Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 538919, 18 pages
doi:10.1155/2010/538919

Research Article

Uncovering Transcriptional Regulatory Networks by Sparse Bayesian Factor Model

Jia Meng,[1] Jianqiu (Michelle) Zhang,[1] Yuan (Alan) Qi,[2] Yidong Chen,[3,4] and Yufei Huang[1,3,4]

[1] *Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249-0669, USA*
[2] *Departments of Computer Science and Statistics, Purdue University, West Lafayette, IN 47907, USA*
[3] *Department of Epidemiology and Biostatistics, UT Health Science Center at San Antonio, San Antonio, TX 78229, USA*
[4] *Greehey Children's Cancer Research Institute, UT Health Science Center at San Antonio, San Antonio, TX 78229, USA*

Correspondence should be addressed to Yufei Huang, yufei.huang@utsa.edu

Received 2 April 2010; Accepted 11 June 2010

Academic Editor: Ulisses Braga-Neto

The problem of uncovering transcriptional regulation by transcription factors (TFs) based on microarray data is considered. A novel Bayesian sparse correlated rectified factor model (BSCRFM) is proposed that models the unknown TF protein level activity, the correlated regulations between TFs, and the sparse nature of TF-regulated genes. The model admits prior knowledge from existing database regarding TF-regulated target genes based on a sparse prior and through a developed Gibbs sampling algorithm, a context-specific transcriptional regulatory network specific to the experimental condition of the microarray data can be obtained. The proposed model and the Gibbs sampling algorithm were evaluated on the simulated systems, and results demonstrated the validity and effectiveness of the proposed approach. The proposed model was then applied to the breast cancer microarray data of patients with Estrogen Receptor positive (ER^+) status and Estrogen Receptor negative (ER^-) status, respectively.

1. Introduction

Response of cells to changing endogenous or exogenous conditions is governed by intricate networks of gene regulations including those by, most notably, transcription factors (TFs) [1]. Understanding how transcription regulatory network (TRN) defines cellular states and eventually phenotypes is a major challenge facing systems biologists.

Computational reconstruction of gene regulation and phenotype prediction based on microarray profiles is a current research focus in computational systems biology [2–7]. Many models have been proposed to infer the transcriptional regulation by TFs including, mostly notably, ordinary differential equations, (probabilistic) Boolean networks, Bayesian networks, information theory, and association models. Ideally, TF protein activity is needed for exact modeling but it is usually difficult to obtain. Currently, due to low protein coverage and poor quantification accuracy of high throughput technologies including protein array and liquid chromatography-mass spectrometry (LC-MS), TF protein abundance measurements are hardly available. As a compromise, most of aforementioned models conveniently yet inappropriately assume the TF's mRNA expression as its protein activity. Given the fact that gene mRNA expression and its protein abundance are poorly correlated, these models cannot accurately model the transcriptional *cis*-regulation and reveal at the best TF *trans*-regulation. In contrast, work based on factor models [8–12] points to a natural and promising direction for TF *cis*-regulation modeling, where TF activities is directly modeled as the unknown, latent factors, and microarray gene expression is modeled as a linear combination of unknown TF abundance, where the loading matrix in this FA model indicates the strength and the type (up- or downregulation) of regulation. However, due to distinct features of TRNs, conventional FA model is not readily applicable. First, since many TFs can share the same protein complex, regulate each, or get involved in the same biological process, the factors should be correlated; while in the existing FA models, factors are typically assumed independent, which, although true in many applications, is not a realistic assumption for TRNs. Secondly, since a TF only regulates a small subset of genes,

the loading matrix should be sparse. While with constructions of TF databases, such as TRANSFAC [13], the knowledge of TF-regulated genes becomes more complete and increasingly available and should be included in the model. The inclusion of prior for sparsity naturally calls for a Bayesian solution. As an added advantage, having this prior knowledge actually resolves the factor order ambiguity of the conventional factor analysis. Thirdly, as suggested in [14–16], the abundance of genes (or TFs) are naturally nonnegative, and also a non-Gaussian factor model should be in place.

In a response to meet these requirements of TRNs, we proposed here a novel Bayesian sparse correlated rectified factor model (BSCRFM). Different from conventional factor analysis models, BSCRFM consists of a sparse loading matrix and a set of correlated nonnegative factors. The sparsity of the loading matrix is constrained by a sparse prior [17] that directly reflects our existing knowledge of TF regulation that is, if a gene is known to be regulated by a TF, then the prior probability that this regulation exists is high, or otherwise, very low due to the generic sparsity nature of the loading matrix. Since TFs can regulate each other, share the same protein complex, or get involved in the same biological process, the factors in this BSCRFM model are considered to be correlated. To model the correlation between factors, a Dirichlet process mixture (DPM) prior [18] was placed on the factors. DPM imposes a natural nonparametric [19] clustering effect on TFs, which, enables automatic determination of the optimal number of clusters. Moreover, since the activities of TFs are nonnegative, they are assumed to follow a (nonnegative) rectified Gaussian distribution [20]. A Gibbs sampling solution is proposed to effectively infer all the relevant variables.

The proposed factor model is different from nonnegative matrix factorization (NMF) [14, 16, 21, 22], which has been reported to be a powerful tool for gene expression data. NMF enforces the constraint that both the loading matrix and the factor matrix must be nonnegative, that is, all elements must be equal to or greater than zero; however, in our method, only the factor matrix is constrained to be nonnegative, and the elements of loading matrix can be either positive or negative, which corresponds to up- or downregulations, respectively.

2. Bayesian Sparse Factor Modeling of Transcription Regulation

Let $\mathbf{y}_n \in \mathcal{R}^{G\times 1}$ for $n = 1,\ldots,N$ represent the nth microarray mRNA expression profile of G genes under a specific context. In practice, microarray data $\mathbf{y}_n$ register the log 2-scaled (fold change of) the expression gene levels under the context of interest relative background expression levels obtained often as the average expression levels among a variety of contexts such as different cell lines and tumors [23, 24]. We assume that the log-scaled expression level $\mathbf{y}_n$ is due to the linear combination of scaled TF protein expressions, or activities and modeled by the following factor model:

$$\mathbf{y}_n = \mathbf{A}\mathbf{x}_n + \mathbf{e}_n, \tag{1}$$

where

$\mathbf{x}_n$ the nth sample vector of the scaled activities of L TFs of interest. Particulary, the nonnegativity of $\mathbf{x}_n$ is modeled by applying the componentwise rectification (or cut) function *cut* to a vector pseudo factors $\mathbf{s}_n$ such that the lth element of $\mathbf{x}_n$ is expressed as

$$x_{l,n} = \mathrm{cut}(s_{l,n}) = \max(s_{l,n}, 0). \tag{2}$$

Since the TFs may share the same protein complex, regulate each, or get involved in the same biological process, the activities of TFs should be correlated. Therefore, pseudofactors $\mathbf{s}_n$ are modeled by a Dirichlet Process Mixture (DPM) of the Gaussian distributions as

$$\begin{aligned} s_{l,n} &\sim \mathcal{N}\left(\mu_{l,n}, \sigma_{l,n}^2\right), \qquad \left(\mu_{l,n}, \sigma_{l,n}^2\right) \sim G, \\ G &\sim \mathrm{DP}(\alpha, \mathrm{NIG}(\mu_0, \kappa_0, \alpha_0, \beta_0)), \end{aligned} \tag{3}$$

where, $\mathcal{N}(\mu_{l,n}, \sigma_{l,n}^2)$ represents the Gaussian distribution with mean $\mu_{l,n}$ and variance $\sigma_{l,n}^2$, DP denotes the Dirichlet process, and NIG is short for the conjugate normal-inverse-gamma (NIG) distribution. This DPM model implies a clustering effect on $\mathbf{s}_n$ such that

$$s_{l,n} \mid \gamma_l, \mu_{\gamma_l,n}, \sigma_{\gamma_l,n}^2 \sim \mathcal{N}\left(\mu_{\gamma_l,n}, \sigma_{\gamma_l,n}^2\right), \tag{4}$$

$$\theta_{\gamma_l,n} \sim \mathrm{NIG}(\lambda_0), \qquad \gamma_l \sim \mathrm{GEM}(\alpha), \tag{5}$$

where $\theta_{.n} = \{\mu_{.n}, \sigma_{.n}^2\}$, $\lambda_0 = \{\mu_0, \kappa_0, \alpha_0, \beta_0\}$, $\gamma_l \in \mathbb{Z}$ represents the cluster label of the lth factor and is governed by a discrete GEM distribution [18], which defines the stick breaking process with parameter α; this implies that the elements of $\mathbf{s}_n$ are correlated. Based on (2) and (4), we have

$$x_{l,n} \mid \gamma_l, \theta_{\gamma_l,n} \sim \mathcal{N}^R\left(\mu_{\gamma_l,n}, \sigma_{\gamma_l,n}^2\right), \tag{6}$$

where, $\mathcal{N}^R$ denotes the rectified Gaussian distribution [20]. Since $\theta_{\gamma_l,n}$ and γ_l are still defined in (5) by the DP, $\mathbf{x}_n$ is hence modeled by the DPM of the rectified Gaussian distributions and the elements of $\mathbf{x}_n$ are accordingly correlated. In contrast to the conventional mixture model, the DPM model enables the number of clusters to be learnt adaptively from the data instead of being predefined.

$\mathbf{A}$ the $G \times L$ loading matrix, whose element $a_{g,l}$ represents the regulatory coefficient of the gth gene by the lth TF. Since a TF is known to regulate only small set of genes, $\mathbf{A}$ should be sparse. In our model, the elements of $\mathbf{A}$ are assumed to be independent and with the *a priori* distribution [17]

$$p\left(a_{g,l}\right) = \left(1 - \pi_{g,l}\right)\delta\left(a_{g,l}\right) + \pi_{g,l}\mathcal{N}\left(a_{g,l} \mid 0, \sigma_{a,0}^2\right), \tag{7}$$

where $\pi_{g,l}$ is the *a priori* probability of $a_{g,l}$ to be nonzero. For instance, if a TF regulates a total of 500

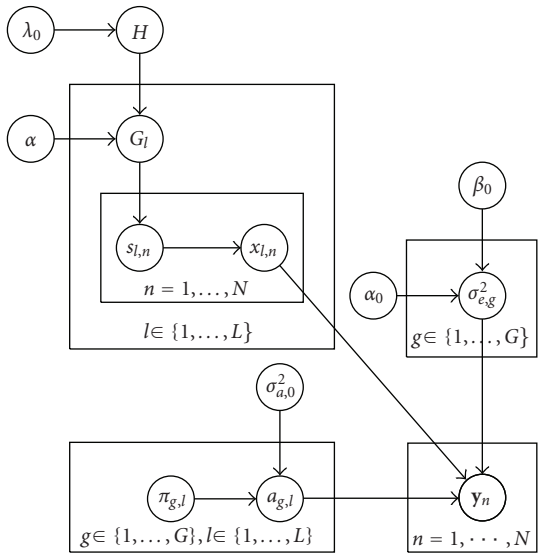


Figure 1: Graphical Model.

genes among the 20000 genes in the human genome, then $\pi_{g,l}$ is equal to

$$\pi_{g,l} = \frac{500}{20000} = 0.025. \tag{8}$$

In most cases, $\pi_{g,l}$ are likely to be smaller than 0.1. In practice, databases such as TRANSFAC [13] and DBD [25] provide information of experimentally validated or predicted target genes of TFs, and this knowledge can be incorporated in the model by setting, for instance, $\pi_{g,l} = 0.9$, if TF l is known to regulate gene g; or otherwise $\pi_{g,l} = 0.025$.

$\mathbf{e}_n$ the $G \times 1$ white Gaussian noise vector with the covariance matrix Σ defined by

$$\Sigma = \text{diag}\left(\sigma^2_{e,1}, \ldots, \sigma^2_{e,G}\right). \tag{9}$$

The overall graphical model is shown in Figure 1. The goal is to obtain the posterior distributions and hence the estimates of $\mathbf{A}$, $\mathbf{x}_n$ for all n, and Σ given the microarray profile $\mathbf{y}_n$ for all n and TF binding database. Since the analytical solution is intractable for the proposed model, we propose in the following a Gibbs sampling solution. For convenience, $\boldsymbol{\Theta}$, $\mathbf{y}_{1:N}$, and $\mathbf{x}_{1:N}$ are introduced to denote the sets of all these unknowns, all the observations, and all the factor activities, respectively. Note that the total number of factor clusters K and θ_k for all k are also unknown but treated as nuisance parameters by the proposed Bayesian solution.

3. The Proposed Gibbs Sampling Solution

The proposed BSCRFA model is high-dimensional and analytically intractable, so the authors proposed a Gibbs sampling solution. Gibbs sampling devises a Markov Chain Monte Carlo scheme to generate random samples of the unknowns from the desired but intractable posterior distributions and then approximate the (marginal) posterior distributions with these samples. The key of Gibbs sampling is to derive the conditional posterior distributions and then draw samples from them iteratively. The proposed Gibbs sampler can be summarized as follows:

Gibbs Sampling for BSCFA.

