BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 15 2012, pages 1990–1997
doi:10.1093/bioinformatics/bts296

Gene expression

Advance Access publication May 17, 2012

Robust identification of transcriptional regulatory networks using a Gibbs sampler on outlier sum statistic

Jinghua Gu[1], Jianhua Xuan[1,*], Rebecca B. Riggins[2], Li Chen[1], Yue Wang[1] and Robert Clarke[2,3]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, [2]Lombardi Comprehensive Cancer Center and Department of Oncology and [3]Department of Physiology and Biophysics, Georgetown University, Washington, DC 20057, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Identification of transcriptional regulatory networks (TRNs) is of significant importance in computational biology for cancer research, providing a critical building block to unravel disease pathways. However, existing methods for TRN identification suffer from the inclusion of excessive 'noise' in microarray data and false-positives in binding data, especially when applied to human tumor-derived cell line studies. More robust methods that can counteract the imperfection of data sources are therefore needed for reliable identification of TRNs in this context.

Results: In this article, we propose to establish a link between the quality of one target gene to represent its regulator and the uncertainty of its expression to represent other target genes. Specifically, an outlier sum statistic was used to measure the aggregated evidence for regulation events between target genes and their corresponding transcription factors. A Gibbs sampling method was then developed to estimate the marginal distribution of the outlier sum statistic, hence, to uncover underlying regulatory relationships. To evaluate the effectiveness of our proposed method, we compared its performance with that of an existing sampling-based method using both simulation data and yeast cell cycle data. The experimental results show that our method consistently outperforms the competing method in different settings of signal-to-noise ratio and network topology, indicating its robustness for biological applications. Finally, we applied our method to breast cancer cell line data and demonstrated its ability to extract biologically meaningful regulatory modules related to estrogen signaling and action in breast cancer.

Availability and implementation: The Gibbs sampler MATLAB package is freely available at http://www.cbil.ece.vt.edu/software.htm.

Contact: xuan@vt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 20, 2011; revised on May 11, 2012; accepted on May 15, 2012

1 INTRODUCTION

Identification of transcriptional regulatory networks (TRNs) is essential to understanding the functional roles of molecules and the mechanisms of complicated biological processes. In cancer research, those identified regulatory networks, also referred to as regulatory modules, can give insight into important biological pathways that drive the development of disease, based on which novel drugs or therapies may be developed. In addition to high-throughput microarray techniques that profile the entire transcriptome of cells, more specific information about regulation can be obtained from protein–DNA interaction data derived from DNA motif sequence analyses or ChIP-on-chip experiments. Binding motif is also called transcription factor binding site, a segment of DNA regulatory sequence, to which a specific regulatory protein binds preferentially. ChIP-on-chip experiments can provide whole-genome analysis of binding site locations of all regulatory proteins. Both binding motif data and ChIP-on-chip data provide an initial structure of the unknown network topology. The availability of the above techniques has promoted extensive researches on regulatory networks in the last decade and some of which are quite successful in applications such as yeast studies.

Early exploration in gene modules depends heavily on statistical methods such as principle component analysis (Alter *et al*., 2000) and independent component analysis (Lee and Batzoglou, 2003). These statistical methods have strong assumptions that the components are mutually orthogonal or statistically independent, which can hardly be satisfied in real biological systems. Other statistical approaches include dynamic Bayesian networks (Friedman *et al*., 2000), probabilistic Boolean networks (Shmulevich *et al*., 2002) and similarity-based methods (Hempel *et al*., 2011). In particular, Hempel *et al*. (2011) has performed a comparative analysis using 21 different measures and 6 scoring schemes, investigating their ability in reconstructing gene regulatory networks. Many of the earlier mentioned methods either utilize information from gene expression data alone (i.e. do not utilize prior network knowledge from protein–DNA binding data) or are infeasible for learning large network structures (Chickering, 1996). On the recent integrated use of gene expression data and binding data such as ChIP-on-chip data, Liao *et al*. (2003) proposed a computational method called network component analysis (NCA), which is based on a log-linear model of the relationship between mRNA abundance and transcription factor activity. Subjected to some mild identifiability conditions referred to as NCA criteria, the NCA method can be used to effectively estimate hidden transcription factor activities. Chang *et al*. (2008) later developed a fast version of the NCA method with a closed-form solution by using matrix factorization techniques.

*To whom correspondence should be addressed.

Despite the earlier mentioned promising progress made for gene regulatory module identification, two major concerns need to be further addressed. First, microarray gene expression data are often corrupted by noise, and the situation is even worse in real microarray data acquired from patients with cancer. Second, protein–DNA binding data, derived from either ChIP-on-chip experiments or motif sequence analyses, often contain false-positive and false-negative connections, further complicated by the discrepancy between binding affinity and true regulation. Recent studies show that transcription factors regulate their target genes in a condition-specific manner, i.e. the regulation is not only determined by the potential of binding strength but also depends on the environmental condition. To address these concerns, especially regarding false-positive connections in the binding data, Brynildsen *et al.* (2006) showed that by selecting only a subset of genes with high confidence, a significant increase in the fitting performance can be achieved and more variations in gene expression data can be explained. Based on the concept of network versatility, Brynildsen *et al.* used 'condition number' (CN) (i.e. ratio of the largest eigenvalue to the smallest eigenvalue) of expression matrix to measure the consistency between the target genes; however, the robustness of CN is not guaranteed; as a matter of fact, it degrades quickly as noise level increases or network topology becomes denser (see Supplementary Material S5), thus limiting its further applications to cancer studies.

