Information Sciences and Computer Engineering, Vol. 1, No. 1, (2010) 68–76

PHARUS

International Journal of
Information Sciences and Computer Engineering

journal homepage: http://www.ijisce.org

Improved Bayesian Network Structure Learning for Breast Cancer Prognosis

Farzana Kabir Ahmad[a,*], Safaai Deris[b], Nor Hayati Othman[c]

[a] *Graduate Department of Information Technology, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia*
[b] *Faculty of Computer Science and Information System, 81300 Skudai, Johor Bharu, Malaysia*
[c] *Clinical Research Platform & Pathologist, 16150 Kubang Kerian, Kelantan, Malaysia*

***Abstract*– Structure learning of Bayesian networks is a well-researched but computationally and NP-hard task. We present an algorithm that integrates a low-order conditional independence approach for learning structures of Bayesian networks. Our algorithm also makes use of basic Bayesian network concepts. We show that the proposed algorithm is capable of handling networks with a large number of variables and small sample size in the case of microarray data analysis. We present the applicability of the proposed algorithm on breast cancer data sets and also compare its performance and computational efficiency with full-order conditional independence method. The experimental results show that our method can efficiently and accurately identify complex network structures from data.**

Keyword: Bayesian network, structure learning, low-order conditional independence, breast cancer.

1. Introduction

Bayesian network is a directed graph that represents the joint probability distribution among a large number of variables and allows for performing probabilistic inference with these variables [1]. It has been applied to a wide range of tasks such as natural spoken dialog systems [2], vision recognition [3], expert systems [4], medical diagnosis [5], and genetic regulatory network inference [6]. Since their introduction in the mid-1980s [7], Bayesian networks have become the prominent technique in biomedical research as it is especially suited for capturing and reasoning with uncertainty data [8]. Several studies have applied this technique to predict future outcome among breast cancer patients and derive a better conclusion that could help oncologists making precise decision regarding the appropriate treatment to be given. Maskery et al. [9] has used Bayesian network to access the co-occurrence from breast pathology information, while Carrivick et al. [10] employed a different approach by applying this method to gain breast cancer prognostic signatures from microarray data, which would provides better insight in breast cancer progression.

A Bayesian network consists of two important components: a directed acyclic graph (DAG) representing the dependency structure among the variables in the network and a conditional probability table (CPT) for each variable in the network given its parent set. Therefore, learning a Bayesian network from data requires both identifying the model structure $\hat{G}$ and identifying the corresponding set of model parameter values.

However, as the number of possible structures grows exponentially with respect to the number of variables, an exhaustive search over all possible structures becomes computationally inhibitive for network structures of moderate size. As a result, in recent years, there has been a growing interest in learning the structures of Bayesian networks in order to identify "good" DAG structures from data. Consequently, many structure-learning methods have been proposed, including methods based on conditional independence tests [11] and methods based on a scoring metric and a search algorithm [12, 13, 14].

These structure-learning methods have been usually presented with less than superexponential complexity with respect to the number of variables and it works well for smaller networks. The key challenge of Bayesian network is to learn the structure of data from insufficient sample with high numbers of variables involved, for instance in the case of microarray data analysis. Thus, this study attempts to address the difficulties in obtaining structure learning in Bayesian network with low-order conditional independence. Breast cancer prognosis has been an ideal problem to cater and microarray data has been utilized for this study.

Breast cancer has been identified as the second most common cause of deaths among women in United Stated. In 2006, it is reported about 212,000 new cases of invasive breast cancer were diagnosed, along with 58,000 cases of non-invasive breast cancer and 40,000 women died due to this disease [15]. The same scenario also occurred among Malaysian population, where breast cancer is discovered as the second cause of death after lung cancer being the common killer. The National Cancer Registry 2002 [16] stated that in the year 2002, 26,089 people were diagnosed with cancer in Peninsular Malaysia and 14,274 (55%) cases were cancers among women and 30.4% from them suffered from breast cancer. These high rates of deaths have stimulated extensive researches in breast cancer.

The major problem in breast cancer is the ability to predict and treat metastatic breast cancer is extremely limited and inadequate [17]. In numerous patients, minuscule clinically evident metas-

[*] Corresponding author:
Email address: farzana58@uum.edu.my, Ph: +60 49284743, Fax: +60 49284753

tases have already occurred by the time the primary tumor is diagnosed. Although chemotherapy or hormonal therapy reduces the risk of distant metastases by one third, but it is estimated that about 70% patients receiving treatment would have survived without it. The intricacy to reliably prognosis the risk of breast cancer metastases for individual patients stems from the fact that cancer is the result of a complex interplay between numerous factors, such as genetic and clinical factors. Current breast cancer indices namely St. Gallen [18] and NIH [19] were discovered contain some limitations in order to predict breast cancer metastases, since patients with identical diagnostic and clinical prognostic profile can have apparently diverse clinical outcome. This phenomenon is due to the missing genes cellular proliferations information in current breast cancer indices and a high reliance on a complex and inexact combination of clinical and histopathological data for instance age, lymph node involvement and grade. Thus, these indices were notified to provide misleading results as it mainly group molecularly distinct patients into alike clinical classes generally based on morphological of disease [20, 21, 22]. Although clinical and histopathological data are proven to be relevance to predict breast cancer metastases, gene cellular proliferation is also essential information that needs to be taken into consideration since breast cancer is a complex and heterogeneous disease.

