Using A Dynamic Bayesian Network-based Model for Inference of *Escherichia coli* SOS Response Pathway from Gene Expression Data

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

Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.

Corresponding author's e-mail: saberi@utm.my

Abstract

Largely due to the technological advances in bioinformatics, researchers are now garnering interests in inferring gene regulatory networks (GRNs) from gene expression data which is otherwise unfeasible in the past. This is because of the need of researchers to uncover the potentially vast information and understand the dynamic behavior of the GRNs. In this regard, dynamic Bayesian network (DBN) has been broadly utilized for the inference of GRNs thanks to its ability to handle time-series microarray data and modeling feedback loops. Unfortunately, the commonly found missing values in gene expression data, and excessive computation time owing to the large search space whereby all genes are treated as potential regulators for a target gene, often impede the effectiveness of DBN in inferring GRNs. This paper proposes a DBN-based model with missing values imputation to improve inference efficiency, and potential regulators selection which intends to decrease computation time by selecting potential regulators based on expression changes. We tested our proposed model on the Escherichia coli SOS response pathway which is responsible for repairing damaged DNA of the bacterium. The experimental results showed reduced computation time and improved efficiency in detecting gene-gene relationships.

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

Introduction

The advance of DNA microarray technology has allowed researchers to design new experimental methods for understanding gene expression and regulations. The output, known as gene expression data or microarray data, contains immense information such as the behaviours 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], thus providing a holistic viewpoint of gene expression to the researchers instead of only a few genes as in the classical experiments.

Motivated by the yearning of researchers to understand the complex phenomena of gene regulations, gene expression data have become very important in the inferring of gene regulatory networks (GRNs) to elucidate the phenotypic behaviours of a particular system. The conventional trial and error method of inferring GRNs from gene expression data is clearly not suitable in handling large-scale data due to the time-consuming nature of repetitive routines as to achieve precise results [2]. To analyse and utilize the immense amount of gene expression data, researchers have already developed many computational methods to automate the inferring process [2, 3]. Specifically, Bayesian network (BN), which models conditional dependencies of a set of variables via probabilistic measure, was extensively utilized by researchers in inferring GRNs from gene expression data.

BN's usefulness in inferring GRNs is primarily due to its ability to handle locally interacting components with a comparatively small number of variables; able to assimilate other mathematical models to avoid the overfitting of data; permits the combination of prior knowledge to reinforce the causal relationship. In spite of the advantages above, BN has two critical limitations in which it does not permit feedback loops and is unable to handle the temporal aspect of time-series microarray data.
Due to the fact that feedback loops represent the importance of homeostasis in living organisms, researchers have developed the dynamic Bayesian network (DBN) as a promising alternate. Ever since the pioneering work of Murphy and Mian [4], DBN has attracted attention from many researchers [5, 6, 7, 8, 9]. Nonetheless, normal DBN typically presumes all genes as potential regulators against target genes, and therefore causes the excessive computational cost which restrains

the efficiency of DBN on large scale gene expression data [8, 9]. Additionally, the missing values commonly found in expression data may influence up to 90% of the genes [10], consequently affecting the inference results. To address the problems above, we proposed a DBN-based model with missing values imputation to improve the inference efficiency, and potential regulators selection which decrease computation time by restricting the numbers of potential regulators for each target gene. The details of our proposed model are discussed in the subsequent section.