As such, more robust methods are needed in order to correctly infer condition-specific gene regulatory modules from noisy data sources. In this article, we propose a new Gibbs sampling-based method for target gene identification and regulatory network reconstruction. Our method is capable of reducing false-positives contained in binding data by iteratively selecting the highly confident genes from an initial pool. In particular, a novel statistic for testing the confidence of target genes, namely outlier sum (OS) of regression *t*-statistic, is specifically designed to pin-down confident target genes; based on this statistic a Gibbs sampling strategy is utilized to sample target genes in a high probability as governed by the underlying distribution. This Gibbs sampler on OS statistic is an effective and robust approach to gene regulatory module identification because of the following features: (i) the regression *t*-statistic is a effective metric to statistically measure the dependency of genes and works particularly well when signal-to-noise ratio (SNR) is low; (ii) the OS of regression *t*-statistic reflects the aggregated evidence of the regulation between one transcription factor and its target genes; (iii) the Gibbs sampling scheme allows us to study the marginal confidence of the target genes, which may be regulated by multiple transcription factors; (iv) in addition to target gene identification, a procedure of significance analysis is also incorporated into our scheme to test the statistical significance of transcription factors being studied.

Experimentally, we have tested the OS-based Gibbs sampler on both simulation and yeast cell cycle data to demonstrate its effectiveness and efficacy. We compared the performance of our method with a sampling method based on CN under different SNRs and network topologies. The results showed that our method exhibited an increasing advantage for regulatory network identification over the competing method as the SNR decreases. The performance also remained robust when the network topology became denser or more false-positive connections were introduced. We then applied our Gibbs sampler to breast cancer cell line data and identified biologically meaningful regulatory network modules associated with condition-specific estrogen actions in breast cancer.

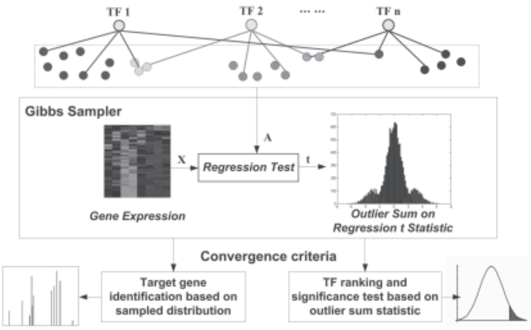
TF 1
TF 2
…… ……
TF n
Gibbs Sampler
A
X
Regression Test
t
Gene Expression
Outlier Sum on Regression t Statistic
Convergence criteria
Target gene identification based on sampled distribution
TF ranking and significance test based on outlier sum statistic

Fig. 1. A general workflow of the Gibbs sampler on OS statistic

2 METHODS

Our Gibbs sampling method based on OS statistic aims to reliably identify regulatory networks consisting of transcription factors and their target genes. Figure 1 shows the workflow of our proposed sampling-based framework. The Gibbs method starts from the initial topology (A) of the TRN by selecting a pool of candidate target genes for each transcription factor. Regression analysis among the candidate gene expression X is performed and the regression t-statistic is then aggregated into an OS. A Gibbs strategy is adopted so that we can sample genes according to their marginal OS of regression t-statistic. After the sampled distribution is converged, both transcription factors and their target genes can be ranked for regulatory module reconstruction. Moreover, a statistical procedure based on OS statistic is designed to test the significance of transcription factors, which can help us prioritize some important regulatory modules for real biological studies.

2.1 A log-linear model for regulatory network identification

The expression levels of genes in live cells are controlled by many activated regulatory proteins in the form of transcription factor activities (TFAs) as binding to the promoters of the genes. Liao *et al.* (2003) have modeled the relationship between gene expression level and TFA as a log-linear model that can be written in the following compact form:

$$E = A \times P, \tag{1}$$

where E is an $N \times M$ gene expression matrix and A is an $N \times L$ connectivity matrix representing the regulation strength of a regulatory network. Here, P(with a size of $L \times M$) is a matrix representing the unknown TFAs.

The log-linear model of transcription regulation depicts a bipartite network constituted by transcription factors and their target genes. An initial skeleton of the network can be constructed from binding data such as ChIP-on-chip data and motif information. For a given transcription factor, we define a group of genes with binding potentials above given threshold as 'candidate genes' or 'candidate pool'. However, due to experimental imperfections and the gap between binding ability and true regulation, the initial skeleton or 'snap-shot' of the regulatory network defined by the candidate genes contains a lot of false connections. Among all the candidate genes for a given transcription factor, those true target genes, with supporting evidence of gene regulation from both gene expression data and protein–DNA interaction data, are referred to as 'foreground genes'; on the other hand, the term, 'background genes', refers to those genes whose gene expression shows little evidence to support true regulation but are included in the initial regulatory network (due to experimental error or technological deficiency in protein–DNA interaction data). It is reasonable to assume that the TFA estimation based on foreground genes is more reliable than that from the entire population of genes or a group of initial target genes contaminated by background genes. Hence, identification of foreground genes becomes a crucial step for reliable reconstruction of gene regulatory modules. With a group of highly confident genes with strong regulatory evidence from both gene expression and binding data, we can assess the significance of any given regulatory module in a condition-specific manner, while gaining an improved estimation of the activity of transcript factors regulating the network.

