Computer Modeling in Engineering & Sciences
DOI:10.32604/cmes.2020.09353
Article

Tech Science Press

A Re-Parametrization-Based Bayesian Differential Analysis Algorithm for Gene Regulatory Networks Modeled with Structural Equation Models

Yan Li[1,2], Dayou Liu[1,2], Yungang Zhu[1,2] and Jie Liu[1,2,*]

[1]College of Computer Science and Technology, Jilin University, Changchun, 130012, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China
*Corresponding Author: Jie Liu. Email: liu_jie@jlu.edu.cn
Received: 05 December 2019; Accepted: 20 March 2020

Abstract: Under different conditions, gene regulatory networks (GRNs) of the same gene set could be similar but different. The differential analysis of GRNs under different conditions is important for understanding condition-specific gene regulatory relationships. In a naive approach, existing GRN inference algorithms can be used to separately estimate two GRNs under different conditions and identify the differences between them. However, in this way, the similarities between the pairwise GRNs are not taken into account. Several joint differential analysis algorithms have been proposed recently, which were proved to outperform the naive approach apparently. In this paper, we model the GRNs under different conditions with structural equation models (SEMs) to integrate gene expression data and genetic perturbations, and re-parameterize the pairwise SEMs to form an integrated model that incorporates the differential structure. Then, a Bayesian inference method is used to make joint differential analysis by solving the integrated model. We evaluated the performance of the proposed re-parametrization-based Bayesian differential analysis (ReBDA) algorithm by running simulations on synthetic data with different settings. The performance of the ReBDA algorithm was demonstrated better than another state-of-the-art joint differential analysis algorithm for SEMs ReDNet obviously. In the end, the ReBDA algorithm was applied to make differential analysis on a real human lung gene data set to illustrate its applicability and practicability.

Keywords: Gene regulatory networks; structural equation models; joint differential analysis; Bayesian analysis

1 Introduction

A GRN is usually a directed network that depicts a set of genes and the regulatory interactions between them. Under different conditions, for example, in different tissues or in diseased and healthy individuals, the GRNs of the same gene set may differ slightly from each other. Identifying inconspicuous changes between condition-specific GRNs is of great significance for discovering molecular mechanisms and biological processes of genes, which helps to understand gene functions and find pathogenic genes [1,2].

A number of models and corresponding computational methods have been developed to infer GRNs from gene expression data and other related data sources under a single condition, such as Boolean networks [3], information theory based networks [4,5], differential equation models [6], Bayesian networks [7,8] and Gaussian graphical models [9]. Our main concern is on GRNs modeled with SEMs, which are inferred from gene expression data and genetic perturbations (e.g., eQTL data). A series of algorithms have been developed to infer GRNs modeled with SEMs successively [10–13].

While it is possible to adopt these existing algorithms to infer condition-specific GRNs separately and then identify their differences from the estimated GRNs, such an approach is not optimal because it doesn't exploit the similarities between the pairwise GRNs [14]. Danaher et al. [15] and Mohan et al. [16] proposed joint inference methods for multiple GRNs modeled with Gaussian graphical models. Danaher et al. [15] proposed the joint graphical lasso to estimate multiple graphical models that share certain characteristics. They formulated the joint graphical lasso problem by introducing penalized log likelihood with a fused lasso penalty or a group lasso penalty and used an ADMM algorithm to maximize the penalized log likelihood. The fused lasso penalty encouraged similar GRN structure and edge values, whereas the graph lasso penalty encouraged a weaker form of similarity. Mohan et al. [16] considered similarities shared across GRNs due to the presence of common hub nodes and differences between GRNs driven by individual nodes that are perturbed across conditions, and formulated two convex optimization problems corresponding to the two problems respectively by using a row-column overlap norm penalty function. Gaussian graphical models can only identify undirected networks, and only exploit gene expression data. Wang et al. [17] developed an efficient proximal gradient algorithm for differential GRN inference based on linear regression models. However, although a linear model can support directed graph, it still models and infers GRNs only from gene expression data.

SEMs provide a systematic framework for GRN inference integrating gene expression data and genetic perturbations conveniently, and can yield more accurate predictive network structure. Motivated by this, we mainly study joint differential analysis method of GRNs modeled with SEMs. Ren et al. [18] proposed a re-parametrization-based joint differential analysis algorithm for SEMs named ReDNet. In the ReDNet algorithm, they re-parameterized two pairwise structural equations (corresponding to two GRNs under different conditions) as an integrated SEM, the commonality and difference are both incorporated into the integrated model. The simulation studies in [18] demonstrated that ReDNet outperforms the naive approach that independently constructs the pairwise GRNs.

In this paper, we introduce a novel joint differential analysis algorithm named ReBDA. In the first stage, we incorporate the two pairwise SEMs into an integrated SEM to consider not only the sparsity of the individual GRN but also the difference between GRNs under different conditions; And then in the second stage, following the Bayesian inference method for sparse SEMs developed by Dong et al. [19], the individual GRN and differential GRN can be directly inferred from the re-parameterized integrated model. The overview of the proposed ReBDA algorithm can be found in Fig. 1.

Tab. 1 lists the detailed comparison between our proposed ReBDA algorithm and other previous proposed related methods around three properties: if the similarities between GRNs under different conditions are considered; if directed networks could be supported; if the genetic perturbations could be incorporated into models. We see that only the ReDNet algorithm supports the joint differential analysis of directed GRNs incorporating genetic perturbations. Therefore, computer simulations are conducted to compare the performance of our proposed ReBDA algorithm and another state-of-the-art joint differential analysis algorithm for SEMs ReDNet. The results demonstrates that ReBDA has apparently better performance for both directed acyclic graphs (DAGs) and directed cyclic graphs (DCGs) in various settings.

