Inferring Gene Networks from Gene Expression Data Using Dynamic Bayesian Network with Different Scoring Metric Approaches

Masarrah Abdul Mutalib[1], Lian En Chai[1], Chuii Khim Chong[1], Yee Wen Choon[1], Safaai Deris[1], Rosli M. Illias[2], and Mohd Saberi Mohamad[1,*]

[1] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia
{masarrah2,lechai2,ckchong2,ywchoon2}@live.utm.my, safaai@utm.my, saberi@utm.my

[2] Department of Bioprocess Engineering, Faculty of Chemical Engineering, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
r-rosli@utm.my

Abstract. Inferring gene networks can be defined as the process of identifying gene interactions from experimental data through computational analysis. The aim is to infer gene network from gene expression data using dynamic Bayesian network (DBN) with different scoring metric approaches. The previous method, Bayesian network has successfully identified those gene networks but there are some limitations. Hence, DBN is able to infer interactions from a data set consisting time series rather than steady-state data. This research is conducted in order to construct and implement gene network and to analyze the effect by applying a different scoring metric approach for modeling gene network. In order to achieve the goals, a discrete model of DBN is used with different scoring metric approaches which are BDe and MDL. The *S. cerevisiae* cell cycle pathway is used for this research. To ensure the gene networks are biologically probable, this research employs previous annotation relative to the dataset. By having all of these implementations, this research is able to identify the effect of different scoring metric approaches, identify biologically meaningful gene network within the gene expression datasets and display the results in convenient representations.

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

1 Introduction

Dynamic Bayesian network (DBN) is well defined as a Bayesian network (BN) that represents sequences of variables. DBN can construct cyclic regulations using time delay information. DBN uses time series data for constructing causal relationships

* Corresponding author.

A.S. Sidhu & S.K. Dhillon (Eds.): *Adv. in Biomedical Infrastructure 2013*, SCI 477, pp. 77–86.
DOI: 10.1007/978-3-642-37137-0_9

among random variables. Friedman *et al*. [1] first applied DBN to the analysis of gene networks. They constructed a discrete DBN model and used the Bayesian Dirichlet equivalence (BDe) scoring metric for learning networks. Ong *et al.* [2] also used a discrete DBN model but combined it with prior biological knowledge and current observations to model the tryptophan metabolism in *E. coli*. They utilized a repetitive EM (Expectation-maximization) algorithm to compute scores in learning network structure. On the other hand, to avoid data loss due to discretization, Kim *et al.* [3] developed a continuous DBN model with non-parametric regression model based on *B*-splines to take into account of linear dependencies. To select the optimal network, Kim *et al.* [3] subsequently defined a scoring metric known as $BNRC_{dynamic}$ based on the Laplace approximation.

Inferring gene networks can be defined as the process of identifying gene interactions from experimental data through computational analysis. Gene expression data from microarray are typically used for this purpose. The aim is to infer gene network from gene expression data using DBN with different scoring metric approaches. In addition, network visualization tools are available to indicate the network surrounding a gene of interest by extracting information from experimental data sets, such as Cytoscape [4]. We evaluated the efficiency of each scoring approach through the analysis of the *S. cerevisiae* gene expression data.

2 Materials and Methods

In previous works, researchers used BN which could not model a feedback loop because it did not have loops or cycles. In this section, we describe the details of the DBN-based model for inferring GRNs from gene expression data. In essence, the proposed model consists of three main steps: missing values imputation, construct gene network and evaluating network structures using scoring metric with respect to the given data. The following sub-sections discuss in detail for each of the three main steps.

2.1 Experimental Data and Missing Values Imputation

After all of the possibly used method and techniques identified, this is the stage where the researcher develops and implements a computational model based on the techniques in the previous steps. The model is implemented using BNFinder software [5]. This software allows for BN reconstruction from experimental data. Besides that, it supports DBN and if the variables are partially ordered, this also applies for static BN. It is written in python, and distributed under GNU GPL Library version 2.

The experimental study is based on the *S. cerevisiae* cell cycle time-series gene expression data [6]. However, the dataset contains missing values which must be processed. Conventional methods of treating missing values include repeating the microarray experiment which is not economically 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 k-nearest neighbor method (kNN) imputation algorithm [7] that is the most fundamental and simple classification

methods, and should be one of the first choices for a classification study when there is little or no prior knowledge about distribution of the data.

