RESEARCH ARTICLE

# Gene Regulatory Network Inferences Using a Maximum-Relevance and Maximum-Significance Strategy

**Wei Liu, Wen Zhu, Bo Liao*, Xiangtao Chen**

College of Information Science and Engineering, Hunan University, Changsha, Hunan, 410082, China

* dragonbw@163.com

OPEN ACCESS

**Citation:** Liu W, Zhu W, Liao B, Chen X (2016) Gene Regulatory Network Inferences Using a Maximum-Relevance and Maximum-Significance Strategy. PLoS ONE 11(11): e0166115. doi:10.1371/journal.pone.0166115

**Editor:** Attila Csikász-Nagy, King's College London, UNITED KINGDOM

**Received:** January 7, 2016

**Accepted:** October 24, 2016

**Published:** November 9, 2016

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This study is supported by the National Nature Science Foundation of China (Grant Nos. 11171369, 61272395, 61370171, 61300128, 61472127 and 61572178), National Nature Science Foundation of Hunan Province (Grant No. 12JJ2041) and the Planned Science and Technology Project of Hunan Province (Grant Nos. 2014FJ3079). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

Recovering gene regulatory networks from expression data is a challenging problem in systems biology that provides valuable information on the regulatory mechanisms of cells. A number of algorithms based on computational models are currently used to recover network topology. However, most of these algorithms have limitations. For example, many models tend to be complicated because of the "large p, small n" problem. In this paper, we propose a novel regulatory network inference method called the maximum-relevance and maximum-significance network (MRMSn) method, which converts the problem of recovering networks into a problem of how to select the regulator genes for each gene. To solve the latter problem, we present an algorithm that is based on information theory and selects the regulator genes for a specific gene by maximizing the relevance and significance. A first-order incremental search algorithm is used to search for regulator genes. Eventually, a strict constraint is adopted to adjust all of the regulatory relationships according to the obtained regulator genes and thus obtain the complete network structure. We performed our method on five different datasets and compared our method to five state-of-the-art methods for network inference based on information theory. The results confirm the effectiveness of our method.

## Introduction

The rapid development of high-throughput technologies has produced extensive gene expression data, and mining useful cell function information from these data has become a crucial goal in systems biology [1,2]. Specific physiological activity in cells occurs at the gene expression level. This physiological activity results from the interaction of a large number of genes and biological molecules and is not controlled by the gene itself. The sophisticated regulatory relationships between genes are often depicted in the form of gene regulatory networks. Therefore, gene network inferences are crucial to identifying regulatory relationships and understanding regulatory mechanisms [3,4].

**Competing Interests:** The authors have declared that no competing interests exist.

A gene network can be represented by a graph, G = {V, E}, where V and E represent the gene sets and the regulatory relationships between genes, respectively [5]. The graph depicts the gene network topology, which makes the interactions between genes more explicit. The main task of gene network inference methods is to recover accurate gene topologies from gene expression data and ensure their consistency with real gene networks. However, network inference is a challenging problem because of limitations in gene expression data; for example, when the number of samples is far less than the number of genes, an ill-posed network structure may result [6,7]. In addition, high noise and non-linear characteristics lead to inaccurate network inferences. Although these limitations may cause difficulties when performing network inferences, many inference methods of gene regulatory networks based on computational models have been reported in recent decades [8–12].

The Boolean network model is a simple model based on natural mechanisms. The expression value of each gene is either 0 or 1, and the interaction relationships between the genes are expressed by abstract Boolean logic, such as AND, OR, and NOT [13,14]. The Boolean network model was first proposed by Kauffman [15], who characterized the framework of the model. To reduce uncertainty in the data and model selection, Shmulevich et al. [16] extended the model and proposed the probabilistic Boolean network. Recently, a variety of computational methods have been introduced to improve Boolean network models, such as information theoretic [17,18], genetic algorithm [19,20], and literature-based methods [9]. Although Boolean networks are easy to implement, capturing complex system behaviors and recovering large-scale gene networks may be difficult using these network models [2,18].

The Bayesian network is a probabilistic graphical model that consists of a directed acyclic graph (DAG) and a conditional probability table (CPT). A DAG describes the causal relationships among a set of genes, and a CPT represents the conditional probability distribution of each gene based on its parent set [21]. The Bayesian network can be static or dynamic according to whether temporal expression profiles are used. Both static and dynamic Bayesian network modeling involve two components: structure learning and parameter learning. In this study, we mainly focused on the first component. Structure learning has become a challenging problem in recent decades. Two categories of methods can be used in structure learning: a constraint-based method [22,23], which adopts condition independence tests to capture the dependent relationships among genes; and a score-based method, which represents the measured probability of each structure from the given data. Recently, several score functions have been used, such as Akaike's information criteria [24], Bayesian information criteria [25], and minimal description length [26,27]. Although Bayesian network modeling has certain advantages with regard to noise data and incomplete data, it also has drawbacks, with the major ones being the exponential growth of the search space size of a DAG with respect to the number of genes and the difficulty of recovering large-scale networks [28].

Network models based on information theory are widely applied in network inference and measure regulatory relationships between genes by the dependency of all gene pairs. Mutual information (MI) is used as the measure of dependency because of its excellent performance in capturing complex dependency. Many gene network inference algorithms based on MI have been proposed [29–36]. When recovering networks based on these methods, the mutual information matrix (MIM) is first calculated, and then different strategies are applied to distinguish between real edges and false edges. One of the earliest of these methods is the relevance network approach [29], in which a proportion of inaccurate edges are eliminated according to a given threshold. However, the method cannot be used to assess indirect interactions between genes. The CLR algorithm presents an MI score based on an empirical distribution [31]. The ARACNE algorithm is used to measure data processing inequality [32] to eliminate indirect interactions. Luo et al. [33] produced a new statistical learning strategy, MI3, for the detection of more complex three-way relationships. Zhang et al. [34] presented a network inference algorithm based on conditional mutual information (CMI) to distinguish non-linear dependence between genes. Meyer et al. [35] employed a new dependency paradigm by introducing a feature selection technique based on maximum relevance/minimum redundancy criteria. Villaverde et al. [36] presented the network inference algorithm MIDER, which is based on entropy reduction. Methods based on information theory can effectively capture non-linear dependency and apply it to large-scale networks. Nevertheless, many models only consider pairwise interactions between genes and ignore other characteristics of the network.

Inspired by the network model based on information theory and a feature selection method known as maximum relevance-maximum significance (MRMS) [37], we propose a novel information—theoretic network inference method based on the MRMS, in which the problem of network recovery is converted into a process whereby the regulator genes for each target gene are selected. Our proposed strategy differs from that of the MRMS in terms of feature selection because it fully considers network topology characteristics, such as relevance, modularity, and sparseness. Therefore, our strategy can effectively select the regulator genes of a target gene, and it can then be applied for network recovery. Our method is compared with typical methods based on information theory, and the results demonstrate that our method outperforms these models.

## Methods

Gene network inference can be implemented by selecting the regulator genes for each gene; however, the key problem with this method is that it fails to present an effective selection strategy. In this section, we propose a new MRMS strategy based on information theory and then present a compact network inference method called the maximum-relevance and maximum-significance network (MRMSn) method, which is based on MRMS. To clearly describe this method, we first introduce the motivations of the method, its relevance and significance in networks and the concepts of information theory, which are the foundation of our proposed method.

### Motivation

Generally, traditional gene network inference methods recover network topologies by using a computational model that ensures that the final network topology is the closest match to the gene expression data. However, building an appropriate model is difficult because of the "large p, small n" problem [38]. The fundamental task of network inference is to identify all of the underlying regulatory relationships between genes. Therefore, if each gene is sequentially treated as the target gene, then the problem of inferring a network involving n genes can be decomposed into n sub-problems, and the task involves selecting the regulator genes for the given target gene. Designing an effective selection strategy is the key to the sub-problem. At present, many selection strategies are being applied with feature selection techniques to resolve the classification problem. To a certain extent, the sub-problem of selecting the regulator genes for a given target gene can be considered a two-class classification problem. Hence, the feature selection technique is used to resolve the sub-problem in this paper.