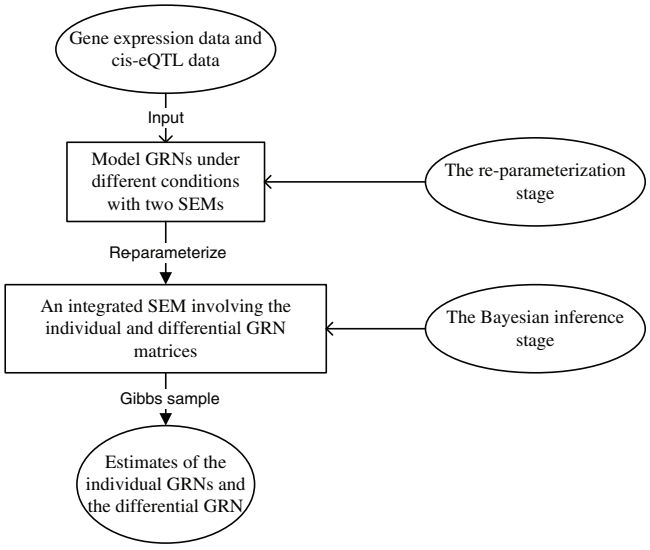

Figure 1: Overview of the ReBDA algorithm

Table 1: Comparison between the proposed method and previous related methods

	Similarities	Directed networks	Genetic perturbations
Naive approach	×	√	√
Danaher et al. [15]	√	×	×
Mohan et al. [16]	√	×	×
Wang et al. [17]	√	√	×
Ren et al. [18] (ReDNet)	√	√	√
Our proposed algorithm (ReBDA)	√	√	√

2 Models and Methods

2.1 GRNs Modeled with SEMs

Consider expression levels of p genes and genotypes of q cis-eQTLs under two different conditions ($k = 1, 2$). Let $\mathbf{Y}^{(k)} = [\mathbf{y}_1^{(k)}, \mathbf{y}_2^{(k)}, \ldots, \mathbf{y}_p^{(k)}]$ be an $n \times p$ gene expression matrix denoting gene expression levels of p genes measured from n individuals under condition k, and let $\mathbf{X}^{(k)} = [\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \ldots, \mathbf{x}_q^{(k)}]$ be an $n \times q$ cis-eQTL matrix denoting genotypes of q cis-eQTLs measured from n individuals under condition k. As in [5,24], we assume each gene has at least one unique cis-eQTLs to ensure the unique identifiable of GRNs, which means $q \geq p$. Then the two GRNs can be modeled with the following SEMs,

$$\mathbf{Y}^{(k)} = \mathbf{Y}^{(k)}\mathbf{B}^{(k)} + \mathbf{X}^{(k)}\mathbf{F}^{(k)} + \mathbf{E}^{(k)}, \ k = 1, 2, \tag{1}$$

where $p \times p$ matrix $\mathbf{B}^{(k)} = [\mathbf{b}_1^{(k)}, \mathbf{b}_2^{(k)}, \ldots, \mathbf{b}_p^{(k)}]$ defines the structure of GRN under condition k, $b_{ij}^{(k)}$ represents the regulatory effect of the ith gene on the jth gene; $q \times p$ matrix $\mathbf{F}^{(k)} = [\mathbf{f}_1^{(k)}, \mathbf{f}_2^{(k)}, \ldots, \mathbf{f}_p^{(k)}]$ is composed of the regulatory effects of the q cis-eQTLs, $f_{ij}^{(k)}$ is the regulatory effect of the ith cis-eQTL on the jth gene; $n \times p$ matrix $\mathbf{E}^{(k)}$ is the error matrix, the entry $e_{ij}^{(k)}$ is often assumed as the ith error term of the jth gene independently normally distributed with mean zero and variance σ^2.

As mentioned in [12,13], it is assumed that there is no self-loop in the GRN, which implies $b_{ii}^{(k)} = 0$ for $i = 1,\ldots, p$. We further assume that the q cis-eQTLs have been identified by an existing eQTL mapping method, but the regulatory effects are still unknown, this is to say, there are q unknown nonzero entries with known locations in $\mathbf{F}^{(k)}$. The known row index set of nonzero entries in $\mathbf{f}_i^{(k)}$, $i = 1,\ldots, p$ is represented as $\mathbf{s}_i$. Therefore, the main task of this paper is to estimate the differential GRN matrix $\Delta\mathbf{B} = \mathbf{B}^{(1)} - \mathbf{B}^{(2)}$, and passingly, the individual GRN matrices $\mathbf{B}^{(k)}$ and the unknown nonzero entries in $\mathbf{F}^{(k)}$ could also be estimated.

With the above definitions, $\mathbf{B}^{(k)}$ and $\mathbf{F}^{(k)}$ can be estimated column by column by decomposing the model in Eq. (1) into

$$\mathbf{y}_i^{(k)} = \mathbf{Y}_{-i}^{(k)}\mathbf{b}_{i,-i}^{(k)} + \mathbf{X}_{\mathbf{s}_i}^{(k)}\mathbf{f}_{i,\mathbf{s}_i}^{(k)} + \mathbf{e}_i^{(k)},\ i = 1,\ 2,\ \ldots,\ p \tag{2}$$

where $(p-1) \times 1$ vector $\mathbf{b}_{i,-i}^{(k)}$ is obtained by excluding the ith entry of $\mathbf{b}_i^{(k)}$; the $n \times (p-1)$ matrix $\mathbf{Y}_{-i}^{(k)}$ refers to the submatrix of $\mathbf{Y}^{(k)}$ excluding the ith column $\mathbf{y}_i^{(k)}$; $\mathbf{f}_{i,\mathbf{s}_i}^{(k)}$ is a reduced form of $\mathbf{f}_i^{(k)}$ excluding the rows whose indices are not in $\mathbf{s}_i$; $\mathbf{X}_{\mathbf{s}_i}^{(k)}$ is a submatrix of $\mathbf{X}^{(k)}$ obtained by only extracting the columns whose indices are in $\mathbf{s}_i$; $\mathbf{e}_i^{(k)}$ is the ith column of $\mathbf{E}^{(k)}$.

