We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Bayesian Inference of Gene Regulatory Network

Xi Chen and Jianhua Xuan

Abstract

Gene regulatory networks (GRN) have been studied by computational scientists and biologists over 20 years to gain a fine map of gene functions. With large-scale genomic and epigenetic data generated under diverse cells, tissues, and diseases, the integrative analysis of multi-omics data plays a key role in identifying casual genes in human disease development. Bayesian inference (or integration) has been successfully applied to inferring GRNs. Learning a posterior distribution than making a single-value prediction of model parameter makes Bayesian inference a more robust approach to identify GRN from noisy biomedical observations. Moreover, given multi-omics data as input and a large number of model parameters to estimate, the automatic preference of Bayesian inference for simple models that sufficiently explain data without unnecessary complexity ensures fast convergence to reliable results. In this chapter, we introduced GRN modeling using hierarchical Bayesian network and then used Gibbs sampling to identify network variables. We applied this model to breast cancer data and identified genes relevant to breast cancer recurrence. In the end, we discussed the potential of Bayesian inference as well as Bayesian deep learning for large-scale and complex GRN inference.

Keywords: gene regulatory network, data integration, Bayesian inference, Gibbs sampling, breast cancer

1. Introduction

The era of "big data" has arrived to the field of computational biology [1]. Biological systems are so complex that in many situations, it is not feasible to directly measure the target signals. Actually, most of biological measurements are noisy and dependent to but not exactly about what we aim to find. This is where probability theory comes to our aid: estimate the true signals from noisy measurements in the presence of uncertainty. Bayesian inference has been widely applied in computational biology field. In certain systems for which we have a good understanding, i.e., gene regulation, behind the observed signals, there exist multiple hidden factors controlling how genes behave under a specific condition. As we are lacking observations on those hidden factors, we model them as parameters in a Bayesian framework, with or without informative prior. Then, for each parameter, Bayesian inference learns a "posterior" distribution, through which we make a final estimation with a confidence interval.

Bayesian inference can update the shape of the learned posterior distributions for model parameters whenever new data observations arrive, providing enough

flexibility for integrative analysis and model extension [2]. Although using more data types means defining more model parameters, Bayesian inference automatically prefers for simple models that sufficiently explain data without unnecessary complexity. This is a very important property for biological data analysis because a simple model is much easier to validate using lab-controlled experiments.

In this chapter, we introduce how to apply Bayesian inference to inferring gene regulatory networks (GRN). GRN is a hierarchical network with regulatory proteins, target genes, and interactions between them [3], playing a key role in mediating cellular functions and signaling pathways in cells [4]. Accurate inference of GRN using data specific for a disease returns disease-associated regulatory proteins and genes, serving as potential targets for drug treatment [5]. In recent years, noncoding DNA analysis reveals more and more noncoding regions with strong regulatory effects on gene transcription [6], which greatly expands the scope of GRN research.

GRN analysis requires an integration of multiple types of measurements including but not limited to gene expression, chromatin accessibility, transcription factor binding, methylation, and histone modification [7]. The challenge of GRN inference is that there exit hundreds of proteins and tens of thousands of genes. One protein can regulate hundreds of target genes, and their regulatory relationship (an interaction in GRN) may vary across different cell types, tissues, or diseases. Experiments of high-throughput target gene measurements for one protein in one specific condition are costive and noisy [8], let alone for hundreds of proteins under diverse conditions. For many tissues or diseases, we need to integrate multiple relevant data types and computationally infer GRNs specific for those conditions.

Bayesian inference is particularly suitable for GRN inference as it is very flexible for large-scale data integration. Moreover, when we have multiple datasets generated from very similar conditions, estimating variables using distribution learning than a single-value prediction makes the final estimation more robust and easier to compare across multiple datasets. We demonstrated this using two breast cancer datasets generated under very similar conditions, in which we also compared a hierarchical Bayesian model with several competing methods. Moreover, using patient data as model input, although they are noisy, we successfully identified a GRN associated with breast cancer recurrence. Finally, we discussed the potential of Bayesian deep learning for large-scale and complex GRN inference.

2. Gene regulatory networks

Human genome can be simply divided into coding (exomes) and noncoding regions. The process of producing an RNA copy from exomes is called transcription, which can be quantitatively measured using microarray or RNA-seq technics [9, 10], producing gene expression data of ~30,000 genes simultaneously. The transcription process is mediated by regulatory regions located in the noncoding genome, including promoters and enhancers [11]. Promoters are proximal to gene transcription starting sites (TSS), usually within 3 kbps (**Figure 1A**), while enhancers are usually located distantly, i.e., 200 kbps (**Figure 1B**), and can be up to 1 Mbps. In general, each gene could be associated with one promoter and multiple enhancers.

Transcription factors (TFs), a special category of proteins, often coordinate with each other as cis-regulatory modules (CRMs) [12] and co-bind at regulatory regions [13]. For example, in **Figure 1A** or **B**, there are three TFs binding at promoter or enhancer regions and functioning together as one CRM to mediate the transcription process of their target genes. It has been known that the association relationships of