To select the correct regulator genes for the target gene in each sub-problem, a selection criterion with a high discriminating power should be designed. Although a large number of different criteria have been applied to feature selection problems and each criterion attempts to select the optimal feature sets that are suitable for all applications, obtaining the optimal features is not realistic. Therefore, the optimal feature selection criterion in the above sub-problem should fully reflect the characteristics of the network topology.

For real network topology, relevance and significance are the most basic characteristics. Regulator genes should have a high relevance with the target gene and directly connect with the target gene. Thus, relevance can be used to search for the regulator genes of the target gene. However, certain genes with high numerical relevance with the target gene may not be directly connected with the target gene. Obviously, these genes, which are called redundant genes, are not actual regulatory genes. This situation may lead to difficulties in selecting regulator genes based on the relevance of the gene network and subsequently degrade the accuracy of the network inference. Hence, redundant genes must be removed. Significance is another variable for determining whether a gene is the true regulator gene of the target gene. Regulator genes should have high significance with the module, which is composed of the target gene and the regulator genes. Therefore, a strategy that combines relevance and significance may effectively remove redundant genes and thereby improve the accuracy of the network inference.

## Relevance and significance

Relevance and significance are the most basic characteristics of a network. Relevance is usually used to measure the dependency between genes in the network, and significance reflects the influence of the genes on the network. For a network with a target gene and corresponding regulator genes, the relevance and significance can be defined as follows.

Let $G = \{g_1,\ldots,g_n\}$ denotes the set of $n$ genes of a given microarray dataset, which is used to infer the network. The relevance of gene $g_i$ with respect to the target gene $g_c$ is defined as follows:

$$\gamma_{g_i} = R(g_i, g_c) \tag{1}$$

where $R(\cdot)$ is a measurement that represents the dependency between two genes. Mutual information is one of the typical measures to define dependency of variables and we use it to characterize the relevance of genes in this paper.

Let $U = \{g_1,\ldots,g_r\}, U \subset G$ be the set of $r$ selected genes. For the target gene $g_c(g_c \in G \backslash U)$, the significance of gene $g_i$ among $U$ has the following form:

$$\varphi_{g_i} = S_U(g_i, g_c) = |E_U(g_c) - E_{U\backslash\{g_i\}}(g_c)| \tag{2}$$

where $E(\cdot)$ represents an energy estimation measure. The description of the energy measure can be different in different contexts. Entropy is an effective measurement to characterize energy dispersal, and it can be used to measure information content of variable in information theory. Thus, we use entropy to measure the energy estimation of genes in this study. Basically, the significance defined in Eq (2) reflects the change in energy estimation when a gene $g_i$ is removed from the gene set $U$.

## Information theory

Information theory was proposed by Claude E. Shannon and has become widely used in applied mathematics, electrical engineering and computer science. Entropy and MI are two fundamental concepts in information theory that are vital for the relationships and interpretation of data.

Entropy is a measure of the average uncertainty of a random variable. Here, $X$ represents a discrete random variable with alphabet $\chi$ and $p(x)$ represents the probability distribution function of $X$. The entropy of a random variable $X$ is defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3}$$

where the log is to the base 2 and the entropy is expressed in bits.

Because of the discrete random variable $Y$ and the conditional distribution function $p(y|x)$, the conditional entropy $H(Y|X)$ is defined as follows:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x, y) \log p(y|x) \tag{4}$$

The definition of conditional entropy can be extended to multiple random $X_1,\ldots,X_i$ and is defined as follows:

$$H(Y|X_1, \ldots, X_i) = -\sum_{y \in Y, x_1, \ldots, x_i \in X} p(y, x_1, \ldots, x_i) \log p(y|x_1, \ldots x_i) \tag{5}$$

$H(Y|X_1,\ldots,X_i)$ is monotonically decreasing, and it follows that

$$H(Y|X_1, \ldots, X_i) \le H(Y|X_1, \ldots, X_{i-1}) \tag{6}$$

MI is used to describe information that a random variable shares with another random variable and represents a measure of the dependency relationship between the two variables. The MI of two random variables $X$ and $Y$ is defined as follows:

$$I(X, Y) = H(Y) - H(Y|X) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) \tag{7}$$

where $p(x,y)$ is the joint probability distribution function of $X$ and $Y$ and $p(x)$ and $p(y)$ are the marginal probability functions of $X$ and $Y$, respectively.

## MRMS strategy

In real network topologies, each regulator gene should have high relevance to the target gene. Hence, the identification of the regulator genes for a target gene is to select a gene set $S$ with $K$ regulator genes $\{g_i\}$that have the greatest relevance to the target gene $g_c$. Let $G_c = \{g_1,\ldots,g_{c-1},g_{c+1},\ldots,g_n\}$ represents the candidate gene set containing all genes in $G$ except for target gene $g_c$. MI is used to represent the relevance between the target gene and the regulator gene. The regulator genes of target gene $g_c$ are the genes that satisfy Eq (8), which is called the maximum relevance criterion.

$$\max R(S, g_c),\ R = I(\{g_i, i = 1, \ldots, K\}, g_c) \tag{8}$$

Obviously, when $K$ equals 1, the selected gene is the gene that maximizes $I(g_i,g_c)(g_i \in G_c)$. When $K>1$, a simple selection scheme is adopted: suppose $V$ is a gene set with m-1 selected regulator genes, the $m$th regulator gene $g_m$ is selected by maximizing $R(\cdot)$ in Eq (9).

$$R(g_m, g_c) = \{I(g_m, g_c); g_m \in G_c \backslash V\} \tag{9}$$

Although the regulatory genes should have high relevance to the target gene, not all genes with high relevance to the target gene are true regulator genes. A likely occurrence is that certain genes selected on the basis of the maximum relevance criterion will not be directly connected with the target gene, and the genes will connect with target gene through a true regulator gene. Obviously, these genes, which are called redundant regulator genes, are not the correct regulator genes. Redundant regulator genes can degrade the accuracy of the network inferences and must therefore be removed.

Modularity is an important characteristic of regulatory network topologies. A module is composed of clustered genes. Compared with other genes outside of the module, the genes inside the module have more associations with each other. Typically, a module is a relatively balanced system in a real network, and the real gene members in this system have different influences on the system according to their differing importance levels. Theoretically, the members outside of the system will have no influence on the system. This influence is called the significance of the gene in the module. A module composed of a target gene and corresponding regulator genes can be a special module in which the target gene is the core of the module, which is regulated by all of the regulator genes. Every regulator gene provides a different contribution to the information of the target gene. Obviously, the significance of the gene can be used to discriminate if the gene is the true regulator gene for the target gene. Thus, defining significance in the special module is key for understanding the modularity of regulator genes and removing redundant regulator genes. Entropy is an effective method of measuring the variable amount of information in information theory and can be used to measure energy dispersal. Based on Eqs (2) and (6), we considered using an entropy reduction method to measure the significance.

