# Research Trends in Microarray Data Analysis: Modelling Gene Regulatory Network by Integrating Transcription Factors Data

Farzana Kabir Ahmad* and Siti Sakira Kamaruddin

*Computational Intelligence Research Cluster,*
*Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia*
**E-mail: farzana58@uum.edu.my*

## ABSTRACT

The invention of microarray technology has enabled expression levels of thousands of genes to be monitored at once. This modernized approach has created large amount of data to be examined. Recently, gene regulatory network has been an interesting topic and generated impressive research goals in computational biology. Better understanding of the genetic regulatory processes would bring significant implications in the biomedical fields and many other pharmaceutical industries. As a result, various mathematical and computational methods have been used to model gene regulatory network from microarray data. Amongst those methods, the Bayesian network model attracts the most attention and has become the prominent technique since it can capture nonlinear and stochastic relationships between variables. However, structure learning of this model is NP-hard and computationally complex as the number of potential edges increase drastically with the number of genes. In addition, most of the studies only focused on the predicted results while neglecting the fact that microarray data is a fragmented information on the whole biological process. Hence, this study proposed a network-based inference model that combined biological knowledge in order to verify the constructed gene regulatory relationships. The gene regulatory network is constructed using Bayesian network based on low-order conditional independence approach. This technique aims to identify from the data the dependencies to construct the network structure, while addressing the structure learning problem. In addition, three main toolkits such as Ensembl, TFSearch and TRANSFAC have been used to determine the false positive edges and verify reliability of

regulatory relationships. The experimental results show that by integrating biological knowledge it could enhance the precision results and reduce the number of false positive edges in the trained gene regulatory network.

***Keywords****: Gene regulatory network, Bayesian Network, Heterogeneous data, Transcription Factors*

## INTRODUCTION

The invention of microarray technology can be considered as the latest technological breakthrough in molecular biology. Microarray experiments allow expression levels of thousands of genes to be monitored at once to provide complete transcription information in the cells. This revolutionized approach has provided a large amount of data from which a lot of knowledge can be explored. Despite the achievement of microarray technology that constantly improves laboratory methods and is prominently being used in biological researches, the major advances of the field is actually derived from the enhanced analysis methods. Due to its high-throughput nature, microarray data usually pose several challenges in terms of data analysis. Thus, computational approaches are generally necessary to divulge the molecular mechanism of cancerous cell and gain holistic view of how all these genes interact. Microarray data analysis generally consists of two major parts, namely the initial stage and the exploratory data analysis stage. The initial stage function is to prepare the raw data for rigorous analyses, as well as playing a crucial role to avoid any key factor that may affect subsequent results. On the other hand, the exploratory data analysis is an approach used to examine microarray data for the purpose of answering research questions.

In conjunction with this invention, network-based classification approach has been used in identifying gene markers that present the maximum discrimination power between cancerous and normal cells. Hence, identifying gene regulatory network has been an impressive research goal and new trend in computational biology. This interesting topic has become one of the vital research areas in microarray data analysis. Gene regulatory network (GRN) is a set of molecular components that includes genes, proteins and other molecules, which collectively accomplish cellular

functions as these molecules interact with each other [1]. The fundamental idea behind GRN analysis is to discover regulator genes by examining gene expression patterns. Notably, some genes regulate other genes, which mean that the amount of a gene expressed at a certain time could activate or inhibit the expression of another gene. Thus, changes in the expression levels of particular genes across a whole process, such as response to certain treatments would provide information that allows reconstruction of GRN using reserve engineering technique. Such data-driven regulatory networks analysis ultimately offers clearer understanding of the genetic regulatory processes, which are normally complex and intricate. Furthermore, it would bring significant implications in the biomedical fields and many other pharmaceutical industries.

Numerous studies have reported that GRN can possibly assist researchers in suggesting and evaluating innovative hypotheses in the context of genetic regulatory processes [2-3]. Various mathematical and computational methods have been used to model GRN from microarray data, including Boolean network, pair-wise comparison, differential equations estimation, Bayesian network and other techniques. Amongst these, the Bayesian network model attracts the most attention and has become the prominent technique because it can capture linear, nonlinear, combinatorial, stochastic and casual relationships between variables. Compared to other methods, Bayesian network model establishes considerable relationships between all genes in the system. Thus, Bayesian network is used in this study to analyze gene regulatory processes and to model gene relationships for breast cancer metastasis. However, the structure learning of Bayesian network is NP hard.

