Chemometrics and Intelligent Laboratory Systems 138 (2014) 193–202

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Constructing metabolic association networks using high-dimensional mass spectrometry data

Imhoi Koo [a], Xiaoli Wei [a], Xue Shi [a], Zhanxiang Zhou [b], Seongho Kim [c,*], Xiang Zhang [a,*]

[a] Department of Chemistry, Center for Regulatory & Environmental Analytical Metabolomics, University of Louisville, Louisville, KY 40292, USA
[b] Department of Nutrition, University of North Carolina at Greensboro, Greensboro, NC 27412, USA
[c] Biostatistics Core, Karmanos Cancer Institute, Department of Oncology, Wayne State University School of Medicine, Detroit, MI 48201, USA

ARTICLE INFO

Article history:
Received 9 April 2014
Received in revised form 28 June 2014
Accepted 1 July 2014
Available online 9 July 2014

Keywords:
Metabolomics
Gaussian graphical model
Partial correlation
Independent component regression
Principal component regression
Partial least squares regression
Extrinsic similarity

ABSTRACT

The goal of metabolic association networks is to identify topology of a metabolic network for a better understanding of molecular mechanisms. An accurate metabolic association network enables investigation of the functional behavior of metabolites in a cell or tissue. 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. The performance of principal component regression (PCR), independent component regression (ICR), shrinkage covariance estimate (SCE), partial least squares regression (PLSR), and extrinsic similarity (ES) methods in constructing metabolic association networks was compared by estimating partial correlation coefficient matrices when the number of variables is larger than the sample size. To do this, the sample size and the network density (complexity) were considered as variables for network construction. Simulation studies show that PCR and ICR are more stable to the sample size and the network density than SCE and PLSR in terms of F1 scores. These methods were further applied to the analysis of experimental metabolomics data acquired from metabolite extract of mouse liver. For the simulated data, the proposed methods PCR and ICR outperform other methods when the network density is large, while PLSR and SCE perform better when the network density is small. As for the experimental metabolomics data, 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 might be more complex and therefore, PCR and ICR have the advantage over PLSR and SCE in constructing the metabolic association networks.

1. Introduction

Metabolomics is a rapidly emerging field to systemically analyze small-molecule metabolites, which are the end products of cellular processes in a biological organism [1]. 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. In this work, we focus on methods of constructing metabolic networks that represent biochemical transformations among metabolites.

A relatively smaller number of studies have been reported for metabolic network construction. Arkin et al. [2] predicted interactions within reaction networks over time for the glycolytic pathway. Steuer et al. [3] examined the relationship between data generated from networks and biochemical pathways using potato plant metabolism. Ursem et al. [4] 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 [5,6]. GGMs have been employed in metabolomics for several studies. Greenberg et al. [7] used the pseudo-inverse method to estimate the partial correlation (PC) for the study of the influence of enzyme evolution on *Drosophila* metabolic pathway. Chan et al. [8] also constructed the metabolic network to quantify metabolites present in *Arabidopsis thaliana* using the first-order correlation in which the effects of only one variable are removed. Theis et al. [9] used GGMs for reconstructing pathway reactions from human population cohort when the size of samples (experiments) was larger than the number of variables (metabolites). None of these studies, however, used dimension-reduced regression to construct the network.

Reconstructing GGMs using high-dimensional data remains as a difficult task, especially when the number of variables is larger than the

* Corresponding authors.
E-mail addresses: imhoi.koo@louisville.edu (I. Koo), jujuxiao@gmail.com (X. Wei), xueshisx@gmail.com (X. Shi), z_zhou@uncg.edu (Z. Zhou), kimse@karmanos.org (S. Kim), xiang.zhang@louisville.edu (X. Zhang).

http://dx.doi.org/10.1016/j.chemolab.2014.07.002
0169-7439/

sample size. The standard estimation of PCs includes either inversion of sample covariance matrices or estimation of p least squares regression problems, where p is the number of variables. If the number of samples (observations) n is much smaller than p, these approaches are inappropriate. One alternative is to use dimension-reduced regression such as the partial least squares regression (PLSR) [10–12]. Its goal is to discover orthogonal components (score matrix) to maximize the covariance of dependent (response) and independent (predictor) variables.

Independent component and principal component regression analyses (ICR and PCR, respectively) were considered in this study, and their performance for the construction of metabolic network was compared with the performance of PLSR, shrinkage covariance estimator (SCE) [13], and extrinsic similarity (ES) [14,15]. Note that PLSR and SCE were included in this comparison based on the previous studies [10,11]. Although some studies have been performed to compare the performance among different GGM-based methods including PLSR and SCE [10,12], none of these studies included PCR and ICR for network construction. PCR finds a score matrix to maximize variance of independent variables, while ICR finds it to maximize independence. It is known that these two methods will produce the same results if a normal distribution is assumed [16]. The main difference between ICR/PCR and PLSR is that ICR and PCR reduce the dimensions of data without using dependent (response) variables.

