Inferring Gene Regulatory Networks from Gene Expression Data by a Dynamic Bayesian Network-Based Model

Lian En Chai, Mohd Saberi Mohamad, Safaai Deris, Chuii Khim Chong, Yee Wen Choon, Zuwairie Ibrahim, and Sigeru Omatu

Abstract. Enabled by recent advances in bioinformatics, the inference of gene regulatory networks (GRNs) from gene expression data has garnered much interest from researchers. This is due to the need of researchers to understand the dynamic behavior and uncover the vast information lay hidden within the networks. In this regard, dynamic Bayesian network (DBN) is extensively used to infer GRNs due to its ability to handle time-series microarray data and modeling feedback loops. However, the efficiency of DBN in inferring GRNs is often hampered by missing values in expression data, and excessive computation time due to the large search space whereby DBN treats all genes as potential regulators for a target gene. In this paper, we proposed a DBN-based model with missing values imputation to improve inference efficiency, and potential regulators detection which aims to lessen computation time by limiting potential regulators based on expression changes. The performance of the proposed model is assessed by using time-series expression data of yeast cell cycle. The experimental results

Lian En Chai · Mohd Saberi Mohamad · Safaai Deris · Chuii Khim Chong · Yee Wen Choon
Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
e-mail: lechai2@live.utm.my, saberi@utm.my, safaai@utm.my, ckchong2@live.utm.my, ywchoon2@live.utm.my

Zuwairie Ibrahim
Department of Mechatronics and Robotics, Center for Artificial Intelligence and Robotics, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
e-mail: zuwairie@fke.utm.my

Sigeru Omatu
Department of Electronics, Information and Communication Engineering, Osaka Institute of Technology, Osaka 535-8585, Japan
e-mail: omatu@rsh.oit.ac.jp

S. Omatu et al. (Eds.): Distributed Computing and Artificial Intelligence, AISC 151, pp. 379–386.
springerlink.com

showed reduced computation time and improved efficiency in detecting gene-gene relationships.

Keywords: Dynamic Bayesian Network, Gene Regulatory Networks, Gene Expression Data, Inference.

1 Introduction

The development of microarray technology has enabled researchers to facilitate new experimental methods for understanding gene expression and regulations. The output, usually referred as gene expression data or microarray data, contains vast information such as the behaviors revealed by the system under normal conditions; abnormalities of the system if certain parts cease to function; the robustness of the system under extreme conditions [1], hence providing a holistic viewpoint of gene expression to the researchers instead of only a few genes as in the classical experiments.

Motivated by the need of researchers to understand the complex phenomena of gene regulations, gene expression data have obtained significant importance in the inferring of GRNs to explain the phenotypic behaviors of a specific system. The traditional trial and error method of inferring GRNs from gene expression data is obviously not feasible in handling large-scale data due to the time-consuming nature of repeating the routine to achieve accurate results [2]. To analyze and utilize the massive amount of gene expression data, researchers have already developed numerous computational methods to automate the inferring procedure [2, 3]. In particular, Bayesian network (BN), which models conditional dependencies of a set of variables via probabilistic measure, was widely utilized by researchers in inferring GRNs from gene expression data.

BN's effectiveness in inferring GRNs is mainly due to its ability to work on locally interacting components with a relatively small number of variables; able to assimilate other mathematical models to avoid the overfitting of data; allows the combination of prior knowledge the strengthen the causal relationship. Despite the advantages stated above, BN has two critical limitations in which it does not allow feedback loops and is unable to handle the temporal aspect of time-series microarray data.

In view of the fact that feedback loops represent the importance of homeostasis in living organisms, researchers have developed the dynamic Bayesian network (DBN) as a promising substitute. Since the pioneering work of Murphy and Mian [4], DBN has attracted particular attention from numerous researchers [5, 6, 7, 8, 9]. Nevertheless, normal DBN usually assumes all genes as potential regulators against target genes, and consequently causes the excessive computational cost which inhibits the efficiency of DBN on large scale gene expression data [8, 9]. In addition, the missing values commonly found in expression data may influence up to 90% of the genes [10], thus affecting the inference results. To tackle the two

problems, we proposed a model of DBN with missing values imputation to improve the inference efficiency, and potential regulators selection which reduce computation time by limiting the numbers of potential regulators for each target gene. The details of our model are discussed in the following section.

