Bioinformatics Advance Access published November 17, 2011

Inferring gene regulatory networks from gene expression data by PC-algorithm based on conditional mutual information

Xiujun Zhang[1,2,3], Xing-Ming Zhao[1], Kun He[4], Le Lu[4], Yongwei Cao[4], Jingdong Liu[4], Jin-Kao Hao[3] , Zhi-Ping Liu[5,*] and Luonan Chen[1,5,*]

[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]Monsanto Company, St. Louis, Missouri 63167, USA, [5] 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

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: Reconstruction of gene regulatory networks (GRNs), which explicitly represent the causality of developmental or regulatory process, is of utmost interest and has become a challenging computational problem for understanding the complex regulatory mechanisms in cellular systems. However, all existing methods of inferring GRNs from gene expression profiles have their strengths and weaknesses. In particular, many properties of GRNs, such as topology sparseness and nonlinear dependence, are general in regulation mechanism but seldom be taken into account simultaneously in one computational method.

Results: In this work, we present a novel method for inferring GRNs from gene expression data considering the nonlinear dependence and topological structure of GRNs by employing path consistency algorithm (PCA) based on conditional mutual information (CMI). In this algorithm, the conditional dependence between a pair of genes is represented by the CMI between them. With the general hypothesis of Gaussian distribution underlying gene expression data, CMI between a pair of genes is computed by a concise formula involving the covariance matrices of the related gene expression profiles. The method is validated on the benchmark GRNs from the DREAM challenge and the widely used SOS DNA repair network in *Escherichia coli*. The cross-validation results confirmed the effectiveness of our method (PCA-CMI), which outperforms significantly other previous methods. Besides its high accuracy, our method is able to distinguish direct (or causal) interactions from indirect associations.

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

Contact: lnchen@sibs.ac.cn; zpliu@sibs.ac.cn.

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

1 INTRODUCTION

An important problem in molecular biology is to identify and understand the gene regulatory networks (GRNs), which explicitly represent the causality of developmental or regulatory process. Microarray technologies have produced tremendous amounts of gene expression data (Hughes *et al.*, 2000) which provide opportunity for understanding the underlying regulatory mechanism. The reconstruction or "reverse engineering" of GRNs, which aims to find the underlying network of gene-gene interactions from the measurement of gene expression, is considered one of most important goals in systems biology (Basso *et al.*, 2005; Margolin *et al.*, 2006). For this, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) program was established to encourage researchers to develop new efficient computation methods to infer robust GRNs (Marbach *et al.*, 2010).

A variety of approaches have been proposed to infer GRNs from gene expression data (Holter *et al.*, 2001; Tegner *et al.*, 2003; Bansal *et al.* 2007), such as discrete models of Boolean networks and Bayesian networks (Kauffman *et al.*, 2003), differential equations (Alter *et al.*, 2000; di Bernardo *et al.*, 2005; Cantone *et al.*, 2009; Honkela *et al.*, 2010), regression method (Tibshirani, 1996; Gardner *et al.*, 2003) and linear programming (Wang *et al.*, 2006). Although many popular network inference algorithms have been investigated (Bansal *et al.* 2007; Altay and Emmert-Streib, 2010), there are still a large space for current models to be improved (Marbach *et al.*, 2010; Smet and Marchal, 2010).

Recently, information-theoretic approaches are increasingly being used for reconstructing GRNs. Several mutual information (MI) based methods (Altay and Emmert-Streib, 2010) have been successfully applied to infer GRNs (Basso *et al.*, 2005), such as ARACNE (Margolin *et al.*, 2006), CLR (Faith *et al.*, 2007) and minet (Meyer *et al.*, 2008). In general, these approaches start by computing the pair-wise MIs between all possible pairs of genes, resulting in an MI matrix. The MI matrix is then manipulated to identify the regulatory relationships (Altay and Emmert-Streib, 2010). Mutual information provides a natural generalization of the correlation since it measures nonlinear dependency (which is common in biology) and therefore attracts much attention (Brunel *et al.*, 2010). Another advantage of these methods is their ability to deal with thousands of variables (genes) in the presence of a limited number of samples (Meyer *et al.*, 2008).

Despite these advantages, mutual information based methods only work when investigating pair-wise regulations in a gene regulatory network. They are unable to discover the joint regulations of a target gene by two or more genes. The three-way mutual information (MI3) was designed to detect the co-regulators

*To whom correspondence should be addressed.

of target genes by scoring the sum of correlative and coordinative regulatory components (Luo *et al.*, 2008). But it can only detect two of the co-regulators while missing the other co-regulators when there are more than two co-regulators with the assumption of regulation relationship. In contrast, conditional mutual information (CMI) is capable of detecting the joint regulations by exploiting the conditional dependency between genes of interest (Wang *et al.*, 2009). Methods based on both MI and CMI have also been proposed to reduce the false positive rate for detecting interactions (Basso *et al.*, 2005). Moreover, CMI can distinguish direct from indirect interactions based on multivariate time series data (Frenzel and Pompe, 2007).