For a target gene $g_c$ and the gene set $V$ with m-1 genes, the significance of gene $g_m$ has the following form:

$$S(g_m, g_c) = S_{VU\{g_m\}}(g_m, g_c) = H(g_c|V) - H(g_c|VU\{g_m\}) \tag{10}$$

Clearly, for target gene $g_c$, the regulator gene should be the top-ranked gene in the significance score (see Eq (10)). This strategy is called the maximum significance method, which is described as follows:

$$\max_{g_m \in G_c \setminus V} S(g_m, g_c) \tag{11}$$

Based on the above description of two criteria, we note that the maximum relevance criterion used to select for regulator genes may introduce redundant regulator genes. To remove the redundant regulator genes, we consider the maximal significance strategy because the correct regulator genes tend to score higher on significance. The strategy that combines the maximal significance strategy with the maximum relevance criterion is called the MRMS. Ideally, the correct regulator genes should have high relevance to the target gene and high significance in the module. Thus, the following form is provided to optimize $R$ and $S$ simultaneously:

$$\max \Phi(R, S), \Phi = R + S \tag{12}$$

For feature selection, combining two strategies to select the optimal regulator genes ensures that the selected regulator genes have maximal relevance and significance to the target gene. Therefore, identifying the regulator genes of a target gene is the same as selecting $K$ regulator genes from $G_c$ by maximizing $\Phi(\cdot)$ in Eq (12). In practice, this goal cannot be achieved, and only near-optimal regulator genes can be obtained. Inspired by the maximal relevance-minimal-redundancy criterion and considering the degree of relative importance between relevance and significance, another form is derived from Eq (12). Suppose a gene set $V$ with m-1 regulator genes has been selected. When selecting the $m$th gene, the selected gene $g_m$ has the following form:

$$\max_{g_m \in G_c \setminus V} [\alpha I(g_m, g_c) + (1 - \alpha)(H(g_c|V) - H(g_c|VU\{g_m\}))] \tag{13}$$

Based on Eq (13), we adopt the following first-order incremental search algorithm to select the regulator genes for the target gene $g_c$.

Step 1: The candidate gene set $G_c$ and the set of the selected regulator genes $V$ are initialized, and $G_c = G - g_c$ and $V = \phi$ are set.

Step 2: The relevance values $I(g_i,g_c)$ between the target gene $g_c$ and each candidate gene $g_i \in G_c$ are calculated, and the relevance values are ranked in descending order.

Step 3: The candidate gene with the largest relevance value is selected as the first gene of $V$ and removed from $G_c$.

Step 4: In the remaining genes of $G_c$, a gene $g_m$ that maximizes Eq (13) is selected. If the score in Eq (13) of the selected gene lies above the given score threshold $T_0$, then the gene is inserted into $V$ and removed from $G_c$. Otherwise, the selection procedure is terminated.

Step 5: Step 4 is performed again until the number of selected genes in $V$ is more than the given number $K$; otherwise, the selection procedure is terminated.

Note that the parameter $\alpha$ represents the weight of the relevance, and it is used to adjust the importance between the relevance and significance. Parameter $K$ represents the number of selected regulator genes. Considering the sparseness of the network, the number $K$ is set to $\lceil \log_2 n \rceil$, where $n$ is the number of genes in the network.

## Gene network inference based on the MRMS strategy

Inferring a gene network involves identifying all of the regulatory relationships between genes. In the above section, we describe an MRMS strategy that helps to identify the regulatory relationships of a target gene. The remaining work involves inferring the gene network topology based on this strategy. We propose a novel regulatory network inference method called MRMSn, which recovers gene networks through the following three stages.

The first stage is the regulatory relationship initialization. In this stage, a MIM is built according to Eq (7), and this MIM can reflect the likelihood of most direct regulatory relationships among genes. Generally, a greater value corresponds to a higher likelihood of a direct regulation relationship. However, the sparsity of the network means that only a few regulatory relationships occur between genes. Hence, a regulatory relationship with small values in the matrix must be removed. The most commonly used removal method is the selection of a unified threshold to remove false regulatory relationships. If the MI value is less than the threshold value, then the MI value is set to zero and the corresponding regulation relationship is removed. However, the numerical range of the MI value for each gene is different. When the unified threshold is too large, all regulatory relationships of certain genes may be lost, which is an undesired result. Therefore, in the network inference algorithm, we provide different thresholds for the regulatory relationships of different genes, and the threshold for the regulatory relationship of a given gene is set according to the MI values between the given gene and other genes. For a MIM $M(n \times n)$, $M_i$ denotes the MI values between gene $g_i$ and other genes. We choose parameter $\theta$ as the threshold for initializing regulatory relationships because it satisfies $\theta = \varepsilon \times \max(M_i)$, where $\varepsilon \in (0,1)$. Parameter $\varepsilon$ is typically set to a small value; specifically, it is set as 0.01 by default.

The second stage involves selecting the regulator genes for all genes. Each gene in a given gene dataset will in turn be a target gene, and the regulator genes of each target gene are selected using the MRMS strategy.

The last stage is to adjust all of the regulatory relationships and build a complete network structure. Although all of the regulatory genes of each target gene have been obtained, we cannot accurately obtain the complete network. A case that should be considered is when gene $g_i$ is the regulator gene of gene $g_j$ and gene $g_j$ is not the regulator gene of gene $g_i$. To resolve this problem, we provide a constraint to adjust the regulation relationships to obtain the complete network structure. The constraint is defined such that the regulatory relationship between gene $g_i$ and gene $g_j$ only occurs if gene $g_i$ is the regulator gene of gene $g_j$ and gene $g_j$ is the regulator gene of gene $g_i$. After the regulation relationships are adjusted according to this constraint, certain target genes may not have a regulator gene. To ensure that each target gene has at least one regulator gene, the first selected regulator gene of the target gene in the MRMS strategy is regarded as the regulated gene of the target gene.

The gene network inference method MRMSn is summarized in Table 1.

## Parameter tuning problem

The parameter $\alpha$ plays an important role in gene regulatory network inferences based on the MRMS strategy because it adjusts the degree of relative importance between the relevance and significance in the model. If $\alpha$ increases, the importance of the relevance increases, but the importance of the significance decreases. Thus, this parameter directly influences the performance of the proposed method.

To determine the optimum value of $\alpha$, an optimization method based on the local density is provided. The local density $\rho_c$ of the target gene $g_c$ is defined as follows:

$$\rho_c = \sum_{j=1}^{n} \chi(I(g_c, g_j) - d) \tag{14}$$

where $\chi(x) = 1$ if $x \geq 0$ and $\chi(x) = 0$ otherwise and $d$ is a cutoff distance. Basically, $\rho_c$ is equal to the number of genes with which the relevance of gene $g_j$ is more than $d$, which reflects the approximate number of regulator genes of gene $g_c$. Clearly, the value of $\rho_c$ will be affected by $d$, and the value of $d$ should be determined according to the datasets. Let $sm_{g_c}$ ($g_c \in G$) denote the sum of the MI between gene $g_c$ and any other genes. For each target gene $g_c \in G$, the $sm_{g_c}$ is calculated and gene $g_\gamma$ that satisfies Eq (15) is selected.

$$g_\gamma = \underset{g_c \in G}{\arg\min}\{sm_{g_c}\} \tag{15}$$

**Table 1. MRMSn for network inference.**

| Input: Microarray data $G = \{g_1, \ldots, g_n\}$ |
|---|
| The number of the selected regulator gene $K$ |
| The weight of network relevance $\alpha$ |
| The threshold of initializing regulatory relationships $\varepsilon$ |
| The threshold of scoring in MRMS $T_0$ |
| Output: A gene network |
| 1: Construct a MI matrix $M$ according to Eq (7). |
| 2: Adjust MI matrix $M$ using the threshold $\varepsilon$ |
| 3: for each gene $g_c$, $c \leftarrow 1$ to n do |
| 4: Select K regulator genes of gene $g_c$ using MRMS criterion |
| 5: end |
| 6: Adjust the regulation relationship using the constraint. |
| 7: Return the gene network. |

doi:10.1371/journal.pone.0166115.t001

Intuitively, the number of regulator genes of gene $g_\gamma$ may be lowest among all of the genes. In a sparse network, this number is typically one, although in a complex method, it is more than one. Thus, the value of $d$ can be calculated as follows:

$$d = \max_{g_j \in G} I(g_\gamma, g_j) \tag{16}$$

Basically, $d$ in Eq (16) is equal to the highest MI value between gene $g_\gamma$ and any other gene. When the value of $d$ is used in Eq (14), the density of certain genes that present the highest MI value among other genes with values lower than $d$ in Eq (16) may be zero. Thus, when the density of certain genes is zero, the value of $d$ is adjusted as follows:

$$d = \frac{1}{n-1}\sum_{j=1}^{n} I(g_\gamma, g_j) \tag{17}$$

After the cutoff distance $d$ is calculated, the density $\rho_c$ of each gene can be obtained. The parameter $\alpha$ is defined as follows:

$$\alpha = \frac{\sum_{i=1}^{n} \Gamma(\lceil \log_2 n \rceil - \rho_i)}{n} \tag{18}$$

where $\Gamma(\eta) = 1$ if $\eta>0$; otherwise, $\Gamma(\eta) = 0$. Considering that relevance and significance are the basic characteristics of the network, a constraint in which the value of $\alpha$ is varied from 0.1 to 0.9 is included. If $\alpha$ is less than 0.1, the final $\alpha$ is set to 0.1, and if $\alpha$ is more than 0.9, the final $\alpha$ is set at 0.9.