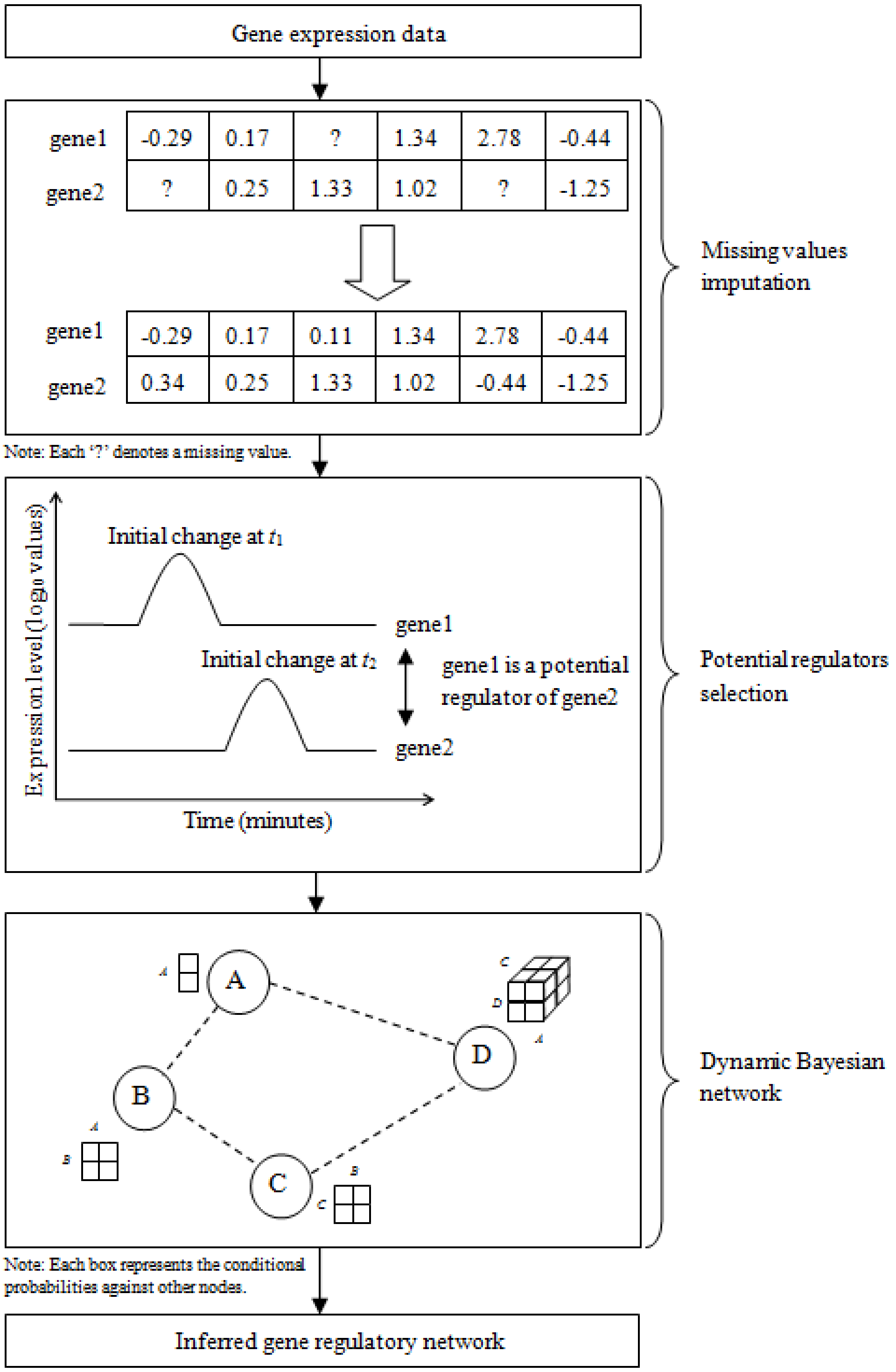
Methods

The proposed model primarily consists of three main steps: missing values imputation, potential regulators selection and dynamic Bayesian network (DBN). Fig. 1 illustrates the overview of our proposed DBN-based model. Table 1 shows the overview of our proposed model and existing DBN-based models.

This experimental study is based on the *E. coli* SOS response pathway gene expression data from Ronen *et al.* [11] This gene network is an error-prone repair system which responses to damaged DNA by arresting cell cycle and inducing DNA repair. Under normal circumstances, the repressor protein, LexA, negatively regulates the SOS genes by binding to the promoter region of these genes. When DNA damage occurs (The accumulation of single-stranded DNA – ssDNA, due to blockage of DNA polymerase), the RecA protein, which acts as a sensor of DNA damage, is activated by binding to these ssDNA. The activated RecA then facilitates the self-cleavage of LexA repressor. The drop in LexA level in turn causes the SOS genes to be de-repressed. This continues until the damage is repaired, whereby the level of activated RecA drops, LexA accumulates and represses the SOS genes again. The dataset contains 8 genes observed at evenly spaced 50 instants with 6 minutes intervals. However, it also contains missing values which must be first processed. Traditional methods of treating missing values include reiterating the microarray experiment which is not economical feasible, or simply substitute the missing values by zero or row average. A better solution is to use imputation algorithms to estimate the missing values by utilising the observed data structure and expression pattern. In view of this, we employed the Bayesian principle component analysis (BPCA) imputation algorithm [12] mainly because of 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 efficiency on large-scale data (1 minutes and 23 seconds on the experimental data with a Core i3 PC).

Yu *et al.* [13] proved that in most cases, transcriptional factors (TFs) experience changes in expression level prior to or concurrently with their target genes. With this in mind, it is possible to devise an algorithm to reduce the search space by limiting the potential regulators of each target genes. Firstly, we determined the cutoff threshold for up-regulation and down-regulation based on the distribution of the gene expression values. After that, we categorised the dataset into three classes (up-, down-regulation and normal) and look for 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 includes a subset of potential regulators which exhibit prior or concurrent expression changes. These are used as the input for the subsequent network inference step using DBN.