2.2 Target gene identification based on regression model and OS statistic

The log-linear model in equation (1) states that a foreground gene's expression is a direct result from activities of corresponding transcription factors. Therefore, we can use foreground genes' expression to indicate hidden TFAs. Suppose that we assign L foreground genes, $[\theta_1,\theta_2,...,\theta_L]$, each to represent one unique transcription factor. These L 'representatives' are referred to as 'seed genes' or 'seeds' for the L transcription factors. We can write the expression of any candidate gene as linear regression form in terms of the seed gene expressions as follows:

$$y=X\beta, \tag{2}$$

where y denotes expression of the candidate gene, $X=[x_1,\ x_2,...x_L]$ is the expression of the L seed genes. $\beta=[\beta_1,\beta_2,...\beta_L]^{\mathrm{T}}$ is the scalar coefficient vector. Under some mild assumptions, it can be derived that $\beta_i=0$ indicates that gene y is not regulated by TF i (see Supplementary Material S3). Hence, we can conduct a significance test on β_i, to determine whether a gene is regulated by TF i. The least square estimate of β (Montgomery, 2006) is given by $\hat{\beta}=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y=CX^{\mathrm{T}}y$. Note that we also denote $C=(X^{\mathrm{T}}X)^{-1}$ in the earlier equation. The hypotheses for testing the significance of any individual coefficient β_i are as follows:

H_0(null hypothesis)$:\beta_i=0$; H_1(alternative hypotheisis)$:\beta_i\neq 0$.

A suitable test statistic for the earlier hypotheses is given by

$$t=\frac{\hat{\beta}_i}{S_{\mathrm{b}}}=\frac{\hat{\beta}_i}{\sqrt{\frac{\mathrm{SSR}}{M-L-1}\cdot C_{ii}}}, \tag{3}$$

where $\mathrm{SSR}=y^{\mathrm{T}}y-\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y$ (squared sum of residuals) and $M-L-1$ is the degree of freedom. C_{ii} is the ith diagonal element of $C=(X^{\mathrm{T}}X)^{-1}$. The test statistic defined in equation (3) is a conditional function, testing the significance of regression between gene y and seed gene θ_i for TF i, based on the seed genes selected for other transcription factors. If the absolute value of t is larger than a given threshold (e.g. 1.96, corresponding to 95% confidence level in Gaussian approximation), it means that candidate gene y is a foreground gene with significant regression effect on seed gene θ_i conditioned on other seeds; otherwise, gene y is more likely a background gene.

The regression test discussed earlier can be utilized to identify foreground genes in the candidate pool for given transcription factors based on selected seed genes. However, we do not know beforehand whether the L seed genes are foreground genes or not. Fortunately, we know that there are multiple foreground genes in the candidate pool, which suggests that we can use foreground genes to support each other. Assume that the candidate pool, $\Phi_i=\{\theta_{i1},...,\theta_{iK_i}\}$, of TF i has cardinality K_i, among which the foreground genes are denoted as set Φ_{iF}; the background genes are denoted as set Φ_{iB}; thus, $\Phi_i=\Phi_{iF}\cup\Phi_{iB}$. To start our regression analysis procedure, we select a random gene $\theta_{ik}\in\Theta_i(1\leq k\leq K_i)$ as seed gene θ_i and carry out regression analysis between each single gene in Θ_i (except θ_{ik} itself) and the L seed genes $[\theta_1,...,\theta_{i-1},\theta_i=\theta_{ik},\theta_{i+1},...,\theta_L]$. Using the remaining candidate genes for TF i, we will obtain a total number of K_i-1 test statistics, $t_{kj}, j=1...K_i$, $j\neq k$. If $\theta_{ik}\in\Theta_{iF}$, because genes in Θ_{iF} are foreground genes, t_{kj} will be significantly larger than 0. This means that t_{kj} corresponding to foreground genes are outliers (represented by red dots in Fig. 2) to the null distribution, which is a Student's t distribution with $M-L-1$ degrees of freedom [denoted as $t\left(\nu=M-L-1\right)$ and shown as the blue curve (probability density function) in Fig. 2]. On the other hand, if $\theta_{ik}\in\Theta_{iB}$, all t_{ki} should follow the null distribution very well because a background gene should not have a strong regression relationship with either true target genes of TF i or other background genes.

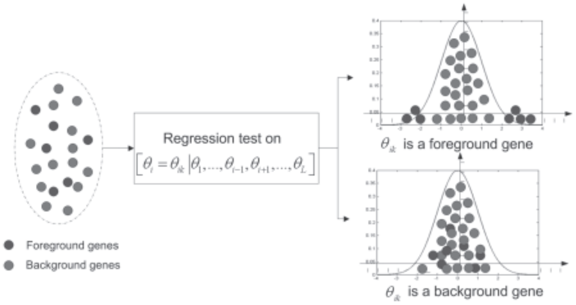
Foreground genes
Background genes
Regression test on
$\left[\theta_i=\theta_{ik}\left|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_L\right.\right]$
θ_{ik} is a foreground gene
θ_{ik} is a background gene

Fig. 2. Target gene identification based on regression analysis

To determine whether a seed gene for TF i is well selected (which means θ_{ik} is a true target gene), we propose to use the OS statistic (Tibshirani, 2007) of θ_{ik} as follows:

$$\mathrm{OS}=\sum_{j\neq k}\left|t_{kj}\right|\cdot I\left(\left|t_{kj}\right|-t_{\alpha/2,M-L-1}\right), \tag{4}$$

where $I(x)=1$ when $x\geq 0$; $I(x)=0$ otherwise. Equation (4) shows that for any selected seed θ_{ik} of TF i, we will sum up all t_{kj} that are 'outliers' to the null distribution using a threshold $t_{\alpha/2,M-L-1}$. If the selected θ_{ik} is a foreground gene, the corresponding OS statistic should be significantly larger than 0. Contrarily, the OS_{ik} will be close to 0 if θ_{ik} is a background gene. The OS statistic represents a conditional test of the quality of a selected seed for TF i, given $L-1$ seed genes for other transcription factors, which will be further discussed in the next subsection.

