Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks

Xiujun Zhang¹,²,³, Juan Zhao¹, Jin-Kao Hao⁴, Xing-Ming Zhao⁵,* and Luonan Chen¹,⁶,*

¹Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, ²Department of Mathematics, Xinyang Normal University, Xinyang 464000, China, ³School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459, Singapore, ⁴LERIA, Department of Computer Science, University of Angers, Angers 49045, France, ⁵Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China and ⁶Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan.

Received September 15, 2014; Revised December 03, 2014; Accepted December 05, 2014

ABSTRACT

Mutual information (MI), a quantity describing the nonlinear dependence between two random variables, has been widely used to construct gene regulatory networks (GRNs). Despite its good performance, MI cannot separate the direct regulations from indirect ones among genes. Although the conditional mutual information (CMI) is able to identify the direct regulations, it generally underestimates the regulation strength, i.e. it may result in false negatives when inferring gene regulations. In this work, to overcome the problems, we propose a novel concept, namely conditional mutual inclusive information (CMI²), to describe the regulations between genes. Furthermore, with CMI², we develop a new approach, namely CMI²NI (CMI²-based network inference), for reverse-engineering GRNs. In CMI²NI, CMI² is used to quantify the mutual information between two genes given a third one through calculating the Kullback–Leibler divergence between the postulated distributions of including and excluding the edge between the two genes. The benchmark results on the GRNs from DREAM challenge as well as the SOS DNA repair network in Escherichia coli demonstrate the superior performance of CMI²NI. CMI²NI was also used to reconstruct cancer-specific GRNs using gene expression data from The Cancer Genome Atlas (TCGA). CMI²NI is freely accessible at http://www.comp-sysbio.org/cmi2ni.

INTRODUCTION

Identifying the causal regulations between genes is the key to understanding the biological processes within cells. Despite the great efforts from the community, such as ENCODE (1) and modENCODE (2), untangling the comprehensive gene regulation networks (GRNs) is still a challenging task (3). With the increasingly accumulated high throughput data, many computational approaches have recently been developed to reconstruct GRNs (4–6). In general, these GRN inference approaches fall into two categories, i.e. model-based and machine learning-based approaches (7,8). In model-based methods, the chemical reactions of transcription and translation as well as other cellular processes are generally described with linear or nonlinear differential equations, where the parameters represent the causal strengths of the corresponding regulations. Popular methods in this category include singular value decomposition (9,10), network component analysis (11), multiple linear regression (12–14), and linear programming (15,16). In machine learning-based approaches, the regulations between genes are described by different indexes (i.e. causal association) (17,18), including Pearson correlation coefficient (19,20), Bayesian network (21), information theory-based mutual information (MI) (22–26) and conditional mutual information (CMI) (27,28).

Among those popular methods, the mutual information (MI) has been widely used to construct GRNs due to its capability of characterizing the nonlinear dependency between genes (23,29). Recent study shows that comparing
with other approaches, MI is a natural way to equitably quantify statistical associations (30). Another advantage of the MI-based methods is their ability to deal with thousands of variables (genes) in the presence of the limited number of samples (14,31,32). Since MI describes the statistical dependencies between two variables, edge in GRN implies possible functional dependency between the two connected genes but not necessarily causal regulation. In other words, the edge detected by MI may be a functional or indirect regulation through one or more intermediaries instead of a direct (or physical) interaction between a transcription factor (TF) and a gene. Therefore, the mutual information overestimates the regulation relationships to some extent and fails to distinguish indirect regulators from direct ones, thereby leading to possible false positives during the inference of GRNs (28,33–35). Recently, the conditional mutual information (CMI) was proposed to infer the causal regulations between genes (36). As an extension of MI, CMI is able to separate the direct regulations from those indirect ones. CMI has also been used to detect the activity of TFs and miRNAs in transcriptional and post-transcriptional regulations (27,37). However, the theoretical analysis shows that CMI tends to underestimate the regulation strength in some cases due to its statistical feature (38–40).