The advance in biomedical research with the invention of microarray technology has modernized the approach of cancer study in such a way thousand of genes can be monitored simultaneously. Microarray-based expressions have led to the promise of cancer prognosis using new molecular-based approaches. It has become a standard tool in many genomic research laboratories. Due to the overwhelming flow of data currently being produced in the biomedical sciences and complex interaction between various factors involves in breast cancer invasion and metastases, a network-based model approach with microarray data is described here. Bayesian network has been proposed in this study as a method to develop a network-based model for breast cancer prognosis. Bayesian network is a well known technique in biomedical and bioinformatics and offers several advantages such as it inherently model the uncertainty in the data. It is also a successful combination between probability theory and graph theory. Furthermore, this technique allows different strategies and data types to be combined.

The remainder of this paper is organized as follows. Section 2 describes representation and modeling issues for learning Bayesian network from microarray data. It includes the mathematical underlying concepts of Bayesian network and two main learning steps to be performed during model implementation, structure and parameters learning. This section also explains the utilization of low-order conditional independence method within Bayesian network approach. We present the empirical evaluation for breast cancer prognosis in Section 3 and it related discussion has been explicitly enlighten in Section 4. Lastly, Section 5 offers concluding and future direction remarks.

2. Bayesian Networks

Bayesian network is a probabilistic graphical model that consists of two major parts; a dependency structure and local probability models. The dependency structure represents a set of variables and their probabilistic independencies. Formally, Bayesian network is a DAG whose nodes represent variables and whose missing edges encode conditional independencies between variables. For example, X_3 is conditional independence of X_4 given X_1, which can be written as $X_3, \perp X_4|X_1$. The second part of this model, the local probability models specifies how the variables depend on their parents. These dependencies can be represent by CPT. Fig. 1 shows the simple Bayesian network with five genes. The X_3 gene in this example has two parents, X_1 and X_2. The CPT for variable X_3 is shown alongside of DAG diagram.

Bayesian network B is defined as a pair $B = (G, P)$, where $G = (V(G), A(G))$ is a DAG with a set of variables (or nodes) $V(G) = X_1, \ldots X_n$ and arcs $A(G) \subseteq V(G) \times V(G)$ and P correspond to joint distribution on the variables. The variables V represent genes or other elements and correspond to random variables X. In the context of this study, V may indicate as a gene, while X is the expression level of V.

If there is arc from node X_1 to another node X_4, X_1 is called a parent of X_4, and X_4 is a child of X_1. The set of parent nodes of a node X_i is denoted by parents (X_i). A DAG is a Bayesian network relative to a set of variables if the joint distribution of the node values can be written as the product of the local distribution of each node and it parents:

$$P(X_1, \ldots X_n) = \prod_{i=1}^{n} P(X_i|parents(X_i)) \tag{1}$$

The joint distribution of Fig. 1 can be obtained by Equation 2. If node X_i has no parents its local probability is said to be unconditional, otherwise it is conditional. If the value of a node is observed, then the node is said to be an evidence node.

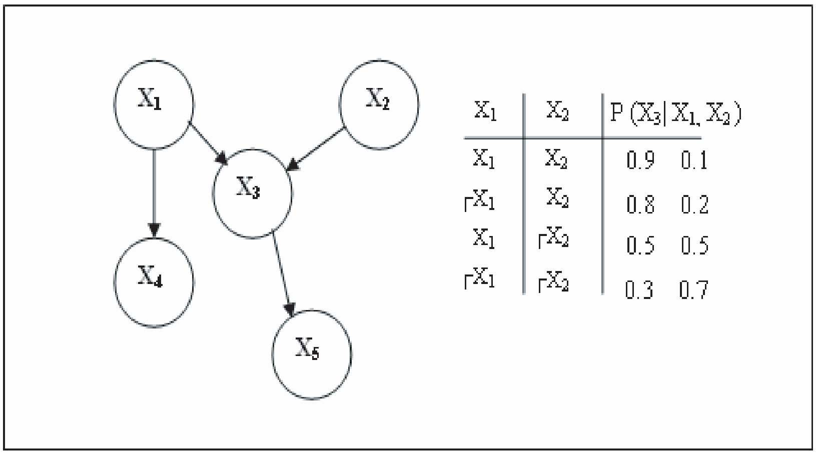
$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_3|X_1, X_2)P(X_2)P(X_4|X_1)P(X_5|X_3) \tag{2}$$

The main objective of Bayesian Network is to allow the probabilistic inference to be performed. Inference is defined as the process of deriving logical conclusions or probabilistic values for each variable when the values of other variables are known. In the fact that, conditional independencies can be recognized through DAG and with the availability of CPT by a graph edge, not all joint probabilities in Bayesian network have to be calculated to make a prediction.

2.1. Learning in Bayesian Networks

The representation and use of probability theory make Bayesian network appropriate for learning from incomplete data sets, expressing causal relationship, combining domain knowledge and data as well as avoiding overfitting in a model. Bayesian network has been applied in numerous applications. Mainly, there are two steps to be performed to build Bayesian network model; parameter and structure learning. Parameters of Bayesian network can be learned from data. For example, the conditional probability tables could be constructed from empirical evidence. The parameters also can be in any form either, discrete or it may also be continuous and be modeled by a probability density function, commonly Gaussian distributions are used. This study applied a continuous form data and no discretization has been applied to data set.

