IET Systems Biology

Research Article

IET Journals
The Institution of Engineering and Technology

Adaptive modelling of gene regulatory network using Bayesian information criterion-guided sparse regression approach

ISSN 1751-8849
Received on 31st January 2016
Revised on 13th June 2016
Accepted on 14th June 2016
doi: 10.1049/iet-syb.2016.0005
www.ietdl.org

Ming Shi[1,2], Weiming Shen[2], Hong-Qiang Wang[1], Yanwen Chong[2] ✉

[1]*Machine Intelligence and Computational Biology Lab, Institute of Intelligent Machines, Chinese Academy of Science, P.O. Box 1130, Hefei 230031, People's Republic of China*
[2]*State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, People's Republic of China*
✉ *E-mail: ywchong@whu.edu.cn*

Abstract: Inferring gene regulatory networks (GRNs) from microarray expression data are an important but challenging issue in systems biology. In this study, the authors propose a Bayesian information criterion (BIC)-guided sparse regression approach for GRN reconstruction. This approach can adaptively model GRNs by optimising the l_1-norm regularisation of sparse regression based on a modified version of BIC. The use of the regularisation strategy ensures the inferred GRNs to be as sparse as natural, while the modified BIC allows incorporating prior knowledge on expression regulation and thus avoids the overestimation of expression regulators as usual. Especially, the proposed method provides a clear interpretation of combinatorial regulations of gene expression by optimally extracting regulation coordination for a given target gene. Experimental results on both simulation data and real-world microarray data demonstrate the competent performance of discovering regulatory relationships in GRN reconstruction.

1 Introduction

Since genome-wide gene interactions maintain and mediate the activity of cells in various environments, modelling gene regulatory networks (GRNs) plays crucial roles in elucidating cellular mechanisms at the molecular level and especially in explaining the causality of gene expression and discovering regulatory pathways [1, 2]. Many computational methods have been proposed to find the regulatory relationships among genes from various types of omics data [3]. However, reverse engineering of GRN from gene expression data still remains a challenging problem both computationally and biologically.

Difficulties in reconstructing GRNs mainly lie in three aspects: non-linearity of regulatory interactions, imbalance between high dimensionality (gene variables) and small sample, as well as high noise and outliers [4]. A successful GRN inference method needs to deal with these issues well. However, most of existing methods have their pros and cons. For example, many mutual information-based methods previously proposed to measure mutual information between genes have the capability of modelling non-linear dependency between genes [5–7]. However, they fail to detect the combinatory regulations involving two or more transcriptional factors (TFs). To cope with the problem, many Bayesian network (BN)-based methods have been proposed [8–10]. The main idea behind these BN methods is to estimate the joint probability distribution of a target (TG) gene and its TFs based on gene expression profiles. The resulted GRN can be represented as a joint probabilistic graph that essentially accounts for the regulatory combinations of TG genes [11]. Other strategies of GRN reconstruction include Pearson correlation, random forest [12–14] and analysis of variance [15]. To our knowledge, few methods can efficiently deal with complex regulatory patterns, *e.g.* feed-forward loops, and adaptively model GRNs with a reasonably sparse network topology [8].

Recently, regression-based methods have been recognised as powerful players for GRN reconstruction due to their clear interpretation and relationship discovery power [8, 16, 17]. A basic assumption behind them is that the expression levels of a TG gene can be viewed as a linear combination of its TFs. Since real GRNs are sparse, a challenging issue for the regression-based methods is how to infer a relatively small number of TFs from a large number of candidate TFs. Toward this problem, some researchers employed regularised regression strategies, *e.g.* lars [18, 19] and group Lasso [20]. A representative sample is the Gene Expression Modeling using Lasso (GEMULA) method, proposed by Geeven *et al.* [19]. In brief, GEMULA assesses a wide range of candidate regression models and selects the optimal one subject to the prior knowledge of TF–TG gene promoter sequence binding. Haury *et al.* [18] proposed another kind of lars-based method named Trustful Inference of Gene REgulation using Stability Selection (TIGRESS). TIGRESS integrates lars with a stability estimator and employs an area under stability selection curve-based measure to recognise the true TFs for a TG gene. Liu *et al.* [20] proposed a Huber group Lasso method for GRNs reconstruction, which uses Huber loss function instead of squared error loss function as usual to optimise GRN reconstruction. Though many regression-based methods have been proposed, the challenges of the topological complexity of GRN and the imbalance of samples and genes still remain unsolved.

It is mandatory and challenging to choose suitable regularised parameters when applying l_1-regularised regression models. Though model selection criteria such as Bayesian information criterion (BIC) [21] and Akaike information criterion (AIC) [22] can be used, substantial evidences have been witnessed that BIC or AIC often overestimates the number of regulators [23]. The complexity of GRNs also makes it even harder to choose the parameter. To deal with these problems, we here propose a novel method, named Bic-guided sparse regression (BicGSR), which can adaptively model GRNs based on a modified BIC (mBIC)-GSR model. In brief, BicGSR decomposes GRN reconstruction into a series of sub-tasks, and each sub-task recognises the coordinated regulatory mechanism of a single TG gene based on an mBIC-GSR model. By collecting the results from each sub-task, the whole GRN can be finally assembled *via* a fusion strategy. To evaluate the proposed method, we generated simulation datasets and collected two real datasets about *Escherichia coli* and *Saccharomyces cerevisiae* from the Dialogue for Reverse Engineering Assessments and Methods (DREAM) [24], and the experiential results demonstrate the superior performance of the proposed method for GRN inference.

The remainder of this paper is organised as follows. In Section 2, we introduce the theoretical framework of BicGSR based on spare regression models used for modelling the regulations in a GRN, and present a mBIC, which is specific to a sparse regression model for inferring GRN biologically meaningfully. In Section 3, we evaluate the performance of BicGSR on simulation and real datasets and compare it with several previous methods. Section 4 concludes this paper.

2 Methods

In general, regulations between a TG gene and its TFs can be depicted in a sparse linear regression model (or a Lasso regression model) and the sparsity parameter plays important roles in efficiently recovering the regulatory relationships. On the basis of the sparse regression model, the proposed method models GRNs by employing an mBIC criterion to optimise the sparsity parameter as a model selection problem. The mBIC can avoid overestimating TFs for a given TG gene. Fig. 1 presents the flowchart of the proposed method, in which the whole GRN inference problem is decomposed into a series of sub-problems each aiming to identify TFs for a TG gene and then a regulatory matrix $\boldsymbol{R}$ is summarised from all the sub-problems to form the whole GRN.

2.1 Sparse linear regression model for GRNs

A GRN is a collection of genes and their regulators including RNA, protein and complexes of these that interact with each other and with other substances in the cell to govern the cell behaviours. Two most important types of regulators are TFs and microRNAs (miRNAs) [25]: TFs regulate genes at the transcriptional level by binding to proximal or distal regulatory elements within gene promoters, whereas miRNAs act at the post-transcriptional levels of genes [26, 27]. In this paper, we focus on inferring TF–gene or TF–TF interactions based on sparse regression models.

