Table of Contents:

Chapter 1.	**Introduction**	**2**
Chapter 2.	**Microarray Data**	**4**
Chapter 3.	**Gene Filtering**	**10**
Chapter 4.	**Cluster Analysis**	**15**
Chapter 5.	**Regression Models with Time Shifts**	**26**
Chapter 6.	**Results**	**29**
Chapter 7.	**Conclusion and Discussion**	**45**
References		**46**

1. Introduction

It is very important to understand the biological process of diauxic shift in fermentation for yeast (DeRisi, Iyer and Brown, 1997, Gasch, Spellman, Kao, Carmel-Harel, Eisen, Storz, Botstein, and Brown, 2000, Schuller 2003). In the laboratory of Dr. Wen-Hsiung Li at Genomics Research Center of Academia Sincia, Dr. Huang-Mo Sung and coworkers have conducted the two-dye oligonucleotide microarray experiments for yeast fermentation to study the biological process of diauxic shift. Two yeast strains of BY4741 and RM11-1a were used. Duplicated spots were used in one microarray. Designs of common reference and dye swapping were used. The microarray experiments were performed at various time points for the period of diauxic shift.

This study will perform the statistical analysis of these microarray data. In particular, we will investigate the selection of differentially expressed genes related to the process of diauxic shift, clustering of differentially expressed genes and time shifts between expression profiles *vs.* glucose consumptions. One example of time shift is illustrated in Figure 1.1.

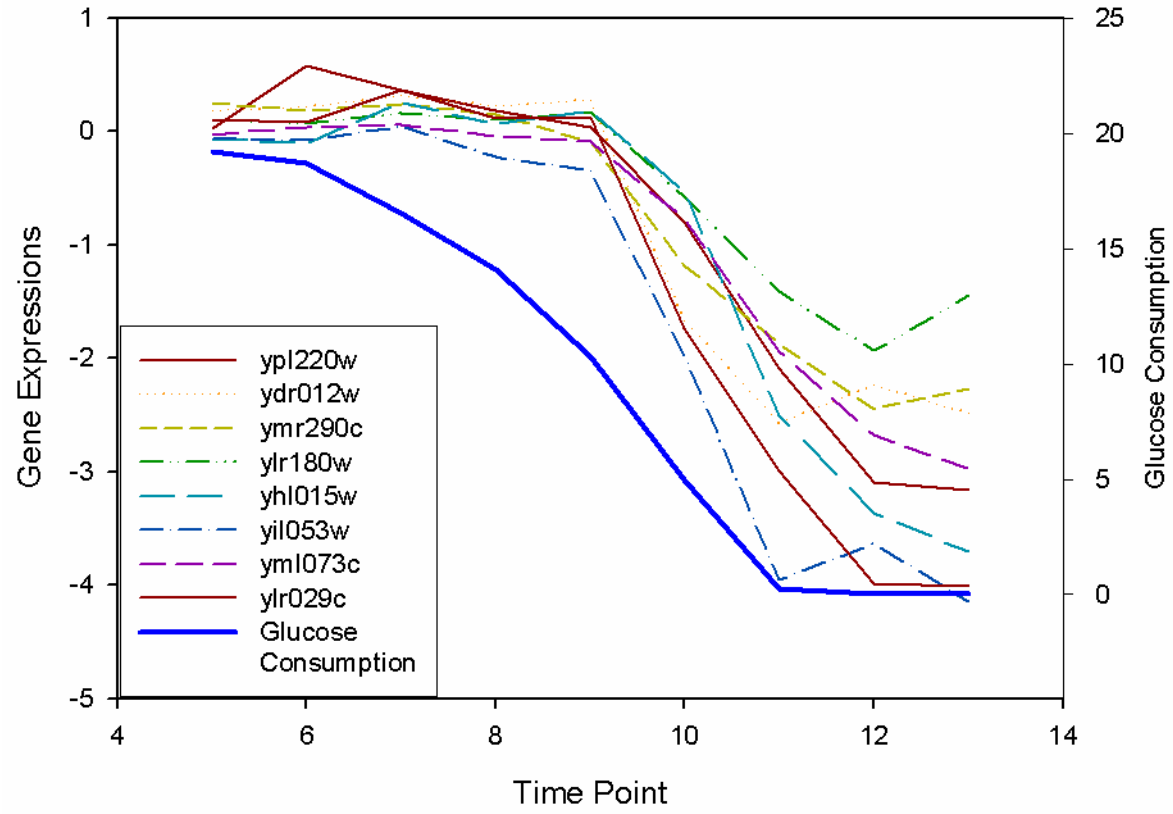


Figure 1.1: The expression profiles change later than the time that glucose consumption drops.