Iterate the following steps and for the tth iteration:

(1) Sample $a^{(t)}_{gl}$ for all g, l from $p(a_{g,l} \mid \boldsymbol{\Theta}_{-a_{g,l}}, \mathbf{y}_{1,N})$;

(2) for $l = 1$ to L

Sample $\gamma^{(t)}_l$ from $p(\gamma_l \mid \boldsymbol{\Theta}_{-\mathbf{x}_l,\gamma_l}, \mathbf{y}_{1:N})$; Set $K = K + 1$ if $\gamma^{(t)}_l = \bar{k}$;

Sample $\mathbf{x}^{(t)}_l$ from $p(\mathbf{x}_l \mid \boldsymbol{\Theta}_{-\mathbf{x}_l}, \mathbf{y}_{1:N})$ given $\gamma^{(t)}_l$;

Sample $s^{(t)}_{l,n}$ from $p(s_{l,n} \mid \boldsymbol{\Theta}_{-s_{l,n}}, \mathbf{y}_{1:N})$ given $\gamma^{(t)}_l$;

(3) Sample $\sigma^2_{e,g}$ for all g from $p(\sigma^2_{e,g} \mid \boldsymbol{\Theta}, \mathbf{y}_{1:N})$.

(4) Remove empty clusters and reduce K accordingly.

Note that θ_k for all k are marginalized and therefore does not need to be sampled. The algorithm iterates until the convergence of samples, which can be assessed by the scheme described in [26, chapter 11.6]. The samples after convergence will be collected to approximate the marginal posterior distributions and the estimates of the unknowns.

The required conditional distributions of the above proposed Gibbs sampling solution are detailed in Appendix A.

4. Result

4.1. Simulation

4.1.1. Test on Small Simulated System. The proposed BSCRFM algorithms was first tested on a small simulated microarray expression profiles of 40 genes and 10 samples. The genes were regulated by 6 TFs that belong to 2 clusters and the noise variance was 0.1. To ensure identifiability, each TF must regulate at least 1 gene, that is, there should be no all zero column in $\mathbf{A}$. Moreover, the sparsity of the loading matrix was set to 20%, that is, a TF regulates an average of 4 genes and a gene is regulated on average by about 1 TFs. The prior $\pi_{g,l}$s of the nonzero elements were assumed to be determined from some database. To mimic the reality that database-recorded regulations may not exist in the specific experiments and unknown regulations could also exist, the precision and the recall of the database records were introduced and both set to 0.9, from which the prior $\pi_{g,l}$ can be obtained. To diagnose the convergence of Gibbs sampler, the scheme described in [26, chapter 11.6] was adopted, where 10 parallel chains were monitored simultaneously.

Figure 2 visually depicts an example that the 10 sample chains of $x_{1,1}$ converges after around 500 iterations. The chains can be seen to converge after around 500 iterations. The estimates of $x_{1,1}$ and $a_{1,1}$ based on the samples after burn-in are summarized in Table 1. Similar results were obtained for other xs and as. Overall, the proposed algorithm

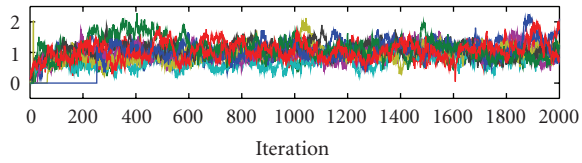


Figure 2: 10 Independent sampling chains of $x_{1,1}$.

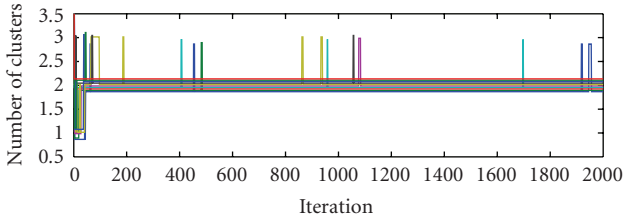


Figure 3: Nonparametric learning of number of clusters.

Table 1: Estimation of parameters $x_{1,1}$ and $a_{1,1}$.

variable	true	mean	median	mode	97.5%	2.5%	variance
$x_{1,1}$	1.08	1.05	1.04	0.97	1.61	0.55	0.07
$a_{1,1}$	0	0.0007	0	0	0	0	0.0005

can successfully recover the loading matrix and factor activities under the given settings.

Figure 3 also shows the number of clusters at each iterations for the 10 chains, which were learned according to the DPM adaptively. As mentioned before, the TFs embedded fall into 2 clusters. It can be seen from Figure 3 that the proposed BSCRFM approach can learn the number of clusters automatically by generating new clusters and eliminating actually nonexisting cluster. After 500 iteration, the chains stay at 2 clusters most of time. In order to systematically evaluate the clustering result in the following tests, a Van Rijsbergen's F metric [27] that combines the BCubed precision and recall [28] was implemented as suggested in [29].

More specifically, let $L(e)$ and $C(e)$ be the category and the cluster of an item e. Then, the correctness of the relation between e and e' is defined by

$$\text{Correctness}(e, e') = \begin{cases} 1, & \text{iff } L(e) = L(e') \longleftrightarrow C(e) = C(e'), \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

That is, two items are correctly related when they share the same cluster. Moreover, the BCubed precision and recall are formally defined as

$$\begin{aligned} &\text{Precision BCubed} \\ &\quad = \text{Avg}_e\Big[\text{Avg}_{e' \cdot C(e)=C(e')}[\text{Correctness}(e, e')]\Big], \\ &\text{Recall BCubed} \\ &\quad = \text{Avg}_e\Big[\text{Avg}_{e' \cdot L(e)=L(e')}[\text{Correctness}(e, e')]\Big], \end{aligned} \tag{11}$$

These two metrics can be further combined using Van Rijsbergen's F metrics

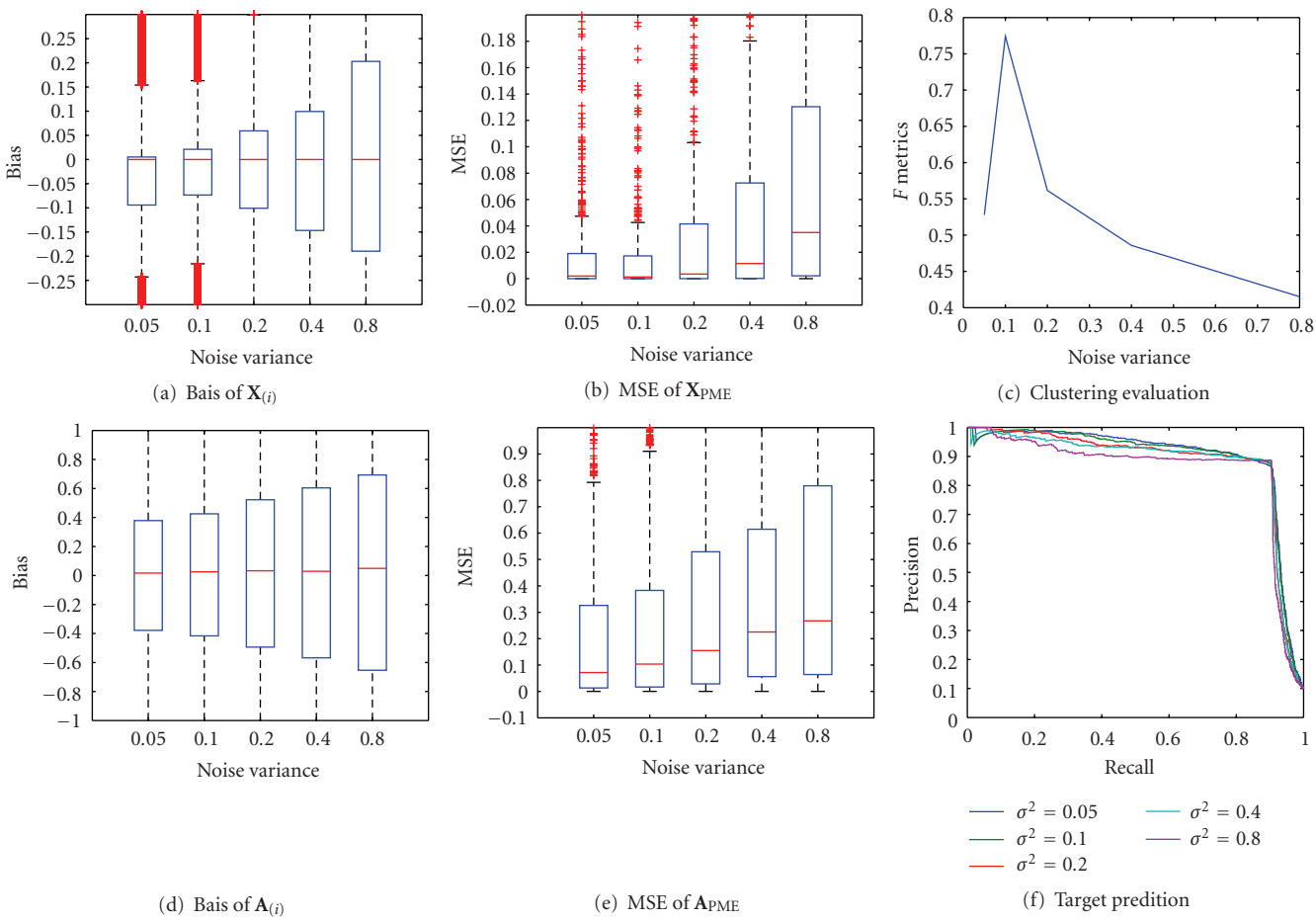
$$F(R, P) = \frac{1}{0.5/P + (1 - 0.5)/R} = \frac{2RP}{R + P}. \tag{12}$$

The F metrics will satisfy all the 4 formal constraints defined in [29], including cluster homogeneity, cluster completeness, rag bag, and cluster size versus quantity. We will use the F metrics to evaluate the clustering result in the following tests.

4.1.2. Test on Larger Simulated System. The proposed BSCRFM model was then tested on a larger simulated system, in which the microarray data consists of the expression profiles of 250 genes with 10 samples, which are regulated by 20 TFs that fall into 3 clusters. The sparsity of loading matrix was 10%, which means on average each gene is regulated by 2 TFs, and each TF regulates 25 genes. The precision and recall of the prior knowledge were still set equal to 0.9 each, indicating again that the recorded regulations may not exist in the experiment, and the unknown regulations could exist. Since this is a relatively large data set involving sampling of many variables, instead of examining convergence based on [26, chapter 11.6], we adopted a more practical strategy by running a single MCMC chain for 10000 iterations with a burn-in period of 2000 iterations [30].

In the first experiment, we tested the impact of noise on the performance of the algorithm, and the result is shown in Figure 4. It can be seen from the Figure that as noise increases, the bias of the minimum mean square estimates (MMSE) of $\mathbf{X}$ increases (Figure 4(a)), the mean squared error (MSE) of the MMSE of $\mathbf{X}$ also increases (Figure 4(b)), and the clustering performance worsens (Figure 4(c)). In general, the performance increases as the noise decreases. However, due to high-dimensionality of the proposed model, the posterior distribution is of multiple modes. When noise is very small, it is more difficult for the sample chains to travel between different modes and instead the sample chains become easily trapped in a local mode [31, 32], resulting in a poor clustering result (Figure 4(c)). Similar result can be observed for the MMSE of $\mathbf{A}$ (Figures 4(d) and 4(e)). Finally, the prediction result of the nonzero elements in $\mathbf{A}$ or targets were evaluated by the precision and recall curve (Figure 4(f)). Since the prior precision and recall are relatively high, the performance of target prediction is similar under all the tested noise conditions; but still, the result is slightly superior when noise is small.

In the last experiment, we tested the impact of prior knowledge. In practice, prior knowledge can be acquired from various databases, and very likely, this information may be imprecise and nonspecific, that is, recorded regulations may not happen in this experiments, and the unknown regulations could also exist. Here, we evaluated the performance of the BSCRFM when prior knowledge is incomplete and with error; the result is shown in Figures 5 and 6. It can be seen from the figures that, as the precision or recall of prior knowledge increases, the MMSE of $\mathbf{X}$ and $\mathbf{A}$, the clustering result and target prediction all improves. Noted that when the precision of prior knowledge is equal to 1,



(a) Bais of $\mathbf{X}_{(i)}$ (b) MSE of $\mathbf{X}_{\text{PME}}$ (c) Clustering evaluation

(d) Bais of $\mathbf{A}_{(i)}$ (e) MSE of $\mathbf{A}_{\text{PME}}$ (f) Target predition

Figure 4: Performance of BSCRFA when noise is different.

