

# 國立交通大學

統計學研究所

博士論文

情緒偵測、基因表現分群  
以及生物網路重建之統計方法

Statistical Approaches for Emotion Detection,  
Gene Expression Clustering and Biological  
Pathway Reconstruction

1896

研究生：闕棟鴻

指導教授：盧鴻興 教授

中華民國九十七年七月

情緒偵測、基因表現分群以及生物網路重建之統計方法  
Statistical Approaches for Emotion Detection, Gene Expression Clustering  
and Biological Pathway Reconstruction

研究生：闕棟鴻

Student : Tung-Hung Chueh

指導教授：盧鴻興

Advisor : Henry Horng-Shing Lu



A Dissertation  
Submitted to Institute of Statistics  
College of Science  
National Chiao Tung University  
in partial Fulfillment of the Requirements  
for the Degree of  
Doctor of Philosophy  
in  
Statistics

Institute of Statistics, National Chiao Tung University  
Hsinchu, Taiwan, R.O.C.

中華民國九十七年七月

# 情緒偵測、基因表現分群以及生物網路重建之統計方法

學生：關棟鴻

指導教授：盧鴻興

國立交通大學統計學研究所 博士班

## 摘 要

本論文主要是利用統計在三個不同的研究上的應用，包括情緒偵測、基因表現分群以及生物網路重建。在第一個研究中，我們致力於發展一種情緒偵測的系統。在人類與電腦的聯繫以及溝通上，發展一種裝置可以辨別人類的情緒狀態，將會是相當具有價值的，在此研究中，我們收集受試者在三種不同的情緒狀態下的生理訊號，包含心電圖、皮膚表面溫度以及皮膚表面電阻，並從中取出三十個特徵值。在藉由多變量變異數分析去除掉因為不同天測量所產生的雜訊後，我們使用六種機器學習的方法來辨別情緒的狀態，最後發現使用邏輯迴歸法即可達到最佳的分辨準確率，同時我們發現在使用多變量變異數分析去除掉每天的雜訊後，即可有效的改善這六種分類方法的準確率。

在本篇論文的第二個研究中，我們藉由生物基因的實驗來探討酵母菌在實驗以及野生品種其在進入發酵生活轉到呼吸生活時基因表現。同時，我們研究在這兩個品種中，表現的不同基因。在使用基因過濾，分群分析以及迴歸模型來偵測此兩個品種擁有不同表現的基因後，我們發現有一群的基因其在野生及實驗品種的表現呈現有負相關的情況，同時，在這群的資料中，其基因的顯著表現的時間比起葡萄糖濃度的下降時間早了一個小時。在我們後續的研究當中，將可利用例如網路分析等工具來研究這種有趣基因其因果的關係。

在生物資訊的研究中，從基因表現的趨勢來推論基因控制網路以及生物的因果路徑是相當重要的一個研究。在本篇論文的第三個研究中，我們提出了一個時間延遲布朗網路來探究生物網路。我們假設每個基因最多是受到  $k$  個基因所影響，同時在推論時，我們假設  $k=2$ ，此外，我們在布朗方程式以及受影響的基因之間，我們考慮兩種關係：相似性以及必要性。在我們推論的方法中，我們將每一個輸出的基因以及成對的輸入基因與八個基本的關係做比較，並且計算其  $p$  分數，我們預期  $p$  分數愈小者，代表其之間的關係愈可能存在，我們將收集所有一致的關係，並找出其最可能出現的關係。最後我們將使用一個模擬的資料例子以

及一個真實的酵母菌基因網路關係來進行分析，其結果呈現，我們所提出的基因網路重建方法可以有效的重建出原本的網路模型。

關鍵字：情緒偵測、生理特徵、機器學習、特徵選取、多變量變異數分析、發酵生活轉到呼吸生活、生物晶片、基因過濾、分群分析、路徑分析、布朗網路、布朗方程式、測量誤差。



# Statistical Approaches for Emotion Detection, Gene Expression Clustering and Biological Pathway Reconstruction

student : Tung-Hung Chueh

Advisors : Henry Horng-Shing Lu

Institute of Statistics  
National Chiao Tung University

## ABSTRACT

This thesis consists of three different researches in the implement of statistical approaches, emotion detection, gene expression clustering and biological pathway reconstruction. In the first research area, we focus on developing an emotion recognition system by the supervised learning. For the importance of communication between human and machine interface, it would be valuable to develop an implement which have the ability to recognize emotion. We propose an approach which can deal with the daily dependence and personal dependence in the data of multiple subjects and samples. Thirty features were extracted from the physiological signals of subject for three statuses of emotion. The physiological signals measured were: electrocardiogram (ECG), skin temperature (SKT) and galvanic skin response (GSR). After removing the daily dependence and subject dependence by the statistical technique of MANOVA, six machine learning including Bayesian network learning, naive Bayesian classification, SVM, decision tree of C4.5, Logistic model and K-nearest-neighbor (KNN) were implement to differentiate the emotional states. The results show that Logistic model gives the best classification accuracy and the statistical technique MANOVA can significant improve the performance of all six machine learning methods in emotion recognition system.

In the second part of this thesis, we explore the expression pattern of yeast genes for diauxic shift in BY and RM strains by Micorarray studies. In particular, we investigate the differential expressed genes between these two strains. After performing gene filtering, cluster analysis and regression model to detect the differential expression patterns of yeast genes for diauxic shift in BY

and RM strains, we find a group of genes which have negative correlation in two strains. Besides, the estimated time shifts of expression time profiles in the group are mainly 1 hour before the time that glucose consumption drops. Further analysis such as network analysis could be used to investigate the causal relationship of these interesting genes based on the framework of current result in the future.

Inference of genetic regulatory networks and biological pathways from gene expression patterns is a critical problem in bioinformatics. In the third part of this thesis, we propose using the structure of Time Delay Boolean networks as a tool for exploring biological pathways. We suppose the indegree of each gene (i.e., the number of input genes to each gene) is bounded by a constant  $K$  and take  $K = 2$  for the instance of inference. In addition, we consider two kinds of relations between the output gene and the Boolean function with input genes: similarity and prerequisite. In our inference strategy, we compare every output gene and all the pairs of input genes with the eight basic relations and calculate their corresponding p-score. Since we expect that the smaller the p-score, the more likely the relation, we combine those consistent relations and find out the most possible relation between output gene and the pair of input genes. We illustrate the method using a simulated example and a published microarray expression dataset of yeast *Saccharomyces cerevisiae* from experiments with regulation of gluconeogenesis by Cat8 and Sip4. The results show that our proposed algorithm is extensible for more realistic network models.

Keywords: Emotion recognition, physiological signals, machine learning, feature selection, MANOVA, diauxic shift, Microarray, gene filtering, cluster analysis, pathway, Boolean network, Boolean function, measurement error, EM algorithm.

## 誌 謝

在求學的這段路程非常的漫長，但又彷彿是昨日才開始，今日我能夠如願的畢業，首先我要感謝我的指導老師盧鴻興教授對我的指導與照顧。在研究上不但給予我許多的幫忙與協助，對於學生的生活也相當的照顧，尤其是在芝加哥大學訪問將近一年的時間裡，更是無時無刻的給予支持及鼓勵，並感謝芝加哥大學李文雄老師的照顧，使我能在出國的這段期間感受到溫暖。同時我也要感謝口試委員撥空參加我的口試，給予我許多寶貴的建議，讓我的論文更臻完善。

其次，我要感謝我在求學中的許多朋友，感謝女友蕙如學妹給予我精神上許多的支持以及生活起居的照顧，感謝泰賓學長時常在研究上的教導，也謝謝研究室裡的各位伙伴的陪伴，使我在這段求學生活中，添加許多樂趣，也更豐富我的求學生活。

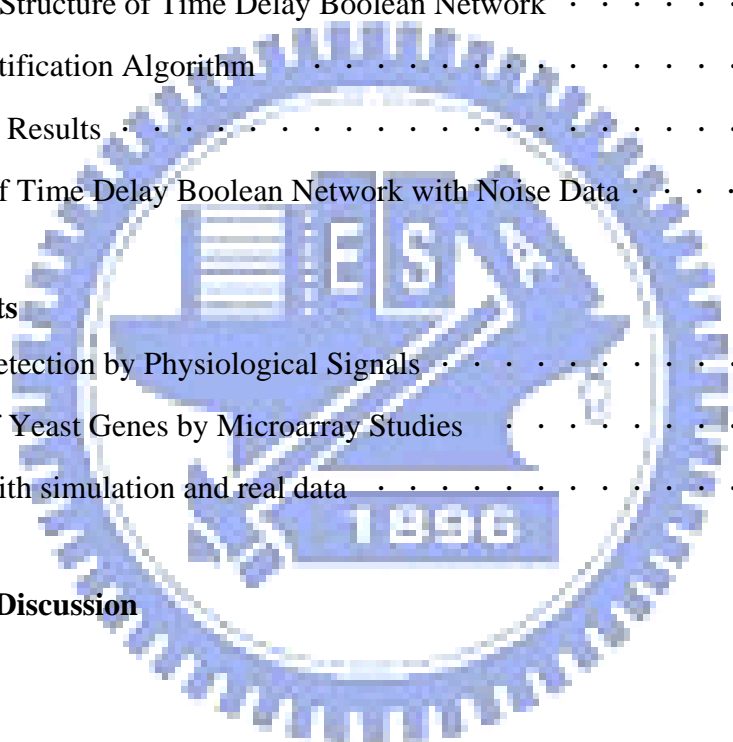
最後，我要感謝的是我的家人，在我求學的這段日子，不斷的給我鼓勵及安慰，以及各種的支持，讓我可以無後顧之憂的全心全力的專注在學習以及研究上。在此，對於所有幫助過我的朋友及師長，獻上最高的敬意與謝忱。願與大家共同分享完成論文的這份喜悅與榮耀。

# Contents

<b>Chinese Abstract</b>	<b>i</b>
<b>English Abstract</b>	<b>iii</b>
<b>Acknowledge</b>	<b>v</b>
<b>Contents</b>	<b>vi</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>1. Motivations and Literature Reviews</b>	<b>1</b>
1.1. Emotion Detection by Physiological Signals	1
1.2. Analysis of Yeast Genes by Microarray Studies	4
1.3. Inference of Biological Pathway by Time Delay Boolean Networks	5
1.4. Organization of The Dissertation	6
<b>2. Data Collection and Statistical Application on Emotion Detection</b>	<b>8</b>
2.1. Collection of the Data	8
2.2. Feature Extraction	11
2.3. Daily and Personal Correction	12
2.3.1. The Problem of Day-effects and Person-effects	12
2.3.2. MANOVA	13
2.4. Pattern Classification	14
2.4.1. Bayesian Network	15
2.4.2. Naive Bayesian	16
2.4.3. Support Vector Machine	17
2.4.4. Decision Tree of C4.5	17
2.4.5. Logistic Model	18
2.4.6. K-Nearest Neighbor (KNN)	19
<b>3. Data Collection and Analysis on Yeast Genes of Microarray Studies</b>	<b>20</b>
3.1. Materials and Microarray Experiment	20
3.2. Data Extraction	21



3.3. Strain Normalization	22
3.4. Gene Filtering	23
3.5. Cluster Analysis	25
3.5.1. Hierarchical Clustering	25
3.5.2. Curve Clustering	29
3.6. Regression models with time shift	32
<b>4. Inference of Biological Pathway by Time Delay Boolean Network</b>	<b>35</b>
4.1. Models	35
4.1.1. Boolean Network	35
4.1.2. The Structure of Time Delay Boolean Network	36
4.1.3. Identification Algorithm	38
4.2. Theoretical Results	41
4.3. Inference of Time Delay Boolean Network with Noise Data	43
<b>5. Empirical Results</b>	<b>48</b>
5.1. Emotion Detection by Physiological Signals	48
5.2. Analysis of Yeast Genes by Microarray Studies	51
5.3. Example with simulation and real data	63
<b>6. Conclusion and Discussion</b>	<b>66</b>
<b>Bibliography</b>	<b>69</b>



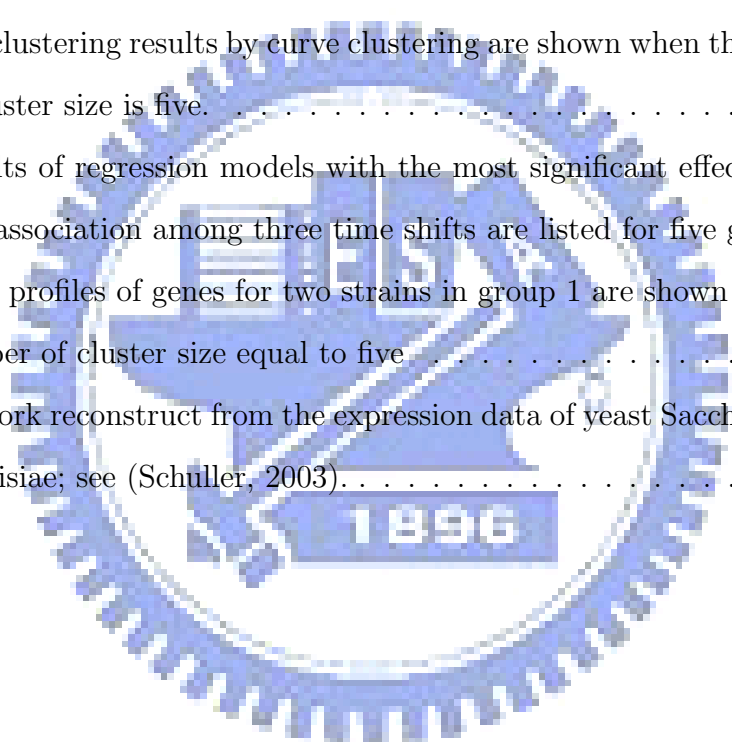
# List of Tables

4.1	Tables for a pair of input gene and one output gene assuming no measurement error . . . . .	39
4.2	Count patterns for the basic eight relations assuming exhaustive sampling and no measurement error . . . . .	40
4.3	The eight basic relations and their corresponding probabilistic hypotheses and scores . . . . .	44
4.4	The $2 \times 4$ count table for a pair of input and output gene and their generated probabilities in the presence of measurement error . . . . .	44
4.5	Splitting counts caused by misclassification error . . . . .	45
5.1	Classification of three emotional statuses by the physiological signals of subject Alice. . . . .	50
5.2	Classification of three emotional statuses by the physiological signals of subject Jane. . . . .	50
5.3	Classification of three emotion status by the physiological signals of 10 subjects and 7 times. . . . .	51
5.4	The detail values of PSE. . . . .	52
5.5	Consistent genes are reported for five groups. . . . .	61
5.6	Degrees of clustering consistency for all genes are tabulated. . . . .	62
5.7	For the Time Delay Boolean network in figure 1, we generate 100 samples, and take $p=0.05$ . . . . .	63

# List of Figures

2.1	Three physiological signals were recorded when the subject was asked to feel neutral. From top to bottom: skin temperature variation (SKT), galvanic skin response (GSR) and electrocardiogram (ECG). The physiological signals were sampling at 256 samples for every second and the measured times were 200 seconds. . . . .	9
2.2	Three physiological signals were recorded when the subject was asked to feel joyful. The physiological signals were sampling at 256 samples for every second and the measured times were 120 seconds. . . . .	9
2.3	Three physiological signals were recorded when the subject was asked to feel angry. The physiological signals were sampling at 256 samples for every second and the measured times were 120 seconds. . . . .	10
2.4	Flow chart in the study of emotion recognition. . . . .	11
2.5	The network structure of the naive Bayesian classifier. . . . .	16
3.1	The flowchart in the study of of yeast genes. . . . .	23
3.2	In this case, gene $g_2$ is considered to have clustering consistency. . . . .	26
3.3	Comparisons of PSEs for different cluster sizes are plotted for hierarchical clustering with different settings. . . . .	28
3.4	The dendrogram of the hierarchical clustering is shown for 30 nodes. . . . .	28
3.5	The typical results of two dimensional expression curves in experiment 1 for five groups are plotted. . . . .	31
3.6	Model selection by BIC is shown for curve clustering. . . . .	32

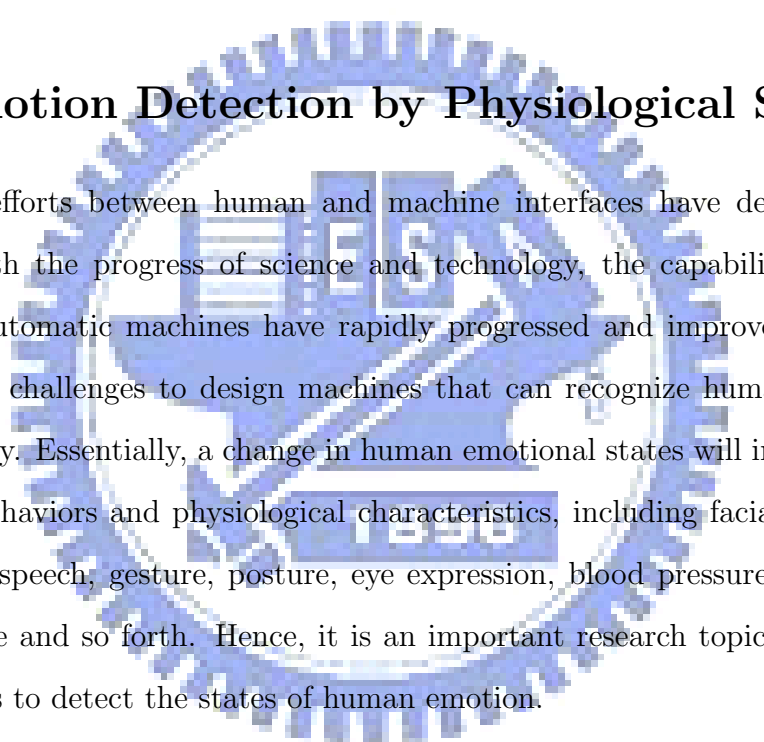
4.1	A Boolean network $G(V, F)$ , its wiring diagram $G'(V', F')$ and the functional dependency table. (Akutsu et al. 1999) . . . . .	36
4.2	One example of Time Delay Boolean network and its Input/Output. . .	38
5.1	The scatter plot of three statuses of emotion without daily correction.	48
5.2	The scatter plot of three statuses of emotion with daily correction. . .	49
5.3	PSE comparisons of different number of clusters are shown for two different clustering methods. . . . .	51
5.4	Mean curves of every group are shown for two clustering methods with different clustering sizes. . . . .	55
5.5	The clustering results by curve clustering are shown when the number of cluster size is five. . . . .	56
5.6	Results of regression models with the most significant effects of glucose association among three time shifts are listed for five groups. . .	58
5.7	Time profiles of genes for two strains in group 1 are shown when the number of cluster size equal to five . . . . .	60
5.8	Network reconstruct from the expression data of yeast <i>Saccharomyces cerevisiae</i> ; see (Schüller, 2003). . . . .	65



# Chapter 1

## Motivations and Literature Reviews

### 1.1 Emotion Detection by Physiological Signals



Research efforts between human and machine interfaces have developed over decades. With the progress of science and technology, the capability and functionality of automatic machines have rapidly progressed and improved. Even so, there are still challenges to design machines that can recognize human emotional states correctly. Essentially, a change in human emotional states will influence a lot of external behaviors and physiological characteristics, including facial expression, intonation of speech, gesture, posture, eye expression, blood pressure, heart beat, skin resistance and so forth. Hence, it is an important research topic to use these characteristics to detect the states of human emotion.

In the field of human and machine interaction, it would be valuable to develop an instrument capable of recognizing a person's emotional status. Emotion recognition has become a critical investigation in emotional intelligence and can be applied in many systems. In 1999, Ark et al. at the laboratory of IBM established a mouse that can distinguish a user's affective states with 75 percent accuracy. A robot with the ability to recognize and determine the underlying emotion of a person can interact with humans using signals in human speech and facial expression (Breazeal

and Aryananda, 2002; Littlewort et al., 2004). Moreover, other applications such as driving safety, training and telemedicine also can implement an emotion recognition system to benefit users (Nasoz et al., 2004).

In previous research about developing an emotion recognition system, features of facial expressions are most commonly used as the determinant attribute and have successfully obtained fairly high rates in emotion recognition (Yacoob and Davis, 1996; Cowie et al., 2001; Hu et al., 2002; Fasel and Luetttin, 2003; Zhou and Lin, 2005). Besides, there are also studies employing signals of speech and vocal intonations to recognize states of emotion (Dellaert et al., 1996; Nwe et al., 2003). Combining facial and voice expression has also been used in distinguishing affective emotional states recently (Busso et al., 2004). However, these two characteristics are sometimes hardly recorded if the subject is moving. Therefore, recognizing emotion using physiological signals, which can be recorded for a moveable subject, is a critical study.

In the study of affective physiological states, Picard et al. (2001) at MIT Media Laboratory have tried to differentiate eight different emotions of a single person using physiological characteristics recorded every day over six weeks, resulting in an 81% overall classification accuracy rate by using a hybrid method involving sequential floating forward search and Fisher projection. For handling the physiological signals with short-term segments, Kim et al. (2004) proposed an algorithm to detect emotional statuses based on their experimental psychosomatic responses for multiple subjects and got the correct classification rate of 78.4% by the machine learning method of support vector machine (SVM). Nasoz et al. (2004) employed three classification methods to discriminate six different emotional states from physiological signals collected via non-invasive technologies. Rani et al. (2006) have applied four different classification methods to determine affective states from physiological signals and have made comparisons of these methods.

Among emotion recognition studies, there are typically two approaches: one

against one (Picard et al., 2001) and one against all (Kim et al., 2004). For the one against one approach, we can collect the labeled psychosomatic signals of a single subject on multiple observations and learn a trainer model out of the same person so that we can decipher the unknown emotional states of that person as a test of his (her) physiological signals. Though it has the benefit of removing the inter-subject difference for subject-based learning, this approach can only recognize one subject's emotion. Alternatively, we can measure the physiological signals of emotion from multiple subjects and learn a trainer model out of them. Hence, we can distinguish other people's emotion status using this system. In practice, this user-independent system is believed to be more convenient in the field of emotional recognition studies. However, the assumption of independence between physiological signals and subjects is not reasonable nor practical.

