Inferring metabolite interactomes via molecular structure informed Bayesian graphical model selection with an application to coronary artery disease

Patrick J. Trainor[1,2,*], Joshua M. Mitchell[3,4,5], Samantha M. Carlisle[6], Hunter N.B. Moseley[3,4,5,7], Andrew P. DeFilippis[1,2], Shesh N. Rai[8,9]

[1]Department of Medicine, Division of Cardiovascular Medicine, University of Louisville
[2]Diabetes and Obesity Center, University of Louisville
[3]Department of Molecular & Cellular Biochemistry, University of Kentucky
[4]Center for Environment and Systems Biochemistry and the Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky
[5]Markey Cancer Center, University of Kentucky
[6]Department of Pharmacology and Toxicology, University of Louisville
[7]Institute for Biomedical Informatics, University of Kentucky
[8]Department of Bioinformatics and Biostatistics, University of Louisville
[9]Biostatistics Shared Facility, James Graham Brown Cancer Center, University of Louisville
*Corresponding Author

Keywords

Systems biology, Metabolomics, Graphical models, Bayesian statistics, Heart disease

Abstract

Introduction
While the generation of reference genomes facilitates the elucidation of gene-phenome associations, reference models of the metabolome that are specific to organism, sample type (e.g. plasma, serum, urine, cell-culture), and state (including disease), remain uncommon. In studying heart disease in humans, a reference model describing the relationships between metabolites in plasma has not been determined but would have great utility as a reference for comparing acute disease states such as myocardial infarction.

Materials and Methods
We present a methodology for deriving probabilistic models that describe the partial correlation structure of metabolite distributions ("interactomes") from metabolomics data. As determining partial correlation structures requires estimating $p*(p-1)/2$ parameters for p metabolites, the dimension of the search space for parameter values is immense. Consequently, we have developed a Bayesian methodology for the penalized estimation of model parameters in which the magnitude of penalization is drawn from probability distributions with hyperparameters linked to molecular structure similarity. In our work, structural similarity was determined as the Tanimoto coefficient of algorithmically-generated "atom colors" that capture the local structure around each atom within each structure. A Gibbs sampler (a Markov chain Monte Carlo technique) was implemented for simulating the posterior distribution of model parameters. We have made software for implementing this methodology publicly available via the R package *BayesianGLasso*.

Results / Conclusions
First, we demonstrate robust performance of our methodology (sensitivity, specificity, and measures of accuracy) for recovering the true underlying partial correlation structure over simulated datasets (with simulated metabolite abundances and simulated known structural similarity). We then present an interactome model for stable heart disease inferred from non-targeted mass spectrometry data via this methodology. Inspection of the local graph topology about cholate reveals probabilistic interactions with other primary bile acids, secondary bile acids, and many steroid hormones sharing the same precursors.

1. Introduction

Untargeted profiling of the metabolome of an organism provides a view into the small molecule determinants of phenotype. While the genome of an organism may be conceptualized as a blueprint for the composition and organization of an organism that is largely immutable (barring epigenetic modifications, DNA damage, and genetic mutations) (Gao, Jia, Zhang, Breitling, & Brenner, 2015; Keating & El-Osta, 2015; Martincorena & Campbell, 2015), the metabolome of an organism is dynamic and variable (Dallmann, Viola, Tarokh, Cajochen, & Brown, 2012; Krycer et al., 2017). Sources of variation within the metabolome of a single organism include tissue-, cell-, and organelle-specific localization of metabolic processes (Shlomi, Cabili, Herrgård, Palsson, & Ruppin, 2008; Voet, Voet, & Pratt, 2013); environmental exposures (Southam et al., 2014); and host-microbe interactions and metabolite exchange (Moriya, Satomi, Murata, Sawada, & Kobayashi, 2017). While the generation of reference human genomes has facilitated the interrogation of gene-phenome associations (including human disease associations), the intrinsic variability and dynamic nature of the metabolome of an organism likely precludes the generation of such a reference model. While a single reference model of the metabolome of an organism may not be sensible, significant efforts such as the HUSERMET project (Dunn et al., 2014) have been undertaken to quantify the repertoire of metabolites in specific biofluids for examining metabolite-metabolite and metabolite-phenotype associations. In order to make systems-level comparisons of the differences in the metabolome across phenotypes, models of the conditional relationships between metabolites are necessary. By the determination of sample media and analytical platform-specific probabilistic interaction models, henceforth called “interactomes”, systems-level comparisons of phenotypes that can be made. A specific use case for such an interactome model is the generation of a plasma interactome for stable heart disease.

Heart disease is the most prevalent cause of death globally (Benjamin et al., 2017). As a disease, heart disease does not represent a uniform condition, but rather a collection of diseases of varying etiologies (Kasper, 2015). Of particular interest in the study of coronary artery disease (CAD) is the elucidation of the precipitants of acute disease events such as myocardial infarction (Arbab-Zadeh & Fuster, 2015) or unstable angina, metabolic pathways associated with disease phenotypes (Y. Fan et al., 2016), and determining the metabolic consequences of acute events (Trainor et al., 2017). To date, an interactome describing the conditional relationships between blood plasma metabolites in humans with heart disease does not exist. If such a reference model were determined, it would facilitate making systems-level inferences regarding metabolic perturbations that accompany acute disease events such as unstable angina or acute myocardial infarction (MI).

While correlation networks have been used to describe the relationships between metabolites in many metabolomics experiments [see for example (Kotze et al., 2013; Madhu et al., 2015; Suarez-Diez et al., 2017; L. Wang et al., 2015)], this approach is limited as the topology learned represents only the pairwise marginal associations between metabolites. Determining a conditional relationship between two metabolites allows for inference regarding how the abundance of a specific metabolite influences the abundance of another metabolite after conditioning on the abundance of other intermediates. In order to model such conditional probabilistic dependencies between metabolite abundances, a Gaussian Graphical Model (GGM) approach may be employed as in the present work. GGMs provide a suitable framework for representing the joint probability distribution of metabolites that are detected in metabolomics experiments and for representing the probabilistic interactions between metabolites and have been employed for such a task previously (Krumsiek, Suhre, Illig, Adamski, & Theis, 2011; Shin et al., 2014).

A significant challenge in evaluating the relationships between metabolites in an untargeted metabolomics experiment is that the dimension of metabolites may be greater than the number of samples. Even given a relatively high ratio of samples to metabolites detected, in the evaluation of pairwise conditional relationships between metabolites, the number of parameters to be estimated can be prohibitive. For example, if $p = 500$ metabolites are detected, an evaluation of all pairwise conditional relationships would require the simultaneous estimation of 124,750 parameters. The use of regularization is a well-established approach for guaranteeing the existence of Gaussian Graphical Model parameters, amenable to the case that the sample size n is less than p (Banerjee, El Ghaoui, & d'Aspremont, 2008; J. Fan, Feng, & Wu, 2009; Friedman, Hastie, & Tibshirani, 2007; Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007).

Penalized estimation of GGM parameters provides a natural mechanism for integrating *a priori* knowledge regarding the molecular structure of metabolites with experimental metabolomics data. The integration of empirical data and scientific knowledge regarding metabolism is common in metabolomics studies. Typically, univariate and/or multivariate analyses first identify sets of metabolite features for which evidence of differences between experimental conditions or phenotypes are observed. After identifying interesting metabolite features, these sets are tested for enrichment of specific metabolic pathways or biological processes greater than that expected by chance (Xia & Wishart, 2010). A promising alternative to pathway analyses

discussed in (Dinesh Kumar Barupal & Fiehn, 2017) is to use structural similarity and chemical ontology as *a priori* knowledge to generate study-specific metabolite sets for contextualizing empirical results. The current work is of a similar paradigm and predicated on the assumption that the individual biochemical reactions that result in statistical dependence between metabolic intermediates also generate statistical dependence in structural similarity between the same intermediates. In our application, structural similarity is determined by an approach that considers overlap in shared local structure between metabolites. Rather than considering fixed sets of metabolites such as pathways, sets, or modules and subsequently quantifying enrichment of these sets in empirical results, we consider *a priori* knowledge of the relationships between metabolites as probabilistic statements about the relatedness of compounds. Thus, the *a priori* scientific knowledge is used to generate prior probability distributions that influence GGM model selection, so that posterior inference probabilistically combines empirical data and prior scientific knowledge to yield an updated model of the probabilistic interactions between metabolites. In the present work, we introduce a methodology for using molecular structure similarity to generate prior distributions that control the degree of penalization in parameter estimation for learning a GGM metabolite interactome from metabolomics data.

We first evaluated the methodology using simulation studies. For these simulation studies, autoregressive processes were simulated for representing linear biological processes in which the correlation between simulated metabolites decreased in tandem with decreasing structural similarity. We evaluated the sensitivity, specificity, AUC, and F_1 measure of the proposed method in recovering the true pairwise conditional correlations structures that were specified in advance. Finally, we applied our methodology to a human plasma dataset, specifically for the development of a reference model for stable heart disease.