DBN, which is derived from BN to describe the stochastic nature of a network against time, is used to infer the network based on the input obtained from the previous step. While BN is restricted to only steady-state data (static data), DBN readily handles time-series data to identify the causal relationships among a set of variables. It also enables the modelling of cyclic network structure while inheriting the advantages of BN. Basically, in modelling network from time-series data, values of a set of random variables are observed at different points in time. Assuming each time point as a 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 only dependent on the previous variable. By applying the chain rule of probabilities and conditional independencies based on Bayes theorem, the joint probability distribution (JPD) of the network 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 created the data matrices of all target genes with their subsets of potential regulators based on the output from previous step. We then updated the data matrices by calculating the conditional probabilities of each target gene against its respective potential regulators. As DBN structure learning is not certainly NP-hard [14], we applied a globally optimal search strategy [15] instead of local search strategy in the structure learning stage.


Gene expression data
Missing values imputation
Initial change at t_1
gene1
Initial change at t_2
gene1 is a potential regulator of gene2
gene2
Expression level (log10 values)
Time (minutes)
Potential regulators selection
A
B
C
D
Dynamic Bayesian network
Inferred gene regulatory network

gene1	-0.29	0.17	?	1.34	2.78	-0.44
gene2	?	0.25	1.33	1.02	?	-1.25

gene1	-0.29	0.17	0.11	1.34	2.78	-0.44
gene2	0.34	0.25	1.33	1.02	-0.44	-1.25

Note: Each '?' denotes a missing value.

Note: Each box represents the conditional probabilities against other nodes.

Fig. 1. Overview of our proposed DBN-based model with missing values imputation and potential regulators selection.

Table 1. Overview of our proposed model and existing DBN-based models for inferring GRNs 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		

Results and Discussions

In this study, we compared the efficiency and computation time of our DBN-based model against normal DBN [15]. The experiment results are evaluated based on the work of Radman [16] and summarised in Table 2. In Table 2, the first row indicates the network predicted by our proposed model and row 2 indicates the network inferred by normal DBN. Our proposed model used 8 minutes and 43 seconds against normal DBN which in turn used 23 minutes and 17 seconds on a Core i3 PC with 4GB main memory. The potential regulators selection before DBN learning helped to reduce the search space by limiting each target gene's number of potential regulators instead of assuming all genes as potential regulators. Our proposed model was able to correctly identify 8 gene-gene relationships in the *E. coli* SOS response pathway (lexA–recA, lexA–polB, lexA–umuD, lexA–uvrY, lexA–uvrA, lexA–ruvA, lexA–lexA, recA–recA) (See Fig. 2) against normal DBN's 4 correctly identified gene-gene relationships. The inferred network showed cyclic regulatory edge of recA, which corresponds to its ability to self-activated when DNA damage is sensed. On the other hand, lexA's cyclic regulatory edge indicates its self-cleavage mechanism when the level of activated recA is raised. Additionally, the proposed model showed relationships between umuD–polB, uvrD–uvrY and ruvA–uvrA. Although we considered them as incorrectly identified relationships, based on data pattern exploited by the model we suggest that there might be regulatory relationships between these genes that could be further investigated on. Nevertheless, 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.

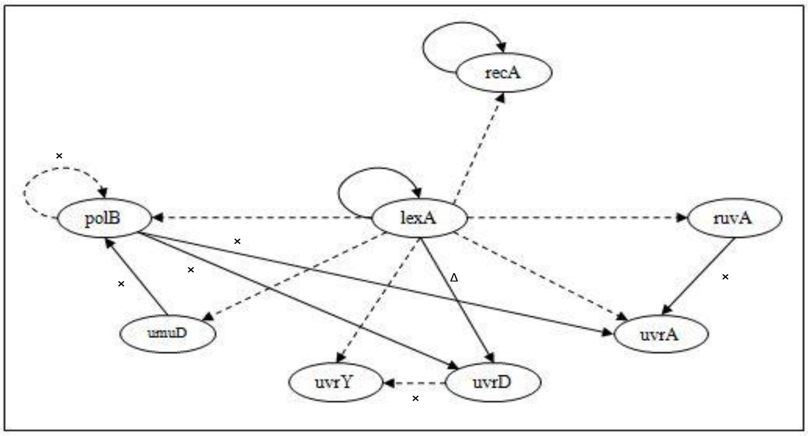


Fig. 2. Inferred SOS response pathway for *E. coli* dataset using our proposed DBN-based model. Dash edges (- - -) represent down-regulations and normal edges (—) represent up-regulations. A cross represents an incorrect inference; a triangle represents a misdirected relationship; an edge without any attachment is a correct inference.

