Gene Network Reconstruction by Integration of Prior Biological Knowledge

Yupeng Li*,†,‡ and Scott A. Jackson*,†,1
*Center for Applied Genetic Technologies, †Institute of Plant Breeding, Genetics and Genomics, and ‡Department of Statistics, University of Georgia, Athens, Georgia 30602
ORCID ID: 0000-0003-0843-7592 (Y.L.)

ABSTRACT With the development of high-throughput genomic technologies, large, genome-wide datasets have been collected, and the integration of these datasets should provide large-scale, multidimensional, and insightful views of biological systems. We developed a method for gene association network construction based on gene expression data that integrate a variety of biological resources. Assuming gene expression data are from a multivariate Gaussian distribution, a graphical lasso (glasso) algorithm is able to estimate the sparse inverse covariance matrix by a lasso (L_1) penalty. The inverse covariance matrix can be seen as direct correlation between gene pairs in the gene association network. In our work, instead of using a single penalty, different penalty values were applied for gene pairs based on *a priori* knowledge as to whether the two genes should be connected. The *a priori* information can be calculated or retrieved from other biological data, *e.g.*, Gene Ontology similarity, protein-protein interaction, gene regulatory network. By incorporating prior knowledge, the weighted graphical lasso (wglasso) outperforms the original glasso both on simulations and on data from Arabidopsis. Simulation studies show that even when some prior knowledge is not correct, the overall quality of the wglasso network was still greater than when not incorporating that information, *e.g.*, glasso.

KEYWORDS
gene network
graphical model
prior knowledge
systems biology
and gene
expression

A key challenge for biology is to understand the complex molecular interactions of genes in a living cell (Barabasi and Oltvai 2004). Analysis of gene networks provides a global view of these interactions and can provide biologists with a better understanding of complex biological systems (Friedman 2004; Karlebach and Shamir 2008). Also, by the use of gene networks, the guilt-by-association paradigm can be applied to infer the biological function of unknown genes (Wolfe *et al.* 2005).

Gene expression data, which is relatively easy to generate or to collect from databases, has been used to infer gene networks. A variety of gene network reconstruction methods based on gene expression data have been developed, *e.g.*, regression, mutual information, correlation Bayesian network, meta predicators, and others (Marbach *et al.* 2012). Here, we focus on the Gaussian graphical model (Dempster 1972). In this approach, gene expression is assumed to follow a multivariate Gaussian distribution $N(\mu, \Sigma)$, and Gaussian Markov random fields have been used to infer the structure of networks from gene expression data (Liu and Ihler 2011). A gene association network can be seen as an undirected graph $G = (V, E)$, where $V = \{v_p\}$ is the vertex set representing genes and $E = \{e_{ij}\}$ is the edge set representing association relations between pairs of genes. In an unweighted network, if $e_{ij} = 0$, genes i and j are conditionally independent given all other genes. The Hammersley–Clifford theorem implies that zeros in the inverse covariance matrix of a multivariate Gaussian distribution indicate absent edges in the corresponding graphical model (Besag 1974; Lauritzen 1996). Therefore, the problem of estimating the gene association network based on gene expression data can be transferred to estimating Σ^{-1} or selecting nonzero entries in Σ^{-1}. Many studies, however, have used Pearson correlation coefficients between pairs of genes to infer network structure; Pearson correlation coefficients correspond to the covariance matrix Σ and cannot infer the true structure.

An accurate inference of biological network using Gaussian graphical model is challenging for two main reasons. The first is that most genome-scale datasets are highly dimensional. Given p genes, there are possible $p(p - 1)/2$ edges, but gene expression data often have a limited number of samples. When the gene or locus number is much higher than the sample size in a dataset,

doi: 10.1534/g3.115.018127
Manuscript received February 5, 2015; accepted for publication March 28, 2015; published Early Online March 30, 2015.

[1]Corresponding author: 111 Riverbend Road, Athens, GA 30602. E-mail: sjackson@uga.edu

a traditional maximum-likelihood estimate for covariance variance is often not appropriate (Schafer and Strimmer 2005; Uhler 2012). Alternative methods have been developed for highly dimensional datasets (Meinshausen and Buhlmann 2006; Friedman *et al.* 2008; Yuan 2010; Cai *et al.* 2011; Ravikumar *et al.* 2011), most of which have used Lasso, a shrinkage and selection method using L_1 regularization, a popular approach to deal with highly dimensional datasets (Tibshirani 1996; Hastie *et al.* 2009). These methods have been shown to be able to asymptotically and consistently estimate the set of non-zero elements of Σ^{-1}.