It was found that regulatory network is sparse as experimentally observed in visual system of primates (Vinje and Gallant, 2000). Among the available methods, optimization methods with penalty or constraint are good alternatives to achieve sparseness (Wang *et al.*, 2006; Banerjee *et al.*, 2010), while they do not perform well when the number of variables is much larger than sample size. Path consistency (PC) algorithm (Spirtes *et al.*, 2000) was recently used to construct networks by calculating partial correlations coefficient (PCC) to estimate the conditional independencies (Kalisch and Bühlmann, 2007; Saito and Horimoto, 2009; Saito *et al.*, 2011). Besides its robustness and uniform consistency, PC-algorithm is computationally feasible and fast especially for sparse problems with a large number of nodes. However, the current PC-algorithm estimates the dependency of gene pairs with PCC, which can only detect the linear correlation between gene pairs while the nonlinear dependency is more common in biological processes.

In this work, we propose a novel method, namely PCA-CMI, for inferring GRNs from gene expression data by employing PC-algorithm based on conditional mutual information. In this algorithm, the conditional independency between a pair of genes is represented by their conditional mutual information instead of PCC with the following advantages. Firstly, MI provides a natural generalization of the correlation since it measures nonlinear dependency. Unlike PCC, it does not assume linearity, continuity, or other specific properties of dependence. Hence, MI has the flexibility to detect regulatory interactions that might be missed by linear measures such as PCC (Faith *et al.*, 2007). Secondly, mutual information is more general than PCC to model the relations between genes, while PCC is more easily distorted when points are not uniformly distributed across the axes (Butte and Kohane, 2000). With the hypothesis of Gaussian distribution for gene expression data, MI and CMI of gene pairs are evaluated by a concise formula involving the covariance matrices of the related gene expression profiles. The method is validated on the benchmark GRNs from DREAM challenge (Marbach *et al.*, 2010) and the widely used SOS DNA repair network in *Escherichia coli* (Shen-Orr *et al.*, 2002; Ronen *et al.*, 2002). The cross-validation results confirmed the effectiveness of our method (PCA-CMI), which outperforms significantly other previous methods. Besides its high accuracy, our method is able to distinguish direct (or causal) interactions from indirect associations.

2 METHODS

In this section, we will introduce some definitions of information theory including entropy, MI, and CMI, as well as the algorithm of PCA-CMI for inferring GRNs.

2.1 Information theory

MI from information theory has been used to construct GRNs from gene expression data (Altay and Emmert-Streib, 2010). In particular, MI is generally used as a powerful criterion for measuring the dependence between two variables (genes) X and Y. For gene expression data, variable X is a vector, in which the elements denote its expression values in different conditions (samples).

For a discrete variable (gene) X, the entropy $H(X)$ is the measure of average uncertainty of variable X and can be defined by

$$H(X) = -\sum_{x\in X} p(x)\log p(x), \tag{1}$$

where $p(x)$ is the probability of each discrete value x in X. The joint entropy $H(X, Y)$ of X and Y can be denoted by

$$H(X,Y) = -\sum_{x\in X,\ y\in Y} p(x,y)\log p(x,y), \tag{2}$$

where $p(x, y)$ is the joint probability of x in X and y in Y.

Mutual information (MI) measures the dependency between two variables. For discrete variables X and Y, MI is defined as

$$I(X,Y) = -\sum_{x\in X,\ y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{3}$$

MI can also be defined in terms of entropies as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{4}$$

where $H(X,Y)$ is joint entropy of X and Y. High MI value indicates that there may be a close relationship between the variables (genes), while low MI value implies their independence.

Conditional mutual information measures conditional dependency between two variables (genes) given other variable(s) (gene(s)). The CMI of variables X and Y given Z is defined as

$$I(X,Y\mid Z) = \sum_{x\in X,\ y\in Y,\ z\in Z} p(x,y,z)\log\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)}. \tag{5}$$

CMI can also be expressed in terms of entropies as

$$H(X,Y\mid Z) = H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z), \tag{6}$$

where $H(X,Z), H(Y,Z), H(X,Y,Z)$ are joint entropies. Similarly, high CMI indicates that there may be a close relationship between the variables X and Y given variable(s) Z.

Here, the entropy is estimated with Gaussian kernel probability density estimator (Basso *et al.*, 2005) as follows.

$$P(X_i) = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{(2\pi)^{n/2}\mid C\mid^{n/2}}\exp\left(-\frac{1}{2}(X_j - X_i)^T C^{-1}(X_j - X_i)\right), \tag{7}$$

where C is the covariance matrix of variable X, $\mid C\mid$ is the determinant of matrix C, N is the number of samples and n is the number of variables (genes) in C.

