Bayesian Networks for Elucidating and Integrating Breast Cancer Knowledge

Farzana Kabir Ahmad
Graduate Department of Computer Science, College of Arts and Sciences
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
phone: +6-04-9284743, fax: +6-04-9284753
Email: farzana58@uum.edu.my

Safaai Deris
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
Email: safaai@utm.my

Nor Hayati Othman
Clinical Research Platform & Pathologist, Health Campus Universiti Science Malaysia
16150 Kubang Kerian, Kelantan, Malaysia
Email: hayati@kb.usm.my

ABSTRACT

Bayesian networks are able to provide a suitable framework for the integration of highly heterogeneous experimental data and domain knowledge. In addition, it can produce interpretable and understandable models for knowledge discovery within complex domains by providing knowledge of casual and other relationships in the data. Accurate prognosis of breast cancer can spare a significant number of breast cancer patients from receiving unnecessary adjuvant systemic treatment and its related expensive medical costs. Recent studies have revealed the potential value of gene expression signatures in examining the risk of disease recurrence. However, most of these studies attempt to implement genetic marker based prognostic models as substitute to existing clinical breast cancer criteria. Clinical data such as grade, lymph node involvement and patient history are frequently underused to guide the cancer clinical management in the presence of microarray data. Given the complexity of breast cancer prognosis, which relies on numerous factors a more practical approach would be to employ both clinical and microarray data that may be complementary. This paper discuss the implementation of integrated modeling approaches to interrogate the massive amount of data being produced in microarray and clinical data to extract useful and valuable answers, which could help physicians to make an appropriate therapeutic decision making.

Key Words: Bayesian networks, breast cancer, prognosis, clinical data

1. Introduction

Bayesian networks are a popular class of graphical probabilistic models for research and application in the field of Artificial Intelligence. Bayesian networks are motivated by Bayes' theorem [1] and are allowed to represent a joint probability distribution over a set of variables. Once known, this joint distribution can be used to calculate the probabilities of any configuration of the variables. Bayesian networks have become the prominent technique in biomedical research since it is especially suited for capturing and reasoning with uncertainty data. This paper focus on the application of Bayesian networks in breast cancer domain.

A growing worldwide health problem is the increased number of those suffering from breast cancer. 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 [2]. 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 [3] 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. This high deaths rate has 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 [4]. In numerous patients, minuscule clinically evident metastases 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 metastasis 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 [5] and NIH [6] 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 [7-9]. Although clinical and histopathological data are proven to be relevance to predict breast cancer metastasis, gene cellular proliferation is also essential information that needs to be taken into consideration since breast cancer is a complex and heterogeneous disease which relies on both, clinical and genetic factors.

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 clinical and gene information in breast cancer invasion and metastasis, an integrated modeling approach with microarray and clinical data is described here. Bayesian network has been proposed as a method to develop an integrated 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 Bayesian network method that has been applied to integrated microarray and clinical 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. The description of data that utilized in this study as well as the pre-processing technique employed is elucidates in section 3. Subsequently, the results and discussion are presented in section 4 and lastly, section 5 offers concluding and future direction remarks.

