Modelling Gene Networks by a Dynamic Bayesian Network-based Model with Time Lag Estimation

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

lechai2@live.utm.my, {saberi, safaai}@utm.my, {ckchong2, ywchoon2}@live.utm.my

*Corresponding author

Abstract. Due to the needs to discover the immense information and understand the underlying mechanism of gene regulations, modelling gene regulatory networks (GRNs) from gene expression data has attracted the interests of numerous researchers. To this end, the dynamic Bayesian network (DBN) has emerged as a popular method in GRNs modelling as it is able to model time-series gene expression data and feedback loops. Nevertheless, the commonly found missing values in gene expression data, the inability to take account of the transcriptional time lag, and the redundant computation time caused by the large search space, frequently inhibits the effectiveness of DBN in modelling GRNs from gene expression data. This paper proposes a DBN-based model (IST-DBN) with missing values imputation, potential regulators selection, and time lag estimation to tackle the aforementioned problems. To evaluate the performance of IST-DBN, we applied the model on the *S. cerevisiae* cell cycle time-series expression data. The experimental results revealed IST-DBN has decreased computation time and better accuracy in identifying gene-gene relationships when compared with existing DBN-based model and conventional DBN. Furthermore, we expect the resultant networks from IST-DBN to be applied as a general framework for potential gene intervention research.

Keywords: Dynamic Bayesian network, missing values imputation, time-series gene expression data, gene regulatory networks, network inference.

1 Introduction

In recent years, the advent of DNA microarray technology has permitted researchers to develop novel experimental approaches for probing into the complicated system of gene regulation. The ensuing output, known as gene expression data, brings forth valuable information such as the robustness, behaviours or anomalies demonstrated by the cellular system under different circumstances [1].

Over the years, different computational approaches have been established to automate GRNs modelling. Particularly, Bayesian network (BN), which relies on probabilistic measure to recognize causal interactions between a set of variables, was widely used to model GRNs. BN has several advantages: ability to work on local components, prevent data overfitting by assimilation of other mathematical models, and capable of merging prior knowledge to reinforce the causal links. However, BN also has two limitations: it cannot model feedback loops and handle time-series gene expression data.

In biological perspective, feedback loops signify the homeostasis process in living organisms. Therefore, the dynamic Bayesian network (DBN) has been introduced as an alternative to counter BN's drawbacks. However, missing values distributed across gene expression data might influence up to 90% of the genes and consequently affecting downstream analysis and modelling methods [2]. Also, in predicting gene-gene relationships, conventional DBN normally includes all genes into the subsets of potential regulators and their target genes, and in turn triggers the large search space and the redundant computational cost which obstructs the usefulness of DBN modelling [3]. In light of these problems, Chai *et al.* [4] proposed a DBN-based model (ISDBN) with missing values imputation and potential regulators selection.

The disadvantage of ISDBN and conventional DBN is that they do not have the capability to tackle transcriptional time lag effectively, whereby the target genes are provided a time delay by their regulators before their expressions in the system. This drawback hinders the accuracy of DBN-based methods in modelling GRNs. To address this problem, we proposed to further enhance the aforementioned model with time lag estimation (IST-DBN) which uses the time difference between the initial changes of expression level of potential regulators against their target genes as an appropriate transcriptional time lag.

2 Methods

In essence, IST-DBN consists of four steps: missing values imputation, potential regulators selection, time lag estimation and DBN modelling. Fig. 1 illustrates the schematic overview of IST-DBN.

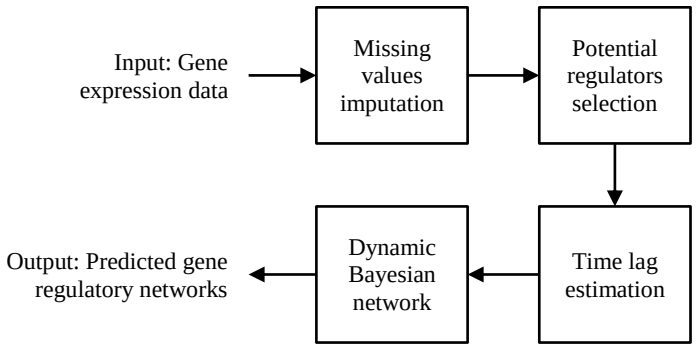

Fig. 1. Schematic overview of IST-DBN.

