Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 1094–1100

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT-2010

Synergy network based inference for breast cancer metastasis

Farzana Kabir Ahmad [a] *, Safaai Deris[b] and Mohd. Syazwan Abdullah[c]

[a, c] *Graduate Department of Computer Science, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.*
[b] *Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.*

Abstract

Breast cancer is a world wide leading cancer and it is characterized by its aggressive metastasis. In many patients, microscopic or clinically evident metastases have already occurred by the time the primary tumor is diagnosed. Chemotherapy or hormonal therapy reduces the risk of distant metastasis by one-third, but it is estimated that about 70% to 80% of patients receiving treatment would have survived without it. Therefore, being able to predict breast cancer metastasis can spare a significant number of breast cancer patients from receiving unnecessary adjuvant systemic treatment and its related expensive medical costs. Current studies have demonstrated the potential value of gene expression signatures in assessing the risk of post-surgical disease recurrence. However, most of these studies attempt to develop genetic marker-based prognostic systems to replace the existing clinical criteria, while ignoring the rich information contained in established clinical markers. Clinical markers, such as patient history and laboratory analysis, which are the basis of day-to-day clinical decision support, are often underused to guide the clinical management of cancer in the presence of microarray data. As a result, given the complexity of breast cancer prognosis, we proposed a novel strategy based on synergy network that utilize both clinical and genetic markers to identify the potential hybrid signatures and investigate their interactions which are associated with breast cancer metastasis. In this study, a computational method is performed on publicly available microarray and clinical data. A rigorous experimental protocol is used to estimate the prognostic performance of the hybrid signature and other prognostic approaches. The hybrid signature performs significantly better than other methods, including the 70-gene signature, clinical makers alone and the St. Gallen consensus criterion. At 90% sensitivity level, the hybrid signature achieves 77% specificity, as compared to 53% for the 70-gene signature and 43% for the clinical makers. The predicted results also showed a strong dependence of regulator genes that are related to cell death in cell development process. These significant gene regulators are useful to understand cancer biology and in producing new drug design.

Keywords: Synergy network; Bayesian network; breast cancer metastasis; inference; conditional independence.

1. Introduction

Breast cancer is a leading cause of cancer-related death and among one of the most aggressive metastasis disease worldwide. The growing mortality rate, with 410,000 deaths each year has yield more than 1.6% of all women deaths worldwide [1]. The major clinical problems of breast cancer are the recurrence of disseminated disease and metastatic behavior. In numerous patients, miniature or clinically evident metastases have already occurred by the time the primary tumor is diagnosed. Although, treatments such as chemotherapy and endocrine therapy could reduce the risk of distant metastasis by approximately one-third, however it is predicted 80% of patient would have survived without receiving these treatments. Being prescribed with highly expensive medicines which turn out to be unnecessary has caused several complications and exacerbates the condition of breast cancer patients. As the results, the study of tumor progression and breast cancer metastasis has become a great interest in biomedical field.

1877-0509
doi:10.1016/j.procs.2010.12.178

Despite significant advances in the treatment of primary breast cancer and enormous studies that have been conducted, the ability to infer the metastatic behavior of tumors remains one of the most clinical challenges in oncology. The main cause for this setback is the complex interactions in the cancer progression and metastasis formation. In early days, three commonly used treatment guidelines such as TNM (tumor, lymph nodes and metastasis) Tumor Staging System, St. Gallen and NIH (National Institute of Health) consensus criteria have been used to determine the distant metastases. These breast cancer indices are based on clinical markers such as tumor size, lymph node involvement, patient age and the aggressiveness of the cancer founded on histopathological parameters. Regardless of the prominent practiced of these indices, it provides inaccurate results in predicting therapy failure with only 10% specificity at 90% sensitivity level. Thus, a more accurate prognostic criterion is urgently needed to avoid unnecessary treatment in newly diagnosed patients.

Recently, the development of genetic marker-based prognostic system has become a breakthrough in cancer progression research and most studies concentrated their efforts solely on this approach. Yet, some researchers do believe that the application of gene expression data to infer cancer progression is often overused in the presence of clinical data [2]. Clinical data which has been used on daily basis has been neglected and the rich information contained in established clinical markers has been ignored. Given the complexity of breast cancer metastasis, a more practical and sensible strategy is to incorporate both clinical and genetic markers that may contain complementary information.

