

# Dynamic Bayesian Network (DBN) with Structure Expectation Maximization (SEM) for Modeling of Gene Network from Time Series Gene Expression Data

Yu Zhang, Zhidong Deng, Hongshan Jiang, Peifa Jia

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science,  
Tsinghua University, Beijing 100084, China.

## Abstract

*Exploring gene regulatory network is a key topic in molecular biology. In this paper, we present a new dynamic Bayesian network (DBN) framework embedded with structural expectation maximization (SEM) to model gene relationship. It is well-suited for analyzing the time-series data and can deal with cyclical structures that can not be tackled by static Bayesian network. We applied the new method to learning the regulatory network and the metabolic pathway from *Saccharomyces Cerevisiae* cell cycle gene expression data. The results show that the proposed method is capable of handling missing values in expression data sets, and the inference accuracy can further be improved.*

**Keyword: Microarrays; Gene regulatory networks; Dynamic Bayesian network; Structural expectation maximization**

## 1. Introduction

The establishment of gene regulatory network is critical to the understanding of the genetic regulation process. This problem has become an important challenge in recent years. The invention of microarray technology is viewed as a milestone, which helps scientist measure expression levels of thousands of genes simultaneously.

Several methods have been presented so far to learn gene network from microarray data, such as Boolean networks [2, 3], differential equations [4, 5], and Bayesian networks [6-8]. Among all of them, Bayesian network based approach has received a lot of attention because of the probabilistic nature of this model. It can be used to learn causal relationship and particularly, combine it with prior knowledge readily. However, there exist a lot of shortcomings for the static version of Bayesian network. First, it is unable to capture the temporal information. Second, it is impossible to model cyclic network, which is often considered to be an accurate description of real gene regulation mechanism.

In this paper, we employ dynamic Bayesian network (DBN) [1, 9,10], instead of its earlier static version, to model a gene network with cyclic regulation. In general, the DBN is well-suited for characterizing time-series gene expression data. Owing to the limitation of experimental condition, there are many missing values in the gene expression data sets, which usually have an impact on the inference accuracy. To address this problem, we propose a new DBN model embedded with structural expectation maximization (SEM) which is capable of efficiently dealing with missing data. Although there have been some literatures that involve the application of the SEM to learning Bayesian network, it is for the first time to introduce the SEM to learn the structure and parameters in the framework of the DBN. Using the gene expression data of *Saccharomyces Cerevisiae*, we carried out two different experiments, and both of them showed our new model had better performance than the previous work. We believe that the incorporation of prior knowledge efficiently improved the inference accuracy.

This paper is organized as follows. In section 2, we present a DBM model with SEM to reconstruct the gene

network. Section 3 implements our new method based on the gene expression data of *Saccharomyces Cerevisiae*. The comparative study of our results with the previous work is done. In section 4, we draw conclusions and suggest some future research topics of interest.

## 2. Method

As a graphic model, Bayesian network is defined by two parts. One is a graphic structure  $S$ , which is a directed acyclic graph (DAG) consisting of nodes and directed acyclic edges. The other is a parameter vector  $\Theta$  comprising a set of conditional probability distributions. Given the parent  $Pa_i$  of one node  $X_i$ , this node is conditionally independent of its non-descendants in Bayesian network. Under the Markov assumption, the joint probability distribution of network can be written as:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa_i). \quad (1)$$

Classical Bayesian network is unable to handle the cyclic edges [12]. Murphy and Mian [9] first employed dynamic Bayesian network (DBN) to build such a gene expression model with cyclic edge, as shown in Fig.1. Apparently, the DBN is able to avoid the ambiguity of the edge directions [14].