2. Methods

2.1. Gaussian Graphical Models (GGM)

We consider Markov Random Fields (MRFs) which are graphical models in which random variables $X_i \in V, i = 1, 2, \dots, p$ are represented as vertices and edges in the edge set $E \subseteq V \times V$ represent probabilistic interactions. Gaussian Graphical Models (GGMs) represent a special class of MRFs in which the underlying joint probability distribution represented by the graph is assumed to be multivariate Gaussian (Koller & Friedman, 2009). In addition to the joint distribution being multivariate Gaussian, the marginal distribution for each X_i is Gaussian, as are the conditional distributions for $X_i|X_j$. Given a multivariate Gaussian distribution $\boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\mu}$ is a vector of means and $\boldsymbol{\Omega}$ is the inverse of the covariance matrix $\boldsymbol{\Sigma}$ (i.e. a concentration matrix), the entries ω_{ij} of the concentration matrix are of particular importance as $\omega_{ij} = 0$ implies that X_i and X_j are conditionally independent and with respect to the graph topology, there does not exist an edge between X_i and X_j. Further from the entries of $\boldsymbol{\Omega}$, the partial correlation coefficient between two random variables X_i and X_j can be computed as: $\rho_{ij|\cdot} = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$.

2.2. GGM parameter estimation

It has been shown previously that if $n < q$ where q represents the maximal clique size of the GGM then a maximum likelihood estimator does not exist (Buhl, 1993). Noting the likelihood function for the concentration matrix $\boldsymbol{\Omega}$:

$$l(\boldsymbol{\Omega}) = \log(\det \boldsymbol{\Omega}) - \left(\frac{\mathbf{S}}{n}\boldsymbol{\Omega}\right),$$

where $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ is the sum of products matrix. As the log-likelihood function is not guaranteed to be convex, regularization of this likelihood has been proposed (Banerjee et al., 2008; Friedman et al., 2007; Meinshausen & Bühlmann, 2006) as a solution for estimating $\boldsymbol{\Omega}$. Friedman et al. (2007) proposed a method, known as the graphical Lasso (Least Absolute Shrinkage and Selection Operator) for finding the maximum of the L_1 norm penalized log-likelihood:

$$l(\boldsymbol{\Omega}) = \log(\det \boldsymbol{\Omega}) - \left(\frac{\mathbf{S}}{n}\boldsymbol{\Omega}\right) - \rho\left|\left|\boldsymbol{\Omega}\right|\right|_1,$$

via a coordinate descent algorithm.

A Bayesian approach has been proposed for the regularized estimation of $\boldsymbol{\Omega}$ (H. Wang, 2012) that provides a natural structure for integrating *a priori* scientific knowledge and high-throughput molecular biology data such as untargeted metabolomics data. H. Wang (2012) introduced a hierarchical Bayesian representation of the regular graphical Lasso as well as the adaptive graphical Lasso (J. Fan et al., 2009). The frequentist adaptive graphical Lasso was devised to link the magnitude of the penalty parameter to the norm of individual concentration matrix entries and proposes the following penalized likelihood for $\boldsymbol{\Omega}$:

$$l(\boldsymbol{\Omega}) = \log(\det \boldsymbol{\Omega}) - \left(\frac{\mathbf{S}}{n}\boldsymbol{\Omega}\right) - \lambda \sum_{1 \le i \le p} \sum_{1 \le j \le p} w_{ij}|\omega_{ij}|$$

with weights $w_{ij} = 1/\left|\tilde{\omega}_{ij}\right|^{\alpha}$ where $\alpha > 0$ and $\tilde{\omega}_{ij}$are estimates for the concentration matrix entries, such as regular graphical Lasso estimates.

As a Bayesian model:

$$p(\mathbf{x}_i|\mathbf{\Omega}) = \boldsymbol{N}(\mathbf{0}, \mathbf{\Omega}^{-1}) \quad i = 1, 2, \dots, n$$

$$p\left(\mathbf{\Omega}|\{\lambda_{ij}\}_{i\leq j}\right) \sim C^{-1}_{\{\lambda_{ij}\}_{i\leq j}} \prod_{i<j} \mathrm{DE}(\omega_{ij}|\lambda_{ij}) \prod_{i=1}^{p} \mathrm{EXP}(\omega_{ii}|\lambda_{ii}/2) \cdot 1_{\mathbf{\Omega}\in M^+},$$

$$p\left(\{\lambda_{ij}\}_{i\leq j}|\,\{\lambda_{ii}\}_{i=1}^{p}\right) \propto C_{\{\lambda_{ij}\}_{i\leq j}} \prod_{i<j} \mathrm{GA}(r, s)$$

In the above model formulation for the density of $\mathbf{\Omega}$ conditional on the λ_{ij}, $\mathrm{DE}(\cdot\,|\lambda_{ij})$ represents the double exponential, or Laplace distribution, and $\mathrm{EXP}(\cdot\,|\lambda_{ij})$ the exponential distribution with scale parameter λ_{ij}. The space of positive definite matrices is represented by $1_{\mathbf{\Omega}\in M^+}$. Finally, C represents the normalizing constant so that $p\left(\mathbf{\Omega}|\{\lambda_{ij}\}_{i\leq j}\right)$ is a proper probability distribution. For the non-adaptive Bayesian graphical Lasso, $\lambda_{ij} = \lambda$ for all i and j, in other words the shrinkage parameter is not specific to each concentration matrix entry. H. Wang (2012) chooses a gamma prior for λ_{ij}, that is $\lambda_{ij} \sim \mathrm{Gamma}(r, s)$, where r and s are hyperparameters and develops a data-augmented block Gibbs sampler for sampling from the posterior distribution of $\mathbf{\Omega}$. Further, it is shown that the conditional distribution of the shrinkage parameter is then $\lambda_{ij}|\mathbf{\Omega} \sim \mathrm{GA}(1 + r, |\widehat{\omega}_{ij}| + s)$. In this case, the scale hyperparameter for the shrinkage varies with the norm of the current MCMC iteration estimate $\widehat{\omega}_{ij}$ for each ω_{ij} allowing for adaptive penalization as introduced by J. Fan et al. (2009). We propose that to incorporate prior knowledge regarding the relatedness of compounds, the scale hyperparameter can be linked to structural similarity, that is by specifying the prior distribution $\lambda_{ij} \sim \mathrm{Gamma}(r, s_{ij})$ where s_{ij} is a measure of structural similarity between compound i and compound j. The conditional expected value of each λ_{ij} is then: $E(\lambda_{ij}|\mathbf{\Omega}) = (1 + r)/(|\widehat{\omega}_{ij}| + s_{ij})$.

2.3. Generating informative priors from molecular structure

To generate informative shrinkage priors for the adaptive Bayesian graphical Lasso, we utilized a local structure similarity metric. This metric was adapted from the previously described Chemically Aware Substructure Search (CASS) algorithm (Mitchell, Fan, Lane, & Moseley, 2014). In this adaptation, the structural similarity between any two chemical structures (A and B) was estimated using strings representing local chemical structure (referred to as the atom's color) centered at every atom in the two structures. The color of every atom was constructed as follows. First, for every bonded atom, its element type and the order of the bond connecting it to the center atom are joined to form a component string that is added to a list of components. For example, if the center atom has a double bonded oxygen, this would contribute a 'O2' component to the component list. Every component represents a portion of the local bonded structure at the center atom. Second, the components strings are then sorted alphanumerically and concatenated to produce a description of the bonded structure one bond away from the center atom. Finally, to the front of this string, the element type of the center atom is then added to yield the atom's color. Each color uniquely maps to a single locally bonded structure (e.g. the 'CC1O1O2' coloring represents a carbon of a carboxylate). Since the component list was first sorted alphanumerically, this color is consistent for all identical local structures regardless of how they are ordered in their representation. Each chemical structure can be represented as the list of its constituent atom's colors and these lists of colors can be compared to determine structural similarity. To