2.2 The Re-Parametrization Stage

Since our main concern is the differential structure of two GRNs under different conditions, that is $\Delta\mathbf{B} = \mathbf{B}^{(1)} - \mathbf{B}^{(2)}$. The original model as in Eq. (2) can be re-parameterized to incorporate $\Delta\mathbf{B}$ into the model. There are several different kinds of re-parametrization methods to construct such model, here we propose a novel one, for all $i = 1, 2,\ldots, p$, we define

$$\mathbf{y}_i = \mathbf{y}_i^{(1)} + \mathbf{y}_i^{(2)},\ \mathbf{e}_i = \mathbf{e}_i^{(1)} + \mathbf{e}_i^{(2)},$$

$$\mathbf{Y}_{-i} = \left[\mathbf{Y}_{-i}^{(1)} + \mathbf{Y}_{-i}^{(2)},\ \mathbf{Y}_{-i}^{(1)}\right],\ \mathbf{b}_i = \left[\mathbf{b}_{i,-i}^{(2)},\ \mathbf{b}_{i,-i}^{(1)} - \mathbf{b}_{i,-i}^{(2)}\right], \tag{3}$$

$$\mathbf{X}_i = \left[\mathbf{X}_{\mathbf{s}_i}^{(1)} + \mathbf{X}_{\mathbf{s}_i}^{(2)},\ \mathbf{X}_{\mathbf{s}_i}^{(1)}\right],\ \mathbf{f}_i = \left[\mathbf{f}_{i,\mathbf{s}_i}^{(2)},\ \mathbf{f}_{i,\mathbf{s}_i}^{(1)} - \mathbf{f}_{i,\mathbf{s}_i}^{(2)}\right],$$

Then model (2) can be rewritten as an integrated model as follows for differential analysis of GRNs,

$$\mathbf{y}_i = \mathbf{Y}_{-i}\mathbf{b}_i + \mathbf{X}_i\mathbf{f}_i + \mathbf{e}_i,\ i = 1,\ldots,p. \tag{4}$$

By estimating the $2(p-1) \times 1$ parameter vector $\mathbf{b}_i$ from the above re-parameterized integrated model for all $i = 1,\ldots, p$, the GRN matrix $\mathbf{B}^{(2)}$ and the differential GRN matrix $\Delta\mathbf{B}$ can be easily obtained, and then the GRN matrix $\mathbf{B}^{(1)}$can be directly computed with $\mathbf{B}^{(1)} = \Delta\mathbf{B} + \mathbf{B}^{(2)}$.

Note that the unknown parameters to be estimated are all contained in vector $[\mathbf{b}_i,\ \mathbf{f}_i]^T$, we further rewrite the integrated model in Eq. (4) as a linear-type model,

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \mathbf{e}_i,\ i = 1,\ldots,p. \tag{5}$$

where $\mathbf{Z}_i = [\mathbf{Y}_{-i}, \mathbf{X}_i]$, $\boldsymbol{\beta}_i = [\mathbf{b}_i,\ \mathbf{f}_i]^T$. Assume the dimension of $\mathbf{f}_i$ is q_i ($q_i \geq 1$), meaning that the expression of gene i is affected by q_i cis-eQTLs, so the dimension of $\boldsymbol{\beta}_i$ can be denoted by $p_i = 2(p-1) + q_i$. With the estimate of $\boldsymbol{\beta}_i$ inferred from Eq. (5), $\mathbf{b}_i$ can be easily recovered according to the re-parametric process.

2.3 The Bayesian Inference Stage

Based on biological characteristics, GRNs or more general biochemical networks are considered sparse [20–22], and the structures of GRNs under different conditions generally differ slightly from each other

[14,18,23,24], that is to say, $\boldsymbol{\beta}_i = [\mathbf{b}_i,\ \mathbf{f}_i]^T$ is sparse. So a sparse inference method for linear regression models can be adopted to infer the individual GRNs and differential GRN. A series of sparse inference algorithms for linear regression models have been developed, such as the lasso [25], the fused lasso [26], the elastic net [27], the SCAD [28], the adaptive lasso [29] and the Bayesian lassos [30,31].

Dong et al. [19] proposed an iterative scheme named LRBI using Bayesian inference to estimate parameters of linear regression models based on SEMs. Proved by simulation studies in [19], the Bayesian inference method in the LRBI algorithm was effective and efficient for inference of GRNs modeled with SEMs. Motivated by this, in what follows, we apply the Bayesian inference method in LRBI on model (5) to deduce our hierarchical conditional posterior distribution to estimate $\boldsymbol{\beta}_i$ consistently and rapidly.

We assume the entries in $\mathbf{e}_i$ are independently identically normally distributed with mean zero and variance σ_i^2, then the likelihood can be expressed as $\mathbf{y}_i|\mathbf{Z}_i,\ \boldsymbol{\beta}_i,\ \sigma_i^2 \sim \text{Normal}\left(\mathbf{Z}_i\boldsymbol{\beta}_i,\ \sigma_i^2\,\mathbf{I}_n\right)$. As in LRBI, an efficient Normal-Gamma prior for $\boldsymbol{\beta}_i$ as follows can be assumed,

$$\begin{aligned}
&\boldsymbol{\beta}_i|\sigma_i^2 \sim \text{Normal}_{p_i}\left(\mathbf{0},\ \sigma_i^2\mathbf{I}_{p_i}\right),\\
&\sigma_i^2 \sim \text{Inverse}-\text{Gamma}(a_{i0},\ b_{i0}),
\end{aligned} \tag{6}$$

where a_{i0}, b_{i0} are hyper parameters that should be preset to fixed values.