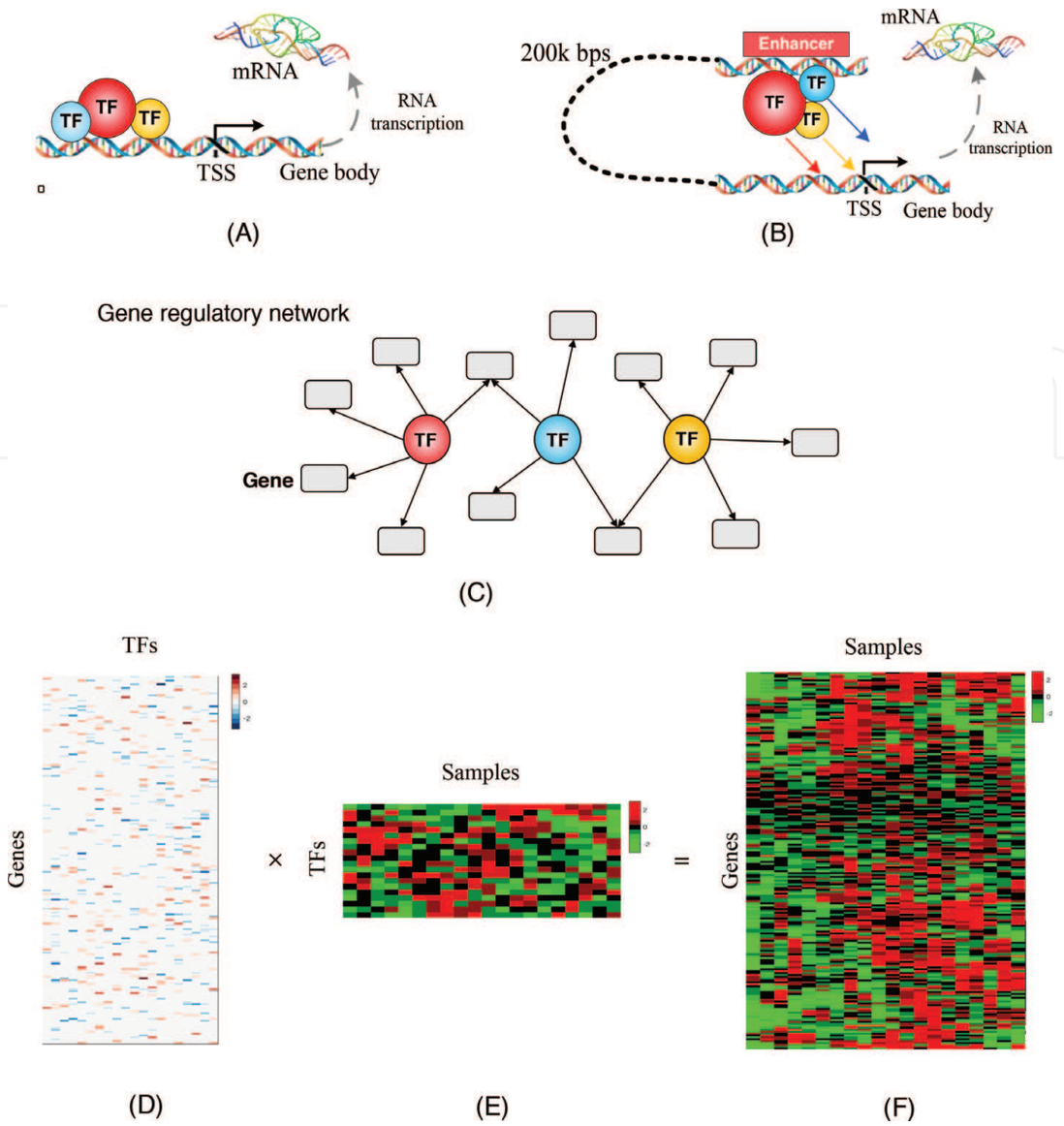

Figure 1.
Illustration of gene regulation: (A) transcription factor (TF)-gene regulation through proximal promoter regions; (B) TF-gene regulation through distal enhancer regions; (C) a gene regulatory network (GRN) including TFs, genes, and their interactions; (D) regulatory effects of TFs on individual genes with "red" as activation, "blue" as depression, and "white" as no regulatory effects; (E) a heatmap of TF protein activities across biological samples of multiple conditions with "red" as enhanced activity, "green" as reduced activity, "black" as no activity; and (F) a heatmap of gene expression across multiple samples, with "red" as up-regulated, "green" as down-regulated, "back" as no change.

TFs are not random [14, 15]. Some TFs tend to co-bind at the same regions more often than with others, i.e. MYC and MAX. One TF can regulate multiple genes, and a target gene can also be regulated by multiple TFs considering the existence of CRMs (**Figure 1C**). For each specific TF-gene interaction in **Figure 1C**, its regulatory effect can be either positive (activating gene expression) or negative (depressing gene expression), as shown in **Figure 1D**. The protein activities of TFs are therefore connected to the dynamic changes of gene expression across multiple samples [13]. To accurately identify GRNs, we need quantitative measures of all types of signals in **Figure 1D–F**. However, due to technical limitations, we can obtain good quality measurements of gene expression, binary measurements (existence or not) of individual TF-gene interactions yet with a high false positive rate, but no measurements of TF activities. To infer GRNs, we must jointly estimate TF activities, TF-gene regulation strengths, and CRMs (TF associations) given gene expression observations.

3. Bayesian inference

Bayesian inference is particularly suitable for inferring GRN as it will learn a posterior distribution for each variable, with a high tolerance on the noise existing in the gene expression data or caused by non-perfect prior assumptions.

3.1 A hierarchical Bayesian model

Given gene expression data under multiple biological samples (conditions), we focus on the expression variation of each gene from its baseline expression because such variation reflects the effects of condition changes. For a specific disease, only genes showing significant expression changes between disease cells and normal cells are interesting candidates. Thus, for gene n, we calculate the log fold change of gene expression under each sample (1,2,3,...,M) to that of baseline condition (0). To model gene expression data of hundreds of genes in the same framework, for genen, we normalize its M log fold change values (indexed by m) to values with 0-mean and 1-standard deviation, denoted by $y_{n,m}$. Then, a linear model is applied to modeling $y_{n,m}$ as follows [16, 17]:

$$y_{n,m} = \sum_t a_{n,t} b_{n,t} x_{t,m} + \varepsilon_n, \tag{1}$$

where variable $a_{n,t}$ denotes the regulation strength of TF t on gene n; $b_{n,t}$ is a binary variable denoting the regulation occurrence of TF t on gene n; TF protein activity variable $x_{t,m}$ under condition (sample) m directly connects to gene expression $y_{n,m}$ under the same condition [16]; and the noise variable ε_n denotes inaccuracy of gene expression marvelments.

Given protein-DNA binding measurements of T TFs and N genes (i.e., ENCODE database), we are able to identify TF binding sites at promoter or enhancer regions within 1 Mbps around individual target genes [18]. Each gene can be associated with several regulatory regions, and at each region, there exit a subset of TFs, as a candidate CRM. Then, we may observe multiple candidate CRMs (in total K_n) for gene n, indexed by $c_n = 1, 2, 3, ..., k, ..., K_n$. Each c_n is associated with a unique set of TF-gene binding events ($b_{c_n,t} = 1$ or $b_{c_n,t} = 0$). We assume c_n a hidden variable controlling how binding variables are associated with each other, with candidate space defined from existing databases.

