Causal gene identification using combinatorial V-structure search

Ruichu Cai a,b,∗, Zhenjie Zhang c, Zhifeng Hao a

a Faculty of Computer Science, Guangdong University of Technology, Guangzhou, PR China
b State Key Laboratory for Novel Software Technology, Nanjing University, PR China
c Advanced Digital Sciences Center, Illinois at Singapore Pte. Ltd., Singapore

ARTICLE INFO

Article history:
Received 15 May 2012
Received in revised form 23 January 2013
Accepted 31 January 2013

Keywords:
Causal gene
V-Structure
Gene expression data
Causality

ABSTRACT

With the advances of biomedical techniques in the last decade, the costs of human genomic sequencing and genomic activity monitoring are coming down rapidly. To support the huge genome-based business in the near future, researchers are eager to find killer applications based on human genome information. Causal gene identification is one of the most promising applications, which may help the potential patients to estimate the risk of certain genetic diseases and locate the target gene for further genetic therapy. Unfortunately, existing pattern recognition techniques, such as Bayesian networks, cannot be directly applied to find the accurate causal relationship between genes and diseases. This is mainly due to the insufficient number of samples and the extremely high dimensionality of the gene space. In this paper, we present the first practical solution to causal gene identification, utilizing a new combinatorial formulation over V-Structures commonly used in conventional Bayesian networks, by exploring the combinations of significant V-Structures. We prove the NP-hardness of the combinatorial search problem under a general settings on the significance measure on the V-Structures, and present a greedy algorithm to find sub-optimal results. Extensive experiments show that our proposal is both scalable and effective, particularly with interesting findings on the causal genes over real human genome data.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

With the advances of biomedical techniques in the last decade, such as microarray (Bassett, Eisen, & Boguski, 1999), the cost of gene activity monitoring is coming down to several hundreds. In the near future, it is likely that microarrays will be used to test the gene activities for every person for disease diagnosis or gene therapy, forming a business market worth billions of dollars. Before the arrival of the new genomic age, biomedical researchers are now eager to look for killer applications in the huge genomic business. Causal gene identification is one of the most promising applications (Noble, 2008), which aims to help potential patients to accurately estimate their risk with respect to certain diseases.

Unfortunately, identification of the causal genes related to genetic diseases is by no means an easy task in the biomedical domain (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009). While traditional biological and pathological methods fail to effectively and efficiently discover the causal genes, computer scientists and statisticians are trying to apply machine learning and data mining techniques to tackle the problem, e.g. Cai, Hao, Yang, and Wen (2009); Cai, Tung, Zhang, and Hao (2011), Kim, Wuchty, and Przytycka (2011) and Schadt et al. (2005). Given the gene expression data from humans with/without certain genetic diseases, algorithms are designed to automatically find out significant genes causing these diseases.

In statistics and learning communities, Bayesian networks (BN) are a common tool used to analyze the correlation and causality relationships between variables. By running statistical significance tests on variable combinations, it is possible to construct a probabilistic graphical model to simulate and evaluate the impact of certain variables over others (Cai, Zhang, & Hao, 2011). However, existing BN methods suffer from three major drawbacks on the causal gene identification problem. Firstly, complete BN construction needs an exponential number of samples to support accurate estimation in statistical tests. Secondly, most of the BN learning methods focus only on building a probabilistic model with high likelihood, instead of finding the exact causality relationship. This potentially leads to a large number of false positive causality connections between genes and diseases, even when the probabilistic model achieves a high likelihood. Therefore, although BN structure learning methods are capable of finding partial causal genes, these methods tend to output more noisy results with diminishing accuracy and low robustness, due to the low signal–noise ratio, the high
considering its interaction with the classifier. In wrapper methods, a classifier is usually built and employed as the evaluation criterion. If the feature selection criterion is derived from the intrinsic properties of a classifier, the corresponding method belongs to the embedded methods category. False discovery (Reiner, Yekutieli, & Benjamini, 2003) and feature set redundancy (Yu & Liu, 2004) are two problems we need to consider for all feature selection problems.

A causality Bayesian network is part of the theoretical background of this work. A causality Bayesian network is a special case of a Bayesian network, whose edge direction presents the causality relations among the nodes (Pearl, 2009). The Causality Bayesian network is different from the Bayesian network used in the regulatory network reconstruction problem, such as Friedman, Linial, Nachman, and Pe’er (2000) and Kim, Imoto, and Miyano (2004).

Structure learning of a Bayesian network is closely related to the algorithmic background of this work, e.g. the well-known PC algorithm (Kalisch & Bühlmann, 2007; Spirtes, Glymour, & Scheines, 2001) and Markov Blanket discovery methods (Zhu, Ong, & Dash, 2007). These methods provide the skeleton of causal structures, i.e. parent–child pairs and Markov Blanket. However, these methods usually cannot distinguish causes from consequences, which mostly relies on other techniques to conduct exact causal discovery.

