BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 106–113
doi:10.1093/bioinformatics/bts619

Systems biology

Advance Access publication October 18, 2012

NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference

Xiujun Zhang[1,2,3], Keqin Liu[1,2,3], Zhi-Ping Liu[4], Béatrice Duval[3], Jean-Michel Richer[3], Xing-Ming Zhao[5,*], Jin-Kao Hao[3,*] and Luonan Chen[1,4,*]

[1]Institute of Systems Biology, Shanghai University, Shanghai 200444, China, [2]School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China, [3]LERIA, University of Angers, Angers 49045, France, [4]Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China and [5]Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Reconstruction of gene regulatory networks (GRNs) is of utmost interest to biologists and is vital for understanding the complex regulatory mechanisms within the cell. Despite various methods developed for reconstruction of GRNs from gene expression profiles, they are notorious for high false positive rate owing to the noise inherited in the data, especially for the dataset with a large number of genes but a small number of samples.

Results: In this work, we present a novel method, namely NARROMI, to improve the accuracy of GRN inference by combining ordinary differential equation-based recursive optimization (RO) and information theory-based mutual information (MI). In the proposed algorithm, the noisy regulations with low pairwise correlations are first removed by using MI, and the redundant regulations from indirect regulators are further excluded by RO to improve the accuracy of inferred GRNs. In particular, the RO step can help to determine regulatory directions without prior knowledge of regulators. The results on benchmark datasets from Dialogue for Reverse Engineering Assessments and Methods challenge and experimentally determined GRN of *Escherichia coli* show that NARROMI significantly outperforms other popular methods in terms of false positive rates and accuracy.

Availability: All the source data and code are available at: http://csb.shu.edu.cn/narromi.htm.

Contact: lnchen@sibs.ac.cn, hao@info.univ-angers.fr and zhaoxingming@gmail.com.

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

Received on May 27, 2012; revised on September 18, 2012; accepted on October 14, 2012

1 INTRODUCTION

A major issue in systems biology is to construct and understand the gene regulatory networks (GRNs), which explicitly characterize regulatory processes in the cell (Basso *et al.*, 2005). The development of high throughput technologies has produced tremendous amounts of gene expression data, which provide insights into the underlying regulatory mechanism of cellular machines (Hughes *et al.*, 2000). The reconstruction or 'reverse engineering' of GRNs, which aims to dissect the underlying network of gene-gene interactions from the measurement of gene expression, is still a challenging task (Margolin *et al.*, 2006a). For this reason, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) project was established to encourage researchers to develop new efficient computation methods to infer robust GRNs (Marbach *et al.*, 2010).

Recently, various approaches have been developed to infer GRNs from gene expression data with the motivation of improving the accuracy and scalability of network inference (Smet and Marchal, 2010). In general, these GRN inference methods fall into two categories, namely model-based approaches and machine learning-based approaches (Bansal *et al.*, 2007; Li *et al.*, 2011). For the model-based methods, chemical reaction of transcription and translation, as well as other cellular processes are described as linear or non-linear differential equations, in which the parameters represent the regulation strengths of the regulators. Representative algorithms in this category include multiple linear regression (Cantone *et al.*, 2009; Gardner *et al.*, 2003; Honkela *et al.*, 2010; Tibshirani, 1996), singular value decomposition method (di Bernardo *et al.*, 2005; Yeung *et al.*, 2002), network component analysis (Chen *et al.*, 2010; Liao *et al.*, 2003) and linear programming (LP) (Wang *et al.*, 2006). For the machine learning-based approaches, the network is inferred through measuring the dependences or causalities between transcriptional factors (TFs) and target genes (Küffner *et al.*, 2012). Popular methods in this category include partial correlation coefficient (De la Fuente *et al.*, 2004, Saito *et al.*, 2011), Bayesian network analysis (Li *et al.*, 2011; Yeung *et al.*, 2011), mutual information (MI) (Basso *et al.*, 2005; Belcastro *et al.*, 2011; Faith *et al.*, 2007; Margolin *et al.*, 2006b; Modi *et al.*, 2011) and conditional mutual information (CMI) (Sumazin *et al.*, 2011; Zhang *et al.*, 2012).

As one of the most popular methods, MI has been widely used to construct GRNs because it provides a natural generalization of correlation owing to its capability of characterizing non-linear dependency (Brunel *et al.*, 2010). Furthermore, MI is able to deal with thousands of variables (genes) in the presence of a limited number of samples (Meyer *et al.*, 2008). Despite these advantages, MI fails to distinguish indirect regulators from direct ones, i.e. it tends to overestimate the number of regulators

*To whom correspondence should be addressed.

targeting the gene. In a GRN, the indirect regulations are the main source of false positives. Although some methods have been developed recently to remove these redundant indirect regulations, such as CMI (Frenzel and Pompe, 2007) and CMI-based path consistensy algorithm (PCA-CMI) (Zhang *et al*., 2012), the high computational complexity makes them infeasible while calculating the high order MIs. Another limitation of MI is that it only describes the correlation between two genes but is unable to determine the regulatory directions.

Different from the methods based on information theory, the model-based methods have the advantage of describing the regulatory dynamics and detecting the direction of regulations (Marbach *et al*., 2010). Furthermore, prior information, such as experimentally verified regulations, can be easily included in these models to improve the accuracy of network inference (Christley *et al*., 2009). Moreover, model-based methods are found useful to remove possible redundant indirect regulations by forcing sparseness on the model (Hurley *et al*., 2011; MacNeil and Walhout, 2011), such as the shortcut removing technique that has been proved to be efficient in GRN inference (Markowetz *et al*., 2007; Wagner, 2001).

