BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 6 2013, pages 765–771
doi:10.1093/bioinformatics/btt048

Data and text mining

Advance Access publication January 30, 2013

Inferring microRNA–mRNA causal regulatory relationships from expression data

Thuc Duy Le[1,*], Lin Liu[1], Anna Tsykin[2], Gregory J. Goodall[2,3,4], Bing Liu[5], Bing-Yu Sun[6] and Jiuyong Li[1,*]

[1]School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, South Australia 5095, [2]Centre for Cancer Biology, SA Pathology, Adelaide, South Australia 5000, [3]School of Molecular and Biomedical Science and [4]Department of Medicine, University of Adelaide, Adelaide, South Australia 5005, [5]Children's Cancer Institute Australia, Randwick, New South Wales 2301, Australia and [6]Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: microRNAs (miRNAs) are known to play an essential role in the post-transcriptional gene regulation in plants and animals. Currently, several computational approaches have been developed with a shared aim to elucidate miRNA–mRNA regulatory relationships. Although these existing computational methods discover the statistical relationships, such as correlations and associations between miRNAs and mRNAs at data level, such statistical relationships are not necessarily the real causal regulatory relationships that would ultimately provide useful insights into the causes of gene regulations. The standard method for determining causal relationships is randomized controlled perturbation experiments. In practice, however, such experiments are expensive and time consuming. Our motivation for this study is to discover the miRNA–mRNA causal regulatory relationships from observational data.

Results: We present a causality discovery-based method to uncover the causal regulatory relationship between miRNAs and mRNAs, using expression profiles of miRNAs and mRNAs without taking into consideration the previous target information. We apply this method to the epithelial-to-mesenchymal transition (EMT) datasets and validate the computational discoveries by a controlled biological experiment for the miR-200 family. A significant portion of the regulatory relationships discovered in data is consistent with those identified by experiments. In addition, the top genes that are causally regulated by miRNAs are highly relevant to the biological conditions of the datasets. The results indicate that the causal discovery method effectively discovers miRNA regulatory relationships in data. Although computational predictions may not completely replace intervention experiments, the accurate and reliable discoveries in data are cost effective for the design of miRNA experiments and the understanding of miRNA–mRNA regulatory relationships.

Availability: The R scripts are in the Supplementary material.

Contact: thuc_duy.le@mymail.unisa.edu.au or jiuyong.li@unisa.edu.au

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

Received on July 17, 2012; revised on January 24, 2013; accepted on January 25, 2013

1 INTRODUCTION

microRNAs (miRNAs) are short (~22 nt) endogenous non-coding RNAs that regulate gene expression by promoting mRNA degradation and repressing translation. They recognize target mRNAs by base pairing to complementary sequences in the 3′-untranslated region (3′-UTR) of the target mRNA (Ambros, 2004; Bartel, 2004; Berezikov *et al*., 2006; Meister and Tuschl, 2004). miRNAs have also been observed to target genes through sites in the 5′-UTR and sometimes in the open reading frames (Bartel, 2009). It has been demonstrated in a body of literature that miRNAs regulate a wide range of biological processes in proliferation (Chen *et al*., 2006; Zhao *et al*., 2005), metabolism (Poy *et al*., 2004), differentiation (Esquela-Kerscher and Slack, 2006), development (Jin *et al*., 2004), apoptosis (Xu *et al*., 2007), cellular signalling (Cui *et al*., 2006) and even cancer development and progression (Bartel, 2004; Kim and Nam, 2006).

Since the advent of microarray data on gene expression programs, various statistical data mining methods have been devised in recent years in an attempt to discover the miRNA–mRNA regulatory relationships. This research into the regulatory relationships between miRNAs and mRNAs can be recognized in two main streams. In the first stream, researchers developed various methods to identify a group of co-expressed miRNAs and mRNAs in data. For instance, Huang *et al*. (2007) applied Bayesian network parameter learning, Joung *et al*. (2007) proposed a population-based probabilistic learning model and Tran *et al*. (2008) used a rule-based method. In the second stream, attempts were made to predict the regulatory networks of miRNAs and mRNAs for specific biological processes. Some highlights in this direction include Joung and Fei (2009) applying a probabilistic graphic model adopted from the author-topic model in information retrieval, Liu *et al*. (2009) using Bayesian network learning and Liu *et al*. (2010) proposing a graphical model inspired by the correspondence latent Dirichlet allocation.

These methods, although varied, all identify only the statistical relationships in the data. These relationships are either correlations or associations between the two types of variables, miRNA and mRNA. However, correlations or associations are not causality. For instance, the expression values of an miRNA

*To whom correspondence should be addressed.

and an mRNA may be strongly correlated across samples, but it is not sufficient to conclude that the miRNA regulates the mRNA. The strong correlation between the miRNA and the mRNA may be a result of the mRNA regulating the miRNA, or a third molecule regulating both the miRNA and the mRNA.

In this article, our ultimate goal is to discover which miRNAs causally regulate which mRNAs. A regulatory relationship between an miRNA and an mRNA means that a change in the expression level of the miRNA will result in a change in the expression level of the mRNA. Associations or correlations are not the right tools to test the causal hypothesis. Therefore, we aim to discover the causal effects of an miRNA on an mRNAs. We refer to this causal effect as *miRNA causal regulatory relationship* to emphasize the difference between causal discovery and statistical discovery.