2.2 Construction of Gene Networks

The DBN is used to construct gene networks, hence producing directed acyclic graphs (DAGs). For this research, we used Cytoscape for visualizing complex network and integrating these with any type of attribute data.

After the gene networks have been constructed, the performance of the gene networks constructed using DBN is evaluated. To evaluate the gene performance, the networks constructed are compared with the sub-networks constructed by Dejori [8]. Dejori [8] has also implemented BN to construct gene networks from *S. cerevisiae* dataset which is the same dataset in this research. Therefore, the sub-networks constructed by Dejori [6] are the benchmarks for this research.

We compared both of the methods by calculating True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Positive is the number of edges that exist in both network constructed by Dejori [6] and in the research. True Negative (TN) is the number of edges that do not exist in both networks (Dejori and this research). False Positive (FP) is the number of edges that exist in this research, but do not exist in the network by Dejori [8], while False Negative (FN) is the number of edges that exist in Dejori [8], but do not exist in this research.

2.3 Evaluating Network Structures

This research applies different scoring metric approaches in order to get the best network structures. The scoring metric approaches used to test in this research are the BDe score and the MDL score.

The BDe scoring criterion originates from Bayesian statistics and corresponds to posterior probability of a network given data. BDe uses Bayesian analysis to evaluate a network given a dataset. The Dirichlet distribution is a multinomial distribution that describes the conditional probability of each variable in the network, and has many properties that are useful for learning.

The MDL scoring criterion originates from information theory and corresponds to the length of the data compressed with the compression model derived from the network structure. Besides that, MDL provides the criterion for the selection, prediction and estimation of models. The purpose of MDL is to discover regularities in observed data. Generally, both BDe and MDL scores were originally designed for evaluating discrete variables.

3 Result and Discussion

The sub-networks that are chosen to be compared are YPL256C sub-network and YOR263C sub-network. TP, TN, FP, and FN are calculated to evaluate the performance of the sub-networks constructed from this research.

3.1 YPL256C Sub-network

Fig. 1 shows the YPL256C sub-network that is constructed by Dejori [8]. It can be seen that, the network consists of 12 nodes (genes) and 9 directed edges. However, the node for YGR108W does not form any edges with other nodes in the network.

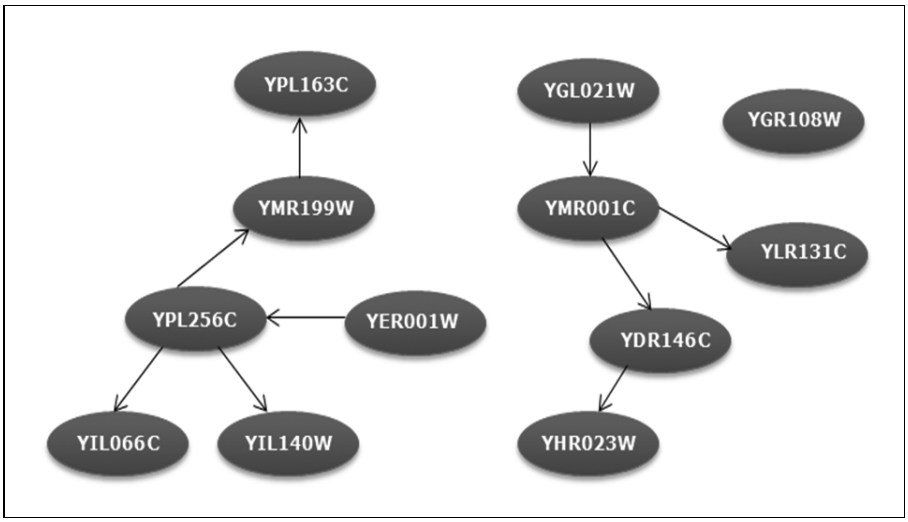

Fig. 1. YPL256C sub-network constructed by Dejori [8]

As shown in Fig. 1, there is two directed edge from gene YPL256C to gene YIL066C and YIL140W. It shows that there is a causal dependency between these three genes. The functions of gene YPL256C are encoding for G1-cyclin which involves in regulation of the cell and activates Cdc28p kinase to promote the G1 to S phase transition. A YIL066C gene is a minor isoform of the large subunit of ribonucleotide-diphosphate reductase which is involved in DNA replication. Whereas, the YIL140W gene is an integral plasma membrane protein that is required for axial budding in haploid cells and has potential to Cdc28p substrate. Therefore, a causal dependence of YIL066C and YIL140W from YPL256C is biologically logical since their functions are correlated.