The second major difficulty is the lack of efficient methods to integrate multiple levels of biological data to enhance model accuracy. Increasing amounts of biological data have been collected, especially with the development of high-throughput technologies, *e.g.*, microarrays and next-generation sequencing, which provide an unprecedented opportunity to explore biological systems. We now have access to genomic, epigenomic, transcriptomic, proteomic, metabolomic, and phenomic data, and careful analysis and integration of these genome-scale datasets should provide large-scale, multidimensional, and insightful views into biological systems (Joyce and Palsson 2006; Hawkins *et al.* 2010). Also, we have a variety of resources that serve as indicators of the functional association of any two genes, *e.g.*, Kyoto Encyclopedia of Genes and Genomes pathway, Gene Ontology (GO) similarity, protein−protein interaction, co-occurrence in literature, gene network generated from other methods, and association of orthologous genes. Integrating potentially reliable information from other sources should increase the accuracy of network reconstruction (Imoto *et al.* 2004; Mostafavi *et al.* 2008; Christley *et al.* 2009; Wang *et al.* 2013; Chen *et al.* 2014). Here, we present a statistical algorithm based on the Gaussian graphical model for gene association network reconstruction using gene expression data, which is able to solve these two challenges to allow a more thorough understanding of complex biological systems.

MATERIALS AND METHODS

Let $\Theta = \Sigma^{-1}$ and S be the empirical covariance matrix, then by applying L_1 penalty to the original log-likelihood for estimating Θ, the problem of estimating Θ becomes finding the Θ which maximize the formula

$$\log(\det(\Theta)) - tr(S\Theta) - \rho\|\Theta\|_1 ,$$

where *tr* is the trace, *i.e.*, the sum of the elements on the matrix diagonal, and $\|\Theta\|_1$ is the L_1 norm, *i.e.*, the sum of the absolute values of the elements of Σ^{-1}, and ρ is the penalty parameter. When $\rho = 0$, it is the normal maximum likelihood. When $\rho > 0$, some elements in Θ will be shrunk to zero. The sparsity of the estimated graph increases when ρ increases. For a fixed ρ, the graphical lasso (glasso) algorithm can be used to quickly solve the equation using a block-coordinate method (Friedman *et al.* 2008). The optimal ρ can be chosen empirically or tuned by cross-validation, Bayesian information criterion (BIC), or other methods (Friedman *et al.* 2008; Foygel and Drton 2010; Liu *et al.* 2010).

The Lasso regression can be interpreted as a Bayesian regression with a Laplace prior distribution, $Lp(0, \rho)$. In Bayesian statistics, prior information can be integrated into the model by changing the parameter in the prior distribution, and the posterior distribution should better approximate the true distribution of the data. Therefore, for the glasso algorithm, instead of using a single penalty parameter, it is reasonable to specify different amounts of penalties for different elements in Θ based on *a priori* information as to whether two genes are associated, or not. A smaller penalty can be given if *a priori* information indicates that they are linked. Then, the log-likelihood becomes

$$\log(\det(\Theta)) - tr(S\Theta) - \rho\|\mathrm{P} * \Theta\|_1 ,$$

where P is the prior information matrix and $*$ indicates component-wise multiplication, and $P \in [0, 1]$. Larger values for an element in