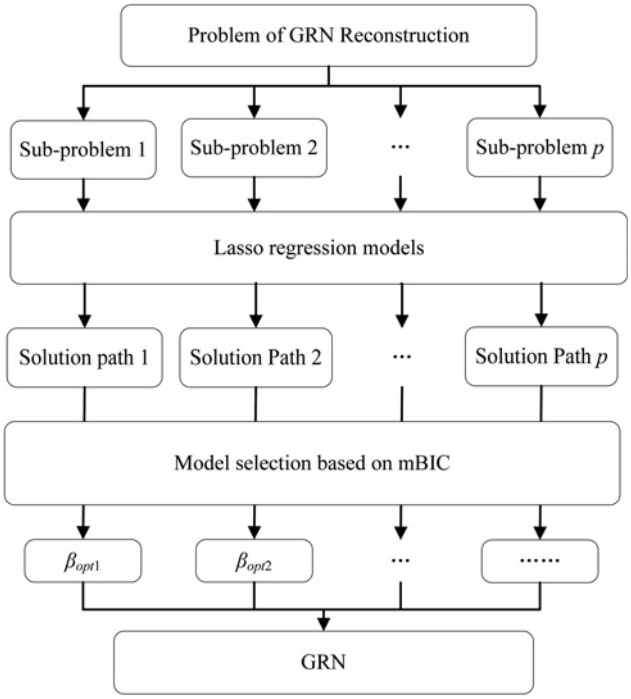

Fig. 1 *Flowchart of BicGSR. BicGSR is a method that uses the 'Lasso' technology integrated with a modified version of BIC (mBIC) to adaptively model the relationships between TG genes and their candidate regulators. First, the problem of GRN inference is decomposed into different sub-problems; second, in each sub-problem the true regulators of a TG gene are recognised based on Lasso regression, and mBIC. Finally, the regression coefficients from all the sub-problems are collected to form a whole gene regulatory matrix for inferring GRN*

Suppose that a vector $\boldsymbol{Y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ represents the expression profile of a TG gene in N samples in a GRN and $\boldsymbol{X}_i = (x_1^i, x_2^i, \ldots, x_N^i)^{\mathrm{T}}$ represents the expression profile of the ith ($i = 1, \ldots, p$) candidate TF in the N samples. The following linear regression model holds

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{X} &= (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p) \\ \boldsymbol{\varepsilon} &= (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^{\mathrm{T}} \sim \tilde{\mathcal{N}}(0, \sigma^2 \boldsymbol{I}_N) \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ represents the unknown regulatory coefficients of the p TFs on the TG gene, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^{\mathrm{T}}$ represents the measurement noise with a $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_N)$ distribution and $\boldsymbol{I}_N$ is an identity matrix of size N.

Generally, the regulatory coefficients can be solved using the following optimisation problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} \|\boldsymbol{\varepsilon}\|_2^2 \tag{2}$$

However, considering that a TG gene is likely regulated by a small number of TFs, most of the coefficients in the vector $\boldsymbol{\beta}$ should be zero in reality. Therefore, one needs to control the sparsity of $\boldsymbol{\beta}$ for an efficient inference of GRN topology. Though the l_0-norm $\|\boldsymbol{\beta}\|_0$ is ideal to be taken as a sparsity penalty term for (2), a non-convex optimisation problem could be raised that is hard to solve in practice [28]. For this reason, we alternatively employ the l_1-norm of $\boldsymbol{\beta}$ as a sparsity penalty term, which has been proven to result in an approximate solution of that by the l_0-norm [29]. Accordingly, the optimisation problem in (2) can be transformed into a l_1-norm-constrained optimisation problem

$$\min_{\beta} \|\boldsymbol{\varepsilon}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_1 \le t \tag{3}$$

where $t \in R^+$ represents the sparsity parameter that controls the sparsity of $\boldsymbol{\beta}$ and consequently the resulted GRN. In the context of this paper, we call the model exhibited in (3) a Lasso regression model. The closer the parameter t is to 0, the fewer the non-zero elements of $\boldsymbol{\beta}$ is. Equation (3) also indicates that different $\boldsymbol{\beta}$ can be obtained with different t parameterised, meaning that $\boldsymbol{\beta}$ can be viewed as a function of the sparsity parameter t. Accordingly, an accurate inference of GRNs depends on the proper choice of t specific to the problem concerned.

2.2 Modified BIC

BIC, which was first introduced by Schwarz [21, 30, 31], is a criterion for model selection in data analysis. Most real-world networks are sparse including GRNs. The original BIC has the tendency to overestimate the degree of model complexity, *i.e.* the number of TFs of a TG gene, in the context of regression-based GRN inference. For improvement, we modify BIC by imposing different prior probabilities on models with different levels of complexity and present an mBIC as follows.

According to Bogdan *et al.* [23] and Claeskens and Hjort [32], maximising the posterior probability of a candidate model M given data $(\boldsymbol{X}, \boldsymbol{Y})$ is equal to maximising

$$\ln(P(\boldsymbol{X}, \boldsymbol{Y}|M)) + \ln(\pi(M)) \tag{4}$$

where $\pi(M)$ represents the prior probability of the model M and $P(\boldsymbol{X}, \boldsymbol{Y}|M)$ represents the conditional probability of the observed data $(\boldsymbol{X}, \boldsymbol{Y})$ given the model M.

Suppose that $g(\boldsymbol{\beta})$ represents the prior probability density function of the parameter $\boldsymbol{\beta}$ and $f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\beta})$ defines the probability density of the data given $\boldsymbol{\beta}$. According to the definition of

conditional probability, we can get

$$P(\boldsymbol{X},\ \boldsymbol{Y}|M)=\int f(\boldsymbol{X},\ \boldsymbol{Y}|\boldsymbol{\beta})g(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} \tag{5}$$

Since $f(\boldsymbol{X},\ \boldsymbol{Y}|\boldsymbol{\beta})g(\boldsymbol{\beta})>0$, (5) can be rewritten as

$$\begin{aligned}P(\boldsymbol{X},\ \boldsymbol{Y}|M)&=\int f(\boldsymbol{X},\ \boldsymbol{Y}|\boldsymbol{\beta})g(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}\\&=\int \exp\ \ln\left[f(\boldsymbol{X},\ \boldsymbol{Y}|\boldsymbol{\beta})g(\boldsymbol{\beta})\right]\mathrm{d}\boldsymbol{\beta}\end{aligned} \tag{6}$$