Pearl is the founder of the causality analysis theory (Pearl, 2009). Most causality inference works simply assume the acquisition of a sufficiently large sample set (Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010a, 2010b), or expensive intervention experiments (He & Geng, 2008). Though there are some works aiming to solve the inference problem when a small number of samples are available (Bromberg & Margaritis, 2009), the exact sample sizes used in their empirical studies remain significantly larger than the scale of gene expression data. To the best of our knowledge, there does not exist a provable method to run robust causal inference on the real gene expression data. In this paper, we present the first practical algorithm to tackle the problems of small sample size and high dimensionality in gene data.

Another concept that is closely related to our work is Granger’s causality (Lozano, Abe, Liu, & Rosset, 2009; Mukhopadhyay & Chatterjee, 2007), which uses Granger’s causality theory to infer the gene regulatory networks from the time series gene expression data. Granger’s work differs from traditional causality inference techniques in two aspects. Firstly, compared with the conventional definition of causality, Granger’s causality is more likely a regression method and does not reflect the true causality mechanism. Secondly, the temporal information is essential for Granger’s causality inference, which is hard to collect in the disease–gene relationship analysis context.

3. Preliminaries

Assume that all samples from the problem domain contain information on m different genes, i.e. $G = \{g_1, g_2, \ldots, g_m\}$, and the disease state of the sample $y$. Let $D = \{x_1, x_2, \ldots, x_n\}$ denote the complete sample set. Each sample $x_i$ is denoted by a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$, where $x_{ij}$ indicates the expression level of the sample $x_i$ on gene $g_j$. And $y_i$ is the disease state associated with the sample $x_i$.

In particular, if $P$ is the distribution defined on all the genes’ expression level and the state of the disease, i.e. $V = G \cup \{y\}$, we assume that there exists a Bayesian network $BN$ faithful to the distribution $P$. A Bayesian network includes a directed acyclic graph which indicates conditional (in)dependent relationships among the variables, and conditional probability functions which simulate conditional probability distribution of each variable given the parent nodes. Following the common assumption of existing studies, we only consider a problem domain with the Faithfulness Condition (Koller & Friedman, 2009) as listed below.
Parents, Children, Spouses are three fundamental relations among the variable nodes in a Bayesian network. Given a target node, e.g., disease state, the parents of the target node are the variable nodes with directed edges pointing to the target node. Similarly, the children of the target node include the variable nodes with directed edges from the target node. Finally, the spouses of the target node are the variable nodes which share at least one common child with the target node. Node AML in Fig. 1, for example, has parents (FLT3,c-KIT), children (STAT,Grb2) and no spouse.

Generally speaking, there can be a large number of Bayesian network structures satisfying a single joint probability. A causal Bayesian network (CBN) is a particular Bayesian network in which each arc is interpreted as a direct causal influence between a parent variable node and a child variable node, conditioned on the other nodes in the network. Given a target node in CBN, its causal node contains its parents. Assume Fig. 1 is the structure of a CBN, FLT3 and c-KIT are thus considered as the direct causes of AML.

In gene–disease relation analysis, we are interested in identifying the genes which determine the state of the disease, i.e., the direct causal genes of a specific disease. The problem of disease-causal gene discovery is thus formally defined as follows.

Definition 2 (Causal Gene Identification). Given sample set D, the problem of disease causal gene identification is to select the smallest group of genes $G' \subseteq G$ which are direct causes of the target node $y$.

4. Combinatorial formulation and algorithm

4.1. V-structure

A Bayesian Network (BN) consists of four types of primitive local structures, as shown in Fig. 2. Fig. 2(a)–(c) are independence-equivalent BNs, because the three local structures imply the same assertion that variable $g_1$ is conditionally independent of variable $g_2$ given variable $y$. However, Fig. 2(d) is not independence-equivalent to the other three. It is called a V-structure or a Collider in BN. A V-structure generally implies a different assertion such that variable $g_1$ is independent of variable $g_2$ not given variable $y$, but variable $g_1$ is conditionally dependent of variable $g_2$ given variable $y$. V-Structure is well known and studied, due to the high importance of V-Structure in Judea’s Inductive Causation methods (Pearl, 2009).

All four local BN structures shown in Fig. 2 can be used to model causality, but independence-equivalent BNs express the same dependency information. In other words, they are equally faithful to the joint probability distribution of the given variables, which are thus indistinguishable unless using domain-expert knowledge. In contrast, given the variables $g_1$, $g_2$, and $y$, V-Structure in Fig. 2(d) can be exclusively and easily identified by testing the following conditional independence conditions, i.e. $(1) \; g_1 \perp g_2 \mid y$; and $(2) \; g_1 \not\perp g_2 \mid y$.