The above optimization method is applied to each dataset. The optimum value of $\alpha$ is 0.25 for Reaction chain with 4 species data, 0.90 and 0.10 for the DREAM3 10 genes data and DREAM3 50 genes data, respectively, and 0.20 and 0.78 for the IRMA benchmark data and SOS data, respectively.

Parameter $T_0$ represents the threshold of the score in Eq (13) and is used to select regulator genes. A gene is considered to be the regulator gene of a target gene when the gene score exceeds $T_0$. During the selection process, all scores in Eq (13) can be observed, and they are obtained in the process of selecting the regulator gene for each target gene. Based on extensive experiments, the value of $T_0$ is set to 70% of the maximum score. Thus, the optimum values of $T_0$ are 0.55, 0.09, 0.12, 0.51 and 0.18 for Reaction chain with 4 species data, DREAM3 10 gene data, DREAM3 50 gene data, IRMA benchmark data and SOS data, respectively.

## Computational complexity

In this section, we detail the computational complexity of the proposed method. Let $n$ be the number of genes in the network and $n'$ be the number of selected regulator genes in already-selected gene set. All of the relevance values of the target genes are obtained by calculating the mutual information matrix, which requires a $o(n^2)$ time complexity. The gene with the largest relevance value to the target gene is selected in step 3 of the MRMS strategy, and this step requires $o(n-1)$. After obtaining the first regulator gene, the genes in the candidate gene set are selected based on the maximum-relevance and maximum-significance strategy in step 4, and this step requires a $o(n-n')$ time complexity. Because the top $K$ regulator genes are selected for the target gene, the selection procedure in step 4 is executed $K-1$ times. In the last stage of the MRMSn, a time complexity of $o(2Kn + n)$ is required to adjust the relationships and build a complete network structure. Thus, when inferring the network with $n$ genes, the overall computational complexity is $o(n^2)+o(n((n-1)+(K-1)(n-n')))+o(2K+\text{n})$. Because the number of total genes is significantly larger than the number of selected regulator genes ($n >> n'$), the computational complexity can be approximated as $o(Kn^2)$.

## Experiments

To evaluate the performance of the proposed MRMSn, we compare the method with five algorithms: CLR, ARACNE, MRNET, MI3 and MIDER. The first two methods are based on information theory and are widely used in network inference. MRNET and MI3 are regarded as two typical methods in which the regulation relationships involving other regulator genes are identified. MIDER can detect interactions with entropy reduction. CLR, ARACNE and MRNET can be implemented in the R package MINET. To avoid unfair comparisons, we chose the default parameter eps = 0.0 in ARACNE [39]. For CLR and MRNET, we calculate the true positive rate (TPR) and the false positive rate (FPR) at different thresholds and select the optimum threshold, for which (TPR-FPR+1) is the maximum in the method. For the datasets in Table 2, the thresholds of CLR are set to 0.4959, 0.0, 0.1233, 0.2505, 0.25 and the thresholds of MRNET are set to 0.0, 0.0114, 0.0105, 0.0577, and 0.0. MI3 is also implemented in R with the package mi3. MIDER and the proposed MRMSn are performed in the MATLAB environment on a personal computer with an Intel core i7 (2.2 GHz) and 16 GB of RAM. The proposed MRMSn does not provide network directions, and the network direction in MI3 and MIDER is not considered.

### Data sets

Five datasets were used in the experiments to evaluate the performance of the method. All of the datasets included simulated data and real data obtained from previous studies [40–43]. Details of the datasets are given in Table 2.

Reaction chain with 4 species data [40] is a small reaction pathway that contains 10 samples for 4 variables. The true network is composed of 4 nodes and 3 edges.

DREAM3 10 gene data [41] represent a yeast network that contains 10 samples for 10 genes. The true network is composed of 10 nodes and 10 edges.

DREAM3 50 gene data [41] represent a yeast network that contains 50 samples for 50 genes. The true network is composed of 50 nodes and 77 edges.

IRMA benchmark data [42] represent a yeast synthetic network that contains 125 samples for 5 genes. The true network is composed of 10 nodes and 33 edges.

S0S data [43] represent an *Escherichia coli* network that contains 9 samples for 9 genes. The true network is composed of 9 nodes and 24 edges.

### Evaluation metrics

Our method must be compared with other methods that use different metrics to assess its performance. Therefore, we used the true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and accuracy (ACC). We let true positives (TP), true negatives (TN),

**Table 2. The details of the data sets used in our experiments.**

| Datasets | Variables | Samples | Type | network nodes | network edges |
|---|---|---|---|---|---|
| Reaction chain with 4 species | 4 | 10 | Simulated | 4 | 3 |
| DREAM3 10 genes | 10 | 10 | Simulated | 10 | 10 |
| DREAM3 50 genes | 50 | 50 | Simulated | 50 | 77 |
| IRMA benchmark | 5 | 125 | Real | 5 | 7 |
| S0S | 9 | 9 | Real | 9 | 24 |

doi:10.1371/journal.pone.0166115.t002

false positives (FP), and false negatives (FN) denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. These measures are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{19}$$

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

$$PPV = \frac{TP}{TP + FP} \tag{21}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{22}$$

## Experimental results

To fairly evaluate the performance of our proposed method, we compared it with other methods using simulation data and real data. It was important to consider network size when choosing these datasets. The number of network nodes varies from 4 to 50 in our datasets. Network sparsity is an important aspect that should be considered. Network datasets that have different degrees of sparsity were chosen to reflect the scalability of our method. The experiment process for different datasets is described in detail in the following subsection.

**Reaction chain with 4 species.** The proposed method was tested on a reaction chain that includes data for 4 species to verify how well it performs in special networks, such as linear chain networks. In this experiment, we chose 0.25 as the weight of the network relevance and 0.55 as the threshold of scoring in the MRMS according to the optimization methods proposed above. We set $2(\lceil \log_2 4 \rceil)$ as the number of the selected regulator genes. Fig 1 shows the network structure of different methods to facilitate a visual observation of the network topology. We found that MRMSn, CLR, ARACNE and MIDER can infer the same network topology with the true network, although redundant edges are produced in MRNET and MI3. To further