The gold standard method for tackling this problem is randomized control experiments. For example, we can use gene knockdown experiments to knockdown miRNAs one by one while measuring changes (i.e. causal effects) in the expression level of mRNAs. However, such experiments are time consuming, expensive and not necessarily definitive. Fortunately, with the foundation created by Pearl (2000), the recent advances in causal discovery research have opened the door to discovering causal relationships from observational data.

Instead of conducting controlled experiments, we can use *do-calculus* (Pearl, 2000) to estimate the causal effects of a variable on other variables based on observational data. The *do-calculus* requires a causal structure of the variables to be given as a DAG (directed acyclic graph); however, such a structure is often unknown in reality. To bridge the gap, Maathuis *et al.* (2009, 2010) proposed a method to estimate causal effects from observational data alone. The method is called intervention calculus when the DAG is absent (IDA), and it includes two main phases: (i) to learn a causal structure from observational data and (ii) to apply *do-calculus* to infer causal effects.

Our method is based on IDA. Given the observational data of variables, IDA can capture the causal effects of the variables on one single response variable. We extend the application of IDA and build our model for multiple response variables, and then apply the model to discover the miRNA causal regulatory relationships. In our problem, miRNAs and mRNAs are nodes or variables in the model, and observational data are the expression profiles of the miRNAs and mRNAs. We can view genes as subjects and miRNAs as analogous to 'treatments', which may have causal effects on the 'responses' (i.e. expression levels) of the genes. Our aim is to measure the causal effect of each miRNA on mRNAs.

Applying this approach, we tackle two drawbacks of current miRNA regulatory relationships research. First, the method discovers causal relationships between miRNAs and mRNAs, not just the statistical relationships. Second, we assume that miRNAs and mRNAs interact with each other in a complex system, for instance, an miRNA can causally regulate mRNAs as well as other miRNAs. This assumption is more reasonable than the commonly used approach that considers only the bipartite of interactions between miRNAs and mRNAs. For example, (Zisoulis *et al.*, 2012) shows that let-7 can regulate other non-coding RNAs including miRNAs.

In this work, we first derive the solution to discover the causal miRNA regulatory relationships based on the IDA approach. Then we apply the method to the epithelial-to-mesenchymal transition (EMT) datasets. We implement a controlled experiment for the miR-200 family to validate the results. The outcome shows that the causal miRNA regulatory relationships discovered using our method largely overlap the findings from the experiment, suggesting that the causal relationships between miRNAs and mRNAs can be identified from expression profiles.

2 METHODS

2.1 Notations

Let $G = (\mathbf{V}, \mathbf{E})$ be a graph consisting a set of vertices $\mathbf{V}$ and a set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. In our context, $\mathbf{V} = \{X_1, \ldots, X_p\}$ is a set of random variables representing the expression levels of miRNAs and mRNAs, and the edges represent relationships between these variables.

In G, X_j is a *parent* of X_i if there is a directed edge $X_j \to X_i$. We use $pa_i(G)$ to represent the set of all parents of X_i. X_k is called a *sibling* of X_i if there is an undirected edge $X_i - X_k$. We denote the set of all siblings of X_i in G, $sib_i(G)$. When the graph G is clear from the context, we use pa_i and sib_i instead of $pa_i(G)$ and $sib_i(G)$.

A *v- structure* is an ordered triple of vertices, (X_i, X_j, X_k), such that in G, there exist directed edges $X_i \to X_j$ and $X_j \leftarrow X_k$, and X_i and X_k are not adjacent. X_j is then known as a *collider* in this *v*-structure.

Graph G is a DAG if G contains only directed edges and has no cycles. The *skeleton* of a DAG G is the undirected graph obtained from G by substituting undirected edges for directed edges. An *equivalence class* of DAGs is the set of DAGs that have the same skeleton and the same *v*-structures.

An equivalence class of DAGs can be uniquely described by a *completed partially directed acyclic graph* (CPDAG). A partially directed acyclic graph (PDAG) is a graph where the edges are either directed or undirected and one cannot trace a cycle by following the directions of the directed edges and any directions of the undirected edges. A PDAG is *completed* if (i) every directed edge exists also in every DAG belonging to the equivalence class and (ii) for every undirected edge, $X_i - X_k$, there exists a DAG with $X_i \leftarrow X_k$ and a DAG with $X_i \to X_k$ in the equivalence class.

