hat{v}_{j,j'} (\hat{v}_{j,j'} \hat{v}_{j,j'})^{-1/2}$.

2.2. Information Incorporation

Qiu and Zhou (2020) showed that $\sqrt{\hat{\mu}_{j,j'}}$ converges to the standard normal distribution under suitable conditions if the true partial correlation between the $j_1$th and $j_2$th variable is zero. Therefore, we could use $\sqrt{\hat{\mu}_{j,j'}}$ as the test statistic to identify nonzero partial correlations among a set of genes. The CLEVELD procedure (Qiu and Zhou, 2020) is built upon the estimate $\hat{\rho}_{j,j'}$. However, this procedure fails to use prior knowledge of the network structure. Motivated by Wang et al. (2013); Benedetti et al. (2020); Yi et al. (2021), we consider a more flexible scheme for gene network inference based on the node-wise regression approach, which can benefit from the prior biological knowledge.

Although the estimator $\hat{v}_{j,j'}$ in Equation (3) is consistent to $v_{j,j'}$, as the $L_1$-norm penalty shrinks the lasso estimates toward zero, $\hat{\alpha}_{j,j'}$ is biased for the true regression coefficient $\alpha_{j,j'}$. This would lead to an estimation error of the estimator $\hat{v}_{j,j'}$ for the residual covariance. To overcome the bias issue of lasso estimation, if we know $\alpha_{j,j'} = 0$, we should not penalize $\alpha_{j,j'}$ in the node-wise regression in Equation (1). In practice, associations between some well-studied genes are often known. To utilize such biological information, we remove the penalty on the corresponding coefficients in the node-wise regressions.
Specifically, given a set of known connections in a gene network \( I = \{(j_1, j_2) : j_1 \text{th, } j_2 \text{th genes are connected from prior knowledge}\} \), let \( I_j = \{j_2 : (j_1, j_2) \in I\} \) be the prior information set of connections with the \( j_1 \)th gene. Let \( I_j^{\perp} \) be the complement set of \( I_j \). We propose to fit the node-wise regression of \( Y_{i,j_1} \), without penalizing the coefficients from \( I_{j_1} \). Let

\[
\hat{\alpha}_{j_1} = \arg \min_{\alpha_{j_1} : \alpha_{j_1} \cdot \alpha_{j_1} = 1} \frac{1}{n} \sum_{i=1}^{n} \left( Y_{i,j_1} - \alpha_{j_1} \cdot \alpha_{j_2} Y_{i,j_2} \right)^2 + \lambda_j \sum_{j_2 \in I_{j_1}^{\perp}} |\alpha_{j_1,j_2}|
\]

be the lasso estimate with the incorporation of prior information, where \( \hat{\alpha}_{j_1} = (\hat{\alpha}_{j_1,1}, \ldots, \hat{\alpha}_{j_1,p})' \) with \( \hat{\alpha}_{j_1,j_1} = -1 \). Intuitively, if the true coefficient \( \alpha_{j_1,j_2} \neq 0 \) and \( j_2 \in I_{j_1} \) is in the prior information set, the un-penalized estimate \( \hat{\alpha}_{j_1,j_2} \) should be more accurate than the lasso estimate \( \hat{\alpha}_{j_1,j_2} \) without using the prior information. On the other hand, if \( j_2 \in I_{j_1} \) but \( \alpha_{j_1,j_2} = 0 \), the estimate \( \hat{\alpha}_{j_1,j_2} \) should still be close to zero.