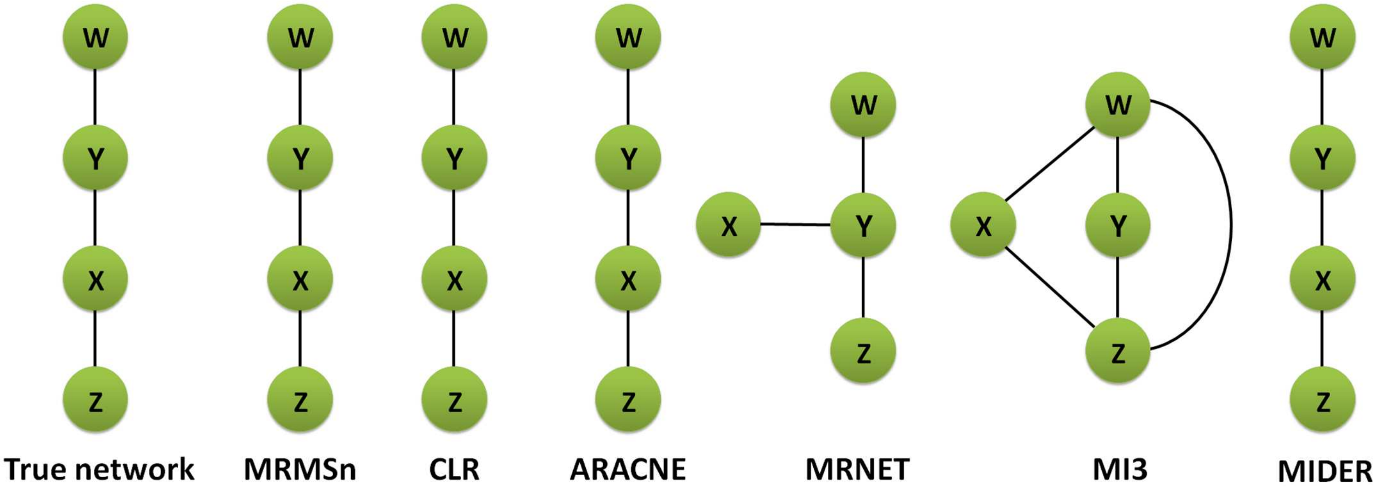

**Fig 1. Comparison of different methods on reaction chain with 4 species dataset.**

doi:10.1371/journal.pone.0166115.g001

**Table 3. Comparison of the performance by different methods on the reaction chain with 4 species dataset.**

| | TP | FP | TPR | FPR | PPV | ACC |
|---|---|---|---|---|---|---|
| CLR | 3 | 0 | 1 | 0 | 1 | 1 |
| ARACNE | 3 | 0 | 1 | 0 | 1 | 1 |
| MRNET | 3 | 1 | 1 | 0.333 | 0.750 | 0.833 |
| MI3 | 2 | 3 | 0.667 | 1 | 0.400 | 0.333 |
| MIDER | 3 | 0 | 1 | 0 | 1 | 1 |
| MRMSn | **3** | **0** | **1** | **0** | **1** | **1** |

doi:10.1371/journal.pone.0166115.t003

reflect the efficiency of our method, the performance of MRMSn with regard to evaluation metrics was compared with the listed methods. Table 3 presents the comparison results and shows that CLR, ARACNE, MIDER and MRMSn can identify all of the correct edges and do not produce redundant edges (TP = 3, FP = 0). For MRNET and MI3, the values of ACC are 0.833 and 0.333, respectively, and these values are less than that of MRMSn (ACC = 1), which indicates that the proposed method performs better than the CLR and MRNET methods.

**DREAM3 challenge network.** The DREAM project provides benchmarks and tools that can be used for the rigorous testing of gene network inference methods. To assess the ability of the MRMSn to analyze simulation data, we tested the proposed method with the DREAM3 challenge, which provides three sub-challenges with networks of sizes 10, 50, and 100. Only the results of the first two sub-challenges are presented here.

First, we tested the MRMSn method with the yeast dataset that contains 10 samples for 10 genes. We set 0.90 as the weight of the network relevance and 0.09 as the threshold of scoring in MRMS. The number of the selected regulator genes for the target gene can be calculated as 4 ($\lceil \log_2 10 \rceil$). The network topologies of the different methods for the dataset are shown in Fig 2, which shows that the MRMSn method can select all of the correct edges except for edge G4–G9. In addition, edge G2–G9 is the only redundant edge. The MRMSn method clearly performs better than the other methods. To support this finding, Table 4 presents a comparison of the results in the evaluation metric among the different methods. MRMSn can obtain 9 correct edges (TP = 9), whereas the other methods only obtain 6–8 correct edges. Note that MIDER cannot be used with the dataset. In addition, MRMSn only produces 1 redundant edge (FP = 1), whereas the other methods produce 6–12 redundant edges. The accuracy of our method is 0.956, which is much higher than the accuracy values of the other methods.

Then, MRMSn was tested on the yeast dataset that contains 50 samples for 50 genes. In the experiment, we chose 0.10 as the weight of the network relevance and 0.12 as the threshold of scoring of the MRMS. The number of the selected regulator genes for the target gene was 6 ($\lceil \log_2 50 \rceil$). In the experiment, we observed that FP increased dramatically as TP increased in the other five methods. The experimental results are shown in Table 5 and indicate that the MRMSn method can identify 21 correct edges and that it includes only 17 redundant edges (TP = 21, FP = 17). In addition, this method can perform well with other metrics, especially PPV and ACC (PPV = 0.553, ACC = 0.940). Our approach clearly performs better than the other tested methods.

**IRMA benchmark network.** To verify the effectiveness of the MRMSn method on real gene expression data, the method was tested on IRMA benchmark data and used to infer the IRMA benchmark network, which is a true yeast synthetic network. When the method was performed, the weighting of network relevance and the threshold of scoring in the MRMS strategy were set as 0.2 and 0.51, respectively. The number of regulator genes selected for the target gene was set to 3($\lceil \log_2 5 \rceil$). Fig 3 presents the network structure of different methods using the IRMA benchmark dataset. The figure shows that the MRMSn correctly selects five of

