Reverse Engineering Regulatory Networks in Cells Using a Dynamic Bayesian Network and Mutual Information Scoring Function

Haodi Jiang, Turki Turki*, Jason T. L. Wang
King Abdulaziz University and New Jersey Institute of Technology
Jeddah, Saudi Arabia and Newark, New Jersey, USA
*Corresponding author: tturki@kau.edu.sa

Abstract—**In systems biology, two important regulatory networks are gene regulatory networks (GRNs) and regulatory networks of microRNAs (RNMs). A GRN is modeled as a directed graph in which a node represents a gene or transcription factor (TF), and an edge from a TF to a gene indicates that the TF regulates the expression of the gene. An RNM is modeled as a bipartite directed graph with two disjoint sets of nodes: a set of nodes that represent microRNAs (miRNAs) and a set of nodes that represent genes or TFs. Directed edges between these two sets of nodes represent miRNA-target interactions or TF-miRNA regulatory relations. In this paper, we present an approach to reverse engineering GRNs and RNMs using a dynamic Bayesian network and mutual information scoring function. Our approach is able to automatically infer both GRNs and RNMs from time series of expression data. Experimental results on different datasets show that our approach is more accurate than other time-series based network inference methods.**

Keywords—data mining; machine learning; Bayesian networks; mutual information; systems biology

I. Introduction

Interactions among genes are mediated by gene products such as DNA-binding proteins, including transcription factors (TFs), and microRNAs (miRNAs). The analyses of gene interactions can be difficult if time-series data are part of the experimental design [1]. Among the previously ignored components of gene networking are miRNAs. In addition to their importance as regulatory elements in gene expression, the capacity of miRNAs to be transported from cell to cell implicates them in a panoply of pathophysiological processes that include antiviral defense, tumorigenesis, lipometabolism, and glucose metabolism. This role in disease complicates our understanding of translational regulation via endogenous miRNAs. Understanding the biogenesis, transport, and mechanisms of action of miRNAs on their target RNA (gene) would result in possible therapies.

In this paper we study two regulatory networks in cells, namely gene regulatory networks (GRNs) [1] and regulatory networks of microRNAs (RNMs) [2]. A GRN is modeled as a directed graph in which a node represents a gene or transcription factor (TF). An edge from a TF to a gene indicates that the TF regulates the expression of the gene. An RNM is modeled as a bipartite directed graph with two disjoint sets of nodes: a set U of nodes that represent microRNAs (miRNAs) and a set V of nodes that represent genes or TFs. An edge from an miRNA node in U to a gene node in V shows an miRNA-target interaction. On the other hand, an edge from a TF node in V to an miRNA node in U shows a TF-miRNA regulatory relation. Notice that a node in V can be both a target gene and a TF.

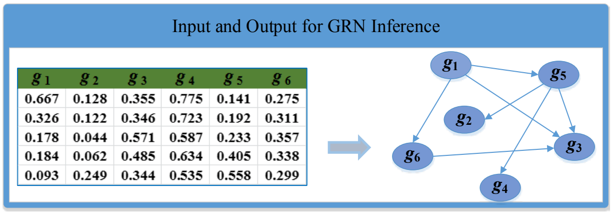
We present an approach to reverse engineering (inferring) GRNs and RNMs from time series of expression data using a Bayesian network and mutual information scoring function. Analyses of time-series expression data can show the chronological expression of specific genes/miRNAs or groups of genes/miRNAs. These temporal patterns can be used to infer or propose causal relationships in gene/miRNA regulation [3]. Thus, genes in the networks can modulate the extent of each other's gene expression over the life span of a cell or a whole organism. Time-series expression data sheds light on a complex but measurable regulatory system, allowing for a more precise inference of gene/miRNA interaction. Although automated GRN inference using time series of expression data has been studied in the past [3, 4], to our knowledge, there is no previous work focusing on automated RNM inference using time-series expression data. Our approach is able to automatically infer both GRNs and RNMs from time series of expression data.

II. Methodology

A. Problem Statement

A time series of expression data for a gene (miRNA, respectively) consists of a series of time points where each time point is associated with an expression value of the gene (miRNA, respectively). The problem at hand is to infer or reconstruct a network from time series of expression data. Specifically, for GRN inference, the input of our approach contains n genes where each gene has a time series of expression data. The output of our approach is an inferred gene regulatory network (GRN).