Structure learning, on the other hand is a learning of network construction from data. When the structure of Bayesian network



X_1	X_2	$P(X_3 \mid X_1, X_2)$	
X_1	X_2	0.9	0.1
⌐X_1	X_2	0.8	0.2
X_1	⌐X_2	0.5	0.5
⌐X_1	⌐X_2	0.3	0.7

Fig. 1. DAG and CPT for a simple Bayesian network.

is unknown, which mean it cannot be specified by prior knowledge, a heuristic search can be implemented to find a 'good' structures. However, as the number of variables relatively high to number of sample size, heuristic search technique over all possible structures becomes computationally inhibitive and untraceable. Therefore, we have applied low-order conditional independence method to improve the structure learning from microarray data. In order to learn the underlying causal model, one needs more than just structure learning, as many network structures are equivalent. Meanwhile, to learn causal relationships between pairs of variables, patterns of dependency in the presence of a third variable must be observed in the context of interventions. Learning in Bayesian network can be used also treated as a point to estimate the parameters to average over possible model structure and parameters to provide an estimate of the posterior distribution of the variables, which is useful to avoid overfitting to the data that might be noisy, limited, incomplete and uncertain.

2.2. *Low-Order Conditional Independence*

Bayesian networks provide a natural representation for (causally induced) conditional independence which is depicted between node/variables that are independent, given the structure of the underlying DAG by testing marginal [23], low-order and full-order conditional independence. The structure learning technique based on conditional independence is generally used to derive the essential pair wise independence of the n-th variables and has become a dominant tool for analyzing full conditional independence between arbitrary variables. Full-order conditional independence is a exact set of edges between successive variable X_i and X_j given the remaining variable X_k, where X_k is the element of all variables except variable X_i and X_j. An edge between vertices i and j$(i \neq j)$ is drawn if the correlation coefficient $\rho ij \neq 0$ and no third gene can explain the correlation. The notation for full-order conditional independence can be defines as;

$$X_i \not\perp X_j | X_k \text{ for all } k \in V/\{i, j\}. \tag{3}$$

The graph encoding the above independence statements for all pairs of nodes is still undirected. It can be shown that knowing independences of all orders gives an even higher resolved representation of the correlation structure. The collection of independence statements already implies a direction of some of the edges in the graph resulting directed probabilistic model, which is called a Bayesian network. Although full-order conditional independence is a dominant technique to learn Bayesian network structure as it offers interactions between variables but full conditional relationships between random variables is difficult and complex to be estimated if the number of observations is relatively large than number of variables as in the case of high-throughput genomic data [24].

In this paper, we proposed a novel approach to reduce complexity of full-order conditional independence according to DAG $\hat{G}$ in Equation 3 used in Bayesian network structure learning by applying low-order conditional independence to implement network-based model from gene expression data. Reverse discovery of DAG $\hat{G}$ requires to determine for each variable X_i, the set of variables X_j are observed on which variable X_i is conditionally dependent given the remaining variables XP_j. As a result, we still encounter the high dimensionality problem since the number of genes p is much higher than the numbers of sample size, n. Therefore to reduce the dimension, we apply DAG $\hat{G}$ by the 1st order conditional dependence, where DAG $\hat{G}^{(q)}(q < p$ and $q = 1)$.

Unlike full-order conditional independence that take all genes that may correlate into account, low-order conditional independence only assesses the condition independence between two genes in the present of single third gene and this approach able to address issue related with high-dimensional data. Let q be smaller than p. In the qth order dependence DAG $\hat{G}^{(q)}$, whenever there exists a subsets X_Q of q variables among the set of $p - 1$ variables X_P such that X_j and X_i are conditionally independence given X_Q, no edge is execute between two successive variables X_j and X_i. DAG $\hat{G}^{(q)}$ is defined as below:

$$\begin{aligned} \forall q < p, \quad & \hat{G}^{(q)} \\ &= \Big(X, \{(X_j, X_i); \forall Q \subseteq P_j, \quad |Q| = q, \quad X_i \not\perp\!\!\!\perp X_j | X_Q\}_{i,j \in P}\Big) \end{aligned} \tag{4}$$

To such an end, we extend Bayesian network approach based on the consideration of low- order independencies introduced by Wille et al. [24] for Graphical Gaussian Model. After obtaining the 1st order conditional dependence DAGs $G(q)$ for Bayesian networks, we applied this result in the manner to allow us to approximate DAG describing full-order conditional independencies to obtain the global independence of the constructed network.

Readers are directed to Malouche [25] for extra technical detail regarding both methods.

3. Empirical Evaluation

In our study, we employed vant Veer et al. [26] dataset, to evaluate the proposed algorithms. The data of [26] was obtained from Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA (2006)). This data set contains expression profile information derived from 97 lymph node negative breast cancer patients, 55 years old or younger. Among the 97 patients, 46 developed distant metastases within 5 years and 51 remained metastases free for at least 5 years. The isolation of RNA from cancerous tissues, labeling of complementary RNA (cRNA), the competing hybridization of labeled cRNA with a reference pool of cRNA from all tumors to arrays containing 24,481 gene probes, quantization and normalization of fluorescence intensities of scanned images are detailed described in the previous publication vant Veer et al. [26]. This study aims to extract genetic markers which are related with metastases and learned genes relationship for breast cancer prognosis from 70 genes signatures.