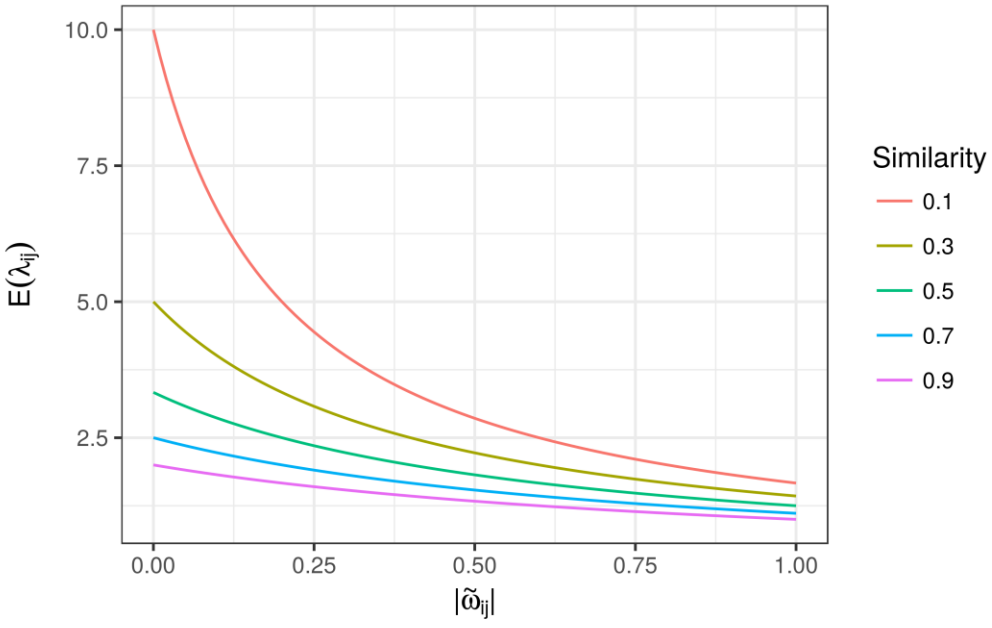

Figure 1: Theoretical relationship between the structural similarity of a pair of metabolites and the expected value of the shrinkage parameter λ_{ij} during model estimation. The horizontal axis represents a current estimate for a concentration matrix entry for a pair of metabolites. The vertical axis represents the expected value of the shrinkage parameter. Color values show the structural similarity between the pair of metabolites.

determine structural similarity between compounds using the color string representations, the Tanimoto coefficient between pairs of compounds was computed, which is defined as (Chen & Reynolds, 2002):

$$s(A,B) = \frac{\sum_{i=1}^{m} \min\left(n_i(A), n_i(B)\right)}{\sum_{i=1}^{m} n_i(A) + \sum_{i=1}^{m} n_i(B) - \sum_{i=1}^{m} \min\left(n_i(A), n_i(B)\right)}$$

where $n_i(A)$ represents the count of unique colored atoms indexed by $i = 1, 2, \ldots, m$ for molecule A. The Tanimoto dissimilarity is then $d(A,B) = 1 - s(A,B)$.

After determining the Tanimoto dissimilarity between each pair of metabolites, the gamma hyperprior distribution for the shrinkage parameter λ can be determined by linking the gamma distribution shape to the dissimilarity, that is by setting $s_{ij} = f\left(1 - d(i,j)\right)$ where i,j index metabolites and $f(x)$ is a monotonic function. The conditional distribution of the shrinkage parameter is then $\lambda_{ij}|\boldsymbol{\Omega} \sim \mathrm{GA}\left(1 + r, |\widehat{\omega}_{ij}| + s_{ij}\right)$. A plot of the relationship between the structural similarity of two hypothetical metabolites and the expected value of the shrinkage parameter is shown in Figure 1.

2.4. Posterior inference of model parameters

To determine a graphical model (or a set of models of high probability) given structural priors samples may be drawn from the posterior distribution of $p(\boldsymbol{\Omega}|\mathbf{X})$, using a Gibb's sampler similar to that introduced by Wang (2012). Previous work has shown that the exponential power family of probability distributions can be represented as a scale mixture of normal distributions with a defined mixing density (West, 1987). Using this fact and introducing the latent scale parameter τ, the unnormalized posterior distribution can be written as:

$$p(\boldsymbol{\Omega}, \boldsymbol{\tau}|\mathbf{X}, \boldsymbol{\Lambda}) \propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left(-\mathbf{tr}\left(\frac{1}{2}\mathbf{S}\boldsymbol{\Omega}\right)\right) \times \prod_{i<j}\left(\tau_{ij}^{-\frac{1}{2}} \exp\left(-\frac{\omega_{ij}^2}{2\tau_{ij}}\right) \exp\left(-\frac{1}{2}\lambda_{ij}^2\tau_{ij}\right)\right) \times \prod_{i=1}^{p} \exp\left(-\frac{1}{2}\lambda_{ij}\omega_{ij}\right) 1_{\boldsymbol{\Omega}\in M^+}$$

The block Gibb's sampler cycles through column-wise partitions of $\boldsymbol{\Omega}$, drawing from the conditional distribution of a single column of the matrix $\boldsymbol{\Omega}$, conditioned on the current values of the remaining columns. We developed an R package, *BayesianGLasso*, for implementing this and other samplers for the Bayesian Graphical Lasso. The underlying sampler was written in C++ using *Rcpp* and *RcppArmadillo* to make use of the Armadillo linear algebra library. In addition to providing Gibb's sampling methods the R package developed by our group includes classes for storing the Markov chains generated by the sampler along with relevant parameters and hyperparameters, and methods for conducting statistical inference over the simulated posterior distributions.

2.5. Efficacy analysis via simulation studies

To evaluate the efficacy of the proposed method, we employed simulation studies. We sought to evaluate the relative performance of the adaptive Bayesian Graphical Lasso (BGL) using informative priors versus (1) the adaptive Bayesian Graphical Lasso using non-informative priors, and (2) the Bayesian Graphical Lasso (non-adaptive). For the informative prior case, we further manipulated the degree to which the priors were accurate relative to the partial correlation structure utilized to generate the data. We evaluated the methods by simulating both simple partial correlation structures as well as more complex structures utilizing two simulation schemas. Under the first schema, a simple autoregressive (AR) process of order 1 was simulated for representing a linear biological process with decreasing structural similarity with increasing process distance. Simulated structural similarity was taken to be deterministically known, that is a structural similarity matrix was defined as: $\boldsymbol{\Sigma} = [\sigma_{ij}]$ where $\sigma_{ij} = \rho^{|i-j|}$. To simulate metabolite abundances, a random matrix was sampled from the multivariate normal distribution $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma})$. In the "accurate" informative prior case, the shrinkage hyperprior s_{ij} was defined as $s_{ij} = \omega_{ij}^{-1}$, where ω_{ij}^{-1} are the elements of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. After generating simulated datasets, the adaptive BGL (with informative and non-informative priors) as well as the non-adaptive BGL were utilized for estimating the concentration matrix and corresponding graph topology. Given the simple dependence structure in the AR(1) case, a measure of ground truth was available as the existence of edges between simulated metabolites was known *a priori*. We evaluated the sensitivity, specificity, and F_1 measure of each method for detecting the presence of edges by utilizing the magnitude of the estimated concentration matrix entries: $|\widehat{\omega}_{ij}|$. In addition, we report the area under the receiver operating characteristic curve for assessing each technique, which considers the range of possible fixed cutoff values of $|\widehat{\omega}_{ij}|$ for estimating the presence or absence of edges. While each technique draws shrinkage parameters from a Gamma distribution, the shape and scale of each distribution depends both on empirical data and hyperparameters. We conducted shape and scale hyperparameter optimization separately for each technique via a grid search over simulated datasets prior to the evaluation of performance.