Furthermore, most of computational methods that were used in classification approach mainly focused on prediction and/or performance results while neglecting the interaction among genes that determine the disease phenotype. Additionally, microarray data only provides fragmented information regarding the whole biological processes. As a result, combining microarray with other biological knowledge is important in order to attain better understanding of cancer-related process and improve predictive power of inference model. Based on above trends and gaps, this study has developed a network-based inference model from gene expression data. Bayesian network has been recognized as an outstanding method to model GRN.

However, structure learning of this model is NP-hard and computationally complex as the number of potential edges increase drastically with the number of genes. Thus, low-order conditional independence method has been proposed to cater the high-throughput data. Although, the proposed method has significantly outperformed compare to other methods and had achieved better performance, the inferred network may still be deficient in terms of biological knowledge. Furthermore, reliability issues and false positive edges are other problems. For these reasons, data integration such as transcription factors has appeared as a right way to revise and check on regulatory relationships. In addition, biological knowledge is combined to further improve the proposed network.

The remainder of this paper is organized as follows. Section 2 describes some previous works, in which different kinds of biological data were utilised to achieve better construction of GRN. Section 3 on the other hand, presents the proposed method. Section 4 meanwhile presents experimental results and discussion. Finally, Section 5 offers concluding and future direction remarks.

## RELATED WORKS

In recent years, the increasing amount of genomic data such as gene expression and proteomic with publicly available databases is another trigger that has opened a new promising approach to combine various data types. Although these data varies in term of size, formats and types, it provides different, partly independent and complementary information on the whole genome. In principle, if cellular biology knowledge is complete, one can infer the genomic interactions given the activity of each molecule at a time. Unfortunately, such network is not yet available for any cell type [4]. Thus, the best option is to integrate diverse biological data that presents fragmented information and seek a better explanation for the development at a system level.

Aligned with these motivations, several studies have combined different types of data to obtain comprehensive network [5-6]. Mainly, there are two different categories to combine data: (1) homologous data integration and (2) heterogeneous data integration. Homologous data integration is

defined as the use of similar data type (for example combination of multiple microarray datasets from different studies). Meanwhile, the latter categories integrate different data types across or within studies to seek for better clarify of information provided by a single data type.

The main idea for homologous data integration is to increase the number of samples to address the issue of high dimensional data. Most studies in homologous data integration have focused on comparing two or more related datasets to identify significant genes that can distinguish different group of samples (e.g. disease and normal samples). For example, [7] have combined multiple microarray datasets to classify common transcription profiles that are universally activated in most cancer types. Generally two main methods have been used in combining homologous dataset namely a) meta-analysis method and b) effect size method.

Unlike homologous data integration, where it used similar data types, heterogeneous data integration mainly focuses on applying various data sources to ensure the reliability of results obtained. Among the popular data integration is gene expression and proteomic data. Protein is the end product of translation process and is also used as a trigger to initiate the expression of other genes. Therefore, the combination of these data type is reasonable to most researchers. Besides that, large number of researchers also utilized transcription factor binding sites (TFBS) to verify the GRN. Like protein, TFBS is another complementary data to measure cellular state. Hence, more recent works have explored data integration of external knowledge to identify transcription factors and their target genes [5-6]. Transcription factors are very essential in regulating gene expression. Motivated by this fundamental concept, transcription factors have been used in this research to discover significant biological information from high-throughput data.

## METHODS

This section describes the Bayesian network with the low-order conditional independence along with integration of transcription factor in examining gene regulatory processes for breast cancer metastasis.