2 Methods

In this section, we describe the details of the proposed DBN-based model for inferring GRNs from gene expression data. In essence, the proposed model consists of three main steps: missing values imputation, potential regulators selection and dynamic Bayesian network. The following sub-sections (2.1 – 2.3) discuss in detail for each of the three main steps. Table 2.1 shows the overview of our proposed model and existing DBN-based models.

Table 2.1 Overview of our proposed model and existing DBN-based models for inferring gene regulatory networks from gene expression data.

Our proposed model	Previous work [8, 9]	Previous work [15]
Missing values imputation	Potential regulators selection	Dynamic Bayesian network
↓	↓	
Potential regulators selection	Dynamic Bayesian network	
↓		
Dynamic Bayesian network		

2.1 Experimental Data and Missing Values Imputation

The experimental study is based on the *S. cerevisiae* cell cycle time-series gene expression data from Spellman *et al.* [11] This dataset contains two short time series (cln3, clb2; both 2 time points) and four medium time series (alpha, cdc15, cdc28 and elu; 18, 24, 17 and 14 time points). However, the dataset contains missing values which must be processed. Conventional methods of treating missing values include repeating the microarray experiment which is not economical feasible, or simply replacing the missing values by zero or row average. A better solution is to use imputation algorithms to estimate the missing values by exploiting the observed data structure and expression pattern. In view of this, we applied the Bayesian principle component analysis (BPCA) imputation

algorithm [12] due to its ability to assumes a global covariance structure of the dataset by iteratively estimating the posterior distribution of the missing values until convergence is achieved, and its effectiveness on large-scale data (9 minutes and 13 seconds for the experimental data on a Core i3 PC).

2.2 Potential Regulators Selection

The work of Yu *et al.* [13] showed that most transcriptional factors (TFs) experience changes in expression level before or simultaneously with their target genes. Following this fact, it is possible to derive an effective algorithm to reduce the search space by limiting the potential regulators of each target genes. Firstly, we determined the cutoff threshold for up-regulation (≥ 1.4) and down-regulation (< -1.1) based on the distribution of the gene expression values. Then, we discretized the dataset into three classes (up-, down-regulation and normal) and search only for the data located in the upper and lower bound classes. A time gap of two time points width is created to slide through the data to group regulation pairs. Thus, each target gene contains a subset of potential regulators which exhibit earlier or simultaneously expression changes. These are used as the input for the subsequent network inference step using DBN.

2.3 A Dynamic Bayesian Network

The network inference step is done by applying DBN, which is actually an extension of BN to describe the stochastic evolution of a network against time. This is mainly because BN is limited to steady-state data (static data), and DBN readily handles time-series data to identify the causal relationships among a set of variables. It also enables the modeling of cyclic structure while inheriting the advantages of BN. In essence, in modeling time-series data, values of a set of random variables are observed at different points in time. Assuming each time point as single variable Y_i, the simplest causal model for asequence of data $\{Y_1,\ldots,Y_t\}$ would be a *first-order Markov chain*, in which the state of the next variable is dependent on the previous variable only. By applying the chain rule of probabilities and conditional independencies based on Bayes theorem, the joint probability distribution (JPD) of the graph has the general form of $P(Y_1, Y_2, \ldots, Y_t) = P(Y_1)P(Y_2|Y_1) \ldots P(Y_t|Y_{t-1})$. DBN consists of two stages: the parameter learning stage followed by the structure learning stage. In the parameter learning stage, we used the results from the previous step to create the data matrices of all target genes with their subsets of potential regulators. We then updated the data matrices by calculating the conditional probabilities of each target gene against each of its respective potential regulators. In light of recent work [14] showed that learning DBN structure is not definitely NP-hard, we employed a globally optimal search strategy [15] instead of using local search strategy in the structure learning stage. Figure 2.1 illustrates the overview of our proposed DBN-based model.