To estimate the abovementioned variables, we develop a hierarchical Bayesian network to model their internal dependency and associations with gene expression, as shown in **Figure 2**. CRM variable c controls the state of each binding variable b. For $b = 1$, regulation strength a can be either positive or negative denoting gene activation or depression by the binding TF. In the meanwhile, through TF-gene regulation, the protein activities of TFs are directly connected to target gene expression, with ε denoting the measurement noise in gene expression data. With Eq. (1) and **Figure 2**, we aim to estimate all these variables using Bayesian inference, which requires a prior assumption (not necessary to be informative) on the distribution of each variable.

Based on prior binding observations from public database, the candidate space of CRM is known, denoted by **C**. Given a gene expression dataset generated from a specific condition, for gene n, we need to estimate which CRM c_n is regulating its gene expression. As the prior data does not tell which CRM is more likely to be true under a specific condition, we assume a discrete uniform prior on c.

Based on data observation, y has a Gaussian-like distribution with 0-mean and 1-standard deviation. The gene expression noise component ε can be assumed to

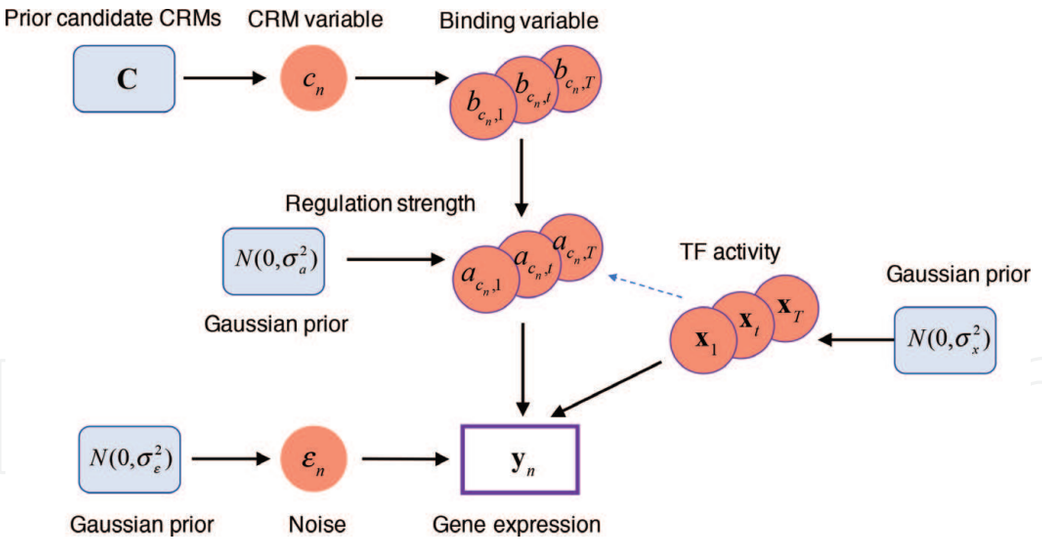

Figure 2.
A hierarchical Bayesian framework for GRN modeling. The number of variables in this framework depends on the numbers of biological samples, TFs, genes, and candidate CRMs. Given gene expression data under different conditions, for the same TF and same gene, their regulatory relationship (variable b) may have very different regulatory strength (variable a). And the TF activity (variable x) can be significantly different as well. Therefore, GRNs are highly context-specific.

follow a 0-mean Gaussian distribution as well, denoted by $N(0, \sigma_\varepsilon^2)$. Although the variance of noise is hard to determine, it should fall in the same scale as gene expression measurements. Therefore, we set $\sigma_\varepsilon^2 = 1$.

The regulation strength variable a is conditional on the state of b (as shown in **Figure 2**): for $b = 0$, we set $a = 0$, denoting the nonexistence of TF-gene regulation; for $b = 1$, a can be either positive or negative so that we assume a 0-mean Gaussian prior on a, as $N\left(0, \sigma_{a,prior}^2\right)$ (the variance $\sigma_{a,prior}^2$ is a hyperparameter). As GRN is a sparse network, most a values would be 0.

We model TF activity x under multiple biological samples using Gaussian random processes. As baseline expression is largely removed from gene expression data during the data normalization process, ideally the baseline activity of each TF is 0. In each sample, x can be either enhanced or reduced with respect to its baseline activity. Thus, we assume a 0-mean Gaussian prior for x, as $N\left(0, \sigma_{x,prior}^2\right)$ (the variance $\sigma_{x,prior}^2$ is also a hyperparameter).

Regarding hyperparameters of the prior mean and variance for a or x, a benefit of assuming 0-mean prior is to control model overfitting. Only when the posterior distribution has a significant non-zero mean value that we will accept that estimation. It is hard to determine the scale of variable values without direct measurements. A conservative way is to assume non-informative prior on them and let the algorithm determine the final posterior distribution, although the non-informative prior will lead to a stickier chain and a posterior with potential multiple modes. Exploring such a posterior is certainly more challenging than exploring a well-behaved unimodal posterior. However, there is really no need to trouble with this multimodal posterior on a or x, as the inferential values of the whole framework are: the discrete posterior distributions of CRMs. For each gene, the posterior distribution of CRMs learned from a data reveals which CRM(s) are regulating this gene. If there are more than one mode in the CRM posterior distribution, this gene will be associated with two or three CRMs. This is quite common in gene regulatory networks as one gene can be regulated by CRMs at multiple regulatory regions simultaneously. $\sigma_{a,prior}^2$ and $\sigma_{x,prior}^2$ should be significantly larger than the variance of

gene expression data to allow a "large" space for the algorithm to generate posterior distributions. As y is already normalized with variance of 1, we set $\sigma^2_{a,prior} = 10$ and $\sigma^2_{x,prior} = 100$.