In many cases, the real regulations between nodes in a GRN are obscured by the noise in the data. Therefore, most network inference methods perform poorly with a high false positive rate. To address this issue, Brarzel et al. (33) and Feizi et al. (35) described easily implemented methods for identifying and removing erroneous links, thereby producing more accurate networks. However, it turns out that both of the two methods are related with the method of partial correlation, which is the correlation between two variables predicted linearly from all other variables (34). Both of the two approaches need to inverse the correlation matrix to achieve the result and they are only different in scaling the inverse correlation matrix. Well known to us, it is difficult to process the inverse of correlation matrix when large scale variables but with small samples are given (41). Hence, approximate methods used by the two approaches to achieve the inverse of correlation matrix will destroy the performance of network inference in many cases.

In this work, to overcome those problems, we propose a concept based on a new measure of causal strength (36), i.e. CMI2 (conditional mutual inclusive information), to quantify the causal associations between variables. To infer GRNs, a new algorithm, namely CMI2NI, is developed by combining CMI2 with path consistency (PC) algorithm. CMI2NI, which alleviates both the overestimation problem of MI and the underestimation problem of CMI based on information theory, gives a quantitative measurement of causal associations between two genes. With the hypothesis of Gaussian distribution for gene expression data, CMI2 can be calculated by a concise formula involving the covariance matrices of the related gene expression profiles. The proposed network inference method CMI2NI can not only accurately quantify causal associations but also reconstruct the correct topological structures of biological networks even with a small number of samples. The experimental results on the benchmark GRNs from DREAM challenge and the widely used SOS DNA repair network in Escherichia coli demonstrated the effectiveness of our CMI2NI. As a case study, CMI2NI was applied to reconstruct cancer-specific GRNs based on gene expression data from The Cancer Genome Atlas (TCGA), where the GRNs provide a global view of the regulatory circuit of the cancer genes.

**MATERIALS AND METHODS**

(Conditional) mutual information

Recently, MI and CMI have been widely used to reconstruct GRNs due to their advantages in measuring dependency between variables. In general, the gene expression data can be described as vectors, in which the elements denote the expression values of genes under different conditions. MI measuring the dependency between two variables (genes) X and Y can be defined as below (42),

\[
MI(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)},
\]

where \(p(x, y)\) is the joint probability distribution of \(X\) and \(Y\), and \(p(x)\) and \(p(y)\) are the marginal probability distributions of \(X\) and \(Y\), respectively. With the widely adopted hypothesis of Gaussian distribution for gene expression data, formula (1) can be easily calculated using the following equivalent formula (43)

\[
MI(X, Y) = \frac{1}{2} \log \frac{|C(X) \cdot |C(Y)|}{|C(X, Y)|},
\]

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

On the other hand, CMI 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

\[
CMI(X, Y|Z) = \sum_{x \in X, y \in Y, z \in Z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)},
\]

where \(p(x, y|z)\), \(p(x|z)\) and \(p(y|z)\) are conditional probability distributions. Similarly, under the assumption of Gaussian distributions (28), formula (3) equals to

\[
MI(X, Y|Z) = \frac{1}{2} \log \frac{|C(X, Z) \cdot |C(Y, Z)|}{|C(Z)| \cdot |C(X, Y, Z)|},
\]

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

Kullback–Leibler divergence-based causal strength measure

Recently, to accurately measure the causal strength between two genes, a measure based on Kullback–Leibler (KL) divergence was proposed (39). Before giving the definition of the measure, the interventional probability is described firstly. In a directed acyclic graph (DAG), if variable \(Y\) is regulated by variable \(X\) both directly and indirectly through
Figure 1. Quantification of association (ASS) by CMI2. (A) The association between two variables (genes) X and Y in a two-node network measured by CMI2 is equal to MI and CMI. (B) The association between two variables (genes) X and Y in a triple-node network measured by CMI2 is equal to CMI. (C) CMI can exclude the indirect information from direct information, but it underestimates the association between X and Y when the expression level of Z is near or equal to the expression level of X or Y. In this case, CMI2 can accurately measure the association prior to both MI and CMI.

variable Z, the interventional probability of moving the link from X to Y is defined as

\[ P_{X\rightarrow Y}(x, y, z) = P(x, z) \sum_y P(y | z, x) P(x), \]  
(5)

where \( P(y | z, x) \) is a conditional probability distribution of Y given Z and X.