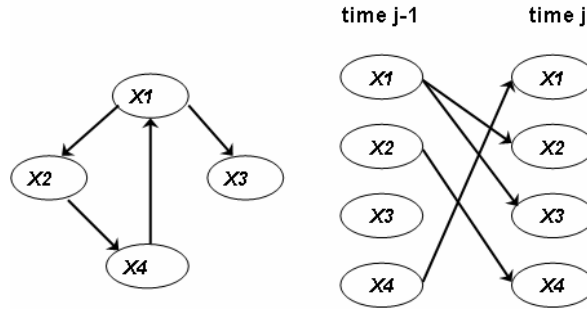


Fig.1. Example of a cyclic network. Bayesian network cannot handle the network (left) that contains a cycle  $X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_3 \rightarrow X_1$ . But the DBN can build a cyclic structure by dividing states of a gene into different time slices (right).

In this case, the joint probability of network can be rewritten below [1]:

$$P(X_{11}, \dots, X_{np}) = P(X_1)P(X_2 | X_1) \dots P(X_n | X_{n-1}) \quad (2)$$

where  $X_i = (X_{i1}, \dots, X_{ip})^T$  is a state vector of the  $p$ th gene at time  $i$ , and

$$P(X_i | X_{i-1}) = P(X_{i1} | P_{i-1,1}) \times \dots \times P(X_{ip} | P_{i-1,p}) \quad (3)$$

where  $P_{i-1,j}$  denotes the state vector of the parent gene of the  $j$ th gene at time  $i-1$ .

The following two assumptions [13] are regarded to be a basis of our transition from static Bayesian networks to the DBN: (1) the genetic regulation process is Markovian; (2) The dynamic casual relationships among genes are invariable over all time slices. Therefore we will search for the DBN that has the highest score. Here we give a new score based on the minimum description length.

$$Score(S, \Theta | D) = \log P(D | \Theta, S) - \frac{|\Theta|}{2} \log p. \quad (4)$$

In the previous works [1, 12], the dataset collected was assumed to be complete. But when the dataset has missing values, we cannot compute the marginal likelihood in closed form. The expectation-maximization (EM)

algorithm is a commonly-used method to cope with missing data. In this article, we use the structural EM (SEM) [15] to learn the network from partially observable gene expression data. The concept is similar to that of the complete data problem, except that the score of the network is found using the expected sufficient statistics from the EM algorithm.

The EM algorithm has two steps. The *E* step assigns some random values to parameter  $\Theta$ , and then the *expected sufficient statistics* for missing values are computed as:

$$E(p_{X_i=k, Pa_i=l}) = \sum_{j=1}^p P(X_i = k, Pa_i = l | d^j, \Theta, S). \quad (5)$$

In the *M* step, the expected sufficient statistics are considered to be real sufficient statistics from a complete dataset  $D'$ . The next step is to estimate the value of  $\Theta$  that maximizes the marginal likelihood  $P(D' | \Theta, S)$ ,

$$\theta_{X_i=k, Pa_i=l} = \frac{E(p_{X_i=k, Pa_i=l})}{\sum_{X_i} E(p_{X_i=k, Pa_i=l})}. \quad (7)$$

In the structural EM,

$$E(p_{X_i=k, Pa_i=l})^{S'} \cong \sum_{j=1}^p P(X_i = k, Pa_i = l | d^j, \Theta, S). \quad (8)$$