2.1 Missing Values Imputation

Missing values in gene expression data happen for various factors. Tiny impurities would corrupt the microarray slides at a number of spots as they are very small and crammed. After scanning and digitalising the array, the problematic spots are labelled as missing. Numerous imputation algorithms have been developed to handle missing values by exploiting the underlying expression data structure and pattern. Particularly, LLSimpute extracts information from local similarity structures by creating a linear combination of similar genes and target genes with missing values via a similarity measure [5]. This algorithm consists of two steps. Firstly, k genes are selected by the L_2-norm, where k is a positive integer that defines the number of coherent genes to the target gene. For instance, to impute a missing value g located at $\boldsymbol{x}_{11}$ in a $m \times n$ matrix $\boldsymbol{X}$, the k-nearest neighbour gene vectors for $\boldsymbol{x}_1$,

$$\boldsymbol{v}_{S_i}^{\mathrm{T}} \in \boldsymbol{X}^{1\times n} \quad 1 \leq i \leq k \tag{1}$$

are first computed, whereby the gene expression data is described as a $m \times n$ matrix $\boldsymbol{X}$ (m is the number of genes, n is the number of observations), and $\boldsymbol{x}_1$ represents the row of the first gene with n observations. s_i is a list of k-nearest neighbour genes vectors, which in turn corresponds to the i-th row of the transpose vector $\boldsymbol{v}^{\mathrm{T}}$. The second step involves regression and estimation of the missing values. A matrix, $\boldsymbol{A} \in \boldsymbol{X}^{k\times(n-1)}$ whereby the k rows of the matrix contains vector $\boldsymbol{v}$, and two vectors, $\boldsymbol{b} \in \boldsymbol{X}^{k\times 1}$ and $\boldsymbol{w} \in \boldsymbol{X}^{(n-1)\times 1}$, are subsequently formed. The vector $\boldsymbol{b}$ contains the first element of k vectors $\boldsymbol{v}^{\mathrm{T}}$, while vector $\boldsymbol{w}$ contains $n - 1$ elements of vector $\boldsymbol{x}_1$. A k-dimensional coefficient vector $\boldsymbol{y}$ is then computed such that the least square problem is minimised as

$$\min_{\boldsymbol{y}} |\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{w}|^2 \tag{2}$$

Let $\boldsymbol{y}^*$ to denote the vector whereby the square is minimised such that

$$\boldsymbol{w} \simeq \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}^* = \boldsymbol{y}_1^*\boldsymbol{a}_1 + \boldsymbol{y}_2^*\boldsymbol{a}_2 + \cdots + \boldsymbol{y}_k^*\boldsymbol{a}_k \tag{3}$$

where $\boldsymbol{a}_i \in \boldsymbol{A}^{k\times 1}$, and therefore, the missing value g can be imputed as a linear combination of coherent genes such that

$$g = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}})'\boldsymbol{w} \tag{4}$$

where $(\boldsymbol{A}^{\mathrm{T}})'$ exists as the pseudoinverse of $\boldsymbol{A}^{\mathrm{T}}$ [5].

2.2 Potential Regulators Selection

In most cases, the expression level of transcriptional factors (TFs) would fluctuate prior to or simultaneously with their target genes [6]. Based on this characteristic, we conceived an algorithm which would decrease the search space by restricting the number of potential regulators for each target genes. The first step is to determine a

threshold, either experimentally or fixed as the average expression level of the genes. In this paper, the threshold for up-regulation and down-regulation are determined based on the baseline cut-off of the gene expression values. As such, for the *S. cerevisiae* dataset used in this paper, the threshold is decided as ≥1.2 for up-regulation and ≤0.7 for down-regulation. The gene expression values are then classified into three states: up-, down- and normal regulation. The three states denote whether the expression value is greater than, lower than or similar to the threshold. After that, the exact time units of initial up-regulation and down-regulation of each gene are determined, and genes with prior changes in expression level are included into the subset of potential regulators against genes with later expression changes. As genes with late expression changes might comprise a large number of potential regulators, we have limited the maximum time gap for prior expression changes to five time units to prevent choosing potential regulators for a target gene from the whole gene expression dataset. To further illustrate this idea, let us first assume two hypothetical genes: gene 1 and gene 2. Gene 1 experienced an initial expression change at time t_1 before gene 2's initial expression change at time t_2, and so gene 1 is included into the subset of potential regulators for gene 2 (See Fig. 2). The same process applies to other up- or down-regulated genes which fulfil the criteria.

