Metabolic Network Construction Using Ensemble Algorithms

Seongho Kim*
Biostatistics Core
Karmanos Cancer Institute
Department of Oncology
Wayne State University
Detroit, MI 48201, USA

Joohyoung Lee
Department of Family
Medicine and Public Health
Sciences
Wayne State University
Detroit, MI 48201, USA

Hyejeong Jang
Biostatistics Core
Karmanos Cancer Institute
Department of Oncology
Wayne State University
Detroit, MI 48201, USA

Xiang Zhang
Department of Chemistry
University of Louisville
Louisville, KY 40292, USA

ABSTRACT

One of the most important and challenging "knowledge extraction" tasks in bioinformatics is the reverse engineering of genes, proteins, and metabolites networks from biological data. Gaussian graphical models (GGMs) have been proven to be a very powerful formalism to infer biological networks. Standard GGM selection techniques can unfortunately not be used in the "small N, large P" data setting. Various methods to overcome this issue have been developed based on regularized estimation, partial least squares method, and limited-order partial correlation graphs. Several studies compared the performances among several network construction algorithms, such as PLSR, SCE, and ES, ICR and PCR, Ridge regression, Lasso and adaptive Lasso, to see which method is the best for biological network constructions. Each comparison analysis resulted in that each construction method has its own advantages as well as disadvantages according to different circumstances, such as the network complexity. However, it is almost impossible to recognize the complexity of the network before estimation. Thus, we develop an Ensemble method which is model averaging to construct a metabolic network. Our simulation studies show that the ensemble averaging based network construction has F1 score larger than these of other methods except only for Adaptive Lasso, reflecting its ability to account for uncertainty of network complexity.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; [**Special Track II**]: Bioinformatics

General Terms

Algorithm

Keywords

Ensemble averaging, Metabolomics, Network construction

1. INTRODUCTION

Construction of metabolic association networks is a critical data analysis step in systems biology. The metabolic association network is a collection of metabolite relations during cellular processes. A relatively smaller number of studies have been reported for metabolic network construction. Arkin et al. [1] predicted interactions within reaction networks over time for the glycolytic pathway. Steuer et al. [13] examined the relationship between data generated from networks and biochemical pathways using potato plant metabolism. Ursem et al. [16] constructed the metabolic networks from metabolite abundance in different tomato genotypes. All of these studies used the Pearson's correlation coefficients to construct the metabolic networks. A major drawback of Pearson's correlation-based networks is unable to distinguish between the direct and the indirect associations. On the other hand, Gaussian graphical models (GGMs) reveal direct associations with conditional independence/dependence among variables, using partial correlation coefficients that are calculated by the correlation of two variables after removing the effect of other variables [3]. GGMs have been employed in metabolomics for several studies. Greenberg et al. [4] used the pseudo-inverse method to estimate the partial correlation for the study of the influence of enzyme evolution on Drosophila metabolic pathway. Chan et al. [2] also constructed the metabolic network to quantify metabolites present in Arabidopsis thaliana using the first-order correlation where the effects of only one variable are removed. Theis et al. [14] used GGMs for reconstructing pathway reactions from human population cohort when the size of samples (experiments) was larger than the number of variables (metabolites).

The key idea behind GGMs is to use partial correlations as a measure of independence of any two variables (metabolite peaks). This makes it straightforward to distinguish the

*Corresponding author's email: kimse@karmanos.org direct interactions from the indirect interactions. Note that the partial correlations are related to the inverse of the correlation matrix and the missing edges indicate conditional independence. Application of GGMs to metabolomic data is quite challenging because the number of metabolites (P) is usually much larger than the number of available samples (N) (“small N, large P”), and the classical/standard GGM theory is not valid in a small sample setting. To resolve this difficulty, some methods have been introduced mainly in gene expression analysis. All of these methods can be categorized into three categories: (i) analysis with classic GGM theory, (ii) using limited order partial correlations, and (iii) application of regularized GGMs including partial least squares based methods. For small N, large P data, the methods from category (iii) are popular [8]. Several studies compared the performances among several network construction algorithms, such as the partial least squares regression (PLSR), shrinkage covariance estimator (SCE), and extrinsic similarity (ES), Independent component and principle component regression analyses (ICR and PCR, respectively), Ridge regression, Lasso and adaptive Lasso, to see which method is the best for biological network constructions [7]. All comparison studies concluded that each construction method has its own advantages and disadvantages according to different circumstances, such as the network complexity. Therefore, it is highly demanded to account for uncertainty of network complexity, giving an insight for an ensemble averaging.

