Towards Gene Network Estimation with Structure Learning

Suhaila Zainudin[1] and Prof Dr Safaai Deris[2]
[1]Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor
[2]Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia, Bangi, 43600, Selangor

suhaila.zainudin@gmail.com

ABSTRACT

Gene network is a representation of gene interactions. A gene usually collaborates with other genes in order to function. Understanding these interactions is a crucial step towards understanding how our body functions. Bayesian Network is a technique that was initially used in Expert System to represent expert knowledge. Since the pioneer work of Friedman et al. that applied this technique to analyse gene expression data, other researchers have enhanced the technique further. This research concentrates on enhancing Bayesian Network technique fro learning gene network. In order to get better results, Bayesian technique will be used with prior knowledge. The tool that is used to learn the gene network is PNL(Probabilistic Network Library). Early results show that PNL can be used to recover gene network for 3 subnetworks for S.Cerevisiae. These 3 subnetworks has been learned using PNL with varying success. The next step in this research is to learn the gene network from the dataset of 800 genes. The knowledge that will be gained will be used to produce a better approach to learning gene network using Bayesian network technique.

KEYWORDS

Gene interactions, gene network, microarray technology, Bayesian Network, Probabilistic Network Library.

1. Introduction

DNA(Deoxyribonucleic acid) contains the genetic instructions for biological development of all forms of life. Recently, there are many developments in the experimental field of DNA microarray technology such as cDNA (complementary DNA) microarrays and oligonucleotide chips. This high-throughput technology has produced high amount of gene expression data that were successfully used in cancer research and drug discovery [1]. The availability of gene expression data has spurred research in gene clustering, gene classification, gene identification and gene regulatory network modeling [2]. Gene regulatory network is an interconnected network of DNA, RNA(Ribonucleic acid), proteins and small molecules. Gene regulatory controls the functioning of organism at the molecular level. Despite the availability of expression data, the process of inferring the complex regulatory network from expression data is not easy [3]. Hence, instead of inferring the complete network, some researches has focused on inferring gene network.

Gene expression is not an independent event. Interactions between genes exist and these interactions are complex. Gene interactions are studied through gene network. Gene network is a group of coordinately expressed genes controlling a particular function of an organism. By reconstructing gene networks, we are trying to find which genes are affecting the network and how the genes are affecting the network. Genes affect the network by being up regulated (over-express) and down regulated (under-express) when exposed to certain conditions.

One of the most important question in biology is how gene expression is turned on or off . Most cells in an organism have an identical genome. However, different levels of gene expression are produced by each cell according to the current need. A type of proteins called

transcription factors play an important role in gene regulations in eukaryotes. Transcription factors bind to specific binding parts of DNA(called transcription factor binding sites) located in promoter region.

Specific promoters are associated with particular genes. Transcription factors functions by binding the gene's promoter and either switch on or switch off the gene's transcription. Transcription factors are actually product of genes and controlled by other transcription factors. Transcription factors can be in control of many genes and most genes are controlled by different combinations of transcription factors. These relations can be represented in gene network.

We can use microarray and computational methods to recover gene networks from experiments. Understanding these interactions leads to better understanding of cellular processes.

2. Problem Background

The main problem is understanding why some genes are expressed while other genes are repressed. In addition, what does the different levels of gene expression at different timepoints signify. It is known that gene expressions are not independent and there exists interactions among genes. Some examples of gene interactions are; the expression of gene A promotes the expression of gene C meanwhile the expression of gene A and gene B inhibits the expression of gene D. The sequence of interactions can be complicated.

The challenge in this research is how to extract meaningful information from expression data and discover interactions between genes based on the microarray measurements. Microarray experiments contains data on thousands of genes, however from past researches, only a handful of genes are responsible for the interactions captured in an experiment.

2.1 Research Objectives

The goal of the research is to produce an algorithm to infer gene network based on Bayesian Network Technique. In order to achieve the goal, several objectives are set. The objectives are:-

1. To identify the genes which are responsible in a biological phenomenon of interest.
2. To produce the gene network model with existing technique.
3. To enhance the technique by incorporating biological knowledge.
4. To verify that the gene network model produced is biologically sound.

3. Yeast Cell-cycle Dataset

Spellman data set contains 76 gene expression measurements of the mRNA(messenger RNA) level of 6177 S.cerevisiae (yeast) genes [4]. The experiments measure six time series under different cell cycle synchronization methods. Yeast cell division is divided into 4 major phases: G1, S, G2 and M, which are the traditional subdivision of cell cycle [5]. Experiments in [4] identified 800 genes as meeting the minimum criterion for cell cycle regulation. A full description and complete data sets are available at cellcycle-www.stanford .edu.

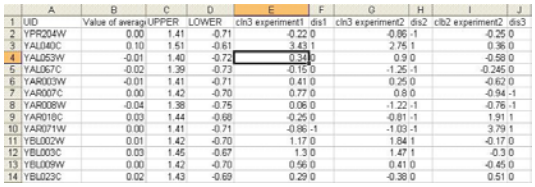
Friedman et al. [6] is the first research that successfully applied Bayesian Network to analyse expression data. The resulting gene network by Friedman is available at www.cs.huji.ac.il/labs/compbio/expression/. Since this research, other researchers has also extended Bayesian Network to produce gene networks for yeast [5][7][8].