As we look further, gene YGL021W contains characteristic motifs for degradation via the APC pathway and phosphorylated in response to DNA damage which is quite similar to A1k2p and to mammalian haspins. Gene YGL021W regulates YMR001C with multiple functions in mitosis and cytokinesis through substrate phosphorylation, also functioning in adaptation to DNA damage during meiosis. An unexpected result is the gene YGR108W does not connect any edge with other nodes. However, it does form edges with other nodes in the research done by Spellman *et al.* [6].

Fig. 2 shows the YPL256C sub-network that is constructed in this research using BDe and MDL scoring metric approaches. It is very clear that both networks consist of 12 nodes and 24 edges. It shows a different number of edges obtained in this research as compared to Dejori [8]. Through this research, we can see that several edges in the network are from cyclic regulation and have at least one directed edge

with other nodes. The network done by Dejori [8] does not show any cyclic regulation and the gene YGR108W failed to construct with any edge. About 20 new edges had been identified in this research. It is two times more compared to the result obtained by Dejori [8]. Hence, it is proven that DBN implemented in this research are able to construct cyclic regulation and form more potential edges between genes in a sub-network.

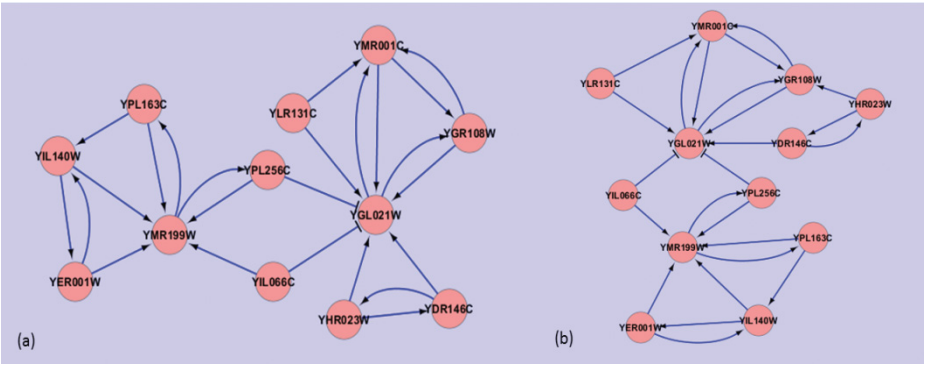

Fig. 2. YPL256C sub-network constructed with (a) BDe, (b) MDL scoring metric approaches

Table 1 shows the comparison of edges formed in YPL256C sub-network between Dejori [8] and this research. True Positive (TP) is the number of edges that exists in both network constructed by Dejori [8] and this research. False Negative (FN) is the number of edges that exist in Dejori sub-network, but does not exist in network of this research. False Positive (FP) is the number of edges that exists in network of this research, but does not exist in Dejori [8]. True Negative (TN) is the number of edges that does not exist in both networks constructed by Dejori [6] and in this research. The sensitivity for this sub-network is 44% whereby 4 directed edges that exist in the network by Dejori [8] have been captured in this research as well. However, there are about 5 directed edges exist in Dejori [8] but it does not exist in the network of this research. The missing edge is between gene YPL256C to YIL140W and YIL066C, gene YER001W to YPL256C, gene YMR001C to YLR131C and YDR146C respectively.

Table 1. Result of YPL256C sub-network

Condition	Number of Edges	Statistical Measures
TP	4	Sensitivity 44.44%
FN	5	
FP	20	Specificity 84.96%
TN	113	

The specificity for this sub-network is approximately 84.96%. Gene YLR131C regulates both genes of YMR001C and YGL021W. Gene YLR131C encodes for transcription factor that activates transcription of genes expressed in the G1 phase of the cell cycle. On the other hand, gene YMR001C is involved in regulation of DNA replication which encodes a protein. Furthermore, gene YIL066C is expressed only after DNA damage occurred in order to cope with the function of YMR199W. Gene YMR199W encodes for G1-cyclin which involved in regulation of the cell cycle. Therefore, it is biologically logical for YIL066C regulating the expression of YMR199W.