Then, the problem of GRN inference is Bayesian formed as estimating posterior probabilistic distributions of $\mathbf{A} = \{a_{c_n,t}\}$, $\mathbf{B} = \{b_{c_n,t} | b_{c_n,t} = 0 \text{ or } 1\}$, and $\mathbf{X} = \{x_{t,m}\}$ given $\mathbf{Y} = \left\{y_{n,m}\right\}$. Considering the dependence relationship of all variables in **Figure 2**, we define a joint posterior probability as follow:

$$\begin{aligned} P(\mathbf{A},\mathbf{B},\mathbf{X}|\mathbf{Y}) &\propto P(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{X}) \times P(\mathbf{A}) \times P(\mathbf{C}) \times P(\mathbf{X}) \\ &\propto \prod_n \prod_m \left(\sigma_\varepsilon^{-1}\right) \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(y_{n,m} - \sum_t a_{c_n,t} b_{c_n,t} x_{t,m}\right)^2\right) \\ &\times \prod_n \prod_t \left(\sigma_{a,prior}^{-1}\right) \exp\left(-\frac{a_{c_n,t}^2}{2\sigma_{a,prior}^2}\right) \\ &\times \prod_n \prod_{c_n} \frac{1}{K_n} \\ &\times \prod_t \prod_m \left(\sigma_{x,prior}^{-1}\right) \exp\left(-\frac{x_{t,m}^2}{2\sigma_{x,prior}^2}\right). \end{aligned} \tag{2}$$

Estimating the joint distribution of above-mentioned variables is difficult. Alternatively, we can approximate the joint posterior distribution by estimating the marginal distribution of each variable. To do that, we iteratively calculate each variable's conditional probability and perform Bayesian estimation using Gibbs sampling. The advantage of using Gibbs sampling is that it is theoretically guaranteed to converge to the posterior distribution [2, 19–21].

3.2 Gibbs sampling

We first sample TF activity variable $x_{t,m}$ for the TF t and sample m, according to its conditional probability (based on Eq. (2)) as follows (**Figure 3**):

$$P(x_{t,m}|\mathbf{Y},\mathbf{A},\mathbf{B}) \propto \prod_n \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(y_{n,m} - \sum_t a_{c_n,t} b_{c_n,t} x_{t,m}\right)^2 - \frac{x_{t,m}^2}{2\sigma_{x,prior}^2}\right). \tag{3}$$

$P(x_{t,m}|\mathbf{Y},\mathbf{A},\mathbf{B})$ is a Gaussian distribution with mean and variance as follows:

$$\begin{cases} \mu_x = \dfrac{\sigma_{x,prior}^2 \sum_n \left(y_{n,m} - \sum_{j \neq t} a_{c_n,j} b_{c_n,j} x_{j,m}\right) a_{c_n,t} b_{c_n,t}}{\sigma_{x,prior}^2 \sum_n a_{c_n,t}^2 b_{c_n,t}^2 + \sigma_\varepsilon^2 N} \\ \sigma_x^2 = \dfrac{\sigma_\varepsilon^2 N \sigma_{x,prior}^2}{\sigma_{x,prior}^2 \sum_n a_{c_n,t}^2 b_{c_n,t}^2 + \sigma_\varepsilon^2 N} \end{cases} \tag{4}$$

As shown in Eq. (4), the estimation of distribution of $x_{t,m}$ is conditional on other TF activities $x_{j,m}(j \neq t)$. Therefore, we iteratively sample $x_{t,m}$ as $x_{t,m}|x_{j,m}(j \neq t)$ one

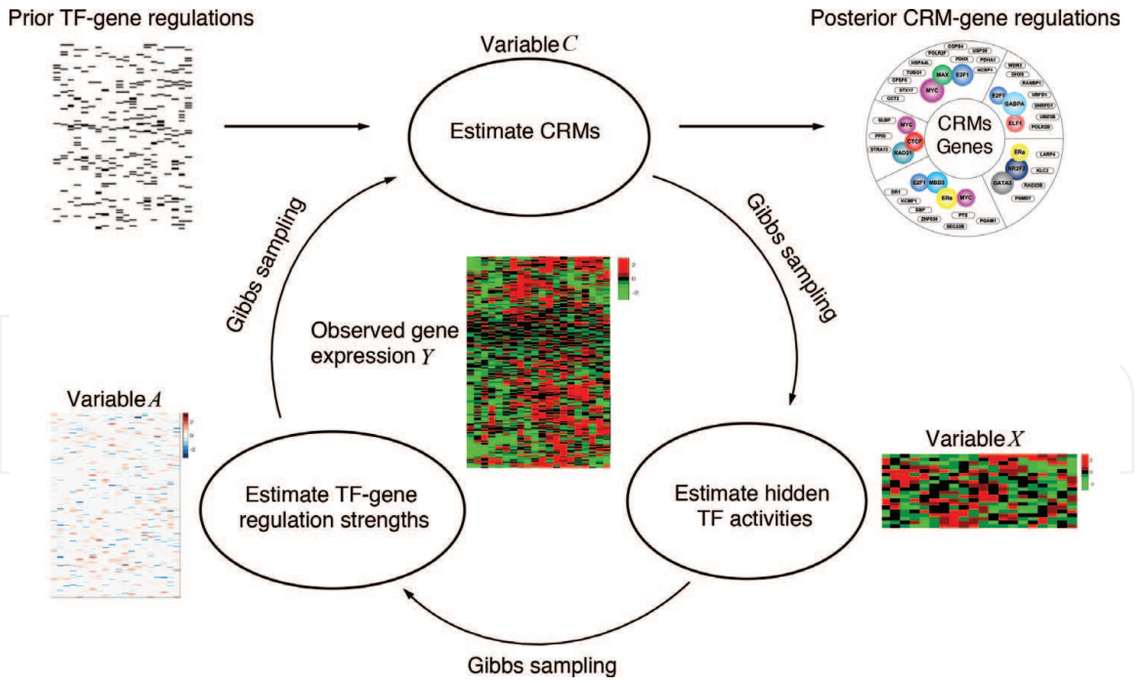

Figure 3.
Gibbs sampling of CRMs, TF activities, and regulation strengths with prior TF-gene regulation and gene expression observations as input.

by one for $t = 1 \sim T$ according to each individual posterior Gaussian distribution $N(\mu_x, \sigma_x^2)$.