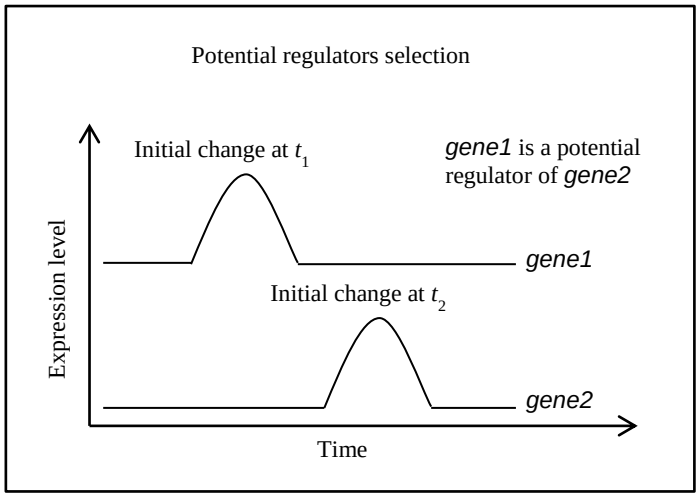

Fig. 2. Schematic overview of potential regulators selection.

2.3 Time Lag Estimation

Transcriptional time lag is defined as the time delay between the expression of regulators and the expression of their target genes to protein products. Using the two hypothetical genes, 1 and 2, and that gene 1 regulates gene 2. Gene 1 initiates expression fluctuation at time t_x and gene 2 has an expression change at t_y. The time difference between t_x and t_y is considered as the transcriptional time lag. In DBN modelling of GRNs, potential regulators are paired up with target genes based on the statistical analysis of their causal strength between time units. However, even though

the actual transcriptional time lag might be several time units, DBN typically aligns regulator-gene pairs by only one time unit. IST-DBN takes into account of the actual transcriptional time lag by pairing up target genes and their potential regulators based on the time difference between their pre-determined initial changes in expression level. For a target gene, potential regulators are divided into separate groups based on the time delay (e.g. one or two time units), mainly due to the fact that a target gene may have several regulators acting upon it in different time unit.

2.4 Dynamic Bayesian Network

DBN models time-series data by observing the values of a set of variables at different time units. DBN modelling usually consists of two steps: parameter learning and structure learning. In parameter learning, the joint probability distribution (JPD) of the variables is computed based on Bayes theorem. Assuming a microarray dataset with m genes and n observations, known as a $m \times n$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ whereby each row, vector $\boldsymbol{x}_m = (\boldsymbol{x}_{m1}, \ldots, \boldsymbol{x}_{mn})$ represents a gene expression vector observed at time t. The relationship is described as a *first-order Markov chain* whereby only forward edges are allowed. The JPD of the model has the general form of:

$$P(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mn}) = P(\boldsymbol{x}_1)P(\boldsymbol{x}_2|\boldsymbol{x}_1) \ldots P(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) \quad (5)$$

Based on the previously defined threshold, we discretised the expression values of the results acquired from the previous steps into three classes: -1, 0 and 1, which correspond to down-, normal and up-regulation respectively. The groups of potential regulators are then further divided in subsets. For example, in a group of potential regulators consisting gene R, and gene S, the subsets would be {R}, {S} and {R, S}. Each of the subset and the target gene, are subsequently organised into a data matrix with their discretised expression values. The conditional probabilities of each subset of potential regulators against their target genes are then computed. The next step is to search for the optimal network structure via a scoring function based on the Bayesian Dirichlet equivalence (BDe). The final results are imported into GraphViz (*http://www.graphviz.org*) for network visualisation and analysis.

3 Result and Discussion

3.1 Experimental Data and Setup

The *S. cerevisiae* cell cycle time-series gene expression data [7] encloses 6178 genes which were observed at four medium time series (alpha, CDC15, CDC28 and elu; 18, 24, 17 and 14 time points) and two short time series (CLN3, CLB2; both 2 time points). It also consists of 5.912% missing values (28,127 out of 475,706 observations).