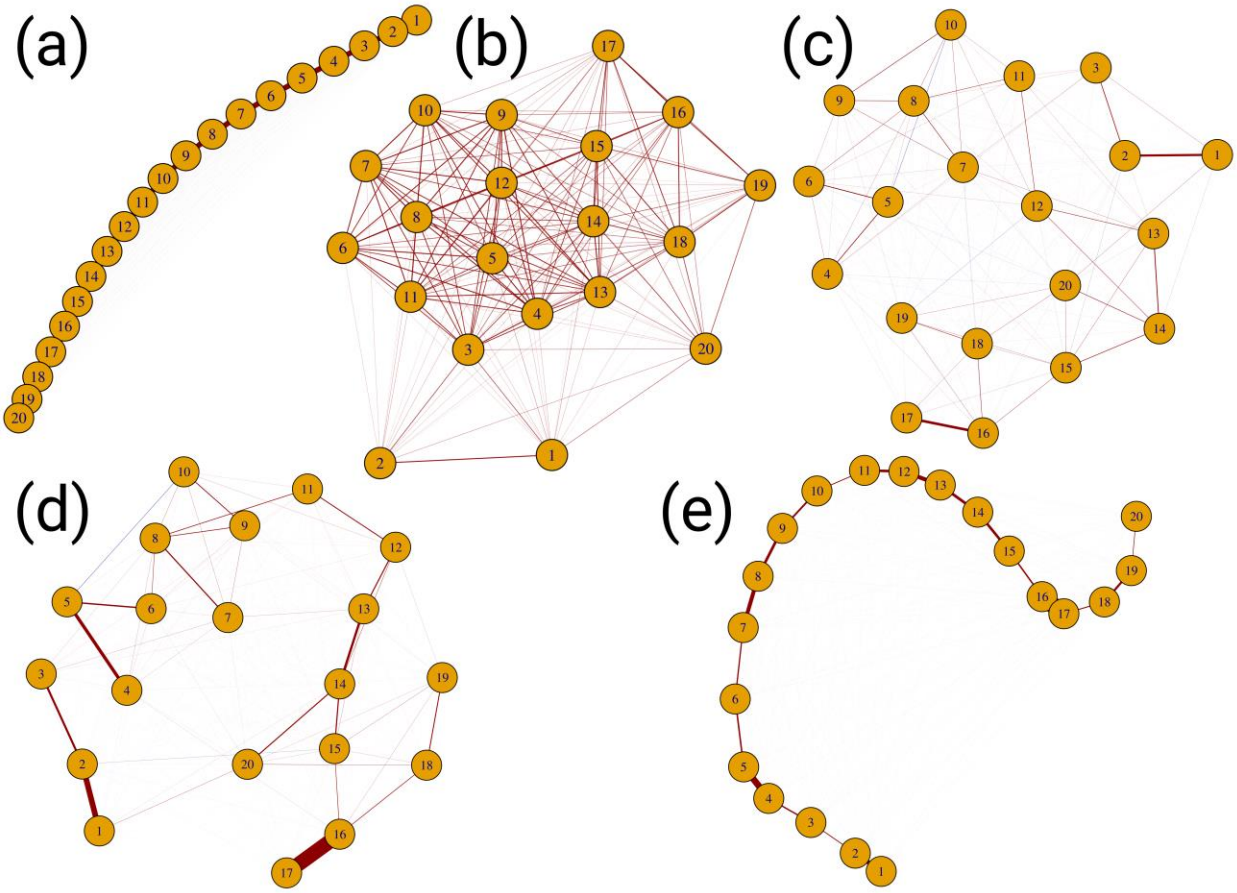

***Figure 2*:** True and estimated concentration matrix graphs for a randomly selected AR(1) simulation study with $p = 20$ simulated random variates (metabolites) and a simulated sample size of $n = 10$. Graphs represent the: (a) true concentration structure given an AR(1) covariance structure with $\rho = 0.95$, (b) sample covariance matrix, (c) concentration matrix estimated by the non-adaptive BGL, (d) concentration matrix estimated by the adaptive BGL with non-informative priors, (e) concentration matrix estimated by the adaptive BGL with chemical structure informative priors.

2.6. A plasma interactome for stable heart disease

In order to determine changes in the plasma metabolome associated with myocardial infarction (MI) characterized by thrombotic etiology versus non-thrombotic etiology, DeFilippis and colleagues assembled a human cohort as previously described (DeFilippis et al., 2016; DeFilippis et al., 2017; Trainor et al., 2017). Briefly, 80 human subjects presenting with suspected acute MI or stable coronary artery disease (CAD) were enrolled. Utilizing a stringent criteria based on clinical presentation, angiographic evidence, and histological evidence, MI subjects were adjudicated as thrombotic MI or non-thrombotic MI. Blood samples were collected at the time of acute presentation (presentation to the coronary artery catheterization lab prior to procedures) and at a follow-up evaluation approximately three months later. To estimate the structure of a stable heart disease plasma interactome, we used the follow-up evaluations from all available MI subjects as well as the evaluations from stable CAD subjects. The analytical sample thus consisted of 47 whole blood samples from human subjects with definitive heart disease who were not experiencing an acute event at the time of sampling.

Details of the metabolite quantification have been described previously (Trainor et al., 2017), but a brief overview is provided as follows. Plasma samples were prepared from whole blood and a recovery standard was added. Vigorous shaking was applied utilizing a GenoGrinder 2000 (Glen Mills, Metuchen, NJ) and methanol was added and to precipitate proteins. The extract containing small molecules was divided into five aliquots, four of which were analyzed using different platforms while the remaining aliquot was reserved. Two aliquots were analyzed by ultra-performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) with negative and positive ion mode electrospray ionization (ESI). A third aliquot was also analyzed by UPLC-MS/MS with negative ion mode ESI and a method optimized for polar metabolite detection. The fourth aliquot was analyzed by gas chromatography-mass spectrometry (GC-MS). 1,032 chemical species were detected utilizing the multiple platforms in the analysis of the plasma samples. Of these, 590 compounds were identified by matching to authentic standards based on retention index, mass to charge ratio, and MS2 data; 73 were identified based on experimental data matched to curated databases; and 369 could not be confidently

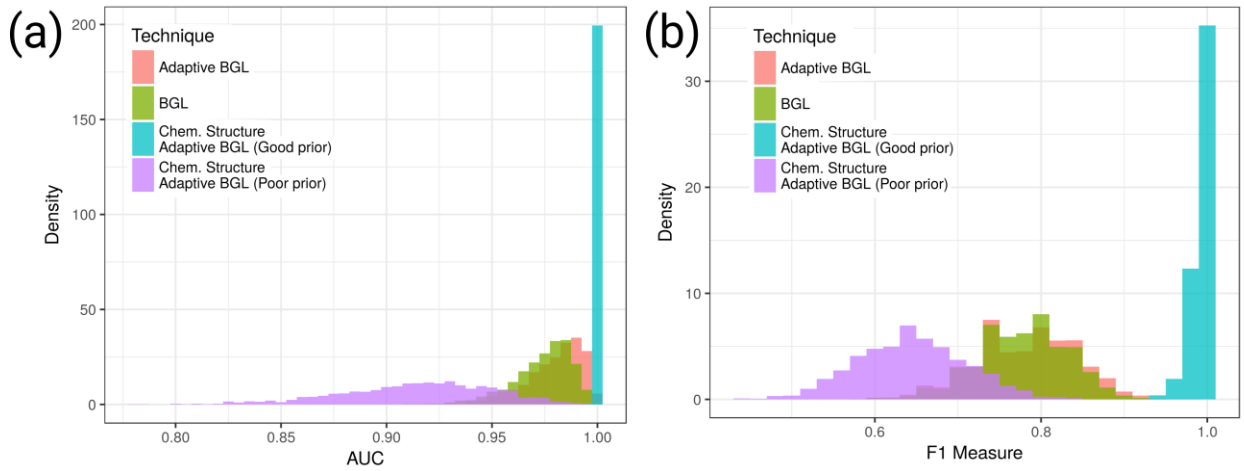

***Figure 3*:** Results of the AR(1) simulation studies. The comparative performance of the techniques (as the ability to detect an edge, given an edge is truly present) is presented. Subfigure (a) shows a histogram of the observed area under the receiver operating characteristic curve (AUC) values, while (b) shows the F1 measure.

***Table 1*:** Results of the AR(1) simulation studies. For each of the compared techniques, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and F_1 measure are reported. Reported values represent the sample mean and standard deviation over the simulation study replicates.

Technique	Sensitivity	Specificity	AUC	F1 Measure
Bayesian Graphical Lasso (BGL)	0.6792 ± 0.073	0.9896 ± 0.007	0.9740 ± 0.014	0.7786 ± 0.054
Adaptive BGL	0.6702 ± 0.079	0.9936 ± 0.006	0.9793 ± 0.014	0.7827 ± 0.062
Chemical Structure Adaptive BGL (Good prior)	0.9894 ± 0.019	0.9997 ± 0.001	1.000 ± 0.000	0.9938 ± 0.011
Chemical Structure Adaptive BGL (Poor prior)	0.8175 ± 0.068	0.8763 ± 0.033	0.9148 ± 0.036	0.6448 ± 0.066

identified. As the original data dependent acquisition was conducted utilizing both acute event samples and stable heart disease samples, metabolites not detected in the stable heart disease samples were removed. Metabolites missing from greater than 70% of the samples or without compound identification were also removed, resulting in a final dataset with 522 metabolites across 47 samples. Minimum values were then imputed for the remaining metabolite relative abundances with missing data. As many of the metabolites exhibited approximately log-normal relative abundance distributions, metabolite abundances were log-transformed. Finally, the data was mean centered so that each metabolite's relative abundance distribution was centered about zero.

To approximate the posterior distribution of $p(\mathbf{\Omega}|\mathbf{X})$, a Markov Chain was generated of length 1,000 with a 250 iteration burn-in period. From each sample from the posterior distribution $p(\mathbf{\Omega}|\mathbf{X})$, a partial correlation coefficient matrix was computed, yielding a simulated posterior distribution for the matrix of partial correlation coefficients.

