Computational Biology and Chemistry xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Node-based differential network analysis in genomics

Xiao-Fei Zhang [a,c], Le Ou-Yang [b,*], Hong Yan [c]

[a] School of Mathematics and Statistics & Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, China
[b] College of Information Engineering, Shenzhen University, Shenzhen, 518060, China
[c] Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 24 March 2017
Accepted 27 March 2017
Available online xxx

Keywords:
Differential network analysis
Gaussian graphical model
Gene dependency network
Hub nodes
Graphical lasso

ABSTRACT

Gene dependency networks often undergo changes in response to different conditions. Understanding how these networks change across two conditions is an important task in genomics research. Most previous differential network analysis approaches assume that the difference between two condition-specific networks is driven by individual edges. Thus, they may fail in detecting key players which might represent important genes whose mutations drive the change of network. In this work, we develop a node-based differential network analysis (N-DNA) model to directly estimate the differential network that is driven by certain hub nodes. We model each condition-specific gene network as a precision matrix and the differential network as the difference between two precision matrices. Then we formulate a convex optimization problem to infer the differential network by combing a D-trace loss function and a row-column overlap norm penalty function. Simulation studies demonstrate that N-DNA provides more accurate estimate of the differential network than previous competing approaches. We apply N-DNA to ovarian cancer and breast cancer gene expression data. The model rediscovers known cancer-related genes and contains interesting predictions.

1. Introduction

A key challenge in genomic research is to characterize complex interactions of molecular entities such as genes and their products (Barabási et al., 2011; Zhang et al., 2015a). Genes within cell signaling pathways interact with each other to form networks that regulate various cellular functions. It is well-established that a gene dependency network can undergo change in response to different conditions such as DNA damage or environmental stress (de la Fuente, 2010; Ideker and Krogan, 2012; Ou-Yang et al., 2014; Grechkin et al., 2016; Zhang et al., 2016a). Thus, it is of great interest to explore how the gene network changes between two conditions. Indeed, differential network analysis has become an important tool in bioinformatics which is complement to differential expression analysis (de la Fuente, 2010; Gill et al., 2010; Ha et al., 2015).

Gene networks are often modeled as Gaussian graphical models (Friedman, 2004; Markowetz and Spang, 2007). These models assume that the gene expression levels are generated from a multivariate Gaussian distribution (Cox and Wermuth, 1996). As a consequence, the conditional dependencies between genes can be determined directly from nonzero elements of the precision matrix (or inverse covariance matrix) (Lauritzen, 1996), where two genes are conditionally independent given the other genes if and only if the corresponding element of the precision matrix is zero. Therefore, the condition-specific gene networks can be modeled as the corresponding precision matrices (Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008). Then, the different network between two conditions can be modeled as the difference between the two condition-specific precision matrices (Danaher et al., 2014; Mohan et al., 2014; Zhao et al., 2014; Huang and Chen, 2015; Yuan et al., 2015a; Zhang et al., 2016b).

There are two main types of approaches to estimate the differential network based on Gaussian graphical models. The most straightforward one is to estimate the condition-specific precision matrices first and then subtract the estimates (Ha et al., 2015; Danaher et al., 2014; Mohan et al., 2014; Zhang et al., 2016b). The inverse of the sample covariance matrix can be a naive estimate of a single precision matrix. However, when the number of genes exceeds the number of subjects, the sample covariance matrix is not invertible. Based on the prior knowledge that many pairs of genes are conditionally independent, graphical lasso models (Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008) have been proposed to obtain sparse estimate of the precision matrix. The

* Corresponding author.
E-mail address: szuouyl@gmail.com (L. Ou-Yang).