The DBN modelling portion of IST-DBN is implemented under the framework of BNFinder [8], whereas the missing values imputation, potential regulators selection and the time lag estimation are implemented in MATLAB environment. For

performance evaluation, we compared the accuracy and computation time of the proposed IST-DBN against ISDBN and DBN (typified by BNFinder). The accuracy is measured by comparing the results from the three models to the established *S. cerevisiae* cell cycle pathway at KEGG (*http://www.kegg.jp*). The computation time of the models are compared on a 3.2GHz Intel Core i3 computer with 2GB main memory. The results are summarised in Table 1, in which the first row represents the network modelled by IST-DBN, the second row represents the network modelled by ISDBN, and the third row represents the network modelled by DBN. An edge indicates a relationship between the two connected genes. 'Correctly predicted relationships' denotes the number of relationships found in the established networks and also in the modelled results, 'sensitivity' is the percentage of correctly predicted relationships out of all predicted relationships, and 'specificity' correspond to the percentage of correct prediction that no relationship exists between two genes.

3.2 Experiment Results

Out of the established 35 gene-gene relationships, IST-DBN managed to identify 32 relationships (Fig. 3.) while ISDBN identified 30 – it failed to detect CLN1-CDC28 and FUS3-CDC28. CLN1 and FUS3 were assigned as the potential regulators of CDC28, however both of them have a transcription time lag exceeding 2 time units, and this caused ISDBN to erroneously dismiss them in the final group of potential regulators for CDC28. IST-DBN took into account of the transcriptional time lag and realigned them accordingly, which in turn increased the strength of their causal relationships with CDC28. Furthermore, by pairing up regulators and genes with a biologically relevant transcriptional time lag, IST-DBN was able to limit down the false positives to two while ISDBN erroneously detected six false positives. Both models were able to identify the feedback loop of the cell cycle pathway, for example, the sub-network of CDC28-SWI4/6-YOX1-MCM1-CLN3-CDC28 which signifies the transcription regulation during G1 phase of the cell cycle. On the other hand, DBN only identified 27 relationships. It missed out YHP1-MCM1, SWI4-CLN1 and CDC28-WHI5. This is more or less attributed to the fact that many missing values were found in the original expression profiles of the six genes, and the lack of an efficient imputation method caused DBN to lose its accuracy. IST-DBN reported 91.43% sensitivity and 98.08% specificity compared to ISDBN's 85.71% sensitivity and 94.06% specificity. DBN performed the worst among the three models, registering 77.14% sensitivity and 93.22% specificity.

While an edge denotes the existence of a relationship between two genes, there are four possible situations: correct direction and regulation type, correct direction but incorrect regulation type, misdirected but correct regulation type, and misdirected and wrong regulation type. One misdirected relationship and two regulation types in ISDBN were correctly reversed in IST-DBN. Also, the search space for both models is relatively small as the number of potential regulators is limited to those which experienced prior expression changes against targeted genes. Hence both models registered similar computation time in which IST-DBN has a slightly faster computation time of 24 minutes and 33 seconds compared to ISDBN's 25 minutes and 9 seconds. In contrast, DBN reported a computation time of 1 hour 8 minutes and

23 seconds, mostly due to the large search space where it includes all genes as potential regulators against target genes.