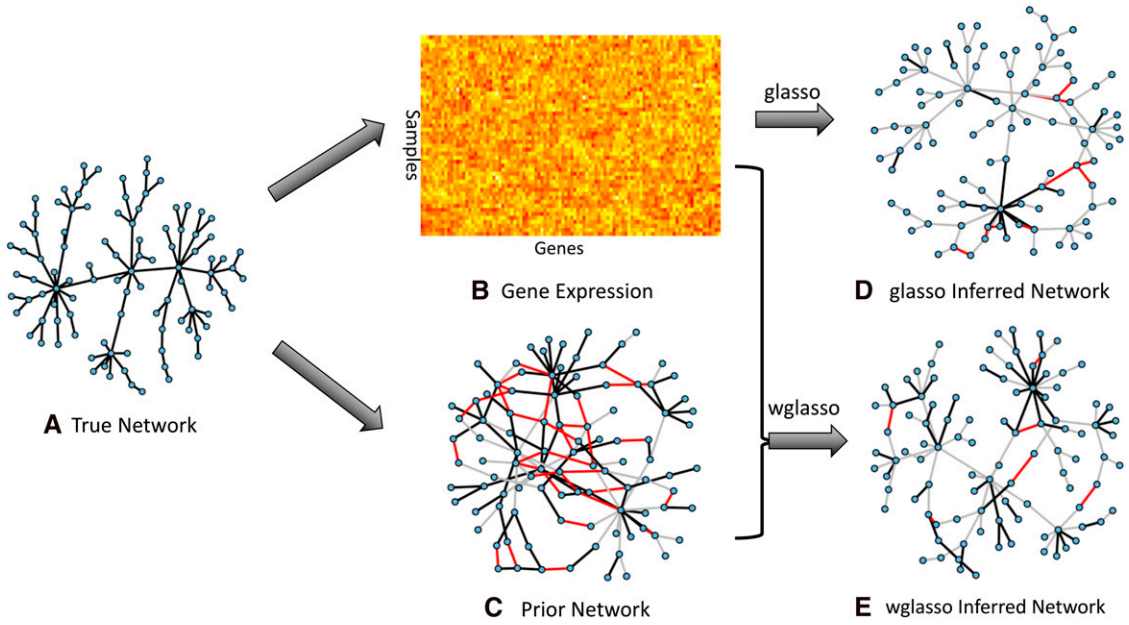

Figure 1 Demonstration of a single simulation using glasso and wglasso. (A) The true network with scale-free property. (B) Heatmap of gene expression data. The X-axis represents genes and the Y-axis represents samples. (C) The prior network, representing prior information of genes' associations with precision ratio 0.7. Black edges are correct association information; edges that exist in prior information but not in the true network are in red, and gray edges are missed associations in the prior information. (D) The estimated network using glasso. The edge colors have the same meaning as prior network. (E) The estimated network using wglasso. The edge colors have the same meaning as prior network.

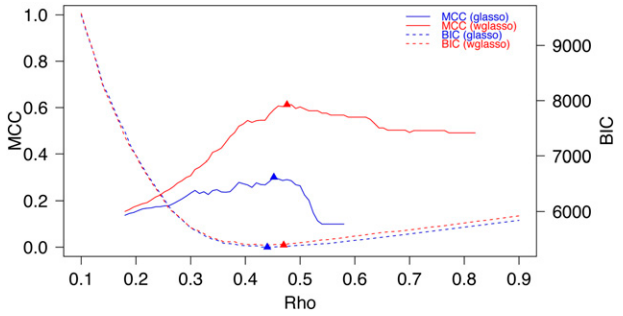

Figure 2 Parameter selection process of glasso and wglasso Matthews correlation coefficient (MCC; solid lines) and Bayesian information criterion (BIC; dashed lines) values of the estimated networks under different penalty parameters (Rho) using glasso (blue lines) and wglasso (red lines). Triangles are points that maximize MCC values and minimize BIC values under the corresponding penalty parameters, considered to be optimal penalty parameters.

P generate larger penalty values and represent weaker association of two genes based on *prior*i information. We name this updated version of glasso as weighted graphical lasso (wglasso), which is not only stable for high-dimensional datasets by utilization of Lasso but also more accurate than the original glasso through the integration of prior information.

The prior matrix can be obtained in numerous ways. For example, the GO semantic similarity between genes can be calculated using tools GOSemSim (Yu *et al.* 2010) or GOssTo (Yu *et al.* 2010), Then, the inverse of similarities can be implemented in the prior matrix, as a high similarity means a low penalty in our model. Other types of networks, *e.g.*, protein-protein interaction and gene regulatory networks, can be directly transferred to a prior matrix after inverse transformation of the weights of the network. Some databases estimate the functional association between genes using data mining and text mining and can be valuable and reliable resources for generating a prior matrix, *e.g.*, STING (Von Mering *et al.* 2005) and AraNet (Lee *et al.* 2010).

Although previous efforts to use lasso for incorporating *a priori* knowledge have been made, none is identical to our proposed method. For example, most of the previously described algorithms partitioned the gene network reconstruction into a number of linear regression problems (Anjum *et al.* 2009; Charbonnier *et al.* 2010; Wang *et al.* 2013); however, it has been shown this type of algorithms is less accurate than glasso (Friedman *et al.* 2008).