3.1 Data Pre-processing

Pre-processing is an important step in data mining. Pre-processing prepares the data by cleaning and formatting data. This will ensure that the data can be used with available software. From visual inspection, 1% to 3 % of data were missing from yeast cell-cycle dataset. Hence KNN (K nearest neighbour) impute method was used to estimate missing data.

After the missing data was estimated, the gene expression is discretised. This research follows the discretisation method from [6]. The values are discretised into three categories: under(-1), normal (0) and over(1). Each value is compared to a control value and from this

comparison, it is categorised into one of the three categories. The control value was set as the average expression level of the gene across experiments. To discretize the data, a threshold was set to the ratio between measured expression and control. In this experiment , the threshold value was set to $\log^2$ 0.5. Any gene expression value with ratio to control less than $2^{-0.5}$ are considered as under-expressed, and values more that $2^{0.5}$ are considered as over-expressed. Any values that fell in between these two values are considered normal. The process of the discretisation is shown in Figure 1 and the resulting discretised values as shown in Figure 2.



	A	B	C	D	E	F	G	H	I	J
1	UID	Value of averag	UPPER	LOWER	cln3 experiment1	dis1	cln3 experiment2	dis2	clb2 experiment2	dis3
2	YPR204W	0.00	1.41	-0.71	-0.22	0	-0.86	-1	-0.25	0
3	YAL040C	0.10	1.51	-0.61	3.43	1	2.75	1	0.36	0
4	YAL053W	-0.01	1.40	-0.72	0.34	0	0.9	0	-0.58	0
5	YAL067C	-0.02	1.39	-0.73	-0.15	0	-1.25	-1	-0.245	0
6	YAR003W	-0.01	1.41	-0.71	0.41	0	0.25	0	-0.62	0
7	YAR007C	0.00	1.42	-0.70	0.77	0	0.8	0	-0.94	-1
8	YAR008W	-0.04	1.38	-0.75	0.06	0	-1.22	-1	-0.76	-1
9	YAR018C	0.03	1.44	-0.68	-0.25	0	-0.81	-1	1.91	1
10	YAR071W	0.00	1.41	-0.71	-0.86	-1	-1.03	-1	3.79	1
11	YBL002W	0.01	1.42	-0.70	1.17	0	1.84	1	-0.17	0
12	YBL003C	0.03	1.45	-0.67	1.3	0	1.47	1	-0.3	0
13	YBL009W	0.00	1.42	-0.70	0.56	0	0.41	0	-0.45	0
14	YBL023C	0.02	1.43	-0.69	0.29	0	-0.38	0	0.51	0

Figure 1 Discretisation Process

	A	B	C	D	E	F	G
1	UID	dis1	dis2	dis3	dis4	dis5	dis6
2	YPR204W	0	-1	0	0	0	0
3	YAL040C	1	1	0	0	0	0
4	YAL053W	0	0	0	-1	0	0
5	YAL067C	0	-1	0	0	0	0
6	YAR003W	0	0	0	0	0	0
7	YAR007C	0	0	-1	0	0	0
8	YAR008W	0	-1	-1	0	0	0
9	YAR018C	0	-1	1	1	0	0
10	YAR071W	-1	-1	1	1	-1	-1
11	YBL002W	0	1	0	-1	-1	-1
12	YBL003C	0	1	0	0	-1	-1
13	YBL009W	0	0	0	0	-1	-1
14	YBL023C	0	0	0	0	0	0

Figure 2 Discretisation Result

4. Probabilistic Network Library

The software tool used in this research is PNL (Probabilistic Network Library) which is an open source tool from Intel for working with graphical models. This tool supports directed and undirected models, discrete and continuous variables, various inference and learning algorithms. Currently, PNL supports three types of learning tasks (Table 1).

Table 1 Types of Learning Tasks

Type of Task	**Graphical Model Structure**	**Observability of Variables**
Type 1	Known	All variables observed
Type 2	Known	Some variables are not observed
Type 3	Unknown	All variables are observed
Type 4	Unknown	Some variables are not observed

The learning task associated with this research is Task 3 (unknown structure, all variables observed). In PNL, learning a model is done by structure learning, where an unknown structure is estimated from given evidences. A graphical model is defined by its structure and parameters (conditional probability distribution). Learning a graphical model is to estimate model factors so as to ensure the best explanation of information for the model [9]. Evidence or input for learning is in a table with columns representing variables and rows representing evidence.

5. Gene Subnetworks

The problem of gene network estimation is NP-hard with super-exponential search space [10]. Attempts made to estimate the whole gene network from yeast has not yield encouraging results, hence a smaller scale approach is taken by estimating yeast gene subnetworks. A past research has identified 3 subnetworks from the yeast cell-cycle dataset(YPL256 subnetwork, YOR263 subnetwork, histone subnetwork) [5]. In order to be familiar with the complexity of PNL, these 3 known subnetworks were estimated. The learning framework is in Figure 3.