Then the conditional posterior distribution can be deduced,

$$\begin{aligned}
&\boldsymbol{\beta}_i|\mathbf{y}_i,\ \mathbf{Z}_i \sim \text{Normal}_{p_i}\left(\boldsymbol{\alpha}_i,\ \sigma_i^2\,\mathbf{A}_i\right),\\
&\sigma_i^2\,|\mathbf{y}_i,\ \mathbf{Z}_i \sim \text{Inverse}-\text{Gamma}\left(a_{i0}+\frac{n}{2},\ b_i\right),
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
&\mathbf{A}_i = \left(\mathbf{Z}_i^T\mathbf{Z}_i+\mathbf{I}\right)^{-1},\\
&\boldsymbol{\alpha}_i = \mathbf{A}_i\left(\mathbf{Z}_i^T\mathbf{y}_i+\boldsymbol{\beta}_i\right),\\
&b_i = b_{i0}+\frac{\mathbf{y}_i^T\mathbf{y}_i+\boldsymbol{\beta}_i^T\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i^T\mathbf{A}_i^{-1}\boldsymbol{\alpha}_i}{2}.
\end{aligned} \tag{8}$$

According to the above conditional posterior, the posterior mean estimate of $\boldsymbol{\beta}_i$ can be obtained via Gibbs sampling with the preset hyper parameters (like $a_{i0} = \frac{n}{5},\ b_{i0} = 1$).

3 Results

3.1 Computer Simulations

In this section, simulations are run on synthetic data with different settings to compare the performance of the proposed ReBDA algorithm and another joint differential analysis algorithm based on re-parametrization: ReDNet [18]. The performance of GRN inference is usually evaluated via power of detection (PD) and false discovery rate (FDR). PD measures the percentage of correctly identified edges in all true edges, and FDR measures the percentage of false identified edges in all detected edges. Let N_t be the number of edges in the reference network, N_d be the number of edges in the estimated network, N_{tp} be the number of correctly identified edges, N_{fp} be the number of false identified edges. Then PD can be calculated by N_{tp}/N_t, and FDR is obtained by N_{fp}/N_d.

We generate DAGs and DCGs with 30 genes under two different conditions referring to the settings in [13]. The sample size n varies from 80 to 600. The average number of edges of each node n_e determines the degree of sparseness, we set it as 1 or 3. Each gene is assumed to have 2 effective cis-eQTLs, that is to say, $q = 2p$. The regulatory effects of all cis-eQTLs are set to 1. The adjacency matrix $\mathbf{A}^{(1)}$ of a DAG or DCG with specified setting is first generated for the GRN under condition 1, the GRN matrix $\mathbf{B}^{(1)}$ is generated

by changing the nonzero entries in $\mathbf{A}^{(1)}$ to a random value sampled from a uniform distribution over $(-1, -0.5)\cup(0.5, 1)$. Then the GRN matrix $\mathbf{B}^{(2)}$ is generated based on $\mathbf{B}^{(1)}$ by randomly modifying a small number of entries in it. In this part, the total number of entries to be modified is set as 30% of p, among which the following three modification patterns share the same proportion: (1) $b_{ij}^{(1)} = 0$ but $b_{ij}^{(2)} \neq 0$; (2) $b_{ij}^{(1)} \neq 0$ but $b_{ij}^{(2)} = 0$; (3) $b_{ij}^{(1)} \neq 0$ and $b_{ij}^{(2)} \neq 0$, but $b_{ij}^{(1)} \neq b_{ij}^{(2)}$. The genotypes of the cis-eQTLs are simulated from an F2 cross. Values 1 and 3 are assigned to two homozygous genotypes, respectively, and value 2 is assigned to heterozygous genotypes. Then ternary random variables $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are generated by sampling from $\{1, 2, 3\}$ with probabilities $\{0.25, 0.5, 0.25\}$, respectively. $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ are simulated by randomly permuting the rows of matrix $[\mathbf{I}_p, \mathbf{I}_p]^T$, where $\mathbf{I}_p$ represents p-dimensional identity matrix. Next, the error terms $\mathbf{E}^{(1)}$ and $\mathbf{E}^{(2)}$ are independently sampled from random variables normally distributed with mean zero and variance $\sigma^2 = 0.01$ or 0.1. Finally, $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are directly calculated via $\mathbf{Y}^{(k)} = \mathbf{Y}^{(k)}\mathbf{B}^{(k)} + \mathbf{X}^{(k)}\mathbf{F}^{(k)} + \mathbf{E}^{(k)},\ k = 1,\ 2$.

For each setting, 20 replicates are generated, ReBDA and ReDNet are applied on each replicate, then the averaged PD and FDR of $[\mathbf{B}^{(1)},\ \mathbf{B}^{(2)}]$ and $\Delta\mathbf{B}$ can be calculated to compare and evaluate the performance. What's more, because the Bayesian penalized regression does not exactly produce zero estimates, a decision threshold t is preset to go from a posterior distribution to a sparse point estimate. We set $t = 0.2$ in the following simulations, that is, all entries in the estimated $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ whose absolute value are smaller than $t = 0.2$ are set to 0.