RESULTS AND DISCUSSIONS

Simulation studies demonstrate that wglasso is more reliable than glasso

Because most biological networks are scale-free (Barabasi and Oltvai 2004), a scale-free network with p genes and gene expression data with n samples based on the network topology are generated using the "huge" package in R (Zhao *et al.* 2012). A prior network with precision ratio x, $x \in [0,1]$ also was generated. A prior network is a weighted network, in which each edge corresponds to the prior information that two genes are functionally associated and the edge weight indicates the strength of the information. Precision ratio = x with $x > 0$ means p^*100 percent of prior edges are correct, and the remaining $(1 - p)^*100$ edges are incorrect. The edge number of the prior network is set to the same number of true edges. The edge weights of the prior network were randomly generated from the uniform distribution, *U(0,1)*. A special case is precision ratio = 0, which indicates no prior information and the prior network will not contain any edges. In this case, wglasso is equivalent to glasso. Once the empirical covariance matrix S is calculated from the expression data and the prior information matrix P retrieved from the prior network, we estimate the true network using glasso and wglasso. In order to find the optimal penalty parameter, ρ, wglasso networks were estimated under a sequence of ρ values from 0 to 1 with 0.01 intervals. Then we selected the network most similar to the true network. The Matthews correlation coefficient (MCC) was used to measure the similarity between the estimated and true networks (Matthews 1975), and was calculated as follows based on 2×2 contingency table,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. A penalty parameter that maximized MCC (maxMCC) was considered optimal. A single simulation with 100 genes, 50 samples and a 0.7 precision ratio is shown in Figure 1 and Figure 2.

Additional simulations were performed with a variety of sample sizes and precision ratios in order to systematically evaluate wglasso. The simulation was repeated 100 times for each combination of sample size and precision ratio. The results show that reconstructed networks with *a priori* information had a significantly greater maxMCC than without *a priori* information, indicating superior performance of wglasso (Figure 3). In most cases, wglasso outperformed glasso, even when most of the prior knowledge was incorrect. If the sample size was high and the precision ratio low, *e.g.*, sample size = 300 and precision ratio = 0.2, it is possible that an excess of incorrect prior information would be harmful. However, in real situations, the sample size is often much lower than the gene number and misleading *a priori* information based on experimental studies is likely to be low. Moreover, highly accurate prior knowledge results in more accurately reconstructed networks. In practice, it would be possible to integrate multiple resources to increase the reliability of prior information.

Tuning the penalty parameter using BIC

In real situations, the true network is unknown; thus, maxMCC cannot be used to select the optimal penalty parameter. Instead, one

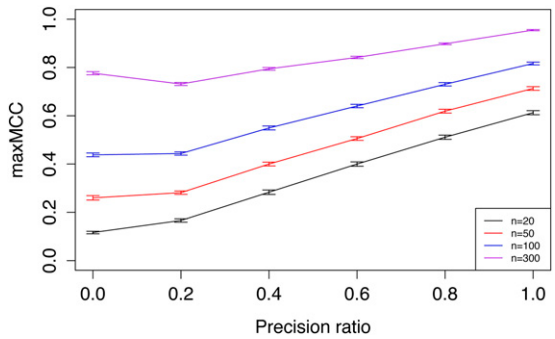

Figure 3 Simulation to compare glasso and wglasso. The simulation was repeated 100 times for each combination of sample sizes (n) and precision ratio of the prior information. Gene number = 100. In each simulation, the maximum Matthews correlation coefficient (maxMCC) of estimated networks from different penalty parameters is recorded. The Y-axis shows the mean of maxMCC from the 100 simulations. The error bars are 95% confidence intervals.