2.3 A Gibbs sampler on the OS statistic

The OS statistic defined in the previous section is specifically designed to test the significance of a seed θ_i by evaluating its suitability to be a true target gene of TF i based on other seeds. Hence, we can view the OS statistic as a conditional function as follows:

$$\mathrm{OS}|_{\theta i}=f_{\mathrm{OS}}\left(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_L\right), \tag{5}$$

where $f_{\mathrm{os}(\cdot)}$ denotes the test function of OS. From equation (5), we can see that the power of the OS statistic is influenced by the quality of the other seeds $\theta_j(j=1,\ldots,L,j\neq i)$. If the conditional seeds for other transcription factors are contaminated by background genes, the performance of test statistic t from regression analysis may degrade, consequently weakening the power of the OS statistic. Hence, the OS statistic that we seek for TF i should be preferably defined only by the seed genes of TF i, i.e. a desirable, univariate function $f_{\mathrm{os}}(\theta_i)$ instead. However, there is no effective way to directly derive $f_{\mathrm{os}}(\theta_i)$ from equation 5 analytically. Fortunately, we can use a Monte Carlo strategy to convert the conditional function defined in equation (5) into a probability density function as follows:

$$p\left(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_L\right)=\frac{1}{K_0}\cdot f_{\mathrm{OS}}$$
$$\times\left(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_L\right), \tag{6}$$

where K_0 is an unknown normalization constant (see Supplementary Material S1 for more details). Consequently, we can re-interpret the problem of calculating $f_{\mathrm{os}}(\theta_i)$ as how to estimate the marginal probability distribution, $p(\theta_i)[\propto\propto_{\mathrm{os}}(\theta_i)]$, from the conditional distribution $p(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_L)$. In particular, we can estimate the marginal density $p(\theta_i)$ by employing a Gibbs sampling strategy (Casella and

George, 1992), generating samples from a set of conditional probability distributions as follows:

$$\theta_i^{(t+1)} \sim p\left(\theta_i | \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_L^{(t)}\right), \quad \text{for } 1 \leq i \leq L, \qquad (7)$$

where t denotes the tth step in the sampling iteration process. At iteration t, we will sequentially sample one seed gene for each TF based on the corresponding conditional probability density function. After we have sampled one gene for TF i, the seed for this transcription factor will be updated by the newest sample. Through a sufficient number of sampling iterations, we will have a group of samples that are drawn from the marginal distribution of seed genes for each TF. From equation (6), we know that during the sampling procedure, we are also estimating the marginal function of the OS statistic because it equals to the probability density function except a normalizing constant. This means that genes with larger marginal probability density have consistently larger marginal OS. Hence, we can approximate the empirical probability density function of $p(\theta_i)$ by calculating the frequency of the samples. The candidate genes for θ_i with large frequency count are more likely the true target genes for TF i. Note that the convergence of Gibbs sampler is theoretically guaranteed and number of iterations needed before convergence has been recommended based on real data analysis (Supplementary Material S2).

Besides for target gene identification, one advantage of the proposed Gibbs sampling method is that it can facilitate to identify true (active) transcription factors. OS statistic is an index showing the possibility that one gene is a true target gene for given TF. If one TF is a background (inactive) TF that presumably does not regulate any target genes, the test statistic t will hardly be significant between any two of its candidate genes. As a result, the OS of all the candidate genes should be close to 0. On the other hand, if one TF is a true regulator, the true target genes of this TF should support each other and yields large OS. Hence, the OS-based sampling method can prioritize both transcription factors and their corresponding target genes to reconstruct regulatory networks from gene expression and binding data.

2.4 Significance test of identified transcriptional regulatory modules

Our Gibbs sampling method is designed to extract a confident set of target genes through a probabilistic manner for condition-specific regulatory module identification. However, when working on the real microarray data such as breast cancer data, we need strong statistical evidence to support and justify the identified regulatory modules before we can proceed to conduct costly biological experiments for validation. To address the concern that the results obtained from certain dataset is not by chance, we propose a workflow to test the statistical significance of any given regulatory module based on the OS distribution associated with the corresponding transcription factor. We summarize the procedure for testing the significance of the regulatory module of TF i as follows:

(1) Generate a null distribution of the regression t-statistic from the entire gene set being studied, denote as N_1.

(2) Suppose TF i has n candidate target genes in the initial pool. Generate the null distribution of the OS statistic of size n based on N_1, denote as $N_{2,n}$.

(3) Test whether the sampled distribution for TF i has the same mean as the null distribution $N_{2,n}$. The statistical significance (i.e. P value) is conventionally denoted as p_i.

Through this significance test procedure, we are able to highlight the modules with strong evidence supporting regulation events under certain condition by both microarray data and binding information. More detailed information regarding the significance test on regulatory modules is provided in the Supplementary Material S4.