A small number of studies have been conducted to determine the possibility of integrating clinical and genetic markers to infer breast cancer metastasis [3, 4]. While some of these approaches show a great promise in incorporating two different markers to infer cancer metastasis, the issue of high dimensionality data has rarely been discussed. One important characteristic of microarray data is the extremely large amount of data in a very small sample size. Thus, by integrating two markers to infer cancer metastasis could be computationally complex as large number of variables may need to be examined. In this paper, we seek to improve the ability to infer breast cancer metastasis using a novel strategy known as synergy network. This method is solely based on Bayesian network that apply two different approaches: an information-theoretic approach and conditional independence approach. Our keen interest is to obtain correctly learnt network in order to examine the two markers, clinical and genetic markers, in the presence of a third variable which represent the state of the cell (metastasis). In addition, we offered scoring markers interactions that provide insights into the tumor progression and indicate markers that highly regulate breast cancer metastasis.

The reminder of this paper is organized as follows. Section 2 describes the method used to develop synergy network based on Bayesian network to integrate two diverse markers. This section also elaborates the approaches taken to implement correct learnt structure learning in order to address the issue of high dimensional data. The empirical results and discussion are presented in Section 3, while Section 4 provides concluding remarks.

2. Methods

2.1 Bayesian network

Bayesian network is a probabilistic graphical model in which vertices represent random variables and the absence of an edge between two vertices represents conditional independence. Consider a finite set $V_n = \{X_1, X_2, \ldots, X_n\}$ of random variables. Bayesian network representation contains two components: a directed acyclic graph (DAG), $G = (V_n, E_G)$ which vertices correspond to random variables, and conditional probability distributions of the random variables, given its dependent variables (parents) in G. The joint distribution of these conditional probability distributions is defined as follows:

$$P(X_1,...,X_n) = \prod_{X_i \in V_n} P(X_i \mid Pa(X_i)) \qquad (1)$$

where $P(X_i \mid Pa(X_i))$ is a set of conditional probabilities for each variables X_i and $Pa(X_i)$ is the set of variables which are the parents of X_i in graph G.

We use an example to illustrate the basic idea of Bayesian networks. Given a Bayesian network specified in Fig. 1 for 5 genes: X_1, X_2, X_3, X_4, and X_5, this structure specifies the parents for genes X_3, X_4, and X_5: $Pa(X_3)=\{X_1, X_2\}$, $Pa(X_4)=\{X_1\}$, $Pa(X_5)=\{X_3\}$, where Pa(V) represents the parent vertex set for vertex V.

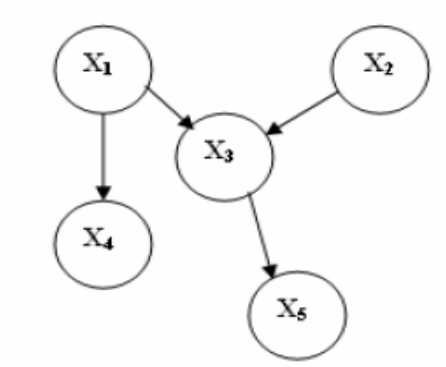

Fig. 1: A simple Bayesian network representation that explicates relationships between five genes

In our context, it can be interpreted that when genes in Z are at fixed expression levels, expression levels of genes in X do not give any information on the expression levels of genes in Y and vice versa. Once the structure of G is specified for a set of genes, we can interpret a directional edge from X to Y in G as a statement that X is the "cause" of Y, or the expression level of X has an effect on the expression level of Y. Therefore, obtaining the correct structure of Bayesian network is essential to perform an efficient inference and correctly represent the dependency relationship.

Determining the optimal network through Bayesian learning structure has been investigated for many decades. In conjunction with the invention of microarray technology, the problem of searching the best fit network given the datasets have become harder. It is due to the fact that analyzing these high-dimensionality data require a large number of variables to be analyzed, which yield to exponential growth in searching space, known as NP hard. Generally there are two approaches to learn the structure of Bayesian networks from data: the search and scoring methods and dependency analysis methods. In the first approach, the learning problem is viewed as searching for a structure that best fits the data. Different scoring methods have been applied to determine the fit between the network structure and the data, including Bayesian scoring, entropy-based, and minimum description length, among others. The dependency analysis approach, on the other hand tries to discover from data the dependencies among variables and then use these dependencies to construct the network structure. Lately, the dependency analysis approach is discovered to be more efficient than the search and scoring approach for sparse networks (the number of edges in the graph is relatively small). In attaining our goal, to infer breast cancer metastasis by integrating clinical and genetic markers, in this paper we proposed a new strategy to find the optimal structure. In the following section 2.2, we introduce the main steps of the proposed method.