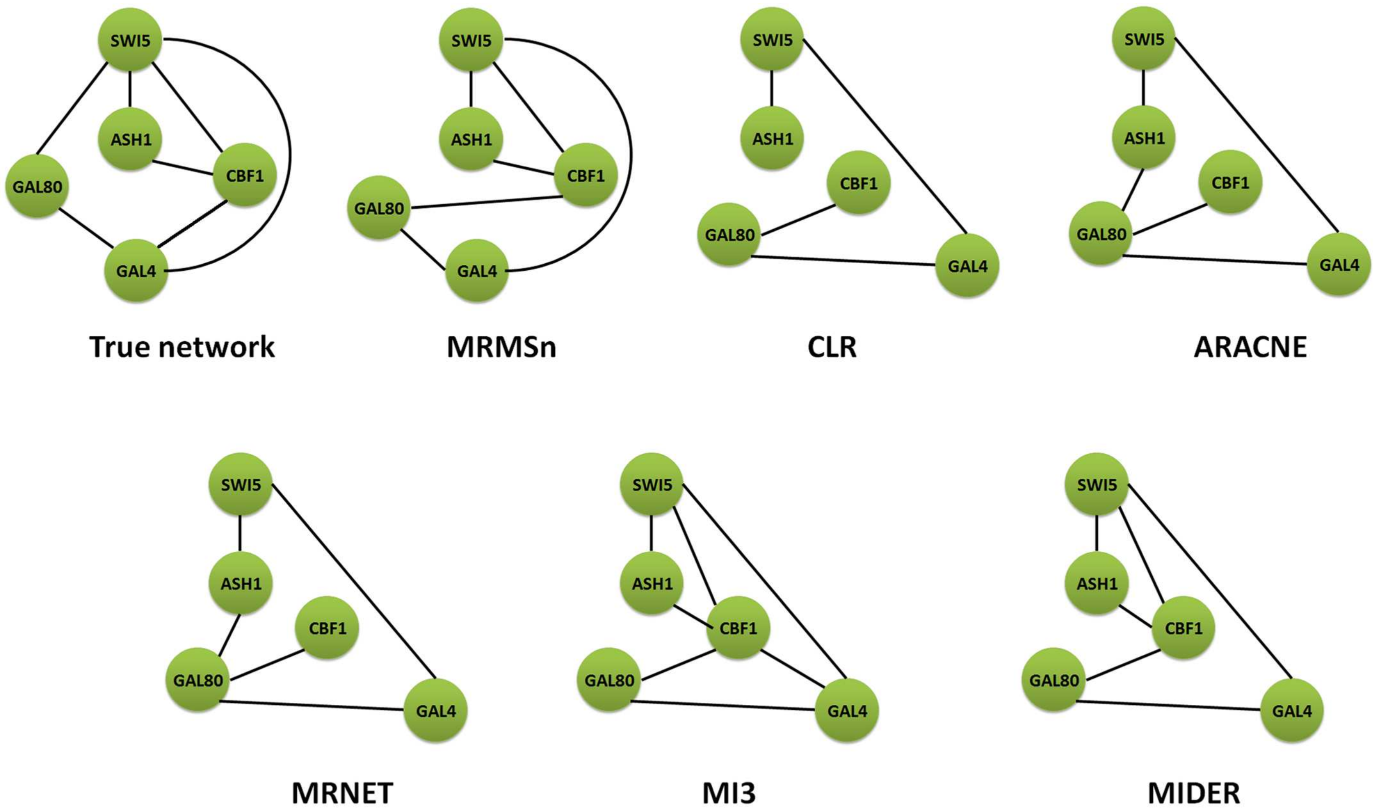

**Fig 2. Comparison of different methods on DREAM3-10genes challenge.**

doi:10.1371/journal.pone.0166115.g002

**Table 4. Comparison of the performance by different methods on DREAM3-10 genes dataset.**

| | TP | FP | TPR | FPR | PPV | ACC |
|---|---|---|---|---|---|---|
| CLR | 6 | 10 | 0.600 | 0.286 | 0.375 | 0.689 |
| ARACNE | 6 | 6 | 0.600 | 0.171 | 0.500 | 0.778 |
| MRNET | 6 | 12 | 0.600 | 0.343 | 0.333 | 0.644 |
| MI3 | 8 | 6 | 0.800 | 0.171 | 0.571 | 0.822 |
| MIDER | ------ | ------ | ------- | ------ | ------- | ------ |
| MRMSn | **9** | **1** | **0.900** | **0.029** | **0.900** | **0.956** |

doi:10.1371/journal.pone.0166115.t004

**Table 5. Comparison of the performance by different methods on DREAM3-50 genes dataset.**

| | TP | FP | TPR | FPR | PPV | ACC |
|---|---|---|---|---|---|---|
| CLR | 19 | 115 | 0.247 | 0.144 | 0.103 | 0.818 |
| ARACNE | 13 | 125 | 0.170 | 0.109 | 0.094 | 0.846 |
| MRNET | 21 | 215 | 0.273 | 0.187 | 0.089 | 0.779 |
| MI3 | 21 | 68 | 0.273 | 0.059 | 0.236 | 0.899 |
| MIDER | 4 | 79 | 0.052 | 0.069 | 0.048 | 0.876 |
| MRMSn | **21** | **17** | **0.273** | **0.015** | **0.553** | **0.940** |

doi:10.1371/journal.pone.0166115.t005

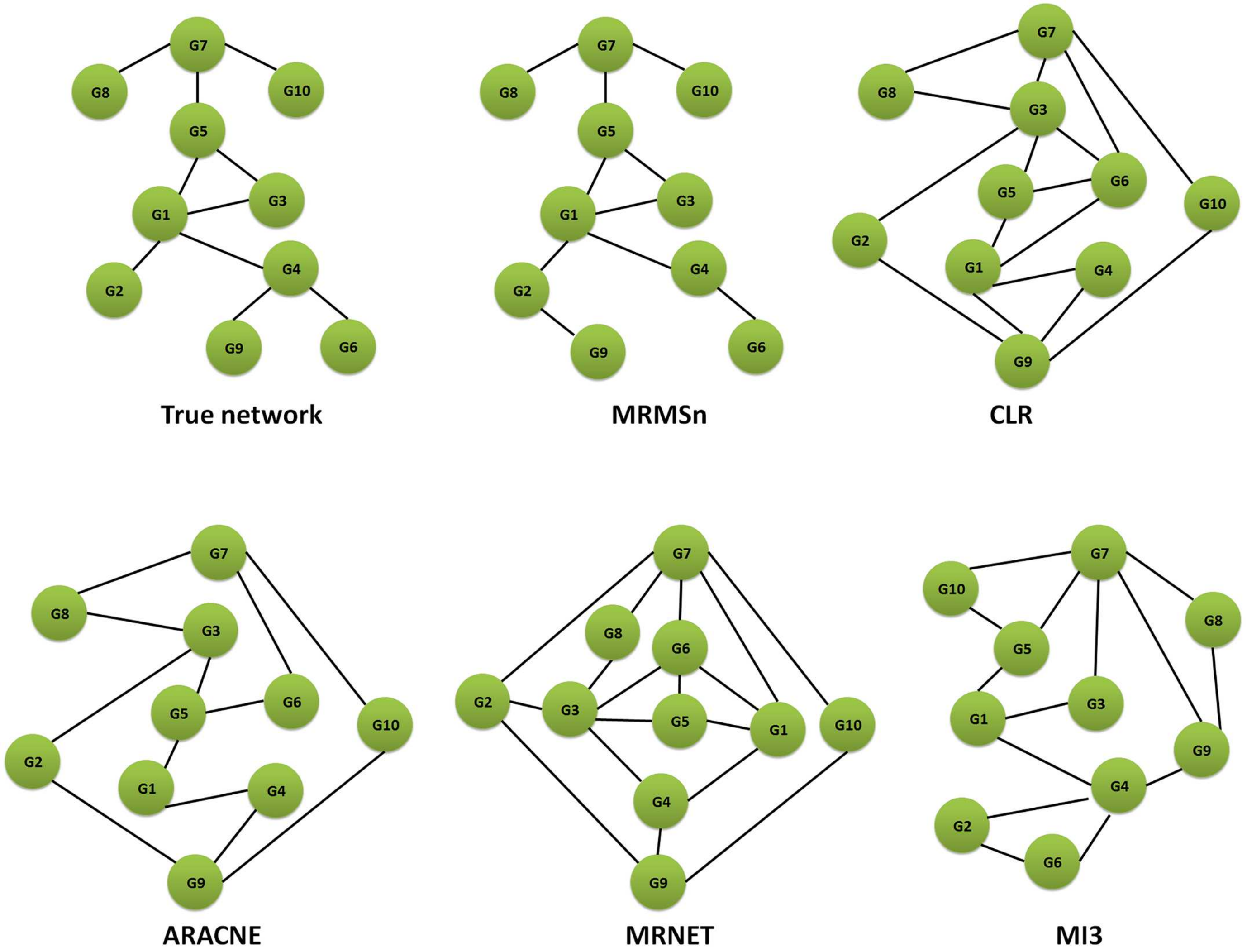

**Fig 3. Comparison of different methods on IRMA benchmark.**

doi:10.1371/journal.pone.0166115.g003

the seven true edges. However, SWI5–GAL80 and CBF1–GAL4 are lost and GAL80–CBF1 is a redundant edge. Similar comparison results with different methods can be obtained for the IRMA benchmark dataset. Table 6 shows that our method and MIDER perform better than CLR, ARACNE and MRNET, whereas MI3 (TPR = 0.857, FPR = 0.333) performs better than

**Table 6. Comparison of the performance by different methods on IRMA benchmark dataset.**

| | TP | FP | TPR | FPR | PPV | ACC |
|---|---|---|---|---|---|---|
| CLR | 3 | 1 | 0.427 | 0.333 | 0.750 | 0.500 |
| ARACNE | 3 | 2 | 0.429 | 0.667 | 0.600 | 0.400 |
| MRNET | 3 | 2 | 0.429 | 0.667 | 0.600 | 0.400 |
| MI3 | 6 | 1 | 0.857 | 0.333 | 0.857 | 0.800 |
| MIDER | 5 | 1 | 0.714 | 0.333 | 0.833 | 0.700 |
| MRMSn | 5 | 1 | 0.714 | 0.333 | 0.833 | 0.700 |

doi:10.1371/journal.pone.0166115.t006

**Table 7. Comparison of the performance by different methods on SOS dataset.**

| | TP | FP | TPR | FPR | PPV | ACC |
|---|---|---|---|---|---|---|
| CLR | 12 | 5 | 0.500 | 0.417 | 0.706 | 0.528 |
| ARACNE | 7 | 3 | 0.292 | 0.250 | 0.700 | 0.444 |
| MRNET | 17 | 6 | 0.708 | 0.500 | 0.739 | 0.639 |
| MI3 | 9 | 5 | 0.375 | 0.417 | 0.643 | 0.444 |
| MIDER | ------- | ------ | ------ | ------ | ------ | ------ |
| MRMSn | 10 | **2** | 0.417 | **0.167** | **0.833** | 0.556 |

doi:10.1371/journal.pone.0166115.t007

the MRMSn (TPR = 0.714, FPR = 0.333). Nevertheless, our MRMSn method performs well with the IRMA benchmark dataset (PPV = 0.833, ACC = 0.700).