With equations (1) and (7), we can get the entropy of variable X as follows.

$$H(X) = \log[(2\pi e)^{n/2}\mid C\mid^{1/2}] = \frac{1}{2}\log(2\pi e)^n\mid C\mid. \tag{8}$$

With equation (8), the formulation (4) can be expressed as follows.

$$I(X,Y) = \frac{1}{2}\log\frac{\mid C(X)\mid\cdot\mid C(Y)\mid}{\mid C(X,Y)\mid}. \tag{9}$$

Similarly, (6) equals to

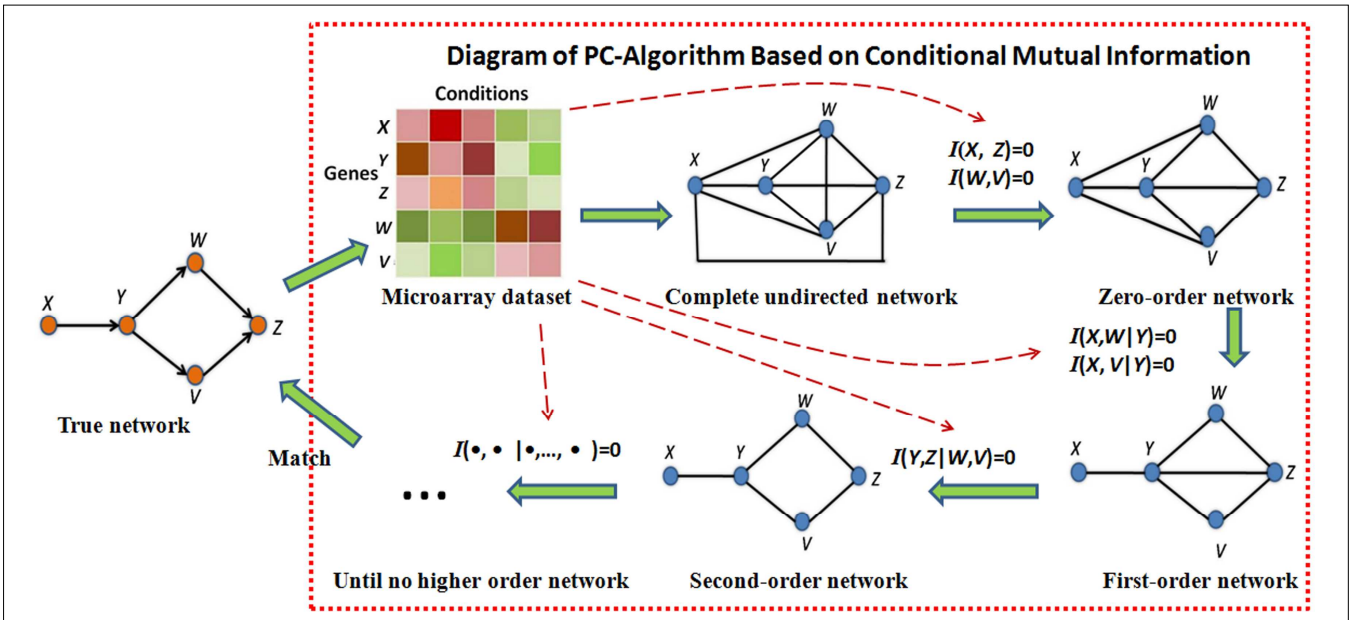

Fig.1. Diagram of method PCA-CMI. In the figure, $I(\cdot,\cdot)$ is the mutual information and $I(\cdot,\cdot|\cdot)$ is the conditional mutual information. They are calculated from gene expression data by a concise formula of computation. The MI and CMI equal to zero or lower than given threshold represent independence between variables (genes). The left graph is the true network of the microarray dataset with gene expression profiles under different conditions (samples). The graph in pink box with dashes is the diagram of PCA-CMI, which detects the true network step by step according to the (conditional) independency of gene pairs.

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

which is an efficient formula to calculate CMI between two variables (genes) given one or more variables (genes). For example, if conditional variable $Z=(Z_1, Z_2)$ is composed of two variables (genes) Z_1 and Z_2, we get second-order CMI.

When variables (genes) X and Y are independent, we get $I(X,Y)=0$. Similarly, if the variables X and Y are conditional independence given Z, we have $I(X, Y|Z)=0$.

To test whether a CMI is zero, it is statistically tested by using the Z-statistic (Kalisch and Bühlmann, 2007; Saito *et al.*, 2011). Firstly, the CMIs are normalized by

$$\hat{I}(X,Y|Z)=\frac{I(X,Y|Z)}{H(X,Z)+H(Y,Z)}, \quad (\hat{I}(X,Y)=\frac{I(X,Y)}{H(X)+H(Y)}, \text{if } Z=\phi).$$

Secondly, Fisher's Z-transforms of CMI are calculated by following equation

$$z_{X,Y|Z}=\frac{1}{2}\log\left(\frac{1+\hat{I}(X,Y|Z)}{1-\hat{I}(X,Y|Z)}\right).$$

Then, classical decision theory yields the following rule when using the significance level α. Reject the null-hypothesis $H_0: z_{X,Y|Z}=0$ against the two-sided alternative $H_1: z_{X,Y|Z}\neq 0$ if

$$\sqrt{n-|Z|-3}\,|z_{X,Y|Z}|>\Phi^{-1}(1-\alpha/2),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal distribution $N(0,1)$ and $|Z|$ is the conditional order of CMI.