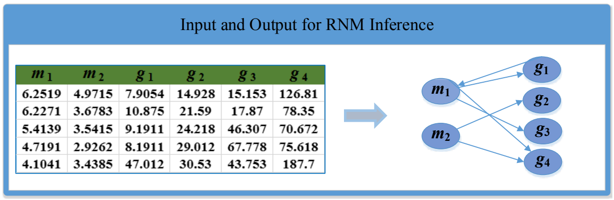
0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.00-67



Input and Output for GRN Inference

g_1	g_2	g_3	g_4	g_5	g_6
0.667	0.128	0.355	0.775	0.141	0.275
0.326	0.122	0.346	0.723	0.192	0.311
0.178	0.044	0.571	0.587	0.233	0.357
0.184	0.062	0.485	0.634	0.405	0.338
0.093	0.249	0.344	0.535	0.558	0.299

Fig. 1. Inferring a GRN (right) from a time-series gene expression dataset (left).



Input and Output for RNM Inference

m_1	m_2	g_1	g_2	g_3	g_4
6.2519	4.9715	7.9054	14.928	15.153	126.81
6.2271	3.6783	10.875	21.59	17.87	78.35
5.4139	3.5415	9.1911	24.218	46.307	70.672
4.7191	2.9262	8.1911	29.012	67.778	75.618
4.1041	3.4385	47.012	30.53	43.753	187.7

Fig. 2. Inferring an RNM (right) from a set of time series of expression data (left) where the expression data are for miRNAs m_1, m_2 and genes g_1, g_2, g_3 and g_4.

Fig. 1 illustrates the GRN inference problem where n = 6. In this figure, there are five time points for each gene.

For RNM inference, the input of our approach contains time series of expression data for m miRNAs and n genes. The output of our approach is an inferred regulatory network of miRNAs (RNM), which is a bipartite directed graph.

Fig. 2 illustrates the RNM inference problem where m = 2 and n = 4. In this figure, there are again five time points for each miRNA and gene. An edge from an miRNA (e.g. m_1) to a gene (e.g. g_1) shows an miRNA-target interaction. On the other hand, an edge from a TF (e.g. g_1) to an miRNA (e.g. m_1) shows a TF-miRNA regulatory relation. Here, g_1 is both a target gene and a TF.

B. Proposed Approach

We present a dynamic Bayesian network approach to network inference. Given time-series expression data, this approach attempts to estimate parameters of a dynamic Bayesian network such that they best fit the given expression data, and produces the best graph topology of an inferred network. The approach is expressed in (1) below [5].

$$\hat{\theta} = \arg\max_{\theta} P(D;\theta,G) \tag{1}$$

Here D represents the given time-series expression data, G is the inferred graph structure, and θ represents the parameters of the dynamic Bayesian network. $\hat{\theta}$ represents the optimal parameter estimates. According to Bayes' rule, we have:

$$P(\theta \mid D,G) = \frac{P(D \mid \theta,G)P(\theta \mid G)}{P(D \mid G)} \tag{2}$$

A dynamic Bayesian network usually constructs two subgraphs: a prior network and a transition network [6]. The prior network contains edges established based on expression values at the same time point. The transition network contains edges established based on expression values at the same or different time points. An edge from g_x to g_y may be established based on the expression value of g_x at time point t and the expression value of g_y at time point t, t+1, t+2, …, or t+d. Thus, first-order Markovianity in the dynamic Bayesian network needs to be replaced by dth-order Markovianity in order to solve the network inference problem at hand [7]. (In the study present here, d is set to 2 and 3 for GRN and RNM inference, respectively.)

To identify the best network structure in the search space, we use a mutual information scoring function [8]. The goodness-of-fit of a network is measured by the total mutual information shared between each node and its parents, penalized by a term which quantifies the degree of statistical significance of this shared information. To calculate the mutual information, we need to discretize the expression values given in the input. To facilitate discussions, in what follows we will focus on GRN inference; RNM inference uses a similar way to calculate the mutual information.