We denote $Q=\ln\left[f(\boldsymbol{X},\ \boldsymbol{Y}|\boldsymbol{\beta})g(\boldsymbol{\beta})\right]$. Suppose that Q reaches its maximum at $\tilde{\boldsymbol{\beta}}$, then we can approximate Q based on Taylor expansion

$$\begin{aligned}Q\simeq&\ln\left[f(\boldsymbol{X},\ \boldsymbol{Y}|\tilde{\boldsymbol{\beta}})g(\tilde{\boldsymbol{\beta}})\right]+(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\nabla_{\beta}Q|_{\tilde{\beta}}\\&+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{H}_{\tilde{\beta}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\end{aligned} \tag{7}$$

where $\boldsymbol{H}_{\tilde{\beta}}$ is a $|\boldsymbol{\beta}|\times|\boldsymbol{\beta}|$ matrix such that $H_{ij}=\left[\partial^2 Q/\left(\partial\beta_i\partial\beta_j\right)\right]\Big|_{\tilde{\beta}}$ and $|\boldsymbol{\beta}|$ represents the dimension of $\boldsymbol{\beta}$. Since Q reaches its maximum at $\tilde{\boldsymbol{\beta}}$, the Hessian matrix $\boldsymbol{H}_{\tilde{\beta}}$ is negative definite. Then approximate $P(\boldsymbol{X},\ \boldsymbol{Y}|M)$

$$\begin{aligned}P(\boldsymbol{X},\ \boldsymbol{Y}|M)\simeq\int\exp\Big\{&Q|_{\tilde{\beta}}+(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\nabla_{\beta}Q|_{\tilde{\beta}}\\&-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\mathrm{T}}(-\boldsymbol{H}_{\tilde{\beta}})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big\}\,\mathrm{d}\boldsymbol{\beta}\end{aligned}$$

Since Q reaches its maximum at $\tilde{\boldsymbol{\beta}}$, we see that $\nabla_{\beta}Q|_{\tilde{\beta}}=0$. Hence

$$P(\boldsymbol{X},\boldsymbol{Y}|M)\simeq\exp(Q|_{\tilde{\beta}})\int\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\mathrm{T}}(-\boldsymbol{H}_{\tilde{\beta}})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right\}\mathrm{d}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \tag{8}$$

Since the matrix $-\boldsymbol{H}_{\tilde{\beta}}$ is symmetric, we can diagonalise it as $-\boldsymbol{H}_{\tilde{\beta}}=\boldsymbol{S}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{S}$. Let us make a substitution $(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})=\boldsymbol{S}^{\mathrm{T}}\boldsymbol{U}$ to evaluate the integral above. The Jacobian matrix $\boldsymbol{J}_{ij}(\boldsymbol{U})=\partial(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})_i/\partial\boldsymbol{U}_j\Rightarrow\boldsymbol{J}(\boldsymbol{U})=\boldsymbol{S}^{\mathrm{T}}$. Thus $\det\boldsymbol{J}(\boldsymbol{U})=1$ and

$$\begin{aligned}P(\boldsymbol{X},\boldsymbol{Y}|M)&\simeq\exp(Q|_{\tilde{\beta}})\int\exp\left\{-\frac{1}{2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{U}\right\}\left(\det\boldsymbol{J}(\boldsymbol{U})\right)\mathrm{d}\boldsymbol{U}\\&=\exp\left(Q|_{\tilde{\beta}}\right)\int\exp\left\{-\frac{1}{2}\sum_{j=1}^{|\boldsymbol{\beta}|}\lambda_j U_j^2\right\}\mathrm{d}\boldsymbol{U}\end{aligned} \tag{9}$$

where λ_j is the jth eigenvalue of the matrix $-\boldsymbol{H}_{\tilde{\beta}}$. According to Laplace's method [33, 34], we rewrite (9) as follows

$$P(\boldsymbol{X},\boldsymbol{Y}|M)\simeq\exp\left(Q|_{\tilde{\beta}}\right)\prod_{j=1}^{|\boldsymbol{\beta}|}\sqrt{\frac{2\pi}{\lambda_j}}=f(\boldsymbol{X},\boldsymbol{Y}|\tilde{\boldsymbol{\beta}})g(\tilde{\boldsymbol{\beta}})\frac{2\pi^{|\boldsymbol{\beta}|/2}}{\left|-\boldsymbol{H}_{\tilde{\beta}}\right|^{1/2}} \tag{10}$$

and, taking logarithm of (10), we get

$$\begin{aligned}2\ln P(\boldsymbol{X},\boldsymbol{Y}|M)=&\,2\ln f(\boldsymbol{X},\boldsymbol{Y}|\tilde{\boldsymbol{\beta}})+2\ln g(\tilde{\boldsymbol{\beta}})+|\boldsymbol{\beta}|\ln(2\pi)\\&+\ln\left|\left(-\boldsymbol{H}_{\tilde{\beta}}\right)^{-1}\right|\end{aligned} \tag{11}$$

If setting $g(\boldsymbol{\beta})\equiv1$, an uninformative flat prior of density of $\boldsymbol{\beta}$ [35], we have

$$\left|-\boldsymbol{H}_{\tilde{\beta}}\right|=N^{|\beta|}\left|\boldsymbol{I}_{\beta}\right| \tag{12}$$

where N is the number of samples in the data and $\boldsymbol{I}_{\beta}$ is the Fisher information matrix for a single sample. Substituting (12) into (11), we can have

$$\ln P(\boldsymbol{X},\boldsymbol{Y}|M)\sim\ln L(\tilde{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{Y})-\frac{|\boldsymbol{\beta}|}{2}\ln N \tag{13}$$

According to (13), (4) can be rewritten as

$$\ln L(\tilde{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{Y})-\frac{|\boldsymbol{\beta}|}{2}\ln N+\ln(\pi(M)) \tag{14}$$

Considering that gene networks are sparse, we assume that the models with less complexity should have a larger prior probability [23, 36, 37]. The prior probability of M is defined as follows

$$\pi(M)=u^{|\beta|}(1-u)^{p-|\beta|} \tag{15}$$

where u is a small positive constant, p is the maximum dimension of candidate models, *i.e.* the number of candidate TFs. By substituting (15) into (14) and omitting the constant terms, we can get

$$\ln L(\tilde{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{Y})-\frac{|\boldsymbol{\beta}|}{2}\ln N+|\boldsymbol{\beta}|\ln\left(\frac{u}{1-u}\right) \tag{16}$$

The formulation in (16) is called the mBIC estimate for the candidate model M. Equation (16) indicates that mBIC has a more additional term $|\boldsymbol{\beta}|\ln(u/(1-u))$ relative to the original BIC, which can help avoid the overestimation of the regulators for a given TG gene. This term will be negative for $u<0.5$ and positive for $u>0.5$, and degrades mBIC to the basic BIC at $u=0.5$. In this paper, we set $u=3/p$, where p is the total number of TFs, according to Zak *et al.* [36].

2.3 mBIC-guided sparsity parameter selection

Suppose $\boldsymbol{\beta}_t$ to be the solution for (3) corresponding to a given t, the likelihood function of $\boldsymbol{\beta}_t$ can be written as

$$\begin{aligned}L(\boldsymbol{\beta}_t|\boldsymbol{X},\ \boldsymbol{Y})&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\exp\left(-\frac{\left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_t\right\|_2^2}{2\sigma^2}\right)\\&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\exp\left(-\frac{\left\|\boldsymbol{\varepsilon}_t\right\|_2^2}{2\sigma^2}\right)\end{aligned} \tag{17}$$

Let k_t be the number of the non-zero elements in β_t, substituting (17) into (16), we can obtain the mBIC estimate as

$$\mathrm{mBIC}_t=-\frac{1}{2}\left[N\ \ln(2\pi\sigma^2)+\frac{\left\|\boldsymbol{\varepsilon}_t\right\|_2^2}{\sigma^2}+k_t\ \ln N+2k_t\ \ln\left(\frac{1-u}{u}\right)\right] \tag{18}$$

By varying t in a range of $t\in[0,+\infty)$, we can obtain a solution path $\{\boldsymbol{\beta}_t\}$. Along the solution path, an optimal regulatory coefficient can be finally determined as that corresponds to the maximum of mBIC.