For the three variables mentioned above, the causal strength, i.e. regulation strength of the arrow from X to Y is defined (39) as

\[ C_{X\rightarrow Y}(X; Y|Z) = D_{KL}(P(X,Y,Z) || P_{X\rightarrow Y}(X,Y,Z)), \]  
(6)

where \( P(X, Y, Z) \) is the joint probability distribution of X, Y, and Z, \( P_{X\rightarrow Y}(X,Y,Z) \) is the interventional probability distribution of X, Y and Z for removing arrow X \( \rightarrow \) Y, and \( D_{KL}(P || P_{X\rightarrow Y}) \) is KL-divergence from \( P(X,Y,Z) \) to \( P_{X\rightarrow Y}(X,Y,Z) \). With above definition, measure \( C_{X\rightarrow Y}(X; Y|Z) \) is unsymmetrical.

**Conditional mutual inclusive information for association measure**

Among the most popular association measures, MI tends to overestimate the regulation strengths between genes (false-positive problem), while CMI tends to underestimate the strengths (false-negative problem). As shown in Figure 1, MI can correctly quantify the regulation strength between genes X and Y for the case of Figure 1A but fails to quantify the association between genes X and Y for the case in Figure 1B due to the indirect regulation mediated by gene Z, i.e. overestimate the strength. Although CMI can successfully quantify the indirect regulation, it fails when the expression level of gene Z is near or equal to that of X (or Y), where the Pearson correlation coefficient between Z and X (or Y) is near or equal to 1. In that case as shown in Figure 1C, CMI will underestimate the regulation strength because the CMI value between genes X and Y given Z is near or equal to zero, which is actually incorrect. The newly proposed measure \( C_{X\rightarrow Y}(X; Y|Z) \) tried to address this issue by calculating the relative entropy distance. However, it can only partially address the issue (see Supplementary Data). In addition, it is difficult to put it into practice for large-scale network inference because of the prior information requirement of network directions (38,39,44).

In this work, to overcome the problems discussed above, we proposed an effective unbiased measure based on the causal strength (39), named CMI2, to quantify causal associations between genes. CMI2 is an association measure using the inclusive information, i.e. entropy relative distance between the postulated edge-existence distribution and edge-non-existence distribution. Next, we describe the definitions of CMI2.

In a DAG, if variable Y is regulated by variable X both directly and indirectly through variable Z, the association between X and Y is defined as

\[ CMI2(X, Y|Z) = (D_{KL}(P || P_{X\rightarrow Y}) + D_{KL}(P || P_{Y\rightarrow X})) / 2, \]  
(7)

where \( P = P(X,Y,Z) \) is the joint probability distribution of X, Y and Z, \( P_{X\rightarrow Y} = P_{X\rightarrow Y}(X,Y,Z) \) and \( P_{Y\rightarrow X} = P_{Y\rightarrow X}(X,Y,Z) \) are the interventional probability distributions of X, Y and Z for removing edges X \( \rightarrow \) Y and Y \( \rightarrow \) X, respectively. \( D_{KL}(P || P_{X\rightarrow Y}) \) and \( D_{KL}(P || P_{Y\rightarrow X}) \) are KL-divergences from \( P \) to \( P_{X\rightarrow Y} \) and \( P_{Y\rightarrow X} \). Similar to CMI, CMI2 has an order number \(|Z|\), i.e. the number of conditional variables Z, and MI can be regarded as zero-order CMI2.

The above quantity can be decomposed into three terms. One of them is CMI and another two are non-negative terms. The decomposition can be derived from the theoretical result as follows. For the three variables defined above, \( CMI2(X, Y|Z) \) between variables X and Y given Z can be decomposed into

\[ CMI2(X; Y|Z) = CMI(X; Y|Z) + \frac{1}{2} D_{KL}(P(Y|Z) || P_{X\rightarrow Y}(Y|Z)) + \frac{1}{2} D_{KL}(P(X|Z) || P_{Y\rightarrow X}(X|Z)). \]  
(8)

The proof of the above result can be found in the Supplementary Data. Equation (8) states that CMI2 is equal to CMI if the second and third terms of Equation (8) are zero, i.e. both X and Y are independent with Z. Since the KL-divergence is non-negative, CMI2 between X and Y given Z is no less than CMI between X and Y given Z.