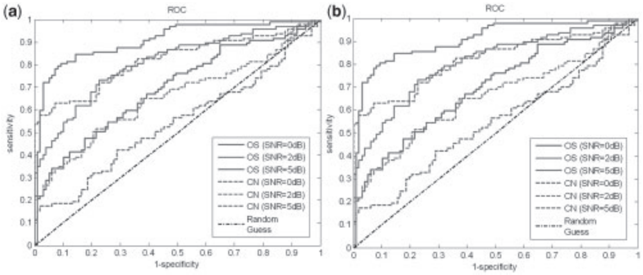

Fig. 3. ROC comparison of Gibbs sampling method based on OS statistic and CN using linear simulation data and SynTReN. (**a**) Experiment on linear simulation. (**b**) Experiment on SynTReN data

3 RESULTS

3.1 Simulation and yeast data

We first tested our sampling method on simulation data to assess its performance for target gene identification. We used Matlab functions to synthesize an initial network that consists of ~200 target genes each with 30 experiments and 20 transcription factors with linearly independent TFAs. Among all the target genes, one-half of them are foreground genes whose expression are generated according to equation (1) to guarantee the linearity, while for the other half, namely background genes, we randomly generate their gene expression profiles for this experiment. We compared the performance of our proposed sampling method for target gene identification with an existing Gibbs sampling method that is based on CN evaluation (Brynildsen *et al.*, 2006). As demonstrated in this article, the competing method outperforms traditional model-fitting methods (e.g. Ordinary Least Squares; Huber, Cauchy and Fair weighted robust regressions) by reducing the over-fitting to background genes. We evaluated the algorithm performance based on several simulated datasets with different SNRs. Figure 3a shows the receiver operating characteristic (ROC) curves for target gene identification using the two Gibbs samplers based on OS and CN, respectively.

From Figure 3a, we can see that for different SNRs (5, 2 and 0 dB), our OS-based sampling method consistently outperforms the CN-based sampling method with a 0.1 increase in area-under-curve (AUC). When SNR decreases to 0 dB, the CN-based sampling method performs slightly better than random guess (AUC = 0.5), while the OS-based sampler still maintains a reasonable performance with the AUC of 0.695.

For a more realistic simulation of transcription regulatory networks, we use the SynTReN software (Van den Bulcke *et al.*, 2006) to generate synthetic networks and gene expression profiles. SynTReN extracts sub-networks from known yeast networks and utilizes Michaelis-Menten and Hill kinetics to model the interaction kinetics. The main difference between the expression data generated by SynTReN and those by equation (1) is that SynTReN models the expression of target genes as a nonlinear function of the expression of their regulators, which is a more realistic representation of biological regulation relationships than equation (1). The network topology and expression data generated by SynTReN are better mimic of the real TRNs than simple linear combinations. We generated a network consisting of 100 foreground genes and 100 background genes regulated by 24 TFs in total. We also set different levels of experimental noise to test the robustness of our proposed method against the quality degradation of gene expression data. Figure 3b shows that as the noise level increases, both methods suffer from performance degradation by a corresponding decrease in the AUCs. However, under the same noise condition, our OS-based sampling method consistently performs better than the Gibbs sampler based on CN with an average AUC increase of 0.1 (Supplementary Tables S5 and S6).

To further demonstrate its improved performance, we compared GibbsOS to four other existing methods [i.e. rank correlation (Hempel *et al*., 2011), linear angle regression (Efron *et al*., 2004), FastNCA (Chang *et al*., 2008) and COGRIM (Chen *et al*., 2007)] on SynTReN-simulated networks. Both ROC and precision-recall analyses were performed; the results have demonstrated that GibbsOS consistently outperforms the existing methods when different levels of noise are introduced (see Supplementary Material S6 for more details).

We also applied the GibbsOS method to Spellman's yeast cell cycle data (Spellman *et al*., 1998) for further algorithm validation. The experimental results are detailed in the Supplementary Material S7 (Table S7); many important transcription factors associated with different cell cycle phases were identified, and their target genes are functionally enriched in cell cycle-related biological processes (note that we used the Gene Ontology term finder from Saccharomyces Genome Database for functional enrichment analysis). These results have clearly showed that our GibbsOS approach can discover more significant clusters associated with mitosis, cell cycle regulation and processes than the CN-based sampler can.

3.2 The impact of network topology and false-positive connections in prior knowledge

As mentioned in the *Introduction* section, protein–DNA interaction data have imperfections with certain false-positive and/or false-negative rate. In this article, we focus on addressing the problem of filtering out false-positive connections as obtained from ChIP-on-chip or binding motif data. Specifically, we aim to separate foreground genes from background genes by integrating gene expression data and protein–DNA interaction data. A high false-positive rate in ChIP-on-chip data means that in the initial candidate pool, a large number of candidate genes are background genes. Following the simulated studies in Section 3.1, we systematically studied the influence on the performance as the density and false-positive rate in initial network topology varied in a certain range. We compared the performance of GibbsOS method with that of the CN-based method for regulatory network identification.

Figure 4 shows the ROC comparison result on simulated networks with different false-positive rates. Note that a parameter, nf:nb, is used to control the proportion of foreground genes contained in the initial candidate pool. More detailed information regarding the comparison study can be found in Supplementary Table S8. From Table S8, we can see that our proposed GibbsOS method is quite robust against the increased number of background genes in initial topology by sustaining a reasonable AUC of 0.86883, when the actual proportion of false-positives (PFPs) in the network reaches 87% (nf:nb ≈ 1:8). Comparatively, we also see that GibbsOS outperforms the CN-based method in our simulation setting.

