Methods 129 (2017) 41–49

Contents lists available at ScienceDirect

Methods

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

Node-based learning of differential networks from multi-platform gene expression data

Le Ou-Yang [a], Xiao-Fei Zhang [b,*], Min Wu [c], Xiao-Li Li [c]

[a] College of Information Engineering & Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China
[b] School of Mathematics and Statistics & Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, China
[c] Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way, Singapore

ARTICLE INFO

Article history:
Received 15 January 2017
Received in revised form 11 April 2017
Accepted 18 May 2017
Available online 1 June 2017

Keywords:
Gaussian graphical model
Differential network analysis
Multi-view learning
Group lasso
Gene expression

ABSTRACT

Recovering gene regulatory networks and exploring the network rewiring between two different disease states are important for revealing the mechanisms behind disease progression. The advent of high-throughput experimental techniques has enabled the possibility of inferring gene regulatory networks and differential networks using computational methods. However, most of existing differential network analysis methods are designed for single-platform data analysis and assume that differences between networks are driven by individual edges. Therefore, they cannot take into account the common information shared across different data platforms and may fail in identifying driver genes that lead to the change of network. In this study, we develop a node-based multi-view differential network analysis model to simultaneously estimate multiple gene regulatory networks and their differences from multi-platform gene expression data. Our model can leverage the strength across multiple data platforms to improve the accuracy of network inference and differential network estimation. Simulation studies demonstrate that our model can obtain more accurate estimations of gene regulatory networks and differential networks than other existing state-of-the-art models. We apply our model on TCGA ovarian cancer samples to identify network rewiring associated with drug resistance. We observe from our experiments that the hub nodes of our identified differential networks include known drug resistance-related genes and potential targets that are useful to improve the treatment of drug resistant tumors.

1. Introduction

Biological processes often involve the interactions of genetic components such as mRNAs and proteins. Characterizing the regulatory interactions between genes is critical for elucidating the structural and functional architecture within cells [1–3]. Moreover, there is strong evidence that gene regulatory networks (GRN) undergo changes in response to different conditions such as cancer progression and drug resistance [4–6]. Therefore, inferring gene regulatory networks and exploring how theses networks change across different conditions are fundamental for understanding the biological mechanisms behind disease development [7].

With the accumulation of gene expression data, an increasing number of computational methods have been proposed for gene regulatory network estimation [8,9]. Gaussian graphical models (GGMs), which can identify conditional dependence (or direct dependence) relationships between genes, have been widely used for network inference [10]. Based on the assumption that the observed gene expression data are generated from a multivariate normal distribution, the gene regulatory network can be determined directly from the precision matrix (or inverse covariance matrix) of GGMs [11]. That is, two genes interact with each other if and only if the corresponding entry of the precision matrix is nonzero. Therefore, based on GGMs, the problem of gene regulatory network estimation can be turned into a problem of precision matrix estimation. However, traditional GGMs typically infer one network for a specific condition, and do not consider the network rewiring between different conditions.

In recent years, several differential network analysis methods have been developed for identifying altered dependencies between genes across different conditions [12–14]. Based on GGMs, the difference between two group-specific networks can be identified by calculating the difference between the two corresponding precision matrices [13]. Thus, most existing differential network analysis methods first estimate each group-specific network separately, and then calculate their difference [15]. However, estimating the group-specific networks separately may lose the global dependencies that preserve across all conditions. To exploit the similarity between the true group-specific networks, several methods have

* Corresponding author.
E-mail address: zhangxf@mail.ccnu.edu.cn (X.-F. Zhang).

http://dx.doi.org/10.1016/j.ymeth.2017.05.014
1046-2023/

been proposed to jointly estimate multiple graphical models that share certain characteristics [13,16]. Most of these methods assume that the differences between networks are driven by individual edges. This is unrealistic in many real-world applications since the difference between gene regulatory networks might be driven by certain genes whose patterns of connectivity to other genes are disrupted across conditions. To provide a more intuitive interpretation of the network differences, Mohan et al. introduced a node-based learning approach to jointly estimate multiple GGMs [14].

Rapidly evolving technologies make it possible to collect gene expression data for same patients from different experimental platforms [17]. As gene expression data collected from different platforms (multi-platform gene expression data) describe the expression levels of genes for same patients from different views, they may share some consistent information. Therefore, integrating multi-platform gene expression data may improve the accuracy of gene regulatory network estimation and differential network analysis [18,13]. However, previous differential network analysis methods focus on analyzing the gene expression data collected from a single platform, which could not effectively leverage the common information provided by multi-platform gene expression data.