Furthermore, daily physiological signals can vary even for the same state of emotion. The daily effect could be removed using the statistical technique of multivariate analysis of variance (MANOVA). Then, typical machine learning methods could be applied to discriminate and predict the emotional state. Hence, the purpose of this work is to advance the improvement of emotion recognition by eliminating inter-subject differences and removing the daily effects by MANOVA with statistical machine learning.

Physiological signals including skin temperature variation (SKT), galvanic skin response (GSR) and electrocardiogram (ECG) were implemented in this study. These physiological signals can be measured conveniently without any annoying sensors attached on the face or scalp. The subjects would induce three different emotional statuses by themselves: anger, joy and neutral. Besides, we would use the techniques of multivariate analysis of variance (MANOVA) and six different classification methods to discriminate various states of emotion.

## 1.2 Analysis of Yeast Genes by Microarray Studies

Although yeast *Saccharomyces cerevisiae* can utilize various carbon substrates as a biomass and energy source, fermentable sugars such as glucose or fructose are clearly the preferred carbon sources over nonfermentable substrates such as ethanol, glycerol, lactate, acetate or oleate (Schuller, 2003). When glucose is present, the enzymes required for the utilization of alternative carbon sources are synthesized at low rates or not at all. This phenomenon is known as carbon catabolite repression, or simply glucose repression (Gancedo, 1998). Analysis of genomic expression has revealed that many genes are differentially transcribed in response to varying glucose levels (DeRisi et al., 1997).

Yeast cells undergo fermentation, which metabolizes sugars (glucose) and produces ethanol when sugars are abundant; as the sugars are depleted, cells undergo a "diauxic shift" in which cells switch to a fully respiratory metabolism (DeRisi et al., 1997; Gasch et al., 2000; Schuller, 2003). It is very important to understand the biological process of diauxic shift in fermentation for yeast. Our major goal is to understand the expression evolution of genes involved in this transition (the diauxic shift) and in non-fermentative metabolism, which is not well understood.

A laboratory strain (BY4741) and a wild strain (RM11-1a) are used in this study. These two strains proliferate rapidly and have propagated under different environmental conditions for decades. These two strains also display substantial divergence in gene expression and are ideal for studying expression divergence within species (Brem et al., 2002). We performed microarray analysis to study the expression profiles of genes during the diauxic shift. In particular, we investigated the differential expressed genes (DEGs) between these two strains. Our results showed that the RM strain may experience the diauxic shift earlier than BY strain and that many of the key genes related to the diauxic shift are turned on earlier in the RM strain.



## 1.3 Inference of Biological Pathway by Time Delay Boolean Networks

In bioinformatics, inference of genetic regulatory networks and biological pathways from gene expression patterns is a crucial issue. Due to the invention of DNA microarray technology, thousands of gene expression can be monitored and measured simultaneously (DeRisi et al., 1997). However, it is still a great challenge to identify complex biological networks, since the number of combinations with the gene interactions is huge. In recent years, there has been a dramatic proliferation of research concerned with network reconstruction problems.

Clustering is such an important method for grouping genes which have similar expression patterns (Eisen et al., 1998). In the framework of clustering, it is an important task to define the degree of similarity between genes. By the method of clustering, we can group genes which have similar expressions. However, we still can not find the causal relationship between genes. Hence, apart from the relations of similarity, we also have to consider another causal relationship between genes.

There have been many methods proposed in the literature for the inference of genetic regulatory networks. Over the past two decades, Bayesian networks is an important technique and has been extensively studied (Pearl, 1988; Jensen, 1996). Bayesian networks is a graphical model that contains directed probabilistic relationships between elements. The structure of a Bayesian network consists of two components. The first component comprises vertices which corresponding to a set of variables and a set of directed edges between variables with Markov properties. The second component describes a conditional distribution for each variable, given its parents. Recently, Bayesian network models have been applied to analyze microarray expression data (Friedman et al., 2000; Heckerman et al., 1995). Although the Bayesian networks are complete models and some algorithms searching for Bayesian networks have been developed, the computational cost is still fairly large. Even there

are only a sparse number of variables, sample sizes of several hundred are required for achieving high accuracy of estimation.

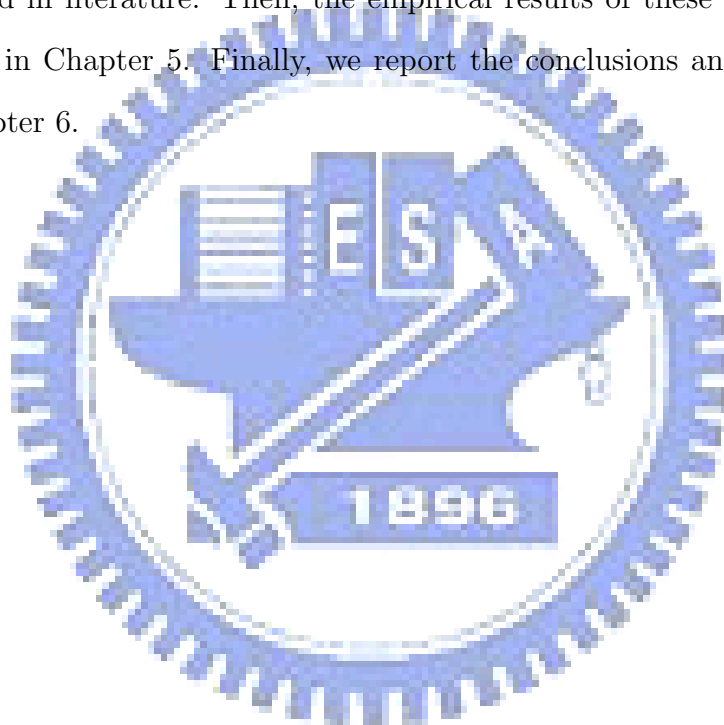
This study is based on a much simpler model: Boolean networks. Boolean networks were originally introduced by Kauffman (Kauffman, 1969) and received much attention for inferencing gene regulatory networks. In Boolean network models, gene expression states are quantized to one of two states: ON and OFF. Under the structure of Boolean networks, the target gene is influenced by a set of genes with a Boolean function. For each gene, if the indegree (i.e., the number of input genes to each gene) is bounded by a constant  $K$ , only  $O(\log n)$  pairs of state transition are necessary to reconstruct the original network with  $n$  nodes (Akutsu and Miyano, 1999). However, the deterministic model predicted by the input genes and Boolean function is criticized.

In 2005, (Li and Lu, 2005) proposed another relationship between two genes: prerequisite under the Boolean network model. If a Boolean function with one or several genes is prerequisite for a target output gene, the target gene will be influenced by the Boolean function with several input genes. However, the target gene may not be expressed right now, but at another future time. Hence, the induction of the Boolean function with input genes is necessary for the expression of the target gene, and we also treat these relations as time delay affection. In this paper, we would infuse these additional relations for more generalized systems.

## 1.4 Organization of The Dissertation

This dissertation is organized as follows. In Chapter 2, we focus on the study of emotion detection. First, we present the procedures of data collection and features extraction from the measured physiological signals. Then, we discuss the problem of day-effects and remove daily effects using MANOVA. Besides, we also consider six classification methods : Bayesian network learning, naive Bayesian classification, support vector machine (SVM), decision tree of C4.5, logistic model and K-nearest

neighbor (KNN). In Chapter 3, we discuss the analysis on yeast genes of microarray studies. We perform gene filtering, cluster analysis and regression model to detect the differential expression patterns of yeast genes for Diauxic shift in BY and RM strains. In Chapter 4, we propose a Time Delay Boolean network model and its identification algorithm for the inference of biological pathway. We also discuss the theoretical results concerning the number of gene expression patterns required to identify the Time Delay Boolean network model. Moreover, we illustrate the method by a simulated example and show some exploratory results on the regulation of gluconeogenesis by Cat8 and Sip4 pathway using the expression dataset that have been published in literature. Then, the empirical results of these three researches are presented in Chapter 5. Finally, we report the conclusions and discuss future works in Chapter 6.



# Chapter 2

## Data Collection and Statistical Application on Emotion Detection

### 2.1 Collection of The Data

In the research of emotion recognition, the collection of physiological signals plays a important role for next analysis. In this study, the database of physiological signals and corresponding emotional states were collected and obtained from the Center for Measurement Standards of the Industrial Technology Research Institute (ITRI) in Taiwan.

The first group included two subjects, Jane and Alice; they are both female and in their twenties. Every morning between 8:30 am to 10:00 am, they were invited to our laboratory. They were asked to feel a neutral emotion for 200 seconds first, followed by an emotion of anger for at least 120 seconds and finish with a emotion of joy for at least 120 seconds. Meanwhile, those physiological signals were measured and recorded by MP100 system in BIOPAC (<http://www.biopac.com>). Regarding the approach in eliciting emotion, the method we used is similar to the efforts pioneered by Picard et al. (2001) with a slight modification. The methodology is subject-elicited instead of event-elicited, open-recording and emotion-purpose. To prevent differences caused by different external stimulations on different days, we do not rely on any auxiliaries to arouse the emotions of subjects. The subjects were

simply asked to feel an emotion without any assistance such as movies, voices or any other outer stimulus; namely, we do not employ a rigorous Clynes protocol as Picard et al. did. Data gathered from 11 days were used in this study. The default sampling rates were 256 points in one second for each state of emotion. An example of every emotional state is given in Figure 2.1-2.3.

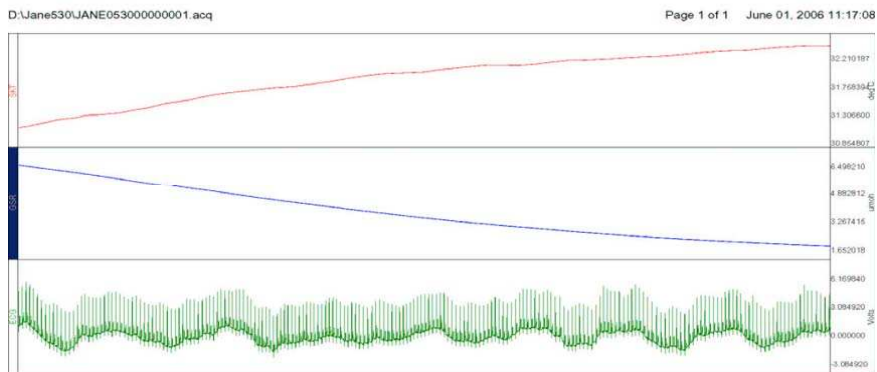


Figure 2.1: Three physiological signals were recorded when the subject was asked to feel neutral. From top to bottom: skin temperature variation (SKT), galvanic skin response (GSR) and electrocardiogram (ECG). The physiological signals were sampling at 256 samples for every second and the measured times were 200 seconds.

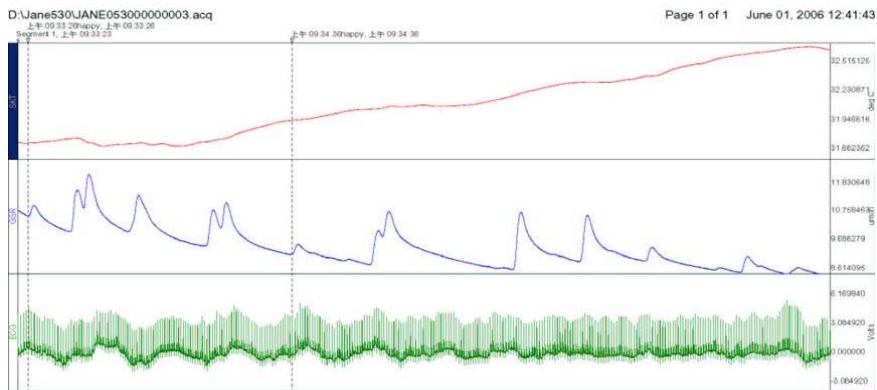


Figure 2.2: Three physiological signals were recorded when the subject was asked to feel joyful. The physiological signals were sampling at 256 samples for every second and the measured times were 120 seconds.

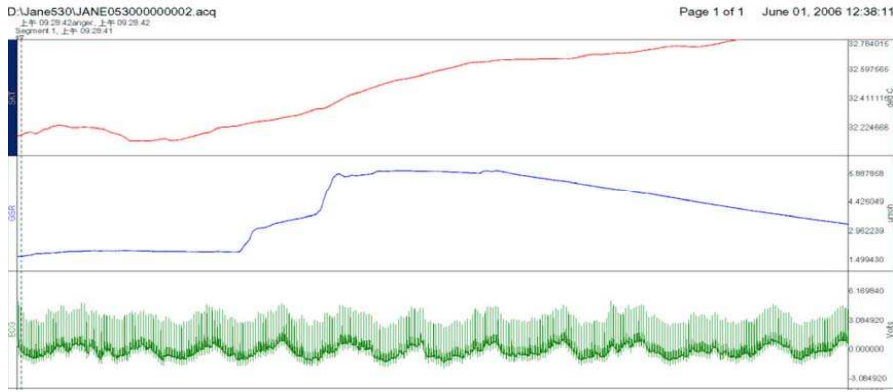


Figure 2.3: Three physiological signals were recorded when the subject was asked to feel angry. The physiological signals were sampling at 256 samples for every second and the measured times were 120 seconds.

For the second dataset, the subjects we used were adults: five men and five women aged from twenty to thirty years. Every morning the subjects were invited to our laboratory with controlled temperature and humidity. At the first practice, the subjects were in a dark place without any voice or music for eliciting a neutral mood within four minutes. Then, to begin the negative emotion eliciting stage, the subjects received eight different pictures with negative expressions, and each picture was broadcasted for thirty seconds. At the same time, the subjects were asked to feel the negative emotion under the stimulus of pictures. Then, the positive emotion eliciting stage was implemented using the same protocol. In the meantime, the physiological signals of the subjects were also measured and recorded by a MP100 system in BIOPAC over the whole experiment. For every subject, the data we gather are from using different pictures over seven days.

After gathering good affective data, the next step was the extraction of representative features from physiological signals. In this study, we would extract 6 features from the collected SKT data, 6 features from the GSR data and 18 features from the ECG data. Then, the daily dependence and personal dependence would be corrected by the statistical technique of MANOVA. Finally, the methodology of

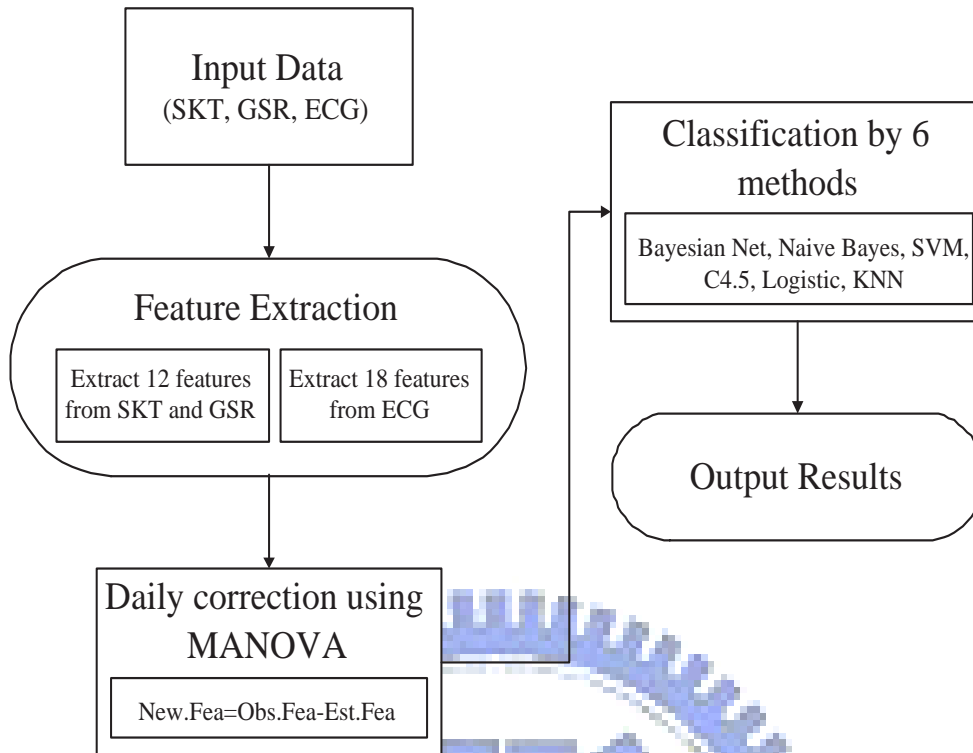


Figure 2.4: Flow chart in the study of emotion recognition.

leave-one-out cross-validation was performed to evaluate the prediction accuracies of six classifiers. The flow chart of the proposed emotion recognition system is given in Figure 2.4.

## 2.2 Features Extraction

Much research has shown significant correlation between physiological signals and emotional status. However, unlike vision or speech recognition, physiological signals in different emotion statuses are not easy to be distinguished by a person immediately. Hence, it is very important to extract representative features that characterize main patterns from the raw physiological signals for classification pattern. For completeness, we would consider most of the features proposed from other literature (Picard et al., 2001; Kim et al., 2004).

For the physiological signals of GSR and SKT, we would use the same features

as Picard et al. (2001). Those six statistic features were the mean, standard deviation, mean of the absolute values of the first difference, standard deviation of the first difference, mean of the absolute values of the second difference and standard deviation of the second difference of the sequence.

The physiological signals of ECG had been calibrated to heart rate variability (HRV) with baseline correction and their R peaks detection. Then, six statistic features were considered as well. In addition, twelve features were extracted from the power spectrum transformation, where the range of the high-frequency (HF) was set as 0.15~0.40 MHz, the median-frequency (MF) was set as 0.08~0.15 MHz and the low-frequency (LF) was set as 0.04~0.08 MHz. In this study, the twelve features we selected were LF, MF, HF, TOTAL (LF+MF+HF), LF/TOTAL, MF/TOTAL, HF/TOTAL, LF/HF, MF/HF, (LF+MF)/HF, (LF+MF)/TOTAL and median of HRV.

## 2.3 Daily and Personal Correction

### 2.3.1 The Problem of Day-effects and Person-effects

There are many external stimuli, such as temperature and humidity, which can affect a person's physiological signals. In addition, a person's diet and sleep patterns can also cause variations in physiology. Hence, a person could have a different expression of the same physiological signal on different days even when he experiences the same emotion. Although we have made an effort to control these annoying factors, there are still some factors, such as hormones or a person's baseline mood, that are not controllable. Therefore, we must remove the day-effects for the emotion recognition study.

In a previous study, Picard proposed some methods to handle the problem of daily variations. Suppose we let the notation  $D$  and  $F$  as the number of experimental days and the number of features, respectively. In the method of day matrix for



handling day-dependence, the method Picard proposed have to enlarge the original  $D \times F$  matrix as  $D \times (F + D - 1)$  matrix. Hence, if the experimental days are long, we must have a large amount of training day data, and consequently the computational overhead would be increased. Even though another method of baseline matrix for handling day-dependence would have avoided the above defect, the state of neutral emotion would be used as the baseline. Hence, we have to lose the opportunity to recognize the neutral emotion, and our number of states of emotion would be reduced.

Besides, in most previous studies of affective status from multiple subjects, the emotion recognition system treated the subjects and physiological signals independently over the same emotional status. However, because of people with different characteristics such as sex, age, weight and so forth, the physiological signals of different subjects would have different expressions even they are experiencing the same emotion. Hence, it is necessary to develop an algorithm or method that can compensate the personal variations and day-to-day variations.

### 2.3.2 MANOVA

Since the problem of personal variations and daily variations would significantly influence the pattern classification in the system of emotion detection, we must remove the day-effects and person-effects for the emotion recognition study. In this project, we use the technique of multivariate analysis of variance (MANOVA), which can be used even on a large number of experimental days; in the meanwhile, it doesn't have to reduce the number of states of emotion. After getting those 30 features from physiological signal of ECG, SKT and GSR, we would transform the features by the statistical technique MANOVA to remove the day-effects and person-effects. The MANOVA in this study is expressed as Eq. (2.1).

$$Z_{ijkl} = \mu_i + \tau_{ij} + \tau_{ik} + \tau_{ijk} + e_{ijkl} \quad (2.1)$$

where  $i = 1, 2, \dots, I$ ,  $j = 1, 2, \dots, J$ ,  $k = 1, 2, \dots, K$ , and  $l = 1, 2, \dots, L$ .

The notation of  $Z_{ijkl}$  represents the value of  $i^{th}$  feature measured in  $j^{th}$  subjects,  $k^{th}$  days, and  $l^{th}$  sample. For the first database, the value of  $I$  is 30,  $J$  is 1,  $K$  is 11 and  $L$  is 3. In the second database, the value of  $I$  is 30,  $J$  is 10,  $K$  is 7 and  $L$  is 3. We let  $Z_{jkl} = (Z_{1jkl}, Z_{2jkl}, \dots, Z_{Ijkl})^T$ ,  $\mu = (\mu_1, \mu_2, \dots, \mu_I)^T$ ,  $\tau_j = (\tau_{1j}, \tau_{2j}, \dots, \tau_{Ij})^T$ ,  $\tau_k = (\tau_{1k}, \tau_{2k}, \dots, \tau_{Ik})^T$ ,  $\tau_{jk} = (\tau_{1jk}, \tau_{2jk}, \dots, \tau_{Ijk})^T$ , and  $e_{jkl} = (e_{1jkl}, e_{2jkl}, \dots, e_{Ijkl})^T$ . Eq. (2.1) can be re-expressed as Eq. (2.2)

$$Z_{jkl} = \mu + \tau_j + \tau_k + \tau_{jk} + e_{jkl} \quad (2.2)$$

where  $j = 1, 2, \dots, J$ ,  $k = 1, 2, \dots, K$ , and  $l = 1, 2, \dots, L$ .