Equation (8) also states that CMI and \( C_{X\rightarrow Y}(X; Y|Z) \) are unbiased estimates of CMI between X and Y given Z. Therefore, The third term in Equation (8) corrects the underestimation. Similarly, if X and Z are similar, Then second term in Equation (8) will correct the underestimation.

**Computation of CMI2**

As described above, CMI2 can be determined by computing the (joint) probabilities of genes X, Y and Z, which can be estimated with kernel density estimator to construct the probability density functions based on gene expression data (23,27,45). In this work, to efficiently estimate CMI2, we assumed that the gene expression profiles follows a multivariate Gaussian distribution, which has been widely accepted and
proved to be reasonable (43,46). Here, to approximate the Gaussian distribution, the log-transformation of gene expression data was adopted. According to the definition of KL-divergence, CMI2(X; Y|Z) can be rewritten as

$$CMI2(X; Y|Z) = \sum_{x,y,z} P(x, y, z) \ln \frac{P(x, y, z)}{P(x|z)P(y|z)P(y)},$$

where $P(y|z, x)$ and $P(x|z, y)$ are the conditional probabilities.

With the hypothesis of Gaussian distribution, CMI2 can be calculated based on the following result.

**Theorem 1.** Let $X$ and $Y$ be 1-dimension variables, $Z$ is a $n_z (n_z \geq 1)$-dimension variable, and $X$, $Y$ and $Z$ follow Gaussian distribution. Then

$$CMI2(X; Y|Z) = \frac{1}{4} \left( \text{tr}(C^{-1}\Sigma) + \text{tr}(C^\dagger) + \ln C_0 + \ln C_0 - 2n \right),$$

(10)

where

$$n = n_z + 2, C_0 = \rho_{xx} \left( (\Sigma^{-1})_{xx} - (\Sigma_1^{-1})_{xx} + \rho_{xx}^{-1} \right),$$

$$\hat{C}_0 = \rho_{yy} \left( (\Sigma^{-1})_{yy} - (\Sigma_1^{-1})_{yy} + \rho_{yy}^{-1} \right),$$

$$C = \begin{pmatrix} C_{xx} & C_{xy} & C_{xz} \\ C_{yx} & C_{yy} & C_{yz} \\ C_{zx} & C_{zy} & C_{zz} \end{pmatrix}^{-1}, \quad \hat{C} = \begin{pmatrix} \hat{C}_{yy} & \hat{C}_{yx} & \hat{C}_{yz} \\ \hat{C}_{xy} & \hat{C}_{xx} & \hat{C}_{xz} \\ \hat{C}_{zy} & \hat{C}_{zx} & \hat{C}_{zz} \end{pmatrix}^{-1},$$

$$C_{xx} = (\Sigma_1^{-1})^{-1}_{xx}, C_{yy} = 0, C_{xx} = (\Sigma_1^{-1})^{-1}_{xx},$$

$$C_{yx} = (\Sigma_1^{-1})^{-1}_{xy}, C_{yy} = 0, C_{xx} = (\Sigma_1^{-1})^{-1}_{xx},$$

$$C_{xz} = (\Sigma_1^{-1})^{-1}_{xz} - (\Sigma_1^{-1})^{-1}_{xx}, C_{yz} = (\Sigma_1^{-1})^{-1}_{yz} - (\Sigma_1^{-1})^{-1}_{yy} + \rho_{yz}^{-1},$$

$$C_{zx} = (\Sigma_1^{-1})^{-1}_{zx}, C_{zy} = (\Sigma_1^{-1})^{-1}_{zy},$$

$$\Sigma_1 = \begin{pmatrix} \rho_{xx} & \rho_{xz} \\ \rho_{xz} & \rho_{zz} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \rho_{yy} & \rho_{yz} \\ \rho_{yz} & \rho_{zz} \end{pmatrix}.$$
pair \(i\) and \(j\) has low or zero \(\text{CMI}_2\), we deleted the edge between them. Subsequently, the higher order \(\text{CMI}_2\) was calculated for a gene pair until there were no further changes in the network topology. The detailed algorithm for inferring a GRN was described in algorithm \(\text{CMI}_2\text{NI}\).