For the aforementioned reasons, we develop a metabolic network construction algorithm using ensemble model averaging (EMA). Using EMA has several benefits such as improved performance over any single network and the lessened risk of overfitting [6]. In particular, Madigan and Raftery[10] demonstrated that EMA performs better average predictive ability than using any single model in terms of a logarithmic scoring rule.

2. METHODS

2.1 Single network construction

Let $X = (x_{ij}) \in R^{n\times p}$ be a data matrix with p metabolites (variables) and n experimental sample size, x_i a column vector as $(x_{i1}, \cdots, x_{ip})^T$, and the sample mean vector $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_{ij} is the ith observation on the jth random variable and $p < n$. The sample variance-covariance matrix is $p \times p$ matrix defined by

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

Then a partial correlation coefficient $\rho_{ij}, i, j = 1, \cdots, p$, is defined from an inverse variance-covariance matrix $S^{-1} = (\theta_{ij})$ as follows [17]:

$$\rho_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}, i \neq j$$

Once all the partial correlation coefficients are estimated, statistical testing should be carried out over each of the estimated partial correlation coefficients to find significant associations. Then, the association network is constructed using the edges with the significant partial correlation coefficients. However, the statistical procedure to identify significant partial correlations involves a multiple comparison problem since many hypotheses are tested simultaneously. To overcome this difficulty, false discovery rate (FDR) control is used in network construction.

Sch$\ddot{a}$fer and Strimmer [12] proposed shrinkage covariance estimator (SCE) to estimate the partial correlation when the variance-covariance matrix S is singular. Under singularity of covariance matrix, an alternative method is to trade off the unbiased sample covariance S and low dimensional shrinkage target matrix T; $\hat{S} = sT + (1-s)S$,where $s \in (0, 1]$ is shrinkage intensity. The optimal value of the tuning parameter s is analytically determined and estimated from the data. For a more detailed description, refer to Sch$\ddot{a}$fer and Strimmer [12].

Principle component regression (PCR) and partial least squares regression (PLSR) [11] circumvent high-dimensional problem by decomposing a data matrix X into orthogonal scores T and loadings P, $X = TP^T + X_R$, and regressing dependent variable Y on the first r important columns of the scores T , where X_R is the remains of decomposition. The differences between PCR and PLSR are as follows: PLSR uses both dependent and independent variables to reduce data dimension, while PCR uses only independent variables, and PCR/PLSR finds orthogonal features based on the normality assumption.

Ridge regression is a shrinkage method which imposes a penalty on the size of regression coefficients. The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\beta} = \arg\min_{\beta}\left\{\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\},$$

where $\lambda \geq 0$, which is a complexity parameter and controls the amount of shrinkage. A larger value of λ results in a great amount of shrinkage. The coefficients are shrunk toward zero (and each other) [5].

The Lasso (Least Absolute Shrinkage and Selection Operator), which was first proposed by [15], is a shrinkage method like ridge, but it has subtle and important differences from the ridge regression. The Lasso is a penalized least squares procedure that minimizes residual sum of squares subject to the non-differentiable constraint expressed in terms of the $L1$ norm of the coefficients [9]. That is, the Lasso estimator is given by

$$\hat{\beta} = \arg\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

This $L1$ norm constraint makes the solutions nonlinear in the y_i, resulting in no analytical solution different from ridge regression.

The adaptive lass was proposed as a means for fitting models sparser than Lasso. The advantages of the Adaptive Lasso are that given the pilot estimates, such as the univariate regression coefficients, its objective function becomes convex in parameters and that it recovers the true model under more general conditions than does the lass if the pilot estimates are $\sqrt{N}$ consistent.