The results of ReBDA and ReDNet for DAGs with 30 genes, $n_e = 1$ or 3 and $\sigma^2 = 0.01$ or 0.1 are depicted in Fig. 2. Let's first see the performance of $[\mathbf{B}^{(1)},\ \mathbf{B}^{(2)}]$, which are shown in the upper panel of Figs. 2(a)–2(d). The PD of ReBDA are slightly lower than that of ReDNet (nearly reach 1 for all settings), and the FDR of ReBDA are obviously lower than that of ReDNet. The performance of $\Delta\mathbf{B}$ can be found in the lower panel of Figs. 2(a)–2(d). The PD of ReBDA are a little better than that of ReDNet for sparse networks and obviously better than that of ReDNet for dense networks. As for FDR, ReBDA significantly outperform ReDNet expect for the networks with $\sigma^2 = 0.1$ at sample size 80 and 100.

The results of ReBDA and ReDNet for DCGs with 30 genes, $n_e = 1$ or 3 and $\sigma^2 = 0.01$ or 0.1 are depicted in Fig. 3. Similarly, the performance of $[\mathbf{B}^{(1)},\ \mathbf{B}^{(2)}]$ are shown in the upper panel of Figs. 3(a)–3(d). ReBDA offers similar PD with ReDNet for sparse networks, and visible better PD than ReDNet for dense networks. The FDR of ReBDA are near or equal to zero for networks at most settings, which are lower than that of ReDNet obviously. The performance of $\Delta\mathbf{B}$ are depicted in the lower panel of Figs. 3(a)–3(d). ReBDA still exhibits better PD than ReDNet, and the difference between them are more significant in dense networks. When it comes to FDR, ReBDA performs much better than ReDNet expect when $\sigma^2 = 0.1$ and $n = 80$.

3.2 Real Data Analysis

Lu et al. [32] measured gene expression levels and genotypes of SNPs in 42 lung tumor tissues and 42 adjacent normal tissues of non-smoking female patients with lung adenocarcinomas with 54,675 probe sets from Affymetrix Human Genome U133 Plus 2.0 arrays and 906,551 SNP probes from Affymetrix GenomeWide Human SNP 6.0 arrays. We preprocessed this data set following the way in [14] with R package affy [33] and MatrixEQTL [34], 1455 genes were found to have at least one cis-eQTLs at FDR < 0.1. Because the number of genes is too much larger than the number of available samples, which may result in less reliable estimates, we further filtered the data set with the GIANT database [35], 15 genes were identified to interact with at least one other genes with high confidence (>0.8), namely: PPP4R2, DBI, DKC1, HSF1, PSMA2, RPS6, USP10, PPP4R3A, ATP5G3, CDC123, MAPKAPK2, PSMD6, RPS16, BTF3, G3BP2.