In a more general setting of when BN contains more than 3 nodes, the above $g_1$ and $g_2$ are likely to be conditional independent with each other, leading to the following formal definition.

Definition 3 (V-Structure). Given three variables $g_1$, $g_2$ and $y$, $g_1 \rightarrow y \leftarrow g_2$ is a V-Structure, if and only if there exists some variable set $Z \subseteq G$– {$g_1$, $g_2$} such that $(1) \; g_1 \perp g_2 \mid Z$; and $(2) \; g_1 \not\perp g_2 \mid \{y, Z\}$.

The definition above illustrates how to generally find V-Structures. We can easily identify all significant V-Structures, following a two-phase strategy. In the first phase, all parent-candidates (PCs) with respect to the target node $y$ are discovered, using the popular growth algorithm, e.g. Spirtes et al. (2001). In the second phase, V-Structures are verified by checking every pair of the PCs ($g_i$, $g_j$) based on Definition 3.

4.2. Conflicts between V-structures

Despite of the stability of individual V-Structures, conflicts could occur between V-Structures, even when evaluating on datasets with a large number of samples. Specifically, a conflict between two V-Structures are observable, if they respectively contain two directed edges between the same pair of variables but with reversed directions. In Fig. 3, we present an example of the conflict between two V-Structures. To get a better understanding to how noisy data raise the conflicts between V-Structure, we hereby describe a simple constructive method which generates conflicted V-Structures following the structure in Fig. 3.

Assume there are $n$ (without loss of generality, $n$ is divisible by 4) gene sequence samples on 4 genes, i.e. {$g_1$, $g_2$, $g_3$, $g_4$}. Our approach first generates the expression levels of the samples on $g_2$ and $g_3$, such that ($g_2 = 0$, $g_3 = 0$) appears $n/4$ times in the samples. Similarly, ($g_2 = 0$, $g_3 = 1$), ($g_2 = 1$, $g_3 = 0$) and ($g_2 = 1$, $g_3 = 1$) also appear $n/4$ times. To generate the expression levels for the samples on $g_1$, our approach lets $g_1 = g_3$ if $g_2 = 0$, otherwise $g_1 = 1 - g_3$. In a similar manner, we set $g_4 = g_3$ if $g_2 = 0$, otherwise $g_4 = 1 - g_2$. In Table 1, we list a concrete example with $n = 8$ samples. It is straightforward to verify that the V-Structures calculated based on the table follow the exact structure in Fig. 3.

We can further extend the capability of the constructive method above, by controlling the degree of conditional independence. This is achieved by replacing the entries on $g_i$ (resp. $g_i$) with new expression levels independent of $g_j$ conditioned on $g_k$. Therefore, we can always simulate V-Structures with arbitrary strength, by running our constructive approach. This is summarized in the following lemma, which can be proved in a straightforward way.

![Figure 2](image1.png)

![Figure 3](image2.png)

Table 1

<table>
<thead>
<tr>
<th>Sample</th>
<th>$g_1$</th>
<th>$g_2$</th>
<th>$g_3$</th>
<th>$g_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x_2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x_3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>x_4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>x_5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>x_6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>x_7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>x_8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Lemma 1. Given a group of V-Structures with arbitrary conditional independence levels, there always exists at least one gene expression sample set accurately rendering all these V-Structures, when the number of samples is large enough.

The analysis above implies that conflicts of V-Structures can happen in arbitrary form. This leads to difficulties in our causal gene selection problem, since we are unable to identify the correct direction of causality when conflicts are observed between two genes. To resolve the issue of conflict, we propose a combinatorial formulation to select V-Structure with maximal significance and zero conflict.

4.3. Combining V-Structures

Both latent variables and noises on data could be the underlying reason behind conflicts between V-Structures. In the gene expression data, a large number of V-Structures with wrong causation relationship may incur, because of the small sample cardinality, high dimensionality and low ratio of signal to noise probably render insignificant results in almost all statistical tests. The focus of this paper is to tackle the problem of conflicted V-Structures for more robust and scalable causation analysis. Fortunately, we observe that the falsely discovered V-Structure can be detected, since they are usually conflicted with the true ones. Moreover, when two conflicted V-Structures are discovered, the one with lower significance measure is usually falsely discovered. Based on the V-Structures, we believe the true causation can be refined by removing the falsely discovered V-Structure. Formally, we transform the causal gene selection problem into another optimization problem, targeting to identify a group of most significant V-Structures with maximal coverage and zero conflict.

Given a set of related V-Structures \( \mathbb{S} = \{v_1, v_2, \ldots, v_k\} \), we assume that their corresponding significance weights \( \{s_1, s_2, \ldots, s_k\} \) are also available, such that each \( s_i \in [0, 1] \) indicates the significance of V-Structure \( v_i \). In particular, we are interested in significance weighting schemes satisfying the following definition.