2.2 Ensemble network construction

Gaussian Graphical model (GGM)-based methods have been widely used in genomics to infer biological networks. However, the performance of various GGM-based methods for the construction of metabolic association networks remains unknown in metabolomics. For this reason, the performance of PCR, ICR, SCE, PLSR, and ES methods in constructing metabolic association networks was compared by estimating partial correlation coefficient matrices when the number of variables was larger than the sample size in our previous study [7]. The previous study demonstrated that PCR and ICR discover more significant edges and perform better than PLSR and SCE, when the discovered edges are evaluated using KEGG pathway. These results suggest that the metabolic network is more complex than the genomic network and therefore, PCR and ICR have the advantage over PLSR and SCE in constructing the metabolic association networks. Overall, this study showed that the network complexity seems to play a more important role in the relative performances among different construction methods, as mentioned before. However, the network complexity cannot be inferred until network construction. Therefore, we propose an ensemble averaging approach to deal with the uncertainty of the network complexity. To this end, we use the aforementioned six methods to incorporate the ensemble averaging method: SCE (1st method), PCR (2nd method), PLSR (3rd method), Lasso (4th method), Adaptive Lasso (5th method), and Ridge regression (6th method). In other words, we take the average of each association over the six partial correlations estimated. That is, for each jth association $\rho(j)$, we estimated the six partial correlations $(\hat{\rho}_1(j), \hat{\rho}_2(j), \cdots, \hat{\rho}_6(j))$, where $\hat{\rho}_k(j)$ is the estimated partial correlation of the jth association by the kth method. Then the average association of the kth pair of two metabolites (peaks) was obtained by

$$\tilde{\rho}(j) = \sum_{k=1}^{6} w_i \hat{\rho}_i(k), j = 1, 2, \cdots, J,$$

where $w_k = \frac{n_k}{\sum_{l=1}^{6} n_l}$, $n_k = \sum_{j=1}^{J} \hat{\rho}_k(j)$, and J is the total number of edges. Then we apply the false discovery rate (FDR) for each $\tilde{\rho}(j)$, $j = 1, \cdots, J$, to find the significant edges.

2.3 Performance criteria

The true positive (TP) is the number of elements whose true value and predicted outcome are positive, the false negative (FN) is the number of elements whose true value is positive but predicted outcome is negative, and the false positive (FP) is the number of elements whose true value is negative but predicted outcome is positive. The performances of all methods were then evaluated using the following four criteria:

- The true positive rate (TPR): TPR is the proportion of the actual positives which are correctly predicted: $TPR = \frac{TP}{TP+FN}$.
- The positive predictive value (PPV): PPV is the proportion of subjects with positive output results which are correctly predicted: $PPV = \frac{TP}{TP+FP}$.
- F1 score: It is a measure of accuracy, which is the harmonic average of TPR and PPV: $F1 = \frac{2 \cdot TPR \cdot PPV}{TPR+PPV}$.

Note that since the true edges of simulated data are known, we could calculate the above measures directly.

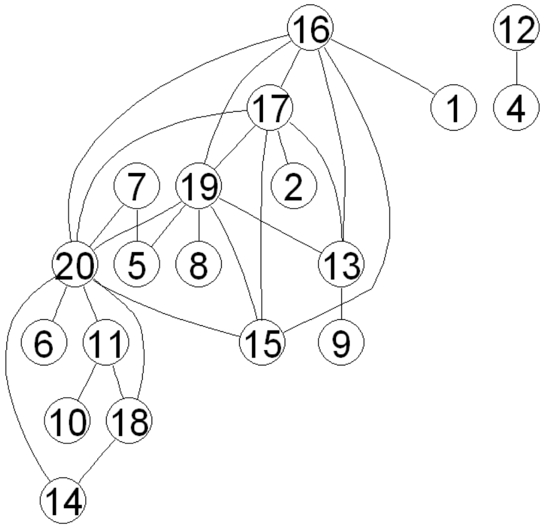

Figure 1: The true network structure. The number of nodes (metabolites) and the network density are 20 and 15 %, respectively.

3. RESULTS

A comparison was performed to deal with the uncertainty of the network complexity using the proposed ensemble averaging approach to compare with the six single network construction methods. To do this, we simulated the true network structure using the R package *igraph* with the number of nodes (metabolites) of 20 and the network density of 15%. The simulated true network structure is depicted in Figure 1. Then using this true network structure, we generated 50 simulated data and then applied to the six single network construction methods as well as the proposed ensemble averaging method. After that, as mentioned before, the performance of each method was evaluated by TPR, PPV, and F1 score.

Table 1 displays the TPRs and PPVs of each method using the 50 simulated data. Lasso performs the best among

Table 1: Ensemble network construction with SCE, PCR, PLSR, Lasso, Adaptive Lasso, and Ridge regression using simulated data. The number of nodes (metabolites), the sample size, and the network density are 20, 50, and 15%, respectively. The mean and standard deviation (SD) of true positive rate (TPR) and positive predictive value (PPV) are calculated after 50 simulations.