The resulting algorithm is shown in Fig. 2

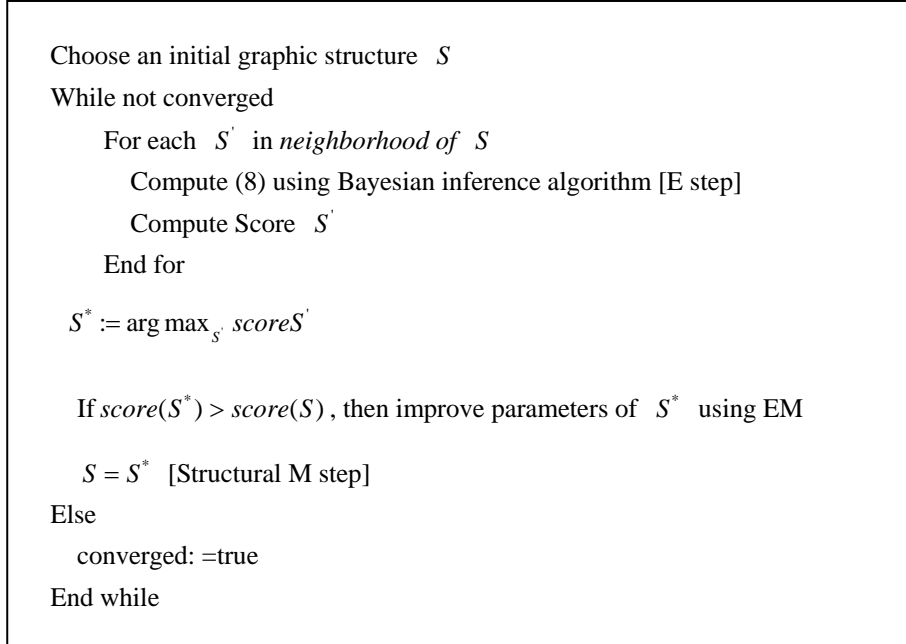


Fig. 2. Pseudo-code for structural EM

### 3. Experimental Results

When the DBN is exploited to analyze the time-series expression data, there are at least two situations [10]. First, one might have some prior knowledge of the regulatory network learned, such as the identification of transcription factor (TF). If we can identify the TFs, the prior knowledge can be used to reduce the search space. Second, we have no prior knowledge of the network and then need to reconstruct regulatory networks by considering all possible gene-pair relationships. In this paper, both the experiments were done.

In experiment 1, three TFs provided by Li *et al.* [12] were treated as the prior knowledge and the regulatory network was learned by our DBN model with SEM. After that, we compared the learned regulatory network with

the published work presented in [1]. In experiment 2, we had no any prior knowledge and learn the metabolic pathway proposed by DeRisi *et al.* [17].

The experiments described below were carried out with Matlab's Murphy's Bayesian Network Toolbox (BNT) [9]. We implemented our new DBN with SEM algorithm based on the framework of BNT.

### 3.1 Experiment 1

In order to compare our model with [1], we applied our approach to the *Saccharomyces Cerevisiae* cell cycle gene expression data that were also adopted by [1]. All these data were originally derived from the work given by Spellman [11], which was processed by four different methods: *cdc15*, *cdc28*, alpha-factor, and elutriation. The target network comes from the KEGG.

In Li's work [12], 11 genes were believed to be yeast TFs (SWI4, SWI6, STB1, MBP1, SKN7, NDD1, FKH1, FKH2, MCM1, SWI5, ACE2), and one cyclin gene (CLN3) were known to activate cell-cycle dependent genes. In our data set, there are 3 TFs, i.e., SWI4, SWI6, and MBP1.

We used circle to represent the correct estimation in Fig.3. Meanwhile, the Christ-cross meant the wrong estimation, and the triangle indicated either a misdirected edge or an edge skipping at most one node. The results are summarized in Table 1 for the accuracy analysis. In Table 1, the DBN-[1] represents the learned network based on [1] and the DBN-SEM-priors indicate our results obtained here. Note that when we calculate the specificity and sensitivity, the total number of pathways in the target network is 19.

Apparently, the number of the correctly identified edges increased from 4 in the DBN-[1] to 8 in the DBN-SEM-priors. The specificity and sensitivity calculated in the DBN-SEM-priors are both better than those from the DBN-[1]. The results showed that the DBN model with SEM when adding prior knowledge had better performance in reconstructing the regulatory network from time-series data than that achieved in [1].