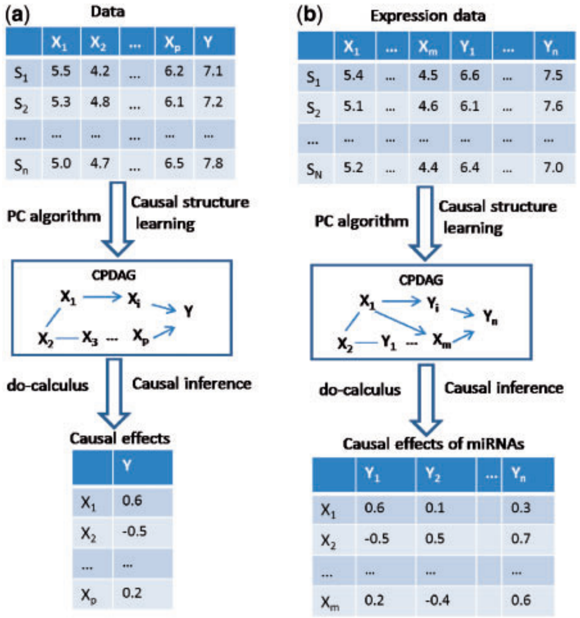
2.2 Method overview

Figure 1 illustrates the method used in our work. Details of the major steps, causal structure learning and causal inference are presented in Sections 2.3 and 2.4, respectively. Section 2.5 summarizes our method as an algorithm with the steps of implementing the method with expression data.

2.3 Causal structure learning

Learning causal structures from data plays a critical role in causality discovery. A popular method used to construct the structure from data is probabilistic graphical models. These graphical models are used to analyse and visualize conditional independence relationships between random variables (Neapolitan, 2004). The structure of conditional independence among the random variables is usually presented as a DAG of which vertices represent random variables and edges encode conditional dependence of the enclosing vertices.

Learning a DAG from data is highly challenging and complex, as the number of possible DAGs is super-exponential in the number of nodes (Robinson, 1971). Nevertheless, a number of methods have been proposed to learn DAGs from data. One approach is to restrict the search space to trees, such as the maximum weight spanning trees (Chow and



	X_1	X_2	...	X_p	Y
S_1	5.5	4.2	...	6.2	7.1
S_2	5.3	4.8	...	6.1	7.2
...	...	...	...	...	...
S_n	5.0	4.7	...	6.5	7.8

	Y
X_1	0.6
X_2	-0.5
...	...
X_p	0.2

	X_1	...	X_m	Y_1	...	Y_n
S_1	5.4	...	4.5	6.6	...	7.5
S_2	5.1	...	4.6	6.1	...	7.6
...	...	...	...	...	...	...
S_N	5.2	...	4.4	6.4	...	7.0

	Y_1	Y_2	...	Y_n
X_1	0.6	0.1		0.3
X_2	-0.5	0.5		0.7
...	...	...		...
X_m	0.2	-0.4		0.6

Fig. 1. (**a**) IDA. $S_1, ..., S_n$ are input data samples of the $p+1$ variables, $X_1, ..., X_p$ and the response variable Y. In a high-dimensional dataset, $p \gg n$. In the first phase, the causal structure in the form of a CPDAG is learnt from data, by applying the PC algorithm. In the second phase, *do-calculus* is used to infer the causal effects between X_i and Y, $i = 1, ..., p$. Y is the only response variable. (**b**) Our method to infer miRNA causal regulatory relationships. $X_1, ..., X_m$ represent the expression levels of miRNAs and $Y_1, ..., Y_n$ the expression levels of mRNAs. Nodes in the CPDAG are all miRNAs and mRNAs in the dataset. $Y_1, ..., Y_n$ are the response variables. The results are the causal effects that each miRNA has on every mRNA

Liu, 1968). Another approach, Greedy Equivalent Search (Chickering, 2002), works for a moderate number of nodes. Adding to these methods are the highly computation intensive Bayesian approaches for DAGs (Heckerman and Chickering, 1995).

An interesting alternative to those greedy or structurally restricted approaches is the Inductive Causation (IC) algorithm proposed by Verma and Pearl (1990). This algorithm is the base of causal structure learning research, and the algorithm is outlined in three phases as follows:

Phase 1—Find the undirected structure: For each pair of X_i and X_j, search for a set $S_{X_iX_j}$ such that $(X_i \perp\!\!\!\perp X_j | S_{X_iX_j})$. Add an undirected edge between X_i and X_j if no set $S_{X_iX_j}$ can be found.

Phase 2—Determine the *v*-structures: For each connected triple $X_i - X_j - X_k$, direct the edges and add a v-structure $X_i \rightarrow X_j \leftarrow X_k$ *if* $X_j \notin S_{X_iX_k}$, i.e. if and only if X_i and X_k are dependent given X_j.

Phase 3—Direct the remaining edges: Follow different rules/constraints to direct the remaining edges whenever possible, avoiding the creation of new *v*-structures and cycles.

The IC algorithm, however, leaves the details of Phases 1 and 3 unspecified and opens the door to further research.

Spirtes and Glymour proposed the PC algorithm (Spirtes *et al*., 2000), named after its inventors Peter Spirtes and Clark Glymour, to implement in detail the IC algorithm. The PC algorithm starts from a complete undirected graph and deletes recursively edges based on conditional independence decisions. This yields an undirected graph that can then be partially directed and further extended to a CPDAG.

With high-dimensional datasets, we need to select the most efficient tool to implement the conditional independence test in the PC algorithm. Lauritzen (1996) proved that when the distribution of variables is multivariate normal, the partial correlation test can be used as a conditional independence test as stated in the following theorem.