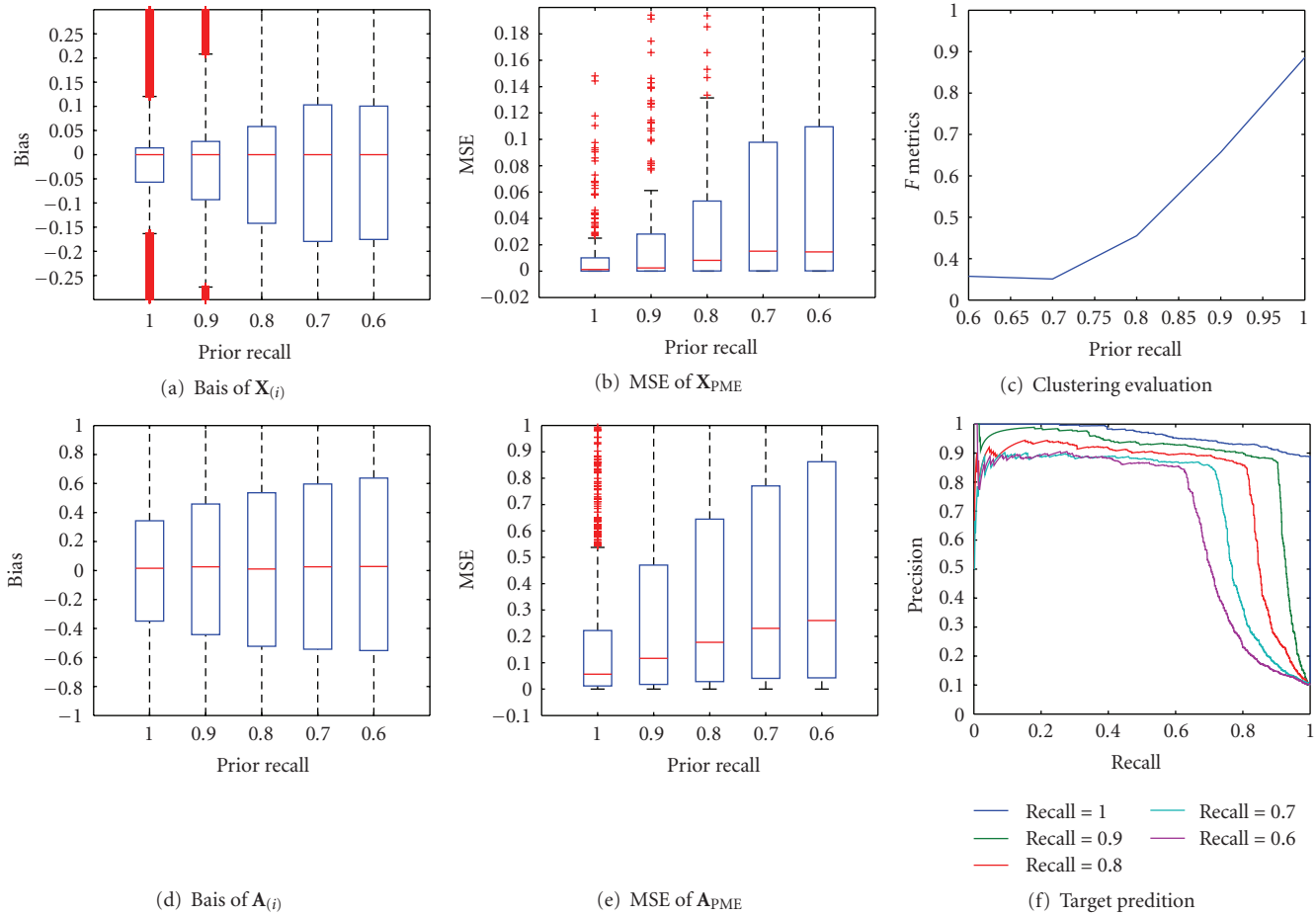
that is, all recorded regulation exist in the text experiment, and the corresponding elements in loading matrix must be nonzero. This may overwhelmingly constrain the loading matrix, resulting the MCMC chain gets trapped in a local mode (Figure 6(c)).

In the next experiment, we test the impact of the sparsity of loading matrix, and the result is shown in Figure 7. It can be seen, the more sparse the loading matrix is, the better the performance is. Since in the experimental setting each TF must regulated at least 1 gene, the more sparse the loading matrix is, a gene is regulated by less number of TFs and thus can be more easily partitioned into the contribution of less number of factors.

In this experiment, we test the impact of the number of genes, and the result is show in 8. When all the other setting are unchanged, the more genes we have, the better estimation result we can get. This is because, the algorithm relies on gene observations to estimate the factors. The more targets a TF has, the better its estimator can be. As the estimation of factor improves, the estimation of loading matrix also improves, but not as significantly Figures 8(b) and 8(d).

4.2. Test on Real Data. The proposed algorithm was then applied to the breast cancer microarray data published in [33–36]. Particularly, we applied the algorithm to two groups of samples independently, that is, 74 samples from patients of Estrogen Receptor positive (ER^+) and 68 samples of Estrogen Receptor negative (ER^-) status. All samples came with gene microarray expression, ER status, and survival time information. For the settings of the algorithm, we first manually selected a total of 11 TFs that are known to highly relevant to breast cancer (see Appendix B) and then retrieved a total of 191 regulated genes (see Appendix C) by these TFs from TRANSFAC database [13] (Release 2009.4). We also assume that TRANSFAC record has a 90% precision and 90% recall, suggesting that the known regulations may be context-specific and unknown regulations could exist. From the precision and the recall, the prior probability of the loading matrix can be determined.

The uncovered GRNs were shown in Figures 10 and 11, with each color corresponding to the predicted regulations oriented from a TF. (Please refer to Appendices B and C for the detailed annotations). It can been seen from Figure 9 that, BSCRFA recovered a total of 295 and 287 regulations respectively from ER^+ and ER^- patient samples, among which 120 are the same. 34 regulations that are recorded in prior knowledge were found in none of the two data sets, and 15 regulations that are not previously recorded



(a) Bais of $\mathbf{X}_{(i)}$ (b) MSE of $\mathbf{X}_{\text{PME}}$ (c) Clustering evaluation

(d) Bais of $\mathbf{A}_{(i)}$ (e) MSE of $\mathbf{A}_{\text{PME}}$ (f) Target predition

Figure 5: Performance of BSCRFA when recall of prior knowledge is different.

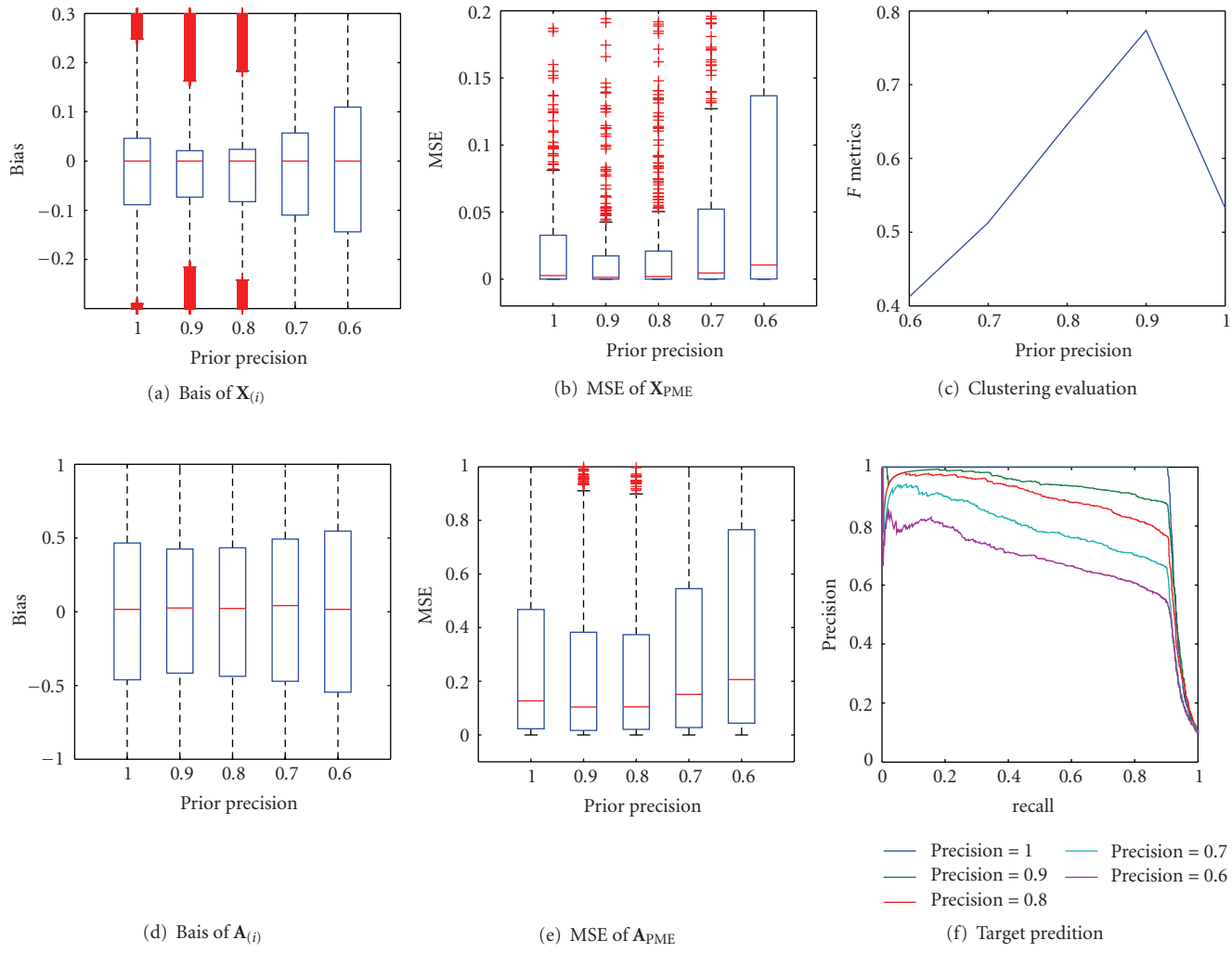
were founded in both data sets, indicating the ability of BSCRFA to recover context-specific and new regulations from microarray expression profiles.

Along with the recovered regulations, the activities of TFs are also estimated and depicted in Figures 12 and 13. In each case, three TF clusters were determined. Interestingly, in both case JUN and FOS were clustered together; this agrees with the fact that JUN and FOS belong to the same TF complex called AP1 and need to regulated collaboratively. The differential activity of each TF in ER^+ and ER^- were investigated using the t-test. The ER transcription factor is the most significantly upregulated TF among the tested 11 TFs in ER^+ samples over ER^- samples ($P = 10^{-5.62}$); also, TFs FOXA1, NFKB, FOS, JUN are shown upregulated in ER^+ samples, while P53, CREB are upregulated in ER^- samples.

For each ER condition, the patients were further classified in two 2 groups according to whether a particular TF is up- ($^+$) or down- ($^-$) regulated, and the survival statuses of each group were estimated by the Kaplan-Meier estimator; the estimated survival curves obtained and compared using the logrank test [37]. The significance levels of the logrank test (not corrected for multiple hypothesis tests) are shown in Table 2. It can be seen from Table 2 that, FOXA1 activities are significant in predicting good survival patients from the poor survival in ER^+ samples ($P = .04$); while those of FOXO3 are significant predictors in ER^- samples ($P = .04$). Their survival curves are plotted in (Figure 14). As a comparison, survival analysis was also performed on the microarray expression of FOXA1 and FOXO3 (Figure 15), and it was determined that they are not significant. These results indicate that the TF activities estimated by the proposed BSCRFM are better predictors for the survival of patients than the mRNA expression, suggesting a potentially more informative and accurate avenue to study phenotypes based on TF activities.

Table 2: Significance level of the logrank test.

TF	ER^+	ER^-	TF	ER^+	ER^-
ER	0.34	0.30	NFκB	0.48	0.28
FOXA1	**0.04**	0.38	Fos	**0.08**	0.49
GATA3	**0.08**	0.39	Jun	0.19	0.47
FOXO3	0.32	**0.04**	ATF2	0.26	0.38
MyC	0.48	0.25	CREB	0.45	0.47
P53	0.45	**0.05**			



(a) Bais of $\mathbf{X}_{(i)}$ (b) MSE of $\mathbf{X}_{\text{PME}}$ (c) Clustering evaluation

(d) Bais of $\mathbf{A}_{(i)}$ (e) MSE of $\mathbf{A}_{\text{PME}}$ (f) Target predition

FIGURE 6: Performance of BSCRFA when precision of prior knowledge is different.

5. Discussion

5.1. Features. BSCRFM is a new approach to reconstruct direct transcriptional regulation from microarray gene expression data. We discuss next a few distinct features of it.