Table 2. The results of experiment study

Inference model	Correctly identified relationships	Misdirected relationships	Incorrectly identified relationships	Computation time (HH:MM:SS)
DBN_prs	8	1	6	00:08:43
DBN_norm [15]	4	2	4	00:23:17

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.

Summary and Future Work

We proposed a DBN-based model with missing values imputation and potential regulators selection to infer GRNs from gene expression data. Based on the dataset of *E. coli* SOS response pathway, our proposed model showed promising results in terms of computation time and efficiency when compared to normal DBN. However, the *E. coli* SOS response pathway dataset was not adequately large to fully examine the potential power of the proposed model. Larger datasets such as *S. cerevisiae* cell cycle pathway could be used for further work. Additionally, we are also interested in taking account of the transcriptional time lag which is commonly ignored in inferring GRNs from gene expression data. As Zou and Conzen [8] pointed out, the lack of an algorithm to handle transcriptional time lag is one of the main factors that contributed to the relatively low accuracy of inferring GRNs using DBN. Researchers have tried to implement time lag mechanism into the potential regulators selection algorithm [8, 9]. Also, it should be noted that presently, our proposed model could only handle 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.* [17] 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. Lastly, in spite of the broad practice of using DBN to infer GRNs from gene expression data, it is in no way to completely substitute gene intervention experiments. 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] G. Karlebach and R. Shamir: *Modelling and analysis of gene regulatory networks* Nature Reviews Molecular Cell Biology Vol. 9(10) (2008), p. 770-780.

[2] W.P. Lee and W.S. Tzou: *Computational methods for discovering gene networks from expression data* Briefing in Bioinformatics Vol. 10(4) (2009), p. 408-423.

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

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

[5] B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, F. d'Alche-Buc: (2003) *Gene networks inference using dynamic Bayesian networks* Bioinformatics Vol. 19(Suppl. 2) (2003), p. ii138-ii148.

[6] S.Y. Kim, S. Imoto, S. Miyano: *Inferring gene networks from time series microarray data using dynamic Bayesian networks* Briefing in Bioinformatics Vol. 4(3) (2003), p. 228-235.

[7] J. Yu, V.A. Smith, P.P. Wang, A.J. Hartemink, E.D. Jarvis: *Advances to Bayesian network inference for generating causal networks from observational biological data* Bioinformatics Vol. 20(18) (2004), p. 3594-3603.

[8] M. Zou and S.D. Conzen: *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data* Bioinformatics Vol. 21(1) (2005), p. 71-79.

[9] Y. Jia and J. Huan: *Constructing non-stationary dynamic Bayesian networks with a flexible lag choosing mechanism* BMC Bioinformatics Vol. 2010(11) (2010), p. S27.

[10] M. Ouyang , W.J. Welsh, P. Geogopoulos: *Gaussian mixture clustering and imputation of microarray data* Bioinformatics Vol. 20 (2004), p. 917-923.

[11] M. Ronen, R. Rosenberg, B.I. Shraiman, U. Alon: *Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinectics* Proceedings of National Academy of Sciences Vol. 99 (2002), p. 10555-10560.

[12] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, S. Ishii: *A Bayesian missing value estimation method for gene expression profile data* Bioinformatics Vol. 19(16) (2003), p. 2088-2096.

[13] H. Yu, N.M. Luscombe, J. Qian, M. Gerstein: *Genomic analysis of gene expression relationships in transcriptional regulatory network* Trends in Genetics Vol. 19(8) (2003), p. 422-427.

[14] N. Dojer: *Learning Bayesian Networks Does Not Have to Be NP-Hard* Proceedings of International Symposium on Mathematical Foundations of Computer Science (2006), p. 305-314.

[15] B. Wilczynski and N. Dojer: *BNFinder: exact and efficient method for learning Bayesian networks* Bioinformatics Vol. 25(2) (2009), p. 286-287.

[16] M. Radman: *Phenomenology of an inducible mutagenic DNA repair pathway in Escherichia coli* Basic Life Sciences Vol. 5A (1975), p. 355-367.

[17] N.X. Vinh, M. Chetty, R. Coppel, P.P. Wangikar: *GlobalMIT: Learning Globally Optimal Dynamic Bayesian Network with the Mutual Information Test (MIT) Criterion*. Bioinformatics (2011), in press.