2.2 Path consistency algorithm (PC-algorithm)

After we obtain MI and CMI through formulation (9) and (10), the path consistency algorithm (PC-algorithm) is used to remove the edges with (conditional) independent correlation from the graph. The inference of GRNs will be performed by deleting the edges with independent correlation recursively, i.e. from low to high order independent correlation until there is no edge can be deleted.

We describe the process of PCA-CMI in detail as follows. Firstly, generate a complete graph according to the number of genes. Secondly, for adjacent gene pair i and j, compute mutual information (zero-order conditional mutual information) $I(i, j)$. If the gene pair i and j has low or zero mutual information, it represents independent correlation, then we delete the edge between genes i and j. Thirdly, for adjacent gene pair i and j, select the adjacent gene k of them and compute first-order conditional mutual information $I(i, j|k)$. If the gene pair i and j has low or zero conditional mutual information which represent their independent correlation, delete the edge between them. The next step is to compute higher order CMI until there are no more adjacent edges.

The following gives the algorithm to infer a gene regulatory network.

PC-Algorithm based on CMI (PCA-CMI)

Step-0: Initialization. Input the gene expression data $\boldsymbol{A}$ and set the parameter θ for deciding the independence. Generate the complete network $\boldsymbol{G}$ for all genes (i.e. the clique graph of all genes). Set $L=-1$.

Step-1: $L=L+1$; For a nonzero edge $\boldsymbol{G}(i, j)\neq 0$, select adjacent genes connected with both genes i and j. Compute the number T of the adjacent genes (not including genes i and j).

Step-2: If $T<L$, stop. If $T\geq L$, select out L genes from these T genes and let them as $\boldsymbol{K}=[k_1,\cdots,k_L]$. The number of all selections for $\boldsymbol{K}$ is C_T^L. Compute the L-order CMI $I(i, j|\boldsymbol{K})$ for all C_T^L selections, and choose the maximal one denoting as $I_{\max}(i, j|\boldsymbol{K})$. If $I_{\max}(i, j|\boldsymbol{K})<\theta$, set $\boldsymbol{G}(i, j)=0$. Return to **Step-1**.

Fig.1 shows a diagram of PC-algorithm based CMI for a five-gene network. Microarray data is the expression data of genes X, Y, Z, W and V. The first step is to generate the complete network with these five genes. Then the independence between gene pairs is decided by the MI between

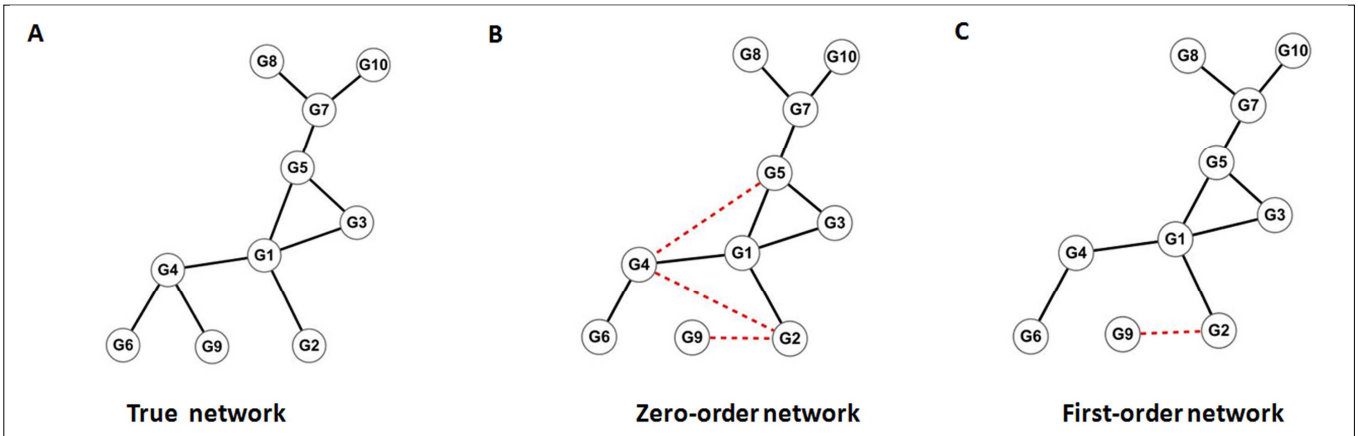

Fig.2. Comparison of a 10-gene network inferred from DREAM3 dataset. (**A**) The true network with 10 nodes and 10 edges. (**B**) Zero-order network inferred by PCA-CMI from gene expression data. The edges with red dotted lines G2-G4, G4-G5 and G2-G9 are false positives, while the edge G4-G9 is false negative. (**C**) First-order network inferred from gene expression data. The false positive edges G2-G4 and G4-G5 in zero-order network were successfully removed by PCA-CMI.