We then use the prior information incorporated lasso estimator \( \hat{\alpha}_{j_1} \) in Equation (3) to estimate the error covariance matrix \( V \) and the partial correlations. The proposed estimate takes the form

\[
\hat{\rho}_{j_1,j_2} = -\hat{v}_{j_1,j_2}(\hat{v}_{j_1,j_2}, \hat{v}_{j_2,j_2})^{-1/2}, \quad \text{where} \quad \hat{v}_{j_1,j_2} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\xi}_{j_1,j_2} + \hat{\alpha}_{j_1,j_2} \hat{\xi}_{j_2,j_2} + \hat{\alpha}_{j_2,j_2} \hat{\xi}_{j_1,j_2}).
\]

The proposed estimate \( \hat{\rho}_{j_1,j_2} \) is in the same formulation as \( \hat{v}_{j_1,j_2} \) in Equation (3), except the fully-penalized lasso coefficient \( \hat{\alpha}_{j_1,j_2} \) is replaced by the information incorporated coefficient \( \hat{\alpha}_{j_1,j_2} \). Note that we still use the residuals \( \{\hat{\xi}_{i,j}\} \) from the fully-penalized lasso regression. This is because the residuals from the information incorporated lasso regression could be small due to removing the penalties in the prior information set.

Under the conditions that data are Gaussian distributed, nonzero elements in each row of the precision matrix \( \Omega \) are at a smaller order of \( n^{1/4} \langle \log p \rangle^{-3/2} \) (a sparse precision matrix), and additionally the size of the prior information set for each variable is bounded, following the proof of Proposition 1 in Qiu and Zhou (2020), it can be shown that the information-incorporated partial correlation \( \sqrt{n} \hat{\rho}_{j_1,j_2} \) still converges to the standard normal distribution under the null hypothesis \( \rho_{j_1,j_2} = 0 \). Therefore, we can use the information-incorporated statistic \( \sqrt{n} \hat{\rho}_{j_1,j_2} \) to test for nonzero \( \rho_{j_1,j_2} \) and to construct the partial correlation graph. Note that the statistics for testing \( \hat{\rho}_{j_1,j_2} = 0 \) based on \( \rho_{j_1,j_2} \) and its Fisher’s transformation are asymptotically equivalent under the null hypothesis by Taylor’s expansion. For testing high-dimensional partial correlations, the classical methods for fixed-dimensional data are no longer applicable due to the rank deficiency of the sample covariance matrix. Our proposed estimator in Equation (4) is designed for high-dimensional data, corrects the biases induced by regularization, and can incorporate prior information.

2.3. FDR Control

We consider two genes to be connected if their partial correlation is nonzero. Recovering the gene network amounts to testing the multiple hypotheses

\[
H_0 : \rho_{j_1,j_2} = 0 \quad \text{v.s.} \quad H_a : \rho_{j_1,j_2} \neq 0.
\]

Based on the proposed estimator \( \hat{\rho}_{j_1,j_2} \) from Equation (4), we propose a multiple testing procedure for the hypotheses (6) to identify nonzero partial correlations with FDR control. The false discovery proportion (FDP) is defined as the ratio of the number of false positives (FP) over the number of discovered connections (denoted as R), i.e., FDP = FP/R. In the context of a gene expression network, the false positives are the identified connections in the network but with the true partial correlation being zero. FDR is the expected value of FDP, which reflects the average proportion of false positives among discoveries over repeated experiments under the same condition. Let \( \Phi(\cdot) \) be the cumulative distribution function of the standard normal distribution and \( B(\tau) = 2 - 2\Phi(\tau \sqrt{\log(p)}) \). Given a threshold level \( \tau \), the candidate set of nonzero partial correlations is

\[
D(\tau) = \{j_1,j_2) : \sqrt{n} \hat{\rho}_{j_1,j_2} > \tau \sqrt{\log(p)}\}.
\]

Since \( \sqrt{n} \hat{\rho}_{j_1,j_2} \) converges to the standard normal if \( \rho_{j_1,j_2} = 0 \) as sample size increases, \( \Pr(\sqrt{n} \hat{\rho}_{j_1,j_2} > \tau \sqrt{\log(p)})|\rho_{j_1,j_2} = 0\approx B(\tau) \). Therefore, the number of false positives in the set \( D(\tau) \) is approximately \( B(\tau) m_0 \) which is bounded by \( B(\tau) (p^2 - p) \), where \( m_0 \) denote the number of true zero partial correlations among a set of \( p \) genes and is bounded by \( p^2 - p \) (the number of off-diagonal elements in a \( p \times p \) matrix). To control FDR at the nominal level \( \alpha \), we could choose

\[
\hat{\tau}_\alpha = \inf \left\{ \tau \in (0,2] : \frac{B(\tau) \times (p^2 - p)}{\max\{1,|D(\tau)|\}} \leq \alpha \right\},
\]

where \( |D(\tau)| \) denotes the size of the discovery set \( D(\tau) \). The proposed PCGII procedure uses \( \hat{\tau}_\alpha \) as the significant set of nonzero partial correlations for the hypotheses (6).