3. Results

3.1. Simulation studies Ω

Results from the simulation studies given an autoregressive covariance structure are shown in Table 1 and Figures 2-3. In Figure 2 a graphical model representation of the underlying covariance structure is shown along with the graphical model representations of the sample covariance matrix and the concentration matrices estimated by the multiple techniques evaluated in this study. These figures were generated from a randomly sampled simulation study. GGM estimation by the Bayesian Graphical Lasso (BGL) and Adaptive BGL exhibited similar performance characteristics with respect to sensitivity, specificity, AUC, and F_1 measure. The performance of the chemical structure informative adaptive BGL varied significantly based on the suitability of the informative prior distribution for shrinkage parameters. In the "good prior" case in which it is assumed that the structural similarity and data generating process were deterministically linked, the structure adaptive BGL demonstrated significantly higher sensitivity, specificity, AUC, and F_1 measure than the other techniques. Conversely, in the "poor prior" case in which the relationship between the simulated structural similarity and the data generating process was masked by gaussian noise, average AUC and F_1 measure were significantly lower for the structure adaptive BGL than the remaining techniques.

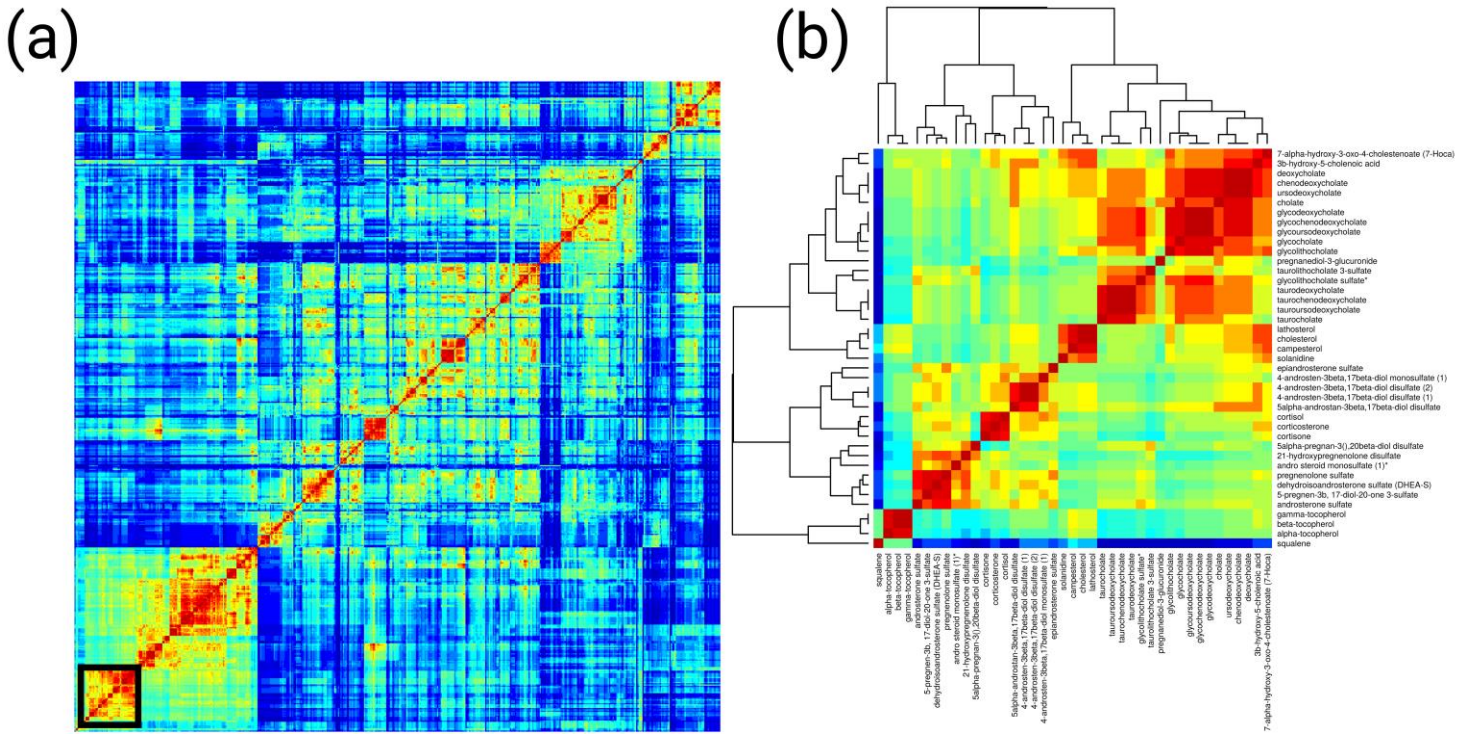

***Figure 4*:** Heatmap showing the structural similarity between metabolites detected in plasma of human subjects presenting with heart disease. A local substructure coloring approach was utilized for quantifying similarity as the degree of overlap (presence and absence) of local structures based on atom types and bonds. Dendrograms were generated via agglomerative hierarchical clustering utilizing Ward's minimum variance criterion and distances defined as 1 - similarity.

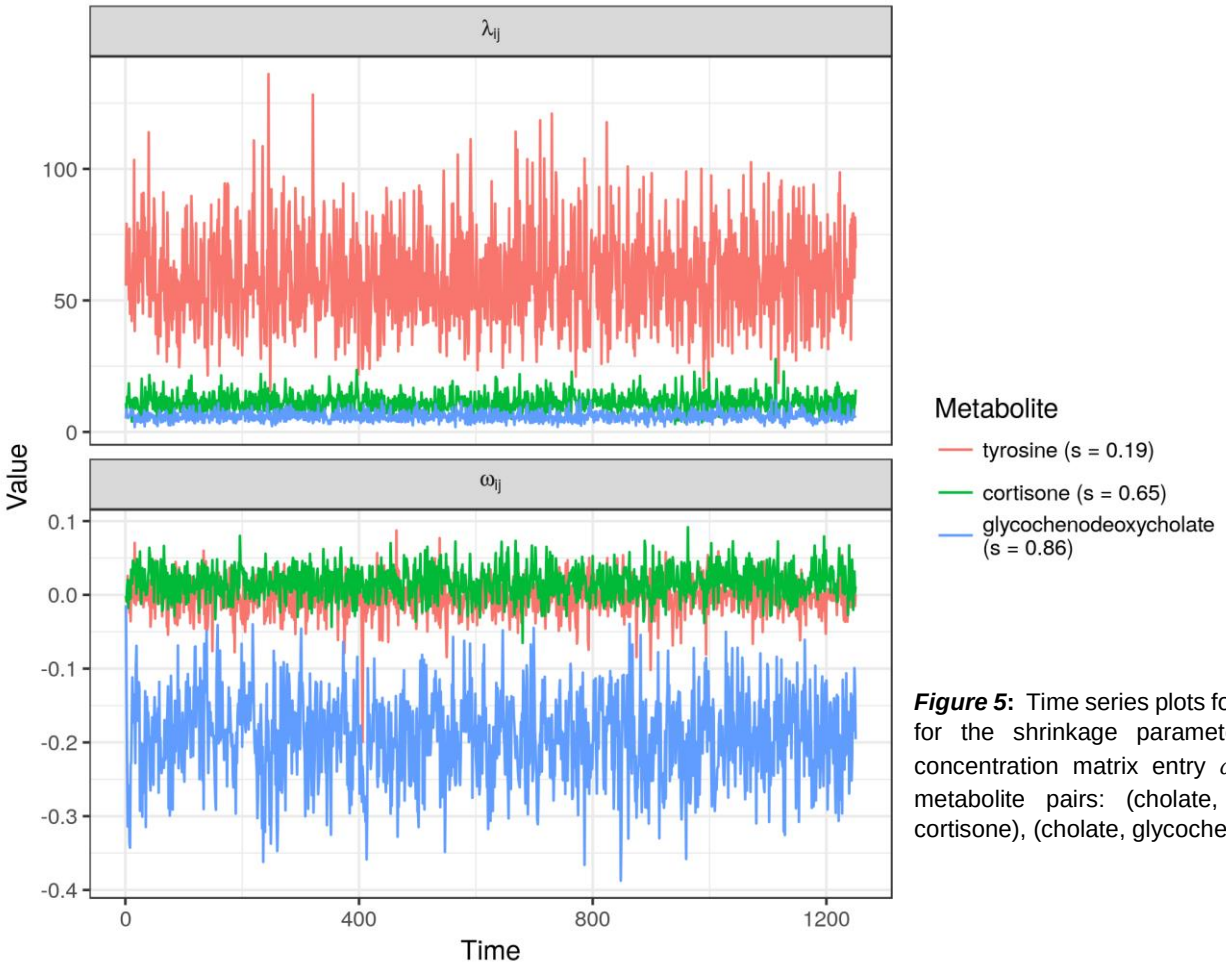

***Figure 5*:** Time series plots for the MCMC sampler for the shrinkage parameter λ_{ij} and for the concentration matrix entry ω_{ij} for the following metabolite pairs: (cholate, tyrosine), (cholate, cortisone), (cholate, glycochenodeoxycholate).