them. If the MI is smaller than a given threshold θ, the edge between the two genes is deleted for the independence. Here, the mutual information $I(X, Z)$ and $I(W, V)$ are approximately equal to zero on assumption, so the edges $E(X, Z)$ and $E(W, V)$ are deleted and the zero-order network is reconstructed. Then, the first-order CMIs between genes with common adjacent edges in the zero-order network are computed, and the conditional mutual information $I(X, W|Y)$ and $I(X, V|Y)$ are assumed to equal to zero, so the edges $E(X, W)$ and $E(X, V)$ are deleted and the first-order network is obtain. Based on the first-order network, the second-order CMIs between genes can be computed and the CMI $I(Y, Z|W, V)$ is assumed approximately equal to zero, so the edge $E(Y, Z)$ is deleted and the second-order network is inferred. There is no third-order CMI, so the algorithm terminates and the second-order network is the inferred GRN (or final GRN).

3 RESULTS

In order to validate our method, PCA-CMI was applied to several simulation datasets as well as real gene expression datasets. As for simulation data, we tested two synthetic datasets from the DREAM3 (Dialogue for Reverse Engineering Assessment and Methods) challenge (Marbach *et al.*, 2010). As for real gene expression data, we applied our method to the well-defined SOS DNA repair network with experiment dataset in *Escherichia coli* (Shen-Orr *et al.*, 2002; Ronen *et al.*, 2002), and also rice (*Oryza sativa* L.) gene expression data (see Supplementary Materials).

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

$$\mathrm{TPR}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}),$$

$$\mathrm{FPR}=\mathrm{FP}/(\mathrm{FP}+\mathrm{TN}),$$

$$\mathrm{PPV}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}),$$

$$\mathrm{ACC}=(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}),$$

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.

3.1 Evaluation on simulation data

In order to assess the effectiveness of methods for constructing GRNs, simulation data was generated based on the benchmarking network. Many tools have been developed for assessing the effectiveness of GRN inference methods (Hache *et al.*, 2009). DREAM challenge introduces a framework for critical performance assessment of methods for GRNs inference and presents an *in silico* benchmark suited as a blinded, community-wide challenge within the context of the DREAM project. In this challenge, the gene expression datasets with noise and their gold standard (benchmark) networks were given. The gold standard networks were selected from source networks of real species. We tested our method on the DREAM3 datasets about *Yeast* knock-out gene expression data in sizes 10 and 50, respectively.

Firstly, we tested method PCA-CMI on the *Yeast* gene expression data with network size 10, sample number 10. Fig.2 shows the inferred networks with different CMI orders from gene expression data. Fig.2 (A) is the true gene regulatory network which contains 10 nodes with 10 edges. The network was selected from an experimental verified network in *Yeast* genomes. We chose 0.03 as the threshold value of mutual information and conditional mutual information to decide independence, and got networks with different CMI orders. Fig.2 (B) is the zero-order network inferred from gene expression data. The edges with red dotted lines are wrongly inferred and the edges with black solid lines are correctly inferred. In the wrongly inferred edges, G2-G4, G3-G4 and G2-G9 are redundant edges, while edge G4-G9 is missed out. Fig.2 (C) is the first-order network inferred from gene expression data which is also the final network because there is no higher-order networks to be inferred based on the algorithm.

Table 1. Results of a 10-gene network in DREAM3 with different CMI orders.

Order	TP	FP	TPR	FPR	PPV	ACC
Zero	9	3	0.900	0.086	0.750	0.911
First	9	1	0.900	0.029	0.900	0.956

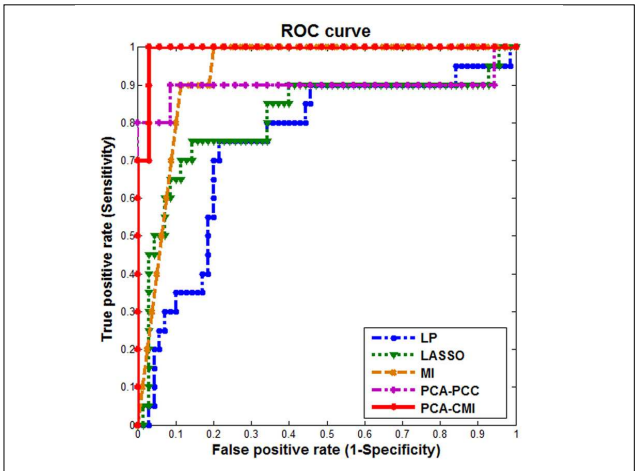

Fig.3. ROC curves of several methods on DREAM3 challenge *Yeast* dataset in size 10. The red solid line with star points is the ROC curve of method PCA-CMI. The AUC value (0.991) of Method PCA-CMI is higher than other methods.