3.1. Data Pre-Processing

The microarray data for each sample has been already preprocessed and log transformed [26]. An initial selection for this study was carried out with 70 gene signatures from the data set. Gene expression experiments can produce data sets with manifold missing expression values. Methods for imputing missing data are required to minimize the effect of incomplete data sets on analyses, and to increase the range of data sets to which these algorithms can be applied. Table 1 shows an example of several missing values in sample 54 from microarray data set. Several imputation methods have been employed to address the problem of missing data, for instance Friedland et al. [27] has used singular value decomposition (SVD) to simultaneously estimate missing data of DNA microarray by addressing the optimization problem using fixed rank approximation algorithm (FRAA). This study has compared the performance of KNNimpute and FRAA algorithms and the results indicate that KNNimpute algorithm more accurate when estimating missing entries had been deleted from the full elutriation matrix, while FRAA might be a feasible option in cases of small number of columns. Another alternative is to use local least imputation proposed by Kim et al. [28] where genes similar to the target gene with missing values are identified based on Euclidean distance or Pearson correlation coefficient. In this paper, we applied knearest neighbors (kNN) imputation method, as it is widely applied to address missing values in DNA microarray gene expression data [29, 30]. The kNN imputation method has been reported as a robust and sensitive approach to estimate missing values in microarray data set through Troyanskaya et al. [31] and Yu et al. [32] studies. We set the value of k equal to 10 to as being proposed by Yu et al. [32]. The result for kNN imputation is illustrated in Table 2. We intend to implement FRAA and local least imputation algorithms and compare its performance with KNN in a future paper.

3.2. Evaluation Results

In this section we describe our approach to analyzing network-based model using Bayesian network and low-order conditional independence. We evaluated the probability of j and X_i an edge (X_j, X_i) by measuring the conditional dependence between the variables X_i and X_j given any variable X_k.

Assuming linear dependencies, we consider the partial regression coefficient $P_{ij}|k$ defined as follows: The conditional dependence between the variables X_j and X_i given any variable X_k has been tested by using the null assumption $H_{i,jk}$, where $P_{ij|k} = 0$. To such an aim, Least Square (LS) estimator was applied to obtain the coefficient. In that case where each $k \neq j$, we compute the estimates $\hat{P}_{ij|k}$ according to LS estimator. Thus, we assign a score $S1(i, j)$ to each potential edge (X_j, X_i) equal to the maximum Max $k \neq j(P_{ij|k})$, that is the most favorable result to low-order conditional independence.

The smallest scores indicate the most significant edges for $\hat{G}(1)$. The inferred DAG $\hat{G}(1)$ contains the edges assigned a score below a chosen threshold $\alpha 1$. Fig. 2 shows the learned network relations for breast cancer prognosis with threshold $\alpha = 0.1$. This learned network reveals some genes which are highly associated in causing metastases, M. The larger nodes in the graph specify the genes which modification in gene expression level lead to a major determination on the status (e.g.: on or off) of other genes, meanwhile the green nodes denote the highly regulated genes. The four genes which are found to dominate the expression levels of other genes are; BBC3, GNAZ, TSPY-like5 (TSPY5), and DCK. The metastases variable, M and its Markov Blanket is shown in Fig. 3 comprehensively with the gene names applied where possible. Six genes has been identifies as regulator genes in controlling metastases and two genes was regulated, and one of them is cyclin E2 (CCNE2).

We also demonstrate the receiver operating characteristics (ROC) curve derived from genetics markers for breast cancer prognosis by comparing its performance with low-order and full-order conditional independence. A ROC curve obtained by varying a decision threshold can give us a direct view on how these algorithms perform at the different sensitivity and specificity levels. In Fig. 4, we plot the ROC curves for low-order and fullorder conditional independence. We observed that the low-order conditional independence outperformed the full conditional independence algorithm for determining network-based analysis in breast cancer prognosis. By following the study of vant Veer and colleagues [26], a threshold of sensitivity has been set to 90%. The corresponding specificity for low-order and full-order conditional independence is 53% and 21% respectively. We also noticed fullorder conditional independence has better true positive rate (TPR) values at beginning of the process (false positive rate (FPR) < 0.33) compared to low-order conditional independence, however this trend has been amended drastically at the point of FRP > 0.33. This is due to increase number of edges in the constructed network and the ability of low-order conditional independence to handle large number of genes as it assessed the condition independence between two genes in the present of single third gene, while full-order conditional independence took all genes that may correlated into account, which decreased the performance of specificity.

Table 1. Missing values in microarray data set

Gene	Sample 50	Sample 51	Sample 52	Sample 53	Sample 54
Gene58	−0.019	0.146	−0.217	0.275	NaN
Gene59	0.188	−0.074	−0.681	−0.081	0.097
Gene60	−0.02	0.383	−0.042	0.128	−0.173
Gene61	0.688	−0.373	−0.173	0.311	NaN
Gene62	0.263	0.074	−0.014	0.238	−0.442
Gene63	−0.357	−0.243	0.116	−0.165	−0.062
Gene64	0.238	−0.033	−0.201	−0.084	−0.385
Gene65	0.398	−0.381	0.009	−0.061	NaN
Gene66	0.569	−0.324	−0.197	0.041	−0.687
Gene67	0.107	−0.069	−0.071	−0.001	−0.324

Table 2. kNN Imputation Method.