Let $r_1,\ldots,r_n$ be the number of discrete states corresponding to the n genes $g_1,\ldots,g_n$ in the input. Let D represent the expression values in the input with N time points. Let G be a dynamic Bayesian network (i.e., an inferred network). Let $Pa_G(g_i) = \left\{g_i^1,\ldots,g_i^{s_i}\right\}$ be the set of parents of g_i in G. The numbers of discrete states corresponding to $g_i^1,\ldots,g_i^{s_i}$ are $r_i^1,\ldots,r_i^{s_i}$ respectively; $s_i = \left|Pa_G(g_i)\right|$. The mutual information scoring function, denoted $F(G:D)$, is calculated as follows [8]:

$$F(G:D)= \sum_{\substack{i=1 \\ |Pa_G(g_i)|\neq 0}}^{n} \left\{ 2N \times MI_D(g_i, Pa_G(g_i)) - \max_{\sigma_i} \sum_{j=1}^{s_i} \chi_{\alpha, w_{i\,\sigma_i(j)}} \right\} \tag{3}$$

where $MI_D(g_i, Pa_G(g_i))$ is the mutual information between g_i and its parents as estimated from D, $\max_{\sigma_i} \sum_{j=1}^{s_i} \chi_{\alpha, w_{i\,\sigma_i(j)}}$ is the penalization component, where $\chi_{\alpha, w_{i\,\sigma_i(j)}}$ is the value such that $P(\chi^2(w_{i\,j}) \le \chi_{\alpha, w_{i\,\sigma_i(j)}}) = \alpha$ (the Chi-square distribution at significance level $1-\alpha$), and the term $w_{i\,\sigma_i(j)}$ is defined as:

$$w_{i\,\sigma_i(j)} = \begin{cases} (r_i - 1)(r_i^{\sigma_i(j)} - 1)\prod_{k=1}^{j-1} r_i^{\sigma_i(k)} & j = 2,\ldots,s_i \\ (r_i - 1)(r_i^{\sigma_i(1)} - 1) & j = 1 \end{cases} \tag{4}$$

where $\sigma_i = \{\sigma_i(1),\ldots,\sigma_i(s_i)\}$ is a permutation of the index set $\{1,\ldots,s_i\}$ of $Pa_G(g_i)$. The permutation that produces the maximum penalization value is the one where the first gene has the largest number of states, the second gene has the second largest number of states, the third gene has the third largest number of states, and so on [8].

We can use a method similar to that in [7] to solve the problem at hand, i.e., to find the best dynamic Bayesian network with the above mutual information scoring function. However, [7] does not deal with regulatory networks of miRNAs and cannot handle bipartite directed graphs. To solve this problem, we exploit the prior knowledge concerning miRNAs and genes, eliminating the edges between miRNAs and the edges between genes from consideration. As our experimental results show later, this approach is promising.

III. Experiments and Results

A. Datasets

We carried out a series of experiments to evaluate the performance of our approach and related methods. Two sets of time series of expression data and their corresponding regulatory networks were used in the experiments.

The first dataset was retrieved from the DREAM (Dialogue for Reverse Engineering Assessments and Methods) initiative website [9]. The DREAM initiative organizes annual reverse engineering competitions called the DREAM challenges. We took the gene regulatory networks (GRNs) with 10 genes from the DREAM4 edition. There are five sets of time-series gene expression data in the DREAM4 challenge. Each dataset contains five time-series, where each time series has 21 time points, for 10 genes. Each time-series dataset is associated with a gold standard file, where the gold standard represents the ground truth of the network structure for the time-series data. Each edge in the gold standard represents a true regulatory relationship between two genes.

The second dataset contains a regulatory network of miRNAs (RNM). This RNM was retrieved from AtmiRNET, which is a web-based resource containing regulatory networks of *Arabidopsis* microRNAs [2]. This network, represented by a bipartite directed graph, consists of two miRNAs, ath-mir156a and ath-mir157d, as well as 14 genes (see Fig. 3). Among the 14 genes, there are two transcription factors (TFs), at1g18860 and at1g14580, and 12 target genes. TF at1g18860 regulates the expression of miRNA ath-mir156a and TF at1g14580 regulates the expression of miRNA ath-mir157d. These TF-miRNA regulatory relations are represented by dotted lines in Fig. 3. In addition, there are 21 miRNA-target interactions, represented by edges in orange and light blue color respectively in Fig. 3.

The time-series expression profiles of *Arabidopsis* miRNAs and genes were retrieved from mirEX and AtGenExpress [2] respectively. There are 12 (8, respectively) time points in the expression dataset of *Arabidopsis* miRNAs (genes, respectively). More precisely, the expression datasets of *Arabidopsis* miRNAs and *Arabidopsis* genes have 8 time points in common. The expression dataset of *Arabidopsis* miRNAs has 4 additional time points, thus having 12 time points in total. Since our approach requires the input data to have the same number of time points (cf. (3)), we use the expression values associated with the 8 common time points in our study.