The multiple testing procedure for PCGII follows Storey’s FDR control procedure (Storey, 2002). Although the estimates \( \{\hat{\rho}_{j_1,j_2}\} \) are dependent, the validity of this FDR control procedure for high-dimensional sparse precision matrices and partial correlation matrices has been shown in Liu (2013) and Qiu and Zhou (2020). As we use \( p^2 - p \) as \( m_0 \), our approach is asymptotically equivalent to the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) could be applied although the theoretical result of this procedure for testing high-dimensional partial correlation matrices has not been well studied.

3. Simulation Studies

In this section, we evaluate the performance of our proposed PCGII procedure and compare it with three existing methods which control FDR but cannot take prior information: E-TEST (Bernal et al., 2019), F-GGM (Ren et al., 2015; Wang et al., 2016), and CLEVEL (Qiu and Zhou, 2020). Other information-assisted methods such as Yi et al. (2021) cannot control FDR and are not included in the comparison.

3.1. Simulation Settings

We simulated gene networks using three frequently used network structures: block diagonal, random nonzero locations, and scale-free.
• Block diagonal structure: The covariance matrix $\Sigma$ of size $p \times p$ is block diagonal with blocks $B_1, \ldots, B_H$ on the main diagonal and zero elsewhere, where each block $B_k$ is compound-symmetric with size $k \times k$, diagonal value $h_{11}$ and off-diagonal value $h_{22}$. The value of $h_{11}$ was generated from $\text{Uniform}(1, 2.5)$, and $h_{22}$ was generated from $\text{Uniform}(0.3, 0.5)$ for $h = 1, \ldots, H$.

• Random structure: The support of the precision matrix $\Omega$ was generated based on the Erdos-Renyi model (Erdős et al., 1960) by the R package igraph (Csarid and Nepusz, 2006) with parameter $\eta$ for the probability of drawing an edge between two arbitrary nodes.

• Scale-free structure: The support of the precision matrix $\Omega$ was generated based on the Barabasi-Albert model (Barabási, 2009) by the R package igraph with parameter $e$ for the density of edges.

For random and scale-free structures, the non-zero elements of the precision matrix in the support set were generated randomly from a uniform distribution with support $(-0.5, -0.2) \cup (0.2, 0.5)$. To ensure positive-definiteness of $\Omega$ for those two structures, diagonal elements of $\Omega$ were equal to the row sums plus a small number, which we set as 0.1. For each network structure, we considered different sparsity levels: the block size was $k = 4, 8$ for the block diagonal structure; the sparsity parameter was $\eta = 0.02, 0.03$ for the random structure; and the number of new edges generated in each step was $e = 1, 2$ for the scale-free structure. The dimension of networks was $p = 100$ and 200. In total, we investigated 12 settings for all combinations of network structure, sparsity, and dimension.

In addition, we simulated data using two moderate sample sizes $n = 60, 80$ and two small sample sizes $n = 10, 15$ for each of the 12 settings. The precision matrix $\Omega$ was fixed once generated for each setting, and 500 expression datasets of size $n$ were simulated for both normal and count data types. For normal data, datasets were generated from a multivariate normal distribution with mean zero and covariance $\Sigma$. For count data, correlated data were drawn from a multivariate normal distribution with non-zero mean and covariance $\Sigma$, and discrete counts were obtained by rounding these continuous data, as suggested in Kruppa et al. (2016). In the main text, we primarily focus on the results for count data with moderate sample sizes. More simulation results for the other scenarios are presented in the supplementary materials. We chose the lasso penalty parameter in the node-wise regression as $\sqrt{2 \log(p/\sqrt{\delta})}/n$ following the suggestion in Wang (2015).