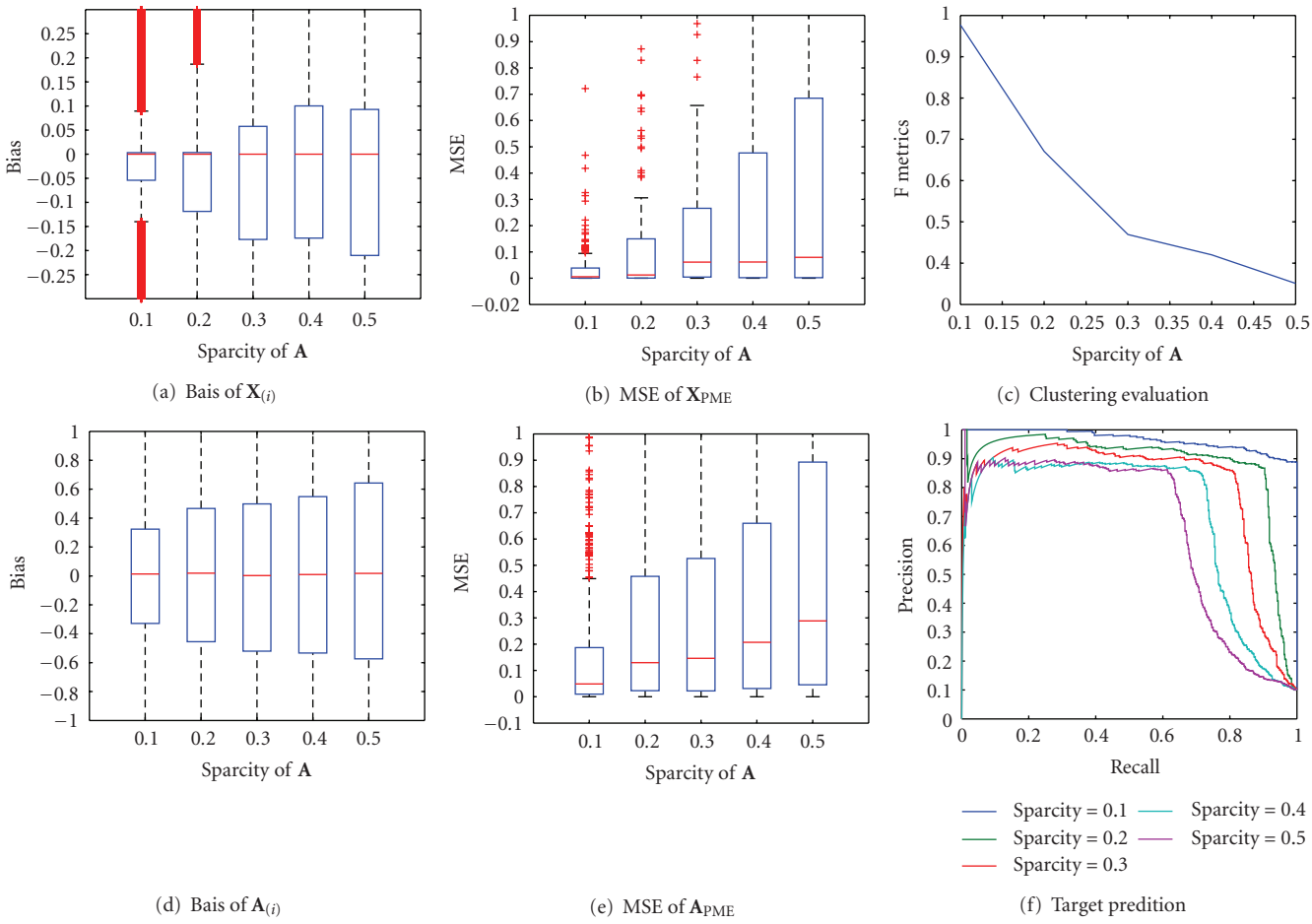
First, in accordance with the fact that a TF only regulates a number of genes in the the genome, the loading matrix of BSCRFM model is constrained by a sparse prior [17], which directly reflects our existing knowledge of the particular TF regulation that is, if the regulation exists according to prior knowledge, then the probability of the corresponding component in the loading matrix to be nonzero is large; otherwise, very small. The introduction of sparsity significantly constrains the factor model, enabling the inference of a set of correlated TF activities.

Second, since the activities of TFs cannot be negative, the factors in BSCRFM are modeled by a nonnegative rectified Gaussian distribution [20], which not only eliminated the sign ambiguity of the factor model, but also is conjugate to the likelihood function, thus greatly facilitating the computation. Noted that a rectified Gaussian distribution $\mathcal{N}^R$ is different from a truncated Gaussian $\mathcal{N}^T$ in that

$$p(x=0)=\begin{cases}0 & \text{if } x\sim\mathcal{N}^T(\mu,\sigma^2),\\ \Phi\left(-\dfrac{\mu}{\sigma}\right) & \text{if } x\sim\mathcal{N}^R(\mu,\sigma^2),\end{cases} \tag{13}$$

which indicates that the rectified Gaussian model can also describe the possible suppressed state of TFs, which cannot be modeled by the truncated Gaussian distribution. A comparison of Gaussian, rectified Gaussian and truncated Gaussian is shown as Figure 16. In our model, the nonnegativity is constrained only on the factor matrix $\mathbf{X}$; and the elements of loading matrix $\mathbf{A}$ can be either positive or negative, which models the corresponding up- or downregulation of TFs.

Third, since TFs can share the same protein complex, regulate each other, or get involved in the same biological process, the factors are assumed correlated and constrained by a Dirichlet process mixture (DPM), which can learn



(a) Bais of $\mathbf{X}_{(i)}$ (b) MSE of $\mathbf{X}_{\text{PME}}$ (c) Clustering evaluation

(d) Bais of $\mathbf{A}_{(i)}$ (e) MSE of $\mathbf{A}_{\text{PME}}$ (f) Target predition

Figure 7: Performance of BSCRFA when the sparcity of loading matrix is different.

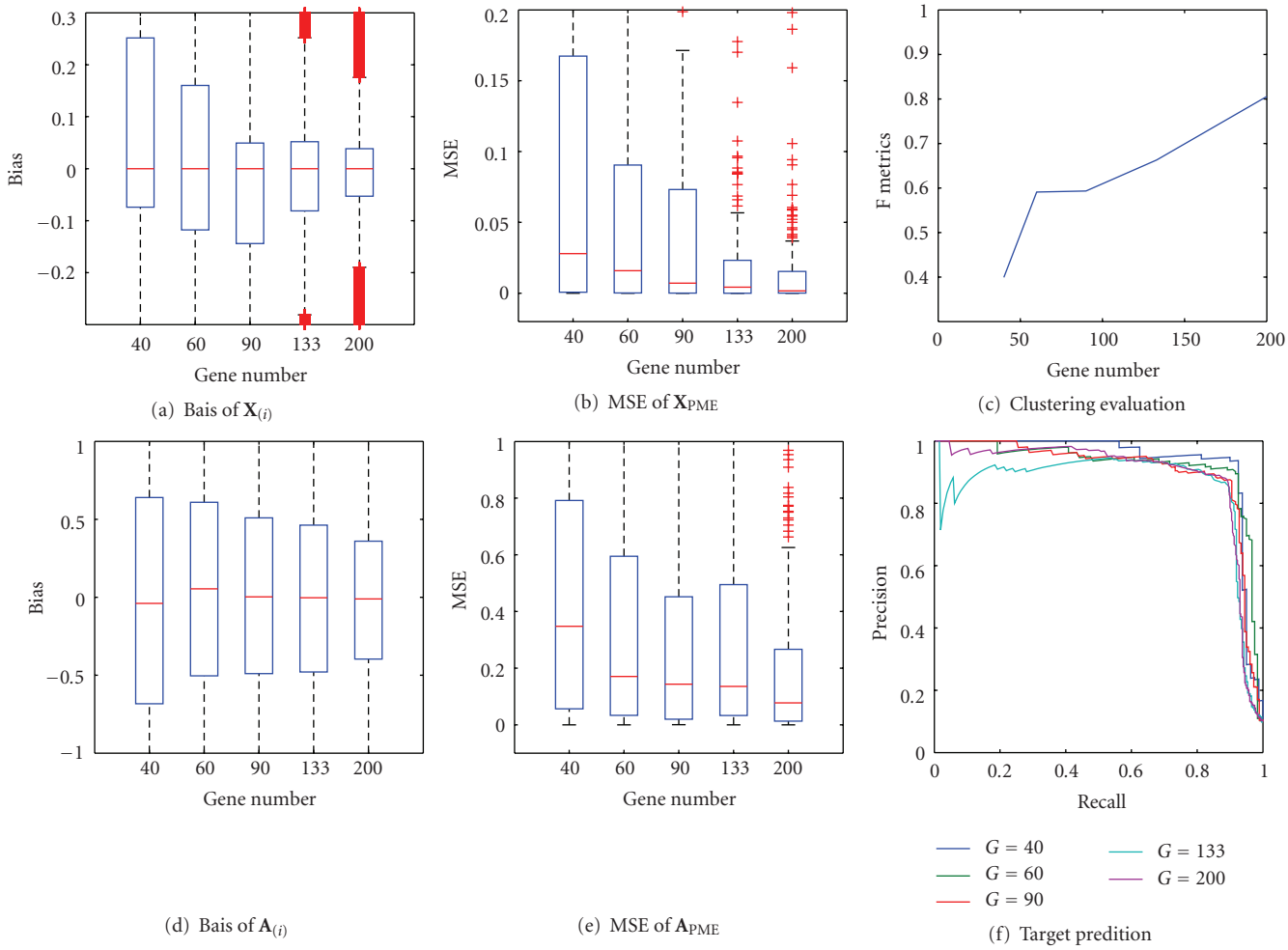
Table 3: Transcription factor list.

ID	Name	Aliases
TF1	ER	ER;ERALPHA;ESR1;ESTRADIOLRECEPTOR;ESTROGENRECEPTOR;NR3A1
TF2	FOXA1	FOXA1;HEPATOCYTENUCLEARFACTOR3ALPHA;HNF3A
TF3	GATA3	GATA3;GATABOXBINDINGFACTOR3;GATA3;NFE1C(CHICK)
TF4	FOXO3	FOXA1;HEPATOCYTENUCLEARFACTOR3ALPHA;HNF3A
TF5	MyC	CMYC;MYC;VMYCMYELOCYTOMATOSISVIRALONCOGENEHOMOLOG(AVIAN)
TF6	P53	ASP53;LFS1;NSP53;P53;P53AS;RSP53;TP53;TRP53;TUMORPROTEINP53
TF7	NFκB	NFKAPPAB;NUCLEARFACTORKAPPAB
TF8	Fos	FOSLIKEANTIGEN1;FOSL1;FRAI
TF9	Jun	AP1;JUNDPROTOONCOGENE;JUND;JUND;TRANSCRIPTIONFACTORJUND
TF10	ATF2	ACTIVATINGTRANSCRIPTIONFACTOR2;ATF2;CREBP1;HB16;TREB7
TF11	CREB	ATF47;CREB;CREB341;CREBA;CREBISOFORM1;CREB1;CREBALPHA;X2BP

automatically the optimal number of TF clusters from data. A sparse Bayesian factor model was proposed in [14], which employs a Dirichlet mixtures to model the correlation of the same factors between samples. In contrast, the proposed BSCRFA model models the correlation between different factors, which is intended to describe the correlation of activities of TFs explicitly. This correlation is a prevalent characteristics in the context of transcriptional regulation, since TFs may share the same protein complex, regulate each other, or get involved in the same biological process. Such modeling has not been investigated in the past and is a modeling focus of this paper. Modeling the additional sample correlations of the same TFs will be a focus of our future research.

Table 4: Gene list.