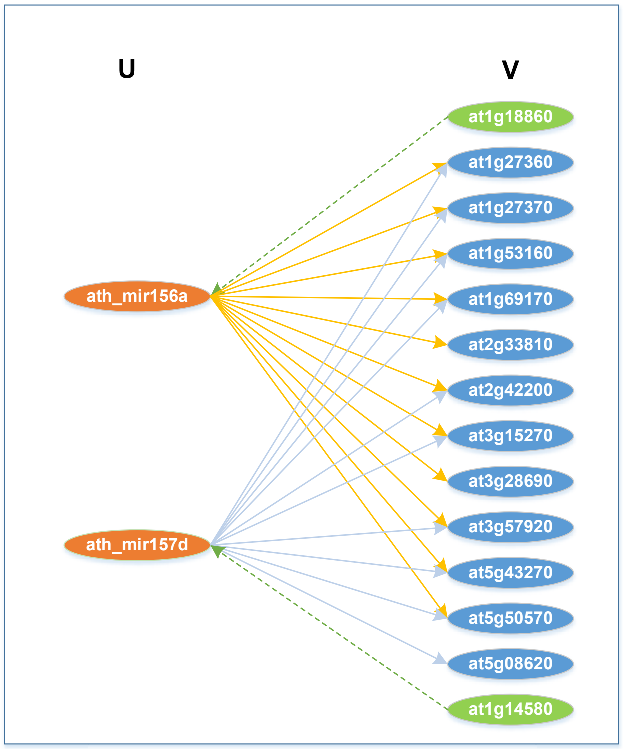


Fig. 3. The regulatory network of *Arabidopsis* microRNAs used in the study.

B. Experiment Setup

We define a true positive to be a true present edge that is an inferred present edge. A false positive is a true absent edge that is an inferred present edge. A true negative is a true absent edge that is an inferred absent edge. A false negative is a true present edge that is an inferred absent edge. Inferred present/absent edges can be obtained from the output of a network inference tool. For GRN inference, true present/absent edges can be found in the ground truth/gold standard in the DREAM4 datasets. For RNM inference, true present edges are shown in the regulatory network of miRNAs in Fig. 3; true absent edges refer to those not shown in Fig. 3. Let TP (FP, TN, FN, respectively) denote the number of true positives (false positives, true negatives, false negatives, respectively). P is the sum of TP and FN. N is the sum of TN and FP. We use accuracy (ACC), defined as (TP+TN)/(P+N), to evaluate the performance of a network inference tool.

We compared our approach with two state-of-the-art network inference methods: TimeDelay-ARACNE (abbreviated as TD-ARACNE) [3] and Jump3 [4]. TD-ARACNE takes an information-theoretic approach combined with Markov random fields to perform network inference. Jump3 uses a non-parametric procedure based on random forests to reconstruct the GRN topology. Both of the two methods, like ours, are time-series based in the sense that they take as input time series of expression data and produce as output a regulatory network. Both are well known and widely used for GRN inference.

C. Experimental Results

Each tool was run on every time series in the first dataset to predict a network. Our approach and TD-ARACNE output a reasonable number of edges, all of which were included in the predicted network. These edges constituted the inferred present edges. The other edges that were not in the output constituted inferred absent edges. On the other hand, Jump3 output a large number of edges, each of which was assigned a weight. We sorted and ranked these output edges from top to bottom based on their weights. The top 15% highest ranked edges with the largest weights were then selected as suggested in the literature [10]. These selected edges were inferred present edges, which constituted the network predicted by Jump3. The other edges that were not selected constituted inferred absent edges. Each predicted network was compared with the corresponding gold standard in the DREAM4 challenge to calculate accuracy. Our approach outperforms the two related methods, TD-ARACNE and Jump3, for GRN inference. The accuracy of our approach is 0.86, compared to the accuracy of 0.79 for TD-ARACNE and 0.76 for Jump3 (see Fig. 4). Jump3 performs the worst among the three tools. A careful examination of the results shows that our approach is very conservative in the sense that it generally predicts or infers few present edges. However, in many cases, these inferred present edges are all in the gold standard, suggesting that the inferred edges are very reliable. On the other hand, our approach also produces several false negatives, which are inferred absent edges but in fact are present in the gold standard. These results show that our approach is capable of predicting highly reliable edges in a GRN while missing several true present edges in the gold standard.