Theorem 1: *(Lauritzen, 1996) Assume that the distribution P of the random vector* $\mathbf{X} = \{X_1, \ldots, X_p\}$ *is multivariate normal. For* $i \neq j, i, j \in \{1, \ldots, p\}, \mathbf{k} \subseteq \{1, \ldots, p\} \setminus \{i, j\}$. $\rho_{i,j|\mathbf{k}}$ *denotes the partial correlation between* X_i *and* X_j, *given* $\{X_r; r \in \mathbf{k}\}$. *Then,* $\rho_{i,j|\mathbf{k}} = 0$ *if and only if* X_i *and* X_j *are conditionally independent, given* $\{X_r; r \in \mathbf{k}\}$.

Recently, Kalisch and Buhlmann (2007) showed that in the high-dimensional context, that is, the number of nodes p may be much larger than the sample size n, the PC algorithm with partial correlation test is uniformly consistent. This sheds light on the discovery of causal structures for high-dimensional data, such as gene expression datasets. In this article, we adapt the PC algorithm as a step of our algorithm in discovering the CPDAG, which includes miRNAs and mRNAs as nodes (see Fig. 1 and Step 2 of Section 2.5).

2.4 Causal inference

Our problem is to infer the causal effects (intervention or knockdown effects) of each single miRNA on the regulations of every mRNA under consideration, using only the observational data, the expression data of miRNAs and mRNAs.

The causal effects can be computed using *do-calculus* (Pearl, 2000), given a set of conditional dependencies from the expression data and a corresponding DAG model. The intervention $do(X_i = x_i')$ is an operation to force X_i to receive the value x_i'. The purpose is to observe how the system reacts with this intervention, or in other words, how the distribution of a variable may change after the intervention. Pearl (2000) formalized the definition of post-intervention distribution by the pre-intervention distribution as shown in the following definition.

Definition 1: *Let* $\mathbf{X} = \{X_1, X_2, \ldots, X_{p+1}\}$ *be a set of variables. The distribution generated by an intervention* $do(X_i = x_i')$, $i \in \{1, \ldots, p+1\}$ *on the set of variables is given by the truncated factorization formula:*

$$P(x_1, \ldots, x_{p+1} | do(X_i = x_i')) = \begin{cases} \prod_{j \neq i} P(x_j | pa_j) & x_i = x_i', \\ 0 & x_i \neq x_i' \end{cases} \quad (1)$$

where pa_j *is a set of parent nodes of* X_j, *and* $P(x_j | pa_j)$ *are the pre-intervention conditional distributions.*

To deal with high-dimensional datasets, we made some assumptions in Theorem 1 and the PC algorithm in the causal structure learning phase. The assumptions are that the distribution of the variables $\mathbf{X}$ is multivariate normal, Markovian and faithful to the true (unknown) causal DAG. With the same assumptions, we can now calculate the intervention effects based on the discussions in Maathuis *et al*. (2009), which are formalized in the following theorem.

Theorem 2: *Let* $X_1, \ldots, X_p, X_{p+1} = Y$ *be jointly Gaussian. The causal effect of* X_i *on* Y *for* $i = 1, \ldots, p$ *can be calculated as:*

$$ef(X_i, Y) = \beta_{i|pa_i} = \begin{cases} 0 & Y \in pa_i, \\ \textit{Coefficient of } X_i \textit{ in } Y \sim X_i + pa_i & Y \notin pa_i \end{cases} \quad (2)$$

in which pa_i *is the set of parent nodes of* X_i, *and* $Y \sim X_i + pa_i$ *is the shorthand for the linear regression of* Y on X_i and pa_i.

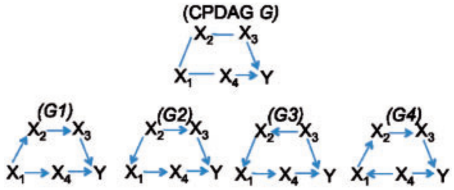


Fig. 2. A CPDAG G with the DAGs $G_1, ..., G_4$ in its equivalence class (Maathuis *et al.*, 2009)

At this point, we have had the tools, the PC algorithm, to learn the CPDAG and Theorem 2 to estimate the causal effects for each DAG in the equivalence class CPDAG. For illustration, consider the CPDAG G in Figure 2, where we are interested in *ef*(X_1, Y). We can first list all the DAGs in the equivalence class, $G_1, \dots, G_4$. The causal effect *ef*(X_1, Y) can then be calculated in each DAG by applying Theorem 2.

However, when the number of nodes in a CPDAG is large, the number of DAGs grows quickly. As the essence of the calculation of *ef*(X_i, Y) is the sets of parent nodes of X_i, Maathuis *et al.* (2009) suggest searching the CPDAG directly for all the valid parental sets of X_i rather than identifying all the DAGs in the equivalence class for the parental sets. We restate this suggestion by the following theorem.