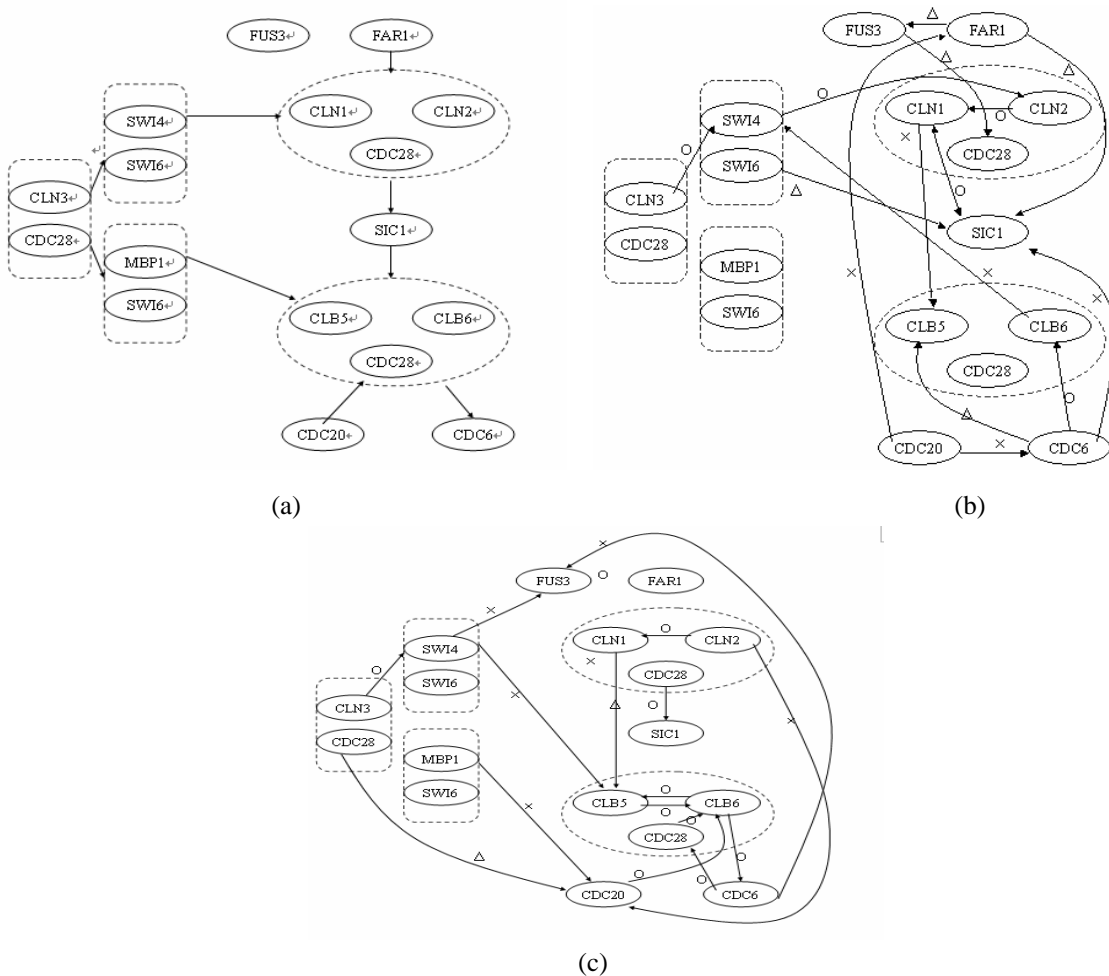


Fig.3. Regulatory network. (a) The correct pathways picked from the KEGG, (b) the result from [1], and (c) our result obtained in experiment 1 when added prior knowledge.

Table 1. Comparison of results achieved by our experiment 1 with that in [1]

	DBN-[1]	DBN-SEM-priors
correct estimation	4	8
wrong estimation	2	5
misdirected and skipping	8	3
specificity	26.7%	50.0%
sensitivity	21.1%	42.1%

### 3.2 Experiment 2

This experiment is to reconstruct the metabolic pathway of DeRisi *et al.* (1997). We chose the data set containing 12 genes and the result is shown in Fig.4. The symbols in the figure have the same interpretation as that in Fig.3. The accuracy analysis of experiment 2 is listed in Table 2, where the DBN-SEM-meta indicates our results and the DBN-[1]-Meta means ones that were reported in [1].