Definition 4 (Consistency Condition). The significance weighting scheme satisfies consistency condition, if given any significance weights \( s = \{s_1, \ldots, s_k\} \), there always exists a group of feasible V-Structures with exactly the weight \( s \) by the weighting scheme.

Specifically, the condition above implies that the weighting scheme is a specific mapping from the conditional independence level to the complete domain \([0.5, 1]\). To identify robust causal genes, given the V-Structure, we aim to find a subset \( R \subseteq \mathbb{S} \) without any conflict pair of V-Structures and maximize the following objective function:

\[
F(R) = \prod_{v_i \in R} s_i \cdot \prod_{v_j \in \mathbb{S} - R} (1 - s_j).
\]  

Intuitively, optimization over Eq. (1) obtains following three good properties: (1) if \( v_i \) does not incur conflict with any other V-Structures, \( v_i \) is contained in the final result, because the significance level \( s_i \) falls in domain \([0.5, 1]\), the contribution of picking up this V-Structure for the result set, \( s_i \), is larger than that of not selecting the V-Structure, \((1 - s_i)\); (2) If two V-Structures are in conflict with each other, the more significant V-Structure is selected, because V-Structures with a larger significance measure are more statistically reliable; (3) when the conflict happens among more than two V-Structures, the objective tries to maximize the significance of the selected V-Structures based on the assumption that each V-Structure is independent of the others.

Optimization over Eq. (1) turns out a binary programming problem and naturally NP-hard. In the following theorem, we prove the above statement rigorously by constructing a polynomial reduction from the problem of Maximal Independent Set.

Theorem 1. It is NP-hard to find the combination of V-Structures maximizing Eq. (1).

Proof. Given a graph \( G(V, E) \), the problem of Maximal Independent Set is to select the maximal subset \( V' \) of vertices in \( V \), such that there is no edge \( (v_1, v_2) \in E \) for each pair of \( v_1, v_2 \in V' \).

To reduce the problem to our combinatorial formulation, for each edge in \( E \), we generate a pair of genes such that conflicted casual relationships with reversed edges are constructed. After the generation of the conflicted edges, we build another group of genes for each vertex in \( V \) and connect them to the conflict pairs correspondingly.

Given the V-Structures, the next step is generating the significance weights \( \{s_1, s_2, \ldots, s_k\} \) for each V-Structure. We hereby simply use the same weight \( s > 0.5 \) for each V-Structure. Using the consistency condition on the significance weights, there always exists a reversed mapping from the weights to the conditional independence level of the V-Structures. Finally, using Lemma 1, we are able to generate a group of samples to simulate the conditional independence levels of the V-Structure among the genes. Fig. 4 presents an example of the reduction, which transforms the original graph with 4 nodes to a gene network with 8 genes.

It is straightforward to show that the result of the optimization formulation with Eq. (1) is always equivalent to the optimal answer to the maximal independent set problem, which completes the proof of the theorem.

In this work, we propose a heuristic solution for this problem, by repeatedly removing the least significant V-Structure from \( \mathbb{S} \) until there is no conflict pair. Moreover, the discovery of the related V-Structure in the high dimensional data set is also computational expensive. Our method thus only considers the V-Structures whose collider is the parent candidate (PC) of the target node. The details of the algorithm are given in Algorithm 1.

It is obvious that the significance measure, i.e. \( s \) for each V-Structure \( v_i \), plays an important role in the performance of our causal gene identification algorithm. In the following, we...
discuss two different measures, ST and SI, based on the analysis on conditional independence and information theory respectively.

**Independence test approach:** From the perspective of probabilistic graphical model, the significance of V-Structure depends on the significance of the following conditional independence statements: (1) firstly, whether \( g_1 \) and \( g_2 \) are independent of each other given the conditional set \( Z \), i.e. \( g_1 \perp \!\!\!\perp g_2 | (Z) \); (2) secondly, whether \( g_1 \) and \( g_2 \) are dependent of each other when the conditional set is \( Z \cup y \), i.e. \( g_1 \not\perp \!\!\!\perp g_2 | (y, Z) \) and \( g_1 \perp g_2 | Z \). To simplify the analysis, we assume these two events are independent of each other, the significance of a V-Structure is thus defined as:

\[
ST(g_1, g_2, y) = p(g_1 \perp \!\!\!\perp g_2 | (y, Z)) \times p(g_1 \perp g_2 | (Z))
\]

(2)

in which \( p(g_1 \perp \!\!\!\perp g_2 | X) \) is the \( p \)-value of the conditional independence test \( g_1 \perp \!\!\!\perp g_2 | X \).