Theorem 3: *Let $S \subseteq sib_i(G)$. A set $pa_i(G) \cup S$ is a parental set of X_i if and only if directing all edges between X_i and vertices in S toward X_i, and all edges between X_i and vertices in $sib_i(G) \setminus S$ away from X_i will not create any new v-structures.*

In Figure 2, $pa_1(G) = \emptyset$, $sib_1(G) = \{X_2, X_4\}$. Therefore, the candidate parental sets of X_1 are $\emptyset$, $\{X_2\}$, $\{X_4\}$ and $\{X_2, X_4\}$. However, the set $\{X_2, X_4\}$ is not the parental set of X_1, as directing edges between $\{X_2, X_4\}$ and X_1 toward X_1 will create a new *v*-structure. Therefore, applying Theorem 2 to the three parental sets, we have the multiset that contains all the causal effects of X_1 on Y, $E_1 = \{\beta_{1|\emptyset}, \beta_{1|X_2}, \beta_{1|X_4}\}$. A multiset is a set in which elements are allowed to appear more than once.

2.5 Algorithm

Now, we are able to present the algorithm for discovering miRNA causal regulatory relationships in four steps as follows:

Step 1: Identify differentially expressed genes, i.e. those genes whose expression values vary significantly across the conditions (categories) of the samples. First, we divide the dataset as per the categories (e.g. normal and cancer) and identify the differentially expressed genes across different categories. We assume that those genes with little or zero change in expression between the categories play a minimum role in the biological processes and are thus omitted. Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ represent miRNAs and mRNAs, respectively, that are identified to be differentially expressed. Combining the expression profiles of the differentially expressed miRNAs and mRNAs, we have a dataset for the $(m+n)$ variables.

Step 2: Use the PC algorithm to estimate the CPDAG G of the $(m+n)$ variables and the conditional dependencies of the variables We use partial correlation as a conditional independence test for the PC algorithm, as the partial correlations are easy to implement in a high-dimensional dataset. The validity of this test has been shown by Theorem 1.

Step 3: Estimate the causal effects of each miRNA on each mRNA Naturally, we can identify all possible DAGs in the CPDAG, and then use Theorem 2 to estimate the causal effects with each DAG. However, when the number of nodes in the CPDAG is large, as described previously, we apply Theorem 3 to reduce the search space of possible DAGs. The causal effects of each miRNA on an mRNA can be achieved by applying Theorem 2 to each of the parental sets of the miRNA.

Step 4: Output the miRNA causal effects. For each miRNA, the outcome of Step 3 is an array of multisets, and each multiset contains all causal effects of the miRNA on an mRNA (note that the causal effects of an miRNA on a particular mRNA may have multiple values, as we estimated the effects from its different parental sets). With each of the multisets, in this step, we select the causal effect value with the smallest absolute value, and output it as the causal effect of the miRNA on the mRNA.

3 RESULTS

In this section, we present the results and analysis of applying our algorithm to the NCI-60 dataset for EMT. The dataset includes the miRNA expression profiles for the NCI-60 panel of 60 cancer cell lines from Søkilde *et al.* (2011). They are available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26375. The mRNA expression profiles for NCI-60 were downloaded from ArrayExpress http://www.ebi.ac.uk/arrayexpress, accession number E-GEOD-5720. Cell lines categorized as epithelial (11 samples) and mesenchymal (36 samples) are used for this work.

3.1 Implementation

In this section, we describe how the algorithm in Section 2.5 is applied to the NCI-60 dataset.

We divide the samples in the dataset into two groups based on their conditions, epithelia and mesenchymal. The differentially expressed gene analysis is performed using the *limma* package of Bioconductor (Smyth, 2005), and 1635 mRNA probes and 43 miRNA probes are identified to be differentially expressed with *p*-value <0.05 (adjusted *p*-value). Detailed results are provided in the Supplementary Material.

The PC algorithm is then used to estimate the CPDAG G. The input for the PC algorithm is a 47×1678 matrix of the 47 samples of miRNA and mRNA expression values for the 1678 probes. The PC algorithm uses partial correlation as the conditional independence test. The significance level of this test, α, is the tuning parameter for the PC algorithm. In this implementation, we use the *R*-package called *pcalg* and set $\alpha = 0.01$ (p-value<0.01).

Then the causal effect *ef*(miRNA, mRNA) is calculated using Theorem 3 for each pair of miRNA and mRNA. As stated in Step 4 of the algorithm, we use the causal effect value whose absolute value is the smallest of all in the obtained multiset as the result for *ef*(miRNA, mRNA). For example, if *ef*(miRNA, mRNA)= $\{0.4, -0.5, 0.4, 0.3\}$, then we choose *ef*(miRNA, mRNA)= 0.3 as the result. This choice assures a consistent result because if the reported result of *ef*(miRNA, mRNA) is high, all the estimations across different DAGs (parental sets) must be high.

To overcome the problem of small number of samples in biological data, we bootstrap the data 100 times and take the median of the 100 estimates for each pair of (miRNA, mRNA) as the final results.

The final result is displayed in a tabular form (as the final table in Fig. 1b). The value in each cell of the table is the causal effect that an miRNA has on an mRNA. It is worth noting that the causal effect of each individual miRNA on each individual mRNA in the result is not necessarily a consequence of the direct interaction between an miRNA and an mRNA. In fact, it indicates the total regulatory effect that the miRNA has on the mRNA, and the regulation pattern can be direct, indirect or both.

We can rank all the genes based on the absolute values of the causal effects that a particular miRNA has on them. In this article, we rank the genes regulated by miR-200a and miR-200b, respectively, and extract the top 20, 50 and 100 genes in each case based on the causal effects that the miRNA has on them for experimental validation, which will be discussed in the next section. The computational results that show the causal effects of miRNAs on mRNAs can be found in the Supplementary Material.