2. Microarray Data

Data Preprocessing

Microarray data were obtained by the GenePix software after scanning and they were saved as *.gpr files. The median intensities of foreground and background in every spot will be extracted from the files for Cy3 and Cy5 dyes. Each file name includes the information of experiment date and experiment design. For example, one typical file name is "20040921-B3-3-BY4741t4c3-BY4741t5c5-460630-g.gpr". This file name indicates this experiment was performed on Sep. 21, 2004 for BY4741 strain. Cy3 (c3) and Cy5 (c5) dyes were applied to the yeast mRNAs that has been fermented for the common reference time at 4 hours (t4) and the experiment time at 5 hours (t5) respectively. The common reference time is always set at t4 in this study. If the reference time t4 is next to c3 dye in the file name, then Cy3 dye was applied for the common reference time at 4 hours (t4) and the swap index is set as 0. Otherwise, the swap index is set as 1 for the swapped array, like the file name of "20040921-B3-3-BY4741t5c3-BY4741t4c5-460630-g.gpr". There are totally four sets of microarray experiments and the detail of total information is listed in Table 2.1:

Experiment	Date	Time Point	Strain	No. of Arrays
1	2004.09	5, 6, 7, 8, 9, 10, 11, 12, 13, 24	BY & RM	10 × 2 × 2 = 40
2	2004.12	4, 6, 8, 9, 10, 11, 12,13, 14, 16, 18, 20	BY & RM	12 × 2 × 2 = 48
3	2005.03	5, 6, 7, 8, 9, 10, 11, 12,13, 14, 24	BY & RM	11 × 2 × 2 = 44
4	2005.09	5, 6, 7, 8, 9, 10, 11, 12,13, 14, 24	BY & RM	11 × 2 × 2 = 44

Table 2.1: The details of four microarray experiments are listed. In the calculation of microarray numbers, the first number is the total number of time points. The first multiplication of two is because two microarrays for BY and RM strains are conducted for one time point. The second multiplication of two is due to the fact that there are two dye swapped arrays for every strain in one time point.

In order to obtain the expression ratios of genes from microarrays, the following preprocessing and normalization are considered in this study. First, the background correction is 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 correction is smaller than one, then the expression intensity is set as one, which will be zero after log transformation. Because the dye efficiencies of Cy3 and Cy5 could be different, 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 for one gene totally for every strain in one time point that are obtained as follows.

$$\text{If Swap} = 0\text{ , Ratio}_{ijr} = \frac{I532_{ijr} / Median_{j=1,\ldots,6367,r=1,2}\{I532_{ijr}\text{ in array i}\}}{I635_{ijr} / Median_{j=1,\ldots,6367,r=1,2}\{I635_{ijr}\text{ in array i}\}};$$

$$\text{If Swap} = 1\text{ , Ratio}_{ijr} = \frac{I635_{ijr} / Median_{j=1,\ldots,6367,r=1,2}\{I635_{ijr}\text{ in array i}\}}{I532_{ijr} / Median_{j=1,\ldots,6367,r=1,2}\{I532_{ijr}\text{ in array i}\}};$$

where

$I532_{ij}$ = F532_Median$_{ij}$-B532_Median$_{ij}$ for Cy3,

$I635_{ij}$ = F635_Median$_{ij}$-B635_Median$_{ij}$ for Cy5,

$i = 1, 2, \ldots, 176$ (176 array files totally),

$j = 1, 2, \ldots, 6367$ (6367 genes totally),

$r = 1, 2$ (two replicated genes in every array).

The average in these four ratios for one gene is used to further normalize the dye and block effects from a pair of two swapped microarrays with two duplicated spots in one array. Thus, we can generate the data matrix for further analyses as Table 2.2.