It is readily observed from these results that the incorporation of prior knowledge is capable of improving the inference accuracy and further reducing the computational cost. It can be concluded that the DBN-SEM model, either with priors or not, performs better than the results obtained in the DBN-[1].

The results from our analysis of yeast cell cycle expression data demonstrated that our method is capable of identifying gene–gene relationships, which can take advantage of the dynamic characteristic of the DBN model to tackle the cyclic structure, and efficiently handle missing data using the SEM algorithm. In particular, the DBN model embedded with SEM can fully combine prior knowledge in order to improve the performance.

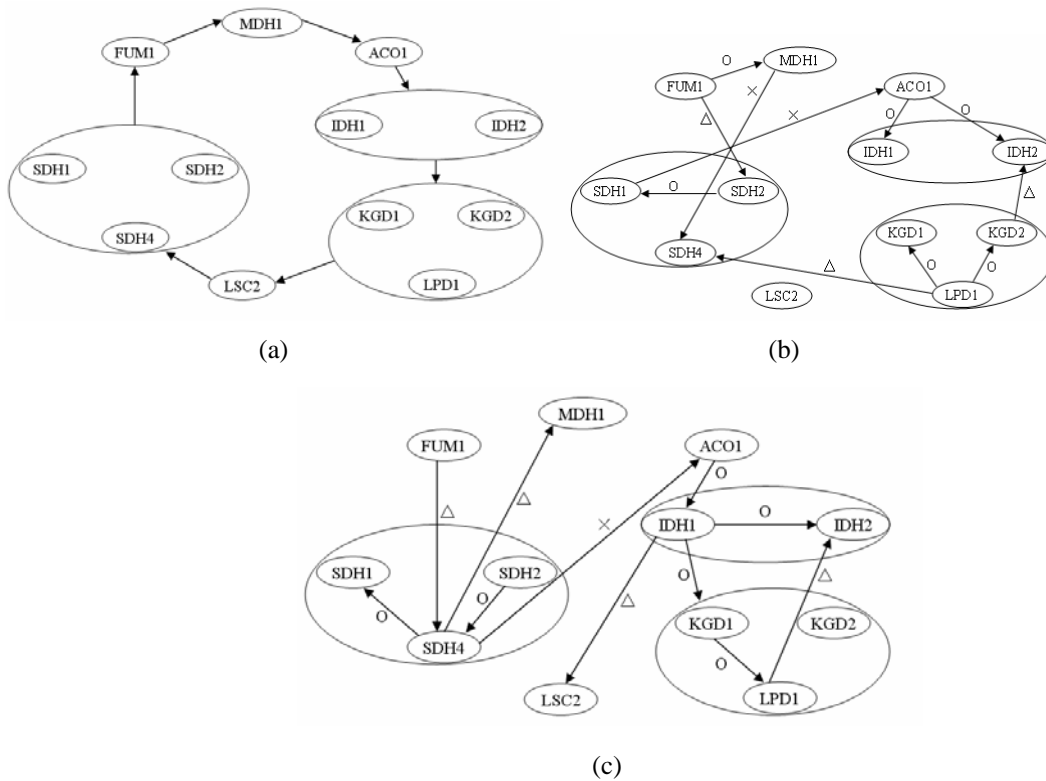


Fig.4. Metabolic pathway. (a) Target pathway [22], (b) the result of the Kim *et al.* [1], and (c) the result of our method obtained in experiment 2.

Table 2. Comparison of results achieved by our experiment 2 with that in [1].

	DBN-[1]-meta	DBN-SEM-meta
correct estimation	6	6
wrong estimation	2	1
misdirected and skipping	3	4
specificity	54.5%	54.5%
sensitivity	42.9%	42.9%

## 4. Conclusions

In this paper we proposed a new model based on the framework of dynamic Bayesian network (DBN) in order to reconstruct the genetic regulatory network. To deal with partial observations of gene expression data, we first added the SEM algorithm to the DBN model. We validated the effectiveness of our method using two experiments. One is based on the real time series microarray data of *S. cerevisiae* cell cycle to find the gene relationships and the other is to predict the metabolic pathway. Compared with the results yielded in [1], the prediction accuracy of our method outperformed the previous work.