3.2 Validation by experiments

The most effective way to validate the results is by biological experiments. In this article, we design the follow-up controlled experiment on the samples of MDA-MB-231.

We measure the gene expression level in the MDA-MB-231 samples transfected with the miR-200 family and in the MDA-MB-231 samples without the miR-200 family (controls). Specifically, we have one sample transfected with miR-200a, two samples transfected with miR-200b and two control samples. The reason for choosing the miR-200 family for experimentation is that the miR-200 family has been confirmed as a biomarker for EMT (Gregory *et al*., 2008). The differentially expressed genes from the controlled and transfected samples are then used to validate the computational results. With the experiment, 345 and 533 genes are identified to be regulated by miR-200a and miR-200b, respectively, with adjusted p-value <0.05 [the p-values are obtained from *limma* (Smyth, 2005) with Benjamini and Hochberg correction method]. The detailed experimental results are in the Supplementary Material, and a summary of validation results is given later in the text in term of the number of genes confirmed by the experiments and the consistency in the signs of regulations:

(1) *A significant number of genes regulated by miR-200 family being confirmed.* There is significant overlap between the top genes regulated by miR-200 that are discovered by our method and the differentially expressed genes identified from the experiment. For miR-200a and miR-200b, we extract respectively from the computational results the top 20, 50 and 100 genes based on the causal effects that the miRNA has on them, for validation. The numbers of genes that have been confirmed by the experiment are shown in Table 1. If we consider the results in Table 1 as precisions for top 20, 50 and 100, the corresponding recalls are 10/345, 22/345 and 42/345 for miR-200a and 14/533, 32/533 and 53/533 for miR-200b. When we select the top- k genes for validation, the recall rates will be improved with the increase of k. In practice, depending on the goal of an experiment, k can be adjusted to have a good balance between the recall rates and precisions.

(2) *Consistency in the signs of regulations.* In the results discovered by our method, a positive/negative causal effect that an miRNA has on an mRNA means that increasing the miRNA expression level will result in an increasing/decreasing expression level of the mRNA. Therefore, the signs of the causal effects, either positive or negative, indicate up- or downregulations in the context of miRNA–mRNA interactions. In the experiments, to identify the up-/downregulations of miR-200a or miR-200b on an mRNA, we compare the expression level of the mRNA between the control samples and the samples transfected by miR-200a or miR-200b. The greater level of the mRNA expression level in the transfected samples implies the upregulation (positive causal effect) and *vice versa*.

We compare the results of the experiments and our method regarding the signs of the regulations of miR-200a and miR-200b on the top 20, 50 and 100 genes, respectively. With miR-200a, for the top 20 and 50 genes, respectively, the signs discovered using our method are 100% consistent with (the same as) those identified by the experiments, and 95% of the regulations discovered by our method on the top 100 genes have the same signs as those discovered by the experiments. With miR-200b, for the top 20, 50 and 100 genes, the consistency of the signs are 100, 94 and 94%, respectively.

Table 1. Number of genes confirmed to be regulated by miR200 family

	Number of confirmed genes for miR200a	p-value	Number of confirmed genes for miR200b	p-value
Top 20	10	0.0544	14	0.0339
Top 50	22	0.0313	32	0.0125
Top 100	42	0.0103	53	0.1442

p-value is calculated based on the probability of the result that can occur by chance.

The significant numbers of genes that have been confirmed by experiments together with the high level of consistency in the signs of causal effects suggest that our method can be used as a new tool to assist in the experimental design for discovering miRNA causal regulatory relationships.

3.3 Functional validation of mRNAs for EMT

As the results are obtained based on the EMT datasets, we extract the top 150 genes based on the total causal effects that all the miRNAs have on them, and we validate the genes against the literature knowledge of pathways and functional biomarkers of EMT.

We use GeneGo Metacore from GeneGo Inc. to identify the pathways previously discovered in the literature that involve the genes in our top 150 list. The results in Table 2 show that the top genes that receive the highest causal effects are highly relevant to the regulation of EMT. For instance, pathways number 6, 7, 8 and 11 are direct pathways of the development of EMT, and others are important pathways involved in the process of EMT. An example of pathway number 1 can be found in the Supplementary Material.

The pathway results suggest that the top genes causally regulated by miRNAs are highly relevant to the biological condition (EMT) of the datasets. Therefore, we can also use the method in this article for identifying functional miRNA regulatory modules for a specific biological condition. This can be done

Table 2. GeneGo mapped pathways for genes in the top 150 list

Id	Maps	*p*-value
1	Cytoskeleton remodelling_Keratin filaments	5.941E-11
2	Cell adhesion_Tight junctions	2.800E-09
3	Development_WNT signalling pathway. Part 2	2.434E-05
4	Cell adhesion_Gap junctions	9.945E-04
5	Cell adhesion_Role of CDK5 in cell adhesion	1.516E-03
6	Development_TGF-β–dependent induction of EMT via SMADs	1.566E-03
7	Development_MicroRNA-dependent inhibition of EMT	1.887E-03
8	Development_TGF-β–dependent induction of EMT via MAPK	3.667E-03
9	Cell adhesion_ECM remodelling	4.883E-03
10	Development_WNT signalling pathway. Part 1. Degradation of β-catenin in the absence WNT signalling	6.902E-03
11	Development_Regulation of EMT	8.703E-03
12	Cell adhesion_Cadherin-mediated cell adhesion	1.273E-02

The pathways are highly relevant to the regulation of EMT.