Exp	Time	T5	T5	T6	T6	…	T13	T13
	Strain	BY	RM	BY	RM	…	BY	RM
	Ratio	BY_t4 /BY_t4	RM_t5 /RM_t4	BY_t6 /BY_t4	RM_t6 /RM_t4	…	BY_t13 / BY_t4	RM_t13 / RM_t4
Exp 1								
Exp 2								
Exp 3								
Exp 4								

Table 2.2: The data matrix of ratios for four experiments is illustrated.

Furthermore, the log2 transformation of ratio is used to evaluate the relative gene expression of one gene in a strain at a specific time referring to the common reference at t4. Those genes names with "-x" are duplicated or other types of genes and they will be regarded as different genes at this stage.

Reference Genes

The purpose of our study is that we try to select genes which have significantly differential expressions over time and related with glucose consumptions. A group of reference genes has been reported in literature and they were summarized in Table 2.3

by Dr. Sung. However, some of reference genes may not be significantly nor consistently expressed in these four microarray experiments. We will perform statistical analysis to select genes with significantly and consistently differential expressions in microarray data firstly. Then, these selected genes will be compared with the reference genes.

ykr097w	ybr072w	ybl045c	ylr340w
ylr377c	yfl014w	ypl012w	yml073c
yal054c	ykl026c	ynl141w	yhl015w
yer065c	ygr043c	ymr290c	ylr029c
yjr095w	yor065w	ylr180w	ydr012w
ylr174w	ynl052w	yil053w	
ynl117w	yhr051w	ygr160w	
ylr258w	ygl191w	ydr398w	
ygr088w	yel024w	ynr069c	
ydr171w	ydr529c	ypl220w	

Table 2.3: Thirst-five reference genes have been reported in literature.

The analysis flow chart is illustrated in Figure 2.1. The microarray data in experiment 1, 3 and 4 are used as the training set because they have common

experiment time points. The microarray data in experiment 2 will be used as the test set to evaluate the performance of analysis results from the training set. Genes will be filtered by the regression coefficients of expression *vs.* time in the training set. These unfiltered genes will be clustering by the methods of hierarchical clustering and curve clustering by the training set of microarray data. One clustering method will be selected based on the performances of clustering results in the training and test sets. The number of cluster will be also determined accordingly. For every cluster, the time shift will be estimated by the regression tests between gene expressions and glucose consumptions. The details of analyses are discussed in next chapters.