3.2 YOL263C Sub-network

In this study, we compared the YOR263C sub-network obtained from this research and YOR263C sub-network by Dejori [8]. Fig. 3 shows the YOR263C sub-network that is constructed by Dejori [8]. It can be seen that, the network consists of 8 nodes (genes) and 6 undirected edges. The undirected edge between YOR263C and YOR264W are the most conspicuous features in the sub-network because both genes are located next to each other on the DNA strand of chromosomes XV. However, the biological and molecular for both genes are still unknown. Gene YNR067C and YGL028C is another feature with high confidence level that is the undirected edge. YNR067C is a daughter cell-specific secreted protein with similarity to glucanases and it degrades cell wall from the daughter side causing daughter to separate from mother. The function of YNR067C is still currently unknown. The function of YGL028C is known to be a soluble cell wall protein and play a role in conjugation during mating based on its regulation by Ste12p. It also has an undirected edge with YER124C which may regulate cross-talk between the mating and filamentation pathways and deletion affects cell separation after division and sensitivity to alpha factor and drugs affecting the cell wall. Gene YGL028C is related to YLR286C which is an endochitinase required for cell separation after mitosis. YER124C has undirected edge with two nodes (YLR286C, YGL028C), and both nodes are functionally related to cell wall biogenesis, therefore it can be assumed that it is involved in cell wall biogenesis. Gene network constructed using BN have provided a testable prediction of an unknown gene function.

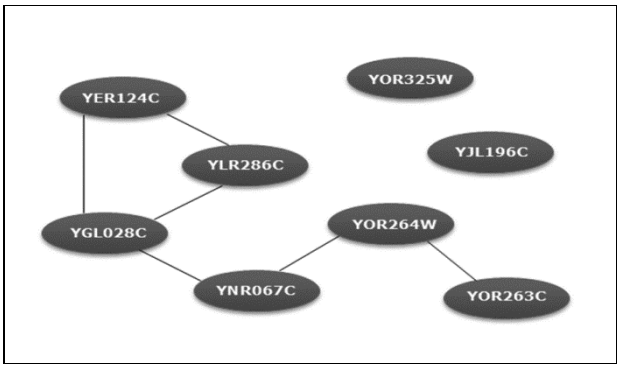

Fig. 3. YOL263C sub-network constructed by Dejori

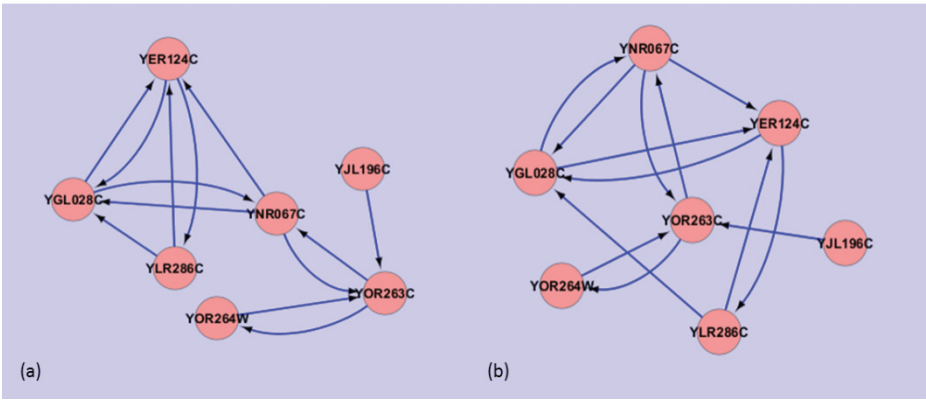

Fig. 4. YOL263C sub-network constructed with (a) BDe, (b) MDL scoring metric approaches

Fig. 4 shows the YOR263C sub-network that is constructed in this research using BDe and MDL scoring metric approaches. It is very clear that both networks consist of 7 nodes and 13 edges. They show a different number of edges obtained in this research as compared to the result obtained by Dejori [8]. Through this research, we can see that several edges in the network are form cyclic regulation and have at least one directed edge with other nodes, while the network done by Dejori [8] does not show any cyclic regulation. The main difference between this research and Dejori [8] are that they can show the interactions between genes clearer. As we can see in the Dejori [8] sub-network, the edge formed between YOR263C and YOR264W cannot show which gene is regulating another. However, this research shown clearly that YOR263C is regulating YOR264W and it is a cyclic regulation. It means that the expression level of YOR264W is depending on YOR263C and YNR067C as well. About three new edges have been identified in this research.