In the following, we discuss the applicability of the **consistency condition** on the measure based on the independence test. Assume that the confidence level of the (conditional) independence test is 0.95. Given a specific significance value \( ST(g_1, g_2, y) \in (0.95^*1, 1) \), we can construct a set of samples with V-Structure whose significance measure is \( ST(g_1, g_2, y) \) by applying the following steps. Firstly, we generate a set of sample on \( g_1, g_2, Z \), ensuring that \( p(g_1 \perp \!\!\!\perp g_2 | (Z)) = \sqrt{ST(g_1, g_2, y)} \). Secondly, we create the variables on value \( y \) for each sample, based on the corresponding values of \( g_1, g_2, Z \), which guarantees that \( p(g_1 \perp \!\!\!\perp g_2 | (y, Z)) = \sqrt{ST(g_1, g_2, y)} \). The feasibility of the above construction is due to the fact that the P-Value of the (conditional) independence test can achieve any value in the interval (0, 1) when one variable is free. This implies that the measure satisfies the **consistency condition**.

**Information theory approach:** From information theory aspect, a V-Structure stands for an information flow in which variances in \( Z \) block the flow from \( g_1 \) to \( g_2 \). And the collider \( y \) reconnects the information flow between \( g_1 \) and \( g_2 \). Based on this observation, the significance of a V-Structure \( g_1 \rightarrow y \leftarrow g_2 \) can be measured based on two factors, including: (1) the degree of \( Z \) blocking the information flow, and (2) the degree of \( y \) reconnecting the flow given the condition \( Z \). In addition, the change of the information flow is also highly related to the information flow between \( g_1 \) and \( g_2 \). In the extreme case, if the entropy with \( g_1 \) or \( g_2 \) is zero, the change of the information flow is also zero. We can thus summarize that the significance of the V-Structure can be modeled as the proportion of the changed information flow, as listed below.

\[
SI(g_1, g_2, y) = \frac{MI(g_1, g_2 | (Z, y)) - MI(g_1, g_2 | (Z))}{\min(H(g_1), H(g_2))}
\]

(3)

In the equation above, \( MI(g_1, g_2 | (Z)) \) and \( MI(g_1, g_2 | (Z, y)) \) are the conditional mutual information between \( g_1 \) and \( g_2 \) conditioned on \( Z \) and \( (Z, y) \) respectively. \( H(g_1) \) and \( H(g_2) \) are the entropies of \( g_1 \) and \( g_2 \).

Similar to the independence test approach, we also verify the consistency condition with the **Information theory approach**. In particular, given any value of significance value \( SI(g_1, g_2, y) \), we can construct a set of samples to get V-Structures with significance value \( SI(g_1, g_2, y) \). Firstly, we simply generate the values of the samples on \( g_1 \) and \( g_2 \) satisfying \( H(g_1) = H(g_2) = 1 \). Secondly, our method generates the value of \( Z \) to ensure \( MI(g_1, g_2 | (Z)) = 0 \). Finally, the values of \( y \) are built, according to the existing value of \( g_1, g_2, Z \) such that \( MI(g_1, g_2 | (Z, y)) = SI(g_1, g_2, y) \). The feasibility of the above construction is based on the fact that the entropy/mutual information function is likely to achieve any value given one free variable. Thus, using the combinatorial problem using the SI significance measure is NP-hard, due to the satisfaction of consistency condition.

**Algorithm 2: Gene Identification**

4.4. **Gene identification**

Given the V-Structures without conflict, the final step of our method is to identify the causal genes related to the target node \( y \). This step is not that straightforward, due to the existence of three types of nodes in the V-Structures, as discussed below.

**Explicit causal nodes:** The explicit causal nodes are the nodes with strong causal association with the target node, which can be easily obtained from the V-Structures. In Fig. 5(a), for example, we have \( g_1 \rightarrow y \leftarrow g_2 \). This tells us that \( g_1 \) and \( g_2 \) are the explicit causal node of \( y \).

**Explicit non-causal nodes:** In Fig. 5(b), \( g_1 \) is an explicit non-causal node of \( y \). Therefore, \( g_1 \) can be removed from the causal candidate node set, since the V-Structure \( y \rightarrow g_1 \leftarrow g_2 \) exists.

**Implicit causal node:** The implicit causal node can be discovered with the help of the V-Structures of target node \( y \)'s PC set. It is based on the MDL principle in the causal inference. As is shown in Fig. 5(c), \( g_2 \rightarrow g_1 \rightarrow y \) exists but neither \( g_2 \rightarrow g_1 \rightarrow y \) nor \( g_1 \rightarrow g_2 \rightarrow y \) exists. This information helps us to infer that there is a direct edge \( g_1 \rightarrow y \). Otherwise, there must be a directed edge \( y \rightarrow g_1 \), which will form a new V-Structure together with the edge \( g_2 \rightarrow g_1 \) and \( g_3 \rightarrow g_1 \).