Clearly, in the first-order network, the non-existing regulations of G2-G4 and G3-G4 were successfully removed.

As for the different order networks, Table 1 gives the results of assessment for the prediction performance. We can find that a higher order network has a higher accuracy (ACC) with a lower false positive rate (FPR) than that of a lower order network, which demonstrates that method PCA-CMI can detect the true network step by step. This provides evidence that PCA-CMI is effective and efficient to infer GRNs.

In order to clearly evaluate the performance of PCA-CMI, the ROC curve was drawn. The red solid line with star points in Fig.3 is the ROC curve of method PCA-CMI. The AUC value reaches 0.991, which indicates the high efficiency of the performance of PCA-CMI. In order to describe the efficiency of PCA-CMI, we also compared the method with linear programming method (LP), multiple linear regression Lasso method (LASSO), mutual information method (MI) and PC-Algorithm based on partial correlation coefficient (PCA-PCC) (Wang *et al.*, 2006; Tibshirani, 1996; Margolin *et al.*, 2006; Kalisch and Bühlmann, 2007). Fig.3 gives the ROC curves of these GRN inference methods and clearly describes the higher performance of method PCA-CMI superior to other methods. Table 2 gives the result about the comparison of different methods for inferring GRNs. From Table 2, we can see that PCA-PCC and MI perform better than LP and LASSO obviously, while our method PCA-CMI performs better than PCA-PCC and MI with the AUC score 0.991.

Table 2. Comparison of the performance of different methods for inferring a 10-gene network in DREAM3.

Method	LP	LASSO	PCA-PCC	MI	**PCA-CMI**
AUC	0.750	0.813	0.897	0.930	**0.991**

Abbreviation: AUC: area under ROC curve; LP: linear programming method; LASSO: Lasso regression method; MI: mutual information method; PCA-PCC: PC-algorithm based on partial correlation coefficient.

Secondly, we tested our PCA-CMI on the *Yeast* gene expression data with network size 50 and sample number 50. The network was selected from real and experimental verified network in *Yeast* genomes. We set the threshold value 0.1 of MI and CMI to decide independence and obtained different order networks. The network structure of the true network and the inferred different order networks from gene expression data are described in Fig.S1 (see Supplementary Materials). Fig.S1 (A) is the true network which contains 50 nodes with 77 edges. Fig.S1 (B-D) gives the network structures with different CMI orders. As for the different order networks, Table S1 gives the results of assessment for the performance of different order networks which demonstrated the efficient performance of PCA-CMI (see Supplementary Materials).

In order to check the advantage of PCA-CMI, we also compared it with some other methods described above. The ROC curves of our PCA-CMI and other methods are analyzed in Fig.S2 (see Supplementary Materials). The AUC value was also considered to test the efficiency of PCA-CMI. Table S2 (see Supplementary Materials) gives the result about the comparison of different methods for inferring GRNs. From the ROC curves and the result of the table, we can see that our method PCA-CMI performs better than all other methods with the AUC value 0.839. All the results demonstrate the high efficiency of our method PCA-CMI.

In addition, to assess the effectiveness of conditional mutual information (CMI) on measuring correlation between genes, the comparison of methods using mutual information including PCA-CMI, MI and MI3 was performed on a 10-gene network with 10 edges in DREAM3 challenge (Marbach *et al.*, 2010) and a synthetic 9-gene network with 13 edges (Luo *et al.*, 2008). The results of comparison listed in Tables S3 and S4 (see Supplementary Materials) show that MI can detect almost all of correlations between genes while CMI can screen out the indirect correlations from direct ones.

3.2 Construction of SOS network in *Escherichia coli*

PCA-CMI was also implemented to identify how well it works in constructing regulatory networks from real gene expression data. We tested our method on the well-known SOS DNA repair network and experiment dataset in *Escherichia coli* (Shen-Orr *et al.*, 2002; Ronen *et al.*, 2002).

In order to evaluate PCA-CMI, the ROC curve was considered. The results were compared with that of LP, LASSO, MI and PCA-PCC methods. The comparison results are shown in Fig. 4. The AUC of PCA-CMI is 0.792, which indicates the proposed method can distinguish most of the true interactions between genes. Moreover, it clearly performed better than other methods. Table 3 gives the results about the comparison of different methods for

Table 3. Comparison of different methods on SOS DNA repair network.