In this work, we propose a novel method, namely NARROMI, to improve the accuracy of GRN inference from gene expression data using a noise and redundancy (NAR)-reduction technology by combining ordinary differential equation (ODE)-based recursive optimization (RO) and information-theory based MI, and therefore it has the advantages of both model-based and machine learning-based methods. Specifically, the noisy regulations and partial significant indirect regulations can be firstly deleted and filtered using MI as a measure of dependence, and the redundant (indirect) regulations are subsequently removed gradually by the RO algorithm, thereby reducing both false positives and false negatives. In addition, our method can determine regulatory directions without prior information of regulators. The results on simulation datasets from DREAM challenge (Marbach *et al*., 2010) and experimentally confirmed network (Gama-Castro *et al*., 2011) in *Escherichia coli* with real gene expression data (Faith *et al*., 2008) show that our method significantly outperforms other popular methods in terms of false positives and accuracy.

2 METHODS

In general, the transcription process can be described by a mathematical model with differential equations based on mass action kinetics and Michaelis–Menten kinetics. However, the noise inherited in the data can decrease the performance of these models. Therefore, we present a new method NARROMI, which first reduces noisy regulations with MI and then uses RO technique to reduce redundant and indirect regulations gradually in the optimization model. The details of NARROMI can be found below.

2.1 Mathematic model of transcription procedure

During transcription, TF(s) binds to DNA sequences so as to recruit RNA polymerase II onto promoter region of DNA to initiate the transcription procedure (Sun *et al*., 2006; Wang *et al*., 2009), which can described as follows.

$$\mathrm{DNA}_i + \beta_i^1 \cdot \mathrm{TF}_1 + \cdots + \beta_i^c \cdot \mathrm{TF}_c + \mathrm{RPII} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII}, \quad (1)$$

where c is the number of TFs that are regulators of gene i ($i=1,2,\ldots,n$), the stoichiometric coefficient β_i^j, $j=1,2,\ldots,c$, represents the effective abundance of TF_j involved in the regulation of gene i, DNA_i is the sequence of gene i, and k_1, k_{-1}, respectively, denotes the rate constant of forward reaction and reverse action. $\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII}$ denotes the immobilized complex formed by TFs and RNA polymerase II. After transcription initiation, mRNAs are synthesized through the following irreversible reaction with rate constant k_2

$$\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII} \xrightarrow{k_2} \mathrm{mRNA}_i, \quad (2)$$

At the translation level, mRNAs are translated into proteins

$$\mathrm{mRNA}_i \xrightarrow{k_3} \mathrm{Protein}_i, \quad (3)$$

where k_s, $s=-1$, 1, 2, 3, are the rate constants of reactions. According to the mass action law, the concentration changes of $\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}$ RPII and mRNA_i can be described with following differential equations.

$$\mathrm{d}[\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII}]/\mathrm{d}t = k_1[\mathrm{DNA}_i][\mathrm{RPII}]\prod_{j=1}^{c}[\mathrm{TF}_j]^{\beta_i^j} - (k_{-1}+k_2)[\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII}], \quad (4)$$

$$\mathrm{d}[\mathrm{mRNA}_i]/\mathrm{d}t = k_2[\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII}] - k_3[\mathrm{mRNA}_i]\,. \quad (5)$$

Assuming that (4) and (5) quickly reach an equilibrium state, i.e. $\mathrm{d}[\mathrm{DNA}_i(\mathrm{TF}_1)_{\beta_i^1} \cdots (\mathrm{TF}_c)_{\beta_i^c}\mathrm{RPII}]/\mathrm{d}t = 0$ and $\mathrm{d}[\mathrm{mRNA}_i]/\mathrm{d}t = 0$, we can get $[\mathrm{mRNA}_i] \propto \prod_{j=1}^{c}[\mathrm{TF}_j]^{\beta_i^j}$. Let $y_i(t) = [\mathrm{mRNA}_i]_t$, $A_j(t) = [\mathrm{TF}_j]_t$, $t = 1, 2, \ldots, m$, and $y_i(t) \propto \prod_{j=1}^{c} A_j(t)^{\beta_i^j}$. With a logarithm transformation, the above model can be described as a log-linear model $\log(y_i(t)/y_i(0)) = \beta_i^j \sum_{j=1}^{c} \log\big(A_j(t)/A_j(0)\big)$. Let $y_i^{(t)} = \log[y_i(t)/y_i(0)]$ and $X_j^{(t)} = \log\big(A_j(t)/A_j(0)\big)$ where $y_i^{(t)}$ represents the expression level of gene i at time t, and $X_j^{(t)}$ represents the activity of TF j at time t, we get a linear model below by dropping t for simplicity.

$$y_i = \beta_i X, i = 1, 2, \ldots, n,$$

where $y_i = (y_i^1, y_i^2, \ldots, y_i^m)$, $\beta_i = (\beta_i^1, \beta_i^2, \ldots, \beta_i^c)$, $X = (x_j^t)_{c\times m}$, n is the number of target genes, c is the number of TFs and m is number of samples.

2.2 RO

For a target gene with expression level y, we intend to define a regulation matrix β that fits well with the experimental data. β can be resolved by minimizing the error between inferred and observed expressions, i.e.

$$\min_{\beta} |y - \beta X| + \lambda|\beta|, \quad (6)$$

where X is the expression matrix of candidate TFs, and λ is a positive parameter that balances the error and sparse term in the objective function.