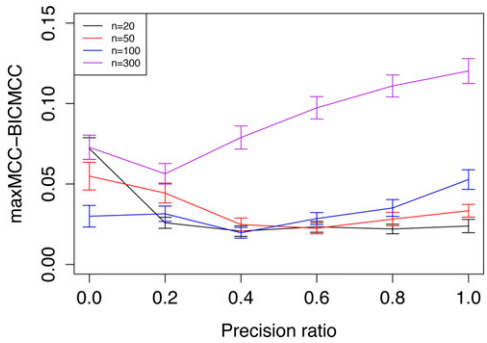

Figure 4 Selection of the optimal penalty parameter based on minimum Bayesian information criterion (BIC) simulation was repeated 100 times for each combination of sample sizes (n) and precision ratio of the prior information. Gene number = 100. For each simulation, the maximum Matthews correlation coefficient (MCC) value (maxMCC) of estimated networks from different penalty parameters is shown. The MCC value of the estimated network based on minimum BIC is also recorded, called BICMCC. The difference between two values (maxMCC − BICMCC) is used to evaluate two parameter selection methods. The Y-axis shows the mean of maxMCC − BICMCC from the 100 simulations. The error bars represent 95% confidence intervals.

can use cross-validation, BIC, extended BIC (Foygel and Drton 2010), and a stability approach for regularization selection (Liu *et al.* 2010). Cross-validation, stability approach for regularization selection, and BIC conduct many subsampling or permutations and thus are computationally intensive. Extended BIC tends to select very sparse networks, and often no edges are found. For a scale-free network, another method is to test whether the log transformed degree distribution has a linear relationship (Langfelder and Horvath 2008). The R-squared values from linear regressions generally increase as the network gets sparser, and the optimal fit is selected at the point where the increase trend slows down. The selection is usually done visually, so the method is not practical for simulations. Also, the optimal parameter varies by dataset and individual interpretation.

BIC is one of the standard methods for choosing the regularization parameter, and works well in our scale-free network simulations (Figure 2 and Figure 4). The BIC for Gaussian graphical model takes the form

$$BIC = -2l_n(\Theta) + |E|\log(n),$$

where $|E|$ is the edge number, n the sample size, and $l_n(\Theta)$ the log-likelihood function simplified from

$$l_n(\Theta) = \frac{n}{2}[\log(\det(\Theta)) - tr(S\Theta)].$$

Simulation showed that the difference between the MCC of reconstructed networks based on BIC and the maxMCC was small, especially when sample size was relatively small (Figure 2 and Figure 4).

Application to experimental data

We applied wglasso to gene expression data from the Eukaryotic species, *Arabidopsis thaliana*. Gene expression data of 795 genes related to isoprenoid pathways from 118 microarray experiments were collected (Wille *et al.* 2004). Prior information was obtained from AraNet, a probabilistic network of functional associations among 19,647 *Arabidopsis* genes (Lee *et al.* 2010). A total of 701 of 795 genes have functional associations in AraNet. The edge weight is the likelihood score, calculated from variety of resources that indicates a functional association between two genes. The likelihood scores range from 0 to 5, so they were rescaled to a range of 0.5 to 1 using the formula,

$$X\_new = 0.5 \times \frac{\max(X) - X}{\max(X) - \min(X)} + 0.5.$$

High likelihood scores will become low penalty values after rescaling. Pairs of genes without prior information have zero values in the prior matrix. The optimal penalty parameter was selected based on BIC. The network reconstructed using wglasso has 16,759 edges, whereas the glasso network has 19,331 edges. The MCC values were calculated by comparing the reconstructed networks with an independent benchmark network, which is the same benchmark set that was used to test the prediction performance of AraNet (Lee *et al.* 2010). The MCC of the estimated network with wglasso was 0.184, higher than that with glasso, which was 0.033.

By incorporating prior knowledge, weighted graphical lasso (wglasso) outperforms glasso both on simulation studies and data from *Arabidopsis*. Simulation studies showed that even when some prior knowledge was incorrect, the overall quality of network from wglasso network was higher than that from glasso. Moreover, the more accurate the prior knowledge, the better the reconstructed network. This method increases gene network reconstruction accuracy and will allow researchers to better study networks in complex biological systems and their interaction with external factors, *e.g.*, the environment.

ACKNOWLEDGMENTS

We would like to acknowledge funding from the National Science Foundation (MCB 1339194).

LITERATURE CITED

Anjum, S., A. Doucet, and C. C. Holmes, 2009 A boosting approach to structure learning of graphs with and without prior knowledge. Bioinformatics 25: 2929–2936.

Barabasi, A. L., and Z. N. Oltvai, 2004 Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5: 101–113.

Besag, J., 1974 Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc., B 36: 192–236.

Cai, T., W. D. Liu, and X. Luo, 2011 A constrained L(1) minimization approach to sparse precision matrix estimation. J. Am. Stat. Assoc. 106: 594–607.