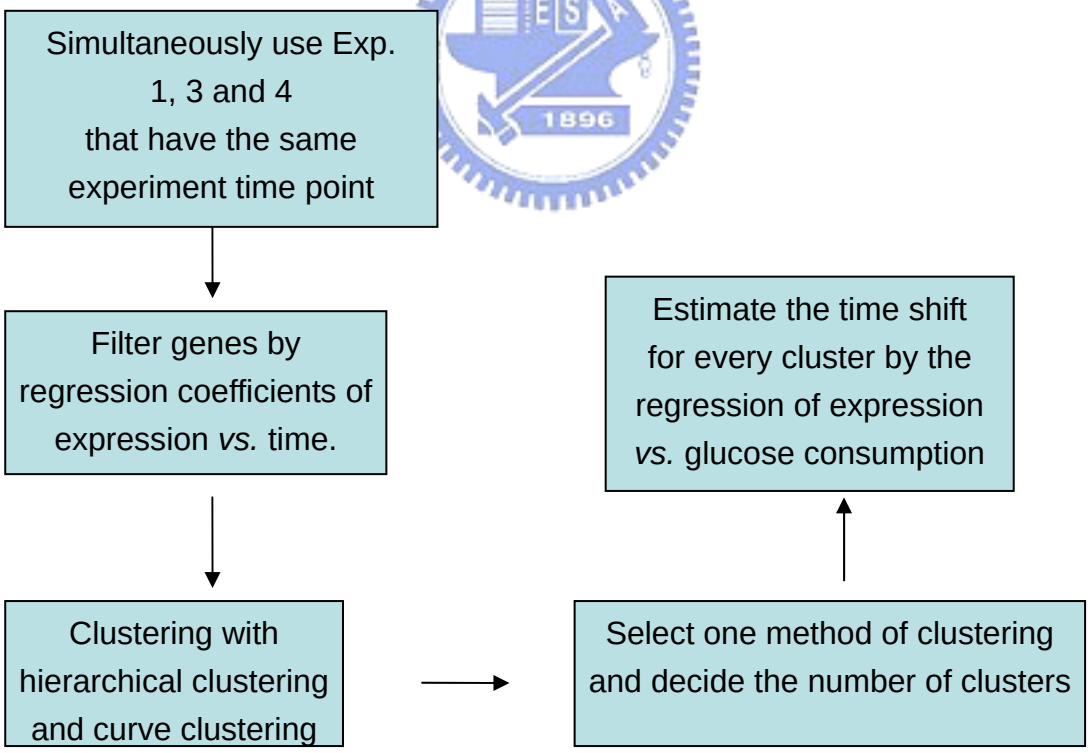


Figure 2.1: the flowchart in our studying.

3. Gene Filtering

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. That is, the following regression model is used for every gene in one strain and one experiment,

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

where log(*Ratio)* is the log ratio of gene expression, *Time* is the time point ranging through 5 to 13 as in Table 2.1, α_0 is the intercept, α_1 is the regression coefficient of slope and ε is the random noise. An example is given in Figure 3.1.

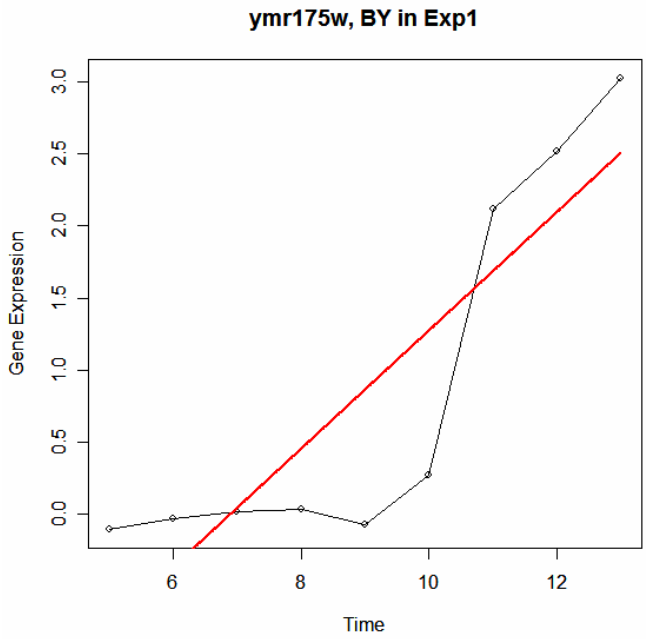
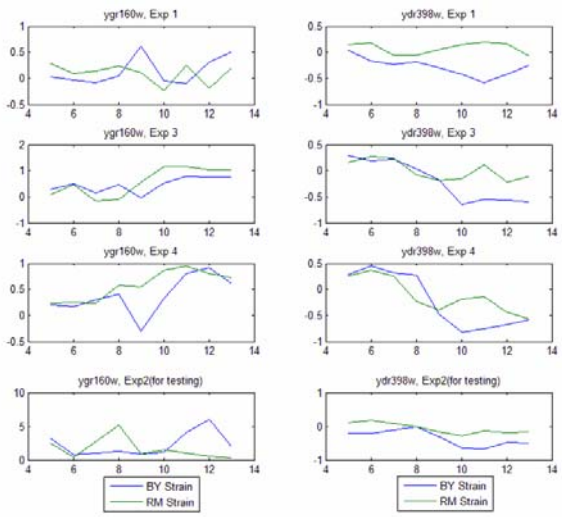