In general, either Bayesian network or dynamic Bayesian network can make use of the prior information when conducting inference. In experiment 1, we first identified transcriptional factors (TFs) before learning network and the TFs were then regarded as prior information good for the learning process. This helped reduce the search space and efficiently improved the result.

There are several research lines for the future work. First, our method is strongly dependent on the quality of the microarray data. The discretization of data may lead to losing useful information and the data noise also has an impact on the result. We are attempting to develop our method to directly cope with continuous expression levels. Second, regulatory network of cell depends on not only the transcriptional regulation but also the post-transcriptional and external signaling events. Learning the genetic regulatory interactions only from expression data is unable to discover the global scene of the genetic regulatory pathways. In the future, one of our goals is to employ the framework reported here to deal with multiple data sources, such as protein-protein interaction, gene annotation, and promoter sequence. How to jointly incorporate all these additional data sources as prior knowledge may be worth trying. Finally, our method can be used for either gene network modeling or many other problems of computational biology. The framework is a good platform to investigate biological process.

## Acknowledgment

This work was supported in part by the National Science Foundation of China under Grant No. 60321002 and the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of MOE (TRAPOYT), China.

## References

- [1] Kim, S., Imoto, S., Miyano, S.: Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems* 75 (2004) 57-65

- [2] Liang, S., Fuhrman, S., Somoyi, R.: REVEAL: a general reverse engineering algorithm for inference of genetic network architectures. In: Proc. Pacific Symposium on Biocomputing (1998) 18-29
- [3] Akutsu, S., Miyano, S., Kuhara, S.: Algorithms for inferring qualitative models of biological networks. In: Proc. Pacific Symposium on Biocomputing (2000) 290-301
- [4] Chen, T., He, H.L., Church, G.M.: Modeling gene expression with differential equation. In: Proc. Pacific Symposium on Biocomputing (1999) 29-40
- [5] D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R: Linear modeling of mRNA expression levels during CNS development and injury. In: Proc .Pacific Symposium on Biocomputing (1999) 41-52
- [6] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. *Computational Biology* 7(3) (2000) 601-620
- [7] Hartemink, A.J., Gifford, D.K., Jaakkola, T.S, and Young, R.A.: Combing location and expression data for principled discovery of genetic regulatory network models. In: Proc .Pacific Symposium on Biocomputing (2002) 437-449
- [8] Imoto, S., Goto, T., Miyano, S.: Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. In: Proc .Pacific Symposium on Biocomputing (2002)-175-186
- [9] Murphy, K., Mian, and S.: Modelling gene expression data using dynamic Bayesian networks. Technology Report, Computer Science Division, University of California Berkeley, CA, (1999)
- [10] Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* Vol.21 (2005) 71-79
- [11] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Aders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccaromyces Cerevisiae* by microarray hybridization. *Mol.Biol.Cell* Vol.9 (1998) 3273-3297.
- [12] Li, S.P., Tseng, J.J., Wang, S.C.: Reconstructing gene regulatory networks from time-series microarray data. *Physica A* Vol.350 (2005) 63-69
- [13] Wu, C.C., Huang, H.C., Juan, H.F., Chen, and S.T.: GeneNetwork: An interactive tool for reconstruction of genetic network using microarray data. Supplementary information, Taiwan. (2003)
- [14] Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* Vol.19 (2003) 2271-2282
- [15] Friedman, N., Murphy, K., and Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proc Conf. Uncertainty in Aritif. Intell. (1998) 139-147
- [16] Home page of KEGG: <http://www.genome.ad.jp/kegg>
- [17] DeRisi, J., Lyer, V., and Brown, P.: Exploring the metabolic and gene control of gene expression on a genomic scale. *Science* Vol.278 (1997) 680-686.