As TF activity can be approximated by the expression level of the gene encoding the TF, we suppose the gene expression level as the TF activity here. The model (6) is equivalent to

$$\min_{\beta^i} \sum_{i=1}^{m} \left|y^i - \sum_{j=1}^{c} \beta^j x_j^i\right| + \lambda \sum_{j=1}^{c} |\beta^j| \quad (7)$$

Let

$$u_i + v_i = \left|y^i - \sum_{j=1}^{c} \beta^j x_j^i\right|,\ u_i - v_i = y^i - \sum_{j=1}^{c} \beta^j x_j^i,$$
$$\xi_j + \eta_j = |\beta^j|,\ \xi_j - \eta_j = \beta^j,$$
$$i = 1, 2, \ldots, m, j = 1, 2, \ldots, c,$$

where $u_i, v_i, \xi_j, \eta_j \geq 0$. Then model (7) can be written as a standard LP model as follows.

$$\begin{aligned} \min_{u_i, v_i, \xi_j, \eta_j} \quad & \sum_{i=1}^{m} (u_i + v_i) + \lambda \sum_{j=1}^{c} (\xi_j + \eta_j) \\ s.t. \quad & u_i - v_i = y^i - \sum_{j=1}^{c} (\xi_j - \eta_j) x_j^i, \\ & u_i, v_i, \xi_j, \eta_j \geq 0. \end{aligned} \tag{8}$$

The above LP model (8) can be solved efficiently by any LP software such as GLPK LP/MIP solver (Wang *et al.*, 2009). The parameter λ in model (8) is a positive value used to balance fitting and sparseness since GRNs are known to be sparse. The method of inferring GRNs simply by model (8) is called LP method in this article. LP is different from regression model-based LASSO, which reaches a least squares solution (Geeven *et al.*, 2012).

Although the sparseness can be controlled by the parameter λ in model (8) to some extent, it is non-trivial to obtain an optimal network structure owing to the noise in the expression data. To ensure sparseness and reduce false positives, we set the variables with low regulation strengths to zero and re-estimate only the non-zero variables using model (8) in the next step. In this way, the accuracy of the network inferred by the second optimization step is improved with the strengthened sparsity by removing NAR regulations. The above procedure is repeated until there are no more non-zero variables. As this technique is composed of a series of optimization procedures, we call it RO. The technique to infer GRNs through solving model (8) recursively is called RO method in this article.

2.3 MI

The gene expression data can be described as vectors, in which the elements denote the expression values of genes under different conditions (samples). MI measuring the dependency between two genes X and Y can be defined as below (Altay and Emmert-Streib, 2010).

$$I(X, Y) = -\sum_{x \in X,\ y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{9}$$

With the widely adopted hypothesis of Gaussian distribution for gene expression data, the formula (9) can be easily calculated using the following equivalent formula (Zhang *et al.*, 2012).

$$I(X, Y) = \frac{1}{2} \log \frac{|C(X)| \cdot |C(Y)|}{|C(X, Y)|}, \tag{10}$$

where C is the covariance matrix of variables, and $|C|$ is the determinant of matrix C. If genes X and Y are independent of each other, $I(X, Y) = 0$.

2.4 NARROMI algorithm

Figure 1 depicts the schematic view of our NARROMI method. The details are addressed as follows.