In Algorithm 2, we list the details of the gene identification procedure, which returns both explicit causal nodes and implicit causal nodes by exploring \( R \), the conflict free V-Structure set which is generated by the function CombineVStructure. Thus we can summarize the full map of our algorithm in Fig. 6.

In the following, we provide some analysis on the completeness of causal gene identification using our algorithm. In particular, we give a sufficient condition on the Bayesian network over the genes, which guarantees the identification of causal genes.

**Theorem 2.** When the noise on the samples are sufficiently small, Algorithm 2 always returns the exact causal genes, if (1) \( y \) has more than one parent in the Bayesian network, or (2) \( y \) has only one parent \( g_1 \), and \( g_2 \) has more than one parent.

**Proof.** We prove the theorem based on the two cases separately. In the first case, when \( y \) has more than two parents in the Bayesian network, for each causal gene \( g_i \), there is at least one another gene \( g_j \), such that \( g_i \rightarrow y \leftarrow g_j \) is a valid V-Structure. Therefore, our algorithm is capable of identifying \( g_i \) by investigating all explicit causal nodes. In the second case, when \( y \) has only one parent \( g_1 \) and \( g_2 \) has at least two parents \( g_1 \) and \( g_2 \), there is a valid V-Structure, i.e. \( g_1 \rightarrow g_2 \leftarrow g_1 \), involving \( g_1, g_2 \), and \( g_2 \). Since \( g_1 \) is the only causal factor for \( y \), it is impossible to find either V-Structures such that \( g_1 \rightarrow g_2 \leftarrow y \) or \( g_2 \rightarrow g_1 \leftarrow y \). This shows that \( g_1 \) must be found as an implicit causal nodes in our algorithm.
Section 4: Methods

5. Experiments

In this section, we evaluate our proposal, called SVS in this section, on both low-dimensional simulated data and high-dimensional real gene expression data. In all the implementations, the same $G^*$ conditional independence test (Spirtes et al., 2001) is employed, with conditional independence threshold at 95%.

5.1. Results on simulated data

The main purpose of the experiments on simulated datasets is to evaluate the accuracy and scalability of our proposal, against existing causal inference techniques applicable only on moderate dimensionality. Specifically, we simulate a network with 10 genes with binary values only, following the causal structure presented in Fig. 7. The same causal structure has been used in Mukhopadhyay and Chatterjee (2007) to test Granger’s causality discovery method. Although the causal network consists of only 10 variables, it contains all types of causal structures discussed in the previous section. For example, $g_2 \rightarrow g_3 \leftarrow g_{10}$ is an explicit Causal Node, $g_4$ can be removed from the causal candidate set with respect to $g_3$ due to Explicit Non-Causal Node, and $g_1 \rightarrow g_2$ is found to be an Implicit Causal Node.

The independent nodes in the network, namely $g_1, g_7, g_8, g_9$, are randomly generated following Bernoulli distribution outputting 0 and 1 with probability 0.5. The states of the other nodes follow certain random conditional probability tables, which assigns probabilities to the variables conditioned on every combination of causal variables. Details of the data generation process is given in Appendix. In the experiments, the results are averaged over 100 runs, i.e. using 100 different sample tables built on different conditional probability tables.

Since the causal structure is known, we are allowed to evaluate the performance using standard recall and precision on the result genes given by the algorithms. Specifically, precision is the fraction of result causal genes which are the true causes of the target node, i.e. $\text{Precision} = \frac{|\{\text{DiscoveredCausalNode} \cap \{\text{CausalNode}\}|}{|\{\text{DiscoveredCausalNode}\}|}$. Similarly, recall is the fraction of true causal genes found by the algorithm, i.e. $\text{Recall} = \frac{|\{\text{DiscoveredCausalNode} \cap \{\text{CausalNode}\}|}{|\{\text{CausalNode}\}|}$.

Comparison on significance measures: In Fig. 8, we report the correct percent, recall and precision of two versions of our algorithm with two different significance measures. Here, correct percent refers to the number of conflict V-Structures pairs whose correct V-Structure’s significance is higher over the number of all conflicted V-Structure pairs. As shown in the Figure, the correct percent of both the two significance measure is higher than 0.5, which proves the usefulness of both significance measures.

In terms of the significance measures, the information theory based approach, $SI$, outperforms the independence test approach $ST$ on all correct percent, recall and precision, when the sample size is small. $SI$ can catch important information even when very few samples are available. When sample size grows, although $ST$ shows slightly better performance on its recall, $SI$ achieves an excellent balance between recall and precision. We conclude that $SI$ is better at identifying significant V-Structures.

