<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Zhang, Xiajun; Zhao, Juan; Hao, Jin-Kao; Zhao, Xing-Ming; Chen, Luonan</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2014-12-24</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/39094">http://hdl.handle.net/10220/39094</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.</td>
</tr>
</tbody>
</table>
Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks

Xiujun Zhang\textsuperscript{1,2,3}, Juan Zhao\textsuperscript{1}, Jin-Kao Hao\textsuperscript{4}, Xing-Ming Zhao\textsuperscript{5,*} and Luonan Chen\textsuperscript{1,6,*}

\textsuperscript{1}Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, \textsuperscript{2}Department of Mathematics, Xinyang Normal University, Xinyang 464000, China, \textsuperscript{3}School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459, Singapore, \textsuperscript{4}LERIA, Department of Computer Science, University of Angers, Angers 49045, France, \textsuperscript{5}Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China and \textsuperscript{6}Collaborative 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 (CMI\textsuperscript{2}), to describe the regulations between genes. Furthermore, with CMI\textsuperscript{2}, we develop a new approach, namely CMI\textsuperscript{2}NI (CMI\textsuperscript{2}-based network inference), for reverse-engineering GRNs. In CMI\textsuperscript{2}NI, CMI\textsuperscript{2} 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 \textit{Escherichia coli} demonstrate the superior performance of CMI\textsuperscript{2}NI. CMI\textsuperscript{2}NI was also used to reconstruct cancer-specific GRNs using gene expression data from The Cancer Genome Atlas (TCGA). CMI\textsuperscript{2}NI is freely accessible at http://www.comp-sysbio.org/cmi2ni.

INTRODUCTION

Identifying the causal regulations between genes is the key to understand 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 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

\textsuperscript{*}To whom correspondence should be addressed. Tel: +86 21 6436 5937; Fax: +86 21 5492 0120; Email: lchen@sibs.ac.cn

Correspondence may also be addressed to Xing-Ming Zhao. Tel: +86 21 6958 3959; Fax: +86 21 6958 3959; Email: xm_zhao@tongji.edu.cn

\textcopyright{} The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
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 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, Brazzel 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. CMII2 (conditional mutual inclusive information), to quantify the causal associations between variables. To infer GRNs, a new algorithm, namely CMII2NI, is developed by combining CMII2 with path consistency (PC) algorithm. CMII2NI, 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, CMII2 can be calculated by a concise formula involving the covariance matrices of the related gene expression profiles. The proposed network inference method CMII2NI 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 CMII2NI. As a case study, CMII2NI 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 quantifies the causal associations between variables. To infer the inverse of correlation matrix will destroy 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). 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. CMII2 (conditional mutual inclusive information), to quantify the causal associations between variables. To infer GRNs, a new algorithm, namely CMII2NI, is developed by combining CMII2 with path consistency (PC) algorithm. CMII2NI, 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, CMII2 can be calculated by a concise formula involving the covariance matrices of the related gene expression profiles. The proposed network inference method CMII2NI 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 CMII2NI. As a case study, CMII2NI 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.

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

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

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

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

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 from $X$ to $Y$, 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

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

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, 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

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

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 $CMI2(X; Y | Z) = CMI(X; Y | Z) + D_{KL}(P(Y | Z) || P_{X\rightarrow Y}(Y | Z))$ underestimate the association strength. For instance, given that $Y$ and $Z$ are almost identical, we can get $CMI(X; Y | Z) \approx 0$ and $D_{KL}(P(Y | Z) || P_{X\rightarrow Y}(Y | Z)) \approx 0$. In other words, strong dependency between $X$ and $Z$ makes the influence of cause $Y$ almost invisible when looking at $CMI(X; Y | Z)$ and $D_{KL}(P(Y | Z) || P_{X\rightarrow Y}(Y | Z))$. Therefore, the third term in Equation (8) corrects the underestimation. Similarly, if $X$ and $Z$ are similar, the 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 follow a multivariate Gaussian distribution, which has been widely accepted and
proven 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