3 Results

To evaluate the performance of BicGSR, we generated two types of simulation data and collected two real gene expression datasets. Simulation data I as linear data were generated by revising the

Table 1 Performance comparison of BicGSR and three previous methods on simulation data I

Network size	Methods	AUROC	AUPR
30	BicGSR	**0.994**	**0.989**
	SA's	0.642	0.192
	SSM	0.481	0.091
	naive LASSO	0.950	0.935
	GL	0.490	0.207
	ARACNE	0.520	0.184
300	BicGSR	**0.985**	**0.955**
	SA's	0.541	0.089
	SSM	0.497	0.063
	naive LASSO	0.923	0.755
	GL	0.49	0.020
	ARACNE	0.49	0.020
1500	BicGSR	**0.947**	**0.956**
	SA's	0.473	0.009
	SSM	0.532	0.051
	naive LASSO	0.913	0.820
	GL	0.50	0.004
	ARACNE	0.51	0.005

procedure depicted in [5]: first, creating background networks with the average in-degree of a TG gene set to be around three; second, simulating the expression levels of the regulators by randomly sampling from a Gaussian distribution and the expression levels of the TG genes as a linear combination of their regulators; third, adding Gaussian noise to the expression levels of TG genes so that the signal-to-noise ratio was 10%. Considering three background networks of sizes 30 (10 regulators), 300 (100 regulators) and 1500 (500 regulators), we finally synthesised three expression datasets with 10, 15 and 20 samples, respectively. Simulation data II as non-linear data were downloaded from DREAM4 (http://www.the-dream-project.org/). There are totally five different network topologies used, denoted as NET1-5, each consisting of 100 genes. All of the five networks are transcriptional regulatory sub-networks of *E. coli* and *S. cerevisiae*. All the expression data in the five cases were generated with 100 samples using GeneNetWeaver [38]. In addition, two real gene expression datasets, which are about two model organisms *E. coli.* and *S. cerevisiae*, were downloaded from DREAM5 (http://www.the-dream-project.org/) for real-world evaluation. The former dataset contains 4511 genes, of which 334 are TFs, and 805 chips, whose background regulatory network consists of 2066 experimentally verified TF–TG gene interactions, while the latter consists of the expression profiles of 5950 TG genes and 333 TFs in 536 samples, whose background regulatory network consists of 3940 verified TF–TG interactions.

Two measures, *i.e.* area under receiver operating characteristic (AUROC) curve and area under precision–recall (AUPR) curve, were used for algorithm evaluation. ROC curves plot false positive rates *versus* true positive rates while PR curves plot P against R. Five popular methods including naive Lasso [29], Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) [39], partial correlation-based graphical Lasso (GL) algorithm [40], Salzman and Almudevar (SA) [41] and state space model (SSM) [42] were adopted for comparison. Compared with BicGSR, naive Lasso fixes the sparsity parameter in (3) to be $t_{\text{fixed}} = e^{-1.5}$, ARACNE measures expression similarity between a gene and its candidate TFs based on mutual information for GRN construction, and GL models GRN based on Gaussian graphical model and

Table 3 Performance comparison of BicGSR and the previous methods on *E. coli* and *S. cerevisiae* datasets

Methods	*E. coli*		*S. cerevisiae*	
	AUROC	AUPR	AUROC	AUPR
BicGSR	**0.623**	0.023	**0.549**	0.003
naive Lasso	0.609	0.017	0.519	0.002
GL	0.553	0.018	0.502	0.004
ARACNE	0.572	**0.069**	0.504	**0.018**

employs a graphical Lasso algorithm for the sparse inverse covariance matrix estimation. The SA's method uses a BIC-based scoring procedure in combination with graphical models for modelling GRNs, which derives an independent estimate of the parametric complexity of the model and then modifies the BIC score. SSM relies on SSMs and uses the original BIC to determine the number of hidden variables in SSMs for inferring GRNs. All the experiments were conducted on a personal computer with Intel Xeon 2.13 GHz and 12.0 GB random access memory.

3.1 Simulation data I

Results of BicGSR, naive Lasso, GL and ARACNE on simulation data I are shown in Table 1. From Table 1, we can clearly see that BicGSR outperforms all the previous methods with highest AUROC and highest AUPR in all the data scenarios. Compared with naive Lasso which fixes the sparsity parameter, BicGSR can optimise the parameter to adaptively model GRNs for a TG gene, which makes it more effective and more efficient in varying scenarios. When network size is very large (e.g. 500), BicGSR still performed well but all of the previous methods degraded much, showing that BicGSR is insensitive to network size and can deal with the complexity of regulatory networks better than the previous methods.

3.2 Simulation data II

Table 2 reports the results of BicGSR and the previous methods on simulation data II. Similar to those for simulation data I, BicGSR surpasses naive Lasso with higher AUROC values on the simulation data II, showing the improved ability. BicGSR also obtained higher AUROC values than those by the two other methods, GL and ARACNE, on the datasets NET3-5, confirming the superior performance of BicGSR in reconstructing large regulatory networks. The degraded performance of ARACNE on AUROC scores maybe because of the ignorance of combinatorial regulations compared with BicGSR.

3.3 Real gene expression data

We next evaluated our method BicGSR on two real-world gene expression data about *E. coli* and *S. cerevisiae*. Table 3 reports the AUROC and AUPR results of the four methods on the two datasets. From this table, it can be clearly seen that BicGSR obtained larger AUROC values than those of all the three previous methods, showing that the superior performance of BicGSR on real-world data.

Table 2 Performance of GRN reconstruction algorithms on simulation data II

Methods	NET1		NET2		NET3		NET4		NET5	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
BicGSR	0.611	0.138	0.615	0.125	**0.695**	0.225	**0.678**	0.208	**0.713**	0.215
SA's	0.591	0.219	0.574	0.281	0.647	0.226	0.638	0.239	0.643	0.193
SSM	0.489	0.001	0.492	0.002	0.499	0.001	0.482	0.001	0.463	0.000
naive Lasso	0.559	0.102	0.573	0.107	0.574	0.185	0.600	0.172	0.538	0.123
GL	**0.652**	0.170	**0.636**	0.173	0.674	0.297	0.678	0.240	0.691	0.218
ARACNE	0.612	**0.203**	0.573	**0.187**	0.667	**0.397**	0.644	**0.345**	0.653	**0.363**