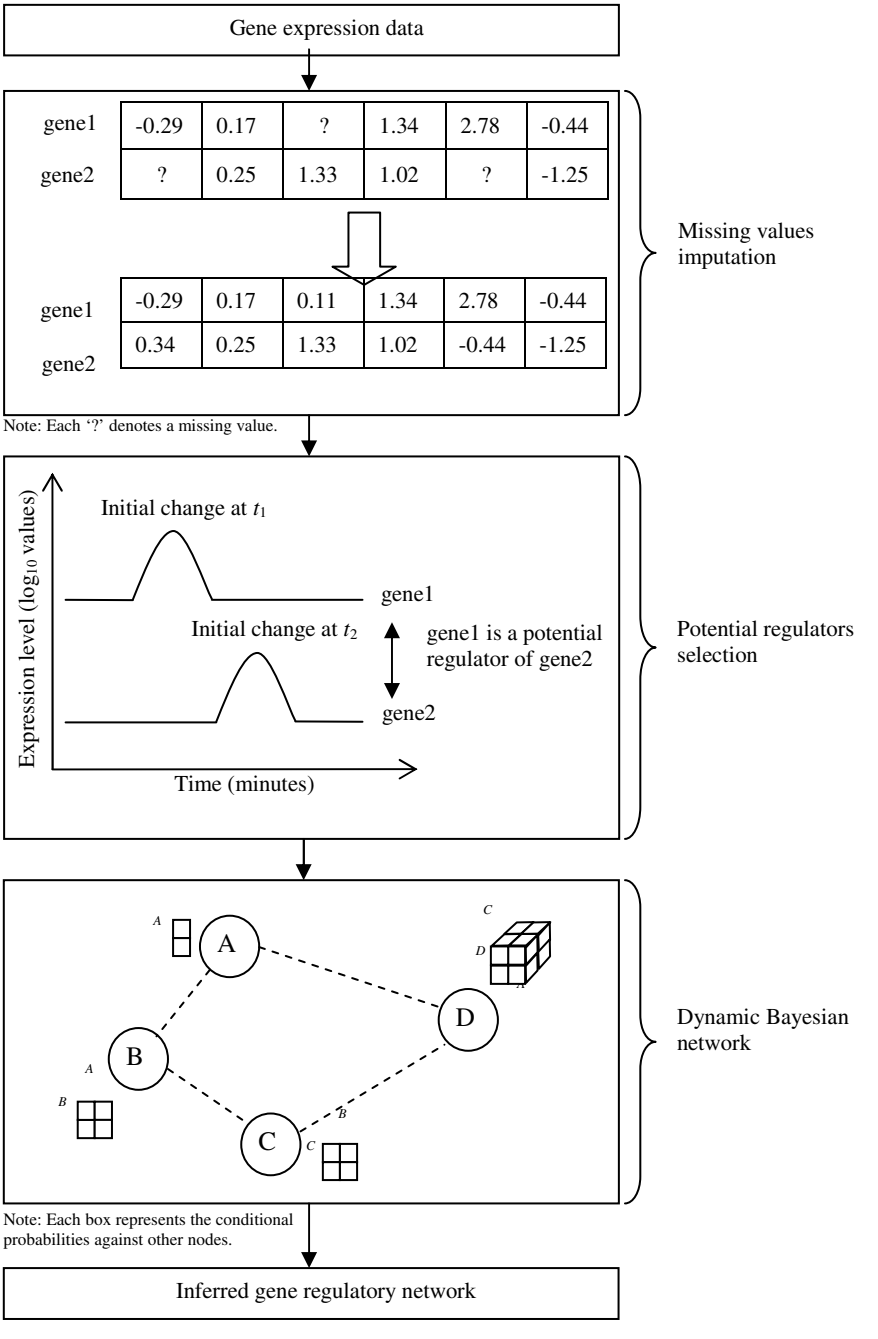

Fig. 2.1 Overview of our proposed DBN-based model with missing values imputation and potential regulators selection.