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

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 \(X\) and \(Y\) follow Gaussian distribution. Then

\[
CMI2(X; Y|Z) = \frac{1}{4} \left( \text{tr}(C^{-1} \Sigma_{xx}) + \text{tr}(\tilde{C}^{-1} \tilde{\Sigma}) + \ln C_0 + \ln \tilde{C}_0 - 2n \right),
\]

where

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

\[
\tilde{C}_0 = \rho_{yy} \left( (\tilde{\Sigma}^{-1})_{yy} - (\tilde{\Sigma}^{-1})_{yy} \right),
\]

\[
C = \left( C_{xx} C_{xy} C_{xz} C_{yx} C_{yy} C_{yz} C_{zx} C_{zz} \right)^{-1},
\]

\[
\tilde{C} = \left( \tilde{C}_{xx} \tilde{C}_{xy} \tilde{C}_{xz} \tilde{C}_{yx} \tilde{C}_{yy} \tilde{C}_{yz} \tilde{C}_{zx} \tilde{C}_{zz} \right)^{-1},
\]

\[
c_{xx} = (\Sigma^{-1})_{xx}, c_{yy} = 0, c_{xy} = (\Sigma^{-1})_{xy},
\]

\[
c_{yx} = (\Sigma^{-1})_{yx}, c_{xz} = (\Sigma^{-1})_{xz} - (\Sigma^{-1})_{xz} + \rho_{xx}^{-1},
\]

\[
c_{zy} = (\Sigma^{-1})_{zy} - (\Sigma^{-1})_{zy} + \rho_{yy}^{-1},
\]

\[
c_{zz} = (\Sigma^{-1})_{zz} - (\Sigma^{-1})_{zz} + \rho_{zz}^{-1},
\]

\[
x_{xy} = (\Sigma^{-1})_{xy} - (\Sigma^{-1})_{xy} + \rho_{xx}^{-1},
\]

\[
x_{yx} = (\Sigma^{-1})_{yx} - (\Sigma^{-1})_{yx} + \rho_{yy}^{-1},
\]

\[
x_{xz} = (\Sigma^{-1})_{xz} - (\Sigma^{-1})_{xz} + \rho_{xx}^{-1},
\]

\[
x_{zy} = (\Sigma^{-1})_{zy} - (\Sigma^{-1})_{zy} + \rho_{yy}^{-1},
\]

\[
x_{zz} = (\Sigma^{-1})_{zz} - (\Sigma^{-1})_{zz} + \rho_{zz}^{-1},
\]

\[
\bar{c}_{xx} = (\tilde{\Sigma}^{-1})_{xx}, \bar{c}_{yy} = 0, \bar{c}_{xy} = (\tilde{\Sigma}^{-1})_{xy},
\]

\[
\bar{c}_{yx} = (\tilde{\Sigma}^{-1})_{yx}, \bar{c}_{xz} = (\tilde{\Sigma}^{-1})_{xz} - (\tilde{\Sigma}^{-1})_{xz} + \rho_{xx}^{-1},
\]

\[
\bar{c}_{zy} = (\tilde{\Sigma}^{-1})_{zy} - (\tilde{\Sigma}^{-1})_{zy} + \rho_{yy}^{-1},
\]

\[
\bar{c}_{zz} = (\tilde{\Sigma}^{-1})_{zz} - (\tilde{\Sigma}^{-1})_{zz} + \rho_{zz}^{-1},
\]

\[
\bar{x}_{xy} = (\tilde{\Sigma}^{-1})_{xy} - (\tilde{\Sigma}^{-1})_{xy} + \rho_{xx}^{-1},
\]

\[
\bar{x}_{yx} = (\tilde{\Sigma}^{-1})_{yx} - (\tilde{\Sigma}^{-1})_{yx} + \rho_{yy}^{-1},
\]

\[
\bar{x}_{xz} = (\tilde{\Sigma}^{-1})_{xz} - (\tilde{\Sigma}^{-1})_{xz} + \rho_{xx}^{-1},
\]

\[
\bar{x}_{zy} = (\tilde{\Sigma}^{-1})_{zy} - (\tilde{\Sigma}^{-1})_{zy} + \rho_{yy}^{-1},
\]

\[
\bar{x}_{zz} = (\tilde{\Sigma}^{-1})_{zz} - (\tilde{\Sigma}^{-1})_{zz} + \rho_{zz}^{-1},
\]

\[
\Sigma_1 = \left( \begin{array}{cc} \rho_{xx} & \rho_{xz} \\ \rho_{zx} & \rho_{zz} \end{array} \right), \Sigma_1 = \left( \begin{array}{cc} \rho_{yy} & \rho_{yz} \\ \rho_{zy} & \rho_{zz} \end{array} \right),
\]

\[
\Sigma = \left( \begin{array}{cc} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{array} \right), \Sigma = \left( \begin{array}{cc} \rho_{xy} & \rho_{yz} \\ \rho_{yx} & \rho_{yy} \end{array} \right).
\]

The proof of Theorem 1 can be found in the Supplementary Data. With Equation \((10)\), the CMI2 can be calculated in a very efficient way with the general hypothesis of Gaussian distribution underlying gene expression data.
pair \( i \) and \( j \) has low or zero CM12, we deleted the edge between them. Subsequently, the higher order CM12 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 CMI2NI.