In practice, we don’t expect the prior information to contain all true connections. To evaluate the effect of prior information size (PIS), we set the prior size to be 70% or 30% of the total number of true edges, denoted as PIS: 0.70 and PIS: 0.30, respectively. In addition, the prior information is not guaranteed to be 100% accurate in real applications. To mimic real situations when applying the PCGII method, we varied the prior accuracy (PA) to be 100% or 50%, where PA is defined as the proportion of true edges of all edges included in the prior information. Our proposed PCGII procedure was evaluated under the four combinations of two PIS levels and two PA levels for each setting.

### 3.2. Simulation Results

To evaluate the performance of PCGII, E-TEST, F-GGM, and CLEVEL, we examined the ROC (receiver operating characteristic) curve, the Precision-Recall curve, and a variety of metrics, including empirical FDR, total power (recall), and excess power, which are calculated as the averages over 500 datasets for FDP, $\text{TP}/m_1$, and $(\text{TP} - \text{TP}_{\text{prior}})/(m_1 - \text{TP}_{\text{prior}})$, respectively, where $m_1$ denote the number of true nonzero partial correlations and $\text{TP}_{\text{prior}}$ is the number of true positives in the prior set. Excess power is discussed in detail in Section 3.3.

The Benjamini-Hochberg method was applied to the $p$-values to control FDR for E-TEST and F-GGM. For CLEVEL and PCGII, we used the method described in Section 2.3 to control FDR. Figure 1 presents results for count data under sparse network conditions and Supplementary Figure S-1 presents the results under dense simulation settings, both for moderate sample sizes. Supplementary Figure S-2 presents all results for normal data with moderate sample sizes. For small sample sizes, Supplementary Figures S-3 and S-4 exhibit results for count and normal data, respectively.

ROC curves plot the true positive rate (TPR) versus the false positive rate (FPR), where TPR is the proportion of true positives among the total number of nonzero partial correlations, and FPR is the proportion of false positives among the total number of zero partial correlations. ROC curves evaluate the ranking of true edges and non-edges for each method, and higher curves indicate better performances. ROC curves in panel (A) of Figure 1 and Supplementary Figures S-1 to S-4 show that PCGII is the best performer or among the best performers for all settings. The other three approaches have comparable performance for the scale-free and random network structures. For the block-diagonal structure, CLEVEL is substantially better than E-TEST and F-GGM, and PCGII is substantially better than CLEVEL.

We also evaluate the FDR control of each method by plotting the empirical FDR versus the nominal level. Lines close to the $y = x$ line (solid black line in the plots) indicate desired FDR control. As shown in panel (B) of Figure 1 and Supplementary Figures S-1 to S-4, for all simulation settings, F-GGM is very liberal and cannot control FDR to the nominal level, while E-TEST tends to be conservative. For dense settings with moderate sample sizes, the empirical FDR of PCGII is very close to the nominal level, indicating that PCGII controls FDR to the desired level. For sparse settings with moderate sample sizes as shown in Figure 1 and Supplementary Figure S-2, CLEVEL and PCGII control FDR very well under the scale-free network structure. Under the random structure, both CLEVEL and PCGII are liberal but to a much lesser degree than F-GGM. Under the block-diagonal structure, PCGII is a little conservative, while CLEVEL controls FDR closer to the nominal level. Results for small sample sizes in Supplementary Figures S-3 and S-4 show that the FDR control tends to be less effective compared to settings with moderate sample sizes for all methods, while PCGII is still the best method with empirical FDR closest to the nominal level. Taking all scenarios into consideration, PCGII provides the best FDR control overall.