The value  $\mu$  is an overall mean value, the value  $\tau_j$  represents the  $j^{th}$  personal effect, the value  $\tau_k$  represents the  $k^{th}$  daily effect, and the value  $\tau_{jk}$  represents the interact effect of daily and personal factor with the constraints that  $\sum_{j=1}^J \tau_j = 0$ ,  $\sum_{k=1}^K \tau_k = 0$ , and  $\sum_{j=1}^J \sum_{k=1}^K \tau_{jk} = 0$ . The I-dimensional error vector  $e_{jkl} = (e_{1jkl}, e_{2jkl}, \dots, e_{Ijkl})^T$  follows an I-dimensional multivariate distribution with a zero mean vector and a positive definite matrix  $\Sigma$ . Hence, the least squared estimates of  $\hat{\mu}$ ,  $\hat{\tau}_j$ ,  $\hat{\tau}_k$  and  $\hat{\tau}_{jk}$  are  $\bar{Z}$ ,  $\bar{Z}_j - \bar{Z}$ ,  $\bar{Z}_k - \bar{Z}$  and  $\bar{Z}_{jk} - \bar{Z}_j - \bar{Z}_k + \bar{Z}$  respectively, where  $\bar{Z} = \frac{1}{JKL} \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L Z_{jkl}$ ,  $\bar{Z}_j = \frac{1}{KL} \sum_{k=1}^K \sum_{l=1}^L Z_{jkl}$ ,  $\bar{Z}_k = \frac{1}{JL} \sum_{j=1}^J \sum_{l=1}^L Z_{jkl}$ ,  $\bar{Z}_{jk} = \frac{1}{L} \sum_{l=1}^L Z_{jkl}$ . Therefore the estimate  $Z_{jkl} - \hat{\tau}_j - \hat{\tau}_k - \hat{\tau}_{jk} = Z_{jkl} - \bar{Z}_{jk} + \bar{Z}$  can be used to represent data after correction and we will use  $X_{ijkl} = Z_{ijkl} - \bar{Z}_{ijk} + \bar{Z}_i$  as our attribute in the following classification methods.

For the comparison of two classification results, we treat the result of discrimination as a Bernoulli trail for every sample. Then, two sample t-tests could be applied in testing the difference between the classifiers. In this study, we use the p-value of the statistical improvement to compare the results of classification with and without daily and personal correction by MANOVA.

## 2.4 Pattern Classification

Tools of machine learning could be applied to discriminate the emotional states

by the physiological signals. After daily and personal correction, we used the estimator  $X_{ijkl} = Z_{ijkl} - \bar{Z}_{ijk} + \bar{Z}_i$  as our attribute for pattern classification of the emotional state  $Y_{jkl}$ , which represented the emotional state in  $j^{th}$  subject, on the  $k^{th}$  day, and for the  $l^{th}$  sample. We let the variable  $Y$  represent the emotional status and the variable  $X_i$  represent the value of  $i^{th}$  feature after removing the daily and personal correction. Six selected classifiers were tested for their performance and accuracy using the method of leave-one-out cross-validation. All of these six classification methods were performed by the software Weka (<http://www.cs.waikato.ac.nz/ml/weka>), and all of the classifiers used the default option in Weka. Further investigation of other options for classifiers in Weka could be studied in the future. The methods of classifiers were described as below.

### 2.4.1 Bayesian Network

A Bayesian network, also called Bayes nets, is a directed acyclic graph (DAG) which consists of two components. The first component  $G$  comprises vertices corresponding to a set of variables  $V = \{V_1, V_2, \dots, V_N\}$  and a set of directed edges between variables with the Markov properties. The second component  $\theta$  is attached the potential table  $P(V_i|U_{V_i})$ , for each variable  $V_i$  in  $V$  with the corresponding parents nodes  $U_{V_i}$  (Pearl, 1988; Jensen, 2001). Given the structure  $G$  and the parameter  $\theta$ , the joint probability distribution can be written as Eq. (2.3):

$$P(V) = \prod_{i=1}^N P(V_i|U_{V_i}). \quad (2.3)$$

For the purpose of learning take place in a Bayesian networks, we have to reconstruct the network structure and the field values. In this study, we apply the hill climbing algorithm and simple estimator to reconstruct the network and estimate the parameters. After getting the network structure, we used junction tree methods which can convert our DAG to a tree by clustering variables (Lauritzen and Spiegelhalt, 1988). Then an efficient algorithm using belief propagation can be applied for our inference. In our study, we would use the estimator  $X_1, X_2, \dots, X_I$

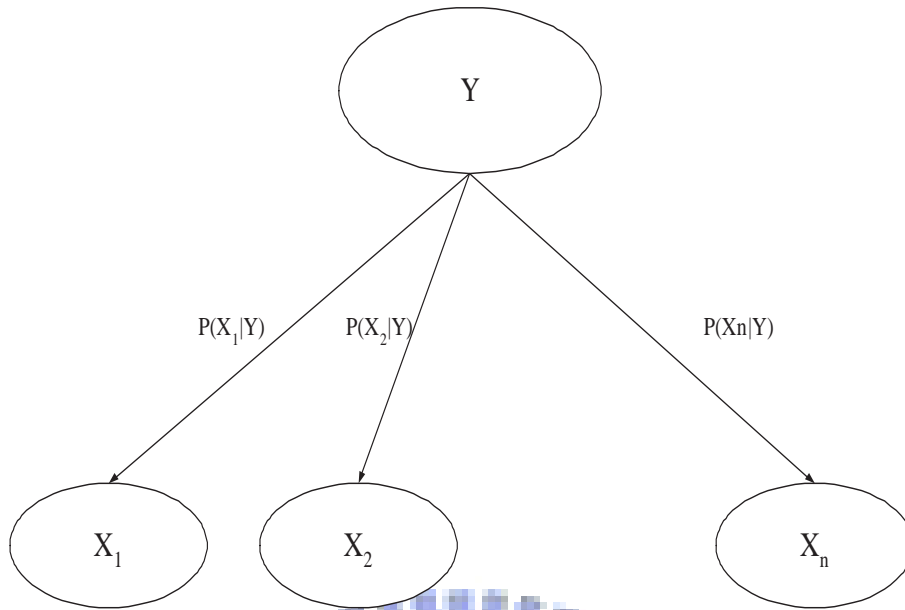


Figure 2.5: The network structure of the naive Bayesian classifier.

and  $Y$  as the prediction variables  $V = \{V_1, V_2, \dots, V_{I+1}\}$  and calculate the conditional distribution of  $Y$  given the observation  $X_1, X_2, \dots, X_I$  in the constructed Bayesian network structure.

## 2.4.2 Naive Bayesian

A naive Bayesian classifier is a simple approach based on the Bayes' theorem. The network structure is illustrated in Figure 2.5. There are two assumptions in the naive Bayesian classifier as follows (John and Langley, 1995). (i) Given the class attribute ( $Y$ ), the predictive attributes ( $X_1, X_2, \dots, X_I$ ) are independent. (ii) There were no other attributes affecting the prediction process. By the Bayes' theorem,

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{P(X = x)}. \quad (2.4)$$

We can predict the class attribute by finding  $y$  that maximizes  $P(Y = y|X = x)$  in Eq. (2.4) given the predictive attributes  $x$ . As the predictive attributes ( $X_1, X_2, \dots, X_I$ ) are assumed to be conditionally independent, we have

$$P(X = x|Y = y) = \prod_{i=1}^I P(X_i = x_i|Y = y). \quad (2.5)$$

For the numeric attributes, we would assume that  $X_i$  is distributed as  $N(\mu_{iy}, \sigma_{iy}^2)$  given the class  $Y = y$  for every  $i = 1, 2, \dots, I$ . Hence, we can estimate the parameters by the maximum likelihood estimates for each class.

### 2.4.3 Support Vector Machine

Support vector machine (SVM) (Vapnik, 1998) is a popular classification method used by a lot of research currently being conducted in the field of emotion recognition (Kim et al., 2004; Chuang and Shih, 2006). Suppose  $\{(x_1^*, y_1^*), (x_2^*, y_2^*), \dots, (x_n^*, y_n^*)\}$  is the training set, where  $y_i^*$  is 1 or -1, denoting whether  $x_i^*$  belongs to one of two classes. In SVM, it is aimed to minimize the cost function  $\frac{1}{2}w^T w + C \sum_{i=1}^n \xi_i$  under the constraints  $y_i^*(w^T x_i^* + b) \geq 1 - \xi_i$  for  $i = 1, 2, \dots, n$ . By using the Lagrange multiplier method, the original problem can be transformed as optimizing  $\alpha_i$ 's in Eq. (2.6).

$$\arg \max_{\alpha} Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i^* y_j^* x_i^{*T} x_j^* \quad s.t. \quad 0 \leq \alpha_i \leq C \quad \forall i; \quad \sum_{i=1}^n \alpha_i y_i^* = 0. \quad (2.6)$$

After obtaining  $\alpha_i$ , we can apply the following decision function for prediction using the new predictive attribute of  $x_{new}^*$ :  $f(x_{new}^*) = \text{sign}(\sum_{i=1}^n y_i^* \alpha_i K(x_{new}, x_i^*) + b)$ , where  $K()$  is the kernel function. In this study, we use the Gaussian kernel and the sequential minimal optimization (SMO) algorithm (Keerthi et al., 2001). Besides, because our case has multiple classes (three emotional statuses), we used the approach of pairwise classification by the one-against-one approach in the SVM classification method.

### 2.4.4 Decision Tree of C4.5

Decision tree is also a common method used in classification (Hunt et al., 1966). C4.5 is a hierarchical data structure using the divide-and-conquer strategy to grow decision trees (Quinlan, 1993). In decision trees, each decision node using a test

function to partition original data  $D$  into subsets  $D_1, D_2, \dots, D_n$ . Suppose the set  $D$  consists of  $C$  numbers of classes and  $p(D, j)$  denotes the proportion of cases in  $D$  that belongs to the  $j$ th class. We can define the information gain by a test  $T$  with  $m$  outcomes as Eq. (2.7):

$$\text{Gain}(D, T) = \text{Info}(D) - \sum_{i=1}^m \frac{|D_i|}{|D|} \times \text{Info}(D_i). \quad (2.7)$$

where  $\text{Info}(D) = -\sum_{j=1}^C p(D, j) \times \log(p(D, j))$  and it can reach its maximal when there is one case left in each subset  $D_i$ . The split information is defined as Eq. (2.8):

$$\text{Split}(D, T) = -\sum_{i=1}^m \frac{|D_i|}{|D|} \times \log\left(\frac{|D_i|}{|D|}\right). \quad (2.8)$$

For every possible test, the ratio of its information gain over its split information is assessed and the test with maximum gain ratio is selected.

## 2.4.5 Logistic Model

Logistic regression is a classical method to model category data for classification (Le Cessie and Van Houwelingen, 1992). Suppose there are  $n$  samples with  $c$  classes and  $I$  attributes. The parameter matrix  $B$  is calculated as an  $I \times (c - 1)$  matrix. The probability that the  $i^{\text{th}}$  sample, given the value of  $x_i^*$ , in the  $j^{\text{th}}$  class but not in the last  $c^{\text{th}}$  class is shown in Eq. (2.9).

$$P_j(x_i^*) = \frac{\exp(x_i^* B_j)}{\sum_{k=1}^{c-1} \exp(x_i^* B_k) + 1}, \text{ where } j = 1, 2, \dots, c - 1. \quad (2.9)$$

The probability that the  $i^{\text{th}}$  sample, given the value of  $x_i^*$ , in the last  $c^{\text{th}}$  class is shown in Eq. (2.10).

$$P_c(x_i^*) = 1 - \sum_{k=1}^{c-1} P_k(x_i^*) = \frac{1}{\sum_{k=1}^{c-1} \exp(x_i^* B_k) + 1}. \quad (2.10)$$

The log-likelihood  $l$  of the data  $(K, X)$  under this model is shown in Eq. (2.11).

$$l(\beta) = \sum_{i=1}^n \left\{ \sum_{k=1}^{c-1} K_{ik}^* \ln(P_k(x_i^*)) + (1 - \sum_{k=1}^{c-1} K_{ik}^*) \ln(1 - \sum_{k=1}^{c-1} P_k(x_i^*)) \right\}. \quad (2.11)$$

The indicator variable  $K_{ij}^* = 1$  if the  $i^{th}$  sample belongs to the  $j^{th}$  class, where  $j \neq c$ . Otherwise,  $K_{ij}^* = 0$  if the  $i^{th}$  sample belongs to the last  $c^{th}$  class. The parameter matrix  $B$  can be estimated by the maximize likelihood estimates of the likelihood function,  $l(\beta)$ .

## 2.4.6 K-Nearest Neighbor (KNN)

The  $k$ -nearest neighbor (KNN) algorithm is one of the classical classification methods that have wide applications (Aha et al., 1991). KNN compares the similarity between testing data and every training data. Then it uses the top  $k$  similarity categories of training data to decide the category of the testing data by a weighted vote. For any testing data of  $H$  and training data of  $\{G_1, G_2, \dots, G_n\}$ , we would classify the category of  $H$  as Eq. (2.12).

$$C(H) = \arg \max_m \sum_{G_i \in S} Sim(H, G_i) I(G_i, C_m). \quad (2.12)$$

The notation of  $Sim(H, G_i)$  is the similarity measure of  $H$  and  $G_i$ . The set  $S = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_k\}$  is the data set closed to the testing point  $H$ , and the notation of  $I(G_i, C_m) \in \{0, 1\}$  indicates whether  $G_i$  belongs to  $C_m$ . If there are tie cases in the classification, we will use the group with a minimal index as the corresponding category of testing data. In this study, we would use the Euclidean distance as the similarity measure and choose the number of nearest neighbors  $k=3$ .

# Chapter 3

## Data Collection and Analysis on Yeast Genes of Microarray Studies

### 3.1 Materials and Microarray Experiment

*S. cerevisiae* was used as the model organism for studying expression evolution because it is experimentally easier to manipulate and its genetics and genomics are better known than most other eukaryotes. A lab strain (BY4741) and a wild isolate (RM11) are used in this project. The BY strain is a direct descendant of S288C, which was generated in the 1960s. The RM strain is a haploid derivative from a California vineyard.

The cultures of BY4741 and RM11-1a were separately started at  $OD_{600}=0.1$  and were grown in YPAD media (which contains 2% glucose) at 30°C with 250 rpm shaking. Overnight cultures of BY4741 and RM11-1a were used for preparing the starting cultures. The yeast cells were harvested at 4hr, 5hr, 6hr, 7hr, 8hr, 9hr, 10hr, 11hr, 12hr, 13hr, 14hr, 16hr, 18hr and 20hr after inoculation and the glucose content of media at each time point will also be measured. Each microarray experiment was conducted with 0.5 $\mu$ g of purified mRNA from each strain. The microarray was scanned with GenePix 4000B microarray scanner (Axon Instruments) with the GenePix 5 software package.

We currently adopt the reference design in the array experiments. The cDNA



sample from the 4hr culture was labeled with Cy3 and used as the reference sample, whereas RNA sampling from different time points were labeled with Cy5 and used as experimental samples. Each experiment was repeated four times. Dye-swapping was also performed in each set of experiments to eliminate dye bias. Theoretically, loop design may be more efficient in finding significant genes when the number of variety in treatment is small or time course experiments are considered. However, we prefer reference design because there is a common reference that is easy to interpret and to include new microarray data with the same reference. Importantly, reference designs are more tolerant to experimental errors.

## 3.2 Data Extraction

We used several statistical analyses to eliminate background noise and to obtain more meaningful expression data. First, the background correction was applied to remove the background median from the foreground median to obtain the expression intensity for every dye in one spot. If the intensity value after background correlation is small than zero, we treated the experimental value of this spot as an invalid value because the dye efficiencies of Cy3 and Cy5 could be different. However, this kind of dye effect can be normalized by the factor between the medians of Cy3 and Cy5 intensities in one microarray. There are two duplicated spots for one gene, and there are two swapped arrays. Therefore, there are four spots in total for one gene per strain at one time point that are obtained as follows.

$$\begin{aligned} \text{If Swap}=0, \text{Ratio}_{ijr} &= \frac{I532_{ijr}/\text{Median}_{j=1,\dots,6368,r=1,2}\{I532_{ijr} \text{ in array } i\}}{I635_{ijr}/\text{Median}_{j=1,\dots,6368,r=1,2}\{I635_{ijr} \text{ in array } i\}}, \\ \text{If Swap}=1, \text{Ratio}_{ijr} &= \frac{I635_{ijr}/\text{Median}_{j=1,\dots,6368,r=1,2}\{I635_{ijr} \text{ in array } i\}}{I532_{ijr}/\text{Median}_{j=1,\dots,6368,r=1,2}\{I532_{ijr} \text{ in array } i\}}, \end{aligned}$$

where

$$I532_{ij} = F532\_Median_{ij} - B532\_Median_{ij} \text{ for Cy3,}$$

$$I635_{ij} = F635\_Median_{ij} - B635\_Median_{ij} \text{ for Cy5,}$$

$$i = 1, 2, \dots, 176 \text{ (176 array files in total),}$$

$j = 1, 2, \dots, 6368$  (6368 genes in total),  
 $r = 1, 2$  (two replicated genes in every array).

The average of the valid ratios in these four ratios for one gene was used for further normalizing the dye and block effects from a pair of two swapped microarrays with two duplicated spots in one array. Furthermore, the  $\log_2$  transformation of ratio was used to evaluate the relative gene expression of one gene in a strain at a specific time referring to the common reference at t4.

### 3.3 Strain Normalization

Because the denominators of expression ratios are different in BY and RM strains, we can adjust them to have the same denominator for further comparisons. We performed another six microarrays with two yeast strains at t4. Hence we can get the ratio of RM\_t4/BY\_t4 for each gene by the average of expression ratios in these six microarrays. This ratio was used to adjust the denominator as the expression of BY\_t4 for every gene in every microarray.

The analysis flow chart is illustrated in Figure 3.1. The microarray data in experiment 1, 3 and 4 were used as the training set because they have common experiment time points. The microarray data in experiment 2 was used as the test set to evaluate the performance of analysis results from the training set. Genes were filtered by the regression coefficients of expression vs. time in the training set. These unfiltered genes were clustered by the methods of hierarchical clustering and curve clustering by the training set of microarray data. The clustering method was selected based on the performance of clustering results in the training and test sets. The numbers of clusters were also determined accordingly. For every cluster, the time shift was estimated by regression tests between gene expressions and glucose consumptions. The details of analyses are discussed in the following sections.

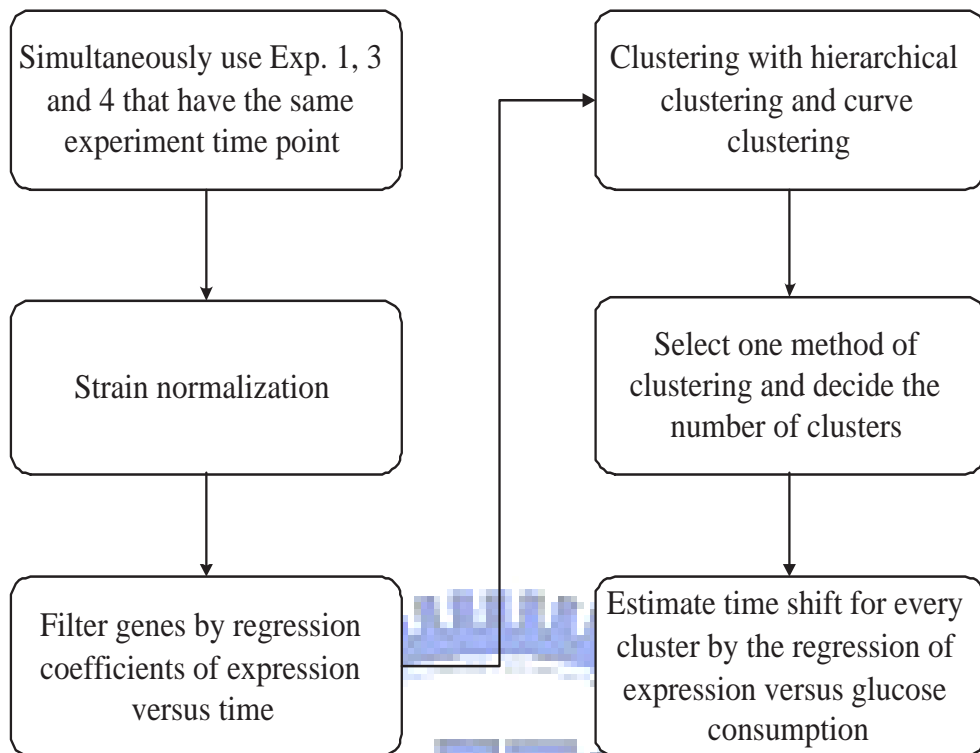


Figure 3.1: The flowchart in the study of of yeast genes.

### 3.4 Gene Filtering

A regression line was used to detect expression trends for gene expression vs. time for every gene in one strain and one experiment in the training set (Exp. 1, 3 and 4). The following regression model was used for every gene in one strain and one experiment,

$$\log(Ratio) = \alpha_0 + \alpha_1 Time + \epsilon \quad (3.1)$$

where  $\log(Ratio)$  is the log ratio of gene expression,  $Time$  is the time point ranging through 5 to 13,  $\alpha_0$  is the intercept,  $\alpha_1$  is the regression coefficient of slope and  $\epsilon$  is the random noise.

The goal of gene filtering is to filter genes that do not have significantly and consistently differential expressions over time in the training set of microarray data. The regression model in (3.1) is used to detect the expression trend for every gene

in one strain and one experiment.