To address the above problems, we propose a novel node-based multi-view learning algorithm called co-perturbed node joint graphical lasso (CPJGL) model, to simultaneously infer multiple gene regulatory networks corresponding to different patient groups and the differential networks between these patient groups based on gene expression data collected from multiple data platforms (Fig. 1). Our model is an extension of the node-based learning approach proposed by Mohan et al. [14] to the case where gene expression data are characterized in terms of two aspect: patient groups and platform types. Instead of assuming that individual edges are shared or differed across disease states, we assume that the differences between networks are driven by certain perturbed regulatory genes. Based on the row-column overlap norm regularizer [14] and the group lasso penalty [19], our model can exploit the characteristics shared by gene expression data collected from different types of platforms. We propose an alternating direction method of multiplier (ADMM) algorithm to solve the optimization problem. In simulation studies, our proposed CPJGL demonstrated better performance than other competing methods in network inference and differential network analysis. To illustrate the effectiveness of CPJGL on real biological data, we apply CPJGL on TCGA ovarian cancer samples to identify network rewiring associated with platinum resistance. We identify three key regulator genes, namely TSC1, IRS1 and PDPK1, from mTOR signaling pathway and two perturbed genes (MYC and BMP7) from TGF-β signaling pathway. By literature search, we find that these five genes play important roles in drug resistance.

2. Methods

2.1. Gaussian graphical models

Gaussian graphical models can encode the conditional dependencies among a set of p genes, where the expression levels (denoted by a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$) of these p genes are assumed to follow a multivariate Gaussian distribution $N(\mu, \Sigma)$ (here $\mu \in \mathbb{R}^p$ and Σ is a positive definite $p \times p$ matrix). Then two genes are conditionally independent if and only if the corresponding entry of the inverse covariance matrix (precision matrix) $\Theta = \Sigma^{-1}$ is zero [11], i.e., genes i and j are independent of each other given all of the other genes if and only if $\Theta_{ij} = 0$. These conditional dependence relationships can be described by a graph in which nodes denote genes and edges connect conditionally dependent pairs of genes. To estimate the conditional dependencies among p genes, it suffices to estimate the sparsity pattern of the corresponding precision matrix Θ. Suppose that we have n observations that are independently drawn from a multivariate Gaussian distribution $N(\mu, \Sigma)$. When $n > p$, we can estimate the precision matrix $\Theta = \Sigma^{-1}$ by maximum likelihood. However, when $p > n$, this approach fails since the empirical covariance matrix is singular and cannot be inverted to yield an estimate of Σ^{-1}. To deal with this problem, a number of studies [20–22] have instead taken a penalized log-likelihood:

$$\max_{\Theta} \frac{n}{2}(\log\det(\Theta) - \mathrm{tr}(S\Theta)) - \lambda\|\Theta\|_1, \quad (1)$$

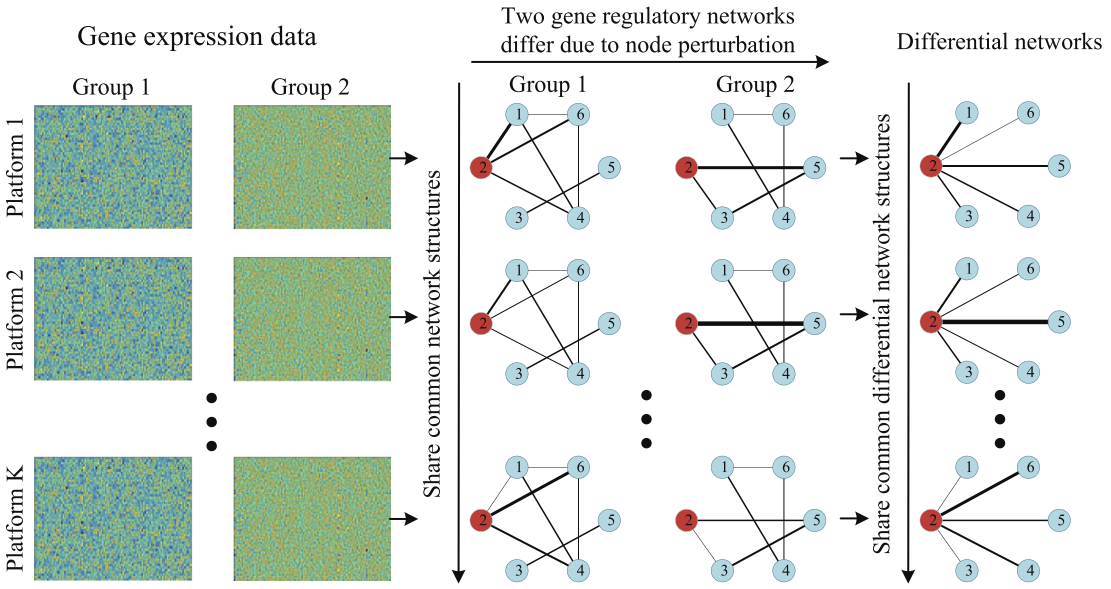

Fig. 1. Motivation and overview of our model. The input data are gene expression data for two different patient groups collected from K data platforms. CPJGL jointly estimates the corresponding $2K$ gene regulatory networks and the K differential networks between these two patient groups by drawing support from the K data platforms. CPJGL encourages the inferred networks and differential networks to share common network structures. It also imposes hub structures on the resulting differential networks. The red node denotes the driver gene that perturbs the network structure.