Recall (statistical power) and Precision (FDR control) are two compromising aspects of a statistical test. We plot Precision-Recall curves in panel (C) of Figure 1 and Supplementary Figures S-1 to S-4. For the same level of Precision, a method achieving higher power than others would be the best-performing method. As shown in panel (C) of Figure 1 and Supplementary Figure S-1, PCGII achieves the highest power across the whole range of empirical FDR for all settings except for the low dimension setting under scale-free structure for the sparse network and moderate sample sizes, where PCGII, CLEVEL, and F-GGM achieve comparable power that is higher than E-TEST. For
all other simulation settings, Supplementary Figures S-2 to S-4 demonstrate that PCGII always achieves the highest power among all methods, and the improvement over the second-best performing method is quite substantial.

Overall, the simulation results show that PCGII provides a better ranking of edges, demonstrates desired FDR control, and achieves higher total power than other methods. PCGII can detect a larger number of true connections while maintaining the FDR controlled to the desired level, providing a more reliable network.

3.3. Excess Power

The simulation results in Section 3.2 show that PCGII does benefit from incorporating prior information compared with CLEVEL. In this sub-section, we evaluate the advantages of incorporating prior connections and the capability of PCGII to identify additional edges besides those in the prior information. To do this, we computed the “excess power” in detecting the remaining connections after excluding edges in the prior set by taking the average of \((TP - TP_{prior})/(m_1 - TP_{prior})\) over 500 simulated datasets for each simulation setting on count data with moderate sample sizes. Such evaluation provides insights into PCGII’s performance in identifying unknown connections in the network, offering a comprehensive understanding of its overall performance in network reconstruction. The excess power of the other three approaches was computed similarly.

Figure 2 presents the results of excess power versus empirical FDR for sparse simulation settings. When prior information is 100% accurate (i.e. PA=1, as in Figure 2) and the prior information includes 70% of true edges (i.e., PIS = 0.7), PCGII achieves the highest excess power than other approaches, indicating PCGII detects more additional true edges outside of the prior set than other methods. For the block-diagonal structure, the difference between PCGII and CLEVEL, the next best-performing method, is quite large. When prior information is still 100% accurate (PA=1) but the prior information includes only 30% of true edges (PIS = 0.3), PCGII achieves the highest excess in most cases but is the second best-performing method for the scale-free setting when \(p = 100\). The results indicate that PCGII’s excess power increases with the inclusion of more correct information. Such benefits from incorporation are more pronounced as the dimension of networks increases. The results from dense networks lead to the same conclusion (Supplementary Figures S-5 to S-7).

When half of the edges incorporated are false connections, i.e. PA=0.5, PCGII is no longer the method with the highest excess power for low levels of empirical FDR (Supplementary Figures S-5 to S-7). When prior information includes a large proportion of false information, the drop in PCGII’s performance is because PCGII prioritizes all incorporated connections regardless of their accuracy, and incorporating false information inflates empirical FDR. Hence, we recommend practitioners only incorporate reliable information in the prior set to improve the prior accuracy and take advantage of PCGII.

4. Analysis of Real Datasets

4.1. FERONIA Network Analysis

FERONIA is a receptor kinase that regulates plant growth and development, stress responses, and reproduction. It has been suggested that FERONIA regulates endoplasmic reticulum (ER) body formation through the transcription factor NAI1 (Wang et al., 2022). Scientists are interested in learning (1) whether other transcription factors are involved in ER body formation mediated by FERONIA, and (2) whether ER body-related genes are directly associated with each other and if so, how FERONIA is involved in such direct associations. To address these questions, we built the gene network of FERONIA and a list of 17 known ER body-associated genes (Wang et al., 2017) based on the gene expression study in Arabidopsis (Wang et al., 2022). This study included two genotypes (wild type and fer-4 loss-of-function mutant) with three biological replicates for each genotype. We centered data by genotype before network analysis to control the treatment effect. When applying PCGII, we included the FERONIA-NAI1 interactions.
connection and five confirmed functional connections between transcription factor NA1I and its target genes JAL34, NA12, BGLU23, MEB1, and GLL23 (Wang et al., 2017). In total, six connections were included in the prior set for PCGII as shown in panel (a) of Supplementary Figure S-8.