ID	Symbol	ID	Symbol	ID	Symbol	ID	Symbol
G1	C3	G51	LTF	G101	GADD45A	G151	PTTG1
G2	CXCR4	G52	TNF	G102	EXO1	G152	MITF
G3	MSH2	G53	TP53INP1	G103	PLAU	G153	APP
G4	GCLM	G54	CYP11B1	G104	DKK1	G154	CD1A
G5	FOS	G55	TNFRSF10B	G105	PTH	G155	SFN
G6	MT2A	G56	MMP1	G106	CDK4	G156	FAS
G7	CCNG2	G57	CD82	G107	POLB	G157	TGM1
G8	IL5	G58	HLA-DRA	G108	ID1	G158	KIR3DL1
G9	DUSP1	G59	VIP	G109	HOXA10	G159	STAT4
G10	DBH	G60	INS	G110	PENK	G160	CD8A
G11	CHEK1	G61	PTGS2	G111	EBAG9	G161	TFF1
G12	SCN3B	G62	JUN	G112	COL1A2	G162	APC
G13	ITGAX	G63	GSTP1	G113	ZNF268	G163	IL6
G14	EIF4E	G64	CCND1	G114	TNFRSF10A	G164	IFNB1
G15	TGFB2	G65	CASP1	G115	AMBP	G165	PTK2
G16	TSHB	G66	TRIM22	G116	TNFRSF10C	G166	SPP1
G17	CDC25A	G67	HBB	G117	PDK4	G167	NPPA
G18	F3	G68	MDM2	G118	CXCL3	G168	TP73
G19	IL2RA	G69	RB1	G119	MICA	G169	SLC3A2
G20	BDNF	G70	NDRG1	G120	TRA@	G170	IL1B
G21	WEE1	G71	NQO1	G121	HLA-DPB1	G171	APOB
G22	CYP11A1	G72	BRCA1	G122	TP53	G172	IL8
G23	NR4A2	G73	SERPINB5	G123	SOX9	G173	VEGFA
G24	TRH	G74	BCL2	G124	PCNA	G174	PBK
G25	CAV1	G75	BAX	G125	NFKB1	G175	TACR1
G26	MUC1	G76	CYP1B1	G126	IL2	G176	RPL10
G27	PGR	G77	TGFA	G127	CRHBP	G177	IVL
G28	GNAI2	G78	ATF2	G128	ERVWE1	G178	FCGR2A
G29	ADRB2	G79	FN1	G129	CRH	G179	MACROD1
G30	GCLC	G80	COX7A2L	G130	FANCC	G180	ERBB2
G31	OPRM1	G81	BCL2L1	G131	RFWD2	G181	CCL2
G32	EPO	G82	GSS	G132	EPHX1	G182	BBC3
G33	ACTA2	G83	TF	G133	YBX1	G183	TP63
G34	KLRC1	G84	GYPB	G134	ATF3	G184	AGER
G35	IFNG	G85	CXCL1	G135	APAF1	G185	SESN1
G36	BCL2A1	G86	CSNK1A1	G136	CYP19A1	G186	GJA1
G37	SLC9A3R1	G87	IL4	G137	CX3CL1	G187	NAT1
G38	CCL5	G88	NR3C1	G138	KRT16	G188	SELE
G39	BCAS3	G89	EGR1	G139	CGA	G189	FASLG
G40	ICAM1	G90	IRF4	G140	SFTPD	G190	HRAS
G41	PSENEN	G91	EDN1	G141	HIF1A	G191	BRCA2
G42	IER2	G92	PRL	G142	CTSD		
G43	HSD17B1	G93	IGFBP3	G143	DDB2		
G44	GNRHR	G94	CFTR	G144	TPT1		
G45	LTA	G95	EGFR	G145	IRS2		
G46	TERT	G96	MYC	G146	DDX18		
G47	OLR1	G97	CYBB	G147	CCNA2		
G48	MMP2	G98	F8	G148	IL13		
G49	APOE	G99	TSC22D3	G149	CDKN1A		
G50	ODC1	G100	LOR	G150	ESR1		



(a) Bais of $\mathbf{X}_{(i)}$ (b) MSE of $\mathbf{X}_{\text{PME}}$ (c) Clustering evaluation

(d) Bais of $\mathbf{A}_{(i)}$ (e) MSE of $\mathbf{A}_{\text{PME}}$ (f) Target predition

Figure 8: Performance of BSCRFA when the number of genes is different.

Forth, other types of data, such as ChIP-chip data [38–40] and DNA methylation data [41] can be conveniently integrated with gene expression data [42] under the proposed BSCRFM by setting a slightly different prior probabilities to the loading matrix. Integrating more data types can potentially improve the performance of the proposed method and will be our future work.

5.2. Limitations. First, this model cannot capture regulation from TFs that are not specified in the prior knowledge database. In reality, it is possible that TFs that are not specified in the prior knowledge actually regulate the gene transcription. However, it is possible to further extend the proposed factor model to capture the contribution of missing factors.

Second, relatively complete and accurate prior knowledge should be present for the approach to be implemented. Since the proposed BSCRFM model assume correlated factors, it is important to have sufficient prior knowledge to constrain the structure (zero and nonzero elements) of the loading matrix. To effectively estimate the relevant variables, relatively complete and accurate prior knowledge must be present. In the absence of such prior knowledge, for example, when studying the transcriptional network of less-studied species, the proposed method is not recommended.

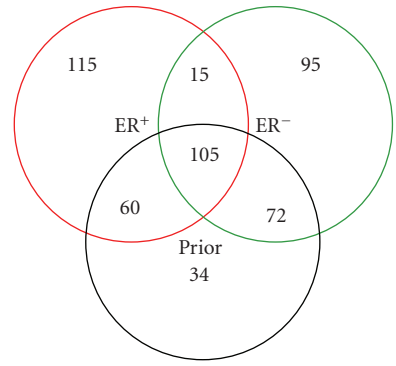


Figure 9: Common and specific recovered regulation.

Third, the algorithm may not converge in a reasonable number of iterations on a large data set, thus cannot be

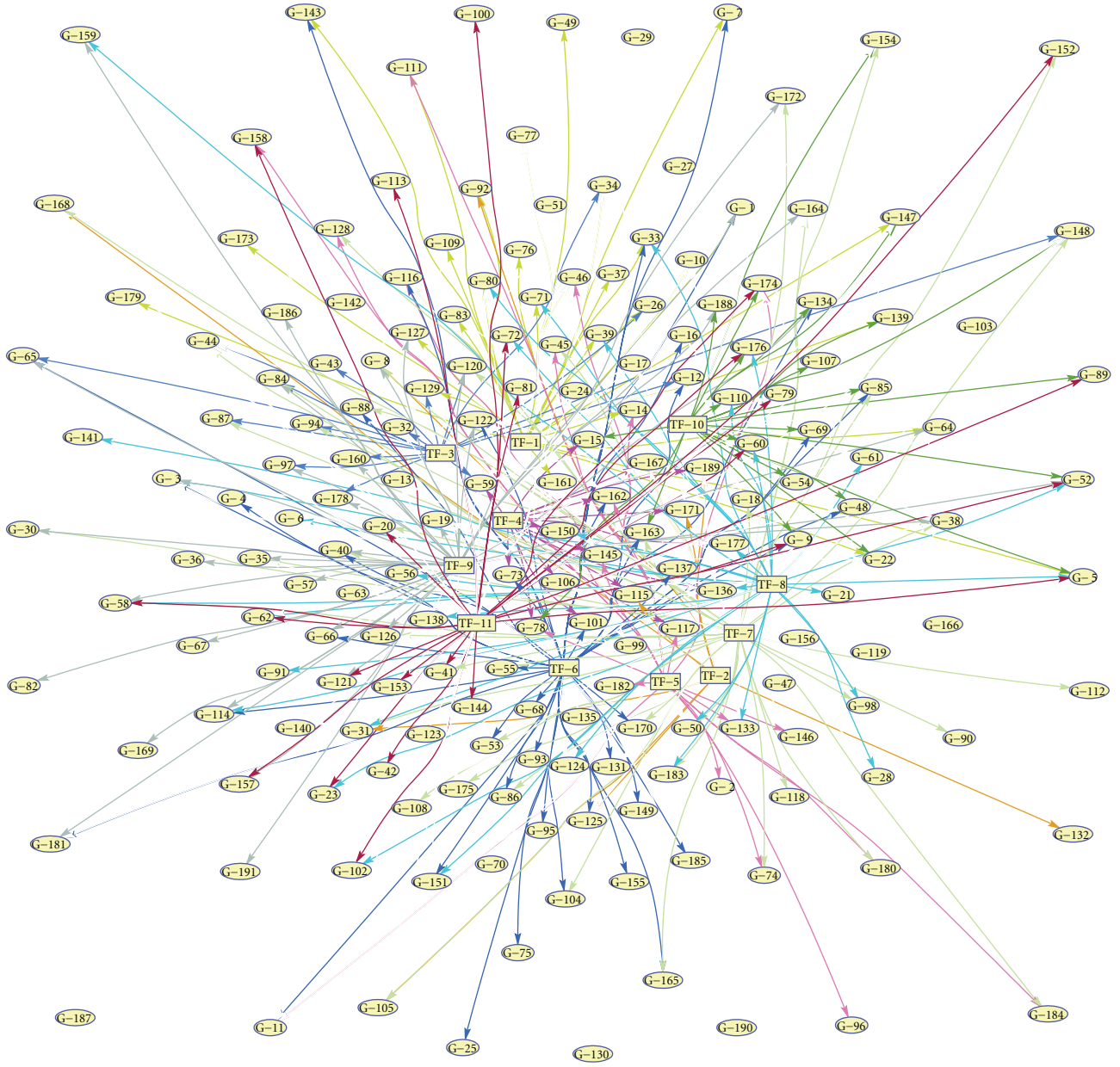


Figure 10: Transcriptional regulatory network in ER^+ samples.

applied to genome wide dataset. Because the model parameters are high-dimensional and highly correlated, the speed of convergence may significantly slow down on a large data set [43, 44]. Moreover, when parameter distribution is bimodal (or multimodal), the Gibbs sampling iterations can easily get trapped in one of the modes, thus reducing the probability of reaching convergence [31, 32]. Even when convergence can be achieved under the criteria defined in [26, chapter 11.6], the narrow mode in the distribution may still not be detected, leading to overestimation of the posterior variance [45]. Currently, the proposed model is intended for analyzing a subset of TFs, for which additional knowledge about their binding and biological relevance is available. Through integrating the prior knowledge, more informative and reliable results can be achieved. In addition, the prior knowledge also makes the interpretation of results easier. We demonstrate in Section 4, how such analysis can be carried out starting from a whole genome microarray data. With the advancement in ChIP-seq technology and increasing knowledge of TFs biological functions, the proposed model could be applied for a genome-wide study in the future.

Forth, prior knowledge may still need to be properly evaluated. If the prior knowledge is considered an estimation of the true TRN, when the precision p, recall r of prior

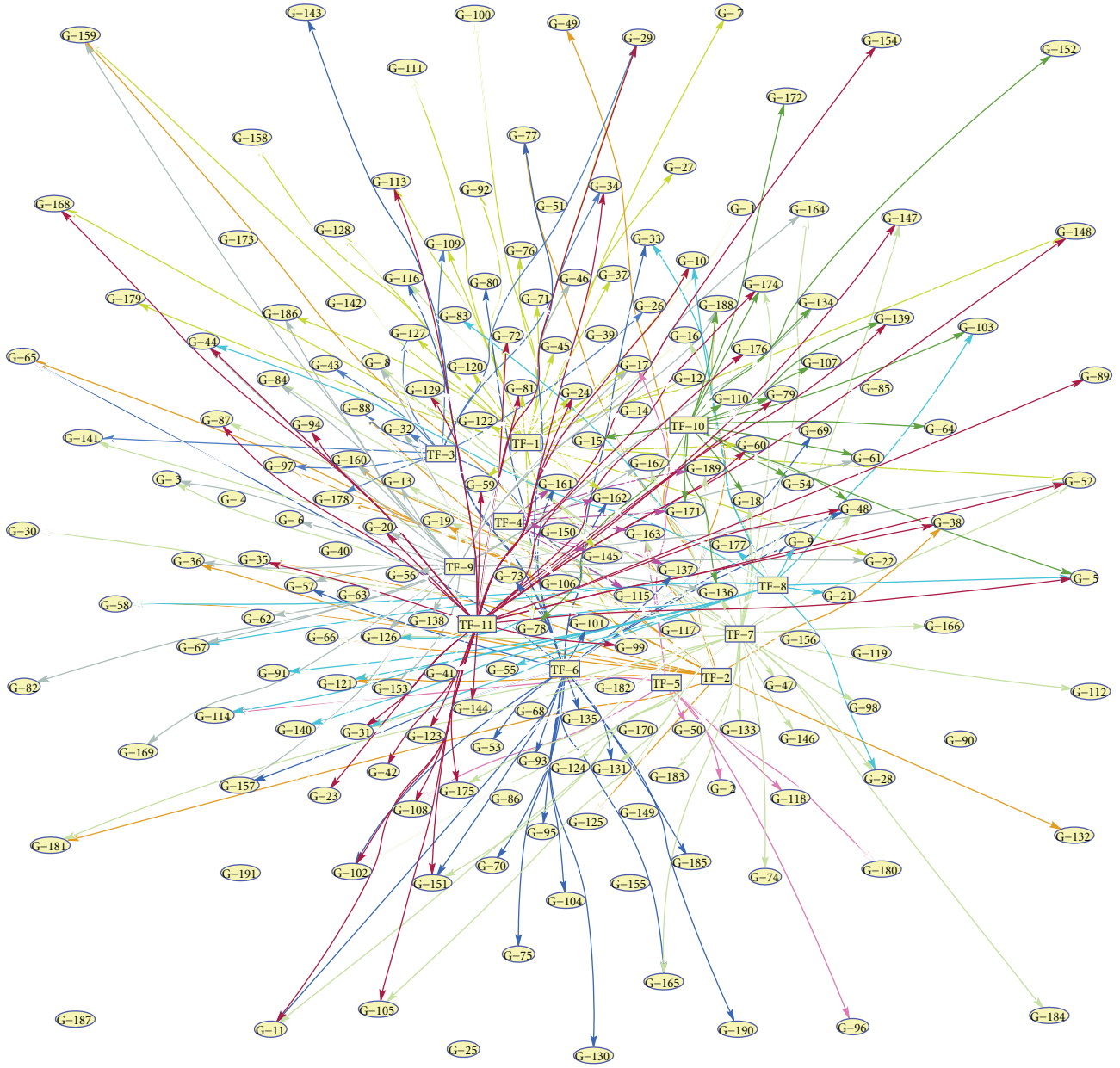


Figure 11: Transcriptional regulatory network in ER$^-$ samples.

information and the sparsity of the loading matrix s is given, the prior probability of the gth gene to be a target of the lth TF $\pi_{g,l}$ can be calculated as follows:

$$\pi_{g,l} = \begin{cases} p, & \text{recorded regulation,} \\ \dfrac{sp(1-r)}{p-sr}, & \text{not recorded regulation.} \end{cases} \tag{14}$$

However, the precision or recall of the prior knowledge database is not available. In practice, the quality of prior knowledge should be evaluated first before more reasonable prior probabilities of regulations can be assigned.