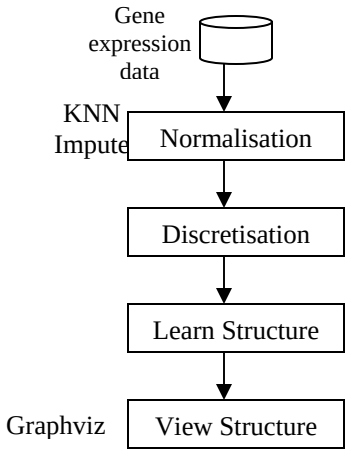


Figure 3 Learning Structure Framework

6. Learning Structure Algorithm

Pseudocode for learning the gene subnetworks is in Figure 4.

```
Declare new network
Add nodes and states  to new network.
Add evidence for nodes from file.
Learn network structure
Save and view network
```

Figure 4 Pseudocode for learn network structure

Learn network structure employs hill-climbing algorithm. For scoring propective structures, BIC (Bayesian Information Criterion) is used. The pseudocode for hill-climbing algorithm is in Figure 5.

```
Starts with initial structure.
    Choose different initial structures as starting points.
    For each structure
        Run hill-climbing search.
        Score each neighbour of current structure using BIC.
        Moves to neighbour with highest score.
    Return best structure
```

Figure 5 Pseudocode for hill-climbing algorithm

7. Subnetwork Learning Results

Data for 3 gene subnetworks were run with PNL tool. The resulting network is viewed using Graphviz, an open source graphical modelling tool. The visualization for YOR263 subnetwork is in Figure 6.

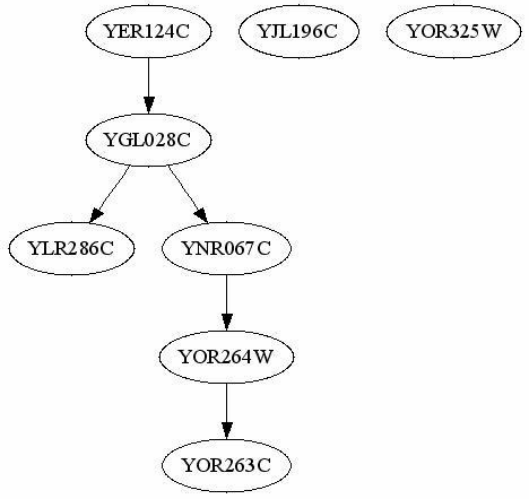


Figure 6 YOR263 Subnetwork

Results for all 3 subnetworks are summarise in Table 2.

Table 2 Results for Subnetworks

Subnetworks	Similar edges	Missing edges
YOR263	5	1
YPL256	4	7
Histone	0	4

Using data for the gene involved with each subnetwork, 3 gene subnetworks were learned. The subnetwork that has the most similarity is YOR263 subnetwork and the subnetwork with no similarity is Histone subnetwork.

Several limitations of PNL were encountered during subnetwork learning such as nodes must be topologically sorted and size of input data is limited to 200 genes.

8. Conclusions and Future Work

From subnetworks results, it is deduced that PNL can be used to estimate gene network from gene expression data. The next major task is to estimate gene network for complete dataset (800 genes). Workaround must also be formulated to overcome the limitations of PNL. An extension to PNL that can display Link Connection Strength for edges in gene network will be implemented.

9. Acknowledgement

The author wishes to thank JPA and UKM for supporting this ongoing research. The author also wishes to thank everyone who has contributed to this research.

References

[1] M.Schena. “Microarray Analysis”. John Wiley. 2003.

[2] I. H. Whitten and E. Frank.“Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”. Morgan Kauffmann. 2000.

[3] L. A. Soinov, M. A.Krestyaninova, and A. Brazma. “Towards reconstruction of gene networks from expression data by supervised learning”.Genome Biology.2003. 4 (1).

[4] P. T. Spellman, G. Sherlock, M.Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O’Brown, D. Botstein, and B. Futcher. “Comprehensive Identification of Cell Cycle-regulated Gene of the Yeast Saccharomyces cerevisiae by Microarray Hybridization”. *Molecular Biology of the Cell*. 1998. 9.3273-3297.

[5] M. Dejori. “Analyzing Gene-Expression Data with Bayesian Networks”. Masters Thesis. 2002. Technical University of Graz.

[6] N. Friedman, M. Linial, I. Nachman, and D. Pe’er. “Using Bayesian Networks to Analyze Expression Data”. Journal of Computational Biology .7(3). 601-620.

[7] S. Imoto, S.Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara, and S. Miyano. “Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network”. Journal of Bioinformatics and Computational Biology.2003. 1(2). 231-252.

[8] N. Nariai, S. Kim, S. Imoto, and S. Miyano. “Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks”. Proceedings of Pacific Symposium on Biocomputing. 2004. 9. pp.336-347.

[9] Probabilistic Network Library:User Guide and Reference Manual.(2003). Intel Corporation.

[10] S.Ott, A.Hansen, S.-Y.Kim, and S.Miyano.”Superiority of Network Motifs over Opitmal Networks and an Application to the Revelation of Gene Network Evolution”. Bioinfomatics.2004.1(1).1-10.