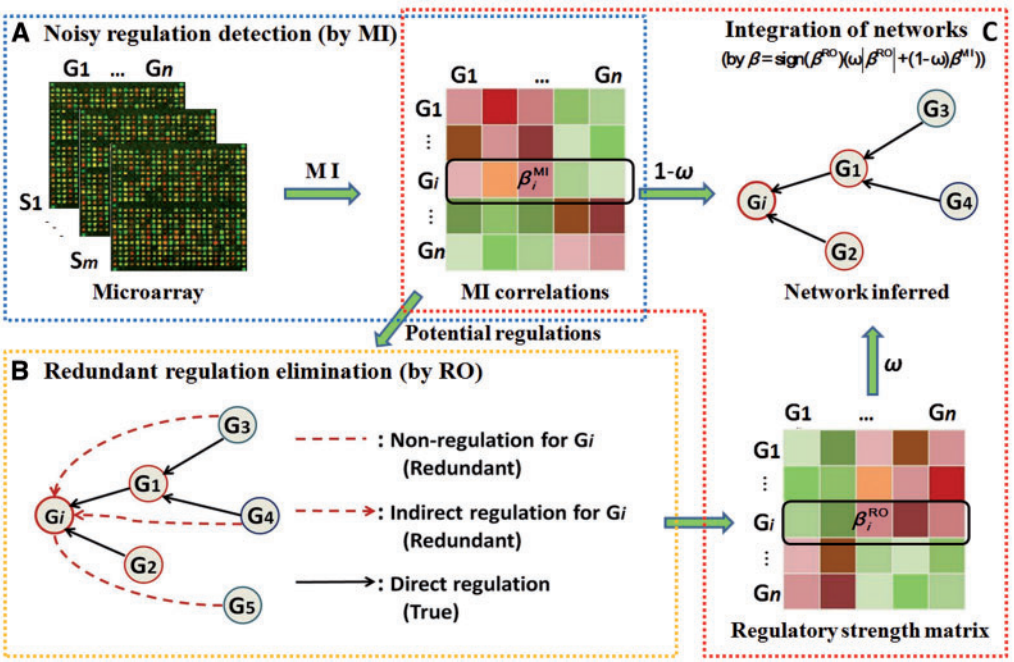

Fig. 1. Overview of the NARROMI method. (**A**) Noisy regulations are detected by MI. In the graph, the genes are labelled with **G**. MI correlations of gene–gene pairs are computed using formula (10) from the microarray data. For each target gene, the candidate regulators with high MI values are selected for the further network inference using RO algorithm. (**B**) The redundant regulations for each target will be removed gradually by RO algorithm, and real regulations will be kept until algorithm finished. In the graph, all the lines (dashed and solid) denote the high MI correlations between two nodes. Moreover, the dashed lines without arrows denote non-regulations (redundant), the dashed arrows denote the indirect regulations (redundant) and the solid arrows denote the true regulations. Take target **G*i*** as an example, edge **G5–G*i*** is a non-regulation, and edges **G3–G*i***, **G4–G*i*** are two indirect regulations. All of them are redundant and will be removed by RO algorithm. Edges **G1–G*i*** and **G2–G*i*** are two true/real direct regulations for **G*i*** and will be kept until RO algorithm terminated. (**C**) The network structure is decided by an integrative technique, in which the regulatory strengths inferred from RO algorithm and MI correlations are mediated by a linear formula to complement the linear and non-linear correlation between regulator and target

Algorithm (NARROMI)

Step 1: Noisy regulation detection

As is well known, the co-expressed genes are more possible to be regulated each other. For a target gene, the genes with high MI scores are co-expressed genes and more possible the true regulators. The MIs between all possible gene–gene (or regulator-target) pairs are firstly computed by formula (10). Given a threshold parameter θ for deciding independence between variables, the regulations with MIs below the threshold are regarded as noisy regulations and removed from further analysis (Fig. 1A). Note that in our algorithm, we select the TFs for each target gene, rather than select targets for each TF. In addition, those regulations with very large MIs will be kept as putative real regulations and integrated into the final GRN. If TFs are known in advance, the possible regulations of genes by these TFs are inferred. Otherwise, all genes are regarded as possible regulators, and all possible regulations are detected.

Step 2: Redundant regulation elimination

In this step, only the candidate regulators selected from Step 1 are used in the subsequent optimization for further inference of the network structure. In this way, the redundant regulators will be removed from the candidate regulators gradually using the RO algorithm for each gene. Usually, a gene is targeted by more than one TF. The regression model in RO can detect these combinatorial regulations simultaneously by selecting those regulators with high coefficients, while the regulators with low regulatory coefficients are eliminated.

Figure 1B gives the overview of the above procedure of RO, where nodes represent target or regulator genes, and arrows represent regulations from regulators to target genes. In more detail, the regulations are divided into three classes: non-regulations, indirect regulations and direct (true) regulations. The first two classes are redundant regulations, which can be removed by RO algorithm, and the real regulations will be kept until algorithm finished. Take target **G*i*** as an example, edge **G**5–**G*i*** is a non-regulation, and edges **R**3–**G*i***, **R**4–**G*i*** are two indirect regulations. All of them are redundant and will be removed by RO algorithm. Finally, edges **G**1–**G*i*** and **G**2–**G*i*** are kept as the two potential direct regulations of target **G*i***.

Step 3: Integration of networks

To combine the linear and non-linear correlations between regulators and targets, the regulatory strengths inferred from RO algorithm, and the MI correlations are integrated in a linear combination way as follows.

$$\beta = \mathrm{sign}(\beta^{\mathrm{RO}})\left(\omega\left|\beta^{\mathrm{RO}}\right| + (1-\omega)\beta^{\mathrm{MI}}\right), \tag{11}$$

where β^{MI} is the MI correlation, which is positive, β^{RO} is the regulatory strength (positive or negative) inferred by RO algorithm, $\mathrm{sign}(\beta^{\mathrm{RO}})$ is the sign ($\pm$) of β^{RO}, $\left|\beta^{\mathrm{RO}}\right|$ is the absolute of β^{RO} and parameter ω is the weighting coefficient for MI and RO (Fig. 1C). The final regulatory strength is decided by the weight parameter β, and the network topology is then determined (Treviño *et al.*, 2012) (Supplementary Material).

3 RESULTS

To validate our method, NARROMI was applied to several simulation datasets and a real gene expression dataset. As for simulation data, the method was tested on the simulated benchmark GRNs with synthetic linear expression data and the widely used reference network in *Yeast* with synthetic non-linear expression data from DREAM3 challenge (Marbach *et al.*, 2010). As for real gene expression data, we applied our method to the experiment confirmed network (Gama-Castro *et al.*, 2011) in *E. coli* with real gene expression data (Faith *et al.*, 2008).

The predictive results were evaluated by following measures, i.e. sensitivity or true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), accuracy (ACC) and Matthews Coefficient Constant (MCC). Mathematically, they are defined as

$$\begin{aligned} \mathrm{TPR} &= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}),\\ \mathrm{FPR} &= \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}),\\ \mathrm{PPV} &= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}),\\ \mathrm{ACC} &= (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}), \end{aligned}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where TP, FP, TN and FN are the numbers of true positives, false positives, true negatives and false negatives, respectively. TPR and FPR are also used to plot the receiver operating characteristic (ROC) curves, and the area under ROC curve (AUC) is calculated.

To evaluate the performance of NARROMI, we compared it with several popular methods including LP, RO, regression model-based LASSO (Geeven *et al.*, 2012), MI-based ARACNE (Margolin *et al.*, 2006b) and random forest-based GENIE3 (Huynh-Thu *et al.*, 2010), where the two alternatives with parameters '*sqrt*' and '*all*' in GENIE3 were considered here, as they performed best in the DREAM challenges (Supplementary Material). For all the methods in comparison, the default values of parameters were set to run the algorithms. For example, the regularization parameter λ of methods LP, RO and NARROMI was set to 1; the ensemble parameter of method GENIE3 was set to 1000; the threshold of MI filtering in method NARROMI was set to 0.05.