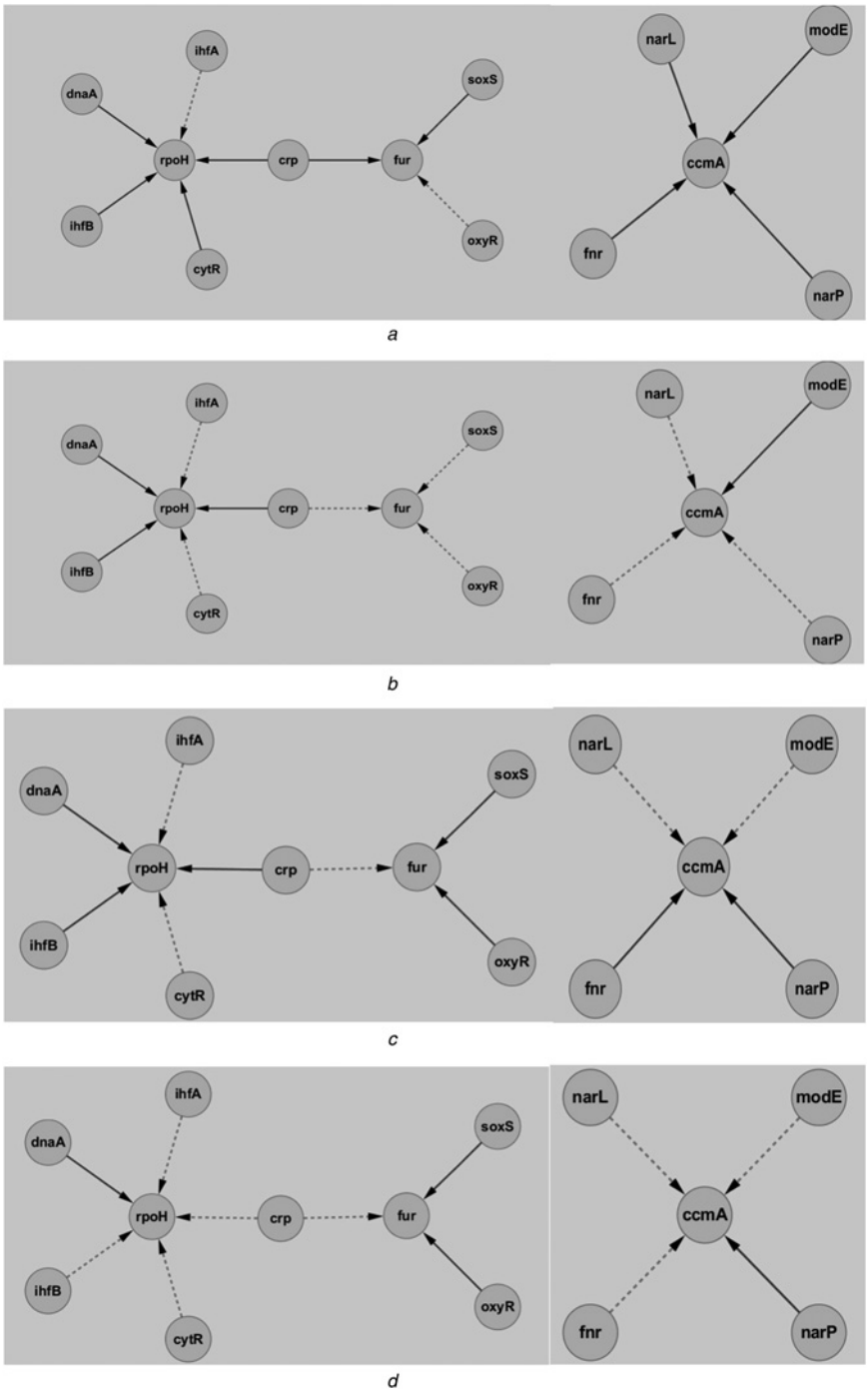

Fig. 2 *Inferred structures of two modules from the E. coli gene network by different methods. Solid arrows represent true regulatory interactions correctly recognised and dashed arrows represent the missing regulatory interactions*

a Network structure estimated by BicGSR
b Network structure estimated by naive Lasso
c Network structure estimated by ARACNE
d Network structure estimated by GL

For further evaluation, we did structure analysis of the GRNs inferred by BicGSR, naive Lasso, ARACNE and GL for the real gene expression data. For fair comparison, all the four GRNs were formed to have the same number of edges by taking different thresholds of regulation coefficients for the different methods. Consider the *E. coli* case and take two hub modules, one simple and another more complex, from the whole gene network. We compare the structures of the two hubs recovered by BicGSR, naive Lasso, ARACNE and GL, as shown in Fig. 2. From Fig. 2, we can clearly see that BicGSR successfully recovered almost all the regulations (10/12), irrespective of the simple or complex module, while the three previous methods, naive Lasso, ARACNE and GL, correctly recognised only 4, 7 and 4 regulations, respectively. The genes in the two modules are correlated with three biological processes: transcription of genetic information from DNA, ferric uptake and cytochrome assembly [43–45]. The product of gene 'rpoH' is an RNA polymerase subunit which is a kind of heat shock sigma factor. The product of gene 'fur' is the ferric uptake regulator protein which functions as a repressor on ferric iron uptake. The product of gene 'ccmA' is the cytochrome c biogenesis protein which is used for cytochrome assembly in bacteria. These results indicate that our method can uncover real GRNs more accurately than the previous methods.

Table 4 Running time of the six methods on simulation data I and II

Networks	Methods					
	BicGSR	SA's	SSM	naive Lasso	GL	ARACNE
Simulation data I						
30	2.962 s	0.733 s	1.356 s	0.157 s	0.351 s	0.020 s
300	47.489 s	31.561 s	1.786 min	2.306 s	4.501 min	0.0392 s
1500	21.753 min	247.147 min	16.945 min	46.766 s	498.331 min	23.430 s
Simulation data II						
NET1	1.010 min	1.463 s	41.381 s	17.940 s	2.668 s	0.034 s
NET2	1.108 min	1.557 s	38.379 s	18.253 s	2.742 s	0.015 s
NET3	58.417 s	1.671 s	37.715 s	16.276 s	3.769 s	0.031 s
NET4	1.075 min	1.592 s	34.852 s	20.012 s	2.986 s	0.028 s
NET5	1.005 min	2.228 s	35.598 s	18.713 s	4.484 s	0.027 s

3.4 Running time comparison of algorithms

We have compared the running time among the six methods on simulation data I–II, as shown in Table 4. Owing to the additional model selection procedure, BicGSR generally spent more running times than the other algorithms, which can be thought of to be a tradeoff with the higher GRN inference accuracies (Tables 1–3). We also noted that for the simulated data I with 1500 genes, the running time of GL sharply increased to 498 min. This maybe due to an exponential increase of network parameters to be estimated in the graphical model with the number of genes increasing.

3.5 Examination of the improvement of BicGSR over naive Lasso

BicGSR can be viewed as a modified version of naive Lasso for GRNs reconstruction. To further assess BicGSR, we examined the improvement of BicGSR over naive Lasso. Fig. 3 shows the PR curves of BicGSR and naive Lasso on the non-linear simulation data II (Net1–Net5, Figs. 3*a*–*e*) and the real gene expression data (*E. coli*, Fig. 3*f*). In pattern recognition, P is defined as the number of true positives over the number of true positives and false positives, and R as the number of true positives over the number of true positives and false negatives. From Fig. 3, it can be found that BicGSR outperforms naive Lasso with an average improvement of 0.04 in AUPR scores on all the six data scenarios, and the smaller the value of R the more the advantage of BicGSR over naive Lasso is. Meanwhile, BicGSR obtained higher P than those of naive Lasso at the same level of R, revealing that BicGSR is more sensitive than naive Lasso for detecting the expression regulators. Specifically, BicGSR had a P of about 1.25 times of naive Lasso when the value of R is among (0, 0.05). Taken altogether, these results indicate that BicGSR is more applicable and more efficient than naive Lasso to various types of gene expression data.