Table 3. Genes in the top 150 list: sub-categories of cellular movement, functional biomarkers of EMT

Functions	mRNAs	Number	*p*-value
Invasion	CCDC88A, CDH1, ANPEP JUP, CLDN4, BMP4, S100P CHST10, MST1R, CLDN3 ST14, CLDN7, ELF3, VIM EPCAM, EPN3, KLK6, ZEB1 ITGB4, SPARC, PLXNB1	21	9.1E-08–2.35E-02
Migration	VIM, KLF5, ANPEP, JUP FERMT1, BMP4, CLDN7 CDH1, MCF2L, CCDC88A DDR1, CLDN4, SPACRC GRB7, KRT8, ZEB1, STAP2 MST1R, PIK3C2B, S100A14 ARHGAP8/PRR5-ARHGAP8 RHOD, ITFB4, KLK6, F11R PLXNB1, CXCL16, S100P	28	2.15E-04–2.36E-02

Genes in invasion and migration, sub-categories of cellular movement, are functional markers of EMT. The results are generated from IPA. The last column is the range of *p*-values for all genes in the group.

by ranking all the miRNAs that regulate the group of genes, e.g. the top 150 genes, and then setting thresholds for the values of causal effects to extract the miRNA–mRNA regulatory modules.

We also conduct enrichment analysis for the top 150 genes to investigate the relevance between the gene functions and the biological condition of the dataset. We use Ingenuity Pathway Analysis (IPA, Ingenuity Systems, www.ingenuity.com), a commercial application that calculates the association between a particular gene set and known functions and pathways, for this purpose. The top 150 genes are significantly enriched for several biological functions. The top five functions listed by IPA are known to be critical for EMT. They are cellular movement, cell-to-cell signalling and interaction, cellular assembly and organization, cellular growth and proliferation and cell morphology. Especially, there are several genes in the sub-categories of cellular movement, invasion and migration, which have been identified as the functional biomarkers of EMT (Zeisberg and Neilson, 2009). Table 3 shows the genes in the top 150 list that are in the classes of invasion and migration, which are functional biomarkers for EMT.

4 CONCLUSIONS AND DISCUSSIONS

miRNAs have been regarded as one of the most important regulators. Identifying their functions and regulatory mechanism is critical in understanding biological processes of organisms. Both biology and computational biology have seen great efforts made to elucidate miRNA functions. However, the precise regulatory relationships between miRNAs and mRNAs remain elusive. The statistical relationships that most computational methods can discover only reveal associations or correlations but not causality.

In this article, we present an alternative approach for revealing the causal relationships between miRNAs and mRNAs. This method is inspired by the recent advances in causality discovery research on large-scale datasets, especially the IDA method (Maathuis *et al.*, 2009, 2010). Based on this method, we derived the solution for miRNA causal regulatory relationship discovery. Our method makes use of the expression profiles of miRNAs and mRNAs and evaluates the causal effects of miRNAs on mRNAs. Our assumption is that miRNAs and mRNAs interact with each other in a complex system. This is a more realistic assumption than those that consider the interactions occur between miRNAs and mRNAs only.

One promising aspect of our approach is that the causal effects we get are similar to those from a randomized controlled experiment. Results from the follow-up experiments have indicated that our method can be used to effectively identify a set of genes causally regulated by miRNAs. Although the discovery based on observational data alone may never replace the actual intervention experiments, these results can serve as a tool for the design of follow-up miRNA experiments.

In this article, we have applied our method to the EMT datasets. The results show that our method has effectively identified the causal relationships between miRNAs and mRNAs. The follow-up experiments confirm the validity of the identified causal relationships for the miR-200 family. The results for other miRNAs remain open for further research and experiments. In addition, the enrichment analysis from the extracted results shows the consistency of the results of our method with the literature regarding biological functions of EMT.

We expect that our method can be applied to different datasets to discover the miRNA causal regulatory relationships and to the complex regulatory networks, including miRNAs, transcription factors and mRNAs. For extension purposes, the results of this method can be further analysed and interpreted in various ways. For instance, we can focus on the miRNA regulatory networks, or discover the level of regulations that a group of miRNAs has on a particular gene, or compare the similarity of two miRNAs based on the causal effects they have on a group of genes.