## Bayesian Network

The Bayesian network is a graphical model that was introduced by Pearl and Wright in 1980s [8]. To deal with a large number of genes in microarray data, this research defines the Bayesian network, $BN$ as: $BN = (G,P)$ where $G= (X, E(G))$ is a DAG with a set of variables $X$ representing $\{X_i; i \in V\}$, and $E(G) \subseteq X_i * X_j$ (set of pairs that represents the dependent among $v$ variables). The element $E$ is an edge from node $X_i$ to $X_j$, indicating $X_i$ is a parent to $X_j$. On the other hand, $P$ corresponds to joint distribution on the variables in the network. The $Pa(V)$ represents the parent for a set of vertex $V$ and can be defined as:

$$Pa(X_i, G) = \{X_j, \text{ such that } (X_i, X_j) \in E(G); j \in V\} \quad (1)$$

where $Pa(X_i, G)$ is the parent of $X_i$ in the graph, $G$ and having node$X_j$ pointing toward $X_i$.

Bayesian structure learning is a NP hard problem [9]. As the number of possible structures in a Bayesian network grows exponentially with respect to the number of variables (large number of genes in microarray dataset), exhaustive search of all possible structures becomes computationally expensive. Thus, structure learning of Bayesian network is currently a challenging task in modelling GRN. In this study we proposed to low-order conditional independence and its variants, full-order conditional independence, to construct a GRN. For additional technical details on this proposed method please refer to Ahmad, Deris and Othman [10].

## Integration of Transcription Factors in GRN

The Bayesian network has emerged as a powerful tool to infer gene regulatory process. However, this method is usually confronted with structure learning problems in handling large-scale of gene expression data. To address such problem, Bayesian network with constraint-based algorithm is proposed as explained in Section 3.1. The basic idea is to develop GRN by measuring the dependencies among nodes of the given data. Low-order conditional independence is used to examine the relationships between genes. Although the proposed method has increased the accuracy of inferred network, such gene network is solely based on the microarray data and is

often insufficient for rigorous analysis. In many cases, microarray data is often daunted by noisy, incomplete data and misleading outliers, which can produce high number of false positive edges. Accordingly, an inferred GRN may contain some incorrect gene regulations that are unreliable from the biological point of view. Thus, integration of biological knowledge into gene network has become necessary to overcome the problem.

This study used transcription factors concept in determining the gene regulatory relationship. Three main bioinformatics toolkits have been used in extracting transcription factor and region binding, which includes (1) Ensembl, (2) TFSearch and (3) TRANSFAC. Activation or inhibition for each regulatory relationship is determined by the following definition:

1. Activation $\left(X \xrightarrow{+} Y\right)$ – IF X is over-expressed (X = positive), THEN Y is over-expressed (Y = positive), IF X is under-expressed (X = negative), THEN Y is under-expressed (Y = negative).

2. Inhibition $\left(X \xrightarrow{-} Y\right)$ – IF X is over-expressed (X = positive), THEN Y is under-expressed (Y = negative), IF X is under-expressed (X = negative), THEN Y is over-expressed (Y = positive).

## Dataset Description

We tested this proposed method using a data set of 97 breast cancer microarray from van't Veer *et al.* [11]. These cohorts of breast cancer patients are 55 years old or younger. We obtained this data from the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA, 2006). Among the remaining 97 samples, 46 developed distant metastasis within 5 years and 51 remained metastasis free for at least 5 years. DNA microarray analysis was used by van't Veer to determine the expression levels of approximately 25,000 genes for each patient.

## EXPERIMENTAL RESULTS AND DISCUSSIONS

To obtain insights into the mechanism of gene regulation and how gene mutations act to turn on tumour development and metastasis progression in a cellular network context, the proposed method is executed on the breast cancer dataset producing a GRN as shown in Figure 1.