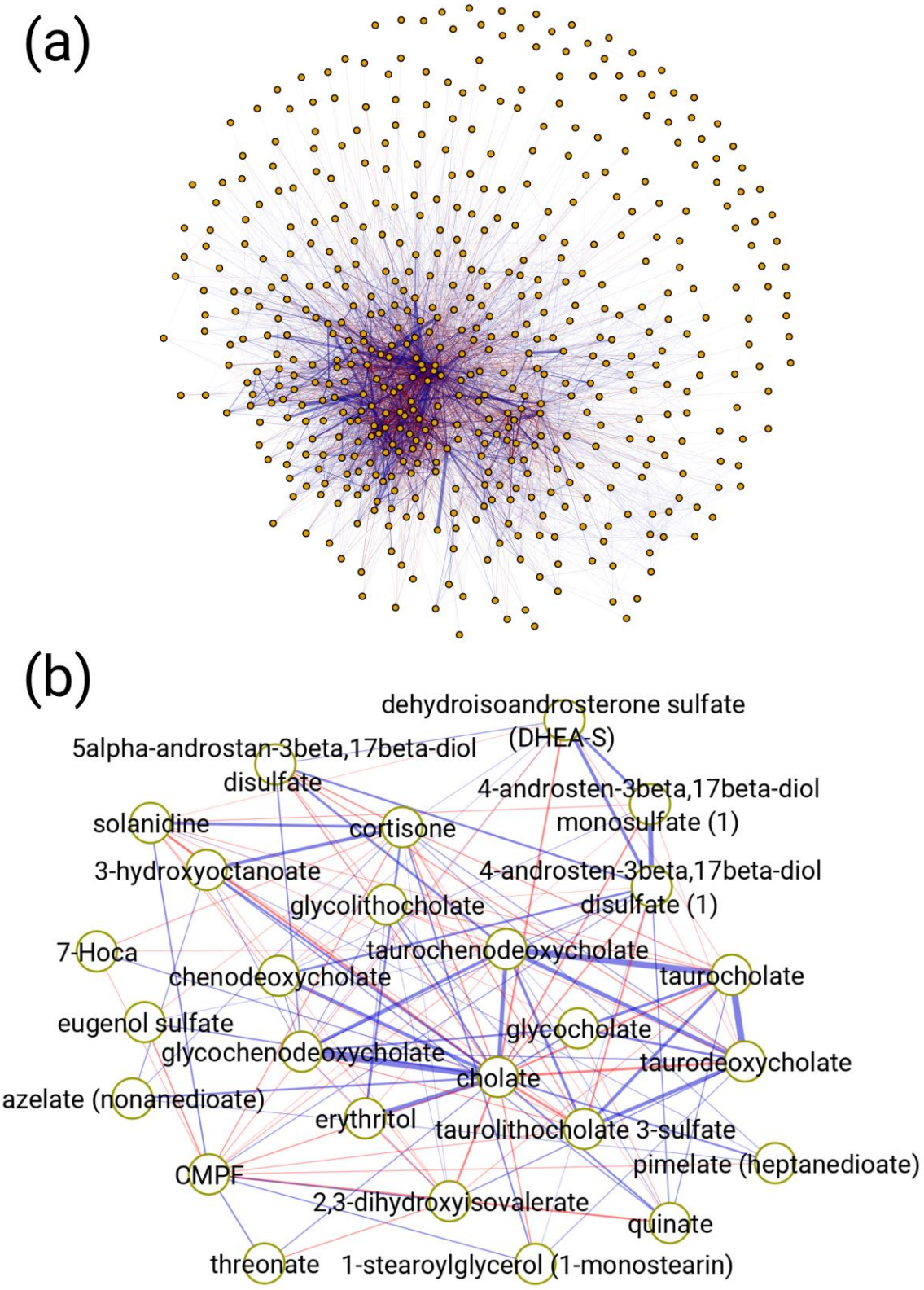

Figure 6: Graphical representation of the plasma metabolite interactome estimated by the chemical structure adaptive Bayesian Graphical Lasso (BGL) for stable heart disease. A simulated posterior distribution for the matrix of partial correlation coefficients was determined from the simulated posterior distribution of the concentration matrix $\mathbf{\Omega}$. Median values of each partial correlation coefficient were then determined and are represented as colored edges (negative values represented in red, positive values in blue). Subfigure (A) shows all metabolites in the interactome along with edges for which $|\rho| > 0.05$.

3.2. Plasma interactome for stable heart disease

A heatmap representation of the structural similarity between metabolites is shown in Figure 4. This heatmap was constructed using agglomerative hierarchical clustering using Ward's method and squared distances (with distance computed as $d_{ij} = 1 - s_{ij}$, where s_{ij} is the structural similarity between compounds). For illustrative purposes, the cluster containing cholate was retrieved from the root dendrogram by extracting the branch with height 0.6, as the structural-adaptive BGL subnetwork generated by cholate is considered later. Considering clusters generated by branches with low merge heights (high structural similarity), cholate was a member of a cluster with other closely related compounds such as deoxycholate, 3b-hydroxy-5-cholenoic acid, and glycocholate. Considering the more inclusive cluster generated by the branch at join height 3, other members included many intermediates in progestagen, androgen, glucocorticoid, and mineralocorticoid steroid metabolic pathways. These steroid hormone metabolites were all members of a cluster with similar within-cluster distances. Finally, at the same branch height that joined steroid hormone and cholate metabolites, a branch consisting of tocopherols cluster and squalene cluster was also joined.

Elements of the MCMC sampling iterations are presented in Figure 5. Continuing with the working example of the metabolite cholate, Markov chains are presented for the estimation of the concentration parameters for the pairs (cholate, tyrosine), (cholate, cortisone), and (cholate, glychochenodeoxycholate). These metabolites are highlighted as exemplars of metabolites with relatively low, medium, and relatively high chemical structure similarity with cholate. The time series of the MCMC sampling for the shrinkage parameter, λ_{ij}, demonstrated differences between the three pairs. Averaged across iterations, more shrinkage was applied with decreasing chemical similarity between the metabolite pairs. While the shrinkage parameter sample values for (cholate, cortisone) tended to be significantly smaller than the sample values for (cholate, cortisone), substantial overlap was observed in the posterior distribution of the concentration parameters for the same pairs.

From the simulated posterior distribution of $\mathbf{\Omega}$, the posterior mean $E(\mathbf{\Omega}|\mathbf{X})$ was estimated after discarding burn in iterations. The posterior mean of the distribution of partial correlation coefficients was also computed. The resulting plasma metabolite interactome inferred by the structure-adaptive Bayesian Graphical Lasso is presented in Figure 6. This figure presents both the entire graph representing the posterior mean partial correlations as well as the subgraph generated by considering the neighbors of cholate. For ease of viewing, only edges for which $|\rho_{ij}| > 0.05$ are plotted in the presentation of the full graph.

Positive partial correlation coefficients were observed between cholic acid the following other primary bile acids: glycocholic acid, chenodeoxycholic acid, and glycochenodeoxycholic acid. Negative partial correlation coefficients were observed between cholate and the following: taurocholic acid, taurodeoxycholic acid, and taurochenodeoxycholic acid. In addition, the bile acid 7-Hoca and the bile acid conjugate taurolithocholate 3-sulfate were first neighbors of cholic acid. Multiple conjugated androsterones were observed to be first neighbors of cholic acid as was the glucocorticoid cortisone and the steroidal alkaloid solanidine. Other metabolites that were first neighbors of cholic acid included: 3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid (CMPF), eugenol sulfate, erythritol, 2,3-dihydroxyisolvalerate, threonate, quinate, pimelate, and azelate.

4. Discussion

Making inferences regarding how metabolic processes differ between phenotypes is the ultimate goal of most metabolomics and systems biology studies. Yet, unlike comparing the concentration or abundance of one metabolite across two or more phenotypes, for which simple statistical tests such as t-tests, Wilcoxon Rank-Sum tests, or multi-group analogues are readily available, a statistical framework for determining if and how metabolic processes differ between phenotypes remains elusive. Both strictly empirical methods (e.g. correlation analyses) and *a priori* knowledge based approaches (e.g. pathway enrichment analyses) suffer from substantial flaws. In terms of empirical methods, the analysis of correlations (such as by the Pearson, Spearman or biweight midcorrelation coefficient) reveals the marginal associations between metabolites; however, these methods do not uncover the relationship between a pair of metabolites conditional on the abundances of the remaining metabolites. Gaussian graphical models have been proposed previously in the context of metabolomics (Krumsiek et al., 2011) as an alternative, as GGMs can be utilized to determine the partial or conditional relationship between metabolites. In their work Krumsiek et al. (2011) show that GGM edges (or concentration matrix entries) estimated from the analysis of blood serum samples from a large human cohort correspond to known metabolic pathway interactions. Consistent with this, our simulation studies illustrate the advantage of analyzing metabolite-metabolite interactions using the partial correlation coefficients from a GGM as opposed to correlation networks. In the case of an autoregressive correlation structure as might be observed given a linear metabolic pathway, correlation networks exhibit extremely high connectivity, and consequently could not be utilized to elucidate the order of reactions. In contrast, we observe high sensitivity and specificity in detecting the true edges using a Bayesian Graphical Lasso estimated GGM. While the approach utilized by Krumsiek et al.