3.1 Evaluation on simulation data

In this section, we show computational results of our method on two types of simulation datasets, i.e. linear and non-linear expression dataset. For linear expression data, we generated the benchmark networks and related expression datasets using the simple conventional method in statistics. For non-linear expression data, we performed our method on the widely used reference network in *Yeast* with synthetic non-linear expression data from DREAM3 challenge (Marbach *et al.*, 2010).

3.1.1 Artificial linear expression data For linear expression data, the benchmark networks and corresponding gene expression data were generated according to the network size, degree of sparseness and rate of noise. First, the benchmark network was generated according to the degree of network, number of target genes and number of candidate regulators. Second, the expression data of the regulators were generated randomly using the Gaussian distribution function. Third, the expression data of the target genes were generated by linear combination with the expression data of regulators. Last, the Gaussian noise was added to the expression data of target genes. In the experiments, we generated different networks with 10, 100, 500, 1000 and 5000 regulators and expression datasets under 5, 10, 15, 20 and 25 samples, respectively (Table 1). The average input degree for each target was set to $\sim$2, and the noise rate was set to 10%.

Figure 2 shows the ROC curves by different methods on datasets of sizes 10, 100 and 1000. For the networks with sizes 500 and 5000, the ROC curves are showed in

Supplementary Figure S1. From the figures, we can clearly see that the performance of our NARROMI method is superior to other methods with AUC score ~0.90. Table 1 summarizes the results obtained by different methods with respect to distinct performance indices. From Table 1, we can see that RO performs better than LP and LASSO, both ARACNE and GENIE3 have good performance on large-scale networks and our method NARROMI performs best with the highest AUC values of 0.992, 0.928 and 0.937 on all three datasets. The results show that NARROMI is more robust than other methods on different network sizes. When the network size is large enough with thousands of genes, the accuracy of NARROMI is still high enough, whereas the large network size degrades the performance of other methods significantly. For the networks with sizes 500 and 5000, the results with respect to other performance indexes are given in Supplementary Tables S1 and S2, where NARROMI performs best.

Table 1. Comparison on networks with sizes 10, 100 and 1000

Method	TPR	FPR	PPV	ACC	MCC	AUC
Size 10						
LASSO	0.667	0.122	0.546	0.840	0.505	0.776
LP	0.778	0.024	0.875	0.940	0.789	0.859
RO	0.778	0.073	0.700	0.900	0.677	0.886
ARACNE	0.556	0.146	0.456	0.800	0.380	0.748
GENIE3_FR_sqrt	0.444	0.171	0.364	0.760	0.254	0.748
GENIE3_FR_all	0.333	0.073	0.500	0.820	0.308	0.753
NARROMI	**1.000**	**0.024**	**0.900**	**0.980**	**0.937**	**0.992**
Size 100						
LASSO	0.474	0.010	0.474	0.980	0.464	0.770
LP	0.474	0.010	0.474	0.980	0.464	0.817
RO	0.474	0.010	0.474	0.980	0.464	0.827
ARACNE	0.421	0.042	0.163	0.948	0.239	0.887
GENIE3_FR_sqrt	0.263	0.015	0.250	0.971	0.241	0.809
GENIE3_FR_all	0.263	0.018	0.217	0.968	0.223	0.778
NARROMI	**0.526**	**0.008**	**0.556**	**0.983**	**0.532**	**0.928**
Size 1000						
LASSO	0.631	0.0006	0.667	0.9987	0.648	0.839
LP	0.684	0.0008	0.619	0.9986	0.650	0.812
RO	**0.684**	0.0007	0.650	0.9987	0.666	0.867
ARACNE	0.211	**0.0006**	0.400	0.9979	0.289	0.929
GENIE3_FR_sqrt	0.316	0.0066	0.061	0.9895	0.136	0.888
GENIE3_FR_all	0.263	0.0092	0.070	0.9920	0.133	0.874
NARROMI	0.578	0.0007	**0.650**	**0.9987**	**0.666**	**0.937**

The best performer for the relative item is noted in bold.
LASSO, regression-based method; LP, linear programming-based method; RO, recursive optimization-based method; ARACNE, MI-based method; GENIE3, random forests-based methods; NARROMI, method based on RO and MI.

3.1.2 Artificial non-linear expression data For non-linear expression data, the widely used benchmark networks with expression datasets from DREAM challenge were adopted here to evaluate our method. The gold standard networks were generated with the non-linear ODE systems in which the network structures were determined with detailed dynamics of both transcriptional and translational processes (Schaffter *et al.*, 2011). In this work, the DREAM3 datasets about *Yeast* knock-out genes with sizes 10 and 50 were used (Marbach *et al.*, 2010).

Firstly, NARROMI was applied to the *Yeast* gene expression data with network sizes 10 and 10 samples. Figure 3 shows the structure of our inferred network and the ROC curves obtained by different methods. Figure 3A shows the true network with 10 genes and 10 edges, and Figure 3B shows the network inferred by NARROMI. From the figure, we can see that most edges were recovered by NARROMI, although some regulations were missed, such as G6-G4 and G9-G4 with dashed lines. In addition, five of eight inferred edges (62.5%) were detected correctly with respect to regulatory directions. This indicates that NARROMI can detect most regulatory directions without the information of TFs. The comparison of NARROMI with other methods was shown in Figure 3C, where NARROMI outperforms other methods significantly with an AUC score of 0.938. The performances of NARROMI and other methods with