By increasing the number of samples used in the testing, SVS achieves the most significant improvements when the sample size grows from 40 to 80. It implies that 80 samples are enough for SVS to capture the important causal relations. Moreover, the samples needed to heavily depend on the local connectivity of the structure. We believe that when SVS is used in problems with similar connectivity structures as the simulated data set, 80 samples are enough to catch the main causal relations.

As results show superiority of $SI$ over $ST$, we will only test our algorithm with $SI$ in the rest of the section.

Comparison on causation search methods: To show the effectiveness of the object function (Formula 1) and our heuristic algorithm, we report the recall and precision of three different methods to find causations based on the conflicted V-Structures in Fig. 9. In terms of search method, SVS denotes the proposed heuristic method; Optimal runs exhaustive search of the object function presented in Formula 1, the same objective function of SVS applies exhaustive search on all possible combinations of causal nodes, while $SI$ runs a simple score test that ignores the causal structure. As results show, our heuristic method with the support of information theoretical heuristics is much more efficient than exhaustive search, while maintaining high recall and precision. In addition, SVS outperforms the independence test approach $ST$ on all correct percent, recall and precision. We believe that when SVS is used in problems with similar connectivity structures as the simulated data set, 80 samples are enough to catch the main causal relations.
search strategy using the same objective function as SVS does; *Rand* randomly removes one of the conflict V-Structures without considering the significance. As shown in the figure, SVS performs much better than *Rand*, and obtains acceptable recall and precision when compared with *Optimal*. It is interesting to observe that SVS achieves the same recall and precision as *Optimal* does when the sample size is larger than 320. This is because, when sample size is larger than 320, the conflicts only occur between two V-Structures, thus SVS can get the optimal solution of the objective function. Moreover, the large improvement of SVS and *Optimal* compared with *Rand* also provides an empirical justification on the objective function employed in our combinatorial formulation.

**Comparison with MMHC**: In the following, our proposed method SVS is compared with MMHC (Tsamardinos, Brown, & Aliferis, 2006), the mainstream causal inference method. Causal Explorer’s implementation of MMHC is used in our experiments. When compared against MMHC, our SVS algorithm with either significance measure is better than MMHC on recall and precision. Since MMHC constructs the complete DAG, the parent node of the target node are all returned as causal genes. Fig. 10 shows that SVS outperforms MMHC by a huge margin, especially when the sample size is not large. It shows that our SVS algorithm is robust even when the samples are insufficient.

**False causality result analysis**: In the following, we provide a detailed case-by-case analysis on the wrong results reported by SVS. In particular, there are three different types of wrong results, including *False Negative*, *False Positive* and *Wrong direction*. Assuming $g_2$ in Fig. 7 is the target node, if $g_1$ is not discovered, the causal relation $g_1 \rightarrow g_2$ is a *False Negative*; if $g_1$ is discovered as the causal node of $g_2$, then $g_2 \rightarrow g_1$ is a *Wrong direction*; if $g_1$ is discovered as the causal node of $g_2$, then the edge $g_1 \rightarrow g_2$ is a *False Positive*. All wrong results with appearance frequency larger than 10 in our experiments are listed in Table 2.

Table 2 shows that SVS reports *Wrong direction* results much less frequently than MMHC, while the frequencies on *False Positive* and *False Negative* results are almost the same. We thus conclude that SVS possesses a strong ability on correctly inferring the directions of the edges, i.e. the causal relations among the variables. Moreover, the direction of $g_1 \rightarrow g_2$ and $g_4 \rightarrow g_5$ is the most frequent wrong result for the SVS method, mainly because the implicit inference step is not sufficiently reliable.

**5.2. Results on real data**

To test the effectiveness of our proposal in real causal gene studies, we test our method on a real dataset and verify the result on the biological pathway database. In particular, we run our analysis on the *prostate cancer* dataset (Singh et al., 2002) and the
Leukemia data set. The prostate dataset contains 52 prostate cancer patients samples and 50 ordinary samples, each of which contains the expression level of 12600 genes. Similarly, The Leukemia dataset contains 52 prostate cancer patients samples and 50 ordinary samples, with expression data on 12 600 genes segments. The Leukemia dataset consists of 47 samples belonging to ALL subtype of Leukemia and 25 samples of AML subtype under Leukemia, each of which contains the expression level of 7129 gene segments. Note that the aim of the studies on the prostate dataset is to find the causal genes deciding target node, i.e. prostate cancer status, with binary values on “normal” or “cancer”; while the goal of Leukemia study is to identify the causes responsible to the state on “ALL” or “AML”. The general information of the datasets is summarized in Table 3.

We emphasize here again that no existing algorithm, such as MMHC, is capable of running causality analysis on such high-dimensional dataset.