6. Conclusion

A Bayesian factor model with sparse-loading matrix and correlated nonnegative factors was proposed to unveil the latent activities of transcription factors and their targeted genes from observed gene mRNA expression profiles. By naturally incorporating the prior knowledge of TF-regulated genes, the sparsity constraint of the loading matrix, and the non-negativity constraints of TF activities, both context-dependent regulation and TF activities can be estimated. A Gibbs sampling solution was proposed. The effectiveness and validity of the model and the proposed Gibbs sampler were evaluated on simulated systems and on real data. The results demonstrated that BSCRFM provides a viable approach to

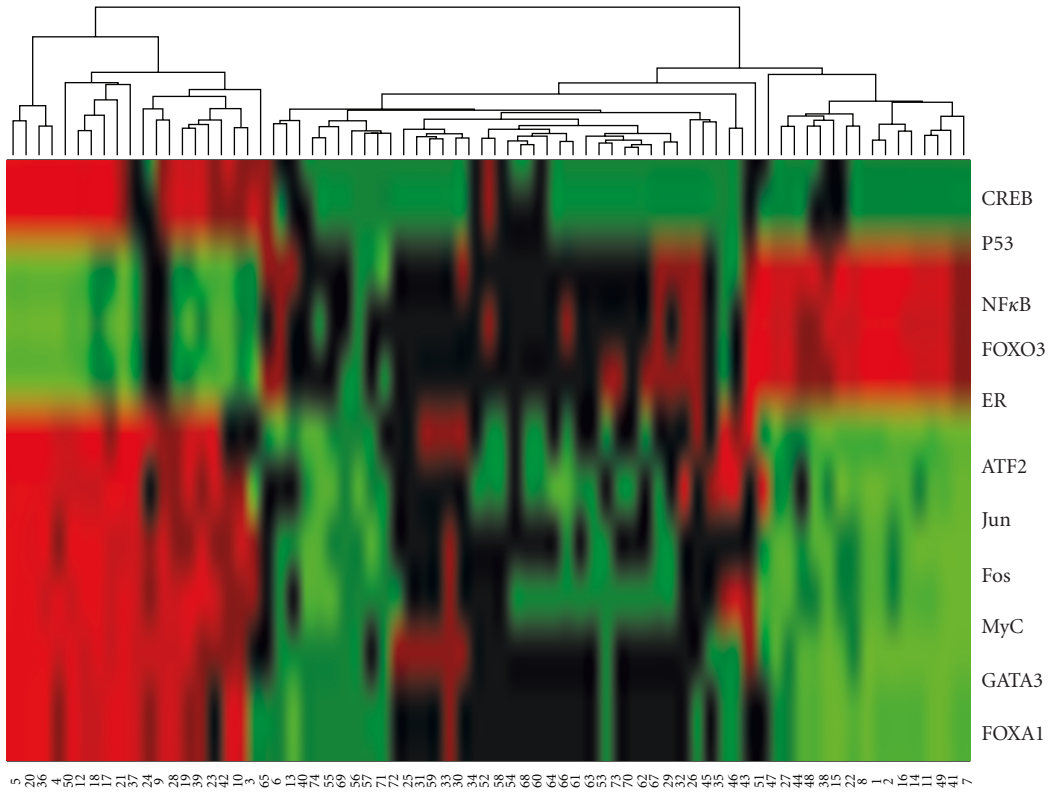


FIGURE 12: Estimated TF activities in ER^+ patients samples. The samples (columns) are arranged according to hierarchical clustering and the TFs (rows) according to the estimated clusters by the Gibbs sampling algorithm.

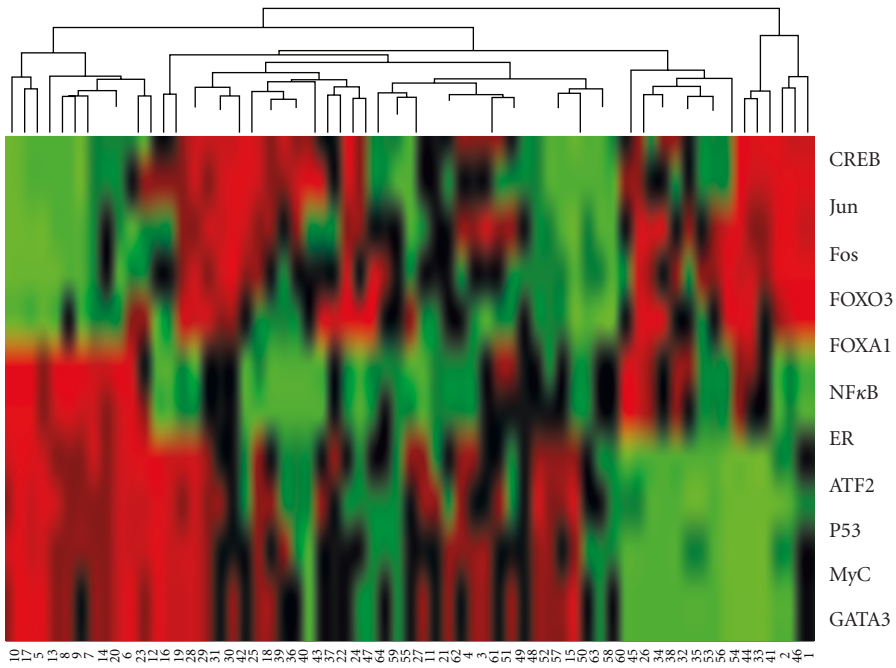
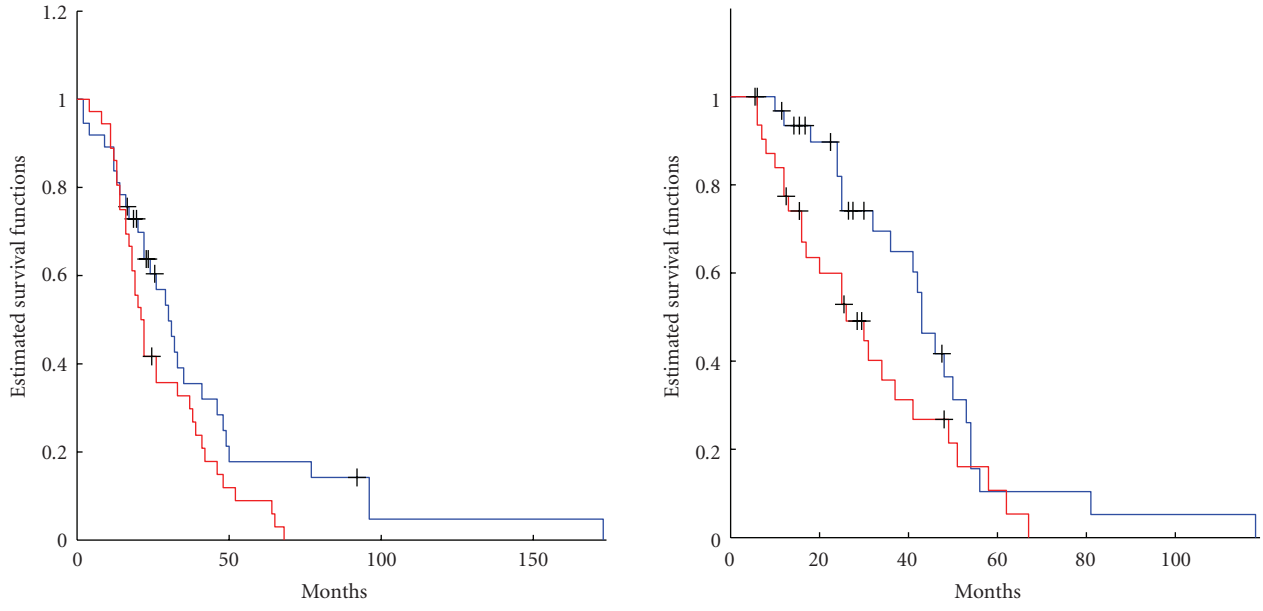


FIGURE 13: Estimated TF activities in ER^- patient samples. The samples (columns) are arranged according to hierarchical clustering and the TFs (rows) according to the estimated clusters by the Gibbs sampling algorithm.



(a) FOXA1 in ER$^+$ Patients ($P = .04$)

(b) FOXO3 in ER$^-$ Patients ($P = .03$)

Figure 14: Kaplan-Meier survival estimates for FOXA1 in ER$^+$ and FOXO3 in ER$^-$ are significantly different.

estimate TF's protein activities and studying phenotypes based on TF's protein activities could yield more informative and accurate results.

Appendix

A. Conditional Distributions of the Proposed Gibbs Sampling Solution

The required conditional distributions of the proposed Gibbs sampling solution are detailed.

A.1. $p(a_{g,l} \mid \boldsymbol{\Theta}_{-a_{g,l}}, \mathbf{y}_{1,N})$. Let $\hat{\mathbf{y}}_{gl} = [\hat{y}_{gl,1}, \ldots, \hat{y}_{gl,N}]^\top$ with $\hat{y}_{gl,n} = y_{g,n} - \sum_{i=1, i\neq l}^{L} a_{g,i} x_{i,n}$ and $\mathbf{x}_l = [x_{l,1}, \ldots, x_{l,n}]^\top$. It then follows $\hat{\mathbf{y}}_{gl} \sim \mathcal{N}(\mathbf{x}_l a_{g,l}, \sigma_{e,g}^2 \mathbf{I}_N)$ and

$$\begin{aligned} p\left(a_{g,l} \mid \boldsymbol{\Theta}_{-a_{g,l}}, \mathbf{y}_{1,N}\right) &= p\left(a_{g,l} \mid \mathbf{x}_l, \hat{\mathbf{y}}_{gl}, \sigma_{e,g}^2\right) \\ &= Z_0 p\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{x}_l, a_{g,l}, \sigma_{e,g}^2\right) p\left(a_{g,l}\right) \\ &= Z_0 \Big[\left(1 - \pi_{g,l}\right) \mathcal{N}\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{x}_l a_{g,l}, \sigma_{e,g}^2 \mathbf{I}_N\right) \delta\left(a_{gj}\right) \\ &\quad + \pi_{g,l} \mathcal{N}\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{x}_l a_{g,l}, \sigma_{e,g}^2 \mathbf{I}_N\right) \mathcal{N}\left(a_{g,l} \mid 0, \sigma_{a,0}^2\right) \Big] \\ &= \left(1 - \hat{\pi}_{g,l}\right) \delta\left(a_{g,l}\right) + \hat{\pi}_{g,l} f\left(a_{g,l}\right), \end{aligned} \tag{A.1}$$

where Z_0 is a normalizing constant, $\hat{\pi}_{g,l} = \pi_{g,l}/[(1 - \pi_{g,l})\mathrm{BF}_{01} + \pi_{g,l}]$ is the posterior probability of $a_{g,l} \neq 0$ and BF_{01} is the Bayes factor of model $a_{g,l} = 0$ versus model $a_{g,l} \neq 0$

$$\mathrm{BF}_{01} = \frac{p\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{x}_l, a_{g,l} = 0, \sigma_{e,g}^2\right)}{p\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{x}_l, a_{g,l} \neq 0, \sigma_{e,g}^2\right)} = \frac{\mathcal{N}\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{0}, \sigma_{e,g}^2 \mathbf{I}_N\right)}{\mathcal{N}\left(\hat{\mathbf{y}}_{gl} \mid \mathbf{0}, \mathbf{C}_{y,gl}\right)}, \tag{A.2}$$

with $\mathbf{C}_{y,gl} = \mathbf{x}_l \mathbf{x}_l^\top \sigma_{a,0}^2 + \sigma_{e,g}^2 \mathbf{I}_N$; $f(a_{g,l})$ is the posterior distribution for $a_{g,l} \neq 0$ and defined by

$$f\left(a_{g,l}\right) = \mathcal{N}\left(a_{g,l} \mid \hat{\mu}_{a,gl}, \hat{\sigma}_{a,gl}^2\right), \tag{A.3}$$

where, $\hat{\mu}_{a,gl} = \hat{\sigma}_{a,gl}^2 \mathbf{x}_l^\top \hat{\mathbf{y}}_{gl} / \sigma_{e,g}^2$ and $(\hat{\sigma}_{a,gl}^2)^{-1} = (\sigma_{a,0}^2)^{-1} + \mathbf{x}_l^\top \mathbf{x}_l / \sigma_{e,g}^2$; $\pi_{g,l}$ is the prior knowledge of the probability of $a_{g,l}$ to be nonzero. When $\pi_{g,l} = 0.5$, that is, a noninformative prior on sparsity is assumed, $\hat{\pi}_{g,l}$ depends only on BF_{01} and $\hat{\pi}_{g,l} < 0.5$ when $\mathrm{BF}_{01} > 1$. Since model selection based BF_{01} favors $a_{g,l} = 0$, it suggests that this Bayesian solution favors sparse model even when $\pi_{g,l} = 0.5$.