For every gene in one strain, there are three regression slopes in experiments 1, 3 and 4. The CV which is calculated as the ratio of standard deviation over the average of three slopes is a measure of SNR. If the CV value is large, then the expression slopes vary a lot, or the average is small among three experiments. Hence, those genes with CV values larger than a threshold can be filtered, and a threshold of 2.1 is used in this study. Then, the average of three slopes is used to partition the unfiltered genes to three groups. If the averages of three slopes in BY and RM strains are of the same signs,  $(+, +)$  or  $(-, -)$ , then they are positively correlated. Otherwise, they are  $(+, -)$  or  $(-, +)$ , which are negatively correlated. A lot of unfiltered genes have patterns of positive correlation in two strains, and few genes have patterns of negative correlation. For the group of positive correlations in two strains, two subgroups are constituted using a threshold for the absolute value of the difference between the average slopes in two strains, like a threshold of 0.3. This partition is considered to keep genes that have a large expression variation in one strain but not in the other strain.

Consequently, there are three different groups we selected. For the first group of positive correlation and large differences of average slopes in two strains, all unfiltered genes are kept because they have a strain with large expression variation but not in the other strain. For the second group of positive correlation and small differences of average slopes in two strains, the maximum of absolute values of average slopes is used to keep genes with large expression variation in one strain, like a threshold of 0.3499. For the third group of negative correlation in two strains, the maximum of the absolute values of average slopes is used to keep genes with large expression variation in one strain, like a threshold of 0.2. As a result, there are 490 genes kept in this study.

The above approach of gene filtering is used to keep genes that could have significant expression patterns in this study. These 490 genes will be further selected after

checking the clustering consistency and will be investigated in the later chapters. Other methods of gene filtering could be studied in the future.

### 3.5 Cluster Analysis

The expression profiles of unfiltered genes will be used to perform cluster analysis. Suppose one gene is clustered into groups g1, g2 and g3 in a training set of three experiments after clustering by one method. Let M1, M2 and M3 be the mean expression value of each group at one time point. Then, the predicted expression value for the gene at that time point is defined to be the average of M1, M2 and M3. Thus the prediction square error (PSE) is the value of the square of the error between a predicted expression and the observed expression of the gene in the test set. Hence the PSE is as follows:

$$PSE = \sum_{i=1}^{490} \sum_{j=1}^{14} \frac{(R_{2,ij} - R_{pred,ij})^2}{14} \quad (3.2)$$

where  $R_{2,ij}$  means the gene expression of the  $i^{th}$  gene in the  $j^{th}$  microarray data, and  $R_{Pred,ij}$  is its predicted value by the clustering method. For every gene, the microarray data contain 14 gene expressions at seven time points for two strains in experiment 2. If the PSE of one clustering method is small, then this clustering method is a good method. Through the comparisons of PSEs, we can select one method from different clustering methods.

The clustering consistency for one gene in the clustering results using three experiments in the training set will be also checked. That is, it will be examined whether the expression time profile of one gene in different experiments will be clustered into the same group. One example is illustrated in Figure 3.2. Genes with clustering consistency will be selected to find the representative curves in every group.

#### 3.5.1 Hierarchical Clustering

Hierarchical clustering is a nonparametric method to cluster data (Eisen et al.,

	selected genes	result of clustering
Exp 1	$g_1$	3
	$g_2$	5
	...	...
	$g_k$	7
Exp 3	$g_1$	1
	$g_2$	5
	...	...
	$g_k$	7
Exp 4	$g_1$	2
	$g_2$	5
	...	...
	$g_k$	4

Figure 3.2: In this case, gene  $g_2$  is considered to have clustering consistency.

1998). The basic idea of hierarchical clustering is to construct a tree based on the similarity (or dissimilarity) among data. If the observations of two data are similar, they will be clustered into the same group. Hierarchical clustering depends on a distance matrix,  $D$ , which records the pairwise distance for expressions of any two data. So, it is a symmetric matrix. The following two distances are commonly used in literature and they will be investigated in this study.

Euclidean distance:

$$d(z^r, z^s) = \left[ \sum_{j=1}^d (z_j^r - z_j^s)^2 \right]^{1/2} \quad (3.3)$$

(Pearson's) Correlation distance:

$$d(z^r, z^s) = 1 - \text{cor}(z^r, z^s) = 1 - \frac{\text{cov}(z^r, z^s)}{\sqrt{\text{var}(z^r) \times \text{var}(z^s)}} \quad (3.4)$$

where  $z^r$  and  $z^s$  are two observation vectors in  $d$ -dimensions,  $z_j^r$  and  $z_j^s$  are  $j^{\text{th}}$  components of two observation vector, and  $\text{cov}$  and  $\text{var}$  are the sample covariance and variance.

In the second step, it is necessary to define the linkage that is the distance between two groups. There are three kinds of linkages that are commonly considered in literature.

Single linkage: the distance is defined as the smallest distance between all possible pairs of elements of the two groups,  $G_i$  and  $G_j$ :

$$d(G_i, G_j) = \min_{z^r \in G_i, z^s \in G_j} d(z^r, z^s). \quad (3.5)$$

Complete linkage: the distance between two groups is taken as the largest distance between all possible pairs:

$$d(G_i, G_j) = \max_{z^r \in G_i, z^s \in G_j} d(z^r, z^s). \quad (3.6)$$

Average linkage: the average of distances between all possible pairs in two groups:

$$d(G_i, G_j) = \text{average}_{z^r \in G_i, z^s \in G_j} d(z^r, z^s). \quad (3.7)$$

The algorithm of agglomerative clustering will be used for hierarchical clustering in this study. First, every observation is treated as a group itself. Then similar groups are merged to form larger groups hierarchically until all groups are merged into a single one.

We will try two kinds of distances and three kinds of linkages (single linkage, complete linkage and average linkage) to investigate which combination is better for the log ratio of expressions obtained from microarray data. Therefore, there will be six different results for hierarchical clustering as shown in Figure 3.3.

By comparing PSEs for different cluster sizes, it is observed that the results of hierarchical clustering by Euclidean distance and the complete linkage have the smallest PSE when the cluster size is large than 2. Hence, the hierarchical clustering by Euclidean distance and the complete linkage will be used in this study. The dendrogram of this hierarchical clustering is shown in Figure 3.4

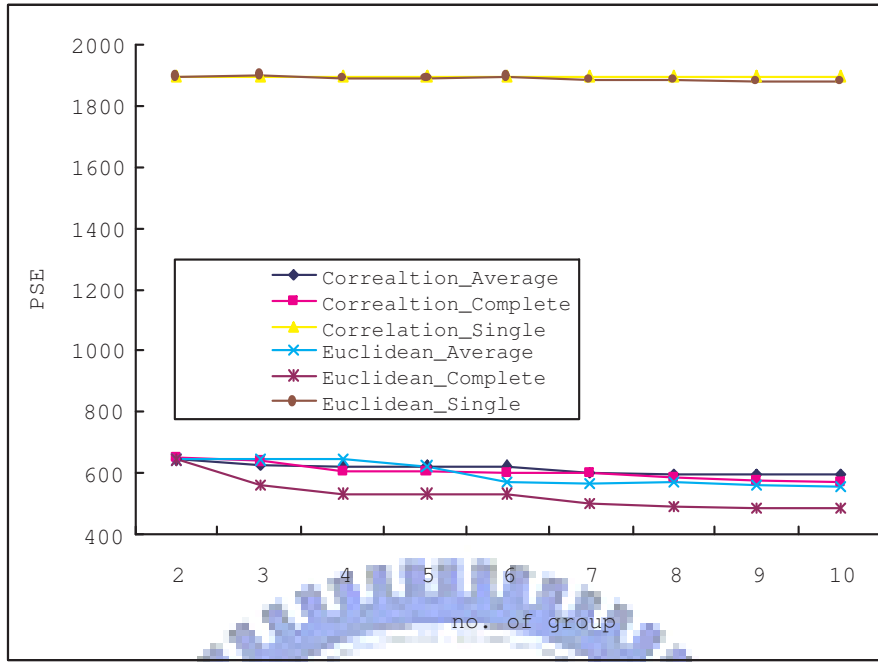


Figure 3.3: Comparisons of PSEs for different cluster sizes are plotted for hierarchical clustering with different settings.

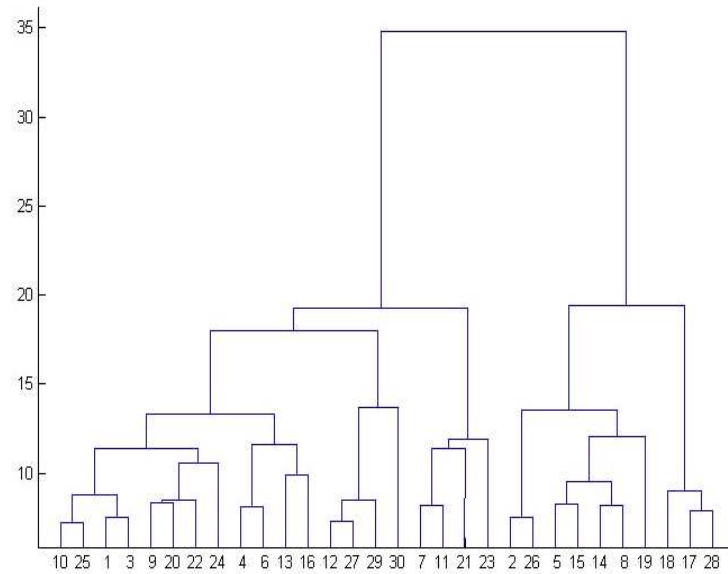


Figure 3.4: The dendrogram of the hierarchical clustering is shown for 30 nodes.



### 3.5.2 Curve Clustering

The clustering method that could be applied to cluster expression profiles can be curve clustering. This method has been proposed to cluster curves based on mixture models (Gaffney, 2004; Gaffney and Smyth, 2004; Gaffney et al., 2007), and the toolbox for matlab is available at (<http://www.ics.uci.edu/~sgaffney/CCT/>). Basically, that method assumed a mixture model with an expectation-maximization (EM) algorithm to estimate parameters in the mixture model, which are reviewed below. Suppose that  $y_i$  is a sequence of curve measurements that are observed at  $n_i$  time points in  $x_i$ . The author defines a cluster-specific conditional probabilistic model, which is denoted as  $p_k(y_i|x_i, \theta_k)$  for the probability distribution in cluster  $k$  with parameters  $\theta_k$ . In this study, the linear polynomial regression model (LRM) is investigated and performed well for the microarray data under investigation. Polynomial regression models of  $y_i$  on  $x_i$  with a Gaussian noise can be summarized with the following equation:

$$y_i = X_i\beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2 I), \quad (3.8)$$

where the  $n_i \times p$  regression matrix  $X_i$  is the Vandermonde matrix evaluated at  $x_i$ ,  $\beta$  is the  $p$ -vector of regression coefficients,  $\epsilon_i$  is the Gaussian noise with mean 0 and covariance matrix  $\sigma^2 I$ . The  $p$ -th order Vandermonde matrix evaluated at  $x_i$  is equal to

$$X_i = \begin{pmatrix} 1 & x_{i1} & x_{i1}^2 & \cdots & x_{i1}^{p-1} \\ 1 & x_{i2} & x_{i2}^2 & \cdots & x_{i2}^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{in} & x_{in}^2 & \cdots & x_{in}^{p-1} \end{pmatrix}.$$

Then, the conditional probability of  $y_i$  give  $x_i$  is distributed as  $N(y_i|X_i\beta, \sigma^2 I)$ .

The polynomial regression mixture model of  $K$  clusters is defined to be:

$$p(y_i|x_i, \theta) = \sum_{k=1}^K \alpha_k p_k(y_i|x_i, \theta_k) = \sum_{k=1}^K \alpha_k N(y_i|X_i\beta_k, \sigma_k^2 I) \quad (3.9)$$

where  $\alpha_k$  is the mixing probability in  $k^{th}$  cluster,  $p_k$  is the conditional probability

of a Gaussian distribution with mean  $X_i\beta_k$  and covariance matrix  $\sigma_k^2 I$ . The log-likelihood function  $N$  observations becomes

$$\log p(\theta|Y, X) = \sum_{i=1}^N \log \left( \sum_{k=1}^K \alpha_k p_k(y_i|x_i, \theta_k) \right) \quad (3.10)$$

The EM algorithm can be applied to obtain the maximum likelihood estimates of parameters of  $\beta_k, \sigma_k^2, \alpha_k, k = 1, 2, \dots, K$  for any fixed cluster size  $K$ . The complete log-likelihood function  $L_c$  can be obtained after assuming a class label variable of the  $i^{th}$  observation,  $z_i$ , as follows:

$$L_c = \sum_{i=1}^N \log \alpha_{z_i} N(y_i|X_i\beta_{z_i}, \sigma_{z_i}^2 I). \quad (3.11)$$

In the E-step, the posterior probability  $p(z_i|y_i, x_i)$  is calculated and denoted as  $w_{ik}$ :

$$w_{ik} = p(z_i = k|y_i, x_i) \propto \alpha_k p_k(y_i|x_i) = \alpha_k N(y_i|X_i\beta_k, \sigma_k^2 I). \quad (3.12)$$

And the conditional expectation  $Q$  is:

$$Q = E[L_c|y_i, x_i] = \sum_{i=1}^N \sum_{k=1}^K w_{ik} \log \alpha_k N(y_i|X_i\beta_k, \sigma_k^2 I). \quad (3.13)$$

In the M-step, we maximize  $Q$  with respect to the parameters  $\beta_k, \sigma_k^2, \alpha_k, k = 1, 2, \dots, K$ . The iterated estimators for parameters turn out to be

$$\hat{\beta}_k = \left[ \sum_{i=1}^N w_{ik} X_i^T X_i \right]^{-1} \sum_{i=1}^N w_{ik} X_i^T y_i, \quad (3.14)$$

$$\hat{\sigma}_k^2 = \frac{1}{\sum_{i=1}^N w_{ik}} \sum_{i=1}^N w_{ik} \| y_i - X_i\beta_k \|^2, \quad (3.15)$$

and

$$\hat{\alpha}_k = \frac{1}{N} \sum_{i=1}^N w_{ik}, \quad (3.16)$$

The method of curve clustering has been applied to cluster observations of latitude and longitude positions in cyclones (Gaffney, 2004; Gaffney and Smyth, 2004). For the analysis of microarray data in this study, we will regard gene expressions of one gene in BY and RM strains at different time points during one experiment as one

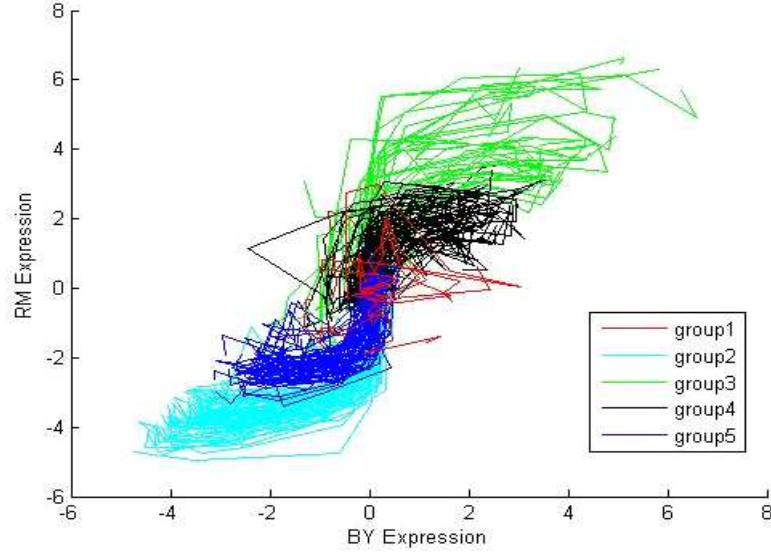


Figure 3.5: The typical results of two dimensional expression curves in experiment 1 for five groups are plotted.

expression curve moved along time in two dimensions of expressions in BY and RM strains. That is, we treat the expression profiles of every gene in one experiment as an observation. The expressions at one time point in BY and RM strain are regarded as a point in two dimensional space for expressions in BY and RM strains. The typical results of two dimensional expression curves for five groups are plotted in Figure 3.5.

The selection for cluster size in curve clustering may be considered by the technique of model selection. A typical method is the Bayesian information criterion (BIC) (Burnham and Anderson, 1998). The value of BIC for the above method of curve clustering is evaluated by the following equation:

$$BIC = -2\log(L_{ML}) + K_a \log N, \quad (3.17)$$

where  $\log(L_{ML})$  is the log-likelihood evaluated at the maximum likelihood estimation,  $K_a$  is the total number of free parameters, and  $N$  is the number of observations. The BIC curve for curve clustering of microarray data in the training set is plotted

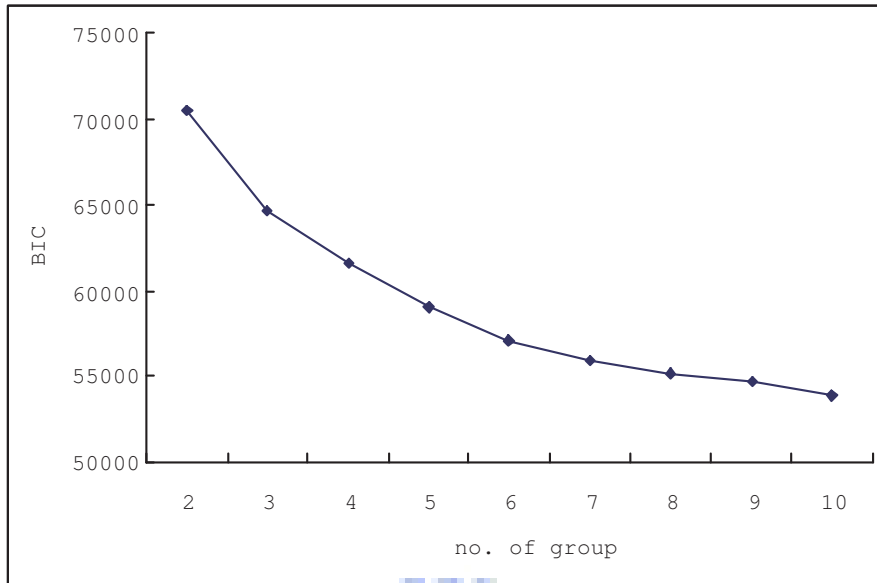


Figure 3.6: Model selection by BIC is shown for curve clustering.

for cluster sizes from 2 to 10 in Figure 3.6. As the BIC curve is decreasing when the cluster size is increasing in Figure 3.6, the method of BIC will tend to select a large cluster size, like 10 in this study. Alternatively, we will also consider other evaluation methods to select a smaller cluster size in this study as reported in Section 5.2.

### 3.6 Regression Models With Time Shift

The analysis of variance (ANOVA) has been applied for microarray data in literature (Kerr and Churchill, 2001; Kerr, 2003; Galindo et al., 2004). In this study, the curves of glucose consumptions can be further incorporated in the model. Furthermore, the time shift between gene expression and glucose consumption shall be considered. Microarray data in different experiments can be combined in statistical models and tests. These statistical models can be applied to every cluster of fewer genes with similar expression profiles to reduce the false errors caused by multiple comparisons of many genes.

The experiment factors of exp, strain, time and gene shall be included in models to investigate the variation of expressions for these factors. The interaction term of gene and time can be included to describe the differences in expression time profiles among genes. The factor of glucose with the parameter of time shift shall be also included to detect the relationship between gene expression and glucose consumption. If the time shift is the same for the expression profiles in both BY and RM strains, we will consider the following regression model for the log ratios of gene expression with other experiment factors:

$$\begin{aligned} \log(\text{Ratio}(\text{time})) &= \mu + \mu_{\text{strain}} + \mu_{\text{time}} + \mu_{\text{exp}} + \mu_{\text{gene}} + \mu_{\text{time} \times \text{gene}} \\ &+ \gamma_{\text{glucose}}(\text{time} + \text{time\_shift}) + \text{error}. \end{aligned} \quad (3.18)$$

If gene expression profiles have different time shifts in BY and RM strains, we will consider estimate the time shift in one strain by using the expression data in one strain only:

$$\begin{aligned} \log(\text{Ratio}(\text{time})) &= \mu + \mu_{\text{time}} + \mu_{\text{exp}} + \mu_{\text{gene}} + \mu_{\text{time} \times \text{gene}} \\ &+ \gamma_{\text{glucose}}(\text{time} + \text{time\_shift}) + \text{error}. \end{aligned} \quad (3.19)$$

With the parameter of time shifts, the above models are nonlinear. For simplicity, we will consider the time shift parameters at fixed values, like -1, 0 and 1. At a fixed value of time shift parameter, the above models become linear and linear regression techniques can be applied. The smallest p-value for testing the hypotheses of  $H_0 : \gamma = 0$  vs.  $H_1 : \gamma \neq 0$  is used to determine the fitted time shift for gene expressions in one cluster. Techniques of nonlinear regression and interpolation may be studied to estimate the shift parameter besides those fixed values in the future.

Different types of hypotheses can be tested based on the above model. For instance, one can consider different regression models with time shifts in glucose separately to investigate whether gene expressions in one group vary before or after the glucose consumption dropped. We set three time shifts as -1, 0, and 1 in this

study. The negative time shift means the gene expression varies after the glucose consumption dropped. The time shift is determined for a group of genes when it will result in a maximum F statistics for testing  $H_0 : \gamma = 0$  vs.  $H_1 : \gamma \neq 0$  among the results of three time shifts as follows:

$$F_{Glucose} = \frac{SS_{Glucose}/1}{SS_{Error}/df_{Error}}, \quad (3.20)$$

The degree of freedom for the sum of squares of Glucose is equal to 1 since the Glucose term is a one-dimensional variable.