Method	LP	LASSO	PCA-PCC	MI	**PCA-CMI**
AUC	0.590	0.618	0.670	0.739	**0.792**

Abbreviation: AUC: area under ROC curve; LP: linear programming method; LASSO: lasso regression method; MI: mutual information method; PCA-PCC: PC-algorithm based on partial correlation coefficient.

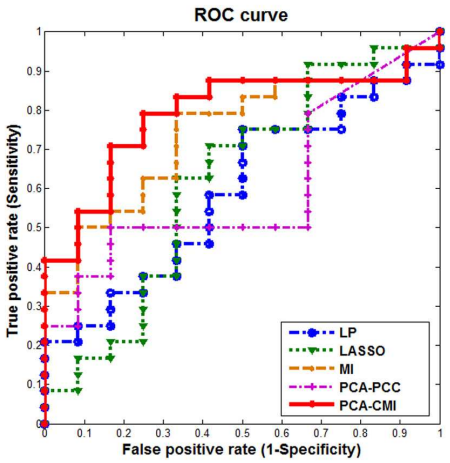

Fig.4. Comparison of ROC curves of PCA-CMI with other methods on SOS real gene expression dataset. The red solid line with star points is the ROC curve of method PCA-CMI. The AUC (Area under ROC curve) value of PCA-CMI is 0.792, which is higher than other methods.

inferring GRNs. In these results, our method PCA-CMI performs better than other methods. We also can find that MI and PCA-CMI with the consideration of nonlinear correlation do better than LP, LASSO and PCA-PCC with the consideration of linear correlation. Furthermore, PCA-CMI is better than MI. For the PC-algorithms, PCA-CMI does better than PCA-PCC. All the results have provided evidence for the effectiveness and efficiency of the proposed method on the real gene expression data.

4 DISCUSSION

In this work, we proposed a novel method PCA-CMI to infer gene regulatory networks. From the benchmark validations, PCA-CMI is an effective method and the comparison results show its advantages. Comparing to linear methods like LP, LASSO and PCA-PCC, our method can cover nonlinear relations between gene pairs based on MI and CMI from information theory. They can detect the nonlinear statistical dependence, which is accordance with the complexity of biology instead of linear assumption.

In addition, PCA-CMI can distinguish the direct interactions or correlations from indirect ones, which are important for causality analysis. For example, consider three genes forming a causal chain: The first gene couples to the second, and the latter to the third gene. In this case, a pair wise mutual information analysis would yield dependencies also between the first and third genes, and we could not decide whether this coupling is made directly or mediated by the second one. An alternative to overcome the problem is to consider partial (conditional) dependence or correlation. To this end, partial mutual information as proposed is a general approach because it relates to nonlinear dependencies, needs no explicit modeling, and further represents the information between two observations that is not contained in a third one. Thus, in this way, we can discover the real underlying coupling or dependence structure.

To illustrate the significance of the inferred networks, the average correlation underling the inferred regulations was compared with that of random gene pairs. We chose the result of the method performed on a rice (*Oryza sativa* L.) gene expression dataset (GEO access number: GSE4471) with 1387 significantly differential expression genes (t-test P-value<0.05). The average mutual information (MI) and Pearson Correlation Coefficient (PCC) between gene pairs from 4459 inferred edges by PCA-CMI were also computed. The average PCC was randomly repeated 5 times. Figs.S3 and S4 describe the histograms of MIs and PCCs (see Supplementary Materials). The result shows that conditional mutual information can significantly quantify the correlations of gene pairs.

For the general PC-algorithm, genes i, j and k are randomly selected in order to reduce the computational complexity. To reduce computational complexity but not sacrifice the accuracy to detect the true regulatory interactions, we adopted an optimal strategy to select L genes from T adjacent genes for randomly selected gene pair i and j, which also ensures the local optimality of the algorithm. For example, suppose that there are T ($T \geq 1$) genes which are adjacent with both genes i and j. When constructing the L-order ($L \leq T$) network, all the L-order CMIs for the possible combinations of L conditional genes from T genes are computed and the maximum one or the geometric mean of them is selected to decide the existence of regulation. It is well known that the path consistency algorithm is not robust for some inputs. In order to improve the robustness and accuracy of the inference in large-scale datasets, modular analysis is adopted in our algorithm. Specifically, the first level network is inferred by PCA-CMI. Then sub-networks are identified from the first level network by module finder methods such as CFinder (Adamcsek *et al.*, 2006; Radicchi *et al.*, 2004). Each modular network can be re-inferred again by PCA-CMI, and the resulting networks are called second level sub-networks. The bi-level integrative method of PCA-CMI can improve accuracy for avoiding the effects of wrong edges in a large scale network. The performance of such bi-level method was tested on DREAM3 challenge InSilico50 datasets and the ACC value was actually improved from 0.9102 to 0.9332 (see Table S5 in Supporting Materials). In addition, the path consistency algorithm based on partial coefficient correlation is not robust for the correlation estimator of PCC, in which a small quantity of outliers suffices to completely distort the resulting network (Kalisch and Bühlmann, 2008). A comparison study on correlation measure for MI and PCC based methods from gene expression datasets showed that MI is more robust than PCC with respect to missing expression values (Priness *et al.*, 2007).