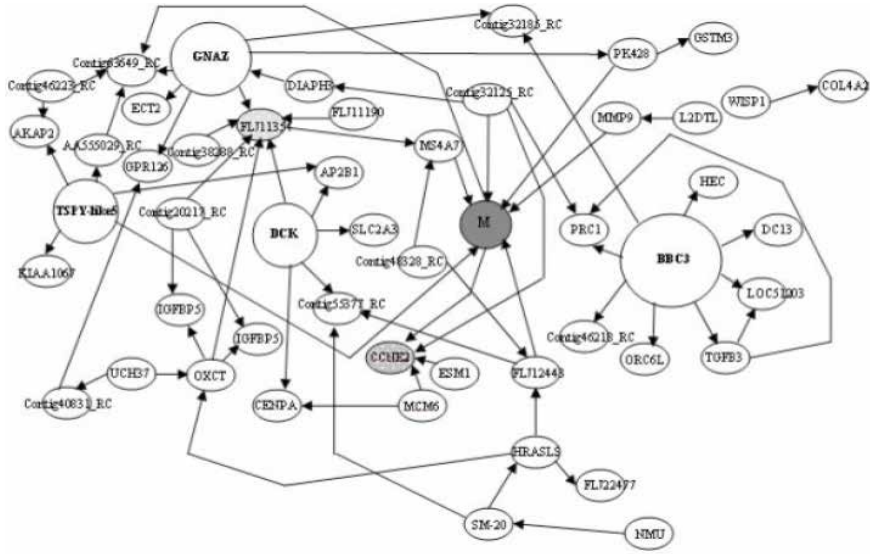


**Figure 1: The GRN for Breast Cancer Metastasis Using the Low-Order Conditional Independence Method**

This learned network revealed a group of genes 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). Meanwhile, the light-shaded nodes denote the 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. This GRN involved 50 genes associated with metastasis, M, and 39 of them are annotated.

To verify the regulatory relationship, this study has tested each relationship using biological integrated data. Three main toolkits such as

Ensembl, TFSearch and TRANSFAC have been used to determine the false positive edges. The experimental results have shown (Table 1) that by integrating heterogeneous data from these sources, the number of false positive edges can be reduced. Accordingly, 258 interactions are found to be biologically related. These interactions have fulfilled the biological test and hypothesis that are set earlier. The results show that the proposed method works better with biological knowledge processing in comparison to network that rely on microarray only since the number of FP edges are discovered decline and the precision result has increased by 2.48% .

**Table 1: Precision Results for Cellular Network Without/with Biological Knowledge Processing. Both Networks are Constructed with 5000 genes (TP = True positive; FP = False positive)**

| Method | Total Edges | FR Edges | TP Edges | Precision % |
|---|---|---|---|---|
| Low-order conditional independence without biological knowledge | 303 | 58 | 245 | 80.85 |
| Low-order conditional independence with biological knowledge | 258 | 43 | 215 | 83.33 |

Table 2 on the other hand illustrates the percentage of activation/ inhibition for each regulatory relationship. Based on the results that have been obtained, this study has discovered that there are five main gene regulators namely BBC3, TGFB3, L2DTL, GNAZ, and TSPY-like 5 that could possibly regulate the expression levels of other genes and highly correlated with breast cancer metastasis. BBC3, TGFB3 and L2DTL are gene regulators that activate the regulation of other genes. These genes mainly inhibit DNA synthesis and induce apoptosis which is the process of programmed cell death (PCD) that causing cells death or damaged. Meanwhile, GNAZ, and TSPY-like 5 are another two genes that most likely activate the expression of other genes which could eventually obstruct the signal transduction pathway.

**Table 2: The Percentage of Activation and Inhibition**

| Regulatory Relationship | Activition % | Inhibition % |
|---|---|---|
| BBC3 → HEC | 34.38 | 65.62 |
| BBC3 → DC13 | 32.81 | 67.19 |
| BBC3 → PRC1 | 34.38 | 65.62 |
| BBC3 → ORC6L | 32.81 | 67.19 |
| BBC3 → Contig46218_RC | 35.94 | 64.06 |
| BBC3 → LOC51203 | 35.94 | 64.06 |
| BBC3 → TGFB3 | 65.63 | 34.37 |
| BBC3 → Contig32185_RC | 31.25 | 68.75 |
| GNAZ → PK428 | 51.56 | 48.44 |
| GNAZ → Contig32185_RC | 65.63 | 34.37 |
| GNAZ → ECT2 | 54.69 | 45.31 |
| GNAZ → Contig63649_RC | 64.06 | 35.94 |
| GNAZ → FLJ11354 | 53.13 | 46.87 |
| GNAZ → GPR126 | 60.93 | 39.07 |
| TSPY-like 5 → AP2B1 | 65.63 | 34.37 |
| TGFB3 → PRC1 | 21.57 | 78.13 |
| TGFB3 → LOC51203 | 20.31 | 79.69 |
| L2DTL → MMP9 | 29.69 | 70.31 |