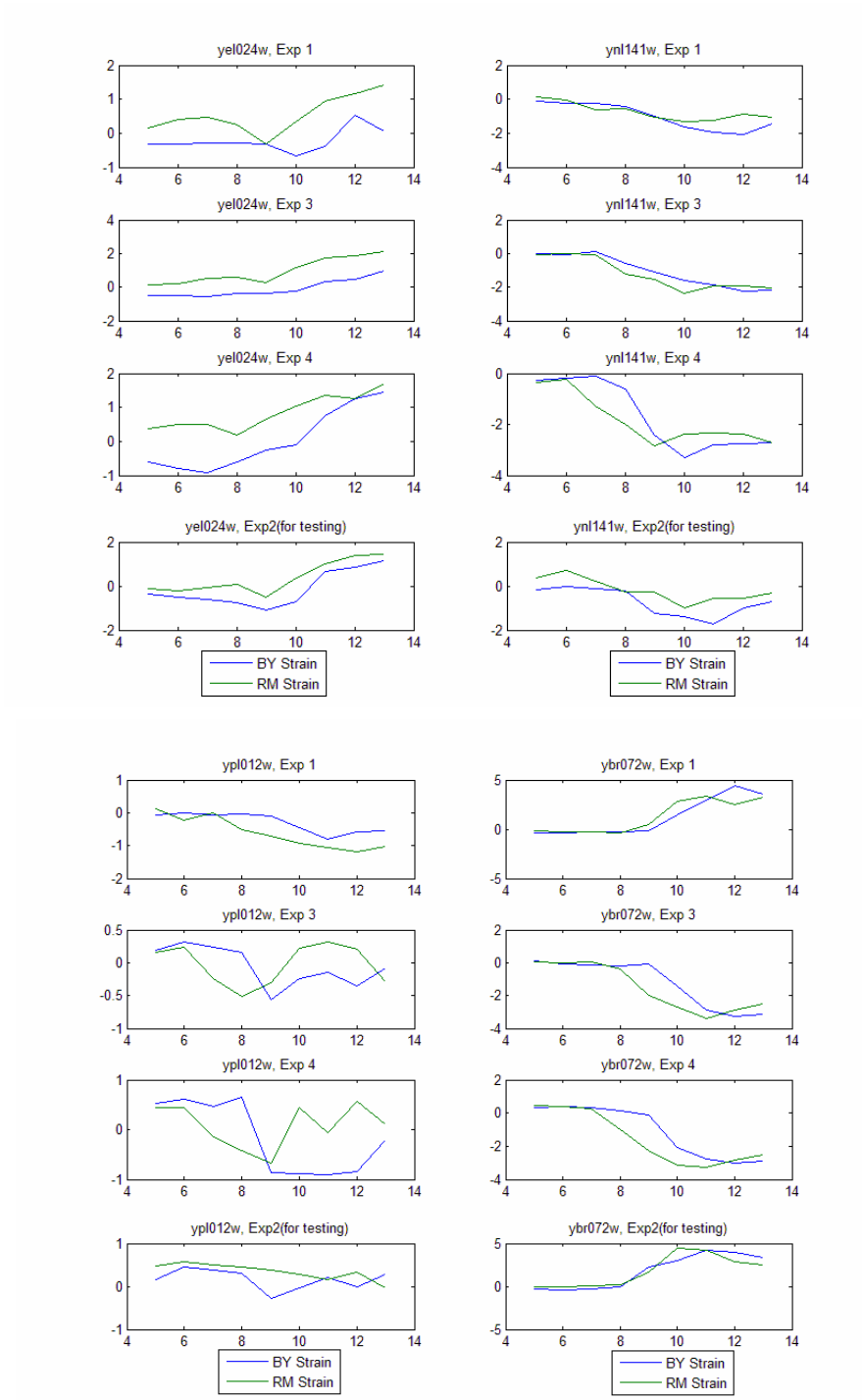
Figure 3.1: An example is given for the expression profile and the fitted regression line of gene expression versus time. If the absolute value of fitted slope, α_1, is large, then gene expression varies much over time.

For every gene in one strain, there are three regression slopes in experiment 1, 3 and 4. The coefficient of variation (CV) is calculated as the ratio of the standard deviation over the average of three slopes. If the CV value is high, then the expression slopes vary a lot or the average is small among three experiments. Hence, those genes with CV values large than a threshold can be filtered and the 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 the patterns of positive correlation in two strains and few genes have the 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 difference between the average slopes in two strains, like the threshold of 0.3 in this study. This partition is considered to keep genes that have large expression variation in one strain but not the other strain. Consequently, there are three groups of remained genes now. For the first group of positive correlation and large differences of average slopes in two strains, all unfiltered genes are kept because they have large expression variation in one strain but not 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 the threshold of 0.3499 in this study. 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 the threshold of 0.2 in this study. As a result, there are 488 genes kept in this study and 26 reference genes are included.