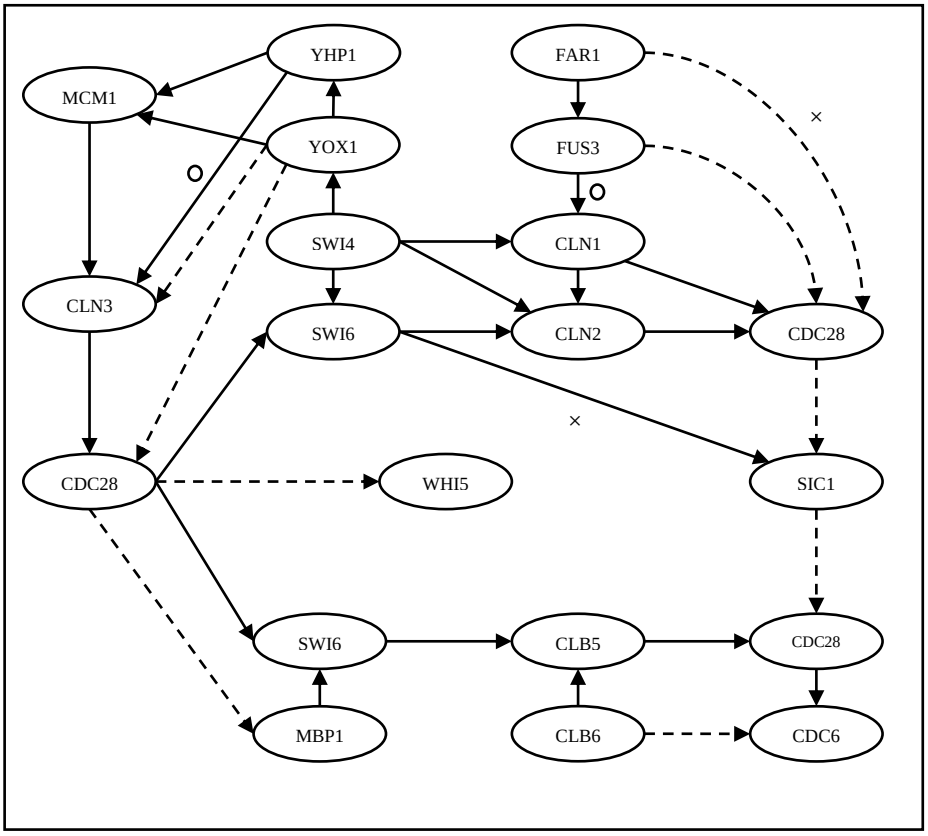

Fig. 3. Predicted cell cycle sub-network for *S. cerevisiae* dataset using IST-DBN. Dash edges (- - -) denote down-regulations and straight-lined edges (—) denote up-regulations. A cross represents an incorrect prediction; a circle represents an incorrect regulation type; an edge without any attachment is a correct prediction.

Table 1. The results of experiment study.

Model	Correctly predicted relationships	Sensitivity	Specificity	Computation time (HH:MM:SS)
IST-DBN	32	91.43%	98.02%	00:24:33
ISDBN	30	85.71%	94.06%	00:25:09
DBN	27	77.14%	93.22%	01:08:23

4 Conclusion

Conventional DBN has been hampered by three problems: the missing values in gene expression data, the relatively large search space caused by comprising all genes into the subset of potential regulators against target genes, and the lack of a way to handle transcriptional time lag. ISDBN was proposed by Chai *et al.* [4] to address the first two problems: Missing values are imputed based on linear combination of similar genes, and the search space is reduced by limiting the subset of potential regulators based on particular criteria. However, this model does not take into account of the transcription time lag. Therefore, in this paper, we proposed an enhanced ISDBN with time lag estimation (known as IST-DBN) to tackle the third problem. Instead of aligning by the default one time unit, IST-DBN uses the time difference between expression changes to pair up regulators and target genes. In this way, IST-DBN is able to capture most of the statistical correlation between genes that have a longer transcriptional time lag. Based on the *S. cerevisiae* cell cycle pathway dataset, IST-DBN showed promising results in terms of accuracy and computation time when compared to ISDBN and conventional DBN. It would be of our upmost interest to apply IST-DBN to other datasets, for instance, *E. coli* and *D. melanogaster*, as the resultant GRNs might be very useful for future gene intervention experiments or hypotheses testing purposes.

Acknowledgments

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

References

1. Karlebach, G., Shamir, R.: Modelling and analysis of gene regulatory networks. Nat. Rev. Mol. Cell Bio. 9(10), 770-780 (2008)
2. Ouyang, M., Welsh, W.J., Geogopoulos P.: Gaussian mixture clustering and imputation of microarray data. Bioinformatics 20(6), 917-923 (2004)
3. Jia, Y., Huan, J.: Constructing non-stationary dynamic Bayesian networks with a flexible lag choosing mechanism. BMC Bioinformatics 2010(11), S27 (2010)
4. Chai, L.E., Mohamad, M.S., Deris, S., Chong, C.K., Choon, Y.W., Ibrahim, Z., Omatu, S.: Inferring gene regulatory networks from gene expression data by a dynamic Bayesian network-based model. In: Omatu, S., De Paz, J.F., Rodriguez, S., Molina, J.M, Bernardos, A.M., Corchado, J.M. (eds.) DCAI 2012. AISC, vol. 151, pp. 379-386. Springer, Heidelberg (2012)
5. Kim, H., Golub, G., Park, H.: Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21(2), 187-198 (2005)

6. Yu, H., Luscombe, N.M., Qian, J., Gerstein, M.: Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. 19, 422-427 (2003)
7. 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. Mol. Biol. Cell 9, 3273-3297 (1998)
8. Wilczynski B., Dojer N.: BNFinder: exact and efficient method for learning Bayesian networks. Bioinformatics 25(2), 286-287 (2009)