2. Methods

2.1 Bayesian Networks

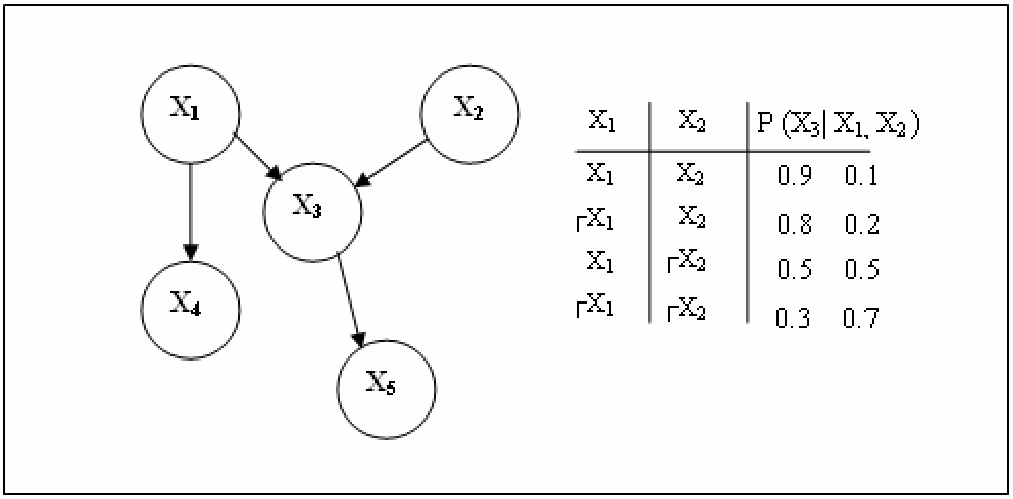
2.1.1 Definition

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 directed acyclic graph (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 Conditionality Distribution Table (CPT). Figure 2 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) = { $X_1, \ldots\ldots X_n$} and arcs A (G) $\subseteq$ V(G) x 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, or V may present a clinical factor such as lymph node, while X is present/ absent of lymph node.

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:

$$P(X_1, \ldots\ldots X_n) = \prod_{i=1}^{n} P\,(X_i \mid \text{parents}\,(X_i)) \quad (1)$$



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

Figure 2: DAG and CPT

The joint distribution of Figure 2 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.

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)\,P(X_3 \mid X_1, X_2)\,P(X_2)\,P(X_4 \mid X_1)\,P(X_5 \mid X_3) \quad (2)$$

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.

2.2 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 (as explained in Figure 2) or it may also be continuous and be modeled by a probability density function, commonly Gaussian distributions are used

Structure learning, on the other hand is a learning of network construction from data. When the structure of Bayesian network is unknown, which mean it cannot be specified by prior knowledge, a heuristic search can be implemented to find a ‘good’ structures. 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.

3. Data

A computational study is executed in this study using Bayesian network for breast cancer prognosis with microarray and clinical data. The data of [10] was obtained from Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA (2006)). This data set contains expression profile information derived from 78 lymph node negative breast cancer patients. These samples belong to 34 patients who had developed distant metastases within 5 years and 44 remained metastases free for at least 5 years. Each record for patient also carried the clinical information namely age, tumor diameter, grade, estrogen and progesterone receptors status, the presence of angioinvasion and lymphocytic infiltration, which together produce the clinical data. The aim of this paper is to implement a computational model to accurately predict the risk of distant recurrence of breast cancer.

3.1 Data Pre-processing

The microarray data for each sample has been already pre-processed and log transformed [10]. 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 several missing values from sample 54 for microarray data set. This paper applied k nearest neighbors (kNN) imputation method with k= 10 to address missing values. The kNN imputation method is the most robust and sensitive approach to estimate missing values in microarray data set. It is proven to be prominent and effective method through Troyanskaya et al.'s research [11]. The result for kNN imputation is illustrated in Table 2.

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

4. Results and Discussion

A computational model to predict the breast cancer prognosis using 70 genes signatures and clinical data has been developed using Bayesian networks as described in section 2.1. The structure learning for this model has been constructed by using conditionally Gaussian distribution. The highest log network score obtained is 95.46 with five genes and 3 clinical data (grade, angioinvasion and prognosis) being the main factors for metastases to occurred.

Histological grading of tumors has been shown in various studies as essential prognostic information in breast cancer prognosis [12]. The grade represents a morphological assessment of the degree of differentiation of the tumor as evaluated by the percentage of tubule formation, the degree of nuclear pleomorphism and the presence of mitoses. Grade 1 tumors correlate with low risk of metastases, meanwhile grade 2 tumors have an intermediate risk of metastases and grade 3 tumors have a high risk of metastases. Thus, patient with grade 1 tumors have a low rate of post-surgical recurrence compare those with grade 2 and 3 tumors. On the other hand, lymph vascular invasion, which is related to angioinvasion has been proven as a crucial prognostic factor in the lymph node negative metastatic cascade. Therefore, the observation of 3 or more tumor cell emboli in tumor-associated vessels has been correlated with poor prognosis in patients with breast cancer [13]-[14]. The five genetic markers which found to be significant are AL080059 (Gene1), CEGP1 (Gene2), PRAME (Gene5), FLJ12443 (Gene3), Contig35251_RC (Gene4). A receiver operating characteristic (ROC) curve obtained by varying a decision threshold can give us a direct view on how a classifier performs at the different sensitivity (true positive rate) and specificity (false positive rate) levels. Figure 3 illustrates the ROC for integrated clinical and gene expression data with specificity 60%.