2.2 Synergy network based on information-theoretic approach and conditional independence

Clinical markers and gene expression profiles play an important role in determining breast cancer metastasis. Integrating and analyzing all this information to discover factors that regulate cancer progression require network-based algorithm. In this study, we employed a novel strategy known as synergy network to achieve the objective. This method is developed solely based on Bayesian network that rely on two structure learning algorithm: information-theoretic approach and conditional independence. Our method generally has two different features. In the first feature, the synergy network is implemented using mutual information that measures the cooperative effect of two variables on the state of a third. The two variables in this case are genes and clinical markers (G_i *and* C_i), and the third state is the binary state variable representing the occurrence of metastasis (M) [M = 1 (Metastasis present) and M = 0 (Metastasis not present)]. Mutual information is used to decide which interactions (edges) are more prominent than the others. Mutual information between random variables X and Y is defined as follows:

$$I(X;Y) = H(X) - H(X \mid Y) \qquad (2)$$

where $H(X)$ is the entropy of X and $H(X \mid Y)$ is conditional entropy of X given Y. By applying the same rule to our study, we formulaically calculated the integration of genes and clinical markers (G_i *and* C_i) as:

$$I(G_i, C_i; M) = I(G_i; M) + I(C_i; M) \quad (3)$$

where $I(G_i, C_i; M)$ is the cooperative effect of two variables and $I(G_i; M) + I(C_i; M)$ is the individual effect.

The proposed algorithm starts from a non-connected network, whereby there is still no edge involved between nodes. Then, we calculate the mutual information for two nodes from the network and based on these mutual information values the edges is ordered. Sequentially, connection between nodes is drawn according to mutual information ordered value, which also offer the highest scoring interactions values. The edges which has the mutual information value less than threshold (threshold = 0.1) are excluded as candidates of correct edges. We only choose edges that have higher values (> threshold) to be the correct edge as it contains better probability of connection.

In the second feature, we further examined the learning structure of constructed network (obtained from the first feature) by using conditional independence approach. Two variables, for instance A and B may have different structures, $A \longrightarrow B$ and $A \longleftarrow B$ but carrying the identical mutual information values. Thus, to overcome this issue, in this algorithm, conditional independence is used to search edges that are incorrect in a triangular structure as depicted in Fig. 2. Conditional independence is defined as follows:

$$P(X_i, X_j \mid X_k) = P(X_i \mid X_k) P(X_j \mid X_k) \quad (4)$$

Hence, once we detected the edge create a triangle loop and hold the same mutual information value, all three edges included in the triangle will be run based on equation and we used the result of these test to update the network.

Fig. 2: Triangular structure of three nodes and two edges

2.3 *Dataset and Pre-processing*

The proposed method is tested and analyzed on van't Veer et al. [5] dataset, which 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 and associated clinical information including age, tumor size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor (ER) and progesterone receptor (PR) status, which all together form clinical markers. Prior the implementation of proposed method, the missing values present in this dataset was addressed. Manifold missing gene expression values is a common problem in microarray dataset. K-nearest neighbors (kNN) imputation method with k = 10 was used to handle these missing values. The kNN imputation method is utilized as it 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 [6]. Subsequently, the processed dataset was used as an input in the proposed method.

3. Results and Discussion

The proposed method was executed on the breast cancer dataset to obtain insights into the cancer development and how various factors may trigger metastasis progression, producing a synergy network as shown in Fig 3. Additionally, the top ten scoring interactions for this particular network are given in Table 1. The learned network reveals a group of genes and clinical markers which are primarily associated with causing metastasis, M. The larger nodes in the graph specify the genes when expressed at different levels lead to a major effect on the status of other genes (e.g., on or off)/clinical marker, and the light-shaded nodes denote highly regulated genes. Four genes that are found to regulate the expression levels of other genes are: BBC3, GNAZ, TSPY-like5 (TSPY5), and DCK. Two genes are highly regulated: FLJ11354 and CCNE2. Meanwhile, angioinvasion has been identified as strong factor in causing breast cancer metastasis. This network involved 50 genes and 6 clinical markers which are closely associated with breast cancer metastasis, M.