(2011) was appropriate for the analysis of their data, it would not be possible to apply this approach in studies in which the sample size is smaller than the number of metabolites, as is common in many metabolomics studies. Frequentist regularization methods represent a class of solutions for ensuring that the concentration matrix, or equivalent GGM topology is estimable. In an implicit manner, frequentist regularization methods for estimating GGMs place a higher *a priori* probability on models with concentration matrix entries of smaller magnitude (H. Wang, 2012). However, this implicit prior cannot incorporate *a priori* knowledge as to whether some metabolites are more likely to be related than others.

In contrast to empirical methods, *a priori* knowledge based approaches such as pathway enrichment analyses consider the relationships between metabolites to be deterministically known which are then used to contextualize empirical results. Previous work (Dinesh Kumar Barupal & Fiehn, 2017; Dinesh K. Barupal et al., 2012) has highlighted that the coverage of metabolites detected in metabolomics studies in commonly utilized metabolic pathway and reaction databases may be extremely low. For example (Dinesh Kumar Barupal & Fiehn, 2017) observe that given 385 metabolites identified from the plasma of non-obese diabetic (NOD) mice, only 135 metabolites (or 35.1%) could be mapped to KEGG pathways (Kanehisa & Goto, 2000; Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016). To address this problem, Dinesh Kumar Barupal and Fiehn (2017) propose an alternative approach that utilizes both existing chemical ontological terms and chemical similarity between metabolites to develop coherent categories of metabolites for enrichment analyses. In the current work, we have sought a framework for balancing the benefits of empiricism with the benefits of *a priori* knowledge based approaches, while seeking to minimize the risks associated with both approaches. As opposed to considering metabolites as deterministically assigned to fixed pathways, our approach assumes that metabolites that are linked by biochemical reactions will exhibit overlap in local substructures. From this, our approach generates prior distributions for shrinkage parameters for the estimation of Gaussian graphical models. The posterior distribution of GGM parameters is thus proportional to the likelihood of the concentration matrix parameters (or the partial correlations between metabolites) times the prior probability of the concentration matrix parameters (which are linked to the structural similarity between metabolites). Similar to the non-informative BGL approach, this approach ensures that the concentration matrix is estimable via the Bayesian analog of regularization, however the regularization is applied given the prior belief that stronger associations are *a priori* more likely given structurally related compounds than unrelated compounds. While we find better justification for using structural similarity to generate prior probability distributions for shrinkage parameters in estimating a GGM, this approach would generalize to the use of priors from metabolic pathway maps. A previous work sought to estimate a GGM using 17 compounds quantified by NMR from 24 microglia cell culture samples using priors determined from KEGG (Peterson et al., 2013).

In addition to evaluation via simulation studies, we have applied the chemical structure adaptive BGL to generate a media-specific (blood plasma) metabolite interactome for stable heart disease. This model may serve as a reference model for comparing how the probabilistic interactions between metabolites in circulation change during acute disease events such as myocardial infarction or unstable angina. From this model, we have observed probabilistic interactions that are consistent with previous research in metabolism, as can be observed by focusing on the metabolite cholate. Bile acids are the major catabolic intermediate of cholesterol (Russell, 2003). Within mammals, the bile acid pool consists of primary bile acids such as cholic acid and chenodeoxycholic acid which are synthesized from cholesterol by enzymes expressed in hepatocytes, as well as secondary bile acids that are synthesized from primary bile acids by bacteria in the gut (García-Cañaveras, Donato, Castell, & Lahoz, 2012; Hofmann, Hagey, & Krasowski, 2010; Russell, 2003). In addition to bile acids aiding in the digestion of nutrients in the gut, bile acids also act as signaling molecules that have been shown to regulate glucose and lipid metabolism (Ferrebee & Dawson, 2015; Khurana, Raufman, & Pallone, 2011). Given the substantial proportion of cholesterol that is converted to bile acids leading to elimination, bile acid metabolism is linked to atherosclerosis (Meissner et al., 2013). In addition bile acids as signaling molecules affect cardiac (Desai et al., 2017; Rainer et al., 2013) and circulatory physiology (Khurana et al., 2011) via direct effects such as taurodeoxycholic acid mediated vasodilation (Khurana, Yamada, Wess, Kennedy, & Raufman, 2005). With respect to the current work, we observed relatively strong partial correlation between cholic acid and other primary bile acids. Additionally, partial correlations were observed between cholic acid and steroid hormones that share cholesterol as a common precursor. Given the importance of bile acids in cholesterol metabolism, atherosclerosis, cardiac physiology and circulatory physiology, a reference model of the probabilistic interactions of bile acids in circulation can help elucidate how acute disease events impact bile acid metabolism.

As with any Bayesian approach, the choice of prior probability distribution has a direct influence on the posterior distribution of model parameters (Gelman, 2014). In the current work, we have utilized informative priors that are linked to chemical structure similarity. This represents a potential limitation of the current work. Over the course of the simulation studies, we observed, unsurprisingly, that by introducing random noise into the

simulated structural similarity the performance of the chemical structure adaptive BGL deteriorated. Further, the performance of the technique given “poor” prior information was, on average, worse than the performance of techniques such as the non-adaptive BGL that rely on non-informative priors. One element of the chemical structure adaptive BGL is worth noting in this context. In our proposed formulation, other monotonic functions for relating structural similarity to the Gamma scale parameter may be employed, as well as different shape and scale hyperparameters can be utilized. In this manner, the experimenter can diminish or strengthen the degree to which structural similarity impacts shrinkage. A second limitation of the current work is the choice of a multivariate Gaussian distribution for representing the joint distribution of metabolite abundances. While transformations in the stable heart disease data were applied over each metabolite to reduce the degree of departure from normality, the underlying intensity data is not normally distributed. Further, there are many cases in which approximate normality is not an achievable aim. A metabolite that is only present in some samples (e.g. acetaminophen metabolites that are present in some human subjects who have taken this medication, but not others) is one such case. Following a missing value imputation procedure, such a metabolite would exhibit a bimodal distribution that would not be well described by a Gaussian model.

References

Arbab-Zadeh, A., & Fuster, V. (2015). The Myth of the “Vulnerable Plaque”. *Journal of the American College of Cardiology, 65*(8), 846-855. doi:10.1016/j.jacc.2014.11.041

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *J. Mach. Learn. Res., 9*, 485-516.

Barupal, D. K., & Fiehn, O. (2017). Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-15231-w

Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E., & Fiehn, O. (2012). MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics, 13*(1), 99. doi:10.1186/1471-2105-13-99

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., . . . Muntner, P. (2017). Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association. *Circulation, 135*(10), e146-e603. doi:10.1161/cir.0000000000000485

Buhl, S. L. (1993). On the Existence of Maximum Likelihood Estimators for Graphical Gaussian Models. *Scandinavian Journal of Statistics, 20*(3), 263-270.