The above approach of gene filtering is used to keep genes that could have significant expression patterns in this study. These 488 genes will be further selected after checking the clustering consistency that will be investigated in the later chapters. Other methods of gene filtering could be studied in the future. There are 9 references genes not included and their time profiles are displayed in Figure 3.2. Most of these filtered reference genes do not have significant and consistent expression profiles in the microarray data.




yel024w, Exp 1
ynl141w, Exp 1
yel024w, Exp 3
ynl141w, Exp 3
yel024w, Exp 4
ynl141w, Exp 4
yel024w, Exp2(for testing)
ynl141w, Exp2(for testing)
BY Strain
RM Strain
ypl012w, Exp 1
ybr072w, Exp 1
ypl012w, Exp 3
ybr072w, Exp 3
ypl012w, Exp 4
ybr072w, Exp 4
ypl012w, Exp2(for testing)
ybr072w, Exp2(for testing)
BY Strain
RM Strain

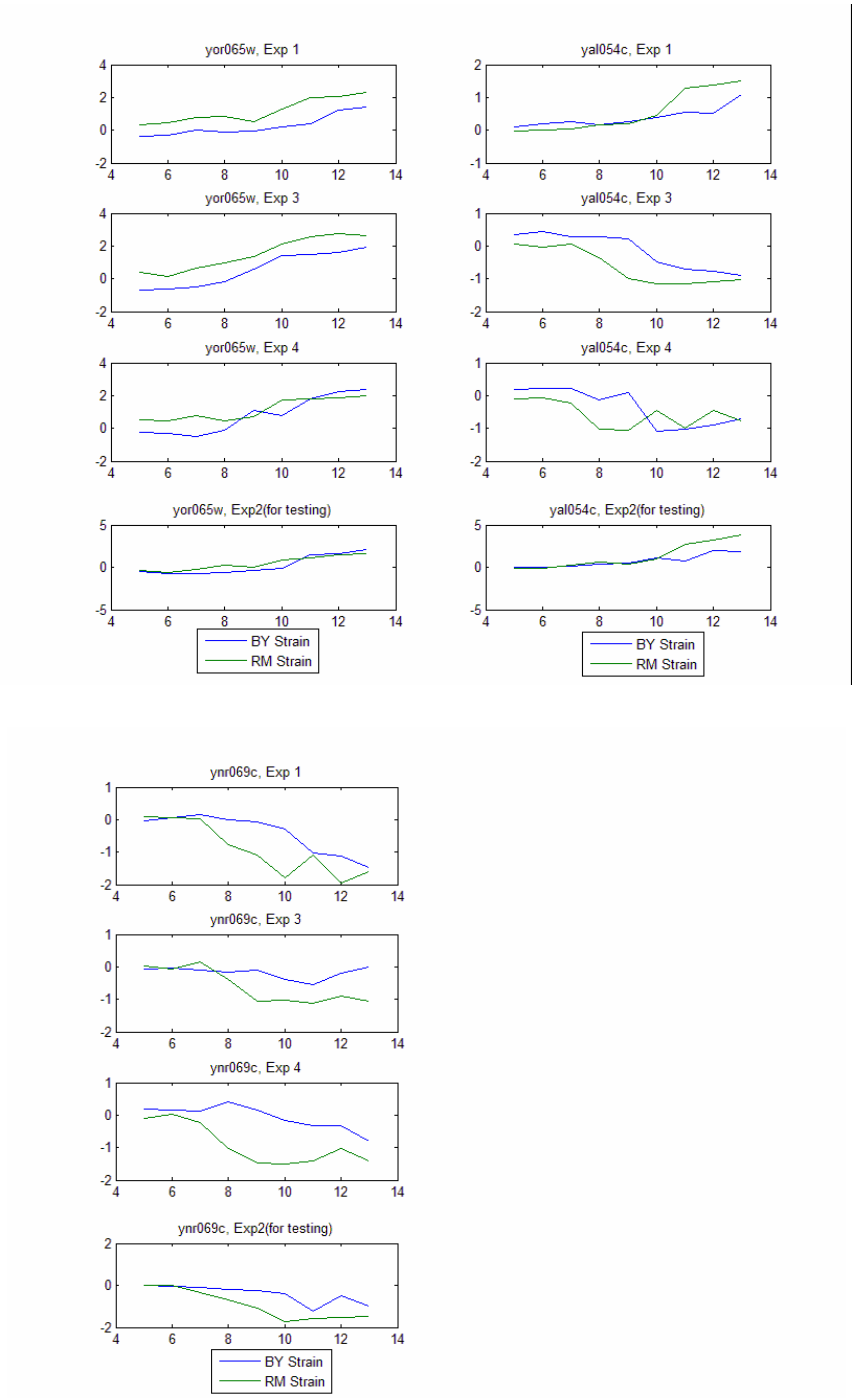


Figure 3.2: The expression profiles of nine filtered reference genes are plotted.

4. Cluster Analysis

The expression profiles of unfiltered genes will be used to perform cluster analysis. Suppose one gene is clustered into group g1, g2 and g3 in the 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. Then, the prediction square error (*PSE*) is the square error between predicted expression and the observed expression of the gene in the test set as follows.

$$PSE = \sum_{i=1}^{488} \sum_{j=1}^{18} \frac{(\log(R_{2,ij}) - \log(R_{Pred,ij}))^2}{18}, \quad (4.1)$$

where $R_{2,ij}$ means the gene expression of i-th gene in j-th microarray data and $R_{Pred,ij}$ is its predicted value by the clustering method. For avery gene, the microarray data contain 18 gene expressions at nine time points for two strains. 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 if the expression time profile of one gene in different experiments will be clustered into the same group or not. One example is illustrated in Figure 4.1. Genes will clustering consistency will be selected to find the representative curves in every group.