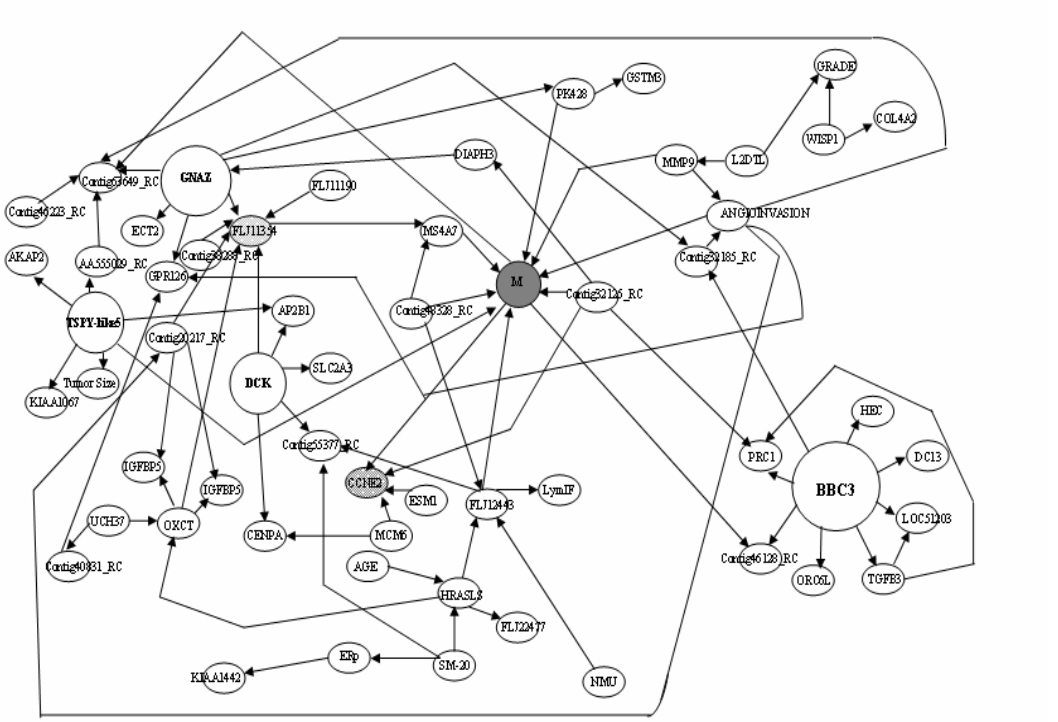

Fig. 3: Synergy network for breast cancer metastasis

The constructed network indicates that the BBC3 gene has a prominent role in regulating others genes. Eight genes are correlated with BBC3. The BBC3 gene, also known as PUMA is activated by the tumor suppressor p53, which is a key regulator of apoptosis and tumorigenesis in breast cancer. On the other hand, there is insufficient information about whether GNAZ could directly regulate the progression of breast cancer, however we discovered that it has an essential role in cellular processes of the nervous system [7]. Meanwhile, TSPYL5 has been identified as a genetic marker for breast cancer in several studies [4, 8]. Lastly, DCK is revealed to be associated with resistance to antiviral and anticancer chemotherapeutic agents, therefore this gene is clinically important because of its relationship to drug resistance and sensitivity. Outside of these regulator genes, two additional highly regulated genes have been identified in the analysis of our proposed method: FLJ11354 and CCNE2. The FLJ11354 gene was discovered by Sun et al. [9], while CCNE2 has been reported to qualify as independent prognostic markers for lymph node–negative breast cancer patients [10]. From the clinical markers point of view, angioinvasion is identified as critical factor that yield to cancer progression compared to other clinical markers

We then further evaluated the performance of constructed synergy network using a receiver operating characteristic (ROC) curve obtained by varying a decision threshold, which can provide a direct view on how this inference network performs at the different sensitivity and specificity levels. By following the study of van't Veer and colleagues [5], a sensitivity is set equal to 90%. The corresponding specificities are computed and reported in Table 2. For the purpose of comparison, the specificities of the TNM Tumor Staging System, St. Gallen and NIH consensus criteria are also compared.