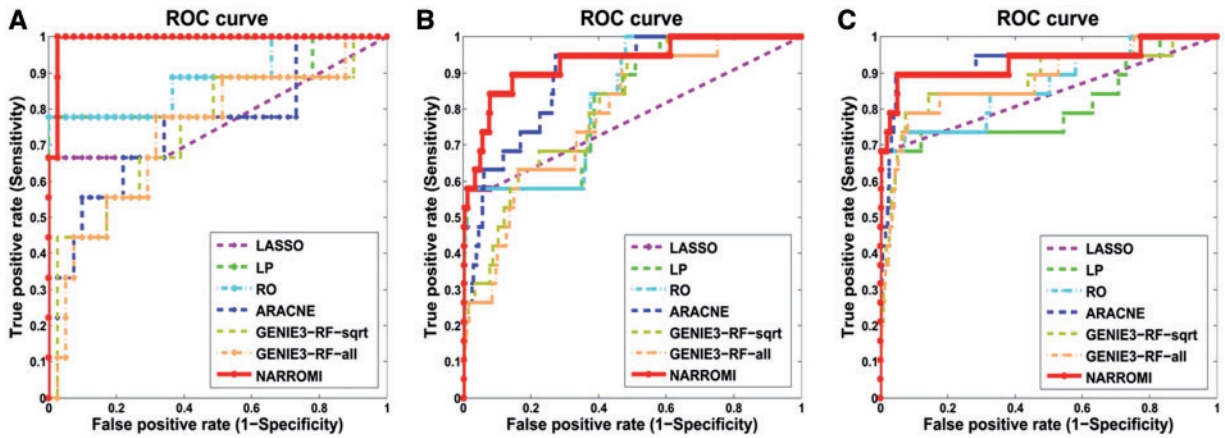

Fig. 2. ROC curves of several methods on networks with different sizes. The solid line with star points is the ROC curve of method NARROMI. The dashed lines with different points are ROC curves of method LASSO, LP, RO, ARACNE, GENIE3_RF_sqrt and GENIE3_RF_all, respectively. (**A**) The ROC curves on network with size 10. (**B**) The ROC curves on network with size 100. (**C**) The ROC curves on network with size 1000

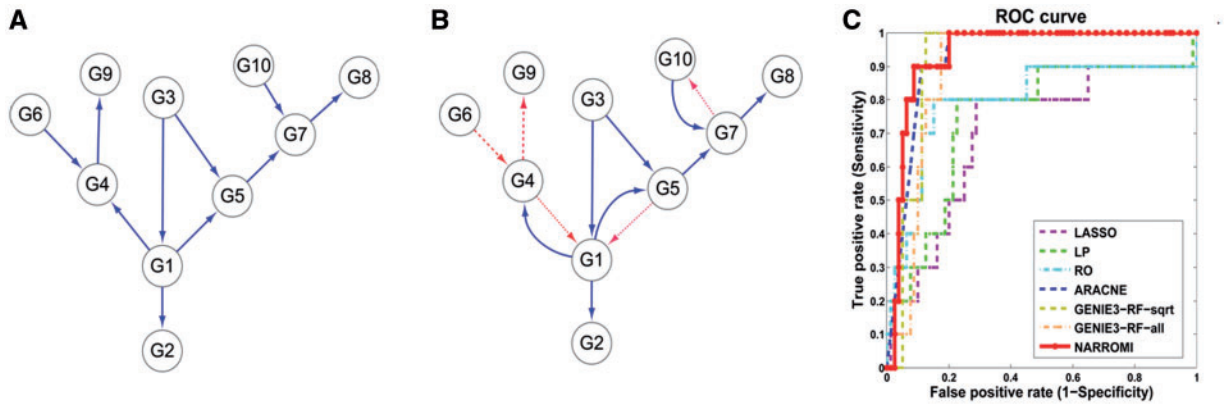

Fig. 3. Comparison of a 10-gene network inferred from DREAM3 dataset. (**A**) The true network with 10 nodes and 10 edges. (**B**) The network inferred by NARROMI. The solid lines are true positive edges of the inferred network. The edges G4-G1, G5-G1 and G7-G10 with dotted lines are false positive edges, which are mistaken for the correlation from opposition, whereas the edges G4-G9 and G6-G4 with dashed lines are missed for the low causalities. (**C**) The comparison of ROC curves of method NARROMI with other methods. The solid line with star points is the ROC curve of method NARROMI, which superior to other methods

Table 2. Comparison of different methods on networks with sizes 10 and 50 in DREAM3

Method	TPR	FPR	PPV	ACC	MCC	AUC
Size 10						
LASSO	0.600	0.837	0.082	0.211	−0.191	0.703
LP	0.100	0.412	0.029	0.533	−0.202	0.738
RO	0.100	0.500	0.024	0.456	−0.252	0.798
ARACNE	**0.900**	0.112	0.500	0.888	0.618	0.930
GENIE3_FR_ sqrt	0.700	0.112	0.437	0.867	0.483	0.919
GENIE3_FR_ all	0.700	0.138	0.389	0.844	0.442	0.894
NARROMI	0.700	**0.050**	**0.636**	**0.922**	**0.623**	**0.938**
Size 50						
LASSO	0.351	0.129	0.081	0.855	0.113	0.711
LP	0.389	0.085	0.130	0.899	0.182	0.669
RO	0.494	0.131	0.109	0.857	0.181	0.727
ARACNE	**0.597**	0.082	0.192	0.908	0.303	0.832
GENIE3_FR_ sqrt	0.481	0.078	0.167	0.908	0.245	**0.843**
GENIE3_FR_ all	0.442	0.073	0.164	0.912	0.231	0.796
NARROMI	0.532	**0.062**	**0.217**	**0.925**	**0.307**	0.839

The best performer for the relative item is noted in bold.
LASSO, regression-based method; LP, linear programming-based method; RO, recursive optimization-based method; ARACNE, MI-based method; GENIE3, random forests-based methods; NARROMI, method based on RO and MI.

respect to PPV, ACC, MCC and AUC are shown in Table 2, where NARROMI is superior to other methods.