Secondly, for gene n, for each $b_{c_n,t} = 1$, we estimate the associated regulation strength $a_{c_n,t}$ according to the following conditional probability:

$$P(a_{c_n,t}|\mathbf{Y},\mathbf{X},\mathbf{B}) \propto \prod_m \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(y_{n,m} - \sum_t a_{c_n,t}x_{t,m}\right)^2 - \frac{a_{c_n,t}^2}{2\sigma_{a,prior}^2}\right). \quad (5)$$

$P(a_{c_n,t}|\mathbf{Y},\mathbf{X},\mathbf{B})$ is a Gaussian distribution, too, with mean and variance calculated as follows:

$$\begin{cases} \mu_a = \dfrac{\sigma_{a,prior}^2 \sum_m \left(y_{n,m} - \sum_{j\neq t} a_{c_n,j}x_{j,m}\right)x_{t,m}}{\sigma_{a,prior}^2 \sum_m x_{t,m}^2 + M\sigma_\varepsilon^2} \\ \sigma_a^2 = \dfrac{\sigma_{a,prior}^2 M\sigma_\varepsilon^2}{\sigma_{a,prior}^2 \sum_m x_{t,m}^2 + \sigma_\varepsilon^2 M} \end{cases} \quad (6)$$

Similar to the estimation process of TF activity variables, the posterior distribution of each $a_{c_n,t}$ also depends on the values of the other $a_{c_n,j}(j \neq t)$. Thus, we iteratively sample $a_{c_n,t}$ for TFs in module c_n one by one according to each individual posterior Gaussian distribution $N(\mu_a, \sigma_a^2)$.

Finally, with sampled TF activity and regulation strength variables, we sample CRM variable c_n for the gene n. It is hard to assume a prior probabilistic distribution shape on the joint distribution of multiple binding variables in c_n. In practice, c_n has a finite number of states as K_n. Therefore we can directly calculate a discrete discrete conditional probability for each $c_n = k$ as follows:

$$P(c_n|\mathbf{y}_n, \mathbf{A}, \mathbf{X}) \propto \prod_t P(\mathbf{y}_n|a_{c_n,t}, \mathbf{x}_t)P(a_{c_n,t}|c_n)P(\mathbf{x}_t)$$
$$\propto \prod_t \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_m\left(y_{n,m} - \sum_t a_{c_n,t}b_{c_n,t}x_{t,m}\right)^2 - \frac{a_{c_n,t}^2}{2\sigma_{a,prior}^2} - \frac{\sum_m x_{m,t}^2}{2\sigma_{x,prior}^2}\right) \quad (7)$$

After calculating Eq. (7) for all possible values of c_n, we sample one value according to the following discrete probability density function:

$$p(c_n = k) = \frac{P(c_n = k|\mathbf{y}_n, \mathbf{X}, \mathbf{A})}{\sum_p P(c_n = p|\mathbf{y}_n, \mathbf{X}, \mathbf{A})} \quad (8)$$

After sampling TFA, TF-gene regulation strength, and cis-regulatory module variables for all N genes, we update binding states in matrix **B** according to the sampled CRMs for individual genes and start the next round of sampling.

Convergence of Gibbs sampling can be monitored based on the ratio (R) of within-variance and between-variance using multiple sequences with different initial states [22]. In each application, we ran five sequences of sampling in parallel. In the i-th round of sampling, for each variable we calculated the within-variance using samples from 1 to i in each sequence and then take the mean value of variances from five sequences. In the meanwhile, we calculate the between-variance of the same variable using its sampled values in the i-th round but from five sequences. For each catalog of variables, the distribution of ratio (R) between within-variance and between-variance is used to monitor the overall sampling convergence. When the sampler converges, values of R would be around "1." We, respectively, monitor the sampling convergence for regulation strengths and TF activities. Once both of them converge, we start to accumulate samples on TF-gene binding variables. As each TF-gene binding variable is binary, its sampling frequency represents the posterior probability of binding occurrence. In the meanwhile, for each gene, a discrete posterior probability distribution of all associated candidate CRMs is inferred, the mode of which reveals the most likely regulatory region associated with current gene.

4. Inferring GRNs for breast cancer

4.1 Application to in vitro breast cancer cell line data

We first applied the hierarchical Bayesian model to gene expression data measured from in vitro breast cancer cell lines. We chose to use cell line data mainly because such data is usually clean and good for validating computational models. Here, we carefully selected two public available breast cancer cell line datasets measured independently but under the same condition (downloadable from the GEO database https://www.ncbi.nlm.nih.gov/geo/, with accession number GSE62789 for Data #1 and accession number GSE51403 for Data #2, both treated by 24 hours of 17b-estradiol (E2) to stimulate breast cancer cells proliferation). The similarity between the two inferred GRNs can be used to evaluate the robustness of GRN inference methods.

For prior TF-gene collection, we checked the ENCODE database (https://www.encodeproject.org/) and selected genome-wide binding profiles of 39 TFs, measured from the same breast cancer cell line. We collected candidate binding events by

examining TF binding signals at promoter and distantly associated enhancers associated with each gene. In total we collected 2,319 candidate TF-gene interactions (**Figure 4A**) between 39 TFs and 275 genes, whose gene expression is consistently upregulated in both datasets when breast cancer cells are stimulated to fast proliferate (**Figure 4B** and **C**). We, respectively, applied the hierarchical Bayesian model to the two gene expression datasets with the same prior settings. To monitor the convergence of the sampling process, we ran five sequences with different initial states and sampled 1000 times in each. As shown in **Figure 4C** and **D** (for Data #1), after 100 rounds of sampling, the model started to converge. The sampling frequency on each TF-gene interaction was calculated as the posterior probabilistic weight. We extracted top ~500 most confident TF-gene interactions as the final GRN estimation for each data set and then focused on common interactions between two relevant GRNs.