	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 4.1: In this case, gene g2 is considered to have clustering consistency.

Hierarchical Clustering

Hierarchical clustering is a nonparametric method to cluster data (Eisen, Spellman, Brown and Botstein 1998). The basic ideal 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 record the pairwise distance for expressions of any two genes. 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:

$$d(z^r, z^s) = \left[\sum_{j=1}^{d} \left(z^r{}_j - z^s{}_j \right)^2 \right]^{1/2}; \quad (4.2)$$

(Pearson's) Correlation distance:

$$d(z^r, z^s) = 1 - \mathrm{cor}(z^r, z^s) = 1 - \frac{\mathrm{cov}(z^r, z^s)}{\sqrt{\mathrm{var}(z^r)\,\mathrm{var}(z^s)}}; \quad (4.3)$$

where z^r and z^s are two observation vectors in d-dimension, $z^r{}_j$ and $z^s{}_j$ are the components of two observation vector in d-dimension, cor and var are the sample variance and covariance.

In the second step, it is necessary to define the *linkage*, which defines the distance between two groups. There are three kinds of linkages that are commonly considered in literature. (Add in the results of single linkage in the comparisons of PSEs!)

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

$$d(G_i, G_j) = \min_{z^r \in G_i, z^s \in G_j} d(z^r, z^s). \quad (4.4)$$

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

$$d(G_i, G_j) = \max_{z^r \in G_i, z^s \in G_j} d(z^r, z^s). \quad (4.5)$$

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

$$d(G_i, G_j) = \underset{z^r \in G_i, z^s \in G_j}{average}\, d(z^r, z^s). \quad (4.6)$$

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

We will try two kinds of distances and three kinds of linkages, 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 four different results for hierarchical clustering as shown in Figure 4.2. By the comparisons of PSEs for different cluster sizes in Figure 4.2, 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 4.3.