Table 2. Result of YOR263C sub-network

Condition	Number of Edges	Statistical Measures
TP	5	Sensitivity 83.33%
FN	1	
FP	3	Specificity 80.00%
TN	12	

Table 2 shows the comparison of edges in YOR263C sub-network between the network constructed by Dejori [8] and this research. The sensitivity of YOR263C sub-network is approximately 83.33%. There are about 5 cyclic edges formed in this sub-network. The specificity for this sub-network is approximately 80%. This shows that the DBN implemented in this research is capable of uncovering more potential edges, interactions and cyclic regulation between genes compared with the study by Dejori [8].

3.3 Performance of Scoring Metrics

Table 3 summarizes the computation time comparison between scoring metric approaches of YPL256C sub-networks. MDL excels in speed as it had a computation time of 1 minute and 10 seconds while BDe took approximately 2 minutes. This concurs with the finding of Vinh *et al.* [9] which discovered that BDe is more time-consuming than MDL. However, both scoring metric approaches obtained the same network results (24 edges and 12 nodes) and accuracy (as summarized in Table 1). On the other hand, Table 4 shows the computation time comparison between scoring metric approaches of YOR263C sub-networks. Both scoring metric approaches gave roughly the same computation time which is 1 second. This is probably due to the fact that YOR263C has a smaller network structure compared to YPL256C. Both scoring metric approaches also computed the same network results (13 edges and 7 nodes) as well as accuracy (refer to Table 2). The experiment with YPL256C showed that MDL has an advantage in computation time without compromising the accuracy for network inference.

Table 3. YPL256C: Comparison of computational time between scoring metrics

Sub-network	Scoring Metric Approaches	Computation Time (HH:MM:SS)
YPL256C	BDe	00:02:01
	MDL	00:01:10

Table 4. YOR263C: Comparison of computational time between scoring metrics

Sub-network	Scoring Metric Approaches	Computation Time (HH:MM:SS)
YOR263C	BDe	00:00:01
	MDL	00:00:01

Table 5. Network scores between scoring metrics for YPL256C and YOR263C sub-networks

Scoring Metric	YPL256C	YOR263C
BDe	*470.257*	*342.084*
MDL	704.546	504.177

Table 5 shows the network scores obtained by both scoring metrics for YPL256C and YOR263C sub-networks respectively. Lower score are said to have optimal network structure. In both sub-networks, BDe performed better than MDL. Nevertheless, this scoring advantage did not influence much on the inference of

optimal network structure as both scoring metric approaches obtained the same network structure for YPL256C and YOR256C.

4 Conclusion

DBN has been widely utilized by researchers in gene networks inference from gene expression data as it is robust, able to handle feedback loops and the temporal aspect of time-series data. To learn the optimal network structure, BDe or MDL scoring metric are often employed in the DBN model. This research is conducted to analyze the influence of both scoring metrics on gene networks inference using DBN. Based on the experiments done on two *S. cerevisiae* cell cycle sub-networks YPL256C and YOR263C, we found that MDL has faster computation speed in larger network structure but BDe has an edge in representing exactness of statistical interpretation. Therefore, we suggest using MDL in exceptionally large networks as exponentially increased computation time would negate the statistical advantage of BDe. BDe is more suitable for smaller networks or in such circumstance whereby accuracy is much sought after. For future work, we would like to apply different scoring function that satisfies the score equivalence property.

Acknowledgments. We would like to thank the Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proc. 14th Conference on the Uncertainty in Artificial Intelligence, San Mateo, pp. 139–147 (1998)
2. Ong, I.M., Glasner, J.D., Page, D.: Modelling regulatory pathways in E. coli from time series expression profiles. Bioinformatics 18, S241–S248 (2002)
3. Kim, S., Imoto, S., Miyano, S.: Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In: Priami, C. (ed.) CMSB 2003. LNCS, vol. 2602, pp. 104–113. Springer, Heidelberg (2003)
4. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003)
5. Wilczynski, B., Dojer, N.: BNFinder: exact and efficient method for learning Bayesian Networks. Bioinformatics 25(2), 286–287 (2009)
6. 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)
7. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520–525 (2001)

8. Dejori, M.: Analyzing Gene Expression Data with Bayesian Networks. Graz University of Technology, Austria (2002)
9. 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 27(19), 2765–2766 (2011)
10. De Campos, L.M.: A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests. J. Mach. Learn. Res. 7, 2149–2187 (2006)