Furthermore, one can also check if there are significant differences in strains, time points, experiments, genes, the interactions between time points and genes by similar test statistics. For example, one can consider the following hypotheses,  $H_0$ : the null hypothesis that gene expressions do not vary by times (the time-gene interaction terms of  $\mu_{time \times gene}$  are all equal to zeros); and  $H_1$ : the alternative hypothesis that gene expressions do vary by times (the time-gene interaction terms of  $\mu_{time \times gene}$  are not all equal to zeros). The F statistics become

$$F_{time \times gene} = \frac{SS_{time \times gene}/df_{time \times gene}}{SS_{Error}/df_{Error}}, \quad (3.21)$$

where  $SS_{time \times gene}$  indicates the sum of squares of  $\mu_{time \times gene}$  terms,  $df_{time \times gene}$  indicates its degree of freedom,  $df_{time \times gene} = (\text{number of time points} - 1) \times (\text{number of genes} - 1)$ ;  $SS_{Error}$  indicates the sum of squares of errors, and  $df_{Error}$  indicates the degree of freedom,  $df_{Error} = (\text{number of observations}) - (\text{degrees of freedom of all terms})$ .

# Chapter 4

## Inference of biological pathway by Time Delay Boolean network

### 4.1 Models

#### 4.1.1 Boolean Network

Boolean networks, introduced by Kauffman was used as model of genetic regulatory networks in thirty years ago (Kauffman, 1969). Following (Akutsu and Miyano, 1999), we are going to review the definition of Boolean networks. A Boolean network  $G(V, F)$  is a directed graph consist of two components. The first component  $V = \{v_1, v_2, \dots, v_n\}$  is a set of nodes representing genes, and the second component  $F = \{f_1, f_2, \dots, f_n\}$  is a list of Boolean functions. For every node  $v_i \in V$ , its expression has only two states, ON and OFF. For every boolean function  $f_i(v_{i_1}, v_{i_2}, \dots, v_{i_k}) \in F$ , the input node  $v_{i_1}, v_{i_2}, \dots, v_{i_k}$  is assigned to the node  $v_i$  in the graph. The state of each node  $v_i \in V$  is determined by the Boolean function  $f_i(v_{i_1}, v_{i_2}, \dots, v_{i_k})$ .

For an element  $U \in V$ , an expression pattern  $\psi$  of  $U$  is a function from  $U$  to  $\{0, 1\}$ . For each node  $v_i$ , the gene expression state at time  $t$  is assumed to take either 0 (not-expressed) or 1 (expressed) and is expressed as  $\psi_t(v_i)$ . In a Boolean network, every gene expression pattern at time  $t + 1$  is determined by the gene expression pattern at time  $t$  and the corresponding Boolean function  $F$  (i.e.,

$$\psi_{t+1}(v_i) = f_i(\psi_t(v_{i_1}), \psi_t(v_{i_2}), \dots, \psi_t(v_{i_k})).$$

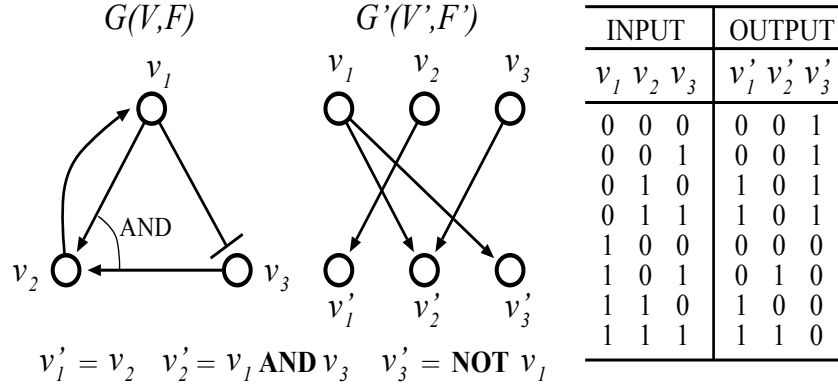


Figure 4.1: A Boolean network  $G(V, F)$ , its wiring diagram  $G'(V', F')$  and the functional dependency table. (Akutsu et al. 1999)

For convenience, we consider the wiring diagram  $G'(V', F')$  of a Boolean network  $G(V, F)$  (See Figure 4.1). For each  $v_i \in V$ , suppose  $v_{i_1}, v_{i_2}, \dots, v_{i_k}$  are the input nodes assigned to  $v_i$ . Then we construct an additional node  $v'_i$  and connected the edge from  $v_{i_j}$  to  $v'_i$  for each  $1 \leq j \leq k$ . That is, the  $\{v_1, \dots, v_n\}$  represent the gene expression pattern at time  $t$  and  $\{v'_1, \dots, v'_n\}$  corresponds to the gene expression pattern at time  $t + 1$ . Hence we can treat the  $\{v_1, \dots, v_n\}$  as the input values and the  $\{v'_1, \dots, v'_n\}$  as the corresponding output values.

#### 4.1.2 The Structure of Time Delay Boolean Network

In the previous subsection, we found that the output gene  $v_i$  at time  $t + 1$  is determined by the input genes  $v_{i_1}, v_{i_2}, \dots, v_{i_k}$  at previous time  $t$ . That is, for every gene  $v_i \in V$ , if the input gene  $\{v_{i_1}, v_{i_2}, \dots, v_{i_k}\}$  at time  $t$  and the Boolean function  $f_i$  is fixed, the gene expression  $v_i$  at the next time  $t + 1$  is determined by  $\psi_{t+1}(v_i) = f_i(\psi_t(v_{i_1}), \psi_t(v_{i_2}), \dots, \psi_t(v_{i_k}))$ . However, in real genetic regulatory situations, the deterministic system always fails because of the misclassification error and noise. Besides, some of the gene express would have the situation of time delay



when the gene influenced by one or several input genes. Hence, it would be much more flexible to use a non-deterministic network system. In this subsection, we consider two relations between the Boolean function and the target gene instead of the deterministic relation.

We define a prerequisite relation between the Boolean function and the target output gene as follows. A Boolean function  $f_i$  with input genes  $v_{i_1}, v_{i_2}, \dots, v_{i_k}$  at time  $t$  is prerequisite for the target gene  $v_i$  at time  $t + 1$ , if the on-status of Boolean function is necessary for the on-status of gene  $v_i$  at time  $t + 1$ , and we denote this by  $f_i(\psi_t(v_{i_1}), \psi_t(v_{i_2}), \dots, \psi_t(v_{i_k})) \prec \psi_{t+1}(v_i)$ . If it does not cause confusion, we omit the notation of  $\psi$  and input genes as denoted by  $f_i \prec v_i$ . Moreover, for every gene  $v_i$ , we use  $\bar{v}_i$  as its dual in this paper. For the prerequisite relation between Boolean function and target gene, we have the following two relations:  $f_i \prec v_i$  and  $f_i \prec \bar{v}_i$ . In this model, we do not consider the situation of a dual Boolean function prerequisite to the target gene, that is  $\bar{f}_i \prec v_i$  and  $\bar{f}_i \prec \bar{v}_i$ . Since for the boolean function whose dual is prerequisite to the target gene, there must be another boolean function which is prerequisite to the target gene. For instance, if  $\bar{f}_i\{v_1, v_2\} \prec v_3$ , where  $f_i\{v_1, v_2\} = (v_1 \text{ AND } v_2)$  then  $f'_i\{v_1, v_2\} \prec v_3$ , where  $f'_i\{v_1, v_2\} = (\bar{v}_1 \text{ OR } \bar{v}_2)$ . Therefore, for the prerequisite relationship, we only consider the Boolean function prerequisite to target gene and Boolean function prerequisite to dual of target gene.

Another relation between Boolean function and target gene is similarity. The Boolean function and target gene are similar if the state of the Boolean function will make the state of the target gene in the same expression, and we denoted this by  $f_i \sim v_i$ . In the same way, we do not consider the negatively similar such as  $f_i \sim \bar{v}_i$  in this study. For the same reasons, if there is one Boolean function which is negatively similar to a target gene, there must exist another Boolean function which is similar to the target gene.

In the model of Time Delay Boolean network we proposed, the output of the gene expression is not totally determined by the input state and Boolean function.

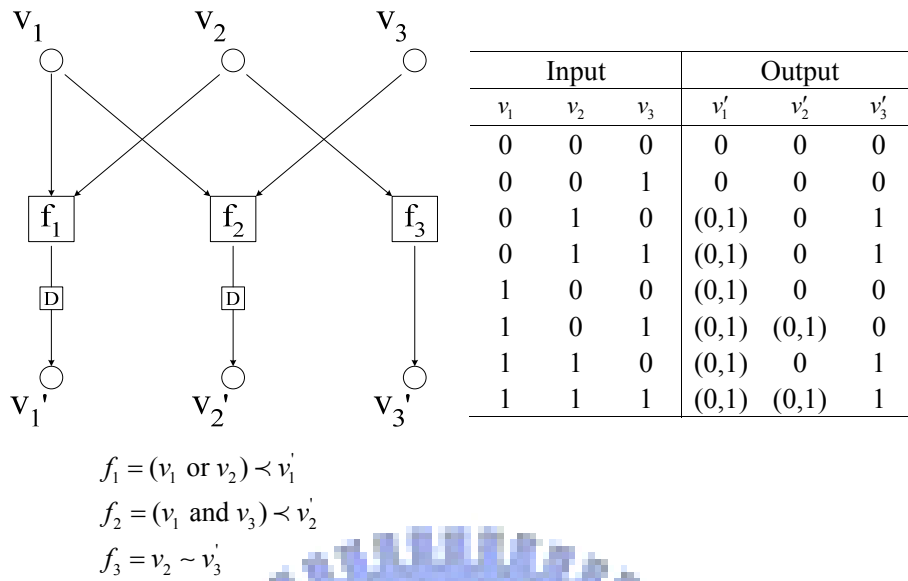


Figure 4.2: One example of Time Delay Boolean network and its Input/Output.

The output expression may have more than one possible result in the Time Delay Boolean network. For example, we consider the graph in Figure 4.2. The possible outputs for every input state are listed in the right part of the graph. If we knew the Boolean network, some of the inputs would have more than one output expression in the Time Delay Boolean network.

### 4.1.3 Identification Algorithm

For convenience, we consider the Boolean network model in which the maximum number of input genes is bounded by constant  $K$  for every target gene. In this subsection, we would consider the case of  $K = 2$ . However, these can be generalized to any  $K$  in a straightforward way. For the inference of the genetic network, we need to clarify the following question for each target output gene.

- Which genes would affect the target gene?
- What kind of Boolean function would be used for combining the input genes?

- What kind of relationship exists between the Boolean function and the target genes?

Table 4.1: Tables for a pair of input gene and one output gene assuming no measurement error

$v'_i/v_jv_h$	00	01	10	11	$v'_i/v_jv_h$	00	01	10	11
0	$m_{000}$	$m_{010}$	$m_{100}$	$m_{110}$	0	$q_{000}$	$q_{010}$	$q_{100}$	$q_{110}$
1	$m_{001}$	$m_{011}$	$m_{101}$	$m_{111}$	1	$q_{001}$	$q_{011}$	$q_{101}$	$q_{111}$

In this subsection, we propose an algorithm to clarify the above questions. The algorithm below is conceptually very simple since it simply uses output Boolean functions with input genes and relations with target genes that are consistent with the data. First, for each output gene expression at time  $t + 1$  such as  $\psi_{t+1}(v_i)$ , we consider all the pairs of elements in  $V$  at time  $t$ , for instance  $\psi_t(v_j)$  and  $\psi_t(v_h)$ . Then we count the eight incidents of  $(v_j, v_h, v'_i)$  being  $(0,0,0)$ ,  $(0,0,1)$ ,  $\dots$ ,  $(1,1,1)$  from the sample and arrange them in a  $2 \times 4$  table ; see the left part of Table 4.1. We mark a cell "+" if the count is positive and mark it "0" otherwise.

For detecting whether there exists a Boolean function which is prerequisite to the target gene, we would compare the  $2 \times 4$  output table with the left four basic relations in Table 4.2. We denote the basic relations are consistent with the output table if the position of 0 cell in the basic relations is also 0 in the output table. By comparing the output table with the four basic relations, we can find the relations which are consistent with the output tables. If there is more than one relation which is consistent with the output tables, we would use the boolean logic gate AND to combine the Boolean function and the result would transfer to another Boolean function. Hence, the final Boolean function is prerequisite to the target gene. Similarly, by comparing the  $2 \times 4$  output table with the right four basic relations in Table 4.2, we may get another Boolean function which is prerequisite to the dual of target gene.

Table 4.2: Count patterns for the basic eight relations assuming exhaustive sampling and no measurement error

$(v_j \text{ or } v_h) \prec v'_i$					$(v_j \text{ or } v_h) \prec \bar{v}'_i$				
$v'_i/v_j v_h$	00	01	10	11	$v'_i/v_j v_h$	00	01	10	11
0	+	+	+	+	0	0	+	+	+
1	0	+	+	+	1	+	+	+	+
$(v_j \text{ or } \bar{v}_h) \prec v'_i$					$(v_j \text{ or } \bar{v}_h) \prec \bar{v}'_i$				
$v'_i/v_j v_h$	00	01	10	11	$v'_i/v_j v_h$	00	01	10	11
0	+	+	+	+	0	+	0	+	+
1	+	0	+	+	1	+	+	+	+
$(\bar{v}_j \text{ or } v_h) \prec v'_i$					$(\bar{v}_j \text{ or } v_h) \prec \bar{v}'_i$				
$v'_i/v_j v_h$	00	01	10	11	$v'_i/v_j v_h$	00	01	10	11
0	+	+	+	+	0	+	+	0	+
1	+	+	0	+	1	+	+	+	+
$(\bar{v}_j \text{ or } \bar{v}_h) \prec v'_i$					$(\bar{v}_j \text{ or } \bar{v}_h) \prec \bar{v}'_i$				
$v'_i/v_j v_h$	00	01	10	11	$v'_i/v_j v_h$	00	01	10	11
0	+	+	+	+	0	+	+	+	0
1	+	+	+	0	1	+	+	+	+

Moreover, if there is only one Boolean function which occurs in above relation, that is, there is no Boolean function prerequisite to the target gene or prerequisite to the dual of target gene, we would treat the relation as our final relation between the Boolean function and the target gene. However, if both of the two prerequisite happen (i.e.  $\exists f_i$  and  $f'_i$  s.t.  $f_i \prec v_i$  and  $f'_i \prec \bar{v}_i$ ), we need to check whether these two relations are in conflict. If the dual of  $f_i$  is equivalent to  $f'_i$ , our conclusion for inference would be  $f_i$  is similar to the target gene (that is,  $f_i \sim v_i$ ); otherwise, we would treat it as if there is no relation between the input genes and the target gene. By the above identification procedure, for every target gene, we can find the corresponding input genes, Boolean function and its relation.

## 4.2 Theoretical Results

In this section, we analyze the number of INPUT/OUTPUT pairs required to identify the Time Delay Boolean network uniquely. The following proposition was obtained from the related paper with a small modification (Akutsu et al., 1998, 2003).

**Proposition 1** *For all subsets of  $V$  with  $2K$  genes, if all assignments (i.e.,  $2^{2K}$  assignments) of Boolean values appear in INPUT expression patterns and all of its possible OUTPUT expression patterns of the target gene are present, the identification of genetic network is determined to be unique, if it exists.*

**Proposition 2** *The probability that one sub-assignment with all of its possible results in the target gene does not appear among  $m$  random INPUT expression pattern is equal to  $2\left(\frac{2^{2K+1}-1}{2^{2K+1}}\right)^m - \left(1 - \frac{1}{2^{2K}}\right)^m$ , and bounded by  $2\left(1 - \frac{1}{2^{2K+1}}\right)^m$ .*

(Proof) For any fixed set of nodes  $\{v_{i_1}, v_{i_2}, \dots, v_{i_{2K}}\}$ , the probability that a sub-assignment  $v_{i_1} = v_{i_2} = \dots = v_{i_{2K}} = 1$  does not appear in one random INPUT expression pattern is  $1 - \frac{1}{2^{2K}}$ . Thus, among the  $m$  random INPUT expressions,

the probability that  $v_{i_1} = v_{i_2} = \dots = v_{i_{2K}} = 1$  appears is  $t$  times is equal to  $\frac{m!}{t!(m-t)!} \left(\frac{1}{2^{2K}}\right)^t \left(1 - \frac{1}{2^{2K}}\right)^{m-t}$ , where  $t \leq m$ , and the probability that all of the possible results in the target gene does not appear among  $t$  times INPUT is smaller than  $\left(\frac{1}{2}\right)^{t-1}$ . Hence the probability that one sub-assignment with all of their possible results does not appear among  $m$  random INPUT expression is smaller than  $\left(1 - \frac{1}{2^{2K}}\right)^m + \sum_{t=1}^m \frac{m!}{t!(m-t)!} \left(\frac{1}{2^{2K}}\right)^t \left(1 - \frac{1}{2^{2K}}\right)^{m-t} \left(\frac{1}{2}\right)^{t-1}$ , and this is equal to  $2\left(\frac{2^{2K+1}-1}{2^{2K+1}}\right)^m - \left(1 - \frac{1}{2^{2K}}\right)^m$ , and bounded by  $2\left(1 - \frac{1}{2^{2K+1}}\right)^m$  by a simple algebra calculation.

Next we prove the main theorem.

**Theorem 1** *For the identification of one Time Delay Boolean network of  $n$  nodes with maximum indegree  $\leq K$ ,  $O(2^{2K+1} \cdot (2K + \alpha) \cdot \log n)$  uniformly randomly sampled Input patterns are sufficient for exact inference with probability at least  $1 - \frac{1}{n^\alpha}$  for  $\alpha > 0$ .*

(Proof) We consider the probability that the condition of Proposition 1 is not satisfied under  $m$  random INPUT expression patterns.

By Proposition 2, the probability that  $v_{i_1} = v_{i_2} = \dots = v_{i_{2K}} = 1$  with all of its possible results in the target gene does not appear among the  $m$  random INPUT expression patterns is bounded by  $2\left(1 - \frac{1}{2^{2K+1}}\right)^m$  for any fixed set of nodes  $\{v_{i_1}, v_{i_2}, \dots, v_{i_{2K}}\}$ . Since the number of combinations of  $2K$  nodes from a set of  $n$  possibilities is bounded by  $2^{2K} \cdot n^{2K}$ , the probability that the condition of Proposition 1 is not satisfied is at most  $2^{2K} \cdot n^{2K} \cdot 2\left(1 - \frac{1}{2^{2K+1}}\right)^m$ . It is not difficult to see that  $2^{2K} \cdot n^{2K} \cdot 2\left(1 - \frac{1}{2^{2K+1}}\right)^m < p$  holds for  $m > \ln 2 \cdot 2^{2K+1} \cdot (2K + 2K \log n + \log 2 + \log \frac{1}{p})$ . Letting  $p = \frac{1}{n^\alpha}$ , we obtain the theorem.

Next we develop an information theoretic lower bound on the number of INPUT/OUTPUT pairs needed for the identification of Time Delay Boolean network. The proof of the theorem is a straightforward adaptation of similar results given in (Akutsu and Miyano, 1999) in the case of Boolean networks.

**Theorem 2** *If the maximum indegree  $\leq K$ , at least  $\Omega(2^K + K \log n)$  INPUT/OUTPUT*

*pairs are required for the identification of Time Delay Boolean network in the worst case.*

(Proof) The number of Time Delay Boolean network is given by all the possible combination of Boolean function with  $k$  nodes from a set of  $n$  possibilities with all possible relations between Boolean function with target node. Since there are  $\Omega(n^K)$  possible combinations of input nodes,  $2^{2^K}$  possible boolean functions and 3 possible relations between Boolean function with each node, there are  $\Omega((2^{2^K} \cdot n^K \cdot 3)^n)$  Boolean networks whose maximum indegree is at most  $K$ . On the other hand, there are at least  $2^n$  possible OUTPUT patterns with one INPUT expression pattern. Therefore,  $\Omega(\log_{2^n}((2^{2^K} \cdot n^K \cdot 3)^n))$  which is the same as  $\Omega(2^K + K \log n)$  INPUT/OUTPUT pairs are required in the worst case.

### 4.3 Inference of Time Delay Boolean Network with Noise data

In Section 4.2, we discussed the identification algorithm for the data without measurement error. In this section we will extend the situation of inference complete the data with measurement error. We assume there exists a measurement error with probability  $p$  in every element, independently, and make the output  $O_j$   $j = 1, 2, \dots, m$  switch to its negation; that is

$$O_{ij}^* = \begin{cases} O_{ij} & \text{with probability } 1 - p; \\ 1 - O_{ij} & \text{with probability } p. \end{cases}$$

This makes the output data  $O_j^*$  with noise are the observations and our goal is to reconstruct the Time Delay Boolean network from the pair data  $(I_j, O_j^*)$ .