## CONCLUSION AND FUTURE WORKS

This paper describes the need to integrate diverse data integration for better interpretation of GRN model. Two types of data integration approaches have been comprehensively explained; (1) homologous data integration and (2) heterogeneous data integration. Since most GRN models are mainly implemented based on microarray data, issues like reliability and quality concern are also debated by many researchers. The best available alternative is to integrate different data to address this problem and obtain a better understanding of the underlying gene regulatory mechanisms. Furthermore, with the currently available and enormous public databases, this effort appears to be the most promising since it utilizes the independent and complementary information to answer research questions. The use of transcription factors to identify relevant regulatory interactions is the key idea in this research. In achieving this, three main bioinformatics toolkits for instance Ensembl, TFSearch and TRANFAC have been used. Each of these tools is used to apprehend the concept of biological intrinsic features of transcription factor and promoter. Based on the experiments that were conducted, 258 out of 303 interactions are identified to be biologically

relevant. Furthermore, this study has discovered that there are five main gene regulators namely BBC3, TGFB3, L2DTL, GNAZ, and TSPY-like 5 that play essential role in breast cancer metastasis. In the future, many more different data types will be integrated to obtain more insightful view of GRN and further facilitate our understanding of cancer growth.

## ACKNOWLEDGEMENT

The authors sincerely thank the Research Management Centre, Universiti Utara Malaysia, for the financial support and facilities provided to complete this study.

## REFERENCES

[1] F. Yavari, F. Towhidkhah and S. Gharibzadeh, 2008. Gene regulatory network modeling using Bayesian networks and cross correlation, *Biomedical Engineering Conference (CIBEC)*.

[2] Z. Huang, J. Li, H. Su, G. Watts and H. Chen, 2007. Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining, *Decision Support Systems, vol. 43,* no. 4, pp. 1207-1225.

[3] W. Zhao, E. Serpedin and E. Dougherty, 2008. Recovering Genetic Regulatory Networks from Chromatin Immunoprecipitation and Steady-State Microarray Data', *EURASIP Journal on Bioinformatics and Systems Biology, vol. 2008,* pp. 1-12.

[4] I. Ulitsky, 2009. Network-based algorithms for analysis of heterogeneous biomedical data, Ph.D., Tel-Aviv University, Israel.

[5] C. Huttenhower, K. Mutungu, N. Indik, W. Yang, M. Schroeder, J. Forman, O. Troyanskaya and H. Coller, 2009. Detailing regulatory networks through large scale data integration, *Bioinformatics, vol. 25, no. 24,* pp. 3267-3274.

[6] C. Kaleta, A. Göhler, S. Schuster, K. Jahreis, R. Guthke and S. Nikolajewa, 2010. Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis, *BMC Syst Biol*, *vol. 4, no. 1,* p. 116, 2010.

[7] D. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. Chinnaiyan, 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, *Proceedings of the National Academy of Sciences, vol. 101, no. 25,* pp. 9309-9314.

[8] J. Pearl, 1998. *Probabilistic reasoning in intelligent systems*. San Fransisco, Cal.: Morgan Kaufmann.

[9] D. Chickering, D. Heckerman and C. Meek, 2004. Large-sample learning of Bayesian Networks is NP-hard, *Journal of Machine Learning Research, vol. 5,* pp. 1287-1330.

[10] F.K. Ahmad, S. Deris and N.H. Othman, 2012. The inference of breast cancer metastasis through gene regulatory networks, *Journal of Biomedical Informatics*, *vol. 45, no. 2,* pp. 350-362.

[11] L. van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart and Mao, 2002. Gene expression profiling predicts clinical outcome of breast cancer, *Nature, vol. 415,* pp. 530-536.