Here, we specifically compared our approach with three competing methods (COGRIM [20], LASSO [23], and NARROMI [24]). COGRIM was a Bayesian inference approach without modeling on CRMs. It treated individual TF-gene binding events independently. Although such an assumption lowered the model complexity, it made the model less robust against the inaccuracy in the TF-gene binding prior. Moreover, for the TF activity, COGRIM simply treated it as an observed value by directly using TF mRNA expression. Although ideally the variation of mRNA transcription is proportional to the activity change of mRNA-translated protein, currently this correlation is very low in most studies using gene expression. These inaccurate assumptions brought a lot of uncertainty to modeling gene expression data. LASSO used a linear regression model to integrate prior TF-gene interactions and gene expression data and predicted one value for each TF-gene interaction. The NARROMI approach inferred GRNs using gene expression data only without any prior on TF-gene interactions, and also, it made single-value prediction for each interaction based on the mutual information between gene and TF expression values. Theoretically, the Bayesian approach described in this chapter should be

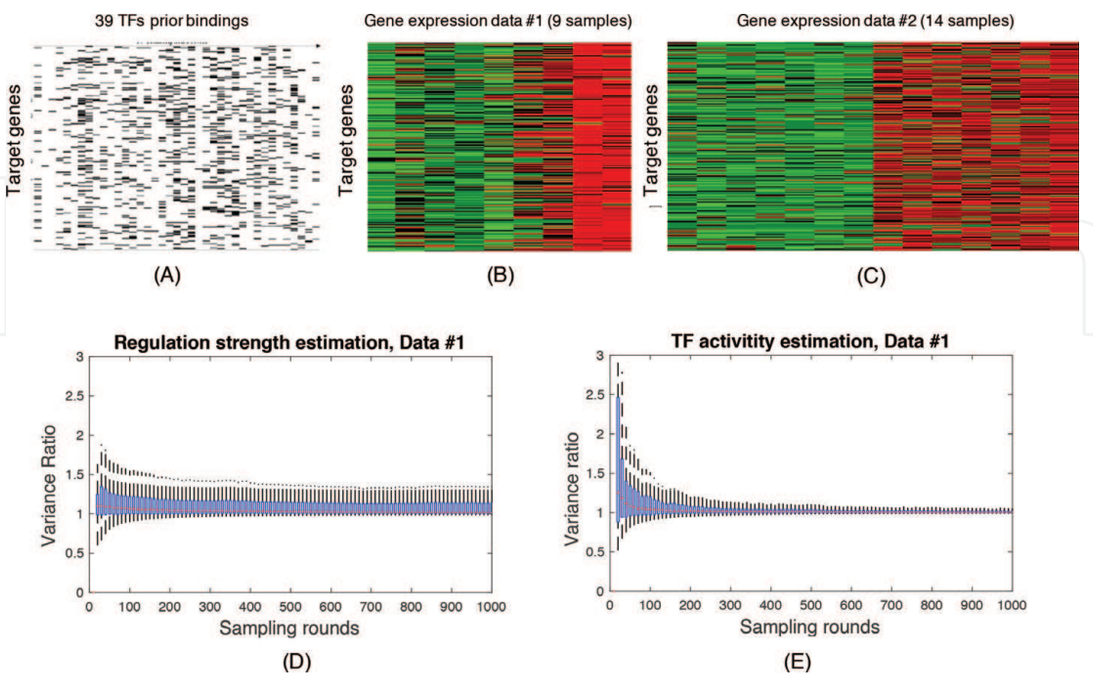

Figure 4.
Input breast cancer cell line data for GRN inference: (A) prior TF-gene interactions ("black" denotes binding occurrence); (B) heatmap of time-course gene expression data; (C) heatmap of steady-state gene expression, all data are from the same breast cancer cell line; (D) convergence of regulatory strength estimation using time-course gene expression data; and (E) convergence of TF activity estimation using time-course gene expression data.

more robust to identify GRNs. We applied the four competing methods to the above two datasets. Indeed, GRNs identified using our Bayesian model were more consistent between two related datasets (**Table 1**).

By analyzing the common 306 TF-gene interactions in **Table 1**, we identified two functional CRMs. The first CRM had five TFs including POL2A, TDRD3, MYC, MAX, and E2F1 (**Figure 5A**). The activities of these TFs, as inferred from both datasets, were shown in **Figure 5B** and **C**, respectively. In total there were 100 genes regulated by this module, and 60 of them were associated with breast cancer through literature survey (selected genes shown in **Figure 5D**). The second CRM had six TFs including ELF1, JUND, JUN, FOXA1, CTCF, and HDAC1. In total, there were 89 genes regulated by this module, and 51 of them were associated with breast cancer (selected genes shown in **Figure 5E**). COGRIM identified fewer genes for the first CRM and failed to identify the second CRM. For the other non-Bayesian approaches, as the number of common TF-gene interactions inferred from two

Methods	GRN edges in Data #1	Similarity with other methods	GRN edges in Data #2	Similarity with other methods	Common GRN for Data #1 and #2
Bayesian	500	**0.878*****	413	**0.822*****	**306*****
COGRIM	516	0.798	457	0.696	239
LASSO	565	0.486	510	0.533	74
NARROMI	514	0.519	591	0.516	44

****denotes hypergeometric p-value < 0.001.*

Table 1.
Comparison of methods for robust GRN inference.