We then compare the performance of the three tools for RNM inference. Here, each tool was run on the time series in the second dataset to predict a network. Again, our approach and TD-ARACNE output a reasonable number of edges, all of which were included in the predicted network. Jump3's output contained a large number of weighted edges, and as in the GRN inference, only the top 15% highest ranked edges were selected and included in the predicted network. Each predicted network was compared with the regulatory network of miRNAs shown in Fig. 3 to calculate prediction accuracy. Our approach continues to be better than the two related methods, TD-ARACNE and Jump3, for RNM inference. The accuracy of our approach is 0.93, compared to the accuracy of 0.80 for TD-ARACNE and 0.77 for Jump3 (see Fig. 4). Jump3 again performs the worst among the three tools. As pointed out above, our approach is very conservative; it rarely produces false positives. All present edges inferred or predicted by our approach occur in the network in Fig. 3. On the other hand, TD-ARACNE and Jump3 produce many false positives, hence lowering their accuracy. Notice that neither TD-ARACNE nor Jump3 was intended for RNM inference. Both of these two tools were mainly designed for GRN inference. In contrast, our approach is developed for reverse engineering both GRNs and RNMs.

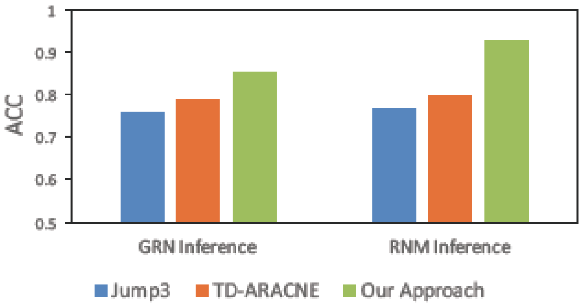


Fig. 4. Performance comparison of three tools for GRN and RNM inference.

IV. CONCLUSIONS

Machine learning in biomedicine has received increasing attention recently. In this paper, we present an approach for reverse engineering gene regulatory networks (GRNs) and regulatory networks of microRNAs (RNMs) by using a dynamic Bayesian network and mutual information scoring function. Our approach exploits the prior knowledge concerning miRNAs and genes when inferring RNMs from time series of expression data. Our experimental results demonstrated the effectiveness of the proposed approach and its superiority over two state-of-the-art network inference methods.

REFERENCES

[1] Y. Abduallah, T. Turki, K. Byron, Z. Du, M. Cervantes-Cervantes, and J. T. L. Wang, "MapReduce algorithms for inferring gene regulatory networks from time-series microarray data using an information-theoretic approach," Biomed Res Int, vol. 2017, p. 6261802, 2017.

[2] C. H. Chien, Y. F. Chiang-Hsieh, Y. A. Chen, C. N. Chow, N. Y. Wu, P. F. Hou, and W. C. Chang, "AtmiRNET: a web-based resource for reconstructing regulatory networks of Arabidopsis microRNAs," Database (Oxford), vol. 2015, p. bav042, 2015.

[3] P. Zoppoli, S. Morganella, and M. Ceccarelli, "TimeDelay-ARACNE: reverse engineering of gene networks from time-course data by an information theoretic approach," BMC Bioinformatics, vol. 11, p. 154, 2010.

[4] V. A. Huynh-Thu and G. Sanguinetti, "Combining tree-based and dynamical systems for the inference of gene regulatory networks," Bioinformatics, vol. 31, pp. 1614-1622, 2015.

[5] M. van der Heijden, M. Velikova, and P. J. Lucas, "Learning Bayesian networks for clinical time series analysis," J Biomed Inform, vol. 48, pp. 94-105, 2014.

[6] N. Friedman, K. P. Murphy, and S. J. Russell, "Learning the structure of dynamic probabilistic networks," in Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998, pp. 139-147.

[7] N. V. Xuan, M. Chetty, R. Coppel, and P. P. Wangikar, "Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network," BMC Bioinformatics, vol. 13, p. 131, 2012.

[8] L. M. de Campos, "A scoring function for learning Bayesian networks based on mutual information and conditional independence tests," Journal of Machine Learning Research, vol. 7, pp. 2149-2187, 2006.

[9] A. Greenfield, A. Madar, H. Ostrer, and R. Bonneau, "DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models," PLoS One, vol. 5, p. e13397, 2010.

[10] R. D. Leclerc, "Survival of the sparsest: robust gene networks are parsimonious," Mol Syst Biol, vol. 4, p. 213, 2008.