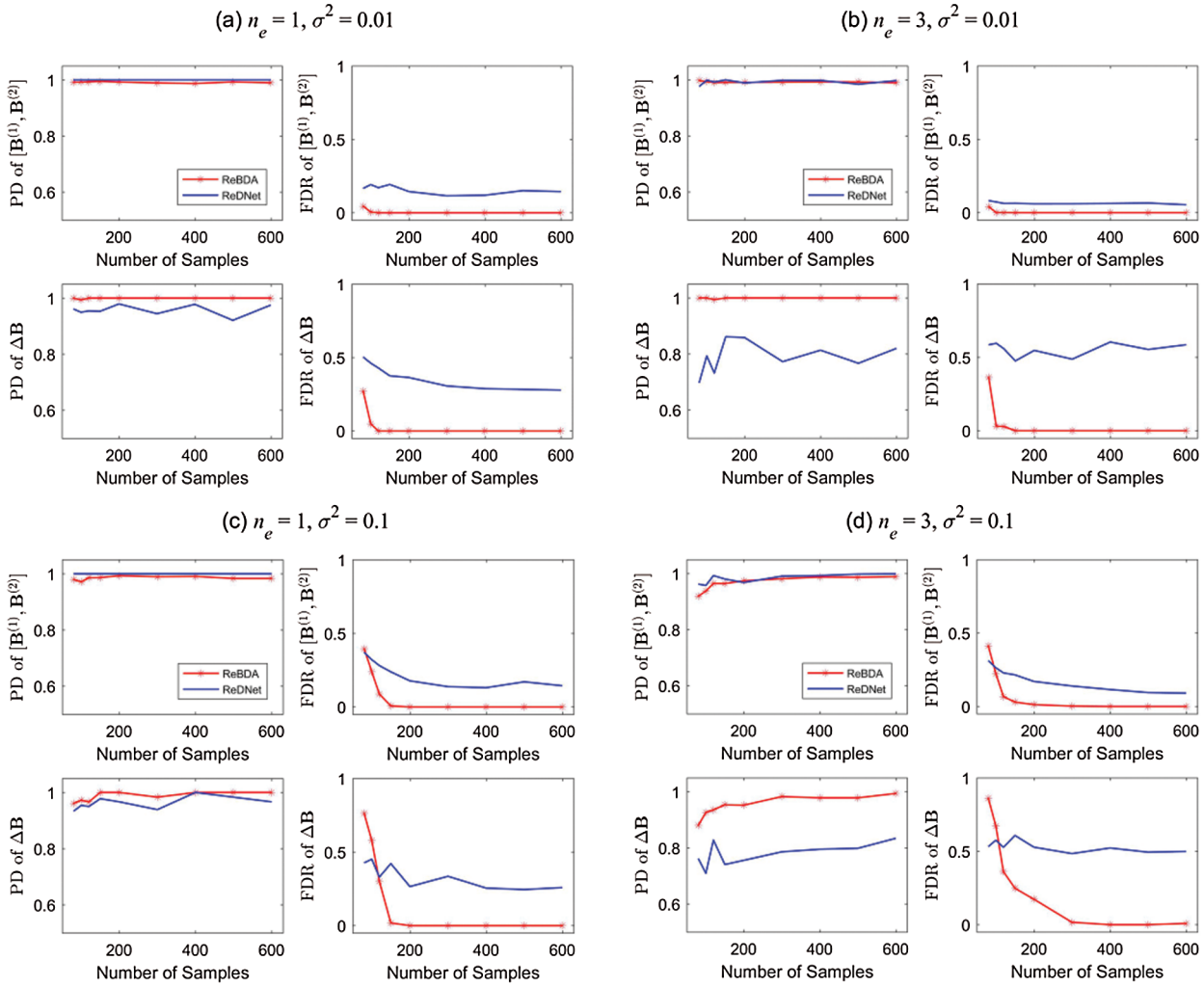

Figure 2: Performance of ReBDA and ReDNet for DAGs with 30 genes

We applied ReBDA to make differential analysis on the filtered data set. 15 genes and 60 regulatory edges were identified in the resulted differential GRN. To evaluate the significance of the identified edges, we re-sampled 100 bootstraps with 42 samples from the original data set and applied ReBDA on each bootstrap data set. As shown in Fig. 4, 4 genes and 2 regulatory edges were identified in over 90% of the differential GRNs inferred from the 100 bootstrap data sets. The weight of each edge was calculated by averaging the results of the 100 bootstraps.

In the above differential GRN inferred from the filtered real data set, 4 highly confidence genes were found to be related to lung tumor. Some previous literature based on experimental approaches have demonstrated that most of these genes are related to lung cancer or other cancers. Lin et al. [36] discovered and demonstrated that USP10 suppresses tumor cell growth through potentiating both SIRT6- and p53-mediated suppression of the oncogene c-myc. The results of [37] unravel the existence of a negative feedback loop of PP4R2 on IKK/NF-κB signaling, that suppresses lung cancer migration/invasion capability. BTF3 was confirmed abnormality in various cancer tissues (such as gastric cancer) [38,39]. Moreover, HSF1 was demonstrated to be associated with gastric cancer [40], breast cancer and two of the studied SNPs correlated significantly with cancer development [41].