Charbonnier, C., J. Chiquet, and C. Ambroise, 2010 Weighted-lasso for structured network inference from time course data. Stat Appl Genet Mol 9: Article 15.

Chen, G. C., M. J. Cairelli, H. Kilicoglu, D. Shin, and T. C. Rindflesch, 2014 Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference. PLoS Comput. Biol. 10: e1003666.

Christley, S., Q. Nie, and X. H. Xie, 2009 Incorporating existing network information into gene network inference. PLoS One 4: e6799.

Dempster, A. P., 1972 Covariance selection. Biometrics 28: 157–175.

Foygel, R., and M. Drton, 2010 Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Info. Process. Syst.* 23: 604–612.

Friedman, J., T. Hastie, and R. Tibshirani, 2008 Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432–441.

Friedman, N., 2004 Inferring cellular networks using probabilistic graphical models. Science 303: 799–805.

Hastie, T., R. Tibshirani, and J. Friedman, 2009 *The Elements of Statistical Learning*. Springer, New York.

Hawkins, R. D., G. C. Hon, and B. Ren, 2010 Next-generation genomics: an integrative approach. Nat. Rev. Genet. 11: 476–486.

Imoto, S., T. Higuchi, T. Goto, K. Tashiro, S. Kuhara *et al.*, 2004 Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. J. Bioinform. Comput. Biol. 2: 77–98.

Joyce, A. R., and B. O. Palsson, 2006 The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell Biol. 7: 198–210.

Karlebach, G., and R. Shamir, 2008 Modelling and analysis of gene regulatory networks. Nat. Rev. Mol. Cell Biol. 9: 770–780.

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

Lauritzen, S. L., 1996 *Graphical Models*. Oxford University Press, Oxford.

Lee, I., B. Ambaru, P. Thakkar, E. M. Marcotte, and S. Y. Rhee, 2010 Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. Nat. Biotechnol. 28: 149–156.

Liu, H., K. Roeder, and L. Wasserman, 2010 Stability approach to regularization selection (StARS) for high dimensional graphical models. Adv. Neural Info. Process. Syst. 23: 1432–1440.

Liu, Q., and A. T. Ihler, 2011 Learning scale free networks by reweighted L1 regularization. J. Machine Learning Res. Proc. Track 15: 40−48.

Marbach, D., J. C. Costello, R. Kuffner, N. M. Vega, R. J. Prill *et al.*, 2012 Wisdom of crowds for robust gene network inference. Nat. Methods 9: 796–804.

Matthews, B. W., 1975 Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim .Biophys. Acta 405: 442–451.

Meinshausen, N., and P. Buhlmann, 2006 High-dimensional graphs and variable selection with the lasso. Ann. Stat. 34: 1436–1462.

Mostafavi, S., D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris, 2008 GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 9: S4.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu, 2011 High-dimensional covariance estimation by minimizing L(1)-penalized log-determinant divergence. Electron J Stat 5: 935–980.

Schafer, J., and K. Strimmer, 2005 A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol 4: Article 32.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J Roy Stat Soc B Met 58: 267–288.

Uhler, C., 2012 Geometry of maximum likelihood estimation in Gaussian graphical models. Ann. Stat. 40: 238–261.

von Mering, C., L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp *et al.*, 2005 STRING: known and predicted protein−protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33: D433–D437.

Wang, Z. X., W. L. Xu, F. A. S. Lucas, and Y. Liu, 2013 Incorporating prior knowledge into gene network study. Bioinformatics 29: 2633–2640.

Wille, A., P. Zimmermann, E. Vranova, A. Furholz, O. Laule *et al.*, 2004 Sparse graphical gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. Genome Biol. 5: R92.

Wolfe, C. J., I. S. Kohane, and A. J. Butte, 2005 Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics 6: 227.

Yu, G. C., F. Li, Y. D. Qin, X. C. Bo, Y. B. Wu *et al.*, 2010 GOSemSim: an R package for measuring semantic similarity among go terms and gene products. Bioinformatics 26: 976–978.

Yuan, M., 2010 High dimensional inverse covariance matrix estimation via linear programming. J. Mach. Learn. Res. 11: 2261–2286.

Zhao, T., H. Liu, K. Roeder, J. Lafferty, and L. Wasserman, 2012 The huge package for high-dimensional undirected graph estimation in R. J. Mach. Learn. Res. 13: 1059–1062.

Communicating editor: B. J. Andrews