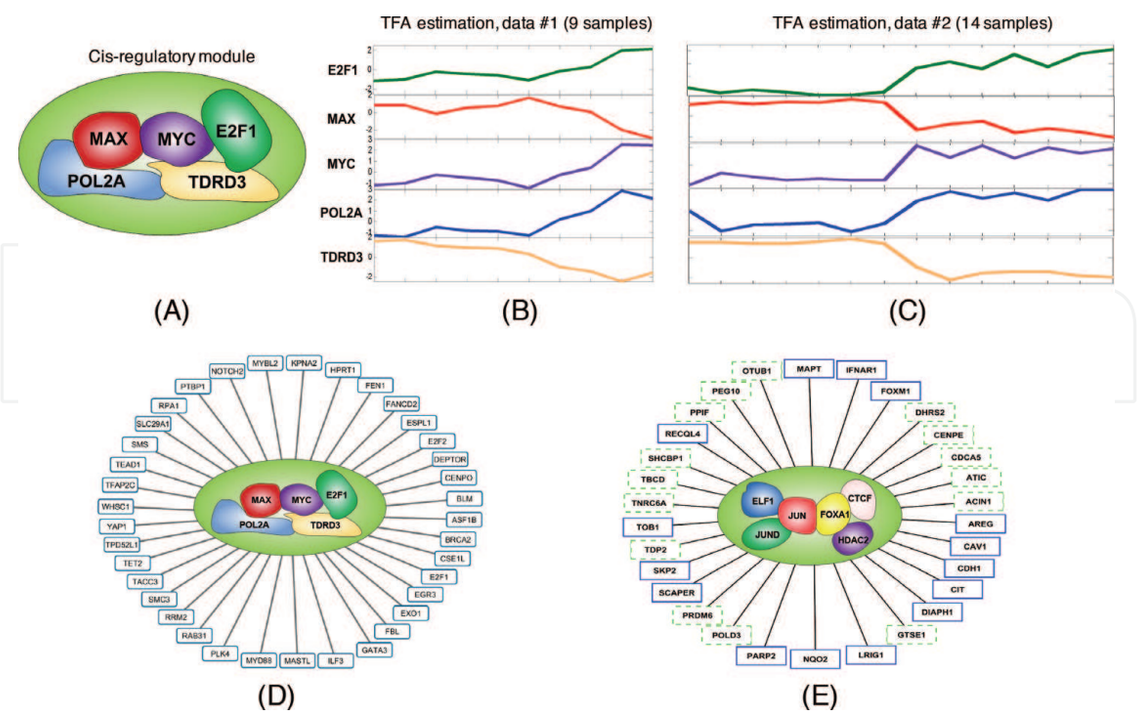

Figure 5.
Key CRMs inferred from breast cancer cell line data: (A) CRM #1 and their TF components; (B) estimated TF activities from Data #1 (time-course); (C) estimated TF activities from Data #2 (steady state); (D) target genes regulated by CRM with MAX, MYC, E2F1, POL2A and TDRD3; (E) target genes regulated by CRM with ELF1, JUND, JUN, FOXA1, CTCF, and HDAC2. Target genes in D and E are associated with breast cancer as supported by literature survey. "Blue" block represents genes showing up in at least two literatures, while "green" block represents genes with one literature support.

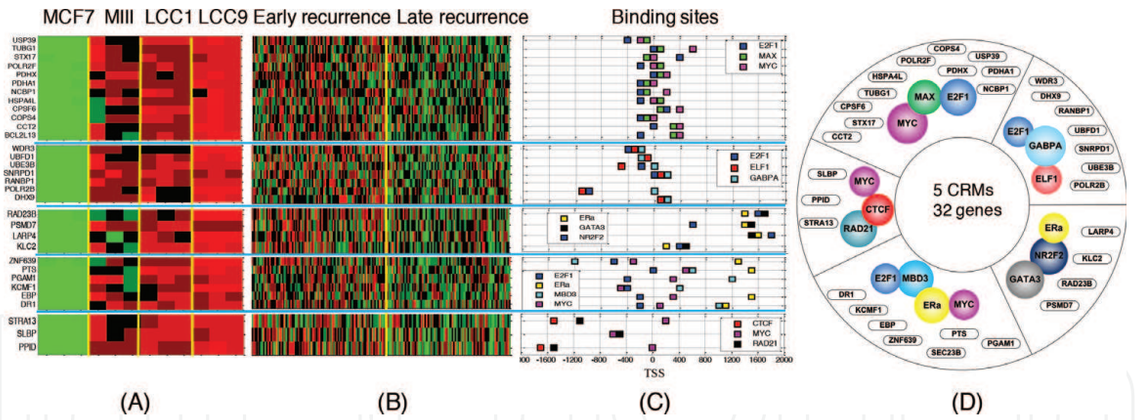

Figure 6.
Breast cancer recurrence-associated GRN: (A) heatmap of gene expression in breast cancer cell lines including MCF7, MIII, LCC1, and LCC9, where "red" represents overexpression and "green" represents lower expression; (B) heatmap of gene expression of breast cancer patients in "Early recurrence" and "Late recurrence" groups, divided by 5-year survival; (C) binding sites of 11 TFs on 32 target genes; and (D) association of 5 CRMs and 32 target genes.

datasets was small, size reduced by over 75%. We did not identify the two key CRMs using either approach.

4.2 Application to breast cancer patient data

We finally applied the Bayesian approach to breast cancer patient data downloaded from the TCGA database (https://portal.gdc.cancer.gov/). Survival time distribution of 93 breast cancer patients treated by *tamoxifen* revealed two modes with 5-year survival as division. Accordingly, we defined an "Early recurrence" group including patients with survival time <5 years and a "Late recurrence" group including patients with survival time longer than 5 years. Differentially expressed genes between two groups (t-test p-value <0.05) were selected for further GRN analysis. It can be seen from **Figure 6B** that the gene expression data of breast cancer patient is quite noisy. To increase the robustness of GRN results, we used another cell line dataset. Specifically, gene expression data was generated from four cell lines including MCF7, MIII, LCC1, and LCC9, with three replicates for each. MCF7 cells were sensitive to *tamoxifen* treatment, while LCC9 cells were drug-resistant. One hypothesis is that breast cancer recurrence is associated with drug resistance. Thus, we expected that the overexpressed genes in the "Early recurrence" group were also overexpressed in LCC9 cells. For 431 genes with such expression pattern in both patient and cell line data, we collected prior TF-gene interactions from 39 TF binding profiles used in previous sections. We, respectively, inferred GRNs using both datasets and identified a common GRN including interactions between 25 proteins and 161 genes. Analysis of this common CRN revealed 5 key CRMs with 11 proteins and 32 target genes highly relevant to breast cancer recurrence (**Figure 6**).

5. Discussion

5.1 Gene regulatory networks in different cell states

Recent technology advance in single-cell gene transcription makes it feasible to study TF-gene regulation during the cell differentiation process [25]. In sections above, across multiple samples, TF-gene interactions are assumed to hold, and the

gene expression change is connected to the dynamic variation of TF activities across samples. Yet, at the single-cell level, gene expression measurements are very noisy, whose variation across cells may be partially disconnected from the dynamic changes of TF activities [26]. In that situation, the linear model in Eq. (1) will not work with such gene expression input. Moreover, during the cell differentiation process, in fact we do not have prior knowledge on whether GRNs will hold or change between individual cell states. That means TF-gene interaction change can be another causal factor on gene expression variation across different cell states, too. To model GRNs individually for cell states, we need to define more binding variables, which will definitely make the estimation process more complex.

Those cell state-specific GRNs will uncover the regulatory mechanism that drives cell differentiation. This would be particularly useful for cancer treatment. If any regulation changes at a very early cell state eventually lead to cancer cell fast proliferation, we can engineeringly target those TFs, binding regions, or genes for cancer prevention. Currently inference of cell-state-specific GRN is either through enrichment analysis of TF binding signals in each cell state [27] or regression modeling of gene expression using the matched measurements of regulatory region activities [28]. When the single-cell expression measurements become more accurate, we hope the connection between gene expression and TF activities still holds. Then, the model in Eq. (1) with proper improvement can be used to infer cell-state-specific GRNs.