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 CMIs 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 CMI2NI algorithm and also Figure 2. Theoretically, for a network with \( n \) genes or molecules (e.g. \( n = 20,000 \) genes), it requires at least \( n \) independent samples to derive their direct regulations due to the computation of CM12\((.,...,n−2)\), which clearly is not available for most of real cases. However, if the algorithm is converged to an \( L \)th-order network, it actually requires only \( L + 2 \) independent samples due to CM12\((...,L)\). In other words, CMI2NI 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 (CMI2NI)

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. \( L := L + 1 \). 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 \( G := G_0 \). If \( T < L \), stop. If \( T \geq L \), select out \( L \) genes from these \( T \) genes and let them as \( K = [k_1, \ldots, k_L] \). The number of all selections for \( K \) is \( C_T^K \). Compute the \( L \)th-order CM12(i, j|K) for all \( C_T^K \) selections, and choose the maximal one denoting as CM12max(i, j|K). If CM12max(i, j|K) < \( \theta \), set \( G(i, j) = 0 \).

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.

Datasets

In order to validate our method, CMI2NI 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 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 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

\[
\text{TPR} = \frac{TP}{TP + FN}, \\
\text{FPR} = \frac{FP}{FP + TN}, \\
\text{PPV} = \frac{TP}{TP + FP}, \\
\text{ACC} = \frac{(TP + TN)}{(TP + FN + FP + TN)}, \\
\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.
\]

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. In addition, we compared CMI2NI 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 Yeast knock-out gene expression data with sizes 10, 50 and 100 were used (4).

Firstly, we tested CMI2NI on the 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 CM12 to be computed. In order to evaluate the performance of CM12NI, the AUC score was adopted. As shown in Figure 3A, CM12NI performs best with an AUC score of 0.994, implying the efficiency of CM12NI. In addition, we also compared CM12NI with partial correlation coefficient-based PC algorithm (pcalg) (49) and conditional mutual information-based PC-algorithm (PCA-CMI) (28). The results can be found in Figure 3A, where we can see that CM12NI 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 CM12NI and PCA-CMI perform comparatively well with respect to PPV, ACC and MCC, and both approaches outperform pcalg.

Secondly, we tested CM12NI 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 CM12 to decide independence and the order index was not constrained. As shown in Figure 3B, CM12NI outperforms other reference methods with the highest AUC score of 0.834. From Table 1, we can see that CM12NI performs better than the other methods with respect to all the metrics listed. For example, CM12NI achieved 0.492, 0.396, 0.397 and 0.834 for PPV, ACC, MCC and AUC, respectively.

Thirdly, we tested CM12NI 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 of MI and CM12 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, CM12NI 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 CM12NI 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 CM12 but missed by CMI. Furthermore, CM12 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 CM12NI. Furthermore, the good performance of CM12NI indicates that CM12, as a new measure of causal regulation strength, is superior to CMI.

Reconstruction of SOS network in Escherichia coli

Besides the above simulation datasets, CM12NI was also applied to reconstruct gene networks from real gene expression data. We evaluated our CM12NI 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 recA 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 un-perturbed cells was accordingly measured using quantitative real-time polymerase chain reaction (qPCR).

The performance of CM12NI 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 CM12. To show the performances of CM12NI 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 pcalg, PCA-CMI and CM12NI, respectively, with corresponding ACC of 0.583, 0.694 and 0.722. From the results shown in Figure 4E, we can see that CM12NI outperforms the other two methods significantly, where CM12NI 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 CM12NI performs best on most metrics. Although pcalg achieves higher TPR with the same parameter, the higher FPR disables it as a good performer. Both PCA-CMI and CM12NI outperform pcalg, implying the efficiency of CMI. The superior performance of CM12NI over PCA-CMI demonstrates that CM12 is better than CMI when quantifying the causal strength between variables.