To reduce computational complexity but not sacrifice the accuracy for detecting 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 \(L\)-th order \((L \leq T)\) network, all the \(L\)-th order \(\text{CMI}_2\)s for the possible combinations of \(L\) conditional genes from \(T\) genes are computed and the maximal one or the geometric mean of them is selected to decide the existence of regulation. Generally, after a few number of iterations \(L\), the computation will terminate due to no change on the network topology, i.e. \(L < m < n\). In other words, we can obtain the network without resorting to any approximations in the computational procedure, as indicated in the \(\text{CMI}_2\text{NI}\) algorithm and also Figure 2. Theoretically, for a network with \(n\) genes or molecules \((\text{e.g.} \ n = 20\ 000\) genes\), it requires at least \(n\) independent samples to derive their direct regulations due to the computation of \(\text{CMI}_2(.,.,.,.,.,.,.,.,..)|\leq 2\) independent samples due to \(\text{CMI}_2(.,.,.,..)|\leq 2\). However, if the algorithm is converged to an \(L\)-th order network, it actually requires only \(L + 2\) independent samples due to \(\text{CMI}_2(.,..)|\leq L\). In other words, \(\text{CMI}_2\text{NI}\) can infer the network without any approximation in the above process even with a small number of samples, i.e. \(L + 2\) independent samples when converged to an \(L\)-th order network, where \(L\) is usually between 3 and 5 for many real networks or datasets, far less than \(n\). Thus, comparing to the traditional requirement of \(n\) independent samples for computing all statistic dependency among variables, our algorithm can obtain the network without the approximation with a small number of samples \(L + 2\).

---

**Algorithm (CMI\text{II}NI)**

**Input:**
- Gene expression matrix \(A\).
- Parameter for dependence threshold \(\theta\).

**Output:**
- Inferred gene network \(G\).
- Order of inferred network \(L\).

**Step-1.** Initialization. Generate the complete connected network \(G_0\) for all genes (i.e. the clique graph of all genes). Set \(L := 1\).

**Step-2.** For a nonzero edge \(G_0(i,j) \neq 0\), 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-3.** Set \(G := G_0\), \(G \neq G_0\). If \(T < L\), stop. If \(T \geq L\), select \(L\) genes from these \(T\) genes and let them as \(K = [k_1, \ldots, k_L]\). The number of all selections for \(K\) is \(C_T^L\). Compute the \(L\)-th order \(\text{CMI}_2(l,j|K)\) for all \(C_T^L\) selections, and choose the maximal one denoting as \(\text{CMI}_2\text{NI}(l,j|K)\). If \(\text{CMI}_2\text{NI}(l,j|K) < \theta\), set \(G(l,j) = 0\).

**Step-4.** If \(G = G_0\), stop; If \(G \neq G_0\), set \(G_0 := G\) and return to Step-2.

Sometimes there is no need to run the algorithm with high order networks, so we set a parameter for deciding the maximal order of the network in the software to terminate the algorithm according to the user’s need. More importantly, this will greatly reduce the computational complexity. The MATLAB implementation of the algorithm described above with detailed tutorials are freely available at [http://www.comp-sysbio.org/cmi2ni](http://www.comp-sysbio.org/cmi2ni).

**Datasets**

In order to validate our method, \(\text{CMI}_2\text{NI}\) was applied to simulation dataset as well as a real gene expression dataset. As for simulation data, the method was tested on the widely used reference network in \textit{Yeast} with synthetic nonlinear expression data from DREAM challenge (4). As for real gene expression data, we applied our method to the well-known SOS DNA repair network with the experimental dataset in \textit{E. coli} (47,48).

**Metrics for evaluation**

The performance of the proposed method was evaluated by the following measures, i.e. sensitivity (SN) 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

\[
\begin{align*}
\text{TPR} & = \frac{TP}{TP+FN}, \\
\text{FPR} & = \frac{FP}{FP+TN}, \\
\text{PPV} & = \frac{TP}{TP+FP}, \\
\text{ACC} & = \frac{(TP+TN)}{(TP+FP+TN+FN)}, \\
\text{MCC} & = \sqrt{\text{ACC}·\text{PPV}-(\text{TPR}·\text{FPR}+\text{TPR}·\text{FPR}+\text{TPR}·\text{FPR})/\sqrt{\text{ACC}·\text{PPV}}}
\end{align*}
\]

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. In addition, we compared \(\text{CMI}_2\text{NI}\) with several PC algorithm-based methods, such as pcalg (49) and PCA-CMI (28).