However, there is some limitation for PCA-CMI. As same as that of ARACNE, PCA-CMI cannot directly infer edge directionality, which is also a general limitation of many other methods, especially for these methods that do not use time series data (Margolin *et al.*, 2006). Another limitation of PCA-CMI (the same as other methods based on general mutual information) is that it cannot detect all regulatory relations, in particular those with time delay between transcriptional factors and their target genes. Although mutual information is a symmetric measure, which cannot derive the direction of an edge (Meyer *et al.*, 2008), this limitation can be relieved by a two-tier approach in which an undirected GRN is inferred firstly, and then edge directionality is accessed via other method like multiple linear regression method (Carrera *et al.*, 2009) or specific biochemical perturbation method (Margolin *et al.*, 2006).

5 CONCLUSION

In this work, we proposed a novel method PCA-CMI for inferring GRNs from gene expression data by taking into account the nonlinear dependence and sparse structure of GRNs. In this algorithm, the conditional independence between a pair of genes is represented by conditional mutual information between this gene-pair given certain other genes. With the hypothesis of Gaussian distribution for gene expression data, CMIs between gene pairs are calculated by a concise formula involving the covariance matrices of the related gene expression profiles. The proposed method performed superior to other methods on the benchmark gene regulatory networks from the DREAM3 challenge, real SOS DNA repair network in *Escherichia coli*, and real rice networks. Both the cross-validation results and comparison studies demonstrate the effectiveness and efficiency of our method. In addition, PCA-CMI is able to distinguish the direct regulatory relationships from indirect ones.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their constructive comments, which greatly help us improve our manuscript.

Funding: The study was supported by a research grant from Monsanto Company. This work was also partially supported by the Innovation Program of Shanghai Municipal Education Commission (10YZ01), Shanghai Pujiang Program (11PJ1410500), Shanghai Rising-Star Program (10QA1402700), Sino-French Cai Yuanpei Program from CSC (20106050), Chief Scientist Program of SIBS from CAS (2009CSP002); by the Knowledge Innovation Program of SIBS of CAS (2011KIP203) and SA-SIBS Scholarship Program, the Knowledge Innovation Program of CAS (KSCX2-EW-R-01), National Center for Mathematics and Interdisciplinary Sciences of CAS; by NSFC (61103075, 61072149, 31100949 and 91029301), Japan (JSPS) FIRST Program initiated by CSTP.

REFERENCES

Adamcsek,B. *et al.* (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 2006, **22**, 1021-1023.

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.

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

Banerjee,I. *et al.* (2010) An integer programming formulation to identify the sparse network architecture governing differentiation of embryonic stem cells. *Bioinformatics*, **26**, 1332-1339.

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.

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.

Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **5**, 415–426.

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

Carrera,J. *et al.* (2009) Model-based redesign of global transcription regulation. *Nucleic Acids Res.*, **37**, e38.

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.

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

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

Hache,H. *et al.* (2009) GeNGe: systematic generation of gene regulatory networks. *Bioinformatics*, **25**, 1205-1207.

Holter,N.S. *et al.* (2001) Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. USA*, **98**, 1693-1698.

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.

Kalisch,M. and Bühlmann,P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**, 613-636.

Kalisch,M. and Bühlmann,P. (2008) Robustification of the PC-algorithm for directed acyclicgraphs. *J.Comput. Graph. Stat.*, **17**, 773–789.

Kauffman,S. *et al.* (2003) Random Boolean network models and the yeast transcriptional network. *Proc. Natl Acad. Sci. USA*, **100**, 14796-14799.

Luo,W., *et al.* (2008) Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinformatics*, **9**, 467.

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

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

Margolin,A.A. *et al.* (2006) 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.

Priness,I. *et al.* (2007) Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, **8**, 111.

Radicchi,F. *et al.* (2004) Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 2658-2663.

Ronen,M. *et al.* (2002) Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA*, **99**, 10555-10560.

Saito,S. and Horimoto,K. (2009) Co-expressed gene assessment based on the path consistency algorithm: operon detention in *Escherichia coli*. *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 4280-4286.

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.

Saito,S. *et al.* (2011) A procedure for identifying master regulators in conjunction with network screening and inference. *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 296-301.

Shen-Orr,S. *et al.* (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, **31**, 64-68.

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

Spirtes,P. *et al.* (2001) Causation, Prediction, and Search. Boston, The MIT Press, 2nd edition.

Tegner,J. *et al.* (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA*, **100**, 5944-5949.

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

Vinje,W.E and Gallant,J.L. (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, **287**, 1273-1276.

Wang,K. *et al.* (2009) Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat. biotechnol.*, **27**, 829-839.

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