http://dx.doi.org/10.1016/j.compbiolchem.2017.03.010
1476-9271/

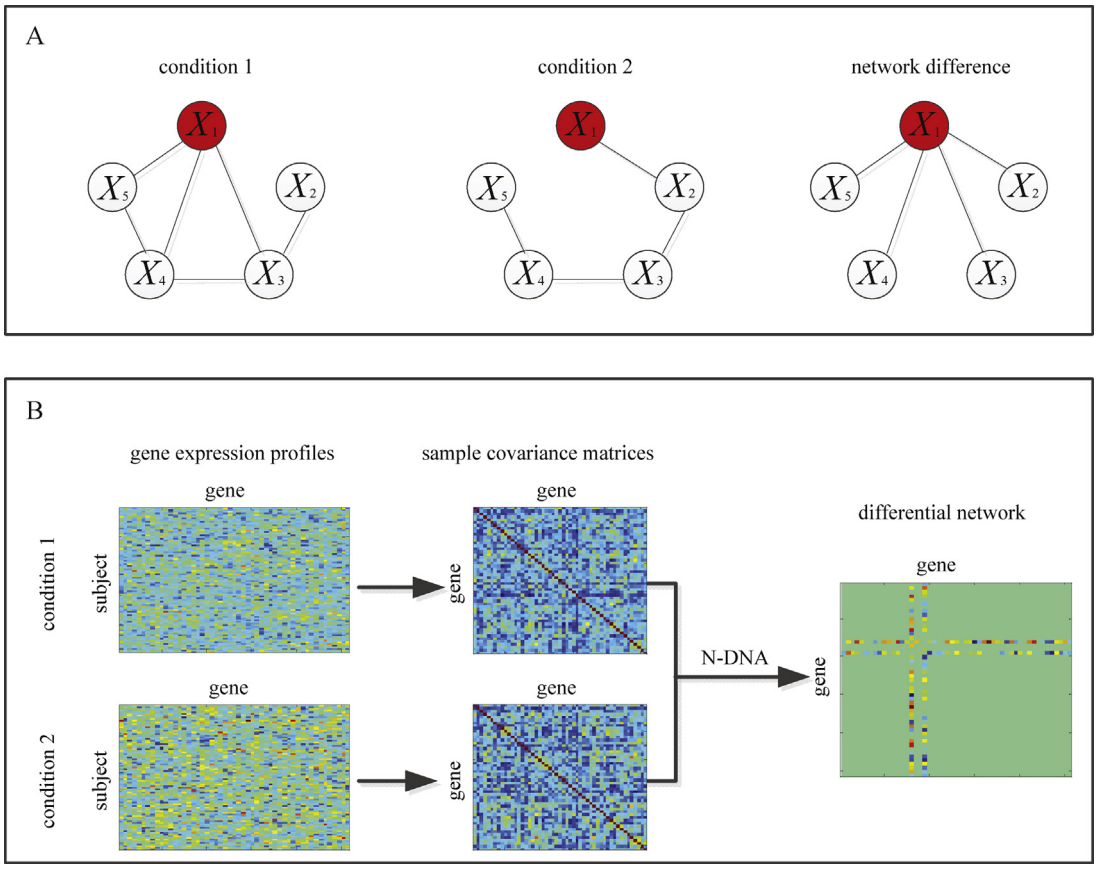

Fig. 1. Motivation and overview of our work. (A) A toy example that illustrates the change of a network of 5 genes between two different conditions. The perturbation of network might be due to a driver mutation on gene X_1 that alters its regulatory interactions with the other four genes. Gene X_1 might represent a key regulator (e.g., transcription factor) that is mutated or abnormally expressed in one condition, in turn changing the underlying gene network. Therefore, the differential network might be driven by certain hub genes. (B) An overview of our method. The input data are two condition-specific gene expression data. Then the condition-specific sample covariance matrices are computed. Based on the estimated sample covariance matrices, we directly estimate the differential network using the proposed N-DNA model. Our N-NDA does not require to estimate the individual gene networks and can impose hub structure on the resulting differential network.

standard graphical lasso models separately estimate the condition-specific precision matrices using subjects within each condition. They might be suboptimal when there exist common structures shared by different conditions (Danaher et al., 2014; Ha et al., 2015; Mohan et al., 2014). To deal with this problem, several joint estimation approaches have been proposed by employing various group penalties (Danaher et al., 2014; Ha et al., 2015; Mohan et al., 2014; Huang and Chen, 2015; Guo et al., 2011; Lee and Liu, 2015; Zhang et al., 2016b). Most of these approaches assume that the individual precision matrices are sparse. However, the sparsity assumption can be violated if the gene network contains hub nodes which connect with many other nodes (Zhao et al., 2014). In addition, their power in detecting weak interactions is limited since the corresponding elements in precision matrices might be shrunk toward zero. There are many interactions whose strengths are weak in single condition but their changes between two conditions are large. Simply subtracting two estimated precision matrices would overlook these changes (Yuan et al., 2015a).

The second type of approach is to directly estimate the difference between the condition-specific precision matrices. Zhao et al. (2014) proposed a differential network analysis approach which does not require the individual precision matrices to be sparse. Their approach outperforms fused graphical lasso (Danaher et al., 2014) when the individual networks include hubs. However, their approach is computationally prohibitive when the number of genes is large. Based on the D-trace loss function which is developed to estimate precision matrix (Zhang and Zou, 2014), Yuan et al. (2015a) developed a new loss function to directly estimate the differential network, which can be solved more efficiently than the loss function used by Zhao et al. (2014). Tian et al. (2016) also proposed a direct estimation method that is is similar to (Yuan et al., 2015a) but use a different algorithm to solve the model. All the three approaches apply a lasso penalty to each edge, which means each edge is treated equally and independent of the others (Mohan et al., 2014; Tan et al., 2014). This is unrealistic in many real-world applications, where the network difference might be due to certain nodes (e.g., regulator genes) that are perturbed across conditions, completely disrupting its dependence relationships with other genes (Mohan et al., 2014; Zhang et al., 2015b; Grechkin et al., 2016). Fig. 1A presents a toy example where the change of network is driven by a particular node, X_1. Here X_1 might represent a transcription factor that is mutated or abnormally expressed in a particular condition, in turn changing its regulatory interactions with other genes. Therefore, there might exist certain hub nodes in the differential network that represent the perturbed regulator genes. Even though the three direct estimation approaches allow the presence of hub nodes in each condition-specific network, they are not designed for encouraging the appearance of hub nodes in the differential network.

In this paper, we estimate the differential network between two conditions, which is driven by certain perturbed regulatory genes. By combining the D-trace loss function (Yuan et al., 2015a) and the row-column overlap norm regularizer (Mohan et al., 2014), we develop a node-based differential network analysis (N-DNA)

Please cite this article in press as: Zhang, X.-F., et al., Node-based differential network analysis in genomics. Comput. Biol. Chem. (2017), http://dx.doi.org/10.1016/j.compbiolchem.2017.03.010