**RESULTS**

**Simulation study**

For simulation expression data, the widely used benchmark networks along with expression datasets from DREAM challenge were adopted here to evaluate our method. The gold standard networks were selected from source networks of real species. The expression data were generated with the nonlinear ordinary differential equation (ODE) systems in which the network structures were determined with detailed dynamics of both transcriptional and translational processes (50). In this work, the DREAM3 datasets about \textit{Yeast} knock-out gene expression data with sizes 10, 50 and 100 were used (4).

Firstly, we tested \(\text{CMI}_2\text{NI}\) on the \textit{Yeast} gene expression data with network size 10 and sample number 10. We chose 0.03 as the threshold value of mutual information and conditional mutual inclusive information to decide independence, and the order index was not constrained, i.e. the
algorithm terminated until there was no more higher order CMID to be computed. In order to evaluate the performance of CMIDNI, the AUC score was adopted. As shown in Figure 3A, CMIDNI performs best with an AUC score of 0.994, implying the efficiency of CMIDNI. In addition, we also compared CMIDNI with partial correlation coefficient-based PC algorithm (pcalg) (28) and conditional mutual information-based PC-algorithm (PCA-CMI) (28). The results can be found in Figure 3A, where we can see that CMIDNI is superior to the other methods. The detailed results of different approaches can be found in Table 1 with respect to PPV, ACC, MCC and AUC. From Table 1, we can see that CMIDNI and PCA-CMI perform comparatively well with respect to PPV, ACC and MCC, and both approaches outperform pcalc.

Secondly, we tested CMIDNI on the Yeast gene expression data with network size 50 and sample number 50. The network containing 50 nodes with 77 edges was selected from real and experimental verified networks. We set the threshold value 0.05 for MI and CMID to decide independence and the order index was not constrained. As shown in Figure 3B, CMIDNI outperforms other reference methods with the highest AUC score of 0.834. From Table 1, we can see that CMIDNI performs better than the other methods with respect to all the metrics listed. For example, CMIDNI achieved 0.492, 0.396, 0.397 and 0.834 for PPV, ACC, MCC and AUC, respectively.

Thirdly, we tested CMIDNI on the Yeast gene expression data with network size 100 and sample number 100. The reference network contains 100 nodes with 166 edges. We set the threshold value 0.03 for MI and CMID to decide independence and the order index was set to the first order which is a usually adopted order for large-scale networks to reduce the computational burden. As shown in Figure 3C, CMIDNI performs better than other reference methods with the highest AUC value of 0.856. The detailed results of different approaches can be found in Table 1 with respect to various metrics. From Table 1, we can see that CMIDNI performs better than the other two methods with the highest values 0.628, 0.972, 0.479 and 0.856 for PPV, ACC, MCC and AUC, respectively. In the 100-gene network from DREAM dataset, 14 edges were detected by CMID but missed by CMI. Furthermore, CMID successfully silenced 20 edges overestimated by CMI without reducing the true positive rate in the 100-gene network.

The results on all the three datasets with different network sizes from DREAM challenge demonstrated the effectiveness of our CMIDNI. Furthermore, the good performance of CMIDNI indicates that CMID, as a new measure of causal regulation strength, is superior to CMI.

Reconstruction of SOS network in Escherichia coli

Besides the above simulation datasets, CMIDNI was also applied to reconstruct gene networks from real gene expression data. We evaluated our CMIDNI on the well-known SOS DNA repair network which is an experimentally verified network in E. coli with real gene expression data (47,48).

The network is a 9-gene sub-network of SOS pathway in E. coli. The SOS pathway, which regulates cell survival and repair after DNA damage, involves the lexA and recA genes. There are more than 30 genes that are directly regulated by lexA and recA, while tens or even hundreds of other genes that are indirectly regulated by the two genes. Here, the nine transcripts in the test network include the principle mediators of the SOS response (lexA and recA), four other regulatory genes with known involvement in the SOS response (ssb, recF, dinI and unnuDC), and three sigma factor genes (rpoD, rpoH and rpoS) whose regulations play important roles in the SOS response. For the expression data, we chose the perturbation data, which were obtained after perturbations were applied to the test network in E. coli. In the perturbation, each of the nine genes in the test network is overexpressed with arabinose-controlled episomal expression plasmid, and the change in expression of each transcript relative to the unperturbed cells was accordingly measured using quantitative real-time polymerase chain reaction (qPCR).