Secondly, the *Yeast* gene expression data with network size 50 was used to evaluate NARROMI and other methods. Table 2 shows the results obtained by different methods with respect to distinct performance indexes. The ROC curves by these methods can be found in Supplementary Figure S2. From the results, we can observe that NARROMI performs better than most methods except the method GENIE3 with parameter 'sqrt' for one case. This is also consistent with the analysis that no single inference method performs optimally across all datasets (Marbach *et al.*, 2012).

To evaluate the effect of Step 2 (RO) on the performance of NARROMI, we compared the results of NARROMI with or without RO. The results can be found in Supplementary Tables S3 and S4, from which we can see that Step 2 of NARROMI can indeed remove most indirect regulations, and the false positives can be reduced significantly from 0.188 to 0.037 and from 0.109 to 0.032, respectively.

3.2 Identification of gene regulatory interactions in *E. coli*

Except the above simulation datasets, NARROMI was also applied to construct regulatory networks from real gene expression data. We evaluated our NARROMI on the experimentally verified reference network in *E. coli* (Gama-Castro *et al.*, 2011). The expression data was drawn from the well known *E. coli* data bank (Faith *et al.*, 2008). In the experimentally verified networks, there are 2675 edges between 160 regulators and 1258 targets that can be found in the expression dataset. For each target, there are ~2 regulators on average, which is consistency with the setting for the regulation degrees of simulation dataset in Section 3.1.1.

To evaluate the performance of our method, the AUC scores were computed. Moreover, the number and proportion of TFs and target genes that were predicted correctly were also recorded. We do not compare NARROMI against GENIE3 with parameter '*all*' that is time consuming in this case. From the results in Table 3, we can see that NARROMI performs better than other methods with the highest average AUC scores for both TFs and target genes of 0.754 and 0.735, respectively. Figure 4 and Supplementary Figure S3 show our inferred network structures, which indicate high overlap with the reference network. In Figure 4A, the inferred target genes of TF AppY were shown, and those overlapped with the reference network were marked in gray. Figure 4B shows the inferred regulators of target gene gadB, and the regulators that were predicted correctly were marked in gray.

Table 3. Comparison of different methods on the network in *E. coli*

Method	LASSO	LP	RO	ARACNE	GENIE3_RF_sqrt	NARROMI
AveAUC_TF	0.708	0.733	0.730	0.749	0.684	**0.754**
#AUC>0.7 (rate)	72 (0.450)	75 (0.484)	84 (0.525)	86 (0.554)	78 (0.503)	**93 (0.600)**
#AUC>0.8 (rate)	49 (0.306)	58 (0.374)	58 (0.374)	68 (0.438)	60 (0.387)	**71 (0.458)**
AveAUC_TG	0.713	0.703	0.729	0.733	0.723	**0.735**
#AUC>0.7 (rate)	568 (0.452)	603 (0.479)	665 (0.528)	691 (0.549)	484 (0.385)	**694 (0.552)**
#AUC>0.8 (rate)	355 (0.282)	367 (0.291)	439 (0.349)	484 (0.385)	428 (0.340)	**485 (0.386)**

The best performer for the relative item is noted in bold.
LASSO, regression-based method; LP, linear programming-based method; RO, recursive optimization-based method; ARACNE, MI-based method; GENIE3_RF_sqrt, random forests-based method with '*sqrt*'; NARROMI, method based on RO and MI; AUC, area under ROC curve; #**, the number of **; AveAUC_TF, Average AUC for TFs; AveAUC_TG, Average AUC for target genes (TGs).

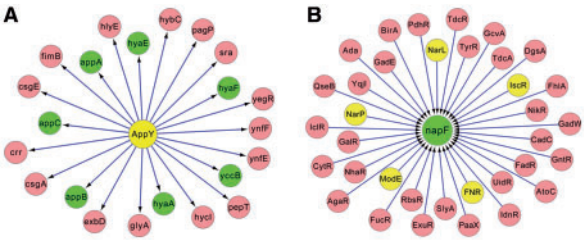

Fig. 4. Network examples inferred by NARROMI. (**A**) Target genes of AppY in the inferred network and the overlapped genes are shown in gray. (**B**) The regulators of gene napF in the inferred network and the overlapped regulators are shown in gray

4 DISCUSSION

In this article, we proposed a novel method NARROMI to infer GRNs from gene expression data by combining ODE-based RO and information theory-based MI. From results on the benchmark datasets, NARROMI is effective and outperforms other methods significantly. The good performance of NARROMI may be contributed by following factors.

First, it reduces the NAR regulation through two steps, i.e. MI filtering and RO sparseness. As the first step, MI removes the noisy regulations and thus provides clear preliminary structure to the following optimizations (RO). The subsequent RO procedure reduces the redundant (indirect) regulations gradually, thereby improving the performance of network inference.

Second, the regulatory networks inferred by NARROMI consist of both linear and non-linear correlations between regulators and targets. The RO technique detects the regulators for target genes using linear systems, which describe the chemical reactions of transcription and translation as linear differential equations. The integration with MI in the last step of the algorithm takes into account the non-linear correlations between regulators and targets, which makes NARROMI superior to general linear model based methods.