5.2 Bayesian neural network

Although theoretically there is no upper limit on the number of model parameters in the Bayesian framework (**Figure 2**), the more variables we have, the slower the convergence will be. Moreover, given a complex network with many states, the dependence of different variables will be hard to model, and the estimation process is more easily to stuck into a local state. In recent years, neural network is widely applied to variable estimation in complex systems. Neural network is an end-to-end system that mimics the human brain and tries to learn complex representation within the dataset to provide an output. Similar to conventional machine learning, deep neural networks make a single-value prediction for each model parameter, without measuring uncertainty. That means the model performance relies heavily on the prediction accuracy, and even one overconfident decision can result in a big problem. A Bayesian approach to neural networks can naturally solve this problem by learning a distribution accounting for the uncertainty in parameter estimates [29].

Unlike Bayesian inference discussed in previous sections, inferring model posterior in a Bayesian neural network is much more difficult as there are many parameters to estimate in neural networks. Direct inference of variable posterior distribution is hard so that approximations to the posterior are often used, i.e., the variational inference. The posterior can be modelled using a simple variational distribution such as a Gaussian distribution, and the distribution's parameters are fitted to approximate the true posterior as close as possible by minimizing the Kullback-Leibler divergence between this simple variational distribution and the true posterior. In earlier sections, we have demonstrated that modeling variables in GRN using Gaussian distribution provided robust performance. To infer large-scale GRN with thousands of genes and hundreds of TFs, Bayesian neural network can be a solution in which posterior distributions of all variables can be approximated by Gaussian distribution.

6. Conclusion

In this chapter, we mathematically illustrated how Bayesian inference can be used to infer gene regulatory networks. Using several breast cancer-specific datasets, we demonstrated the effectiveness of Bayesian network modeling in biological meaningful signal discovery, in comparison with methods of linear regression. Potentially, Bayesian inference can be used to infer dynamic GRN during cell differentiation using new types of gene expression data. For very large-scale GRN inference in complex systems, the big number of variables may degrade conventional Bayesian inference performance. Bayesian neural networks using variational inference can be a good solution.

Acknowledgements

Funding for open access charge: Virginia Tech's Open Access Subvention Found (VT OASF).

Author details

Xi Chen[1,2*] and Jianhua Xuan[1]

1 Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA

2 Center for Computational Biology, Flatiron Institute, New York, NY, USA

*Address all correspondence to: xichen86@vt.edu

References

[1] Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods. 2008;**5**(1):16-18

[2] Chen X et al. CRNET: An efficient sampling approach to infer functional regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data. Bioinformatics. 2018; **34**(10):1733-1740

[3] Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nature Reviews. Genetics. 2004;**5**(2):101-113

[4] Blais A, Dynlacht BD. Constructing transcriptional regulatory networks. Genes & Development. 2005;**19**(13): 1499-1511

[5] van 't Veer LJ et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;**415**(6871): 530-536

[6] Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. Nature Reviews. Genetics. 2016;**17**(4):207-223

[7] Bock C, Lengauer T. Computational epigenetics. Bioinformatics. 2008;**24**(1): 1-10

[8] Landt SG et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research. 2012;**22**(9):1813-1831

[9] Quackenbush J. Microarray data normalization and transformation. Nature Genetics. 2002;**32**(Suppl): 496-501

[10] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews. Genetics. 2009;**10**(1):57-63

[11] Riethoven JJ. Regulatory regions in DNA: Promoters, enhancers, silencers, and insulators. Methods in Molecular Biology. 2010;**674**:33-42

[12] Chen X, Xuan J, Shi X, Shajahan-Haq AN, Hilakivi-Clarke L, Clarke R. A novel statistical approach to identify co-regulatory gene modules. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine; 2013. pp. 16-18

[13] Spitz F, Furlong EE. Transcription factors: From enhancer binding to developmental control. Nature Reviews. Genetics. 2012;**13**(9):613-626

[14] Wang J et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research. 2012;**22**(9):1798-1812

[15] Chen X, Shi X, Shajahan-Haq AN, Hilakivi-Clarke L, Clarke R, Xuan J. Statistical identification of co-regulatory gene modules using multiple ChIP-seq experiments. In: Presented at the International Conference on Bioinformatics Models, Methods and Algorithms (Bioinformatics); 2014

[16] Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: Reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences of the United States of America. 2003;**100**(26): 15522-15527

[17] Chen X, Xuan J, Wang C, Shajahan AN, Riggins RB, Clarke R. Reconstruction of transcriptional regulatory networks by stability-based network component analysis. IEEE/ ACM Transactions on Computational Biology and Bioinformatics. 2013;**10**(6): 1347-1358

[18] Chen X et al. ChIP-BIT: Bayesian inference of target genes using a novel joint probabilistic model of ChIP-seq profiles. Nucleic Acids Research. 2016; **44**(7):e65

[19] Sabatti C, James GM. Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics. 2006;**22**(6):739-746

[20] Chen G, Jensen ST, Stoeckert CJ Jr. Clustering of genes into regulons using integrated modeling-COGRIM. Genome Biology. 2007;**8**(1):R4

[21] Shi X et al. mAPC-GibbsOS: An integrated approach for robust identification of gene regulatory networks. BMC Systems Biology. 2013;**7** (Suppl 5):S4

[22] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992;**7**(4): 457-472

[23] Qin J, Hu Y, Xu F, Yalamanchili HK, Wang J. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. Methods. 2014; **67**(3):294-303

[24] Zhang X et al. NARROMI: A noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics. 2013;**29**(1):106-113

[25] Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology. 2016;**34**(11):1145-1160

[26] Raj A, van Oudenaarden A. Nature, nurture, or chance: Stochastic gene expression and its consequences. Cell. 2008;**135**(2):216-226

[27] Aibar S et al. SCENIC: Single-cell regulatory network inference and clustering. Nature Methods. 2017; **14**(11):10831-11086

[28] Cao J et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;**361**(6409):1380-1385

[29] Crucianu M, Bone R, de Beauville JPA. Bayesian learning for recurrent neural networks. Neurocomputing. 2001;**36**:235-242