Gene	Sample 50	Sample 51	Sample 52	Sample 53	Sample 54
Gene58	−0.019	0.146	−0.217	0.275	0.037
Gene59	0.188	−0.074	0.681	0.081	0.097
Gene60	−0.02	0.383	−0.042	0.128	−0.173
Gene61	0.688	−0.373	−0.173	0.311	−0.2454
Gene62	0.263	0.074	−0.014	0.238	−0.442
Gene63	−0.357	−0.243	0.116	−0.165	−0.062
Gene64	0.238	−0.033	−0.201	−0.084	−0.385
Gene65	0.398	−0.381	0.009	−0.061	−0.1495
Gene66	0.569	−0.324	−0.197	0.041	−0.687
Gene67	0.107	−0.069	−0.071	−0.001	−0.324

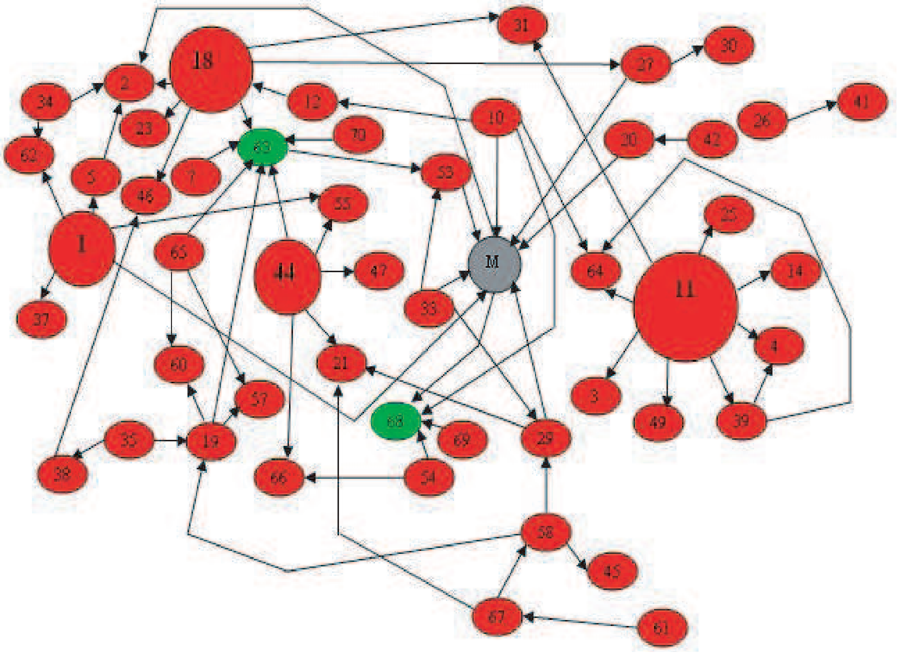


Fig. 2. Network-based model for breast cancer prognosis using Bayesian network and low-order conditional independence.

4. Discussion

We present some discussion on the optimality of genes regulator and regulated genes in the context of breast cancer prognosis. Fig. 2 showed the network learned for gene interactions analysis on breast cancer metastases. Four genes were identified as the optimal regulator in this analysis, including BBC3, GNAZ, TSPY-like5 (TSPY5), and DCK. The BBC3 gene (also known as JFY1, PUMA) is located on human chromosome 19q13.3–q13.4 and has homology to the Bcl2 family member. The biological role for BBC3 is to induce apoptosis via the mitochondrial apoptotic pathway. BBC3 is transcriptionally activated by p53 and it is also up-regulated after endoplasmic reticulum stress, independently to P53 status. The expression of PUMA was being ob-

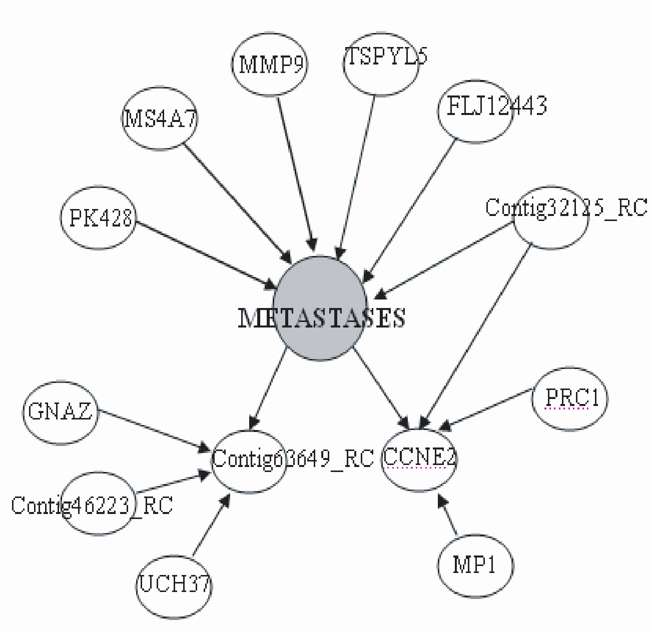


Fig. 3. Markov blanket for metastases variable.