Biologically, genes that are regulated by the same regulator tend to be co-expressed [46–48]. The co-expression of the genes to which a regulator regulates in a regulatory network can provide an indicator about reconstruction accuracy [48]. So, we then verified BicGSR by examining the co-expression of the TG genes regulated by the same regulator based on Pearson correlation. Fig. 4 shows the distributions of the absolute Pearson correlation coefficients obtained by BicGSR and naive Lasso on simulation data II NET1-5 (Figs. 4*a*–*e*) and real-world data *E. coli* (Fig. 4*f*). From Fig. 4, we can clearly see that BicGSR resulted in larger average co-expression levels than those of naive Lasso on all the six data scenarios, suggesting a

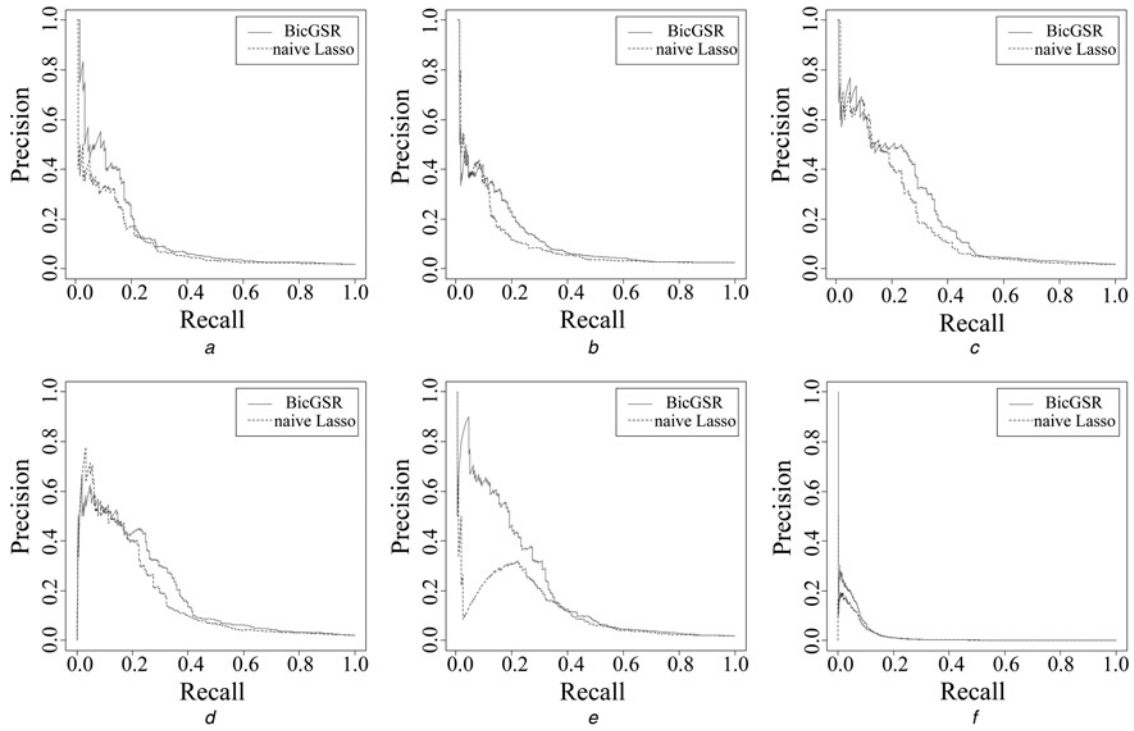

Fig. 3 *PR curves of BicGSR (solid lines) and naive Lasso (dashed lines) for the simulation data II NET1-5 and the real E. coli data*

a–e Simulation data II NET1-5
f Real-world *E. coli* data

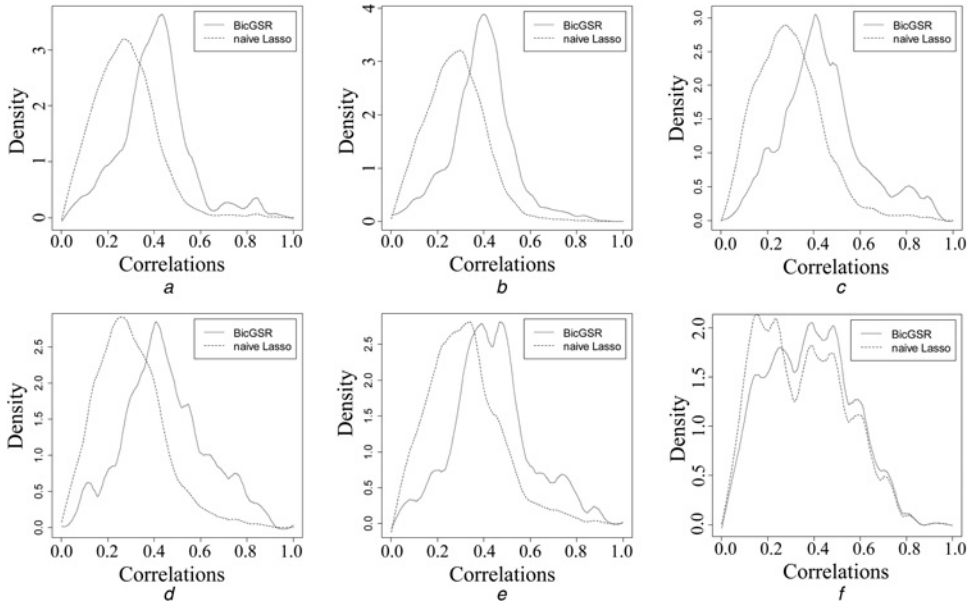

Fig. 4 *Comparison of co-expression distribution of the resulted co-regulated genes between BicGSR and naive Lasso predictions*

a–e Simulation data II NET1-5
f Real-world *E. coli* data

significant improvement of the regulation recovery capability by BicGSR.

3.6 Influence of sparsity parameter t on GRN reconstruction

It has been evidenced that biological network often exhibits low connectivity and high clustering coefficients [49, 50]. In BicGSR, the sparsity parameter t controls the sparseness of the resulted GRN, and the larger the parameter t is the denser the resulted GRN is. In theory, setting t as the full least-square estimate will lead to a complete graph [29].

To demonstrate how the parameter t impacts the accuracy of GRN reconstruction, we examined the AUROC changes of BicGSR with $t \in [e^{-8}, e^{-1}]$ on the simulation and real-world datasets, as shown in Fig. 5. From this figure, it can be seen that for the simulation data II (NET1), AUROC first increases and then drops, and reaches its maximum value (0.619) at $t \simeq e^{-4.5}$, which is very close to that of BicGSR (0.611). For the real expression data (*E. coli*), the maximum AUROC value is also close to that of BicGSR. From the results, a conclusion can be drawn: the optimum value of t changes with data scenarios, and it can be efficiently approximated using BicGSR *via* adaptive modelling.