The performance of CMIDNI was evaluated with the network of size 9 and the expression dataset generated under perturbations. We chose 0.01 as the threshold of both MI and CMID. To show the performances of CMIDNI and other methods, the true network and inferred networks were visualized with Cytoscape (51). Figure 4A shows the true network with 24 edges. Figure 4B–D shows the networks inferred by pcalc, PCA-CMI and CMIDNI, respectively, with corresponding ACC of 0.583, 0.694 and 0.722. From the results shown in Figure 4E, we can see that CMIDNI outperforms the other two methods significantly, where CMIDNI achieves the highest AUC of 0.802.

The detailed results of different methods can be found in Table 2 with respect to some indexes, such as PPV, ACC, MCC, etc. From Table 2, we can see that CMIDNI performs best on most metrics. Although pcalc achieves higher TPR with the same parameter, the higher FPR disables it as a good performer. Both PCA-CMI and CMIDNI outperform pcalc, implying the efficiency of CMI. The superior performance of CMIDNI over PCA-CMI demonstrates that CMID is better than CMI when quantifying the causal strength between variables.

Performance of CMIDNI

The theoretical analysis of CMID has proven that CMID can be decomposed into CMI and a non-negative term. This implies that CMID value is bigger than CMI. Hence, CMID can
Table 1. Comparison of different methods on networks with sizes 10, 50 and 100 in DREAM3 challenge

<table>
<thead>
<tr>
<th>Method</th>
<th>Size 10</th>
<th>Size 50</th>
<th>Size 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>pcalg</td>
<td>20</td>
<td>8</td>
<td>62</td>
</tr>
<tr>
<td>PCA-CMI</td>
<td>18</td>
<td>2</td>
<td>68</td>
</tr>
<tr>
<td>CMI2NI</td>
<td>18</td>
<td>2</td>
<td>68</td>
</tr>
</tbody>
</table>

Table 2. Comparison of different methods on SOS DNA repair network

<table>
<thead>
<tr>
<th>Method</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>TPR</th>
<th>FPR</th>
<th>PPV</th>
<th>MCC</th>
<th>ACC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>pcalg</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>12</td>
<td>0.500</td>
<td>0.250</td>
<td>0.800</td>
<td>0.239</td>
<td>0.583</td>
<td>0.701</td>
</tr>
<tr>
<td>PCA-CMI</td>
<td>16</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>0.667</td>
<td>0.250</td>
<td>0.842</td>
<td>0.393</td>
<td>0.694</td>
<td>0.792</td>
</tr>
<tr>
<td>CMI2NI</td>
<td>17</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>0.708</td>
<td>0.250</td>
<td>0.850</td>
<td>0.435</td>
<td>0.722</td>
<td>0.802</td>
</tr>
</tbody>
</table>

Figure 4. Comparison of CMI2NI with pcalg and PCA-CMI on the SOS network. (A) Benchmark (true) network. (B) Network inferred by pcalg. (C) Network inferred by PCS-CMI. (D) Network inferred by CMI2NI. In the network, a pink node represents a gene. (E) ROC curves of three methods, where the AUC score of CMI2NI is 0.802.

address the underestimation problem of CMI. To further test this theoretical conclusion, we investigated the value of CMI2 with both simulation and real expression datasets. Moreover, we compared values of CMI2 and CMI. For simulation dataset, we used a random dataset as well as the dataset from DREAM challenge. For real gene expression dataset, we used the expression data of SOS network in *E. coli*. Secondly, we investigated the values of CMI2 on the datasets from DREAM challenge and the real expression dataset for SOS network in *E. coli*. When computing the causal strength, we only calculated CMI2 or CMI for the edges in the network, where the means of all edges for CMI2s or CMIs were used for comparison. Figure 5B shows the values of CMI2 and CMI on the five datasets, from which we can clearly see that CMI2 is indeed higher than CMI.