Discretization is an important preprocessing step in the causal gene discovery. Entropy discretization procedure (Dougherty, Kohavi, & Sahami, 1995) is well accepted in gene expression data analysis. However, entropy discretization may cause over-fitting problems in our high-dimensional setting. Instead, we employ two-state Gaussian Mixture Model to discretize the data (Xing, Jordan, & Karp, 2001), this strategy exploits the expressed/suppressed two stage assumption of the gene.

Prostate cancer: We found that all the result causal genes in Table 4 have been reported in previous prostate cancer researches. EGFR, i.e. epidermal growth factor receptor, has been reported highly correlated to prostate cancer (Mimeault, Pommery, & Hénichart, 2003). EGFR’s role in prostate’s gene therapy also shows that it is a direct causal gene of prostate cancer. RPS27, i.e. ribosomal protein S27, also plays an important role in genetic information processing.3 It is believed that the misregulation of its genetic processing is generally the cause of cancer. CHN2 encodes a protein of the chimerin family, which plays a role in the proliferation and migration of smooth muscle cells.4 MLF1, the myeloid leukemia factor 1, was firstly discovered as a factor of Leukemia; recent research has found that MLF1 is also involved in other cancers, because of its role in transcriptional misregulation (Badano, Teslovich, & Katjasnis, 2005).

In order to provide further biological interpretation of the discovered causal genes, we try to locate the results in the known pathway of prostate cancer. We verify our results with KEGG pathways,5 which provides an easy-to-use visualization interface. One of the discovered causal gene, EGFR, is found in the Prostate pathway. Fig. 11 shows the location of EGFR on the skeleton of prostate pathway. EGFR appears in the very origin of the prostate pathway, which controls almost all the functions related to prostate cancer, such as Cell cycle procession, Cell survival, Cell proliferation and Tumor growth. Thus, we can believe that EGFR is the causal gene of prostate and verifies our results.

Leukemia: In this group of studies, we try to find genes distinguishing between two subclasses of Leukemia, i.e. Acute myeloid leukemia (AML) and Acute lymphoblastic leukemia (ALL). The causal genes making the difference of AML and ALL are listed in Table 5. Among all result genes, MLF2 is the myeloid leukemia factor, which has been verified in existing research as one of the causes deciding if AML or ALL happens (http://www.wikigenes.org/e/gene/e/8079.html, 0000). MEF2A is one of MEF2 family. The other version of MEF2, MEF2C has also been found as an important factor with respect to AML, as Schler calls ‘A causal role of up-regulated MEF2C expression in myelomonocytic acute myeloid leukemia (AML) has recently been demonstrated’ (Schler et al., 2008). CD22 plays an important role in B cell receptor signaling pathway, which is important pathway in the lymphoblastic system (Hasler & Zouali, 2001). Thus it is reasonable that CD22 is one of the causes to the difference between AML and ALL. Histone H2AX is reported to be responsible to render different classes of Leukemia, because of its impact on the promotion of B-cell tumorigenesis (Walsh & Rosenquist, 2005).

6. Conclusions and discussion

In this work, we presented the first practical solution to causal gene identification problem, using a new formulation of combinatorial V-Structure search. To fully utilize the framework, we discuss how to choose significance measures and how to pick up causal genes based on the V-Structure combination. Our proposal facilitates us to run analysis on extremely high-dimensional gene expression data. Our results on synthetic data and real data show that our proposal is much more effective than any existing solution in the literature to tackle the problems of high dimensionality and small sample size at the same time.

While our methods turn out to be effective, there remains room for more improvements. Our current solution to the combinatorial issues will be explored in the future.

---

formulation only applies a simple greedy selection heuristic. It is interesting to attempt other combinatorial algorithms with performance guarantees. We are also keen to test on more human genomic data for new findings of causal genes.

Acknowledgments

This work is financially supported by the Natural Science Foundation of China (61070033, 61100148), the Natural Science Foundation of Guangdong province (S2011040004804), Key Technology Research and Development Programs of Guangdong Province (2010B050400011), Opening Project of the State Key Laboratory for Novel Software Technology (KFKT2011B19), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (LYM11060), Science and Technology Plan Project of Guangzhou City (12C42111607), Science and Technology Plan Project of Panyu District Guangzhou (2012-Z-03-67).

Appendix. Pseudocodes of synthetic data generation

Given the parents set of a variable \( g \), the Conditional Probability Table is presented in Algorithm 3.

\[
\text{Algorithm 3: Conditional Probability Table Generator.}
\]

With the states of parents variables \( X_{\text{parents}} \), the conditional probability table \( CPT \) and the noise ratio \( p_{\text{noise}} \), Algorithm 4 gives the generation of \( x \). In this work, \( p_{\text{noise}} = 0.05 \), which means 5% samples are noise.

\[
\text{DataGenerator} (CPT, X_{\text{parents}}, x) \]

\[
\text{Algorithm 4: Random Data Generator}
\]

References

http://www.wikigenes.org/e/gene/e30879.html