PCGII, CLEVEL, E-TEST, and F-GGM were applied to construct the network, and they identified 28 (PCGII), 28 (CLEVEL), 0 (E-TEST), and 36 (F-GGM) edges, respectively, when FDR was controlled at 0.05. See Figure 3 and Supplementary Figure S-9. Similar to the performance in simulation studies, E-TEST was very conservative and did not recover any connections in this network.

Comparing the constructed networks with confirmed functional connections, these methods respectively affirmed 5 (PCGII), 1 (CLEVEL), and 0 (F-GGM) out of the 6 priorly confirmed connections. CLEVEL confirmed one validated connection (FERONIA-NAI1) (Wang et al., 2022) without incorporating information and PCGII recovered most of the confirmed connections benefiting from information incorporation. PCGII and CLEVEL both detected another edge (NAI1-ML3) that was not in the prior. We later found that this connection was a confirmed functional relationship in Wang et al. (2022). PCGII also uniquely detected FERONIA-JAL35 and FERONIA-MVP connections. Because jacalin-related lectins (JALs) regulate the size of ER body-type β-glucosidase complexes (Nagano et al., 2008), our modeling suggests that additional transcription factor(s), possibly regulated by FERONIA, may function to mediate FERONIA-JAL35/MVP connections. This hypothesis can be examined by future biological experiments.

PCGII, CLEVEL, and F-GGM shared 17 common edges among the ER body-associated genes (see panel (b) of Supplementary Figure S-8). By analyzing the dataset presented in Zander et al. (2020), we found that 8 out of the 17 common edges can be potentially explained by the corresponding genes’ co-expression with the transcription factor MYC2, which was not included in the network construction. MYC2 is an established phosphorylation target of FERONIA (Guo et al., 2018) and mediates jasmonic acid signaling to control ER body formation (Wang et al., 2017). So our results also suggest a role of MYC2 in ER body related gene expression.

4.2. Network Construction When No Prior Information Is Available

In practice, there are gene expression studies with no prior information on gene-gene direct associations. In this subsection, we illustrate how to apply PCGII to such studies by introducing a “pseudogene” that has connections with some of the originally observed genes and using these connections as prior information. More specifically, we construct a supplemented dataset by generating \( \tilde{Y}_{i,j} = Y_{i,j} + a_j Z_i \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, p \), where \( Y_{i,j} \) denote the original gene expression data for the \( i \)th sample and \( j \)th gene, \( a_j \) is a coefficient that controls the connection between the pseudogene and the original set of genes, and \( Z_1, \ldots, Z_n \) are the generated “expression levels” of the pseudogene and are i.i.d. from the standard normal distribution. Denote the supplemented data by \( \bar{Y}_i = (\bar{Y}_{i,1}, \ldots, \bar{Y}_{i,p}, Z_i)' \). It can be shown that the upper \( p \times p \) block \( \bar{\Psi} \) of the partial correlation matrix of the supplemented data \( \bar{Y}_i \) is the same as the partial correlation matrix of the original data. Furthermore, the true partial correlation between \( Y_{i,j} \) and \( Z_i \) is determined by the coefficients \( a_j \).s. If \( a_j \neq 0 \), the partial correlation between the pseudogene and the \( j \)th gene is nonzero and there is an edge between them in the network. A more detailed description of the method is presented in Section 4.1 of supplementary materials.

We applied such a method to an E. coli dataset obtained by Schmidt-Heck et al. (2004) to study the stress response during the
production of recombinant protein. The dataset comprised 102
genes measured at 9 time points. We conducted network analysis
on the original data as suggested by the previous studies (Schäfer
and Strimmer, 2005; Bernal et al., 2019). No prior information is
available for this dataset.