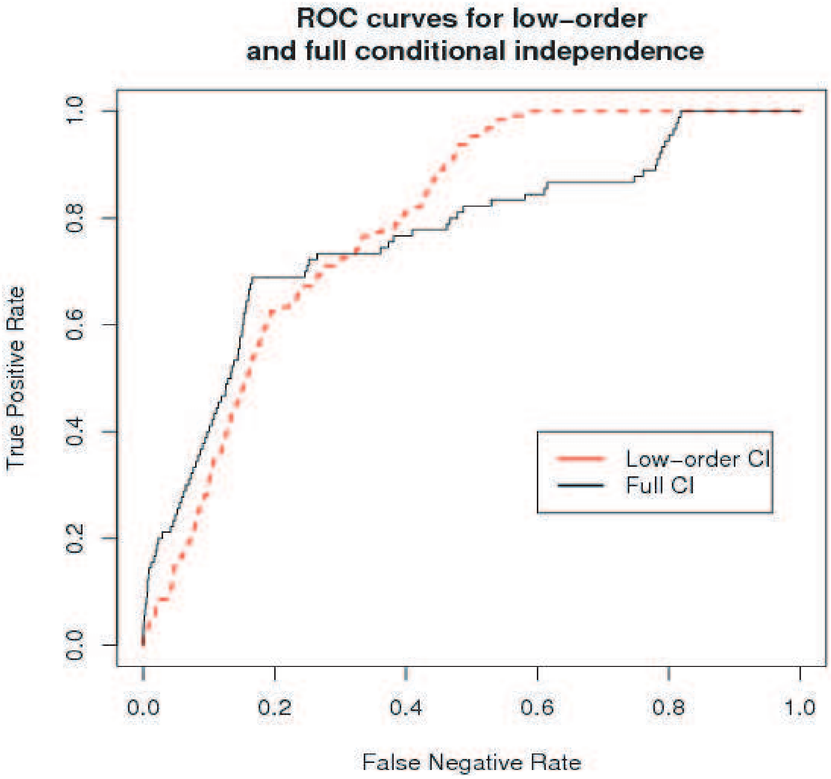


Fig. 4. ROC curves for low-order and full conditional independence.

served on human cancer cells which associated with P53 in many studies [33, 34], and recently has been reported to be associated with estrogen receptor alpha (ER) in breast cancer study [35].

On the other hand, the protein encoded by gene GNAZ (AI847979, Gz) is Guanine nucleotide-binding proteins which involved as modulators or transducers in various transmembrane signaling systems. This encoded protein may play a role in maintaining the ionic balance of perilymphatic and endolymphatic cochlear fluids. However, there is insufficient information about this protein could cause cancer. The TSPY-like 5 (TSPYL5) gene also known as KIAA1750is involved in nucleosome assembly, a process which, if destabilized, can alter the regulatory mechanisms of a cell, which is likely to occur in cancer. The TSPYL5 has been identified as a genetic marker in several studies [36, 37] and has been identified as a significant gene in luteinizing hormone (LH) [38]. Deoxycytidine kinase (DCK), in contrast is required for the phosphorylation of several deoxyribonucleosides and their nucleoside analogs. Deficiency of DCK is associated with resistance to antiviral and anticancer chemotherapeutic agents. DCK is clinically important because of its relationship to drug resistance and sensitivity. DCK gene has been used to study the resistance to chemotherapy among myeloid leukemia (AML) patients [39] and was found active in sporadic breast cancer [40].

Beside these regulator genes, two highly regulated genes have been revealed in the gene regulatory analysis; FLJ11354 and CCNE2. FLJ11354 gene was discovered by Sun et al. [41], however the exact function of this gene is still unknown. Meanwhile, CCNE2 is protein coding gene type belong to cyclin that act as regulators of Cyclin Dependent Kinase (CDK). Different cyclins exhibit distinct expression and degradation patterns which contribute to the temporal coordination of each mitotic event. This cyclin forms a complex with and functions as a regulatory subunit of CDK2 and plays a role in cell cycle G1/S transition. The expression of this gene peaks at the G1-S phase and exhibits a pattern of tissue specificity distinct from that of cyclin E1. A significantly increased expression level of this gene was observed in tumor-derived cells. CCNE2 also has been reported to be qualify as independent prognostic markers for lymph nodenegative breast cancer patients [42] and appear to have a predictive value in ER positive among breast cancer patients [43].

5. Conclusion and Future Directions

In this study, we examined Bayesian network with low-order conditional independence for network-based learning for breast cancer prognosis. We conducted evaluation study to assess the practical value of this technique in helping researchers analyze gene interaction analysis from huge amounts of gene expression data. The results indicated that the low-order conditional independence technique could enhance inference and outperformed full-order conditional independence as well as provide an alternative to cater problem associated with high dimensional microarray data. Empirical evaluations also showed the top relation learned by this technique, contained 20–30% “interesting” or “perhaps interesting” relations that had potential as experiment hypotheses for further investigation. Our general conclusion is that network-based analysis on microarray data can capture large portions of underlying gene interaction structures and it can be modeled using Bayesian network. Since learning the structure of Bayesian network is NP hard and computational complex, low-order conditional independence is an optional way. In our future research, we will perform formal analysis of the effect of the modifications we have introduced to the Bayesian network learning in comparison with other commonly technique such as heuristic search. We will also perform large-scale simulations to further verify the robustness of the network accuracy performance measures we have reported in this paper as well as verify the gene relations we obtained.

References

[1] U. B. Kjaerulff and A. L. Madsen, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer, 2008.

[2] O. Pietquin and T. Dutoit, “A probabilistic framework for dialog simulation and optimal strategy learning,” *IEEE Trans. Speech and Audio Processing*, vol. 14, pp. 589–599, 2005.

[3] J. Agosta and T. L. R. Shacter, “The structure of bayes networks for visual recognition,” *Uncertainty in Artificial Intelligence*, vol. 4, pp. 397–405, 1990.

[4] D. Heckerman, “Probabilistic similarity networks,” *Networks*, vol. 20, pp. 607–636, 1990.

[5] D. Nikovski, “Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics,” *IEEE Trans. Knowledge and Data Eng*, vol. 12, pp. 509–516, 2000.