Method	TPR		PPV	
	Mean	SD	Mean	SD
SCE	0.4886	0.1144	0.8290	0.1057
PCR	0.4807	0.1261	0.6748	0.1306
PLSR	0.5464	0.1227	0.6892	0.1237
Lasso	0.6871	0.0818	0.5792	0.0921
Alasso	0.5293	0.0897	0.8548	0.0980
Ridge	0.5207	0.1043	0.7530	0.1188
Ensemble	0.6000	0.1258	0.6855	0.1160

Table 2: Ensemble network construction with SCE, PCR, PLSR, Lasso, Adaptive Lasso, and Ridge regression using simulated data. The number of nodes (metabolites), the sample size, and the network density are 20, 50, and 15%, respectively. The mean and standard deviation (SD) of F1 socres are calculated after 50 simulations.

Method	Mean	SD	Method	Mean	SD
SCE	0.6015	0.0904	PCR	0.5443	0.0941
PLSR	0.5936	0.0755	Lasso	0.6219	0.0595
Alasso	0.6472	0.0747	Ridge	0.6049	0.0759
Ensemble	0.6259	0.0910			

the seven methods in terms of TPR, while Adaptive lasso (Alasso) achieves the best performance in terms of PPV. On the other hand, the proposed ensemble method has the second largest TPR but the third smallest PPV. In general, the performance of ensemble method is ranked middle among the seven methods.

The empirical means and standard deviations (SDs) of F1 scores are displayed for each method in Table 2. Interestingly, the ensemble averaging based network construction has F1 score larger than these of other methods except only for Adaptive Lasso, reflecting its ability to account for uncertainty of network complexity.

4. CONCLUSIONS

We propose an ensemble network construction using the six existing methods. As mentioned before, the performance of the network construction algorithm is highly depending on the network complexity. It is almost impossible to recognize the complexity of the network before estimation, however. In addition, it is not always guaranteed that the method should be performed better for other circumstances even though it worked well with a certain circumstance. Therefore, ignoring uncertainty of network complexity can impair the predictive performance and lead to false statements of the associations. Thus, we proposed ensemble model averaging to account for uncertainty of network complexity. As shown in the simulation study that the network complexity is highly related to the construction method, we took care of the network complexity uncertainty by averaging the associations estimated by several construction methods. Our simulation studies show that the ensemble averaging based network construction has F1 score larger than these of other methods except only for Adaptive Lasso, reflecting its ability to account for uncertainty of network complexity.

5. ACKNOWLEDGMENTS

The research reported in this paper was partially supported by the National Science Foundation (NSF) under Award Number DMS-1312603. The Biostatistics Core is supported in part by NIH Cancer Center Support Grant P30 CA022453 to the Karmanos Cancer Institute at Wayne State University. The contents in this paper are solely the responsibility of the authors and do not necessarily represent the official views of NIH and NSF.

6. REFERENCES

[1] A. Arkin, P. Shen, and J. Ross. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277(5330):1275–1279, November 1997.

[2] E. Chan, H. Rowe, B. Hansen, and D. Kliebenstein. The complex genetic architecture of the metabolome. *Plos Genet*, 6(11):e1001198, 2010.

[3] A. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *J Multivariate Anal*, 90(1):196–212, 2004.

[4] A. Greenberg, S. Stockwell, and A. Clark. Evolutionary constraint and adaptation in the metabolic network of drosophila. *Mol Biol Evol*, 25(12):2537–2546, 2008.

[5] T. Hasite, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: The lasso and generalizations*. CRC Press, Boca Raton, FL, 2015.

[6] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science.*, 14:382–401, 1999.

[7] I. Koo, X. Zhang, and S. Kim. Constructing metabolic association networks using high-dimensional mass spectrometry data. *Chemometr Intell Lab Syst.*, 15(138):193–202, 2014.

[8] N. Kramer, J. Schafer, and A. Boulesteix. Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10:384, 2009.

[9] M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–411, 2010.

[10] D. Madigan and A. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1335–1346, 1994.

[11] B. Mevik and R. Wehrens. The pls package: Principal component and partial least squares regression in r. *J Stat Softw*, 18, 2007.

[12] J. Schafer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21:754–764, 2005.

[13] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026, June 2003.

[14] F. Theis, J. Krumsiek, K. Suhre, T. Illig, and J. Adamski. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *Bmc Syst Biol*, 5:21, 2011.

[15] R. Tibshirani. Regression shrinkage and selection via the lasso. *J Royal Satist Soc B*, 58(1):267–288, 1996.

[16] R. Ursem, Y. Tikunov, A. Bovy, R. van Berloo, and F. van Eeuwijk. A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica*, 161(1-2):181–193, June 2008.

[17] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.