4 Discussion and conclusions

In this paper, we have proposed a novel method BicGSR for GRN inference. Motivated by the fact that gene networks are sparse, our method uses sparse regression to infer the relationships between genes and their TFs. Specifically, an improved BIC criterion, mBIC, is employed to optimise the sparsity parameter of the sparse regression model, which allows GRNs to be adaptively modelled. We evaluated the proposed method on simulation and real gene expression data and compared with previous methods, and the experimental results show the effectiveness and efficiency of BicGSR in GRN reconstruction.

The sparsity parameter makes an important impact on GRN inference: too small or too large t will degrade the accuracy of GRN reconstruction. Experiments demonstrated that different datasets are corresponding to different optima t, which are often missed by naive Lasso, and so a robust rule, such as mBIC, is

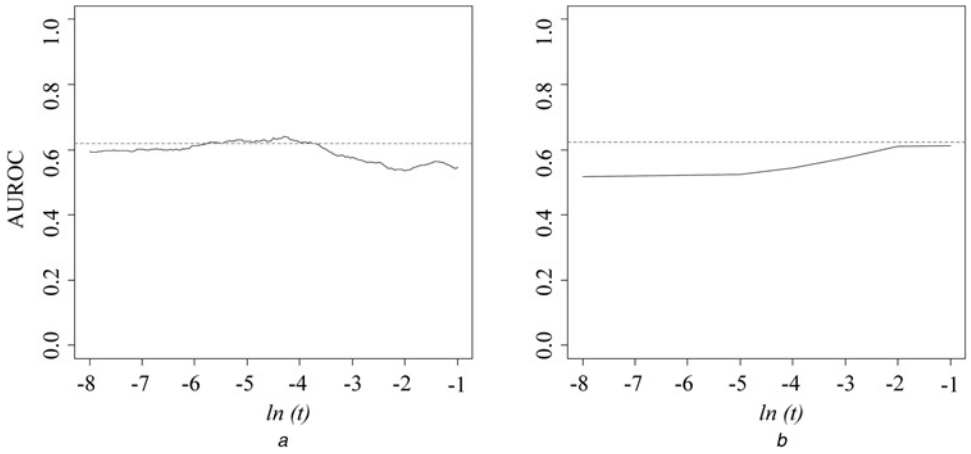

Fig. 5 *Changing curves of AUROC with the sparsity parameter t. Dashed lines indicate the AUROC values of BicGSR with optimal t*

a Results on simulation data II (NET1)
b Results on the *E. coli* data

needed in practice. A strong correlation between GRN reconstruction accuracy and sparsity setting suggests the justification of the sparse regression models in GRN reconstruction. Compared with other modifications of BIC proposed in [23, 36, 37], our mBIC focuses on GRNs inference and made three distinctive modifications: (i) imposing unequal prior probabilities to candidate regression models with different regulatory complexity; (ii) replacing the ordinary linear regression model with a l_1-regularised version in modelling TF–TG gene relationships; (iii) combining BIC with a sparsity parameter for implicitly controlling the structural sparseness of the resulted GRNs. Even so, considering that all the methods have their own advantages and biological systems are very complex, we would like to recommend trying as many methods as possible for GRN inference in practice. On the other hand, we are still aware of the complexity of real gene regulation mechanisms, which maybe more reasonably recovered by non-linear models, and future work will be improving BicGSR by introducing non-linear kernel techniques to address the complexity.

5 Acknowledgments

Funding: This work was supported in part by the National Natural Science Foundation of China (61374181, 61272339, 61402010, 61572372, 41271398); the Anhui Province Natural Science Foundation (1408085MF133); the Shanghai Aerospace Science and Technology Innovation Fund Projects (SAST201425); the state key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) Special Research Funding; and the K.C. Wong education foundation.