Table 1: The top ten scoring interactions. No. Rel indicates the number of relation involved while Pred referred to predictor genes.

No. Rel	Pred	Target	Score
1	Metastasis	Contig63649_RC	0.000487
2	Metastasis	CCNE2	0.001081
3	UCH37	Contig40831_RC	0.003260
4	BBC3	PRC1	0.006655
5	BBC3	ORC6L	0.014038

6	WISP1	COL4A2	0.014892
7	DIAPH3	GNAZ	0.015032
8	MCM6	CCNE2	0.015653
9	DCK	Contig55377_RC	0.017289
10	HRASLS	FLJ22477	0.017314

Table 2: Inference network at sensitivity of 90%

Methods	Specificity	AUC	std
NIH 2000	0%	0.61905	0.16234
TNM Staging System	18%	0.71429	0.12747
St. Gallen	43%	0.73810	0.36204
Genetic markers	53%	0.79203	0.03245
Clinical and genetic markers (hybrid signatures)	**77%**	**0.86438**	**0.02928**

We observed that the St. Gallen criterion significantly outperformed both the TMN staging system and the NIH 2000 consensus, whereas the latter approach (the NIH 2000 consensus) was worse than the TNM staging system. The St. Gallen criterion achieved a specificity of 43%, while the TNM staging system and the NIH 2000 consensus obtained a specificity of 18% and 0%, respectively. This result is consistent with previous reports in the literature [11] whereby the specificity of the St. Gallen criterion outperforms the other clinical indices. On the other hand, the clinical and genetic markers (hybrid signatures) improve the specificities of the genetic markers and the clinical markers (St. Gallen criterion) approximately by 20%-30%. We point out that our estimation of the specificity of the 70-gene signature is worse than that reported in [5](43% versus 73%), but is consistent with that in the follow-up validation done on a larger dataset [12] (53%). Furthermore, we measured the area under curve (AUC) for all five methods, where the highest AUC suggesting a better inference network. Therefore, our results have shown that clinical and genetic markers improved the specificity of inference network compared to network those based on genetic and clinical marker alone.

4. Conclusion

Understanding the breast cancer progression network structure reveals the inherent biological information flow and interactions of various factors which will lead to more effective therapies and disease treatments. In this paper, we applied computation model which was implemented based on synergy network to study the breast cancer metastasis using genetic and clinical markers. Different genes and clinical markers were found to have high correlation in causing metastasis. For future work, we intend to validate our discovery by using biological knowledge. This attempt could arm biologist with information regarding up-stream and down-stream of gene mechanisms, which further enlighten the interactions in tumor progression.

References

[1] P. Boyle and B. Levin, "World cancer report 2008," International Agency for Research on Cancer, World Health Organization 2008.

[2] P. Edén, C. Ritz, C. Rose, M. Fernö, and C. Peterson, "“Good Old” clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers," *European Journal of Cancer*, vol. 40, pp. 1837-1841, 2004.

[3] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, pp. e184–e190, 2006.

[4] 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.

[5] L. J. van't Veer, H. Dai, M. J. van De Vijver, Y. D. He, A. M. Hart, M. Mao, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530 - 536, 2002.

[6] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, et al., "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, pp. 520–525, 2001.

[7] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Research*, pp. 1-6, 2009.

[8] G. Alexe, S. Alexe, D. E. Axelrod, T. O. Bonates, Lozina, M. Reiss, et al., "Breast cancer prognosis by combinatorial analysis of gene expression data," *Breast Cancer Research*, vol. 8, pp. R41, 2006.

[9] Y. Sun, V. Urquidi, and S. Goodison, "Derivation of molecular signatures for breast cancer recurrence prediction using a two-way validation approach," in *Breast Cancer Research Treatment*: Springer Netherlands, 2009.

[10] A. M. Sieuwerts, M. P. Look, M. E. Gelder, M. Timmermans, A. A. C. Trapman, R. RodriguezGarcia, et al. "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.

[11] C. Lohrisch, J. Jackson, A. Jones, D. Mates, and I. A. Olivotto, "Relationship between tumor location and relapse in 6,781 women with early invasive breast cancer," *Journal of Clinical Oncology*, vol. 18, pp. 2828-2835, 2000.

[12] M. J. van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. M. Hart, D. W. Voskuil, and et al., "A gene-expression signature as a predict of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, pp. 1999-2009, 2002.