Despite the advantages of NARROMI, there is still room to improve it. For example, as a popular correlation measure, MI can identify most linear and non-linear correlations, but there are also some special non-linear correlations such as sinusoidal that MI cannot detect. Recently, a meaningful measurement, maximal information coefficient, has been proposed to detect associations from large dataset (Reshef *et al.*, 2011), which may help to improve the performance of NARROMI. Although the RO step in NARROMI is able to determine regulatory directions, other popular techniques for causal regulations may obtain better results and will be considered in NARROMI in the future.

5 CONCLUSION

We proposed a novel method NARROMI to improve the accuracy of GRN inference by simultaneously implementing the NAR regulation reduction. In this algorithm, the noisy regulations with low pair-wise correlations and the redundant regulations from indirect regulators are removed with MI and RO, respectively. Moreover, the dimension shrinking of the technique improves the efficiency of optimization and further increases the accuracy of network inference. The method was validated on the simulated benchmark GRNs from DREAM challenge and the experimentally verified network in *E. coli* with real gene expression data. The cross-validation results confirmed the effectiveness of our method (NARROMI), which outperformed previous methods.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their constructive comments, which greatly help them improve their manuscripts. They also thank Xiaodong Zhang in their group for the technical help and Dr. Vân Anh Huynh-Thu of University of Liège, Belgium, for the help with the use of software GENIE3.

Funding: Sino-French Cai Yuanpei Program from CSC (20106050), NSFC (91029301, 61134013, 61072149, 91130032, 61103075, 31100949), Innovation Program of Shanghai Municipal Education Commission (13ZZ072), Shanghai Pujiang Program, Chief Scientist Program of SIBS from CAS (2009CSP002), Knowledge Innovation Program of SIBS of CAS (2011KIP203), the Knowledge Innovation Program of CAS (KSCX2-EW-R-01), National Center for Mathematics and Interdisciplinary Sciences of CAS, and the FIRST Program initiated by CSTP, Radapop and LigeRO Projects (2009-2013, Pays de La Loire Region, France).

Conflict of Interest: none declared.

REFERENCES

Altay,G. and Emmert-Streib,F. (2010) Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, **26**, 1738–1744.

Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**, 78.

Basso,K. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.

Belcastro,V. *et al.* (2011) Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic Acids Res.*, **39**, 8677–8688.

Brunel,H. *et al.* (2010) MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics*, **26**, 1811–1818.

Cantone,I. *et al.* (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

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

Christley,S. *et al.* (2009) Incorporating existing network information into gene network inference. *PLoS One*, **4**, e6799.

De la Fuente,A. *et al.* (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**, 3565–3574.

di Bernardo,D. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, **23**, 377–383.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, 54–66.

Faith,J.J. *et al.* (2008) Many microbe microarrays database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

Frenzel,S. and Pompe,B. (2007) Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, **99**, 204101.

Gama-Castro,S. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.

Gardner,T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Geeven,G. *et al.* (2012) Identification of context-specific gene regulatory networks with GEMULA-gene expression modeling using Lasso. *Bioinformatics*, **28**, 214–221.

Honkela,A. *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. USA*, **107**, 7793–7798.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Huynh-Thu,V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Hurley,D. *et al.* (2011) Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. *Nucleic Acids Res.*, **40**, 2377–2398.

Küffner,R. *et al.* (2012) Inferring gene regulatory networks by ANOVA. *Bioinformatics*, **28**, 1376–1382.

Li,Z. *et al.* (2011) Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics*, **27**, 2686–2691.

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

MacNeil,L.T. and Walhout,A.J.M. (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.*, **21**, 645–657.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA*, **107**, 6286–6291.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Markowetz,F. *et al.* (2007) Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, **23**, i305–i312.

Margolin,A.A. *et al.* (2006a) Reverse engineering cellular networks. *Nat Protoc.*, **1**, 663–672.

Margolin,A.A. *et al.* (2006b) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Meyer,P.E. *et al.* (2008) minet: a R/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Modi,S.R. *et al.* (2011) Functional characterization of bacterial sRNAs using a network biology approach. *Proc. Natl. Acad. Sci. USA*, **108**, 15522–15527.

Reshef,D.N. *et al.* (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–1524.

Saito,S. *et al.* (2011) Discovery of chemical compound groups with common structures by a network analysis approach. *J. Chem. Inf. Model.*, **51**, 61–68.

Schaffter,T. *et al.* (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.

Smet,R.D. and Marchal,K. (2010) Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, **8**, 717–729.

Sumazin,P. *et al.* (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, **147**, 370–381.

Sun,N. *et al.* (2006) Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proc. Natl. Acad. Sci. USA*, **103**, 7988–7993.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Treviño,S. III *et al.* (2012) Robust detection of hierarchical communities from *Escherichia coli* gene expression data. *PLoS Comput. Biol.*, **8**, e1002391.

Wagner,A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, **17**, 1183–1197.

Wang,R.S. *et al.* (2009) Modeling post-transcriptional regulation activity of small non-coding RNAs in *Escherichia coli*. *BMC Bioinformatics*, **10**, S6.

Wang,Y. *et al.* (2006) Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **22**, 2413–2420.

Yeung,K.Y. *et al.* (2011) Construction of regulatory networks using expression time-series data of a genotyped population. *Proc. Natl. Acad. Sci. USA*, **108**, 19436–19441.

Yeung,M.K.S. *et al.* (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA*, **99**, 6163–6168.

Zhang,X. *et al.* (2012) Inferring gene regulatory networks from gene expression profiles by path consistency algorithm based on conditional mutual information. *Bioinformatics*, **28**, 47–54.