A.2. $p(\gamma_l \mid \boldsymbol{\Theta}_{-x_l,\gamma_l}, \mathbf{y}_{1:N})$. It should be noted that γ_l does not depend on $\mathbf{x}_l$ in the distribution. It is intended that samples of γ_l from this distribution are not affected by the immediate sample of $\mathbf{x}_l$, thus achieving faster convergence of the sample Markov chains. To derive this distribution, first let

$\hat{\mathbf{y}}_{l,n} = \mathbf{y}_n - \mathbf{A}\mathbf{x}_n + \mathbf{a}_l x_{l,n}$ with $\mathbf{a}_l$ being the lth column of $\mathbf{A}$ and hence $\hat{\mathbf{y}}_{l,n} \sim \mathcal{N}(\mathbf{a}_l x_{l,n}, \Sigma)$. Then,

$$\begin{aligned} p\left(\gamma_l \mid \boldsymbol{\Theta}_{-\mathbf{x}_l,\gamma_l}, \mathbf{y}_{1:N}\right) &= p\left(\gamma_l \mid \boldsymbol{\gamma}_{-l}, \hat{\mathbf{y}}_{l,1:N}\right) \\ &= \int p\left(\gamma_l, \mathbf{x}_l \mid \boldsymbol{\gamma}_{-l}, \hat{\mathbf{y}}_{l,1:N}\right) d\mathbf{x}_l \\ &= \frac{1}{Z_0} \int p(\hat{\mathbf{y}}_{l,1:N} \mid \mathbf{x}_l)\, p\left(\mathbf{x}_l, \gamma_l \mid \mathbf{x}_{-l}, \boldsymbol{\gamma}_{-l}\right) d\mathbf{x}_l \\ &= \frac{1}{Z_0}\left(\sum_{k=1}^{K} N_{-l,k} g_{l,k} \delta(\gamma_l - k) + \alpha g_{l,\bar{k}} \delta\left(\gamma_l - \bar{k}\right)\right), \end{aligned} \tag{A.4}$$

where $\bar{k}$ denotes a new cluster other than the existing K, $\mathcal{S}_{-l,k} = \{i \mid i \neq l, \gamma_i = k\}$ represents the set of the pseudo factors besides s_l that also belong to cluster k, $N_{-l,k}$ is size of $\mathcal{S}_{-l,k}$

$$\begin{aligned} Z_0 &= \sum_{k=1}^{K} N_{-l,k} g_{l,k} + \alpha g_{l,\bar{k}}, \\ g_{l,k} &= \prod_{n=1}^{N}\left(\mathcal{N}(\hat{\mathbf{y}}_{l,n} \mid \mathbf{0}, \Sigma)\Phi\left(\frac{-\hat{\mu}_{l,n}}{\hat{\sigma}_{l,n}}\right)\right. \\ &\qquad \left. + \mathcal{N}\left(\hat{\mathbf{y}}_{l,n} \mid \boldsymbol{\mu}_{\hat{\mathbf{y}}_{l,n}}, \Sigma_{\hat{\mathbf{y}}_{l,n}}\right)\Phi\left(\frac{\hat{\mu}_{x_{l,n}}}{\hat{\sigma}_{x_{l,n}}}\right)\right), \end{aligned} \tag{A.5}$$

with

$$\begin{aligned} \boldsymbol{\mu}_{\hat{\mathbf{y}}_{l,n}} &= \mathbf{a}_l \hat{\mu}_{l,n}, \\ \Sigma_{\hat{\mathbf{y}}_{l,n}} &= \mathbf{a}_l \mathbf{a}_l^{\top} \hat{\sigma}_{l,n}^2 + \Sigma, \\ \mu_{x_{l,n}} &= \hat{\mu}_{l,n} + \hat{\sigma}_{l,n}^2 \mathbf{a}_l^{\top}\left(\mathbf{a}_l \mathbf{a}_l^{\top} \hat{\sigma}_{l,n}^2 + \Sigma\right)^{-1}(\hat{\mathbf{y}}_{l,n} - \mathbf{a}_l \hat{\mu}_{l,n}), \\ \sigma_{x_{l,n}}^2 &= \hat{\sigma}_{l,n}^2 - \hat{\sigma}_{l,n}^2 \mathbf{a}_l^{\top}\left(\mathbf{a}_l \mathbf{a}_l^{\top} \hat{\sigma}_{l,n}^2 + \Sigma\right)^{-1} \mathbf{a}_l \hat{\sigma}_{l,n}^2, \\ \hat{\mu}_{l,n} &= \frac{\mu_0 \kappa_0 + \sum_{i \in \mathcal{S}_{-l,k}} s_{i,n}}{\overline{\kappa}}, \\ \overline{\kappa} &= \kappa_0 + N_{-l,k}, \\ \hat{\sigma}_{l,n}^2 &= \frac{(\overline{\kappa}+1)\overline{\beta}}{\overline{\kappa}(\alpha_0 + N_{-l,k}/2 - 1)}, \\ \overline{\beta} &= \beta_0 + \frac{\sum_{i \in \mathcal{S}_{-l,k}} s_{i,n}^2 + \kappa_0 \mu_0^2 - \overline{\kappa}\hat{\mu}_{l,n}^2}{2}. \end{aligned} \tag{A.6}$$

Noted that for a new cluster, $k = \bar{k}$, $\mathcal{S}_{-l,k} = \phi$ and $N_{-l,k} = 0$, and $g_{l,\bar{k}}$ can be derived from $g_{l,k}$ for $k = \bar{k}$.

A.3. $p(\mathbf{x}_l \mid \boldsymbol{\Theta}_{-x_l}, \mathbf{y}_{1:N})$. This distribution can be expressed as

$$\begin{aligned} p(\mathbf{x}_l \mid \boldsymbol{\Theta}_{-\mathbf{x}_l}, \mathbf{y}_{1:N}) &= p\left(\mathbf{x}_l \mid \boldsymbol{\gamma}_{-l}, \mathbf{s}_{-l}, \mathbf{y}_{1:N}, \Sigma\right) \\ &= Z_0 \prod_{n=1}^{N} p(\mathbf{y}_n \mid x_{l,n})\, p\left(x_{l,n} \mid \mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}\right) \\ &= Z_0 \prod_{n=1}^{N} p(\hat{\mathbf{y}}_{l,n} \mid x_{l,n}) \times \left(\sum_{k=1}^{K} p\left(x_{l,n} \mid \mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l = k\right)\right. \\ &\qquad \left. + p\left(x_{l,n} \mid \mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l = \bar{k}\right)\right) \\ &= Z_0 \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{y}}_{l,n} \mid \mathbf{a}_l x_{l,n}, \Sigma) \\ &\quad \times \left(\sum_{k=1}^{K}\left[p(x_{l,n} \mid s_{i,n} \forall i \in \mathcal{S}_{-l,k}, \gamma_l)\delta(\gamma_l - k)\right]\right. \\ &\qquad \left. + p(x_{l,n})\delta\left(\gamma_l - \bar{k}\right)\right) \\ &= \prod_{n=1}^{N} \hat{\pi}_{l,n}\delta(x_{l,n}) + (1 - \hat{\pi}_{l,n}) \frac{\mathcal{N}\left(x_{l,n} \mid \mu_{x_{l,n}}, \sigma_{x_{l,n}}^2\right) U(x_{l,n})}{\Phi\left(\mu_{x_{l,n}}/\sigma_{x_{l,n}}\right)}, \end{aligned} \tag{A.7}$$

where

$$\hat{\pi}_{l,n} = \frac{\mathcal{N}(\hat{\mathbf{y}}_{l,n} \mid \mathbf{0}, \Sigma)\Phi(-\hat{\mu}_{l,n}/\hat{\sigma}_{l,n})}{\mathcal{N}(\hat{\mathbf{y}}_{l,n} \mid \mathbf{0}, \Sigma)\Phi(-\hat{\mu}_{l,n}/\hat{\sigma}_{l,n}) + \mathcal{N}\left(\hat{\mathbf{y}}_{l,n} \mid \boldsymbol{\mu}_{\hat{\mathbf{y}}_{l,n}}, \Sigma_{\hat{\mathbf{y}}_{l,n}}\right)\Phi\left(\mu_{x_{l,n}}/\sigma_{x_{l,n}}\right)}. \tag{A.8}$$

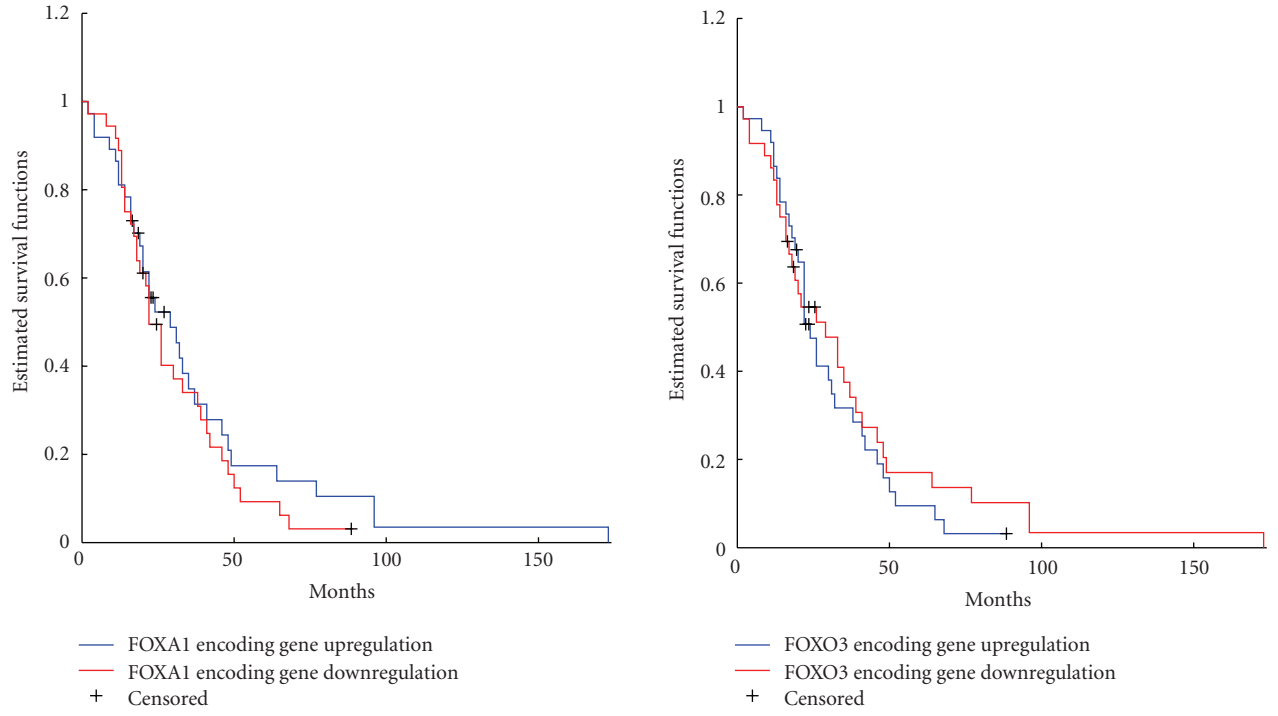
A.4. $p(s_{l,n} \mid \boldsymbol{\Theta}_{-s_{l,n}}, \mathbf{y}_{1:N})$. According to the graphical model, given $x_{l,n}$, the conditional distribution of $s_{l,n}$ does not depend on $\mathbf{y}_{1:N}$; therefore this conditional distribution can be expressed as

$$\begin{aligned} p\left(s_{l,n} \mid \boldsymbol{\Theta}_{-s_{l,n}}, \mathbf{y}_{1:N}\right) &= p\left(s_{l,n} \mid x_{l,n}, \mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l\right) \\ &\propto p(x_{l,n} \mid s_{l,n})\, p(s_{l,n} \mid \mathbf{s}_{-l,n}, \boldsymbol{\gamma}). \end{aligned} \tag{A.9}$$