Performance of CM12NI

The theoretical analysis of CM12 has proven that CM12 can be decomposed into CMI and a non-negative term. This implies that CM12 value is bigger than CMI. Hence, CM12 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>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>TPR</th>
<th>FPR</th>
<th>PPV</th>
<th>ACC</th>
<th>MCC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pcalg</td>
<td>20</td>
<td>8</td>
<td>62</td>
<td>0</td>
<td>1.000</td>
<td>0.114</td>
<td>0.714</td>
<td>0.911</td>
<td>0.795</td>
<td>0.991</td>
</tr>
<tr>
<td>PCA-CMI</td>
<td>18</td>
<td>2</td>
<td>68</td>
<td>2</td>
<td>0.900</td>
<td>0.028</td>
<td>0.900</td>
<td>0.956</td>
<td>0.871</td>
<td>0.991</td>
</tr>
<tr>
<td>CM12NI</td>
<td>18</td>
<td>2</td>
<td>68</td>
<td>2</td>
<td>0.900</td>
<td>0.028</td>
<td>0.900</td>
<td>0.956</td>
<td>0.871</td>
<td>0.991</td>
</tr>
<tr>
<td>Size 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pcalg</td>
<td>66</td>
<td>162</td>
<td>2134</td>
<td>88</td>
<td>0.428</td>
<td>0.071</td>
<td>0.289</td>
<td>0.898</td>
<td>0.299</td>
<td>0.782</td>
</tr>
<tr>
<td>PCA-CMI</td>
<td>72</td>
<td>78</td>
<td>2218</td>
<td>82</td>
<td>0.467</td>
<td>0.034</td>
<td>0.480</td>
<td>0.934</td>
<td>0.438</td>
<td>0.810</td>
</tr>
<tr>
<td>CM12NI</td>
<td>78</td>
<td>80</td>
<td>2216</td>
<td>76</td>
<td>0.506</td>
<td>0.035</td>
<td>0.493</td>
<td>0.936</td>
<td>0.466</td>
<td>0.834</td>
</tr>
<tr>
<td>Size 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pcalg</td>
<td>152</td>
<td>252</td>
<td>9316</td>
<td>180</td>
<td>0.457</td>
<td>0.026</td>
<td>0.376</td>
<td>0.956</td>
<td>0.392</td>
<td>0.829</td>
</tr>
<tr>
<td>PCA-CMI</td>
<td>114</td>
<td>96</td>
<td>9472</td>
<td>218</td>
<td>0.343</td>
<td>0.010</td>
<td>0.542</td>
<td>0.968</td>
<td>0.416</td>
<td>0.849</td>
</tr>
<tr>
<td>CM12NI</td>
<td>128</td>
<td>76</td>
<td>9492</td>
<td>204</td>
<td>0.385</td>
<td>0.007</td>
<td>0.627</td>
<td>0.971</td>
<td>0.478</td>
<td>0.855</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>CM12NI</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 CM12NI 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 CM12NI. In the network, a pink node represents a gene. (E) ROC curves of three methods, where the AUC score of CM12NI is 0.802.

Figure 5. Comparison of methods CM12 and CMI in quantifying causal strengths in different datasets. Blue and red bars represent CM12 and CMI respectively. (A) Comparison on five replicate random datasets. (B) Comparison on random, DREAM and SOS datasets.

address the underestimation problem of CMI. To further test this theoretical conclusion, we investigated the value of CM12 with both simulation and real expression datasets. Moreover, we compared values of CM12 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*.

Firstly, we generated random variables with different samples 5, 10, 100 and 500, we took one or more of them as conditional variable(s). We computed their CM12 and CMI values using these datasets. The calculation was performed 100 times on each dataset, and the mean of the CM12 and CMI values were used for comparison. Figure 5A gives the results of CM12 and CMI values. From the histogram, we can find that for all the types of datasets with different samples, CM12 values are always larger than CMI values, which is consistent with theoretical analysis of CM12.

Secondly, we investigated the values of CM12 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 CM12 or CMI for the edges in the network, where the means of all edges for CM12s or CMIs were used for comparison. Figure 5B shows the values of CM12 and CMI on the five datasets, from which we can clearly see that CM12 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 CM12NI to build a GRN for acute myeloid leukemia (AML) based on the RNA sequencing data of a large cohort of AML patients from TCGA (http://cancergenome.nih.gov).
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 sparseness 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 \(T (T \geq 1)\) genes which are adjacent with both genes \(i, j\) and their neighbours are randomly selected when calculating association. In the 1st-order \((L = 1)\) network, all the 1st-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 regression 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)

Figure 6. AML-specific gene regulatory network reconstructed by CMI2NI.
REFERENCES