Similar to section 4.2, we consider the maximum number of input genes is bounded by 2 for every target gene. Instead of using full model including every element, we consider the pair of input genes with every output gene as our model and use probabilistic models to compute the measurement error. We treat the data

Table 4.3: The eight basic relations and their corresponding probabilistic hypotheses and scores

Relation	Hypothesis	Scores
$(v_j \text{ or } v_h) \prec v'_i$	$q_{000} = 0$	$p(v_j \text{ or } v_h) \prec v'_i$
$(v_j \text{ or } \bar{v}_h) \prec v'_i$	$q_{010} = 0$	$p(v_j \text{ or } \bar{v}_h) \prec v'_i$
$(\bar{v}_j \text{ or } v_h) \prec v'_i$	$q_{100} = 0$	$p(\bar{v}_j \text{ or } v_h) \prec v'_i$
$(\bar{v}_j \text{ or } \bar{v}_h) \prec v'_i$	$q_{110} = 0$	$p(\bar{v}_j \text{ or } \bar{v}_h) \prec v'_i$
$(v_j \text{ or } v_h) \prec v'_i$	$q_{001} = 0$	$p(v_j \text{ or } v_h) \prec v'_i$
$(v_j \text{ or } \bar{v}_h) \prec v'_i$	$q_{011} = 0$	$p(v_j \text{ or } \bar{v}_h) \prec v'_i$
$(\bar{v}_j \text{ or } v_h) \prec v'_i$	$q_{101} = 0$	$p(\bar{v}_j \text{ or } v_h) \prec v'_i$
$(\bar{v}_j \text{ or } \bar{v}_h) \prec v'_i$	$q_{111} = 0$	$p(\bar{v}_j \text{ or } \bar{v}_h) \prec v'_i$

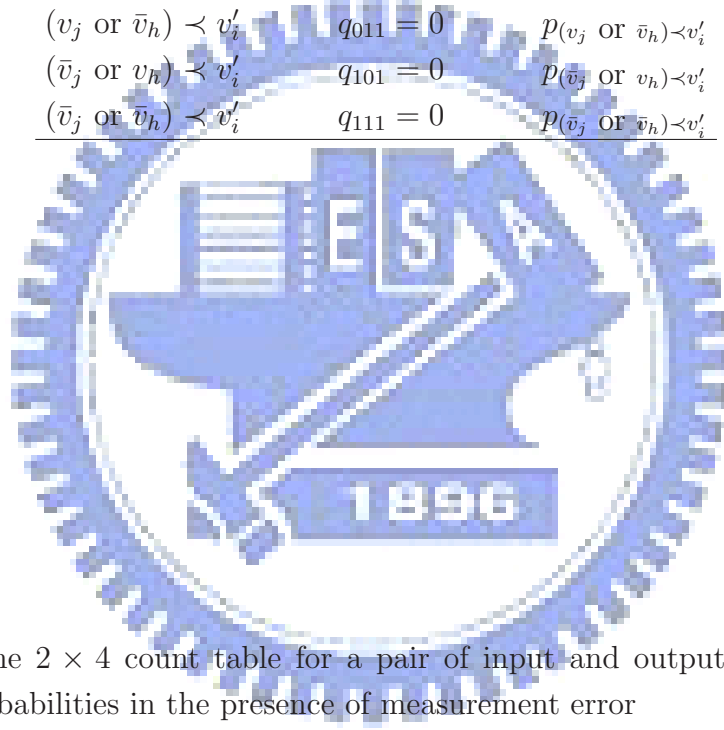


Table 4.4: The  $2 \times 4$  count table for a pair of input and output gene and their generated probabilities in the presence of measurement error

$v'_i/v_j v_h$	00	01	10	11	$v'_i/v_j v_h$	00	01	10	11
0	$n_{000}$	$n_{010}$	$n_{100}$	$n_{110}$	0	$r_{000}$	$r_{010}$	$r_{100}$	$r_{110}$
1	$n_{001}$	$n_{011}$	$n_{101}$	$n_{111}$	1	$r_{001}$	$r_{011}$	$r_{101}$	$r_{111}$



in the  $2 \times 4$  table as a multinomial distribution with eight cells whose probabilities are  $q_{000}, q_{001}, \dots, q_{111}$  as shown in the right part of Table 4.1, where  $q_{000} + q_{001} + \dots + q_{111} = 1$ . Similarly, we extract the data with measurement error for every output gene and each pair of input genes as the  $2 \times 4$  table. Now the counts  $n_{000}, n_{001}, \dots, n_{111}$  are not generated from the multinomial  $q_{000}, q_{001}, \dots, q_{111}$ , but from another multinomial  $r_{000}, r_{001}, \dots, r_{111}$  as shown in Table 4.4, where  $r_{000} + r_{001} + \dots + r_{111} = 1$ .

Table 4.5: Splitting counts caused by misclassification error

$v'_i/v_j v_h$	00	01	10	11				
0	$m_{000,000}$	$m_{000,001}$	$m_{010,000}$	$m_{010,001}$	$m_{100,000}$	$m_{100,001}$	$m_{110,000}$	$m_{110,001}$
	$m_{000,010}$	$m_{000,011}$	$m_{010,010}$	$m_{010,011}$	$m_{100,010}$	$m_{100,011}$	$m_{110,010}$	$m_{110,011}$
	$m_{000,100}$	$m_{000,101}$	$m_{010,100}$	$m_{010,101}$	$m_{100,100}$	$m_{100,101}$	$m_{110,100}$	$m_{110,101}$
	$m_{000,110}$	$m_{000,111}$	$m_{010,110}$	$m_{010,111}$	$m_{100,110}$	$m_{100,111}$	$m_{110,110}$	$m_{110,111}$
1	$m_{001,000}$	$m_{001,001}$	$m_{011,000}$	$m_{011,001}$	$m_{101,000}$	$m_{101,001}$	$m_{111,000}$	$m_{111,001}$
	$m_{001,010}$	$m_{001,011}$	$m_{011,010}$	$m_{011,011}$	$m_{101,010}$	$m_{101,011}$	$m_{111,010}$	$m_{111,011}$
	$m_{001,100}$	$m_{001,101}$	$m_{011,100}$	$m_{011,101}$	$m_{101,100}$	$m_{101,101}$	$m_{111,100}$	$m_{111,101}$
	$m_{001,110}$	$m_{001,111}$	$m_{011,110}$	$m_{011,111}$	$m_{101,110}$	$m_{101,111}$	$m_{111,110}$	$m_{111,111}$

Because of the measurement error, some samples of  $m_{000}$  may translate to other seven cells. We use the notation  $m_{000,000}, m_{000,001}, \dots, m_{000,111}$  represent the counts of eight cells translated from  $m_{000}$ . Analogous notation is defined for  $m_{001}, m_{010}, \dots, m_{111}$ . The splitting is shown in Table 4.5. Corresponding, their generating probabilities ( $q_{000}, q_{001}, \dots, q_{111}$ ) are redistributed as follows:  $q_{i_1 i_2 i_3, j_1 j_2 j_3} = p^{I(i,j)} (1-p)^{3-I(i,j)} q_{i_1 i_2 i_3}$ , where  $I(i, j) = \sum_{k=1}^3 |i_k - j_k|$ . Here, we adopt the notation  $q_{i_1 i_2 i_3, j_1 j_2 j_3}$  analogous to  $m_{i_1 i_2 i_3, j_1 j_2 j_3}$ . These two sets of counts and probabilities are linked as follows:

$$\begin{cases} n_{j_1 j_2 j_3} = \sum_{i_1, i_2, i_3=0,1} m_{i_1 i_2 i_3, j_1 j_2 j_3} \\ r_{j_1 j_2 j_3} = \sum_{i_1, i_2, i_3=0,1} q_{i_1 i_2 i_3, j_1 j_2 j_3} \end{cases}$$

and

(4.1)

$$\begin{cases} m_{i_1 i_2 i_3} = \sum_{j_1, j_2, j_3=0,1} m_{i_1 i_2 i_3, j_1 j_2 j_3} \\ q_{i_1 i_2 i_3} = \sum_{j_1, j_2, j_3=0,1} q_{i_1 i_2 i_3, j_1 j_2 j_3} \end{cases}$$

Under the full data  $\{m_{i_1 i_2 i_3, j_1 j_2 j_3}\}$ , the log-likelihood is given by

$$L = \sum_{i_1, i_2, i_3, j_1, j_2, j_3=0,1} m_{i_1 i_2 i_3, j_1 j_2 j_3} \log q_{i_1 i_2 i_3, j_1 j_2 j_3} \quad (4.2)$$

where  $q_{i_1 i_2 i_3, j_1 j_2 j_3}$  are those splitting probabilities. Later we define p-scores via maximum likelihood estimates (MLE). Since the full data  $\{m_{i_1 i_2 i_3, j_1 j_2 j_3}\}$  is not observable, we use the famous E-M algorithm to maximize the likelihood of full data (2) to estimate MLE. In the E-step, the splitting counts of full data  $\{m_{i_1 i_2 i_3, j_1 j_2 j_3}\}$  is evaluated by the conditional expectations calculated at the current value of the parameter by the formula

$$E_{p, q_{000}, q_{001}, \dots, q_{111}}(m_{i_1 i_2 i_3, j_1 j_2 j_3} | n_{j_1 j_2 j_3}) = \frac{n_{j_1 j_2 j_3} q_{i_1 i_2 i_3, j_1 j_2 j_3}}{\sum_{i'_1, i'_2, i'_3=0,1} q_{i'_1 i'_2 i'_3, j_1 j_2 j_3}} \quad (4.3)$$

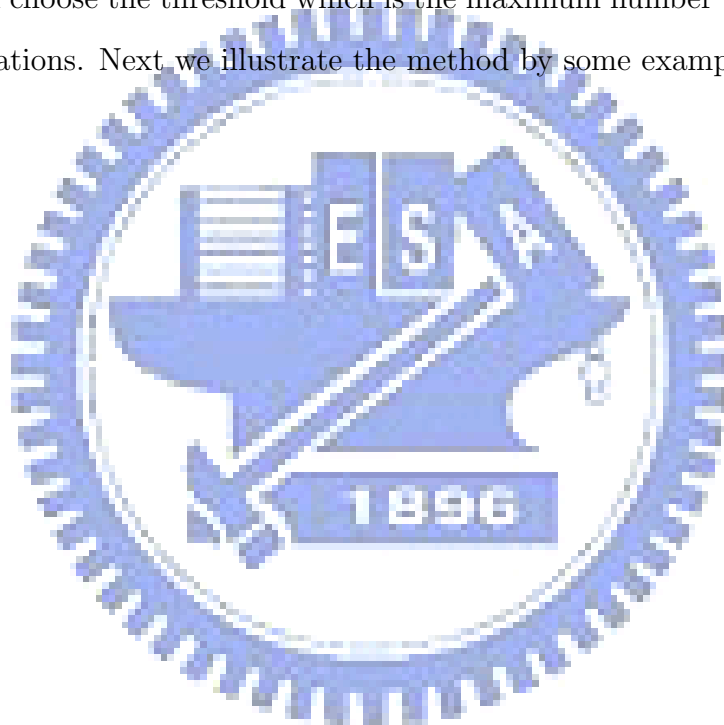
where  $i_1, i_2, i_3, j_1, j_2, j_3 = 0, 1$ . One probabilities of  $q_{000}, q_{001}, \dots, q_{111}$  are zero in the different hypotheses specified in Table 4.3. In the M-step, we maximize the conditional expectation of the log-likelihood for the full data to calculate the MLE of the parameters.

e first consider a problem simpler than reconstructing a Time Delay Boolean network: what is the most likely relation for one output gene and a pair of input genes?

**Definition 1** For one output gene  $v'_i$  and a pair of input genes  $v_j$  and  $v_k$ , the p-scores  $p_{(v_j \text{ or } v_k) \prec v'_i}, p_{(\bar{v}_j \text{ or } v_k) \prec v'_i}, p_{(v_j \text{ or } \bar{v}_k) \prec v'_i}, p_{(\bar{v}_j \text{ or } \bar{v}_k) \prec v'_i}, p_{(v_j \text{ or } v_k) \prec v'_i}, p_{(\bar{v}_j \text{ or } v_k) \prec v'_i}, p_{(v_j \text{ or } \bar{v}_k) \prec v'_i}, p_{(\bar{v}_j \text{ or } \bar{v}_k) \prec v'_i}$  are, respectively, the maximum likelihood estimates of  $p$  under the triangular model:  $q_{000} = 0, q_{010} = 0, q_{100} = 0, q_{110} = 0, q_{001} = 0, q_{011} = 0, q_{101} = 0, q_{111} = 0$

We compute p-scores by the E-M algorithm described earlier. The heuristic of the definition is that we use the MLE  $\hat{p}$  to measure how well each hypothesis fits: the smaller the score, the more evidence that the corresponding hypothesis is true.

The p-scores are more meaningful if they are generated from a Time Delay Boolean network because we may discover significant relations by ranking the scores in the ascending order. Here we use the *maximum compatibility criterion*: choose the maximum threshold value so that the selected relations contain no conflicts (Li and Lu, 2005). We collect those relations whose p-scores are smaller than a threshold. Known biological results are helpful for the determination of a threshold. For example, if we know the relation  $(v_1 \text{ or } v_2) \prec v_3$  is true, then the p-scores smaller than  $p_{(v_1 \text{ or } v_2) \prec v_3}$  should be in our watch list. Please notice that as more relations are included in the watch list, the more likely we are to observe incompatible ones. Hence, we can choose the threshold which is the maximum number that contains no conflicting relations. Next we illustrate the method by some examples.



# Chapter 5

## Empirical Results

### 5.1 Emotion Detection by Physiological Signals

Due to the existence of day-dependency of the features of physiological signals, some of the features would have a large discrepancy even they are in the same state of emotion. Besides, some of the value of features would be quite near even for different states of emotion, as shown in Figure 5.1. However, after removing the daily effects by MANOVA, the scatter of the features in the same state of emotion would be more tight and become differentiable for distinct states of emotion, as shown in Figure 5.2. Therefore, the statistical technique of MANOVA would be helpful for emotion classification and recognition.

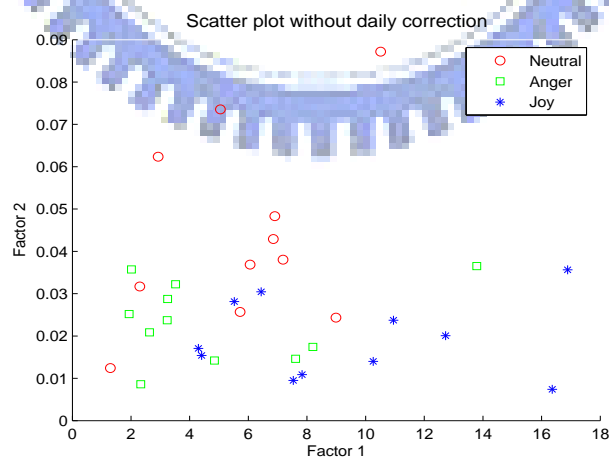


Figure 5.1: The scatter plot of three statuses of emotion without daily correction.

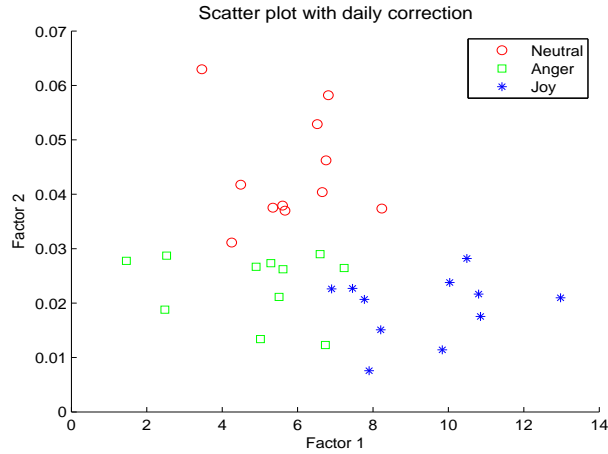


Figure 5.2: The scatter plot of three statuses of emotion with daily correction.

After getting the features from physiological signals and removing daily effects by MANOVA, six classification methods would be applied as discussed before. For the first database, the classification results without daily correction and with daily correction by leave-one-out cross-validation in three emotional statuses by six classification methods are listed in Table 5.1 and Table 5.2. From the classification result of subject Alice, the highest correct recognition rate is 90.91% using the classification method of logistic model. As for the classification result of subject Jane, the highest correct recognition rate is 93.94% by the same classification method of logistic model. Hence, the leave-one-out cross-validation can be used to evaluate the prediction accuracy and make comparisons. Besides, the p-values of statistical improvement are all significant in most classification methods, except the subject Jane with the classification method of C4.5.

For second database, in order to see how the statistical technique of MANOVA influence classification, we compare the classification accuracy with and without removing the daily and subject dependence in Table 5.3. In additional, the p-value with statistical significance of the difference in classification between with and without the technique of MANOVA for every classification method is also attached in Table 5.3 as well. Among these six classification methods, the highest correct

Table 5.1: Classification of three emotional statuses by the physiological signals of subject Alice.

Method	Alice		
	Without daily correction	With daily correction	p-value of improvement
Bayesian Network	45.45%	81.82%	0.001
Naive Bayesian	48.48%	75.76%	0.011
SVM	45.45%	78.79%	0.003
C4.5	45.45%	78.79%	0.003
Logistic Model	57.58%	90.91%	0.001
KNN	39.40%	75.76%	0.001

Table 5.2: Classification of three emotional statuses by the physiological signals of subject Jane.

Method	Jane		
	Without daily correction	With daily correction	p-value of improvement
Bayesian Network	51.52%	90.91%	0.000
Naive Bayesian	48.48%	78.79%	0.005
SVM	69.70%	84.85%	0.073
C4.5	66.67%	75.76%	0.211
Logistic Model	60.61%	93.94%	0.000
KNN	63.64%	87.88%	0.011

Table 5.3: Classification of three emotion status by the physiological signals of 10 subjects and 7 times.

Method	Without daily correction	With daily correction	p-value of improvement
Bayesian Network	49.05%	64.76%	0.001
Naive Bayes	48.57%	65.71%	0.000
SVM	45.24%	70.48%	0.000
C4.5	50.00%	61.90%	0.007
Logistic Model	54.29%	74.76%	0.000
KNN	42.38%	58.10%	0.001

recognition rate is 74.76% using the classification method of logistic model and all of these classification method can significant improve the overall accuracy rate after removing daily and personal effects by MANOVA.

## 5.2 Analysis of Yeast Genes by Microarray Studies

In this section, the results by two different kinds of clustering methods in the microarray study are compared. Firstly, the PSE is considered. The results are plotted and tabulated in Figure 5.3 and Table 5.4.

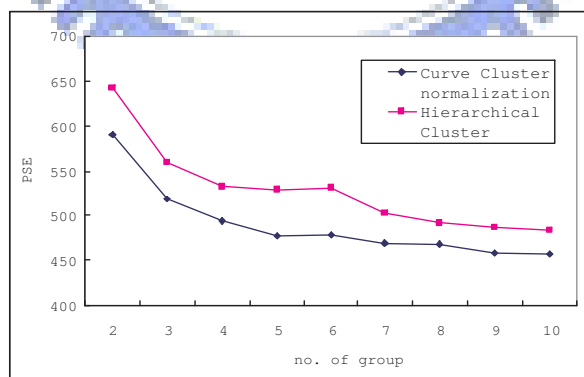


Figure 5.3: PSE comparisons of different number of clusters are shown for two different clustering methods.

Table 5.4: The detail values of PSE.

No. of groups	2	3	4	5	6	7	8	9	10
H. Cluster	642.1	559.9	532.4	529.8	531.2	502.2	491.2	486.6	483.4
C. Cluster	590.1	518.3	493.3	477.5	478.4	468.7	468.2	457.6	456.9

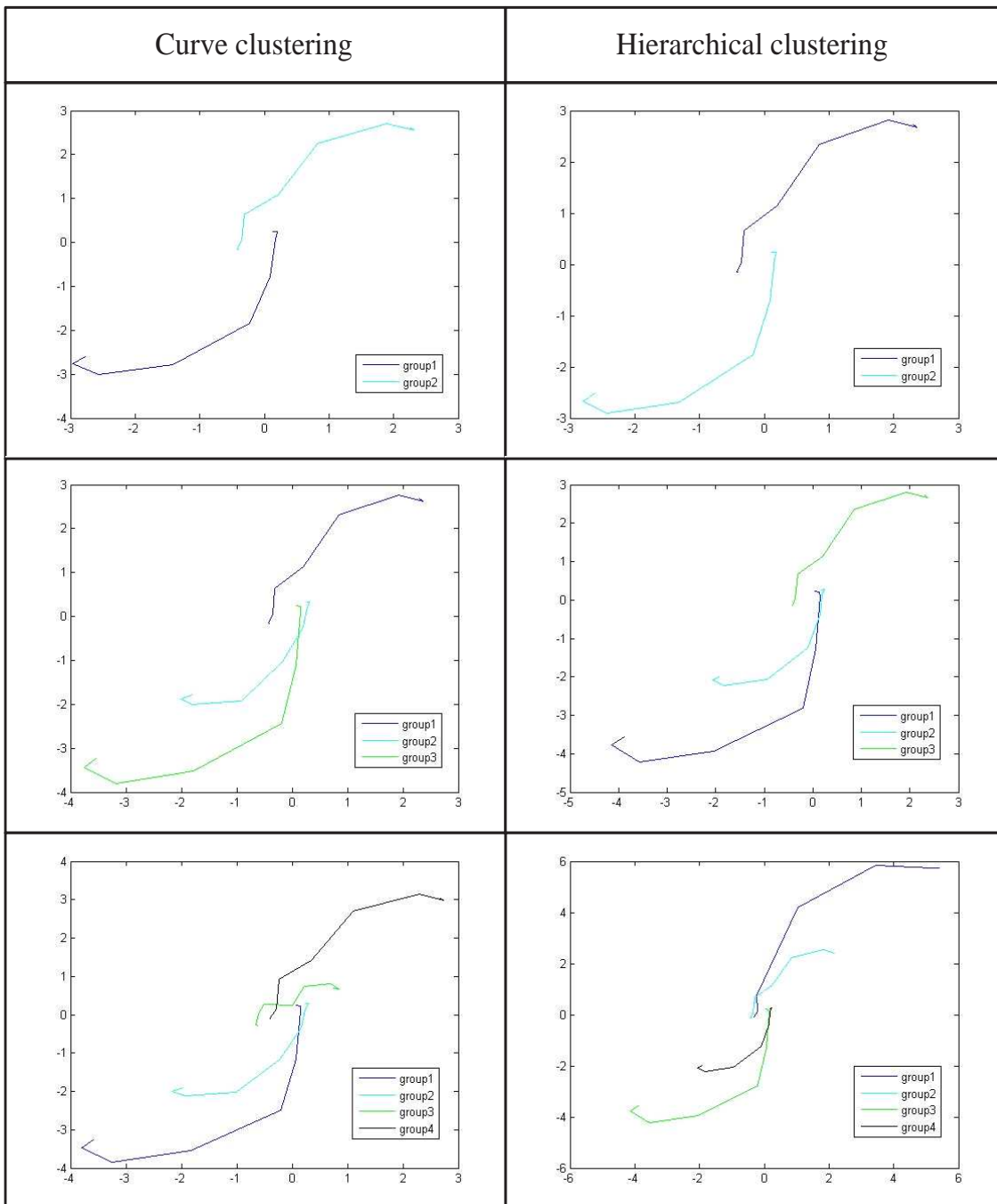
From the above comparisons, the results by curve clustering have smaller PSE than those by hierarchical cluster do. In addition, we will check the consistency for two clustering method as the mean curves shown in Figure 5.4.