3 Results

In this study, we compared the efficiency and computation time of our DBN-based model against normal DBN [15]. The experiment results are evaluated by comparing with the yeast cell cycle pathway compiled at KEGG (Figure 3.1) and summarized in Table 3.1. In Table 3.1, row 1 represents the network inferred by our proposed model and row 2 represents the network predicted by normal DBN. Our proposed model used 45 minutes and 9 seconds against normal DBN which in turn used 1 hour 38 minutes and 23 seconds on a Core i3 PC with 4GB main memory. This is due to the reduced search space by applying potential regulators selection prior to DBN learning. Each target gene has a limited number of potential regulators instead of assuming all genes as potential regulators. Our proposed model was able to identify 14 gene-gene relationships against normal DBN which identified 12 gene-gene relationships. The missing values imputation helped to improve the efficiency of our proposed model by making use of the data structure and pattern to impute missing values. Interestingly, our proposed model also incorrectly identified more relationships (6 against 3). Putting aside that, the results of this study proved that the performance of DBN in inferring GRNs can be improved by imputing missing values and potential regulators selection.

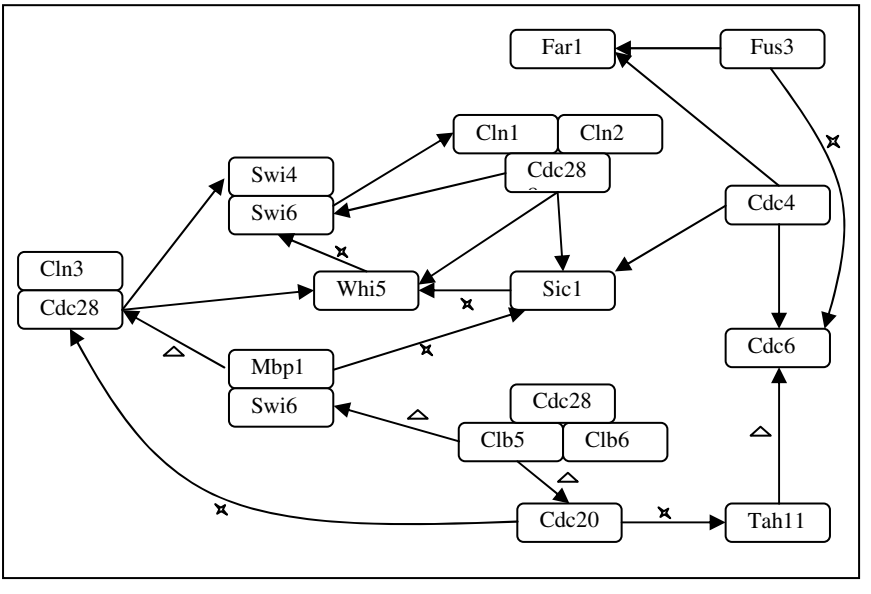

Fig. 3.1 Predicted cell cycle pathway for *S. cerevisiae* dataset using our proposed DBN-based model. A cross represents an incorrect inference; a triangle represents a misdirected relationship; an edge without any attachment is a correct inference. Genes that are grouped closely together represent a complex.

Table 3.1 The results of experiment study

Inference model	Correctly identified relationships	Misdirected relationships	Incorrectly identified relationships	Computation time (HH:MM:SS)
DBN_prs	14	4	6	00:45:09
DBN_norm [15]	12	8	3	01:38:23

Note: Shaded row represents the network inferred by our proposed model (DBN_prs) and unshaded row represents the network predicted by normal DBN (DBN_norm). Relationships refer to the gene-gene relationships.