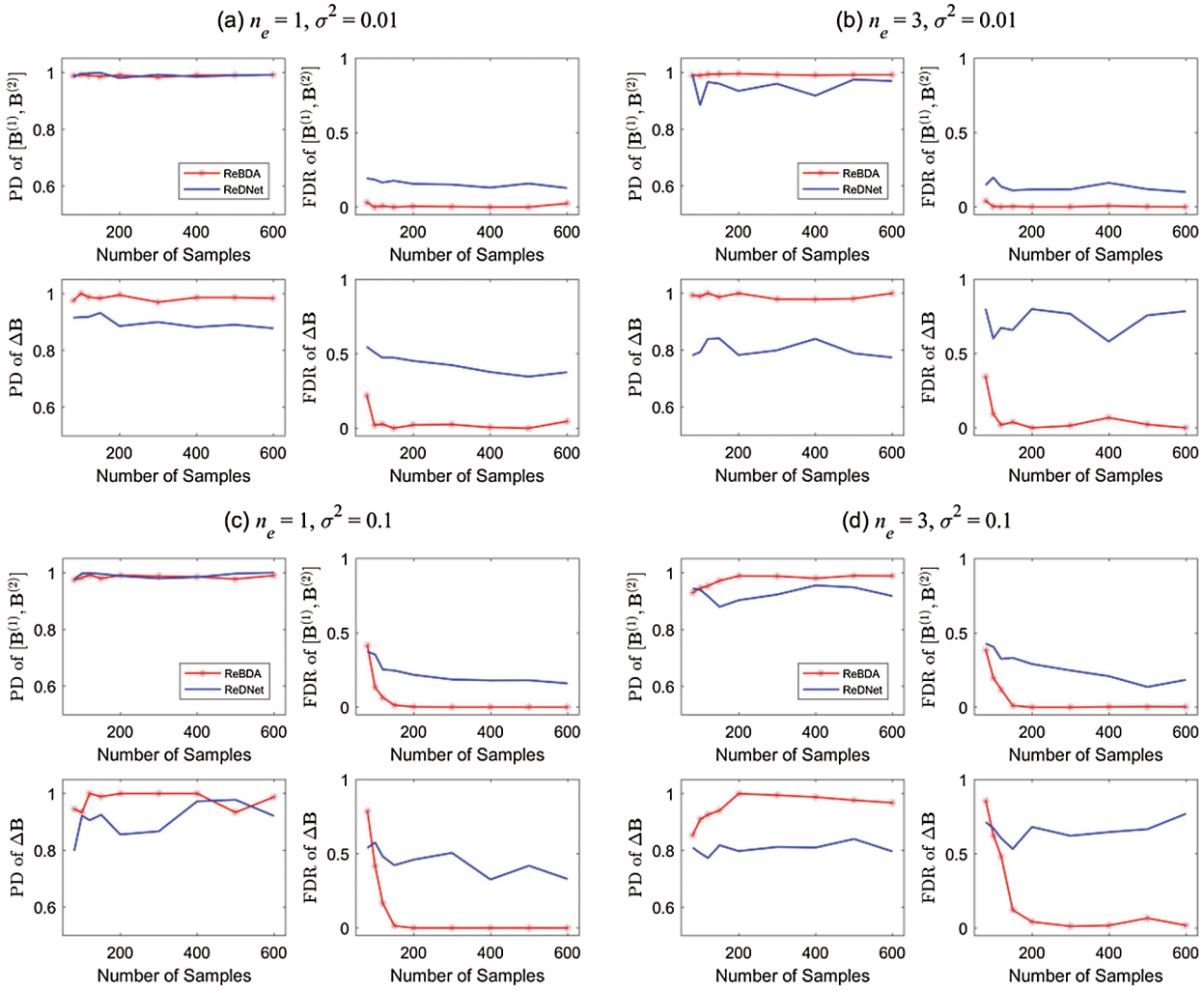

Figure 3: Performance of ReBDA and ReDNet for DCGs with 30 genes

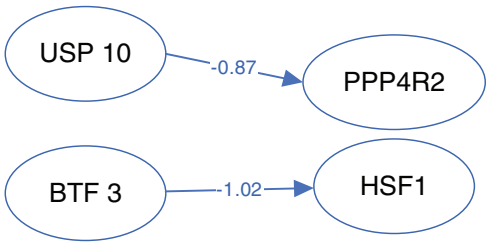

Figure 4: The differential GRN constructed by regulatory edges identified in over 90% of the 100 bootstraps

4 Discussion and Conclusion

In the ReDNet algorithm, Ren et al. [18] proposed a re-parametrization-based differential analysis method for SEMs, they re-parameterized the pairwise SEMs to form a joint model, and then applied adaptive lasso to estimate the summed structure and differential structure simultaneously. In this paper, we developed a novel differential analysis method for GRNs under different conditions based on re-parametrization named ReBDA. The ReBDA algorithm was also developed for GRNs modeled with

SEMs, which are widely used for GRN inference integrating gene expression data with genetic perturbations. At the first stage, the original two pairwise SEMs corresponding to the condition-specific GRNs are re-parameterized as an integrated SEM incorporating individual GRN and differential GRN. The re-parameterization method in this stage is different from that in ReDNet, which could get effective improvement on the predictive accuracy. Then at the second stage, a Bayesian inference method following the idea in LRBI [19] is developed to solve the re-parameterized integrated SEM. In ReDNet, Chen et al. [42] adopted a two-stage penalized least squares (2SPLS) method to solve its re-parameterized SEM. We compared the performance of LRBI and 2SPLS on SEMs with synthetic data, the results demonstrates that LRBI has better PD than 2SPLS and slightly better FDR in most cases, only when the sample size is relatively smaller (e.g., less than 100), the FDR of LRBI may be a little worse than that of 2SPLS. Therefore, we adopt LRBI to infer the re-parameterized SEM in our ReBDA algorithm to further improve the performance.

Computer simulations have proved that the ReBDA algorithm outperforms the ReDNet algorithm for networks with different settings in terms of both PD and FDR in general. The analysis of a real data set with 15 genes measured from 42 lung tumors and 42 adjacent normal tissues further demonstrated the availability and efficiency of ReBDA in practical biological applications.

Although the ReBDA algorithm could offer better inference accuracy in the differential analysis of GRNs, it still has several limitations: Firstly, it is mainly developed for differential analysis of pairwise GRNs, some extensive method could be explored to make comparison of more than two GRNs. Secondly, when with very limited samples and relatively higher σ^2, the FDR of ReBDA may be not that well, it is worthwhile to try some other different priors such as spike and slab prior to further improve the FDR. Thirdly, The reparameterization method adopted in ReBDA only supports pairwise datasets with the same sample size, in our future work, a more general re-parameterization method may be more applicable.

Funding Statement: The work was supported by grants from National Natural Science Foundation of China (Nos. 61502198, 61572226, 61472161, 61876069). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Lewis, C. C., Yang, J. Y. H., Huang, X., Banerjee, S. K., Blackburn, M. R. et al. (2012). Disease-specific gene expression profiling in multiple models of lung disease. *American Journal of Respiratory and Critical Care Medicine, 177(4),* 376–387. DOI 10.1164/rccm.200702-333OC.
2. Guo, W., Zhu, L., Deng, S., Zhao, X., Huang, D. (2016). Understanding tissue-specificity with human tissue-specific regulatory networks. *Science China-Information Sciences, 59(7),* e070105.
3. Claussen, J. C., Skiecevičienė, J., Wang, J., Rausch, P., Karlsen, T. H. et al. (2017). Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Computational Biology, 13(6),* e1005361. DOI 10.1371/journal.pcbi.1005361.
4. Butte, A. J., Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 5,* 418–429.
5. Meyer, P. E., Kontos, K., Lafitte, F., Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *Eurasip Journal on Bioinformatics & Systems Biology, 2007(1),* e79879.

6. Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S., Ko, S. B. et al. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics, 33(15),* 2314–2321. DOI 10.1093/bioinformatics/btx194.
7. Acerbi, E., Zelante, T., Narang, V., Stella, F. (2014). Gene network inference using continuous time Bayesian networks: a comparative study and application to Th17 cell differentiation. *BMC Bioinformatics, 15(1),* 387. DOI 10.1186/s12859-014-0387-x.
8. Larjo, A., Shmulevich, I., Lähdesmäki, H. (2013). Structure learning for Bayesian networks as models of biological networks. *Methods in Molecular Biology, 939,* 35–45.
9. Shimamura, T., Imoto, S., Yamaquchi, R., Miyano, S. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics, 19,* 142–153.
10. Xiong, M., Li, J., Fang, X. (2004). Identification of genetic networks. *Genetics, 166(2),* 1037–1052. DOI 10.1534/genetics.166.2.1037.
11. Liu, B., de la Fuente, A., Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics, 178(3),* 1763–1776. DOI 10.1534/genetics.107.080069.
12. Logsdon, B. A., Mezey, J. (2010). Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology, 6(12),* e1001014. DOI 10.1371/journal.pcbi.1001014.
13. Cai, X., Andrés, B. J., Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology, 9(5),* e1003068. DOI 10.1371/journal.pcbi.1003068.
14. Zhou, X., Cai, X., Kelso, J. (2020). Inference of differential gene regulatory networks based on gene expression and genetic perturbation data. *Bioinformatics, 36(1),* 197–204. DOI 10.1093/bioinformatics/btz529.
15. Danaher, P., Wang, P., Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, 76(2),* 373–397. DOI 10.1111/rssb.12033.
16. Mohan, K., Palma London, M. F., Witten, D., Lee, S. I. (2014). Node-based learning of multiple Gaussian graphical models. *Journal of Machine Learning Research, 15(1),* 445–488.
17. Wang, C., Gao, F., Giannakis, G. B., D'Urso, G., Cai, X. (2019). Efficient proximal gradient algorithm for inference of differential gene networks. *BMC Bioinformatics, 20(1),* 565. DOI 10.1186/s12859-019-2749-x.
18. Ren, M., Zhang, D. (2018). Differential analysis of directed networks. *Conference on Uncertainty in Artificial Intelligence, Monterey.*
19. Dong, Z., Song, T., Yuan, C. (2013). Inference of gene regulatory networks from genetic perturbations with linear regression model. *PLoS One, 8(12),* e83263. DOI 10.1371/journal.pone.0083263.
20. Gardner, T., di Bernardo, D., Lorenz, D., Collins, J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science, 301(5629),* 102–105. DOI 10.1126/science.1081900.
21. Brazhnik, P., de la Fuente, A., Mendes, P. (2002). Gene networks: how to put the function in genomics. *Trends in Biotechnology, 20(11),* 467–472. DOI 10.1016/S0167-7799(02)02053-X.
22. Thieffry, D., Huerta, A. M., Pérez-Rueda, E., Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays News and Reviews in Molecular Cellular and Developmental Biology, 20(5),* 433. DOI 10.1002/(SICI)1521-1878(199805)20:5<433::AID-BIES10>3.0.CO;2-2.
23. West, J., Bianconi, G., Severini, S., Teschendorff, A. E. (2002). Differential network entropy reveals cancer system hallmarks. *Scientific Reports, 2(46),* 802. DOI 10.1038/srep00802.
24. de la Fuente, A. (2010). From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases. *Trends in Genetics, 26(7),* 326–333. DOI 10.1016/j.tig.2010.05.001.
25. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58(1),* 267–288. DOI 10.1111/j.2517-6161.1996.tb02080.x.

26. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 67(1),* 91–108. DOI 10.1111/j.1467-9868.2005.00490.x.
27. Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological), 67(2),* 301–320. DOI 10.1111/j.1467-9868.2005.00503.x.
28. Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96(456),* 1348–1360. DOI 10.1198/016214501753382273.
29. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101 (476),* 1418–1429. DOI 10.1198/016214506000000735.
30. Park, T., Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association, 103(482),* 681–686. DOI 10.1198/016214508000000337.
31. Kyung, M. J., Gill, J., Ghosh, M., Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis, 5(2),* 369–412. DOI 10.1214/10-BA607.
32. Lu, T. P., Lai, L. C., Tsai, M. H., Chen, P. C., Hsu, C. P. et al. (2011). Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PLoS One, 6(9),* e24829. DOI 10.1371/journal.pone.0024829.
33. Gautier, L., Cope, L., Bolstad, B. M., Irizarry, R. A. (2004). Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics, 20(3),* 307–315. DOI 10.1093/bioinformatics/btg405.
34. Shabalin, A. A. (2015). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics, 28(10),* 1353–1358. DOI 10.1093/bioinformatics/bts163.
35. Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A. et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics, 47(6),* 569–576. DOI 10.1038/ng.3259.
36. Lin, Z., Yang, H., Tan, C., Li, J., Liu, Z. et al. (2013). USP10 antagonizes c-Myc transcriptional activation through SIRT6 stabilization to suppress tumor formation. *Cell Reports, 5(6),* 1639–1649. DOI 10.1016/j.celrep.2013.11.029.
37. Ho, M., Liang, C., Liang, S. (2016). PATZ1 induces PP4R2 to form a negative feedback loop on IKK/NF-κB signaling in lung cancer. *Oncotarget, 7(32),* 52255–52269.
38. Liu, Q., Zhou, J. P., Li, B., Huang, Z. C., Dong, H. Y. et al. (2013). Basic transcription factor 3 is involved in gastric cancer development and progression. *World Journal of Gastroenterology, 19(28),* 4495–4503. DOI 10.3748/wjg.v19.i28.4495.
39. Zhang, D. Z., Chen, B. H., Zhang, L. F., Cheng, M. K., Fang, X. J. et al. (2017). Basic transcription factor 3 is required for proliferation and epithelial-mesenchymal transition via regulation of FOXM1 and JAK2/STAT3 signaling in gastric cancer. *Oncology Research, 25(9),* 1453–1462. DOI 10.3727/096504017X14886494526344.
40. Kim, S. J., Lee, S. C., Kang, H. G., Gim, J., Lee, K. H. et al. (2018). Heat shock factor 1 predicts poor prognosis of gastric cancer. *Yonsei Medical Journal, 59(9),* 1041–1048. DOI 10.3349/ymj.2018.59.9.1041.
41. Almotwaa, S., Elrobh, M., AbdulKarim, H., Alanazi, M., Aldaihan, S. et al. (2018). Genetic polymorphism and expression of HSF1 gene is significantly associated with breast cancer in Saudi females. *PLoS One, 13(3),* e0193095. DOI 10.1371/journal.pone.0193095.
42. Chen, C., Zhang, M., Zhang, D. (2018). Two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research, 19(2),* 1–34.