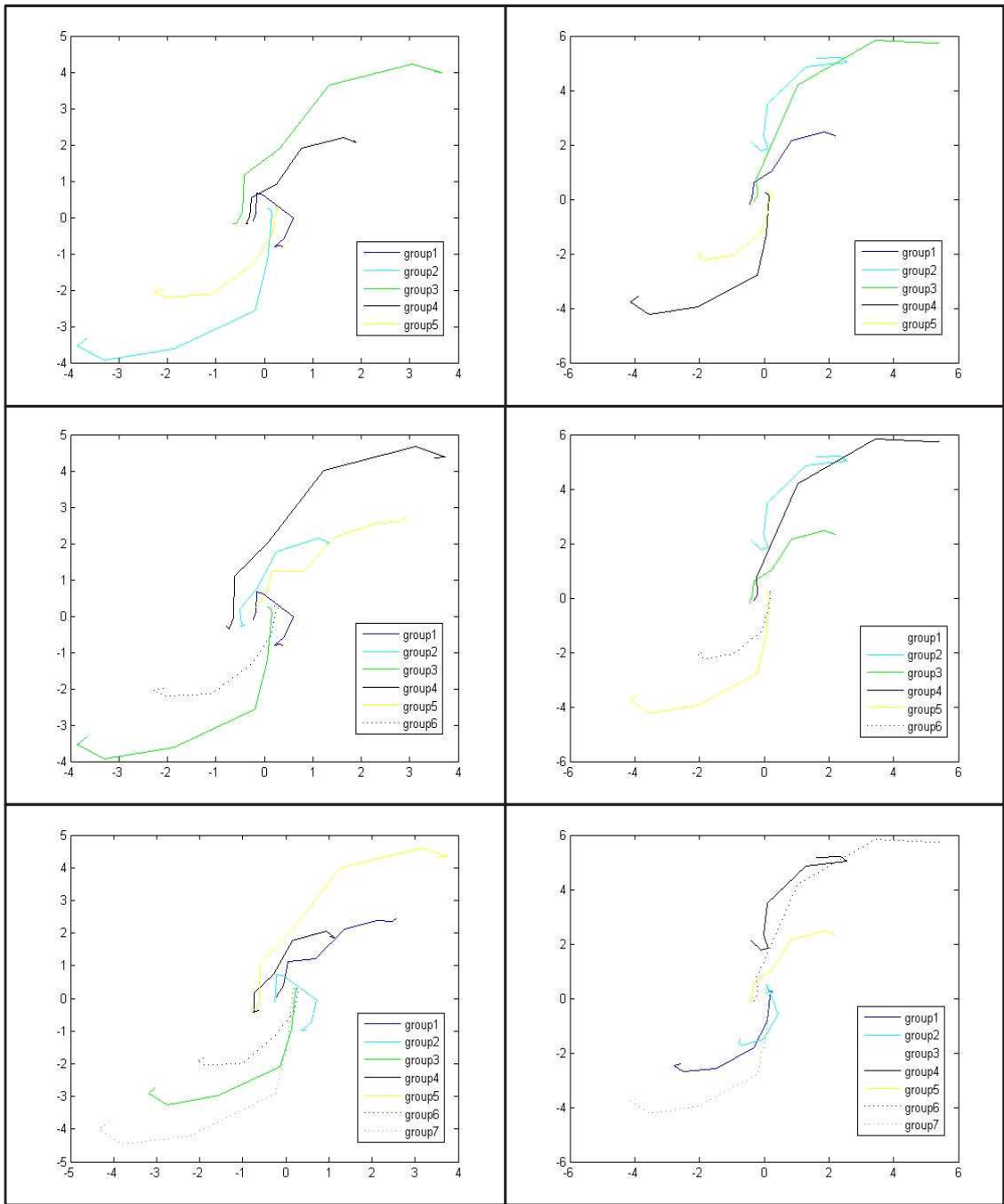
From the above results for two clustering method, it often exist groups in hierarchical clustering that do not have consistent gene expression profiles in three experiments when the number of clusters is large. By these viewpoints of prediction errors and consistency, the results by curve clustering are preferred. Then, it is necessary to decide the cluster size. When the number of cluster size equals to five, there will be one group that gene expressions appear negative correlation between BY and RM strains. As the cluster size increases, patterns of negative correlation are recurrent. However, the number of genes with consistent expression profiles in every group becomes fewer as the cluster size increases. Hence, we will consider the cluster size of five in this study.

The expression profiles of consistent genes are listed in Figure 5.5. Expression profiles in group 2, 3, 4 and 5 show similar time trends and positive correlations in two strains. However, consistent genes in group 1 show different time trends and patterns that will be explored below.

The results of time shifts determined by regression models in these five clusters are listed in Figure 5.6. From Figure 5.6, gene expressions appear to vary later than glucose consumption do in most groups, except for group 1. Genes in group 1 are interesting because there are negative correlations between gene expressions in BY and RM strains as shown in the mean curves in Figure 5.4 when the cluster size is five. The regression results show that the gene expression profiles in group 1







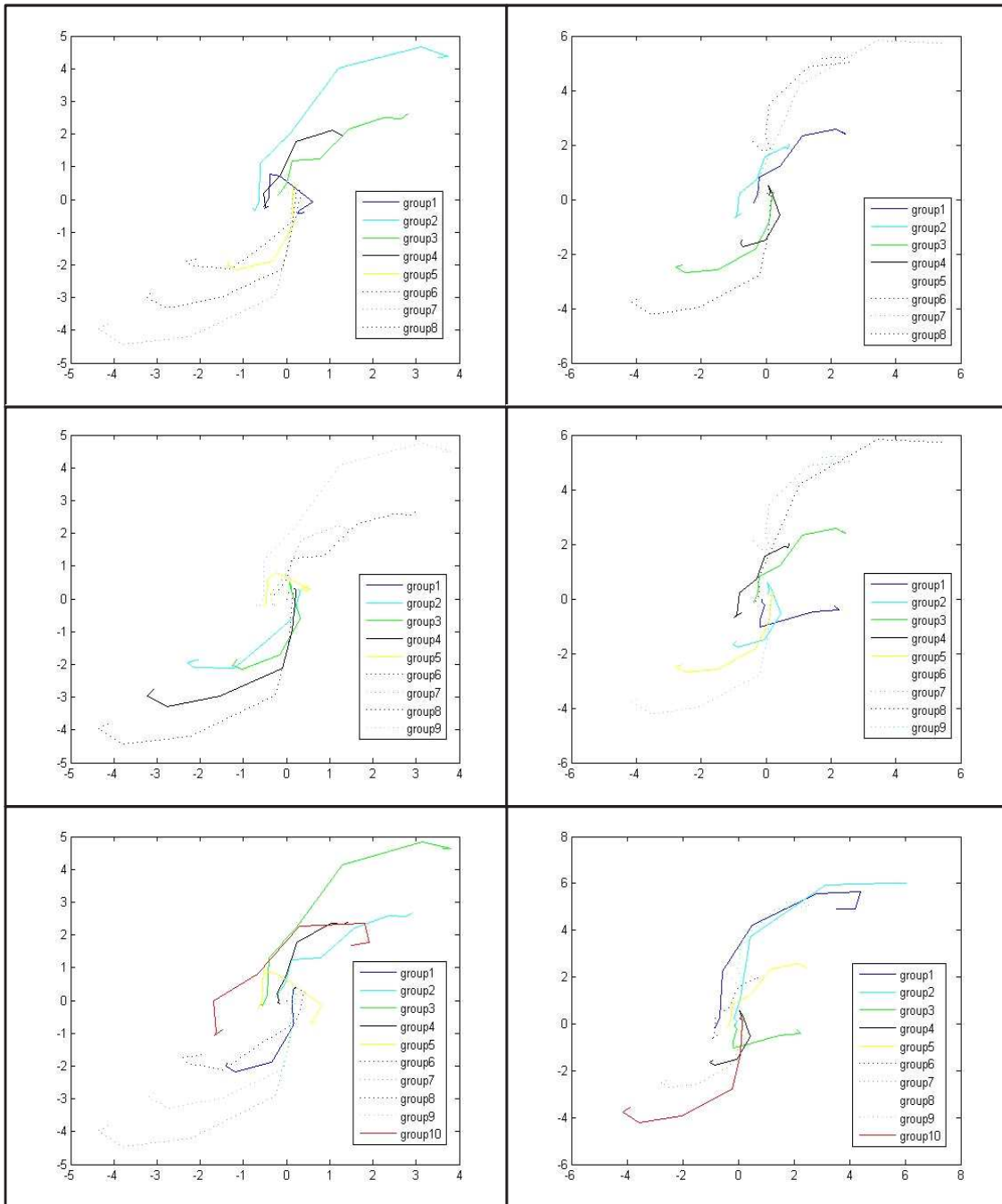


Figure 5.4: Mean curves of every group are shown for two clustering methods with different clustering sizes.

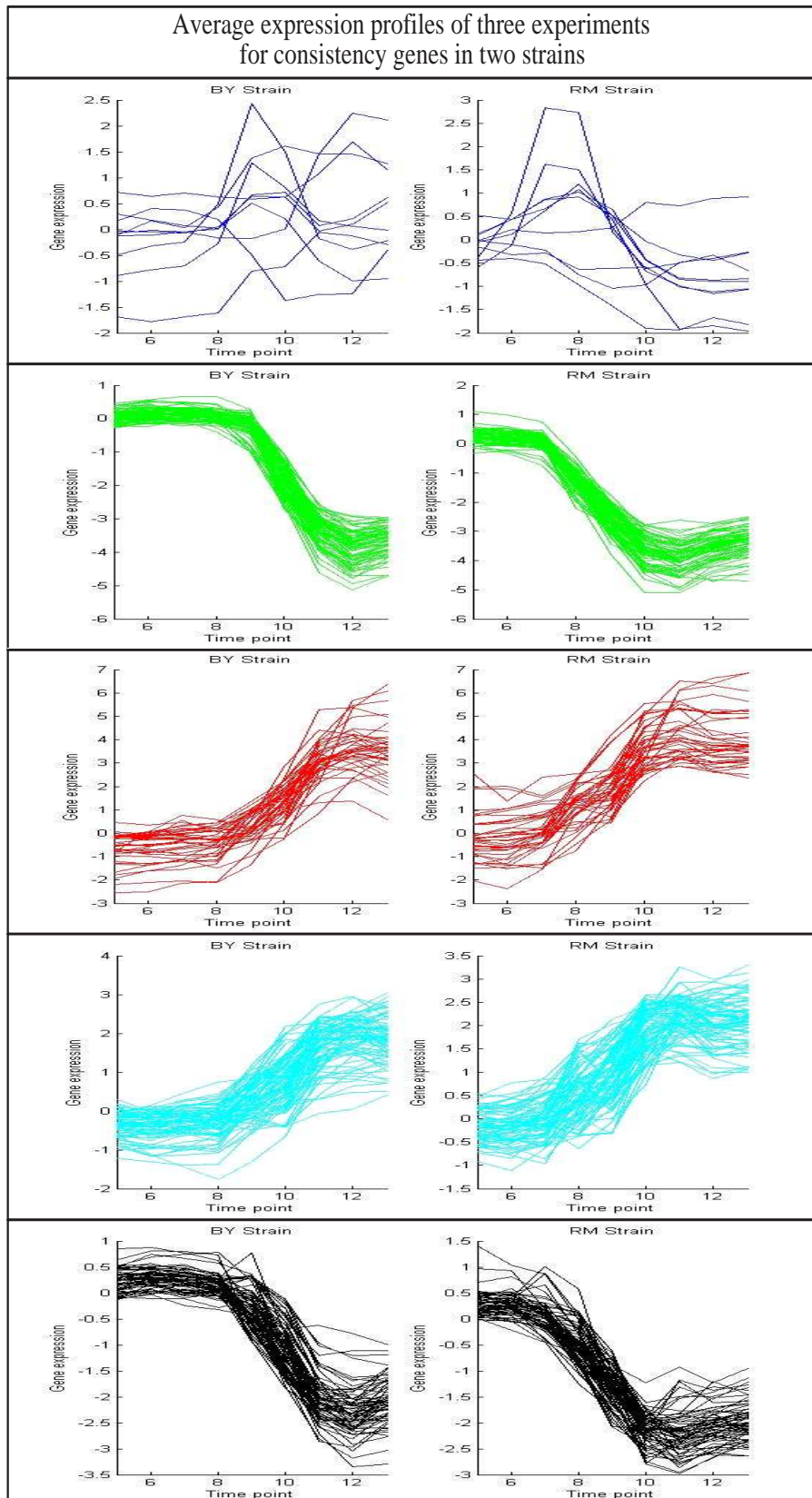


Figure 5.5: The clustering results by curve clustering are shown when the number of cluster size is five.

**Tests of Between-Subjects Effects (Group 1)**

Dependent Variable: Log\_Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	266.063 <sup>a</sup>	93	2.861	4.024	.000
Intercept	48.393	1	48.393	68.060	.000
Exp	26.697	2	13.348	18.773	.000
Gene_Name	62.734	9	6.970	9.803	.000
Time	79.745	8	9.968	14.019	.000
Strain	53.011	1	53.011	74.554	.000
Gene_Name * Time	103.123	72	1.432	2.014	.000
Glucose_Time1	52.333	1	52.333	73.602	.000
Error	317.122	446	.711		
Total	583.414	540			
Corrected Total	583.185	539			

a. R Squared = .456 (Adjusted R Squared = .343) (Group time shift =1)

**Tests of Between-Subjects Effects (Group 2)**

Dependent Variable: Log\_Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12727.843 <sup>a</sup>	696	18.287	98.080	.000
Intercept	1870.725	1	1870.725	10033.340	.000
Gene_Name	235.350	76	3.097	16.609	.000
Exp	64.590	2	32.295	173.208	.000
Strain	15.156	1	15.156	81.289	.000
Time	334.420	8	41.803	224.201	.000
Gene_Name * Time	156.854	608	.258	1.384	.000
Glucose_Time1_A	942.628	1	942.628	5055.636	.000
Error	645.307	3461	.186		
Total	25044.593	4158			
Corrected Total	13373.149	4157			

a. R Squared = .952 (Adjusted R Squared = .942) (Group time shift =-1)

**Tests of Between-Subjects Effects (Group 3)**

Dependent Variable: Log\_Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	8287.881 <sup>a</sup>	345	24.023	40.464	.000
Intercept	582.225	1	582.225	980.708	.000
Gene_Name	426.397	37	11.524	19.412	.000
Exp	53.377	2	26.689	44.955	.000
Time	69.501	8	8.688	14.634	.000
Strain	75.522	1	75.522	127.210	.000
Gene_Name * Time	682.426	296	2.305	3.883	.000
Glucose_Time1_A	247.515	1	247.515	416.917	.000
Error	1044.874	1760	.594		
Total	14641.449	2106			
Corrected Total	9332.754	2105			

a. R Squared = .888 (Adjusted R Squared = .866) (Group time shift=-1)

**Tests of Between-Subjects Effects (Group 4)**

Dependent Variable: Log\_Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4710.936 <sup>a</sup>	723	6.516	24.122	.000
Intercept	280.220	1	280.220	1037.403	.000
Gene_Name	240.983	79	3.050	11.293	.000
Exp	33.762	2	16.881	62.495	.000
Time	29.905	8	3.738	13.839	.000
Strain	36.177	1	36.177	133.930	.000
Gene_Name * Time	296.629	632	.469	1.738	.000
Glucose_Time1_A	111.298	1	111.298	412.037	.000
Error	971.338	3596	.270		
Total	8492.808	4320			
Corrected Total	5682.274	4319			

a. R Squared = .829 (Adjusted R Squared = .795) (Group time shift=-1)

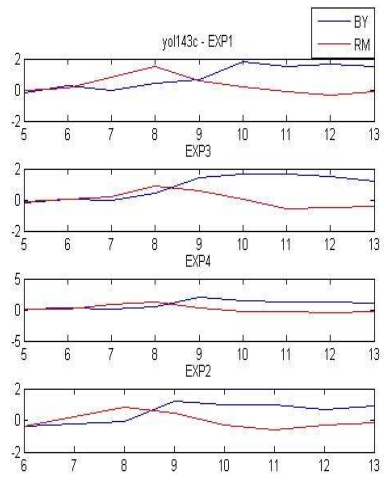
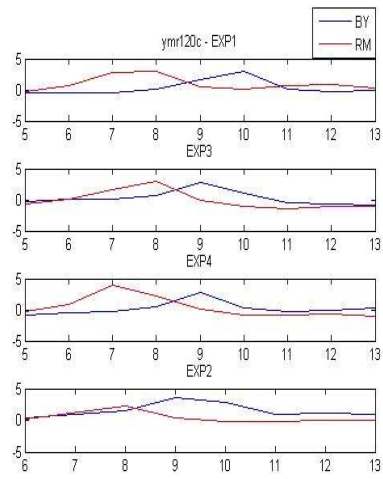
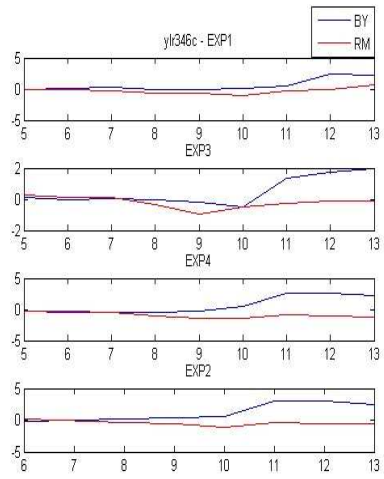
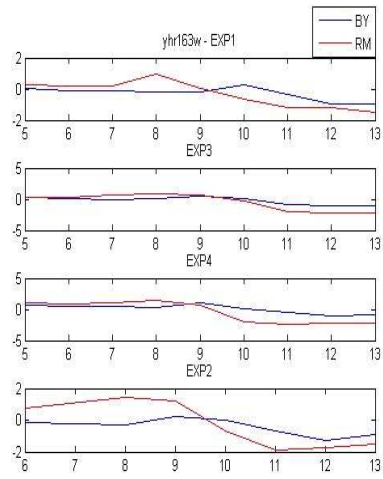
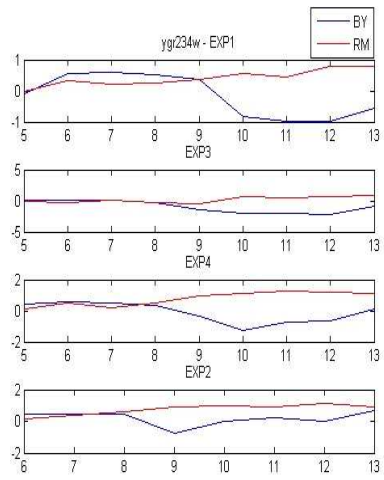
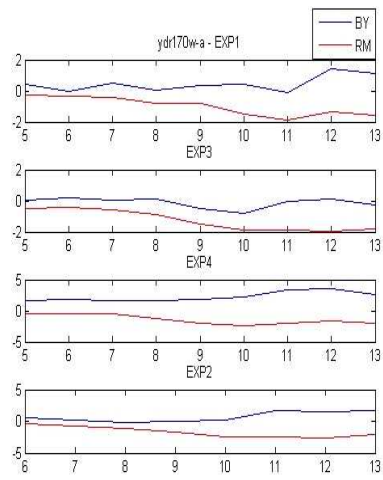
**Tests of Between-Subjects Effects (Group 5)**

Dependent Variable: Log\_Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4397.232 <sup>a</sup>	624	7.047	44.617	.000
Intercept	476.824	1	476.824	3019.020	.000
Gene_Name	105.873	68	1.557	9.858	.000
Exp	41.536	2	20.768	131.492	.000
Time	74.690	8	9.336	59.112	.000
Strain	4.357	1	4.357	27.584	.000
Gene_Name * Time	144.470	544	.266	1.681	.000
Glucose_Time1_A	237.392	1	237.392	1503.050	.000
Error	489.772	3101	.158		
Total	7947.709	3726			
Corrected Total	4887.004	3725			

a. R Squared = .900 (Adjusted R Squared = .880) (Group time shift=-1)

Figure 5.6: Results of regression models with the most significant effects of glucose association among three time shifts are listed for five groups.