6 References

1 Bansal, M., Belcastro, V., Ambesi-Impiombato, A., *et al.*: 'How to infer gene networks from expression profiles', *Mol. Syst. Biol.*, 2007, **3**, (1), p. 78
2 Davidson, E., Levine, M.: 'Gene regulatory networks', *Proc. Natl. Acad. Sci. USA*, 2005, **102**, (14), p. 4935
3 Cai, X.D., Bazerque, J.A., Giannakis, G.B.: 'Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations', *PLoS Comput. Biol.*, 2013, **9**, (8), pp. 796–804
4 Liu, L.Z., Wu, F.X., Zhang, W.J.: 'Reverse engineering of gene regulatory networks from biological data', *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2012, **2**, (5), pp. 365–385
5 Zhang, X.J., Liu, K.Q., Liu, Z.P., *et al.*: 'NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference', *Bioinformatics*, 2013, **29**, (1), pp. 106–113
6 Zhang, X.J., Zhao, X.M., He, K., *et al.*: 'Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information', *Bioinformatics*, 2012, **28**, (1), pp. 98–104
7 Villaverde, A.F., Ross, J., Moran, F., *et al.*: 'MIDER: network inference with mutual information distance and entropy reduction', *PLoS One*, 2014, **9**, (5), p. e96732
8 Marbach, D., Costello, J.C., Kuffner, R., *et al.*: 'Wisdom of crowds for robust gene network inference', *Nat. Methods*, 2012, **9**, (8), pp. 796–804
9 Njah, H., Jamoussi, S.: 'Weighted ensemble learning of Bayesian network for gene regulatory networks', *Neurocomputing*, 2015, **150**, pp. 404–416
10 Zou, M., Conzen, S.D.: 'A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data', *Bioinformatics*, 2005, **21**, (1), pp. 71–79
11 Belcastro, V., Gregoretti, F., Siciliano, V., *et al.*: 'Reverse engineering and analysis of genome-wide gene regulatory networks from gene expression profiles using high-performance computing', *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2012, **9**, (3), pp. 668–678
12 Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., *et al.*: 'Inferring regulatory networks from expression data using tree-based methods', *PLoS One*, 2010, **5**, (9), p. e12776
13 Huynh-Thu, V.A., Sanguinetti, G.: 'Combining tree-based and dynamical systems for the inference of gene regulatory networks', *Bioinformatics*, 2015, **31**, (10), pp. 1614–1622
14 Petralia, F., Wang, P., Yang, J.L., *et al.*: 'Integrative random forest for gene regulatory network inference', *Bioinformatics*, 2015, **31**, (12), pp. 197–205
15 Kuffner, R., Petri, T., Tavakkolkhah, P., *et al.*: 'Inferring gene regulatory networks by ANOVA', *Bioinformatics*, 2012, **28**, (10), pp. 1376–1382
16 Fujita, A., Sato, J.R., Garay-Malpartida, H.M., *et al.*: 'Modeling gene expression regulatory networks with the sparse vector autoregressive model', *BMC Syst. Biol.*, 2007, **1**, (1), p. 39
17 Abegaz, F., Wit, E.: 'Sparse time series chain graphical models for reconstructing genetic networks', *Biostatistics*, 2013, **14**, (3), pp. 586–599
18 Haury, A.C., Mordelet, F., Vera-Licona, P., *et al.*: 'TIGRESS: trustful inference of gene regulation using stability selection', *BMC Syst. Biol.*, 2012, **6**, (1), p. 145
19 Geeven, G., van Kesteren, R.E., Smit, A.B., *et al.*: 'Identification of context-specific gene regulatory networks with GEMULA – gene expression modeling using Lasso', *Bioinformatics*, 2012, **28**, (2), pp. 214–221
20 Liu, L.Z., Wu, F.X., Zhang, W.J.: 'A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets', *BMC Syst. Biol.*, 2014, **8**, (Suppl 3), p. S1
21 Schwarz, G.: 'Estimating the dimension of a model', *Ann. Stat.*, 1978, **6**, (2), pp. 461–464
22 Akaike, H.: 'Information theory and an extension of the maximum likelihood principle', in Parzen, E., Tanabe, K., Kitagawa, G. (EDs.): 'Selected papers of Hirotugu Akaike' (Springer, New York, 1998), pp. 199–213
23 Bogdan, M., Ghosh, J.K., Doerge, R.W.: 'Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci', *Genetics*, 2004, **167**, (2), pp. 989–999
24 Marbach, D., Prill, R.J., Schaffter, T., *et al.*: 'Revealing strengths and weaknesses of methods for gene network inference', *Proc. Natl. Acad. Sci. USA*, 2010, **107**, (14), pp. 6286–6291
25 Ristevski, B.: 'A survey of models for inference of gene regulatory networks', *Nonlinear Anal. Model. Control*, 2013, **18**, (4), pp. 444–465
26 Ristevski, B.: 'Overview of computational approaches for inference of microRNA-mediated and gene regulatory networks', in Memon, A.M. (ED.): 'Advances in computers' (Elsevier Academic Press, San Diego, 2015), pp. 111–145
27 Caffrey, B., Marsico, A.: 'Computational modeling of miRNA biogenesis', in Zazzu, V. (ED.): 'Mathematical models in biology' (Springer, Cham, 2015), pp. 85–98
28 Candes, E.J., Romberg, J.K., Tao, T.: 'Stable signal recovery from incomplete and inaccurate measurements', *Commun. Pure Appl. Math.*, 2006, **59**, (8), pp. 1207–1223
29 Tibshirani, R.: 'Regression shrinkage and selection via the Lasso: a retrospective', *J. R. Stat. Soc. Ser. B. (Stat. Methodol.)*, 2011, **73**, pp. 273–282
30 Liddle, A.R.: 'Information criteria for astrophysical model selection', *Mon. Not. R. Astron. Soc.*, 2007, **377**, (1), pp. L74–L78
31 Lee, E.R., Noh, H., Park, B.U.: 'Model selection via Bayesian information criterion for quantile regression models', *J. Am. Stat. Assoc.*, 2014, **109**, (505), pp. 216–229
32 Claeskens, G., Hjort, N.L.: 'Model selection and model averaging' (Cambridge University Press Cambridge, Cambridge, 2008)
33 MacKay, D.J.C.: 'Information theory, inference, and learning algorithms' (Cambridge University Press, Cambridge, 2003)
34 Lopez, J.L., Pagola, P., Sinusia, E.P.: 'A simplification of Laplace's method: applications to the gamma function and gauss hypergeometric function', *J. Approx. Theory*, 2009, **161**, (1), pp. 280–291
35 Yuan, M., Lin, Y.: 'Model selection and estimation in regression with grouped variables', *J. R. Stat. Soc. Ser. B. (Stat. Methodol.)*, 2006, **68**, pp. 49–67
36 Zak, M., Baierl, A., Bogdan, M., *et al.*: 'Locating multiple interacting quantitative trait Loci using rank-based model selection', *Genetics*, 2007, **176**, (3), pp. 1845–1854
37 Frommlet, F., Ruhaltinger, F., Twarog, P., *et al.*: 'Modified versions of Bayesian information criterion for genome-wide association studies', *Comput. Stat. Data Anal.*, 2012, **56**, (5), pp. 1038–1051
38 Marbach, D., Schaffter, T., Mattiussi, C., *et al.*: 'Generating realistic in silico gene networks for performance assessment of reverse engineering methods', J. Comput. Biol., 2009, 16, (2), pp. 229–239
39 Margolin, A.A., Nemenman, I., Basso, K., *et al.*: 'ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context', *BMC Bioinf.*, 2006, **7**, (Suppl 1), p. S7
40 Menendez, P., Kourmpetis, Y.A.I., ter Braak, C.J.F., *et al.*: 'Gene regulatory networks from multifactorial perturbations using graphical Lasso: application to the DREAM4 challenge', *PLoS One*, 2010, **5**, (12), p. e14147
41 Salzman, P., Almudevar, A.: 'Using complexity for the estimation of Bayesian networks', *Stat. Appl. Genet. Mol. Biol.*, 2006, **5**, (1), pp. 1–23
42 Wu, X., Li, P., Wang, N., *et al.*: 'State space model with hidden variables for reconstruction of gene regulatory networks', *BMC Syst. Biol.*, 2011, **5**, (Suppl 3), p. S3
43 Erickson, J.W., Vaughn, V., Walter, W.A., *et al.*: 'Regulation of the promoters and transcripts of rpoH, the Escherichia coli heat-shock regulatory gene', *Genes Dev.*, 1987, **1**, (5), pp. 419–432
44 Prince, R.W., Storey, D.G., Vasil, A.I., *et al.*: 'Regulation of Toxa and Rega by the Escherichia coli fur gene and identification of a fur homolog in Pseudomonas aeruginosa Pa103 and Pa01', *Mol. Microbiol.*, 1991, **5**, (11), pp. 2823–2831
45 Thonymeyer, L., Fischer, F., Kunzler, P., *et al.*: 'Escherichia coli genes required for cytochrome-C maturation', *J. Bacteriol.*, 1995, **177**, (15), pp. 4321–4326
46 Shalgi, R., Lieber, D., Oren, M., *et al.*: 'Global and local architecture of the mammalian microRNA-transcription factor regulatory network', *PLoS Comput. Biol.*, 2007, **3**, (7), pp. 1291–1304
47 Su, N.F., Wang, Y.F., Qian, M.P., *et al.*: 'Combinatorial regulation of transcription factors and microRNAs', *BMC Syst. Biol.*, 2010, **4**, (1), p. 150
48 Su, N.F., Dai, D., Deng, C., *et al.*: 'Using graphical adaptive Lasso approach to construct transcription factor and microRNA's combinatorial regulatory network in breast cancer', *IET Syst. Biol.*, 2014, **8**, (3), pp. 87–95
49 Thieffry, D., Huerta, A.M., Perez-Rueda, E., *et al.*: 'From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli', *Bioessays*, 1998, **20**, (5), pp. 433–440
50 Liu, L.Z., Wu, F.X., Zhang, W.J.: 'Properties of sparse penalties on inferring gene regulatory networks from time-course gene expression data', *IET Syst. Biol.*, 2015, **9**, (1), pp. 16–24