Table 1. Results on E. coli dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th># Total Discoveries</th>
<th># True Discoveries</th>
<th># False Discoveries</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCGII</td>
<td>19</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>CLEVEL</td>
<td>7</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>E-TEST</td>
<td>10</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>F-GGM</td>
<td>13</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

We applied E-TEST, F-GGM, CLEVEL, and PCGII to the
supplemented data \{\mathbf{Y}'_i\} that was constructed with 20 non-zero
\(a_{ij}\)s for 20 randomly selected genes. Section 4.2 of supplementary
materials presents data supplementation and analysis details. We
first evaluate the performance of four approaches in recovering
the network involving the pseudogene. Table 1 shows the numbers
of total discovered connections, discovered true connections, and
discovered false connections, respectively. PCGII identified 19 out
of 20 true edges and none of the null edges, providing high power
with no type I errors. The other three approaches all discovered
some false connections while identifying fewer true connections.

We next report the identified connections of the upper \(p \times p\)
block \(\mathbf{\Phi}\), which is the same as the partial correlation matrix of
the original data. When controlling FDR at 0.05, E-TEST exhibited a
conservative nature, identifying only 8 edges among 102 nodes.
Conversely, F-GGM detected a large number of edges (4342),
potentially compromising the interpretability and meaningfulness
of the network. In contrast, CLEVEL and PCGII discovered 245
and 240 edges. The varied performances aligned with our findings
in simulation studies. Analysis of the original data without the
pseudogene led to 4356, 9, and 245 detected edges for F-GGM, E-
TEST, and CLEVEL, respectively, which is highly consistent with
the analysis of the supplemented dataset (Supplementary Figure
S-10). This indicates that supplementing the observed data with
the pseudogene did not seem to change the main results of the
other methods.

5. Discussion

In this paper, we develop a new approach called PCGII to
construct gene networks by incorporating biological information
into regularized node-wise regression when estimating partial
correlations. The proposed approach allows prior biological
knowledge to assist network construction, which is not feasible
for other GGM-based network inference methods. In addition,
PCGII controls FDR much better than other methods, which
can be either too conservative (E-TEST) or too liberal (F-GGM).
Benefiting from the prior information, PCGII exhibits enhanced
signal detection ability in both simulation and case studies.

For well-studied organisms, some gene-gene direct associations
have been validated by independent experiments, establishing
highly reliable prior information. Such prior information provides
confident network connections. PCGII effectively utilizes such
information to reconstruct the network. Our analysis of excess
power shows that when prior information is accurate, prior
information incorporation not only maintains those confirmed
connections but also helps identify new connections not in the
prior set, benefiting the construction of the whole network.

In some cases, it is possible that some known direct associations
may be experimental artifacts and are not as reliable. To explore
the impact of prior information accuracy and size on PCGII,
we conducted a sensitivity analysis. The results are presented
in Section S-5 of supplementary materials. In general, PCGII’s
power positively correlates with the inclusion of correct prior
information. However, incorporating false information could lead
to an inflation of the empirical FDR. As a result, we strongly
recommend incorporating only high-confidence information in real
data applications.

Our proposed PCGII can handle large and complex network
structures with efficient computation. For a dataset with 200
nodes, it takes about 1 second on a MacBook Pro with Apple
M1, 10-core CPU, 24-core GPU, 14GB of RAM, and 1TB SSD.
PCGII utilizes penalized node-wise regression and can handle
high-dimensional data. In addition, PCGII can be used not
only to analyze transcriptomic data but also to analyze other
omics data such as metabolomics, proteomics, and microbiome
data. The R package PCGII is freely available for download at
https://cran.r-project.org/package=PCGII.

Acknowledgments

The authors thank the reviewers for their valuable suggestions.

Funding

This research was partially supported by the Office of Science
(BER) US Department of Energy (DE-SC0014395) and the
Iowa State University Plant Sciences Institute Scholars Program.
This paper is also a product of the Iowa Agriculture and
Home Economics Experiment Station, Ames, Iowa. Project No.
IOW05658 is supported by USDA/NIFA and State of Iowa
funds. Any opinions, findings, conclusions, or recommendations
expressed in this publication are those of the author(s) and do not
necessarily reflect the views of the U.S. Department of Agriculture.