To obtain the predictive density $p(s_{l,n} \mid \mathbf{s}_{-l,n},)$, first notice, based on the DPM of Gaussian model of $s_{l,n}$ that the joint conditional distribution of $s_{l,n}$, and γ_l is

$$p(s_{l,n}, \gamma_l \mid \mathbf{s}_{-l,n}, \gamma_{-l}) = \frac{\sum_{k=1}^{K} N_{-l,k}\, p(s_{l,n} \mid s_{i,n} \forall i \in \mathcal{S}_{-l,k}, \gamma_l)\delta(\gamma_l - k) + \alpha p(s_{l,n})\delta\left(\gamma_l - \bar{k}\right)}{(\alpha + L - 1)} \tag{A.10}$$

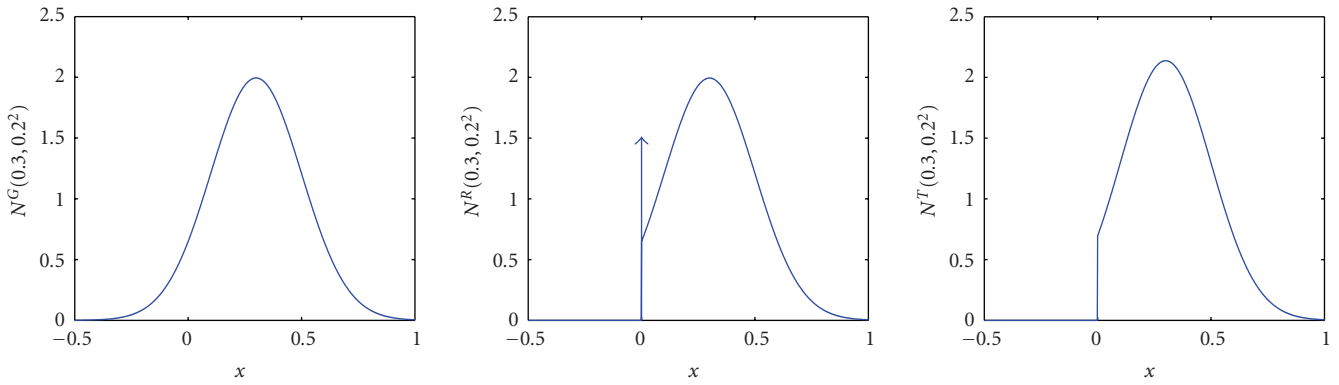
The distribution (A.10) demonstrates the correlation between pseudo factors—$s_{l,n}$ depends only on other pseudo



(a) Encoding gene of FOXA1 in ER$^+$ ($P = .26$)

(b) Encoding Gene of FOXO3 in ER$^-$ ($P = .32$)

FIGURE 15: Kaplan-meier survival estimates for the encoding gene of FOXA1 in ER$^+$ and the encoding gene of FOXO3 in ER$^-$.



(a) Gaussian

(b) Rectified Gaussian

(c) Truncated Gaussian

FIGURE 16: Comparison of the Gaussian, rectified Gaussian, and truncated Gaussian.

factors belonging to the same cluster. As such, the predictive density $p(s_{l,n} \mid \mathbf{s}_{-l,n}, \gamma_l)$ is shown to be a Student-t distribution, which can be conveniently approximated as a normal distribution when $N_{-l,k}$ is large

$$p(s_{l,n} \mid \mathbf{s}_{-l,n}, \boldsymbol{\gamma}) \approx \mathcal{N}\left(\hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n}\right), \tag{A.11}$$

where denotes a vector of all γ_l; $k \in \{1, 2, \ldots, K, \bar{k}\}$ Moveover, $p(x_{l,n}|s_{l,n})$ can be shown as

$$\begin{aligned} p(x_{l,n} \mid s_{l,n}) &= \delta(x_{l,n})U(-s_{l,n}) + \delta(x_{l,n} - s_{l,n})U(s_{l,n}) \\ &= \widetilde{\pi}_{x_{l,n}}\delta(x_{l,n}) + \left(1 - \widetilde{\pi}_{x_{l,n}}\right)\delta(x_{l,n} - s_{l,n}), \end{aligned} \tag{A.12}$$

where

$$\widetilde{\pi}_{x_{l,n}} = U(-s_{l,n}). \tag{A.13}$$

Taking together, the conditional distribution can be shown as

$$\begin{aligned} &p\left(s_{l,n} \mid x_{l,n}, \mathbf{s}_{-l,n}, \boldsymbol{\gamma}_{-l}, \gamma_l\right) \\ &\quad = \overline{\overline{\pi}}_{x_{l,n}}\delta(s_{l,n} - x_{l,n}) \\ &\quad + \left(1 - \overline{\overline{\pi}}_{x_{l,n}}\right)\frac{\mathcal{N}\left(s_{l,n} \mid \hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n}\right)U(-s_{l,n})}{\Phi(-\hat{\mu}_{l,n}/\hat{\sigma}_{l,n})}, \end{aligned} \tag{A.14}$$

where

$$\bar{\bar{\pi}}_{x_{l,n}} = \frac{\mathcal{N}\left(x_{l,n} \mid \hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n}\right)}{\delta(x_{l,n})Q(-\hat{\mu}_{l,n}/\hat{\sigma}_{l,n}) + \mathcal{N}\left(x_{l,n} \mid \hat{\mu}_{l,n}, \hat{\sigma}^2_{l,n}\right)U(x_{l,n})} = \operatorname{sgn}(x_{l,n}). \tag{A.15}$$

Samples of $s_{l,n}$ can be generated from (A.14).

A.5. $p(\sigma^2_{e,g} \mid \boldsymbol{\Theta}, \mathbf{y}_{1:N})$. Let $\mathbf{E} = \mathbf{Y} - \mathbf{AX}$, and thus

$$\mathbf{e}_g \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{e,g}\mathbf{I}_N\right). \tag{A.16}$$

Given the conjugate Inverse-Gamma prior, we have

$$p\left(\sigma^2_{e,g} \mid \boldsymbol{\Theta}, \mathbf{y}_{1:N}\right) = p\left(\sigma^2_{e,g} \mid \mathbf{e}_g\right) = \mathrm{IG}\left(\alpha_g, \beta_g\right), \tag{A.17}$$

where IG represents the Inverse-Gamma distribution and

$$\alpha_g = \alpha_0 + \frac{N}{2}, \qquad \beta_g = \beta_0 + \sum_{n=1}^{N} \frac{e^2_{g,n}}{2}. \tag{A.18}$$

B. Transcription Factor List

See Table 3.

C. Gene List

See Table 4.

Acknowledgments

This work is supported by a San Antonio Life Science Institute Award to J. Zhang, NSF IIS-0916443 to Y. Qi, NCI Cancer Center Grant P30 CA054174-17 and NIH CTSA 1UL1RR025767-01 to Y. Chen, and NSF CCF-0546345 to Y. Huang.

References

[1] O. Hobert, "Gene regulation by transcription factors and MicroRNAs," *Science*, vol. 319, no. 5871, pp. 1785–1786, 2008.

[2] H. Kitano, Ed., *Foundations of System Biology*, The MIT Press, Cambridge, Mass, USA, 2001.

[3] A. Levchenko, "Computational cell biology in the post-genomic era," *Molecular Biology Reports*, vol. 28, no. 2, pp. 83–89, 2001.

[4] H. Kitano, "Looking beyond that details: a rise in system-oriented approaches in genetics and molecular biology," *Current Genetics*, vol. 41, no. 1, pp. 1–10, 2002.

[5] H. Kitano, "Computational systems biology," *Nature*, vol. 420, no. 6912, pp. 206–210, 2002.

[6] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[7] D. W. Selinger, M. A. Wright, and G. M. Church, "On the complete determination of biological systems," *Trends in Biotechnology*, vol. 21, no. 6, pp. 251–254, 2003.

[8] C. Sabatti and G. M. James, "Bayesian sparse hidden components analysis for transcription regulation networks," *Bioinformatics*, vol. 22, no. 6, pp. 739–746, 2006.

[9] G. Sanguinetti, N. D. Lawrence, and M. Rattray, "Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities," *Bioinformatics*, vol. 22, no. 22, pp. 2775–2781, 2006.

[10] T. Yu and K.-C. Li, "Inference of transcriptional regulatory network by two-stage constrained space factor analysis," *Bioinformatics*, vol. 21, no. 21, pp. 4033–4038, 2005.

[11] A.-L. Boulesteix and K. Strimmer, "Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach," *Theoretical Biology and Medical Modelling*, vol. 2, no. 1, article no. 23, 2005.

[12] K. C. Kao, Y.-L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury, and J. C. Liao, "Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 2, pp. 641–646, 2004.

[13] V. Matys, E. Fricke, R. Geffers et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003.

[14] Q. Qi, Y. Zhao, M. Li, and R. Simon, "Non-negative matrix factorization of gene expression profiles: a plug-in for BRB-ArrayTools," *Bioinformatics*, vol. 25, no. 4, pp. 545–547, 2009.

[15] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, p. 1469, 2004.

[16] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 12, pp. 4164–4169, 2004.

[17] C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modeling: applications in gene expression genomics," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1438–1456, 2008.

[18] E. Sudderth, *Graphical models for visual object recognition and tracking*, Ph.D. thesis, Massachusetts Institute of Technology, 2006.

[19] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[20] N. Socci, D. Lee, and H. Sebastian Seung, "The rectified Gaussian distribution," in *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 350–356, Denver, Colo, US, 1998.

[21] P. M. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data," *Genome Research*, vol. 13, no. 7, pp. 1706–1718, 2003.

[22] T. Li and C. Ding, "The relationships among various non-negative matrix factorization methods for clustering," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, pp. 362–371, Hong Kong, December 2006.

[23] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology*, vol. 4, no. 4, article no. 210, 2003.

[24] C. Wong, Differential Expression and Annotation.

[25] D. Wilson, V. Charoensawan, S. Kummerfeld, and S. Teichmann, "DBD—taxonomically broad transcription factor predictions: new content and functionality," *Nucleic Acids Research*, vol. 36, pp. D88–D92, 2008.

[26] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, CRC Press, Boca Raton, Fla, USA, 2003.

[27] C. Van Rijsbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[28] A. Bagga and B. Baldwin, "Entity-based cross-document coreferencing using the vector space model," in *Proceedings of the 17th International Conference on Computational Linguistics*, vol. 1, pp. 79–85, Association for Computational Linguistics, Morristown, NJ, USA, 1998.

[29] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information Retrieval*, vol. 12, no. 4, pp. 461–486, 2009.

[30] W. A. Thompson, L. A. Newberg, S. Conlan, L. A. McCue, and C. E. Lawrence, "The Gibbs centroid sampler," *Nucleic Acids Research*, vol. 35, pp. W232–W237, 2007.

[31] A. Smith and G. Roberts, "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society. Series B*, vol. 55, no. 1, pp. 3–23, 1993.

[32] G. Celeux, M. Hurn, and C. P. Robert, "Computational and inferential difficulties with mixture posterior distributions," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 957–970, 2000.

[33] K. A. Hoadley, V. J. Weigman, C. Fan et al., "EGFR associated expression profiles vary with breast tumor subtype," *BMC Genomics*, vol. 8, article no. 258, 2007.

[34] M. Mullins, L. Perreard, J. Quackenbush, et al., "Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues," *Clinical Chemistry*, vol. 53, no. 7, p. 1273, 2007.

[35] J. I. Herschkowitz, K. Simin, V. J. Weigman et al., "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors," *Genome Biology*, vol. 8, no. 5, article no. R76, 2007.

[36] J. I. Herschkowitz, X. He, C. Fan, and C. M. Perou, "The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas," *Breast Cancer Research*, vol. 10, no. 5, p. R75, 2008.

[37] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports. Part 1*, vol. 50, no. 3, pp. 163–170, 1966.

[38] J. D. Lieb, X. Liu, D. Botstein, and P. O. Brown, "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association," *Nature Genetics*, vol. 28, no. 4, pp. 327–334, 2001.

[39] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, vol. 409, no. 6819, pp. 533–538, 2001.

[40] B. Ren, F. Robert, J. J. Wyrick et al., "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, no. 5500, pp. 2306–2309, 2000.

[41] R. Jaenisch and A. Bird, "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals," *Nature Genetics*, vol. 33, pp. 245–254, 2003.

[42] E. S. Tasheva, B. Klocke, and G. W. Conrad, "Analysis of transcriptional regulation of the small leucine rich proteoglycans," *Molecular Vision*, vol. 10, pp. 758–772, 2004.

[43] A. Justel and D. Peña, "Gibbs sampling will fail in outlier problems with strong masking," *Journal of Computational and Graphical Statistics*, vol. 5, no. 2, pp. 176–189, 1996.

[44] C. Borgs, J. T. Chayes, A. Frieze et al., "Torpid mixing of some Monte Carlo Markov chain algorithms in statistical physics," in *Proceedings of the 1999 IEEE 40th Annual Conference on Foundations of Computer Science*, pp. 218–229, October 1999.

[45] D. Woodard, "Detecting poor convergence of posterior samplers due to multimodality," Tech. Rep., Citeseer, 2007.