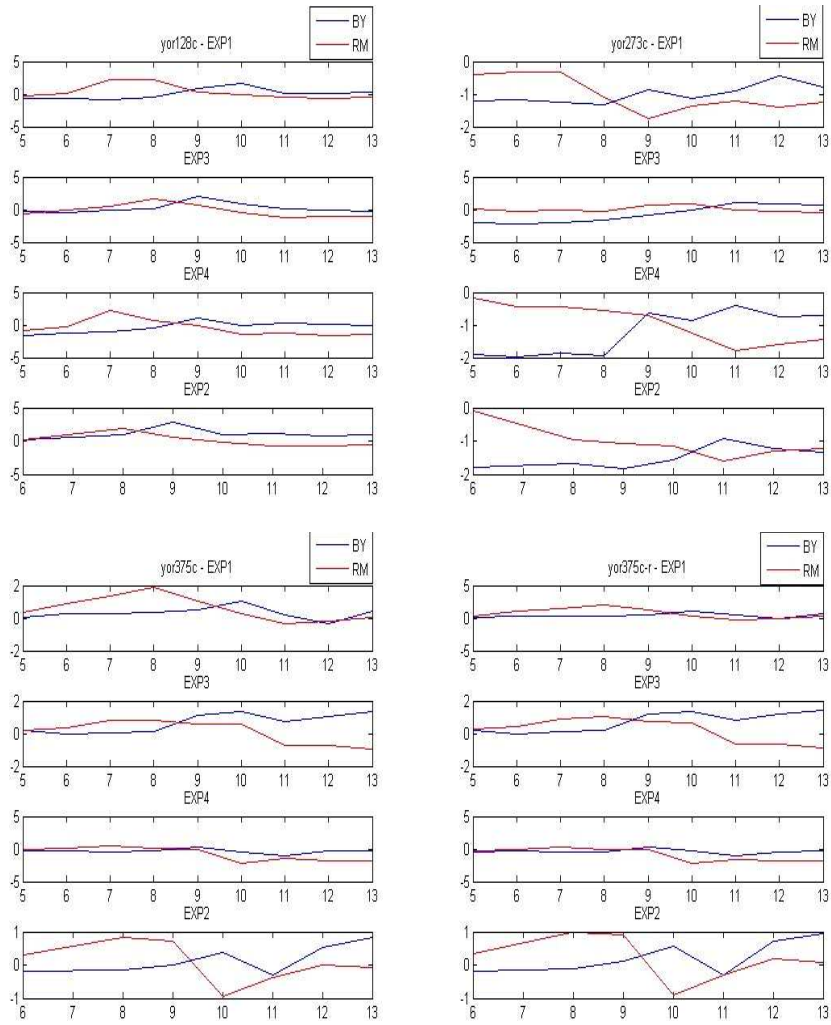


Figure 5.7: Time profiles of genes for two strains in group 1 are shown when the number of cluster size equal to five



Table 5.5: Consistent genes are reported for five groups.

Group	Consistent genes included in the group						
1	yor128c yhr163w	ydr170w-a yor273c	ymr120c ygr234w	yol143c	yor375c-r	ylr346c	yor375c
2	ygr148c ylr388w yml024w yol127w ymr230w ydr447c ylr101c ynl209w ymr098c yol040c yel054c	ylr344w ynl096c ypl081w ypr132w yhl015w yfr032c-a ypl249c-a ypl079w yor312c yol121c yer056c-a	yml063w ymr142c ydr064w yhl001w yjl190c ygr085c ygl123w ydr418w ydr450w ydl082w ygr214w	yol120c ynl301c ymr116c yil018w ygl147c yil069c ygr034w ygl030w yhr010w ydr025w ylr029c	yfr031c-a yor096w ynl178w yhr141c yil053w ypl143w ygl103w ygr162w yjl136c yfl034c-a ynl302c	yjr123w yor369c ynl162w yjr145c ylr061w ydl083c yjl189w yil052c ymr121c yer131w yor293w	ylr075w yjl191w yor063w ymr242c yor234c ydl075w yml026c ylr048w ynl069c ykr094c ypr102c
3	ymr105c yhr001w-a ynl015w ylr366w ynl117w yer150w	ydr178w yer053c-a-r ylr038c ymr250w q0080 yer053c-a	ykl217w ylr327c ynl134c ynr034w-a yil160c ymr256c	yfl052w ynr002c ydr529c yol052c-a yml054c yer394w	ygl121c ypr160w ylr149c yer067w ylr178c	ygr043c ygl191w ymr107w yel039c ynl160w	ykl148c yjl166w ygr183c ymr175w yfl030w
4	ynl055c ymr251w-a ymr181c ydr530c yhr138c ynl237w ypl222w yml120c ykl016c yol048c ypl154c yor285w	ykr076w yol152w ymr271c ykl026c ypr184w ypl123c ydr018c yol084w yll020c ypl134c yhl021c ypl271w	ylr270w ybl045c yor120w ygr063w yir039c yor317w ydl067c ypr006c yor136w ydl168w yll009c ypl078c	ylr356w ydr513w ypr193c ylr164w ymr081c yer015w yil113w ypr002w ypl201c ygr194c yml131w	ygl188c ydr322c-a ydl222c ylr258w yol077w-a yml081c-a yjl161w ypl186c ygr174c yjl164c ynl037c	yil111w yel060c ydl124w ycl064c ydr343c yol126c ykl142w ypl087w yjl144w ylr294c yol083w	yjl163c yhl032c ydl021w ydr377w ykr049c ypl135w ylr295c ydl110c ylr080w yor374w yor289w

Group	Consistent genes included in the group						
5	ydr341c	ykr059w	yjr016c	ymr307w	ynl175c	ydl081c	yer036c
	yor182c	yor272w	ygl029w	ykl081w	ykr059w-r	ymr075c-a	yor108w
	ymr290c	ypl211w	yor310c	ydr087c	yer110c	ygl076c	ygr272c
	yjl158c	yjl138c	yfl045c-r	ynl182c	yor344c	ydr101c	ykl056c
	ynl110c	ypl160w	ydr324c	yer055c	ygl120c	ykl006w	yjr063w
	yol097c	ypl273w	ykr057w	ykl009w	ydl229w	yhr167w	yor254c
	ypl131w	yjl177w	yhr287c-a	ypr187w	yhr064c	yjr070c	yhr121c
	yhr406c	yml022w	yol077c	ypr069c	yhr170w	yhr432w	ypl043w
	yor340c	ypl126w	ygr118w	yil096c	yhr216w	ykr081c	ydl192w
	yfl045c	ykl153w	ynl132w	ypl090c	yor247w-r	ypr190c	

Table 5.6: Degrees of clustering consistency for all genes are tabulated.

Max. no. of occurrence in one group among three experiments	3	2	1
No. of genes	276	205	9
	(56.33%)	(41.84%)	(1.84%)

are inhomogeneous. The time profiles of consistent genes for two strains in group 1 are further investigated in Figure 5.7. From Figure 5.7, it is observed that the negative correlations between two strains may be due to the differences in time shifts or time trends of time profiles in BY and RM strains. Therefore, the regression results of group 1 in Figure 5.6 show the mixing effects of these two types. These interesting phenomena occur not only in three experiments of the training set but also the experiment in the test set. These are interesting observations that need more investigations in the future.

The lists of consistent genes in these five groups are reported in Table 5.5. The clustering consistency for all genes can be further evaluated by Table 5.6. From Table 5.6, the probabilities of consistent genes in three experiments of the training set among all and reference genes are over 56%. This is very high because the consistent

probability is only  $5/125 = 4\%$  when one gene is randomly clustered 5 clusters for three experiments. Hence, these consistent genes have consistent patterns among three experiments in the training set.

### 5.3 Example with simulation and real data

For the pair of samples consisting of three elements as list in the right part of Fig. 4.2, we uniformly generate 100 input samples and their corresponding possible output samples with misclassification probability  $p = 0.05$ . For the prerequisite relation, if the state of Boolean with input genes is ON, then we let the output value have equal probability with ON and OFF. The data can be arranged as input/output sample similar to that obtained from micorarray data with time. Namely, the input of each sample can represent the gene expression at time  $t$  and the output can represent the gene expression at time  $t + 1$ . For each pair of input and output genes, we compute the 8 p-scores which represent the 8 basic hypotheses in Table 4.3 for all of pair input genes and output genes. After calculating, the results are shown in Table 5.7.

Table 5.7: For the Time Delay Boolean network in figure 1, we generate 100 samples, and take  $p=0.05$

Samples		Hypotheses								Relation
Input	Output	$q_{000}=0$	$q_{010}=0$	$q_{100}=0$	$q_{110}=0$	$q_{001}=0$	$q_{011}=0$	$q_{101}=0$	$q_{111}=0$	
$v_1, v_2$	$v'_1$	0.569	0.192	0.213	0.230	0.016	0.251	0.419	0.210	
$v_1, v_3$	$v'_1$	0.459	0.419	0.226	0.253	0.218	0.089	0.344	0.435	$(v_1 \text{ or } v_2) \prec v'_1$
$v_2, v_3$	$v'_1$	0.547	0.411	0.297	0.422	0.194	0.315	0.432	0.244	
$v_1, v_2$	$v'_2$	0.327	0.272	0.331	0.266	0.018	0.075	0.172	0.214	
$v_1, v_3$	$v'_2$	0.337	0.235	0.323	0.248	0.042	0.081	0.056	0.293	$(v_1 \text{ and } v_3) \prec v'_2$
$v_2, v_3$	$v'_2$	0.367	0.283	0.316	0.218	0.017	0.169	0.072	0.150	
$v_1, v_2$	$v'_3$	0.210	0.038	0.361	0.015	0.047	0.211	0.034	0.346	
$v_1, v_3$	$v'_3$	0.339	0.478	0.386	0.644	0.640	0.260	0.374	0.467	$v_2 \sim v'_3$
$v_2, v_3$	$v'_3$	0.274	0.293	0.029	0.029	0.049	0.040	0.291	0.264	

Next, we have to decide the threshold for choosing the relations. When we increase the threshold of the p-score, there are some relations whose p-scores are smaller than the threshold, and the relations would be chosen. Moreover, when the number is 0.089, the conflict occurs, since we have  $(v_1 \text{ or } v_2) \prec v'_1$  and  $(v_1 \text{ or } \bar{v}_3) \prec v'_1$ . However, in our model, there are at most two genes which would affect an output gene. Therefore, we can choose 0.089 as our threshold and include the relations whose p-score is smaller than the threshold. By these procedures, we can reconstruct the Time Delay Boolean network identical with Fig. 4.2.

In the area of gene regulatory network study, (Schuller, 2003) summarized regulatory cis-acting elements of structural genes of the nonfermentative metabolism and described the molecular interactions among general regulators and pathway-specific factors. In the gene regulation of gluconeogenesis by Sip4 and Cat8 pathway, the carbon source control could be identified for the regulator Cat8; see (Figure 6) in (Schuller, 2003). For the experimental data collection, we use the microarray expression dataset of yeast *Saccharomyces cerevisiae* produced by (Spellman et al., 1998) and (DeRisi et al., 1997). By these data sets, we can reconstruct the biological pathway using our proposed method. Under the Time Delay Boolean network model, we reconstruct the genetic regulation network as shown in Figure 5.8. The result is consistent with the genetic network in literature. That is, the restraint of Mig1 or activation of Snf1 is prerequisite for the decreasing of Cat8. Moreover, the restraint of Snf1 or Cat8 is prerequisite for the decreasing of Mls1. However, the negative similarity between Snf1 and Mig1 is undetectable in our current model.

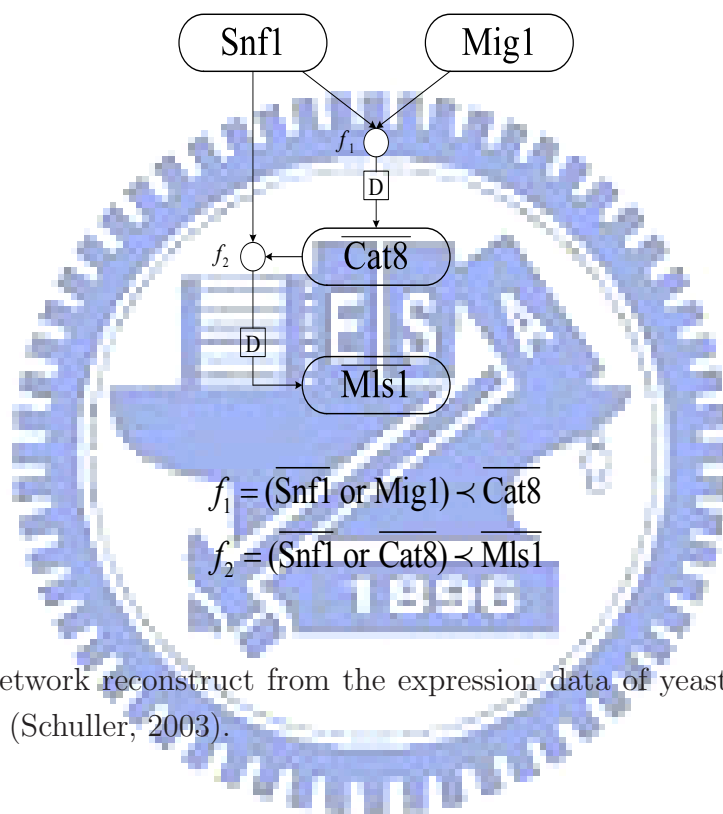


Figure 5.8: Network reconstruct from the expression data of yeast *Saccharomyces cerevisiae*; see (Schuller, 2003).

# Chapter 6

## Conclusion and Discussion

In the study of emotion recognition, we have compared six typical classifiers by their performances in emotion recognition using physiological signals with daily and personal correction by MANOVA. As the results mentioned above, these classification methods can be very useful to perform emotion recognition by using the physiological signals with daily and personal correction. In particular, we can successfully correct daily effects using the statistical techniques of MANOVA.

There are still challenges for future studies. For example, we could investigate and determine significant features using feature selection and dimensional reduction methods. In addition, more data collection could be performed in future studies to improve the accuracy. Real-time applications could be further investigated for the prediction of emotional states based on the physiological signals with daily correction. Further adjustments of parameters in classification methods could be investigated. These are interesting topics that we plan to study in the future based on the framework of the current research results.

Regarding the analysis of yeast, five major clusters of gene expression time profiles were discovered in this study. Four clusters show positive correlations between gene expression profiles in BY and RM strains. The estimated time shifts of expression time profiles in these four clusters are mainly 1 hour after the time that glucose consumption drops. The first cluster shows very interesting pattern of negative cor-

relations between gene expression profiles in BY and RM strains. In this group, the estimated time shift of expression time profiles are mainly 1 hour before the time that glucose consumption drops. These consistent genes show negative correlations in two strains are: yor128c, ymr120c, ydr170w-a, yol143c, yor375c-r, ylr346c, yor375c, yhr163w, yor273c, ygr234w. The negative correlations in two strains could be due to the differences of time shifts or the differences in expression shapes in two strains according to the time profiles from microarray data. The experiment data by RT-PCR can be studied to confirm the time profiles of consistent genes in the group of negative correlation of expressions in BY and RM strains in the future. Other models are possible to analyze these microarray data. For instance, time series models with dependent errors, longitudinal models, models of functional data analyses and so forth. Besides, network analysis such as Boolean network or Bayesian network could be used to investigate the causal relationship of these interesting genes. These will be of interest to investigate in future studies.

For the study of Time Delay Boolean network, we introduced the Time Delay Boolean network which generalizes the Boolean network model in order to cope dependencies that have time delay relationships. The approach to genetic network inference from gene expression data rely on the assumption that only the expression of a gene is likely to be controlled by a relatively small number (say  $k$ ) of genes. Some bounds on the size of data needed for the identification of the Time Delay Boolean networks under constant of indegree are stated. Moreover, the algorithm of the network reconstruction from data with noise are developed.

In practice, there exists differences between real biological systems and Boolean networks. Nodes in a Boolean network take binary values updated synchronously. In contrast, quantities of gene expression in real cells are continuous and vary with time. Hence, we need to discretize them. The gene expression which is increasing or decreasing with time is also a possible discretization choice.

Work in progress is aimed at evaluating the effectiveness of the described ap-

proach for inferring genetic networks from biological gene expression time series data. Besides, implementation on some other real biological data is also an important task.

For the implement of the inference algorithm, the most complexity is the computation of p-score for each of the  $\frac{n!}{k!(n-k)!}$  input elements and  $n$  output elements, where  $n$  is the number of elements and  $k$  is the number of indegree. It is an iterative algorithm to compute the MLE for the p-scores by E-M procedure and the common practice is setting an upper bound for iterations in numerical implementation. Consequently, this keeps the  $O(n^{k+1})$  complexity for the computation of MLE. Moreover, the sorting algorithm for the  $\frac{n!}{k!(n-k)!}n$  data cost  $O(n^{k+1} \log(n))$  in time. Hence, the overall time complexity is  $O(n^{k+1} \log(n))$  in this algorithm.





# Bibliography

- Aha, D. W., Kibler, D., Albert, M. K., 1991. Instance-based learning algorithms. *Machine Learning* 6 (1), 37–66.
- Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S., 1998. Identification of gene regulatory networks by strategic gene disruptions and gene overexpression. *Proc. 9th ACM-SIAM Symp. Discrete Algorithms*, 695–702.
- Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S., 2003. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model. *Theoretical computer science* 298 (1), 235–251.
- Akutsu, T., Miyano, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Proc. Pacific Symposium on Biocomputing*, 17–28.
- Ark, W. S., Dryer, D. C., Lu, D. J., 1999. The emotion mouse. *International conference on Human-computer Interaction*, 818–823.
- Breazeal, C., Aryananda, L., 2002. Recognition of affective communicative intent in robot directed speech. *Autonomous Robots* 12, 83–104.
- Brem, R. B., Yvert, G., Clinton, R., Kruglyak, L., 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755.
- Burnham, K. P., Anderson, D. R., 1998. *Model selection and inference*. Springer, New York.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., 2004. Analysis of emotion recognition using facial expressions, speech and multinodal information. *International Conference on Multimodal Interfaces*, 205–211.
- Chuang, C. F., Shih, F. Y., 2006. Recognizing facial action units using independent component analysis and support vector machine. *Pattern Recognition* 39 (9), 1795–1799.

- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18 (1), 32–80.
- Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. *International Conference on Spoken Language Proceedings* 3, 1970–1973.
- DeRisi, J. L., Iyer, V. R., Brown, P. O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278 (24), 680–686.
- Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95, 14863–14868.
- Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: a survey. *Pattern Recognition* 36 (1), 259–275.
- Friedman, N., Linial, M., Nachman, I., Pe’er, D., 2000. Using bayesian networks to analyze expression data. *J. Comp. Biol.* 7, 601–620.
- Gaffney, S., 2004. Probabilistic curve-aligned clustering and prediction with mixture models. Ph.D. thesis, Department of Computer Science in University of California, Irvine.
- Gaffney, S., Robertson, A. W., Smyth, P., Camargo, S. J., Ghil, M., 2007. Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dynamics* 29 (4), 423–440.
- Gaffney, S., Smyth, P., 2004. Joint probabilistic curve clustering and alignment. *Advances in Neural Information Processing Systems* 17, 473–480.
- Galindo, C. L., Gadl, A. A., Sha, J., Chopra, A. K., 2004. Microarray analysis of aeromonas hydrophila cytotoxic enterotoxin-treated murine primary macrophages. *Infection and Immunity* 72 (9), 5439–5445.
- Gancedo, J. M., 1998. Yeast carbon catabolite repression. *Microbiology and Molecular Biology Reviews* 62 (2), 334–361.
- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O., 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* 11, 4241–4257.
- Heckerman, D., Geiger, D., Chickering, D. M., 1995. Learning bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 20, 197–243.

- Hu, T., De Silva, L. C., Sengupta, K., 2002. A hybrid approach of nn and hmm for facial emotion classification. *Pattern Recognition Letters* 23 (11), 1303–1310.
- Hunt, E., Marin, J., Stone, P., 1966. *Experiments in induction*. Academic Press, New York.
- Jensen, F., 2001. *Bayesian networks and decision graphs*. Springer, New York.
- Jensen, F. V., 1996. *An introduction to bayesian networks*. University College London Press, London.
- John, G. H., Langley, P., 1995. Estimating continuous distributions in bayesian classifiers. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 338–345.
- Kauffman, S. A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology* 22 (3), 437–467.
- Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K., 2001. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation* 13 (3), 637–649.
- Kerr, M. K., 2003. Design considerations for efficient and effective microarray studies. *Biometrics* 59 (4), 822–828.
- Kerr, M. K., Churchill, G. A., 2001. Experimental design for gene expression microarrays. *Biostatistics* 2 (2), 183–201.
- Kim, K., Bang, S., Kim, S., 2004. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing* 42 (3), 419–427.
- Lauritzen, S., Spiegelhalt, D., 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistical Society Series B* 50, 157–224.
- Le Cessie, S., Van Houwelingen, J., 1992. Ridge estimators in logistic regression. *Applied Statistics* 41 (1), 191–201.
- Li, L. M., Lu, H. S., 2005. Explore biological pathways from noisy array data by directed acyclic boolean networks. *Journal of Computational Biology* 12 (2), 170–185.
- Littlewort, G., Bartlett, M. S., Fasel, I. R., Chenu, J., Kanda, T., Ishiguro, H., Movellan, J. R., 2004. Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. *Advances in Neural Information Processing Systems* 16, 1563–1570.

- Nasoz, F., Alvarex, K., Lisetti, C. L., Finkelstein, N., 2004. Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology and Work* 6 (1), 4–14.
- Nwe, T., Foo, S., De Silva, L., 2003. Speech emotion recognition using hidden markov models. *Speech communication* 41, 603–623.
- Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo.
- Picard, R. W., Vyzas, E., Healy, J., 2001. Toward machine emotional intelligence: analysis of affective psychological states. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (10), 1175–1191.
- Quinlan, J., 1993. C4.5: Programs for machine learning.
- Rani, P., Liu, C., Sarkar, N., Vanman, E., 2006. An empirical study of machine learning techniques for affect recognition in human-robot interaction. *Pattern Analysis and Applications* 9, 58–69.
- Schuller, H.-J., 2003. Transcriptional control of nonfermentative metabolism in the yeast *saccharomyces cerevisiae*. *Current Genetics* 43 (3), 139–160.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of Cell* 9, 3273–3297.
- Vapnik, V. N., 1998. *Statistical learning theory*. Wiley, New York.
- Yacoob, Y., Davis, L. S., 1996. Recognizing human facial expression from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18 (6), 636–642.
- Zhou, C., Lin, X., 2005. Facial expressional image synthesis controlled by emotional parameters. *Pattern Recognition Letters* 26 (16), 2611–2627.