[6] G. Bastos and K. S. Guimaraes, “A simpler bayesian network model for genetic regulatory network inference,” *Proceedings IEEE International Joint Conference on Neural Networks*, vol. 1, pp. 304–309, 2005.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1988.

[8] P. J. F. Lucas, *Biomedical Applications of Bayesian Networks*, vol. 214. Springer Berlin/Heidelberg, 2007.

[9] S. M. Maskery, J. H. H. Hu, C. D. Shriver, and M. N. Liebman, “A bayesian derived network of breast pathology co-occurrence,” *Journal of Biomedical Informatics*, vol. 41, pp. 242–250, 2008.

[10] L. Carrivick, S. Rogers, J. Clark, C. Campbell, M. Girolami, and C. Cooper, “Identification of prognostic signatures in breast cancer microarray data using bayesian techniques,” *Journal of The Royal Society*, pp. 1–15, 2005.

[11] L. d. Campos and J. Huete, “A new approach for learning belief networks using independence criteria,” *Int’l J. Approximate Reasoning*, vol. 24, pp. 11–37, 2000.

[12] D. Heckerman, D. Geiger, and D. Chickering, “Learning bayesian networks: The combination of knowledge and statistical data,” *Machine Learning*, vol. 20, pp. 197–243, 1995.

[13] S. Acid, L. M. de Campos, and J. G. Castellano, “Learning bayesian network classifiers: searching in a space of partially directed acyclic graphs,” *Machine Learning*, vol. 59, pp. 213–235, 2005.

[14] B. Sierra, E. Lazkano, E. Jauregi, and I. Irigoien, “Histogram distance-based bayesian network structure learning: A supervised classification specific approach,” *Decision Support Systems*, vol. 48, pp. 180–190, 2009.

[15] “Cancer facts and figures,” 2006. American Cancer Society.

[16] G. C. C. Lim, H. Yahaya, and T. O. Lim, “The first report of the national cancer registry cancer incidence in malaysia 2002, national cancer registry, ministry of health malaysia,” 2002.

[17] J. M. Reuben, S. Krishnamurthy, W. Woodward, and M. Cristofanilli, “The role of circulating tumor cells in breast cancer diagnosis and prediction of therapy response,” *Expert Opinion on Medical Diagnostics*, vol. 2, pp. 339–348, 2008.

[18] A. Goldhirsch, “Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer,” *J. Clin. Oncol.*, vol. 21, pp. 3357–3365, 2003.

[19] P. Eifel, J. A. Axelson, J. Costa, J. Crowley, W. J. Curran, A. Deshler, S. Fulton, C. B. Hendricks, M. Kemeny, A. B. Kornblith, T. A. Louis, M. Markman, R. Mayer, and D. Roter, “National institutes of health consensus development conference statement: adjuvant therapy for breast cancer,” *Journal of the National Cancer Institute*, vol. 93, pp. 979–989, 2000.

[20] F. Andre and L. Pusztai, “Molecular classification of breast cancer: implications for selection of adjuvant chemotherapy,” *Nature Clinical Practice Oncology*, vol. 3, pp. 621–632, 2006.

[21] E. Andreopoulou and G. N. Hortobagyi, "Prognostic factors in metastatic breast cancer: Successes and challenges toward individualized therapy," *Journal of Clinical Oncology*, vol. 26, pp. 3660–3662, 2008.

[22] E. P. Diamandis and M. K. Schwartz, "Tumor markers: Physiology, pathobiology, technology, and clinical applications," *Amer. Assoc. for Clinical Chemistry*, 2002.

[23] P. Kontkanen, P. Myllymki, and H. Tirri, "Classifier learning with supervised marginal likelihood," presented at.

[24] A. Wille and P. Buhlmann, "Low-order conditional independence graphs for inferring genetic networks," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, 2006.

[25] D. Malouche, "Determining full conditional independence by low order conditioning." In print, 2009.

[26] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. Hart, M. Mao, H. L. Peterse, K. V. d. Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.

[27] S. Friedland, A. Niknejad, and L. Chihara, "A simultaneous reconstruction of missing data in dna microarrays," *Linear Algebra and its Applications*, vol. 416, pp. 8–28, 2006.

[28] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for dna microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, pp. 187–198, 2005.

[29] P. Meesad and K. Hengpraprohm, "Combination of knn-based feature selection and knnbased missing-value imputation of microarray data," *presented at The 3rd International Conference on Innovative Computing Information and Control (ICICIC'08)*, 2008.

[30] K. Y. Kim, B. J. Kim, and G. S. Yi, "Reuse of imputed data in microarray analysis increases imputation efficiency," *Bioinformatics*, vol. 5, pp. 1–9, 2004.

[31] O. Troyanskaya, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, pp. 520–525, 2001.

[32] C. Yu, R. Zhang, Y. Huang, and H. Xiong, "High-dimensional knn joins with incremental updates," *GeoInformatica*, vol. 14, pp. 55–82, 2009.

[33] M. Lacroix, R.-A. Toillon, and G. Leclercq, "p53 and breast cancer, an update," *Endocrine-Related Cancer*, vol. 13, pp. 293–325, 2006.

[34] J.-W. Han, C. Flemington, A. B. Houghton, Z. Gu, G. P. Zambetti, R. J. Lutz, L. Zhu, and T. Chittenden, "Expression of bbc3, a proapoptotic bh3-only gene, is regulated by diverse cell [35] death and survival signals," *Proceedings of the National Academy of Sciences*, vol. 98, pp. 11318–11323, 2001.