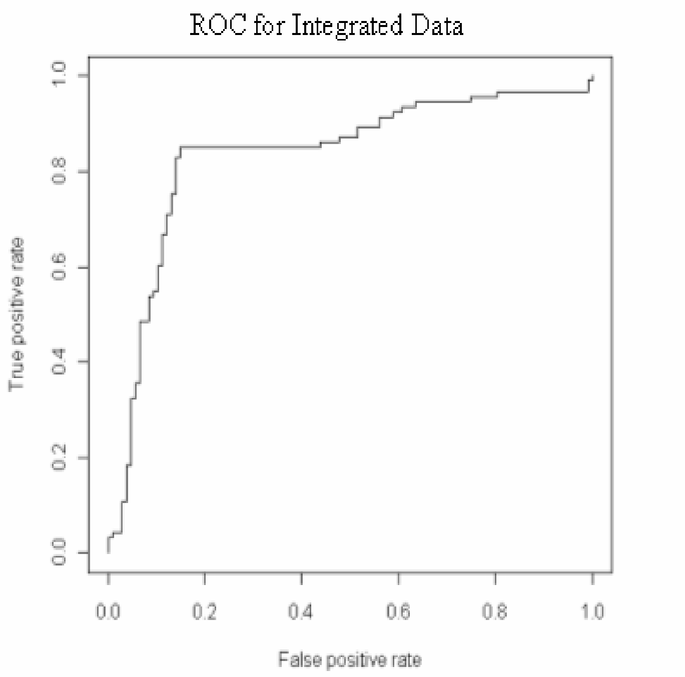


Figure 3: ROC curve for integrated clinical and gene expression data

5. Conclusion

This paper described an integrated model to predict the likelihood of disease recurrence and metastases in breast cancer using Bayesian network. Our preliminary result has shown that microarray and clinical data can provide significant improvement in breast cancer prognosis. Informative discussion on the issue of Bayesian network in the context of breast cancer prognosis also has been described in detail. It is believed that biologists and oncology community are facing great challenge to prognosis breast cancer metastasis since it is the outcome of various interactions between enormous factors. Therefore this study attempt to combine both data types in order to obtain accurate prognosis. To fully address the matter of what is the pre-eminent can be performed in breast cancer prognosis given all available information, larger-scale computational studies which involve different learning algorithms and more patient data are under progress in our laboratory.

6. *References*

[1] T. Bayes, "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philosophical Trans. Royal Soc. of London*, 1763.

[2] American Cancer Society, "Cancer Facts and Figures," 2006.

[3] 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.

[4] 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.

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

[6] P. Eifel et al., "National Institutes of Health consensus development conference statement: adjuvant therapy for breast cancer," *J. Natl. Cancer Inst*, vol. 93, pp. 979–989, 2000.

[7] 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.

[8] 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.

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

[10] van't Veer L. J., H. Dai, van de Vijver M. J., 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.

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

[12] C. Elston et al., "Pathological prognostic factors in breast cancer. I.The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, pp. 403–410, 1991.

[13] de Mascarel et al, "Obvious peritumorous emboli: an elusive prognostic factor reappraised: multivariate analysis of 1320 node-negative breast cancers," *Eur. J. Cancer*, vol. 34, pp. 58–65, 1998.

[14] S. Pinderet al., "Pathological prognostic factors in breast cancer. III. Vascular invasion: relationship with recurrence and survival in a

large study with a long-term follow-up.," *Histopathology*, vol. 24, pp. 41–47, 1994.