Chen, X., & Reynolds, C. H. (2002). Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *Journal of Chemical Information and Computer Sciences, 42*(6), 1407-1414. doi:10.1021/ci025531g

Dallmann, R., Viola, A. U., Tarokh, L., Cajochen, C., & Brown, S. A. (2012). The human circadian metabolome. *Proceedings of the National Academy of Sciences, 109*(7), 2625-2629. doi:10.1073/pnas.1114410109

DeFilippis, A. P., Chernyavskiy, I., Amraotkar, A. R., Trainor, P. J., Kothari, S., Ismail, I., . . . Bhatnagar, A. (2016). Circulating levels of plasminogen and oxidized phospholipids bound to plasminogen distinguish between atherothrombotic and non-atherothrombotic myocardial infarction. *J Thromb Thrombolysis, 42*(1), 61-76. doi:10.1007/s11239-015-1292-5

DeFilippis, A. P., Trainor, P. J., Hill, B. G., Amraotkar, A. R., Rai, S. N., Hirsch, G. A., . . . Bhatnagar, A. (2017). Identification of a plasma metabolomic signature of thrombotic myocardial infarction that is distinct from non-thrombotic myocardial infarction and stable coronary artery disease. *PLoS One, 12*(4), e0175591. doi:10.1371/journal.pone.0175591

Desai, M. S., Mathur, B., Eblimit, Z., Vasquez, H., Taegtmeyer, H., Karpen, S. J., . . . Anakk, S. (2017). Bile acid excess induces cardiomyopathy and metabolic dysfunctions in the heart. *Hepatology, 65*(1), 189-201. doi:10.1002/hep.28890

Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelena, E., . . . Kell, D. B. (2014). Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics, 11*(1), 9-26. doi:10.1007/s11306-014-0707-1

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics, 3*(2), 521-541. doi:10.1214/08-aoas215

Fan, Y., Li, Y., Chen, Y., Zhao, Y.-J., Liu, L.-W., Li, J., . . . Qi, L.-W. (2016). Comprehensive Metabolomic Characterization of Coronary Artery Diseases. *Journal of the American College of Cardiology, 68*(12), 1281-1293. doi:10.1016/j.jacc.2016.06.044

Ferrebee, C. B., & Dawson, P. A. (2015). Metabolic effects of intestinal absorption and enterohepatic cycling of bile acids. *Acta Pharmaceutica Sinica B, 5*(2), 129-134. doi:10.1016/j.apsb.2015.01.001

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432-441. doi:10.1093/biostatistics/kxm045

Gao, X., Jia, M., Zhang, Y., Breitling, L. P., & Brenner, H. (2015). DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clinical Epigenetics, 7*(1). doi:10.1186/s13148-015-0148-3

García-Cañaveras, J. C., Donato, M. T., Castell, J. V., & Lahoz, A. (2012). Targeted profiling of circulating and hepatic bile acids in human, mouse, and rat using a UPLC-MRM-MS-validated method. *J Lipid Res, 53*(10), 2231-2241. doi:10.1194/jlr.D028803

Gelman, A. (2014). *Bayesian data analysis* (Third edition. ed.). Boca Raton: CRC Press.

Hofmann, A. F., Hagey, L. R., & Krasowski, M. D. (2010). Bile salts of vertebrates: structural variation and possible evolutionary significance. *J Lipid Res, 51*(2), 226-246. doi:10.1194/jlr.R000042

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res, 44*(D1), D457-462. doi:10.1093/nar/gkv1070

Kasper, D. L. (2015). *Harrison's principles of internal medicine* (19th edition / editors, Dennis L. Kasper, MD, William Ellery Channing, Professor of Medicine, Professor of Microbiology, Department of Microbiology and Immunobiology, Harvard Medical School, Division of Infectious Diseases, Brigham and Women's Hospital, Boston, Massachusetts and five others . ed.). New York: McGraw Hill Education.

Keating, S. T., & El-Osta, A. (2015). Epigenetics and Metabolism. *Circulation Research, 116*(4), 715-736. doi:10.1161/circresaha.116.303936

Khurana, S., Raufman, J.-P., & Pallone, T. L. (2011). Bile Acids Regulate Cardiovascular Function. *Clinical and Translational Science, 4*(3), 210-218. doi:10.1111/j.1752-8062.2011.00272.x

Khurana, S., Yamada, M., Wess, J., Kennedy, R. H., & Raufman, J.-P. (2005). Deoxycholyltaurine-induced vasodilation of rodent aorta is nitric oxide- and muscarinic M3 receptor-dependent. *European Journal of Pharmacology, 517*(1-2), 103-110. doi:10.1016/j.ejphar.2005.05.037

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. Cambridge, MA: MIT Press.

Kotze, H. L., Armitage, E. G., Sharkey, K. J., Allwood, J. W., Dunn, W. B., Williams, K. J., & Goodacre, R. (2013). A novel untargeted metabolomics correlation-based network analysis incorporating human metabolic reconstructions. *BMC Systems Biology, 7*(1), 107. doi:10.1186/1752-0509-7-107

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology, 5*(1), 21. doi:10.1186/1752-0509-5-21

Krycer, J. R., Yugi, K., Hirayama, A., Fazakerley, D. J., Quek, L.-E., Scalzo, R., . . . James, D. E. (2017). Dynamic Metabolomics Reveals that Insulin Primes the Adipocyte for Glucose Metabolism. *Cell Reports, 21*(12), 3536-3547. doi:10.1016/j.celrep.2017.11.085

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989-2001. doi:10.1016/j.jmva.2009.04.008

Madhu, B., Narita, M., Jauhiainen, A., Menon, S., Stubbs, M., Tavaré, S., . . . Griffiths, J. R. (2015). Metabolomic changes during cellular transformation monitored by metabolite–metabolite correlation analysis and correlated with gene expression. *Metabolomics, 11*(6), 1848-1863. doi:10.1007/s11306-015-0838-z

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science, 349*(6255), 1483-1489. doi:10.1126/science.aab4082

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics, 34*(3), 1436-1462. doi:10.1214/009053606000000281

Meissner, M., Wolters, H., de Boer, R. A., Havinga, R., Boverhof, R., Bloks, V. W., . . . Groen, A. K. (2013). Bile acid sequestration normalizes plasma cholesterol and reduces atherosclerosis in hypercholesterolemic mice. No additional effect of physical activity. *Atherosclerosis, 228*(1), 117-123. doi:10.1016/j.atherosclerosis.2013.02.021

Mitchell, J. M., Fan, T. W. M., Lane, A. N., & Moseley, H. N. B. (2014). Development and in silico evaluation of large-scale metabolite identification methods using functional group detection for metabolomics. *Frontiers in Genetics, 5*. doi:10.3389/fgene.2014.00237

Moriya, T., Satomi, Y., Murata, S., Sawada, H., & Kobayashi, H. (2017). Effect of gut microbiota on host whole metabolome. *Metabolomics, 13*(9). doi:10.1007/s11306-017-1240-9

Peterson, C., Vannucci, M., Karakas, C., Choi, W., Ma, L., & Maletic-Savatic, M. (2013). Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. *Statistics and Its Interface, 6*(4), 547-558. doi:10.4310/SII.2013.v6.n4.a12

Rainer, P. P., Primessnig, U., Harenkamp, S., Doleschal, B., Wallner, M., Fauler, G., . . . von Lewinski, D. (2013). Bile acids induce arrhythmias in human atrial myocardium—implications for altered serum bile acid composition in patients with atrial fibrillation. *Heart, 99*(22), 1685-1692. doi:10.1136/heartjnl-2013-304163

Russell, D. W. (2003). The Enzymes, Regulation, and Genetics of Bile Acid Synthesis. *Annual Review of Biochemistry, 72*(1), 137-174. doi:10.1146/annurev.biochem.72.121801.161712

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., . . . Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nature Genetics, 46*(6), 543-550. doi:10.1038/ng.2982

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., & Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology, 26*(9), 1003-1010. doi:10.1038/nbt.1487

Southam, A. D., Lange, A., Al-Salhi, R., Hill, E. M., Tyler, C. R., & Viant, M. R. (2014). Distinguishing between the metabolome and xenobiotic exposome in environmental field samples analysed by direct-infusion mass

spectrometry based metabolomics and lipidomics. *Metabolomics, 10*(6), 1050-1058. doi:10.1007/s11306-014-0693-3

Suarez-Diez, M., Adam, J., Adamski, J., Chasapi, S. A., Luchinat, C., Peters, A., . . . Saccenti, E. (2017). Plasma and Serum Metabolite Association Networks: Comparability within and between Studies Using NMR and MS Profiling. *Journal of Proteome Research, 16*(7), 2547-2559. doi:10.1021/acs.jproteome.7b00106

Trainor, P. J., Hill, B. G., Carlisle, S. M., Rouchka, E. C., Rai, S. N., Bhatnagar, A., & DeFilippis, A. P. (2017). Systems characterization of differential plasma metabolome perturbations following thrombotic and non-thrombotic myocardial infarction. *Journal of Proteomics*. doi:10.1016/j.jprot.2017.03.014

Voet, D., Voet, J. G., & Pratt, C. W. (2013). *Fundamentals of biochemistry : life at the molecular level* (4th ed.). Hoboken, NJ: Wiley.

Wang, H. (2012). Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis, 7*(4), 867-886. doi:10.1214/12-ba729

Wang, L., Hou, E., Wang, L., Wang, Y., Yang, L., Zheng, X., . . . Tian, Z. (2015). Reconstruction and analysis of correlation networks based on GC–MS metabolomics data for young hypertensive men. *Analytica Chimica Acta, 854*, 95-105. doi:10.1016/j.aca.2014.11.009

West, M. (1987). On Scale Mixtures of Normal Distributions. *Biometrika, 74*(3), 646. doi:10.1093/biomet/74.3.646

Xia, J., & Wishart, D. S. (2010). MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics, 26*(18), 2342-2344. doi:10.1093/bioinformatics/btq418

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika, 94*(1), 19-35. doi:10.1093/biomet/asm018