[35] S. Tozlu, I. Girault, S. Vacher, J. Vendrell, C. Andrieu, F. Spyratos, P. Cohen, R. Lidereau, and I. Bieche, "Identification of novel genes that co-cluster with estrogen receptor alpha in breast tumor biopsy specimens, using a large scale real-time reverse transcription-pcr approach," *Endocrine-Related Cancer*, vol. 13, pp. 1109–1120, 2006.

[36] G. Alexe, S. Alexe, D. E. Axelrod, T. O. Bonates, L. I. Lozina, M. Reiss, and P. L. Hammer, "Breast cancer prognosis by combinatorial analysis of gene expression data," *Breast Cancer Research*, vol. 8, no. R41, 2006.

[37] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," *Bioinformatics*, vol. 23, pp. 30–37, 2007.

[38] V. K. Yadav, P. Muraly, and R. Medhamurthy, "Identification of novel genes regulated by lh in the primate corpus luteum: insight into their regulation during the late luteal phase," *Molecular Human Reproduction*, vol. 10, pp. 629–639, 2004.

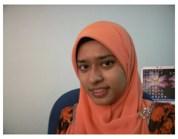
[39] M. van den Heuvel-Eibrink, E. A. C. Wiemer, M. Kuijpers, R. Pieters, and P. Sonneveld, "Absence of mutations in the deoxycytidine kinase (dck) gene in patients with relapsed and/or refractory acute myeloid leukemia (aml)," *Leukemia*, vol. 15, 2001.

[40] C. Rodriguez, L. Hughes-Davies, H. Valle's, B. Orsetti, M. Cuny, L. Ursule, T. Kouzarides, and C. Theillet, "Amplification of the brca2 pathway gene emsy in sporadic breast cancer is related to negative outcome," *Clinical Cancer Research*, vol. 10, pp. 5785–5791, 2004.

[41] Y. Sun, V. Urquidi, and S. Goodison, *Derivation of molecular signatures for breast cancer recurrence prediction using a two-way validation approach*. Springer Netherlands, 2009.

[42] A. M. Sieuwerts, M. P. Look, M. E. M. v. Gelder, M. Timmermans, A. A. C. Trapman, R. RodriguezGarcia, M. Arnold, A. J. W. Goedheer, V. d. Weerd, H. Portengen, J. M. Klijn, and J. A. Foekens, "Which cyclin e prevails as prognostic marker for breast cancer? results from a retrospective study involving 635 lymph node negative breast cancer patients,," *Clinical Cancer Research*, vol. 12, pp. 3319–3328, 2006.

[43] C. Sotiriou, M. Paesmans, A. Harris, M. A. Colozza, S. Fox, M. Taylor, A. Sorre, P. Martiat, F. Cardoso, and M. Piccart, "Cyclin e1 (ccne1) and e2 (ccne2) as prognostic and predictive markers for endocrine therapy (et) in early breast cancer," *Journal of Clinical Oncology*, vol. 22, 2004.



Farzana Kabir Ahmad is a doctoral student in the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia. She earned her MSc in Computer Science from the School of Computer Science, Universiti Sains Malaysia in 2005. Her research interest resides in biological and medical data mining and knowledge discovery and their application in bioinformatics and genomics. Mainly, she is interested in the integration of biological data from various sources, development of data-mining tools for the prediction of gene functions and their interaction in genetic data, modeling and reconstruction of gene regulatory network and identification of tumor progression.

Safaai Deris is a Professor of Artificial Intelligence and Software Engineering at the Faculty of Computer Science and Information Systems, Deputy Dean at the School of Graduate Studies, and Director of Laboratory of Artificial Intelligence and Bioinformatics at Universiti Teknologi Malaysia. His current academic interests include data mining for bioinformatics applications and the application and development of intelligent techniques in planning and scheduling. His articles have appeared in the Journal of Systems Architecture, Elsevier Science, Journal of Biomedical Informatics, IEEJ Transactions on Electrical and Electronic Engineering, International Journal of Biological and Medical Sciences and other refereed journals. He received the MEng degree in Industrial Engineering, and the DEng degree in Computer and System Sciences, from the Osaka Prefecture University, Japan, in 1989 and 1997 respectively.

Nor Hayati Othman is a Professor at Universiti Sains Malaysia, obtained her medical degree, MBBS, from University of Malaya in 1981 and Master of Pathology from University of Malaya in 1987. She is a general surgical pathologist and has undergone various sub-specialty pathology trainings; Neuropathology from the University of Western Australia in 1989; Dermatolopathology from the University of Sydney in 1994, Molecular Pathology from the University of Toronto in 1999 and the University of Cape Town in 2002. She is currently the Dean [Clinical Science Research] for Universiti Sains Malaysia (USM). Her research interests are in cancer research particularly in thyroid and cervical cancers. She has 102 publications in peer-reviewed journals, 42 past and current research projects as main or co-investigator and has supervised/co-supervised/is supervising/-co-supervising a total of 36 students at MSc and PhD levels to date. Together with electronic and electric engineers from the USM engineering campus, she invented *NeuralPap* [diagnostic software to diagnose cervical cancer], *DataPap* [management system for cervical cancer] and *Neuralmammo* [diagnostic software to diagnose breast cancer]. These 3 inventions won several awards at national and international invention competitions. She is currently a Council member for the College of Pathologist, Academy of Medicine, Malaysia, and a member of the Editorial Board of the Medical Journal of Malaysia.