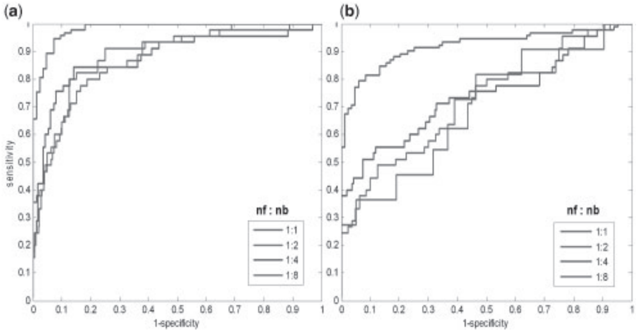

Fig. 4. Performance comparison between the proposed GibbsOS method (**a**) and the CN-based method (**b**), when false-positive rate in the initial topology increases. 'nf:nb' is a parameter used to control the ratio of foreground genes and background genes. 'nf:nb = 1:1' means that the number of foreground genes is equal to that of background genes

Besides the PFPs, network topology will affect the performance of the proposed method for regulatory network identification through other factors such as network size (cardinality), network degree (connection density) and number of transcription factors. To comprehensively evaluate the robustness of our GibbsOS method, we studied the ROC/AUC performance under three typical network configurations with different network sizes and degrees (as shown in Supplementary Fig. S14). The majority of genes in Network A have degree 1 or 2, which means that the overall network topology is quite sparse. The average degree in Network B increases to more than 3 and Network C is densely connected with an average degree of ∼5. Network A and C are similar to the example networks in (Boscolo *et al*., 2005), and network B is similar to the example in (Tran *et al*., 2005). We have carried out a quite comprehensive study to evaluate the robustness of GibbsOS on these synthetic networks that differ in size, average degree, PFP and number of transcription factors. The detailed results can be found in Supplementary Material S8 (Figs. S15–S16 and Tables S2–S3); the experimental results have demonstrated that our proposed GibbsOS method maintains a reasonably good power for regulatory network identification (AUC $>$ 0.8). GibbsOS only shows minor performance degradation as the topology of network becomes denser or the number of background genes increases (see Supplementary Material S8).

4 BREAST CANCER DATA

We applied our Gibbs sampler to two breast cancer cell line datasets to identify regulatory modules associated with two different estrogen-associated conditions: estrogen induced (Creighton *et al*., 2006) and long-term estrogen deprived (LTED) (Aguilar *et al*., 2010). The estrogen-induced dataset consists of three estrogen-dependent breast cancer cell lines (MCF-7, T47D and BT-474) treated with 17β-estradiol (E2 or estrogen) *in vitro* over a time course ranging from 0 to 24 h. The original paper reported four E2-induced clusters with clear over-expression patterns at different starting or ending points. The LTED MCF-7 dataset contains eight time points ranging from day 0, when the cells were first cultured in media depleted of estrogen, up to 6 months when the cells are fully estrogen independent. We used Hierarchical Clustering Explorer to identify two clusters that are over-expressed or under-expressed after day 90, which is a critical time point when these (and other) LTED cell models have recovered from the initial effects of estrogen deprivation and begin to proliferate again (Aguilar *et al.*, 2010; Martin *et al.*, 2003; Masamura *et al.*, 1995) (Supplementary Material S9).

To understand the role that estrogen signaling plays in transcriptional regulation that leads to estrogen independence, and ultimately to breast cancer progression, we selected 26 estrogen receptor (official gene symbol: ESR1) related transcription factors (Supplementary Material S10). These transcription factors are known to be directly or indirectly associated with ER signaling and participate in important cellular functions such as apoptosis and cell cycle regulation. Because comprehensive ChIP-on-chip data are not fully available for the human transcriptome, we used binding motif information for these 26 transcription factors to construct an initial pool of target genes. A threshold (e.g. 10 percentile of binding motif scores) was used and genes with binding motif scores larger than this threshold were regarded as candidate targets. To study gene regulatory modules in a condition-specific manner, we then divided the genes identified in each condition into two groups, namely 'early up-regulated' and 'late up-regulated'. We applied our Gibbs sampler to both groups of genes under different E2 conditions and identified significant transcriptional regulatory modules associated with each condition. The *P*-values of the binding motifs are listed in Supplementary Tables S9 and S10.

In summary, we observed that ESR1 has strong regulatory activity in the estrogen-induced condition but is no longer a significant transcriptional modulator in LTED cells, which is consistent with global transcriptional reprogramming and decreased reliance on pro-proliferative ESR1 signaling during the acquisition of estrogen independence. V$AP1_Q2_01, V$AP1_Q4_01 and V$CREB_Q2 were all highly enriched in the early E2 up-regulated condition, while V$NFKB_Q6_01 had strong regulatory activity under LTED conditions. The latter observation echoes a recent study in which up-regulation of NFκB contributes to endocrine resistance in ER-positive breast cancer cells (Chen *et al.*, 2010; Pratt *et al.*, 2003; Riggins *et al.*, 2005).

When we compared the early up-regulated group under estrogen-induced conditions (indicative of the acute response to E2) to the late up-regulated group under LTED conditions (indicative of E2 independence), we uncovered several underlying regulatory networks (Fig. 5). AP1 and ESR1 were uniquely activated under estrogen-induced conditions while expression of NFκB, TP53 and CEBPA target genes were associated with the LTED condition. STAT family transcription factors and SP1 showed strong regulation under both conditions. For these regulatory networks, the expression patterns and corresponding motif binding sites of target genes can be found in Supplementary Material S12. We further obtained network information of the transcription factors from Ingenuity Pathway Analysis software. From Figure 5, we can see that ESR1 is extensively connected to other transcription factors in a variety of ways, indicating its crucial role in mediating transactional regulation under E2 conditions. ESR1 can either interact with SP1 and TP53 through protein–DNA interactions and protein–protein interactions, or promote/inhibit activities of NFκB, CEBPA and C-fos, which is the component of AP1 protein complex. Although STAT family proteins do not directly interact with ESR1, their association with the estrogen receptor may be indirectly obtained from interactions with SP1, NFκB and AP1. Hence, we conclude that ESR1 remains highly active under estrogen-induced conditions while its canonical transcriptional function is weakened in LTED conditions. Concomitantly, alternate transcriptional pathways are activated through either direct or indirect crosstalk with estrogen receptor.