Case study: reconstruction of cancer-specific gene regulatory network

It is well recognized that most complex traits are caused by the dysfunction of certain functional modules and pathways (52), where the gene regulatory circuit may be rewired in the diseases. Since the gene regulatory network provides a global view of the gene regulations, we hereby investigated how the cancer genes are regulated by constructing a gene regulatory network for cancer. As a case study, we applied CMI2NI to build a GRN for acute myeloid leukemia (AML) based on the RNA sequencing data of a large cohort of AML patients from TCGA.
Figure 6. AML-specific gene regulatory network reconstructed by CMI2NI.

The Level-3 processed data was used here, and the RPKM (read per kilobase of exon per million mapped reads) value was used as the gene expression value.

Figure 6 shows the AML-specific GRN constructed by CMI2NI. In particular, we considered the 81 cancer genes involved in a network built by RACER (55), where the network was constructed based on the same AML gene expression dataset used here. In the AML-specific GRN, there are 16 regulators and 65 target genes, and we inferred the regulation relationships between the regulators and genes with CMI2NI. In total, we detected 550 regulations, among which 113 regulations have been reported by RACER. Compared with the network inferred by RACER, there are two hub regulators NRSF and BCLAF1 in the GRN constructed by CMI2NI, where these two regulators respectively target 39 and 35 genes. In the GRN built by RACER, the degrees of these two regulators are small. By investigating the target genes of NRSF and BCLAF1 from our constructed GRN, we want to see which pathways these two regulators are involved in and how they are related to cancer. With the cancer gene annotation system CaGe (http://mgrc.kribb.re.kr/cage/), we noticed that the genes targeted by NRSF are significantly enriched in cancer pathways and the top two ranked pathways are leukemia-specific pathways (detailed results can be found in Supplementary Tables S1). Similarly, BCLAF1 target genes are also significantly enriched in leukemia-specific pathways (detailed results can be found in Supplementary Tables S2). Despite these two regulators have not been reported directly to be related to leukemia, the analysis of their target genes implies that they play important roles in leukemia. In the future, these two regulators can be considered when designing target therapy for leukemia.

DISCUSSION

The information theory-based association measures, e.g. MI and CMI have been widely used to infer gene regulatory networks. However, these measurements either under-estimate or overestimate the associations between variables (5,56). In this paper, we proposed a novel association concept, namely CMI2, to accurately quantify the dependency between a pair of variables. CMI2 provides a natural generalization of correlation and is capable of characterizing the nonlinear dependency between variables that is common in biology. Furthermore, CMI2NI can accurately quantify the causal strengths or correlations between gene-pairs so that the indirect regulation can be eliminated, which is the key point to improve the accuracy of GRN inference. With CMI2, we developed a network inference algorithm, namely CMI2NI, to infer gene regulation networks. With the power of PC algorithm, CMI2NI can also keep the natural sparsity of biological networks. The benchmark results show that CMI2NI outperforms other popular approaches, implying the effectiveness of CMI2NI. Considering the complex functional relationships among genes (57), we also constructed an AML-specific GRN with CMI2NI to see how the cancers genes are regulated. Similar to other network-based approaches, such as network differentiation (58–63) and dynamical network biomarkers (64–66), investigating the regulatory circuit of cancer genes provides new insights into the cancer pathogenesis.

Despite the advantages of CMI2NI, we notice there is still room to improve it. Firstly, similar to PCA-CMI, CMI2NI cannot directly infer edge directionality, which is also a general problem of many other methods, especially for those not working on time series data (23). Secondly, it is still a challenge task to select the conditional genes in an optimization way. In the PC-algorithm, genes i, j and their neighbours are randomly selected when calculating association. For example, there are T (T ≥ 1) genes which are adjacent with both genes i and j. When constructing the Lth-order (L ≤ T) network, all the Lth-order CMI2s for the possible combinations of L conditional genes from T genes are computed, and the maximal one or the geometric mean of them is used to describe the regulation strength. However, the selection of conditional genes may affect the performance of PC algorithm.

AVAILABILITY

The software CMI2NI is freely accessible at http://www.comp-sysbio.org/cmi2ni.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their constructive comments, which greatly help us improve our manuscript, Dr Dominik Janzing for the discussion on the model of causal strength, Dr Yue Li for providing the leukemia dataset and Ms Peipei Xiao for the technique help.

FUNDING

Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences (CAS).
REFERENCES