**SOS network in *E. coli*.** The SOS network is a signal pathway in the DNA repair system. It is inferred from real gene expression data and is frequently used to test the effectiveness of network inference methods. Here, we test the MRMSn on the network of *E. coli*. We set the weighting parameter of network relevance as 0.78 and the threshold of scoring as 0.18. The number of regulator genes selected for the target gene was $4(\lceil \log_2 9 \rceil)$. Table 7 presents the results of the six methods applied to the SOS network in the *E. coli* dataset. The results show that the MRMSn performs better than all of the other methods except MRNET in terms of accuracy. Although our method did not identify the highest number of correct edges, it produced the fewest redundant edges of all of the methods. In addition, MIDER cannot be used with the SOS dataset. Although the performance of our method is poor for the TPR (TPR = 0.417), it performs better than the other methods for the FPR and PPV (FPR = 0.167, PPV = 0.833). Compared with previous experiments with other datasets, the effectiveness of the MRMSn on the SOS network is not ideal. The principal reason for this finding is related to the characteristics of the SOS network because most genes have 6–8 edges. However, the number of regulator genes selected for the target gene in the MRMSn is set to 4, which causes our method to overlook correct edges. Although the experimental results are not ideal, Table 7 shows that our method is nonetheless more effective than all of the other tested methods except MRNET.

## More performance evaluation

MRMSn was implemented to evaluate the performance based on the optimum threshold. To test the efficiency of the method on a variety of thresholds, we calculated the area under the receiver operating characteristic curve (AUROC) of the method on five datasets. We also compared the AUROC value of MRMSn with that of CLR, ARACNE, MRNET, MIDER and MI3. Except for MI3, these inference methods have some tunable parameters. Consequently, we did not show results for MI3. Table 8 presents the AUROC value results of the five methods for five datasets. From the table, we can see MRMSn, CLR, ARACNE and MIDER perform very

**Table 8. AUROC scores for five datasets using different methods.**

| Datasets | CLR | ARACNE | MRNET | MIDER | MRMSn |
|---|---|---|---|---|---|
| Reaction chain with 4 species | 1 | 1 | 0.889 | 1 | 1 |
| DREAM3 10 genes | 0.654 | 0.709 | 0.629 | ------ | 0.944 |
| DREAM3 50 genes | 0.542 | 0.531 | 0.530 | 0.509 | 0.690 |
| IRMA benchmark | 0.476 | 0.476 | 0.500 | 0.667 | 0.667 |
| SOS | 0.559 | 0.519 | 0.559 | ------ | 0.660 |

doi:10.1371/journal.pone.0166115.t008

well on Reaction chain with 4 species data, and the AUROC values reach 1. For other datasets, we can observe that our method performs better than all other methods. In particular, the AUROC value on DREAM3 10 gene data reaches 0.944, which is significantly more than other methods. All the results show the effectiveness and efficiency of the MRMSn method.

## Discussion

In this paper, we emphasize that the feature selection technique can be effectively applied to recover gene regulatory networks and obtain satisfactory accuracy. Our method converted the recovery of the gene regulatory network to the selection of the regulator genes of all of the target genes based on a feature selection strategy. The process of selecting the regulator genes is independent for each target gene. Therefore, this approach enables the use of parallel technologies to recover gene regulatory networks, and it is advantageous for inferring large-scale gene networks.

The MRMS selection criterion was proposed to measure the relevance between a target gene and regulator genes and simultaneously maintain the significance in local network structures. MIDER is also a network inference method based on the significance strategy, which refines the existing edges based on the significance strategy (entropy reduction) and infers the network structures. However, MRMSn refines the edges among genes according to a score function, which combine the relevance and significance strategy. Furthermore, there are some difference between MRMSn and MIDER: for each target genes, MRMSn stops adding genes when a predefined maximum K is met, while MIDER uses a different stopping criterion; and MSMRn uses local density to calculate the threshold for different datasets, while MIDER gives the threshold based on entropy reduction. All the experimental results show that the MRMSn can effectively infer gene regulatory networks in most cases. However, as the number of genes increases, calculating the criterion, which involves MI and multivariate conditional entropy, will lead to inaccuracies and difficulties when inferring large-scale networks.

Simulation data and real data were applied in our experiments. For the simulation data, the accuracy of the MRMSn is satisfactory and most of the correct regulation relationships can be identified. MRNET is another network inference method based on the feature selection technique. However, our method stresses the importance of significance in modularity in addition to relevance. The comparison indicates that the MRMSn is better than MRNET and the other methods for certain evaluation metrics. Significance can also eliminate redundant edges caused by the maximum relevance strategy. For the real data, the performance of the MRMSn outperformed most of the other methods, although not with perfect accuracy. These drawbacks may have been due to various causes, including the strong noise that may be contained in real data, which might lead to failures in selecting the real relationships between genes. Another cause involves the selection of parameter *K* in the MRMSn scheme. The number of regulator genes for a target gene in certain real networks may be more or less than the parameter provided in our paper, which will affect the accuracy of the regulator selection. For example, in the SOS network of *E. coli*, the maximum number of regulator genes for the target gene can reach 6, which is more than the given parameter. Thus, for the given parameter, certain regulation relationships will be lost, which was confirmed by the experimental results.

In the experiments, we found that the MRMSn works better in simple networks than in complex networks, especially for predicting certain special networks. For example, all of the regulatory relationships can be found in the reaction chain with 4 species. However, our results show that certain regulatory relationships can be lost in more complex networks, such as in the complicated SOS networks, in which almost all of the genes have connections with each other. This finding is still determined by parameter *K*. Although this type of network is uncommon and most networks still follow the sparsity of networks, the MRMSn is limited because it may not be able to estimate the optimal parameter *K*. Therefore, future work will focus on developing theoretical estimations for the number of regulators for target genes.

## Conclusions

In this paper, we developed the novel method MRMSn for inferring gene regulatory networks. The method treats the problem of network recovery as a problem of selecting regulator genes for each gene. We proposed an MRMS algorithm based on information theory. The definition of relevance and significance is the key to this algorithm, and these parameters are measured by MI and the reduced entropy of the target gene. A first-order incremental search algorithm can be used to search for regulator genes. After obtaining regulator genes for all of the target genes, a strict constraint is adopted to adjust the regulatory relationships and obtain the complete network structure. Five standard datasets were applied to test the methods, and we found that the MRMSn method can infer network structures efficiently on both simulated and real data. We compared our method with the CLR, ARACNE, MRNET, MI3 and MIDER methods and found that the performance of our method was superior.

