Journal of Biomedical Informatics 46 (2013) 894–904

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

A comparative study of covariance selection models for the inference of gene regulatory networks ☆

Patrizia F. Stifanelli [a], Teresa M. Creanza [a], Roberto Anglani [a], Vania C. Liuzzi [a], Sayan Mukherjee [c], Francesco P. Schena [b], Nicola Ancona [a,*]

[a] Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR-ISSIA, Via Amendola 122/D-O, I-70126 Bari, Italy
[b] Dipartimento Emergenza e Trapianti di Organi, DETO, Università di Bari, I-70124 Bari, Italy
[c] Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 21 February 2013
Accepted 8 July 2013
Available online 20 July 2013

Keywords:
Gaussian graphical models
Gene networks
Pathway analysis
Covariance selection

ABSTRACT

Motivation: The inference, or 'reverse-engineering', of gene regulatory networks from expression data and the description of the complex dependency structures among genes are open issues in modern molecular biology.

Results: In this paper we compared three regularized methods of covariance selection for the inference of gene regulatory networks, developed to circumvent the problems raising when the number of observations n is smaller than the number of genes p. The examined approaches provided three alternative estimates of the inverse covariance matrix: (a) the 'PINV' method is based on the Moore–Penrose pseudoinverse, (b) the 'RCM' method performs correlation between regression residuals and (c) 'ℓ_{2C}' method maximizes a properly regularized log-likelihood function. Our extensive simulation studies showed that ℓ_{2C} outperformed the other two methods having the most predictive partial correlation estimates and the highest values of sensitivity to infer conditional dependencies between genes even when a few number of observations was available. The application of this method for inferring gene networks of the isoprenoid biosynthesis pathways in *Arabidopsis thaliana* allowed to enlighten a negative partial correlation coefficient between the two hubs in the two isoprenoid pathways and, more importantly, provided an evidence of cross-talk between genes in the plastidial and the cytosolic pathways. When applied to gene expression data relative to a signature of *HRAS* oncogene in human cell cultures, the method revealed 9 genes (p-value < 0.0005) directly interacting with *HRAS*, sharing the same Ras-responsive binding site for the transcription factor RREB1. This result suggests that the transcriptional activation of these genes is mediated by a common transcription factor downstream of Ras signaling.

Availability: Software implementing the methods in the form of Matlab scripts are available at: http://users.ba.cnr.it/issia/iesina18/CovSelModelsCodes.zip.

1. Introduction

A challenging goal of systems biology is to provide quantitative models for the study of complex interaction patterns among genes and their products that are the result of many biological processes in the cell, such as biochemical interactions and regulatory activities [23]. Among these models, gene regulatory networks (GRNs) are essential representations for the comprehension of the development, functioning and pathology of biological organisms. Indeed, it is widely believed that the GRNs embody the comprehensive information of the mechanisms that govern the expression of the genes in the cell [28]. In particular, the GRNs inferred by genome-wide expression data depend on environmental factors, tissue type, disease-state and experimental conditions. This condition-specificity of GRNs play a major role for the study of biological processes in distinct phenotypical conditions. Indeed, under different conditions, networks exhibit different interaction patterns that can enlighten the understanding of cell development and the identification of key drivers such as disease-related genes or altered biological processes [28,51,31].

One of the simplest and most popular approaches in bioinformatics is to compute the sample Pearson correlation between every pair of genes [7]. The resulting *relevance network* considers two genes 'not-linked' in the case of *marginal* independence. This method, although useful for unveiling co-expression of genes implicated in the same biological process, has important shortcomings for the investigation of GRNs. For assessing co-expression

* Corresponding author.
E-mail address: ancona@ba.issia.cnr.it (N. Ancona).

1532-0464/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.002

between two genes, the Pearson correlation does not take into account the activities of the remaining genes in the cell. Moreover, this method does not distinguish between direct and indirect interactions, and is not able to highlight regulations by a common gene.

These drawbacks may be overcome exploiting *partial correlation*, a more sophisticated statistical model which is able to infer relations of *conditional* dependences among random variables [10,47]. In this framework, Gaussian Graphical Models (GGMs) have been exploited to study and describe dependency structures between random variables [14,26]. In our context, partial correlation assesses association between two genes by removing the effects of a set of controlling genes. Moreover, in a GGM an edge uniquely indicates a direct interaction between a gene *A* and a gene *B*, that can be interpreted biologically as one of the following mechanisms [32]:

- *A* and *B* are regulated by the same transcription factor (TF) which is not included in the network;
- *A* encodes a TF which directly regulates B;
- *A* encondes a TF which directly regulates an intermediate gene *C* which encodes a TF that in turn regulates gene *B*, and *C* is not included in the network;
- *A* encodes a protein which interacts with the TF encoded by an intermediate gene, and modifies its action on the transcription of gene *B*.