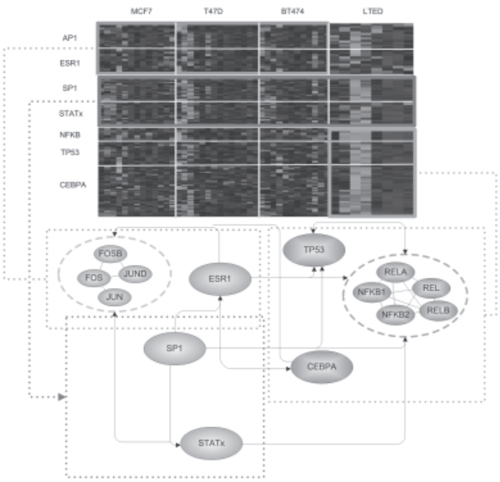

Fig. 5. Identified transcriptional modules and their interactions associated with different estrogen conditions. Note that the interactions between different transcription factors were obtained from Ingenuity Pathway Analysis (IPA) (www.ingenuity.com). AP1 and ESR1 are significantly enriched in early up-regulated estrogen-induced group while NFκB, CEBPA and TP53 have significant regulatory activities with late up-regulated long-term deprived group. Meanwhile, there is strong evidence showing that SP1 and STAT family proteins are important mediators for transcriptional regulation under both groups. Identified significant target genes are listed in Supplementary Tables S11–S18

Interestingly, the TRANSFAC motif V$STAT_01 was activated in all four groups, suggesting that this family of transcription factors is particularly important in breast cancer progression and estrogen independence. This is a general STAT-binding motif that does not preferentially recruit any particular family member. However, under LTED conditions the family member STAT5B is the most significantly up-regulated, suggesting that this member of the STAT family is responsible for driving the network in this setting. This is consistent with the known activities of STAT5B in breast cancer; a naturally occurring mutant of STAT5B functions as a dominant-inhibitory molecule that blocks ESR1 transcriptional activity (Yamashita *et al.*, 2003), STAT5B is an essential target of BCAR1/c-Src signaling in Tamoxifen-resistant breast cancer (Riggins *et al.*, 2006), and a constitutively active variant of STAT5B can independently induce Tamoxifen resistance in ER+ breast cancer cell lines (Fox *et al.*, 2008).

Given the apparent importance of STAT5B signaling in this context, we more closely examined the V$STAT_01 target genes that our method identified (Table S16). GCSH (Hungermann *et al.*, 2011), CYP1B1 (Angus *et al.*, 1999; Han *et al.*, 2010), CISH (Fang *et al.*, 2008), and PHB (He *et al.*, 2008) have all previously been shown to be either associated with or directly regulated by estrogen/ESR1 signaling, yet all four of these genes are strongly up-regulated under late LTED conditions. Our findings therefore suggest that STAT signaling via STAT5B may functionally replace ESR1 activation of key genes that support the pro-proliferative and/or pro-survival phenotype of LTED cells, ultimately contributing to the acquisition of estrogen independence and breast cancer progression.

5 DISCUSSION

Identification of transcriptional regulatory modules plays a key role in understanding the mechanism of cancer progression. However, intrinsic defects of microarray data and ChIP-on-chip/binding-motif that caused by noise and the discrepancy between gene profiles and prior binding knowledge make it challenging to correctly recover hidden regulation patterns. The CN-based sampling method proposed by Brynildsen *et al.* is endeavored to filter out the 'background genes' with no regulation patterns but are falsely introduced by binding data. The CN is defined as the ratio of the largest eigenvalue to the smallest eigenvalue, which may be affected by noise because the smallest eigenvalue typically represents the noise power while the larger eigenvalues represent the signal. Hence, as SNR decreases, the algorithm suffers from increasingly large performance degradation. On the other hand, the CN of a gene profile matrix measures the linear independency of gene expression levels regardless of which specific genes having linear dependency. Thus, the evaluation of the seed gene for a given TF may be impaired by highly correlated seeds for other TFs.

For more robust identification of gene regulatory modules including both target genes and transcription factors, we propose a Gibbs sampler based on OS of regression t-statistic. The robustness of the new statistic against experimental noise in microarray data is inherited from that the regression t-statistic maintains a reasonable statistic power when the data are relatively noisy.

In real biological applications such as breast cancer cell line study, we are fully aware of the fact that a considerable number of genes may be co-expressed through some unknown mechanism while only a small portion of them should be the true drivers of cancer progression under certain condition. By using a significance test on OS, we are able to identify the TFs that have the maximum consistency between its binding knowledge and the target genes' expression pattern. We generate different null distributions associated with different target gene pool sizes to prevent the algorithm from being biased to TFs with more candidate genes. However, we are also aware of one limitation of GibbsOS for its relying on a hard threshold to screen binding data for an initial pool of target genes. Therefore, the proposed Gibbs sampling framework is only aimed to address the problem of false-positive connections, not the one of false-negative connections.