Conflict of Interest: none declared.

References

science, 325(5939), 412–413.
understanding the cell’s functional organization. Nature reviews
genetics, 5(2), 101.
Benedetti, E., Pučić-Baković, M., Keser, T., Gerstner, N.,
Büyüközkkan, M., Štambuk, T., Selman, M. H., Rudan, I.,
Polasék, O., Hayward, C., et al. (2020). A strategy to incorporate prior knowledge into correlation network cutoff
discovery rate: a practical and powerful approach to multiple
testing. Journal of the Royal statistical society: series B
(Methodological), 57(1), 289–300.
discovery rate in multiple testing under dependency. Annals of
statistics, pages 1165–1188.
Bernal, V., Bischoff, R., Guryev, V., Grzegorczyk, M., and
Horvathovich, P. (2019). Exact hypothesis testing for shrinkage
based gaussian graphical models. Bioinformatics,
World Scientific.
Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R.
complex network research. InterJournal, Complex Systems,
1695.
covariance estimation with the graphical lasso. Biostatistics,
9(3), 432–441.
Guo, H., Nolan, T. M., Song, G., Liu, S., Xie, Z., Chen, J.,
receptor kinase contributes to plant immunity by suppressing
jasmonic acid signaling in arabidopsis thaliana. Current Biology,
28(20), 3316–3324.
Kruppa, J., Kramer, F., Beißbarth, T., and Jung, K.
(2016). A simulation framework for correlated count data of
features subsets in high-throughput sequencing or proteomics
experiments. Statistical Applications in Genetics and Molecular
discovery rate control. The Annals of Statistics, 41(6), 2948–
2978.
Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C.,
Aracne: an algorithm for the reconstruction of gene regulatory
networks in a mammalian cellular context. In BMC bioinformatics,
graphs and variable selection with the lasso. The annals of
Nagano, A. J., Fukao, Y., Fujiwara, M., Nishimura, M., and
regulate the size of er body-type β-glucosidase complexes in
arabidopsis thaliana. Plant and cell physiology, 49(6), 969–980.
correlation estimation by joint sparse regression models. Journal
of the American Statistical Association, 104(486), 735–746.
partial correlation graphs with application to brain imaging.
Asymptotic normality and optimalities in estimation of large
1026.
large-scale covariance matrix estimation and implications for
functional genomics. Statistical applications in genetics and
molecular biology, 4(1).
Schmidt-Heck, W., Guthke, R., Toepfer, S., Reisser, H.,
Duerrschmid, K., and Bayer, K. (2004). Reverse engineering of
the stress response during expression of a recombinant protein.
Journal of the Royal Statistical Society Series B: Statistical
Methodology, 64(3), 479–498.
Wang, J.-Z., Li, B., Xiao, Y., Ni, Y., Ke, H., Yang, P., de Souza,
er body formation and indole glucosinolate metabolism by the
plastidial retrograde signaling metabolite, mecpp. Molecular
plant, 10(11), 1400–1416.
Wang, P., Clark, N. M., Nolan, T. M., Song, G., Bartz, P. M.,
Liao, C.-Y., Montes-Serey, C., Katz, E., Polko, J. K., Kieber,
J. J., et al. (2022). Integrated omics reveal novel functions and
underlying mechanisms of the receptor kinase feronia in
arabidopsis thaliana. The Plant Cell.
Graphical Model. R package version 1.0.
Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L.,
efficient algorithm for the inference of gaussian graphical model in
biological networks. PLoS computational biology, 12(2),
e1004755.
(2013). Incorporating prior knowledge into gene network study.
Bioinformatics, 29(20), 2633–2640.
Information-incorporated gaussian graphical model for gene expression data.
Biometrics.
Zander, M., Lewsey, M. G., Clark, N. M., Yin, L., Bartlett, A.,
Saldierna-Guzmán, J. P., Hann, E., Langford, A. F., Jow, B.,
for weighted gene co-expression network analysis. Statistical
applications in genetics and molecular biology, 4(1).
Biological network inference using low order partial correlation.