Several studies compared the performances between PCR/ICR and PLSR but not in metabolomics. For example, Dupret et al. [17] showed that PLSR performs better than ICR, while Funatsu et al. [18] verified that ICR is superior to PLSR when it was applied to a quantitative structure–property relationship analysis. Also, Wentzull and Montoto [19] reported that no significant difference is shown between PLSR and PCR in terms of prediction errors although Yeniay and Göktas [20] urged that PLSR outperforms PCR. It still remains unclear which of these methods provides the precise output for network construction in metabolomics.

2. Methods

2.1. PC

The PC $\rho_{XY\setminus Z}$ between X and Y given a set of n variables $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ is the correlation between the residuals R_X and R_Y resulting from the linear regression of X and Y on $\mathbf{Z}$, respectively. PC can be interpreted as the association between two random variables after eliminating the effect of a set of random variable. Consider x_i, y_i and $\mathbf{z}_i = (z_1^i, \ldots, z_n^i)$ as samples of a joint probability distribution over X and Y on $\mathbf{Z}$, and assume that the multiple regression problems are

$$x_i = w_0^x + w_1^x z_1^i + \cdots + w_n^x z_n^i, \tag{1}$$

$$y_i = w_0^y + w_1^y z_1^i + \cdots + w_n^y z_n^i, \tag{2}$$

where $i = 1, \cdots, N$. Then the least square solutions $\hat{w}_X$, $\hat{w}_Y$ of the regressions find the vectors to minimize the mean squared error of estimators $\hat{x}_i$ and $\hat{y}_i$ with respect to x and y, respectively. The residuals then are

$$r_{X,i} = x_i - \hat{x}_i, \tag{3}$$

$$r_{Y,i} = y_i - \hat{y}_i, \tag{4}$$

and the sample PC is

$$\hat{\rho}_{XY\setminus Z} = \mathrm{Corr}(R_X, R_Y), \tag{5}$$

where $\mathrm{Corr}(\cdot,\cdot)$ denotes the Pearson's correlation coefficient of two random variables, $R_X = (r_{X,1}, \cdots r_{X,N})$ and $R_Y = (r_{Y,1}, \cdots r_{Y,N})$.

The problem often is that X^TX is singular or ill-posed because the sample size is smaller than the number of variables. An alternative solution of this problem is to use dimension reduction methods for linear regression, which transforms the high-dimensional space into a space spanned by fewer components. Also, those methods can be applied to linear regression and machine learning approaches to increase performance [21,22]. It is desirable that the dimension-reduced data ($\hat{p} \leq n$) can make X^TX well-posed as well as increase the performance.

As mentioned before, we employ five methods to resolve this difficulty in this study, which are shrinkage covariance estimation (SCE) [13], PCR, PLSR [23], ICR [24], and ES [15]. Here SCE is a regularized approach with shrinkage intensity, while ES uses mutual information. PCR, ICR, and PLSR are dimension reduction methods with feature extraction. The differences among PCR, ICR, and PLSR are as follows: PLSR uses both dependent and independent variables to reduce data dimension, while PCR/ICR uses only independent variables, and PCR/PLSR finds orthogonal features based on the normality assumption, while ICR finds independent features based on non-normality. Several studies considered PLSR to see the performance on biological network construction [10,11]. However, there is no study to see the effect of differences among these three approaches on network construction. For this reason, we consider PCR, ICR, and PLSR in this comparison study. Furthermore, we employ SCE as a reference based on the previous comparison study [10], and ES is also included to see the effect of mutual information.

2.2. SCE

Schäfer and Strimmer [13] proposed SCE to estimate the PC when the covariance matrix Σ is singular. Under singularity of covariance matrix, an alternative method is to trade off the unbiased sample covariance Σ and low dimensional shrinkage target matrix T;

$$\hat{\Sigma} = \lambda T + (1-\lambda)\Sigma, \tag{6}$$

where $\lambda \in (0, 1]$ is shrinkage intensity. The optimal value of the tuning parameter λ is analytically determined and estimated from the data. For a more detailed description, refer to Schäfer and Strimmer [13].

2.3. PCR and PLSR

PCR and PLSR [23] circumvent high-dimensional problem by decomposing a data matrix X into orthogonal scores T and loadings P

$$X = TP^{\mathrm{T}} + X_R, \tag{7}$$

and regressing dependent variable Y on the first r important columns $\{t_1, t_2, \ldots, t_r\}$ of the scores T, where X_R is the remains of decomposition. In PCR, the orthogonal scores $T(n \times r)$ and loadings $P(p \times r)$ matrices can be calculated by applying the singular value decomposition (SVD) method to a centered data matrix X as follows:

$$X = UDP^{\mathrm{T}}, \tag{8}$$

where $U(n \times r)$ and $P(p \times r)$ are orthogonal matrices corresponding to r singular values. And the scores matrix T is defined by

$$T = UD. \tag{9}$$

After choosing the optimal or suitable number of components, the first r important components of X are preserved by T. Since the matrix T is orthogonal, $T^{\mathrm{T}}T$ is diagonal and nonsingular matrix. Then the coefficient β_T for a linear regression Y on the score matrix T is estimated by

$$\hat{\beta}_T = \left(T^{\mathrm{T}}T\right)^{-1} T^{\mathrm{T}}Y, \tag{10}$$