6 CONCLUSION

We proposed a Gibbs sampler to identify condition-specific transcriptional regulatory modules based on the OS of regression t-statistic. By utilizing a Gibbs strategy, we are able to estimate the marginal distribution of the OS statistic for each transcription factor and evaluate the significance of the corresponding regulatory module. Our proposed method is aimed to extract the true target genes through the integration of both microarray gene expression data and chip-on-chip/binding motif data. The simulation experiments demonstrated the robustness of the method against noise and network structures, while we further validated its efficacy on yeast cell cycle data. Moreover, we showed that our method could successfully identify important regulatory modules related to estrogen signaling and action in breast cancer.

Funding: National Institutes of Health (NIH) [CA139246, CA149653, CA149147, CA109872, CA096483, NS29525-13A and NS29525-18A, in part]; National Institutes of Health/National Cancer Institute/Science Applications International Corporation (NIH/NCI/SAIC) [HHSN261200800001E]; Department of Defense/Congressionally Directed Medical Research Programs (DoD/CDMRP) [BC030280]

Conflict of Interest: none declared.

REFERENCES

Aguilar,H. *et al.* (2010) Biological reprogramming in acquired resistance to endocrine therapy of breast cancer. *Oncogene*, **29**, 6071–6083.

Alter,O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 10101–10106.

Angus,W.G. *et al.* (1999) Expression of CYP1A1 and CYP1B1 depends on cell-specific factors in human breast cancer cell lines: role of estrogen receptor status. *Carcinogenesis*, **20**, 947–955.

Boscolo,R. *et al.* (2005) A generalized framework for network component analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 289–301.

Brynildsen,M.P. *et al.* (2006) A Gibbs sampler for the identification of gene expression and network connectivity consistency. *Bioinformatics*, **22**, 3040–3046.

Casella,G. and George,E.I. (1992) Explaining the Gibbs sampler. *Am. Statistician*, **46**, 167–174.

Chang,C. *et al.* (2008) Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. *Bioinformatics*, **24**, 1349–1358.

Chen,G. *et al.* (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.

Chen,L. *et al.* (2010) Multilevel support vector regression analysis to identify condition-specific regulatory networks. *Bioinformatics*, **26**, 1416–1422.

Chickering,D.M. (1996) Learning Bayesian networks is NP-Complete. In Fisher,D. and Lenz,H. (eds.). *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130. Springer-Verlag.

Creighton,C.J. *et al.* (2006) Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. *Genome Biol.*, **7**, R28.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**.

Fang,F. *et al.* (2008) Quantification of PRL/Stat5 signaling with a novel pGL4-CISH reporter. *BMC Biotechnol.*, **8**, 11.

Fox,E.M. *et al.* (2008) Signal transducer and activator of transcription 5b, c-Src, and epidermal growth factor receptor signaling play integral roles in estrogen-stimulated proliferation of estrogen receptor-positive breast cancer cells. *Mol. Endocrinol.*, **22**, 1781–1796.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Han,E.H. *et al.* (2010) Prostaglandin E2 induces CYP1B1 expression via ligand-independent activation of the ERalpha pathway in human breast cancer cells. *Toxicol. Sci.*, **114**, 204–216.

He,B. *et al.* (2008) A repressive role for prohibitin in estrogen signaling. *Mol. Endocrinol.*, **22**, 344–360.

Hempel,S. *et al.* (2011) Unraveling gene regulatory networks from time-resolved gene expression data—a measures comparison study. *BMC Bioinformatics*, **12**, 292.

Hungermann,D. *et al.* (2011) Influence of whole arm loss of chromosome 16q on gene expression patterns in oestrogen receptor-positive, invasive breast cancer. *J. Pathol.*, **224**, 517–528.

Lee,S.I. and Batzoglou,S. (2003) Application of independent component analysis to microarrays. *Genome Biol.*, **4**, R76.

Liao,J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems, *Proc. Natl. Acad. Sci. USA.*, **100**, 15522–15527.

Martin,L.A. *et al.* (2003) Enhanced estrogen receptor (ER) alpha, ERBB2, and MAPK signal transduction pathways operate during the adaptation of MCF-7 cells to long term estrogen deprivation. *J. Biol. Chem.*, **278**, 30458–30468.

Masamura,S. *et al.* (1995) Estrogen deprivation causes estradiol hypersensitivity in human breast cancer cells. *J. Clin. Endocrinol. Metab.*, **80**, 2918–2925.

Montgomery,D.C. *et al.* (2006) *Introduction to Linear Regression Analysis*. Wiley Interscience, New York.

Pratt,M.A. *et al.* (2003) Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence. *Mol. Cell. Biol.*, **23**, 6887–6900.

Riggins,R.B. *et al.* (2006) Physical and functional interactions between Cas and c-Src induce tamoxifen resistance of breast cancer cells through pathways involving epidermal growth factor receptor and signal transducer and activator of transcription 5b. *Cancer Res.*, **66**, 7007–7015.

Riggins,R.B. *et al.* (2005) The nuclear factor kappa B inhibitor parthenolide restores ICI 182,780 (Faslodex; fulvestrant)-induced apoptosis in antiestrogen-resistant breast cancer cells. *Mol. Cancer Ther.*, **4**, 33–41.

Shmulevich,I. *et al.* (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, **18**, 261–274.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.

Tibshirani,R. (2007) Outlier sums for differential gene expression analysis. *Biostatistics*, **8**, 2–8.

Tran,L.M. *et al.* (2005) gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab. Eng.*, **7**, 128–141.

Van den Bulcke,T. *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7**, 43.

Yamashita,H. *et al.* (2003) Naturally occurring dominant-negative Stat5 suppresses transcriptional activity of estrogen receptors and induces apoptosis in T47D breast cancer cells. *Oncogene*, **22**, 1638–1652.