4 Conclusion and Future Work

As a conclusion, our proposed DBN-based model is showed to perform better than normal DBN in terms of computation time and efficiency. However, it should be noted that our proposed model could only deals with inter-time slice edges. To learn DBN with both inter- and intra-time slice edges remains an interesting point of research. It is suggested by Vinh *et al.* [16] to learn intra-time slice edges separately before combining with the inter-time slice edges and post-processing as an alternative to describe gene-gene interactions. Additionally, we are also interested in taking account of the transcriptional time lag which is commonly found in GRNs. As Zou and Conzen [8] pointed out, the lack of an algorithm to handle transcriptional time lag is one of the main factors that contribute to the relatively low accuracy of inferring GRNs using DBN. Researchers have implemented time lag mechanism in the potential regulators selection algorithm [8, 9]. Lastly, despite the extensive usage of DBN on gene expression data to infer GRNs, it is by no means to replace gene intervention experiments completely. The resultant networks should be treated as a guideline or framework of the studied biological pathways for future hypotheses testing and intervention experiments.

Acknowledgments. This work is financed by Institutional Scholarship MyPhd provided by the Ministry of Higher Education of Malaysia. We also would like to thank Universiti Teknologi Malaysia for supporting this research by the UTM GUP research grants (Vot number: QJ130000.7107.01H29 and QJ130000.7123.00H67).

References

[1] Karlebach, G., Shamir, R.: Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology 9(10), 770–780 (2008), doi:10.1038/nrm2503

[2] Lee, W.P., Tzou, W.S.: Computational methods for discovering gene networks from expression data. Briefing in Bioinformatics 10(4), 408–423 (2009), doi:10.1093/bib/bbp028

[3] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernado, D.: How to infer gene networks from expression profiles. Molecular Systems Biology 3, 78 (2007), doi:10.1038/msb4100120

[4] Murphy, K., Mian, S.: Modelling gene expression data using dynamic Bayesian networks. Technical Report. Computer Science Division, University of California, Berkeley (1999)

[5] Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., d'Alche-Buc, F.: Gene networks inference using dynamic Bayesian networks. Bioinformatics 19(suppl.2), 138–148 (2003), doi:10.1093/bioinformatics

[6] Kim, S.Y., Imoto, S., Miyano, S.: Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefing in Bioinformatics 4(3), 228–235 (2003), doi:10.1093/bib/4.3.228

[7] Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20(18), 3594–3603 (2004), doi:10.1093/bioinformatics/bth448

[8] Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1), 71–79 (2005), doi:10.1093/bioinformatics/bth463

[9] Jia, Y., Huan, J.: Constructing non-stationary dynamic Bayesian networks with a flexible lag choosing mechanism. BMC Bioinformatics (11) S27(11) (2010), doi:10.1186/1471-2105-11-S6-S27

[10] Ouyang, M., Welsh, W.J., Geogopoulos, P.: Gaussian mixture clustering and imputation of microarray data. Bioinformatics 20, 917–923 (2004), doi:10.1093/bioinformatics/bth007

[11] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology Cell 9, 3273–3297 (1998)

[12] Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Ishii, S.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19(16), 2088–2096 (2003), doi:10.1093/bioinformatics/btg287

[13] Yu, H., Luscombe, N.M., Qian, J., Gerstein, M.: Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends in Genetics 19(8), 422–427 (2003), doi:10.1016/S0168-9525(03)00175-6

[14] Dojer, N.: Learning Bayesian Networks Does Not Have to Be NP-Hard. In: Proceedings of International Symposium on Mathematical Foundations of Computer Science, pp. 305–314 (2006), doi:10.1007/11821069_27

[15] Wilczynski, B., Dojer, N.: BNFinder: exact and efficient method for learning Bayesian networks. Bioinformatics 25(2), 286–287 (2009), doi:10.1093/bioinformatics/btn505

[16] Vinh, N.X., Chetty, M., Coppel, R., Wangikar, P.P.: GlobalMIT: Learning Globally Optimal Dynamic Bayesian Network with the Mutual Information Test (MIT) Criterion. Bioinformatics (2011), doi:10.1093/bioinformatics/btr457