In recent years, several reverse-engineering approaches have been proposed for inferring regulatory networks from gene expression data. The nature of the data makes this problem clearly ill-posed. Indeed, the genomic data are typically characterized by a huge number p of genes and by a small number n of samples. The simplest solution proposed to overcome this problem was to reduce the numbers of genes in order to reach the $n > p$ regime [45]. Other solutions have been proposed to circumvent the problem of computing full partial correlation coefficients by using only zero and first order coefficients [48,8,19]. However, these approaches do not take into account all multigene effects on each pair of variables. More sophisticated approaches determine regularized estimates of the covariance matrix and its inverse [50,17,49]. A fundamental assumption usually adopted by these methods in $n < p$ regime is the sparsity of biological networks: only a few edges are supposed to be present in the gene regulatory networks, so that reliable estimates of the graphical model can be inferred also in small sample case [8]. A regularized GGM method based on a Stein-type shrinkage has been applied to genomic data [13] and the network selection has been based on false discovery rate multiple testing. The same procedure to select the network has been adopted, with a Moore–Penrose pseudoinverse method to obtain the precision matrix [39]. Finally, the authors in [34] suggested an attractive and simple approach based on lasso-type regression to select the non-vanishing partial correlations, paving the way to a number of analysis and novel algorithms based on lasso ℓ_1 regularizations [50,17,49,18].

To date, a comparative analysis of these methods is missing. In this work, we focus on recently proposed methods developed in the general framework of regularization and statistical learning theories which provide the state-of-art approaches for the study of ill-posed problems as the ones in which the signal is overwhelmed by the noise and the number of variables is much larger than the number of observations [46]. In particular, we focus on regularized methods for the estimation of the precision matrix in an undirected GGM. We present a comparative study of three methods in terms of AUC (area under the Receiving Operative Characteristic curve), mean square error (MSE), positive predictive values (PPV) and sensitivity (SE). The first method is based on Moore–Penrose pseudoinverse (PINV); the second one provides an estimate of the partial correlation coefficients based on Regularized Least Square regression (RCM); the third method determines an estimate of the precision matrix by maximizing a log-likelihood function properly regularized by an ℓ_2 penalty term (ℓ_{2C}). The conditional dependence between each pair of variables was assessed by using the Efron's bootstrap method [22]. Due to the lack of a perfectly known ground truth related to real biological networks [4], we measured the performance of the three methods by generating simulated data based on golden standard interaction patterns, built according to biological inspired different topologies [18,40]. We found that the ℓ_{2C} method exhibited the most predictive partial correlation estimates. More importantly, this method had the highest values of sensitivity showing its ability to infer true conditional dependencies between genes also when a few number of observations is available.

We assessed the ability of the ℓ_{2C} method to infer GRNs in two real biological contexts: the isoprenoid biosynthesis pathways in *Arabidopsis thaliana* and the *HRAS* oncogenic signature in human cell cultures. In the first case, the method enlightened known relevant pathway properties. In particular, it revealed a negative partial correlation coefficient between the two hubs in the two isoprenoid pathways. This suggests a different response of the pathways to the several tested experimental conditions and, together with the high connectivity of the two hubs, provides an evidence of cross-talk between genes in the plastidial and the cytosolic pathways. In the second case, ℓ_{2C} method highlighted 34 genes directly interacting with *HRAS*. In particular, 9 of these genes (p-value < 0.0005) shared the same Ras responsive transcription factor binding site, suggesting that their transcriptional activation is mediated by a common transcription factor downstream of Ras signaling.

2. Methods

Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be a random vector distributed according a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The interaction structure among these variables can be described by means of a graph $G = (V,E)$, where V is the vertex set and E is the edge set. If vertices of V identify the random variables $X_1, \ldots, X_p$, then the edges of E represent the conditional dependence between the vertices. In other words, the absence of an edge between the ith and jth vertex means a conditional independence between the associated variables X_i and X_j. The structure of a graph is properly described by a $p \times p$ matrix, called adjacency matrix $\boldsymbol{A}$, with elements $a_{ij} = 1$ if the variables X_i and X_j (vertices) are connected by an edge and 0 otherwise.

In this study, we shall consider only undirected Gaussian graphs G with *pairwise Markov property*, such that for all $(i,j) \notin E$ one has

$$X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}} \quad i,j = 1,\ldots,p, \tag{1}$$

i.e. X_i and X_j are conditionally independent being fixed all other variables $X_{V\setminus\{i,j\}}$. Since X follows a $p-$ variate normal distribution, the condition (1) turns out to be $\rho_{ij\cdot V\setminus\{i,j\}} = 0$, where $\rho_{ij\cdot V\setminus\{i,j\}}$ is the partial correlation coefficient between the ith and jth variable, being fixed all other variables. It has been shown [26] that partial correlation matrix elements are related to the *precision matrix* (or inverse covariance matrix) $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, as:

$$\rho_{ij\cdot V\setminus\{i,j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \quad i \neq j, \tag{2}$$

where θ_{ij} are elements of $\boldsymbol{\Theta}$. In general, when the number of observations n is greater than the number of variables p, it is straightforward to evaluate θ_{ij} in Eq. (2) by inverting the sample covariance matrix. Moreover, in this case, a simple parametric test exists for