## Supporting Information

**S1 File. The Matlab implement for the MRMSn method.** The compressed file includes the source code of MRMSn method and all the datasets in experiments.
(ZIP)

## Author Contributions

**Conceptualization:** WL BL.

**Data curation:** XTC.

**Funding acquisition:** BL.

**Investigation:** WZ.

**Methodology:** WL.

**Project administration:** BL.

**Resources:** WZ XTC.

**Software:** WZ.

**Validation:** BL XTC.

**Writing – original draft:** WL BL.

**Writing – review & editing:** WL BL.

## References

1. Bashan A, Bartsch RP, Kantelhardt JW, Havlin S, Ivanov PCh. Network physiology reveals relations between network topology and physiological function. Nat Commun. 2012; 3: 702. doi: 10.1038/ncomms1705 PMID: 22426223
2. Lee WP, Tzou WS. Computational methods for discovering gene networks from expression data. Brief Bioinform. 2009; 10: 408–423. doi: 10.1093/bib/bbp028 PMID: 19505889

3. Chen KC, Wang TY, Tseng HH, Huang CY, Kao CY. A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. BioInformatics. 2005; 21: 2883–2890. doi: 10.1093/bioinformatics/bti415 PMID: 15802287

4. Äijö T, Lähdesmäki H. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. BioInformatics. 2009; 25: 2937–2944. doi: 10.1093/bioinformatics/btp511 PMID: 19706742

5. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLOS ONE. 2010; 5: e12776. doi: 10.1371/journal.pone.0012776 PMID: 20927193

6. Pal R, Ivanov I, Datta A, Bittner ML, Dougherty ER. Generating Boolean networks with a prescribed attractor structure. BioInformatics. 2005; 21: 4021–4025. doi: 10.1093/bioinformatics/bti664 PMID: 16150807

7. Wang Y, Joshi T, Zhang XS, Xu D, Chen L. Inferring gene regulatory networks from multiple microarray datasets. BioInformatics. 2006; 22: 2413–2420. doi: 10.1093/bioinformatics/btl396 PMID: 16864593

8. Longabaugh WJ, Davidson EH, Bolouri H. Computational representation of developmental genetic regulatory networks. Dev Biol. 2005; 283: 1–16. doi: 10.1016/j.ydbio.2005.04.023 PMID: 15907831

9. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008; 9: 770–780. doi: 10.1038/nrm2503 PMID: 18797474

10. Banga JR. Optimization in computational systems biology. BMC Syst Biol. 2008; 2: 47. doi: 10.1186/1752-0509-2-47 PMID: 18507829

11. Ocone A, Millar AJ, Sanguinetti G. Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. BioInformatics. 2013; 29: 910–916. doi: 10.1093/bioinformatics/btt069 PMID: 23407360

12. Meyer P, Cokelaer T, Chandran D, Kim KH, Loh PR, Tucker G, et al. Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. BMC Syst Biol. 2014; 8: 13. doi: 10.1186/1752-0509-8-13 PMID: 24507381

13. Gardner TS, Faith JJ. Reverse-engineering transcription control networks. Phys Life Rev. 2005; 2: 65–88. doi: 10.1016/j.plrev.2005.01.001 PMID: 20416858

14. Markowetz F, Spang R. Inferring cellular networks—a review. BMC Bioinformatics. 2007; 8;Suppl 6: S5. doi: 10.1186/1471-2105-8-S6-S5 PMID: 17903286

15. Kauffman S. The large scale structure and dynamics of gene control circuits: an ensemble approach. J Theor Biol. 1974; 44: 167–190. doi: 10.1016/S0022-5193(74)80037-8 PMID: 4595774

16. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. BioInformatics. 2002; 18: 261–274. doi: 10.1093/bioinformatics/18.2.261 PMID: 11847074

17. Chen CC, Zhong S. Inferring gene regulatory networks by thermodynamic modeling. BMC Genomics. 2008; 9;Suppl 2: S19. doi: 10.1186/1471-2164-9-S2-S19 PMID: 18831784

18. Vasić B, Ravanmehr V, Krishnan AR. An information theoretic approach to constructing robust Boolean gene regulatory networks. IEEE/ACM Trans Comput Biol Bioinform. 2012; 9: 52–65. doi: 10.1109/TCBB.2011.61 PMID: 21464507

19. Iba H, Mimura A. Inference of a gene regulatory network by means of interactive evolutionary computing. Inf Sci. 2002; 145: 225–236. doi: 10.1016/S0020-0255(02)00234-7

20. Keedwell E, Narayanan A. Discovering gene networks with a neural-genetic hybrid. IEEE/ACM Trans Comput Biol Bioinform. 2005; 2: 231–242. doi: 10.1109/TCBB.2005.40 PMID: 17044186

21. Chen XW, Anantha G, Lin X. Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm. IEEE Trans Knowl Data Eng. 2008; 20: 628–640. doi: 10.1109/TKDE.2007.190732

22. Campos LMD. Independency relationships and learning algorithms for singly connected networks. J Exp Theor Artif Intell. 1998; 10: 511–549. doi: 10.1080/095281398146743

23. De Campos LM, Huete JF. A new approach for learning belief networks using independence criteria. Int J Approximate Reasoning. 2000; 24: 11–37. doi: 10.1016/S0888-613X(99)00042-0

24. Steck H. Learning the Bayesian network structure: Dirichlet prior versus data. arXiv preprint arXiv:1206.3287; 2012.

25. Acid S, de Campos LM, Fernández-Luna JM, Rodríguez S, María Rodríguez J, Luis Salcedo J. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. Artif Intell Med. 2004; 30: 215–232. doi: 10.1016/j.artmed.2003.11.002 PMID: 15081073

26. Lam W, Bacchus F. Learning Bayesian belief networks: an approach based on the MDL principle. Comput Intell. 1994; 10: 269–293. doi: 10.1111/j.1467-8640.1994.tb00166.x

27. Friedman N, Goldszmidt M. Learning Bayesian networks with local structure. In: Jordan MI, editor. Learning in graphical models. Dordrecht, Netherlands: Springer Verlag; 1998. pp. 421–459.

28. Watanabe Y, Seno S, Takenaka Y, Matsuda H. An estimation method for inference of gene regulatory net-work using Bayesian network with uniting of partial problems. BMC Genomics. 2012; 13;Suppl 1: S12. doi: 10.1186/1471-2164-13-S1-S12 PMID: 22369509

29. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 2000; 5: 418–429.

30. Villaverde AF, Ross J, Banga JR. Reverse engineering cellular networks with information theoretic methods. Cells. 2013; 2: 306–329. doi: 10.3390/cells2020306 PMID: 24709703

31. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLOS Biol. 2007; 5: e8. doi: 10.1371/journal.pbio.0050008 PMID: 17214507

32. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006; 7 Suppl 1: S7. doi: 10.1186/1471-2105-7-S1-S7 PMID: 16723010

33. Luo W, Hankenson KD, Woolf PJ. Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. BMC Bioinformatics. 2008; 9: 467. doi: 10.1186/1471-2105-9-467 PMID: 18980677

34. Zhang X, Zhao XM, He K, Lu L, Cao Y, Liu J, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. BioInformatics. 2012; 28: 98–104. doi: 10.1093/bioinformatics/btr626 PMID: 22088843

35. Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol. 2007; 2007: 79879. doi: 10.1155/2007/79879 PMID: 18354736

36. Villaverde AF, Ross J, Morán F, Banga JR. MIDER: network inference with mutual information distance and entropy reduction. PLOS ONE. 2014; 9: e96732. doi: 10.1371/journal.pone.0096732 PMID: 24806471

37. Maji P, Paul S. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. Int J Approximate Reasoning. 2011; 52: 408–426. doi: 10.1016/j.ijar.2010.09.006

38. Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, et al. Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Statistics. 2003; 7: 733–742.

39. Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA. Supervised semi-supervised and unsupervised inference of gene regulatory networks. Brief Bioinform. 2014; 15: 195–211. doi: 10.1093/bib/bbt034 PMID: 23698722

40. Samoilov MS. Reconstruction and functional analysis of general chemical reactions and reaction networks. Ph.D. Thesis, Stanford University. 1997.

41. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci U S A. 2010; 107: 6286–6291. doi: 10.1073/pnas.0913357107 PMID: 20308593

42. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell. 2009; 137: 172–181. doi: 10.1016/j.cell.2009.01.055 PMID: 19327819

43. Ronen M, Rosenberg R, Shraiman BI, Alon U. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proc Natl Acad Sci U S A. 2002; 99: 10555–10560. doi: 10.1073/pnas.152046799 PMID: 12145321