Our approach complements existing prediction methods that are based on sequence complementarity. Current understanding of the nature of miRNA–mRNA interactions is still somewhat limited. Predictions based on sequence complementarity and/or structural stability of the putative duplex have a very high rate of both false positives and false negatives. A major reason for this is the role played by RNA folding as well as accessibility because of protein binding. The more sophisticated prediction methods such as mirSVR (Betel *et al*., 2010) incorporate training of the models on experimental data derived from genome-wide mRNA expression changes after miRNA transfection. Our approach predicts miRNA–mRNA interactions based on gene expression data. Some of the interactions have already been experimentally validated, with reporter assays and the like, supporting the validity of our approach. Some interactions from our study are novel and can be good candidates for future investigation. Some of the interactions and regulatory modules are likely to be applied to other biological scenarios in which the component miRNAs and mRNAs are expressed, whereas others, such as those involving miR-200, will be restricted to epithelial cells and the EMT process because the miRNA and/or mRNA targets have cell-specific expression. However, because our approach not only indicates bi-component miRNA–mRNA interactions but also generates multicomponent regulatory modules, the approach has the power of suggesting important regulatory pathways that warrant future study. Some of these are likely to be restricted to the EMT process, but they are worthy of investigation because EMT has a major role in development and in cancer progression.

Using our method (which is based on the IDA method), we can theoretically infer the causal relationships between every two variables in the dataset. However, to make it computationally efficient for a specific biological problem, in practice, we should restrict the number of 'cause' variables and/or the number of 'effect' variables. For instance, in our method, we only focus on the causal relationships between miRNAs (causes) and mRNAs (effects). Other relationships, such as miRNA–miRNA, mRNA–mRNA or mRNA–miRNA, although in theory can be inferred, are beyond the scope of this article, and the complexity of the algorithm would increase significantly.

Funding: Australian Research Council Discovery Project (DP130104090, in part) and National Health and Medical Research Council Project Grant (APP1008327, in part).

Conflict of Interest: none declared.

REFERENCES

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Berezikov,E. *et al*. (2006) Approaches to microRNA discovery. *Nat. Genet*., **38**, 2–8.

Betel,D. *et al*. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*., **11**, R90.

Chen,J.F. *et al*. (2006) The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat. Genet*., **38**, 228–233.

Chickering,D.M. (2002) Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res*., **2**, 445–498.

Chow,K. and Liu,N.C. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, **14**, 462–467.

Cui,Q. *et al*. (2006) Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol*., **2**, 1–7.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Gregory,P.A. *et al*. (2008) The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat. Cell Biol*., **10**, 593–601.

Heckerman,D. and Chickering,D.M. (1995) Learning Bayesian networks: the combination of knowledge and statistical data metrics for belief networks. Mach. Learn., **20**, 197–243.

Huang,J.C. *et al*. (2007) Using expression profiling data to identify human microRNA targets. *Nat. Methods*, **4**, 1045–1050.

Jin,P. *et al*. (2004) Biochemical and genetic interaction between the fragile X mental retardation protein and the microRNA pathway. *Nat. Neurosci*., **7**, 113–117.

Joung,J.G. and Fei,Z. (2009) Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model. *Bioinformatics*, **25**, 387–393.

Joung,J.G. *et al*. (2007) Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics*, **23**, 1141–1147.

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

Kim,V.N. and Nam,J.W. (2006) Genomics of microRNA. *Trends Genet*., **22**, 165–173.

Lauritzen,S. (1996) *Graphical Models*. Oxford University Press, Oxford.

Liu,B. *et al*. (2009) Discovery of functional miRNA mRNA regulatory modules with computational methods. *J. Biomed. Inf*, **42**, 685–691.

Liu,B. *et al*. (2010) Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, **26**, 3105–3111.

Maathuis,H.M. *et al*. (2009) Estimating high-dimensional intervention effects from observational data. *Ann. Stat*., **37**, 3133–3164.

Maathuis,H.M. *et al*. (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Methods*, **7**, 247–249.

Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.

Neapolitan,R. (2004) *Learning Bayesian Networks*. Pearson Prenctice Hall, Upper Saddle River, NJ.

Pearl,J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Poy,M.N. *et al*. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, **432**, 226–230.

Robinson,J. (1971) Counting labeled acyclic digraphs. In: *New Directions in the Theory of Graphs: Proceedings of the Third Ann Arbor Conference on Graph Theory*. Academic Press, New York, pp. 239–273.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al*. (eds) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. pp. 397–420.

Søkilde,R. *et al*. (2011) Global microRNA analysis of the NCI-60 cancer cell panel. *Mol. Cancer Ther*., **10**, 375–384.

Spirtes,P. *et al*. (2000) *Causation, Prediction, and Search,* 2nd edn. MIT Press, Cambridge, MA.

Tran,D.H. *et al*. (2008) Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics*, **9** (**Suppl. 12**), S5.

Verma,T. and Pearl,J. (1990) Equivalence and synthesis of causal models. In: *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, pp. 220–227.

Xu,C. *et al*. (2007) The muscle-specific microRNAs miR-1 and miR-133 produce opposing effects on apoptosis by targeting HSP60, HSP70 and caspase-9 in cardiomyocytes. *J. Cell Sci*., **120** (Pt. 17), 3045–3052.

Zeisberg,M. and Neilson,E.G. (2009) Biomarkers for epithelial-mesenchymal transitions. *J. Clin. Invest*., **119**, 1429–1437.

Zhao,Y. *et al*. (2005) Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, **436**, 214–220.

Zisoulis,D.G. *et al*. (2012) Autoregulation of microRNA biogenesis by let-7 and Argonaute. *Nature*, **486**, 541–544.