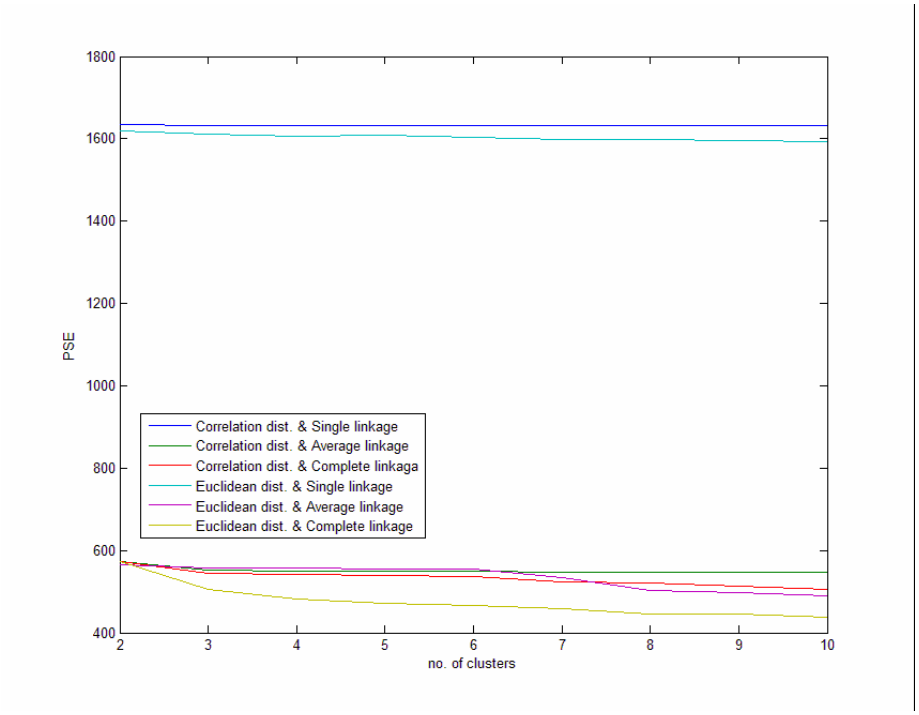


Figure 4.2: Comparisons of PSEs for different cluster sizes are plotted for hierarchical clustering with different settings.

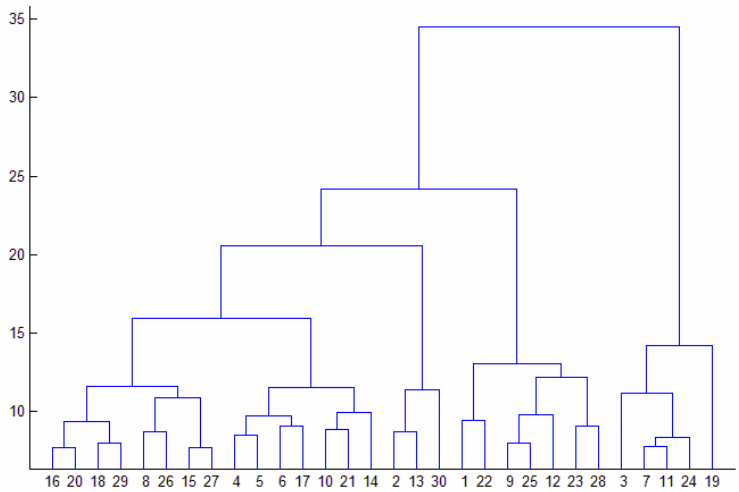


Figure 4.5: The dendrogram of the hierarchical clustering is shown for 30 nodes.

Curve Clustering

The alternative clustering method that could be applied to cluster expression

profiles can be the method of curve clustering. This method has been proposed to cluster curves based on mixture models (Gaffney, 2004, Gaffney and Smyth, 2004) and the toolbox for matlab is available (http://www.ics.uci.edu/~sgaffney/CCT/). Basically, that method assumed a mixture model with expectation-maximization (EM) algorithm to estimate parameters in the mixture model, which are reviewed below. Suppose that $\mathbf{y_i}$ is a sequence of curve measurements that are observed at the $\boldsymbol{n_i}$ time points in $\boldsymbol{x_i}$. He defines a cluster-specific conditional probabilistic model, which is denoted as $p_k(y_i \mid x_i, \theta_k)$ for the probability distribution in cluster k with parameters θ_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:

$$y_i = \mathbf{X_i}\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2\mathbf{I}), \quad (4.7)$$

where the $n_i \times p$ regression matrix $\mathbf{X_i}$ is the Vandermonde matrix evaluated at x_i, β is the p-vector of regression coefficients, ε_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

$$\mathbf{X_i} = \begin{bmatrix} 1 & x_{i1} & x_{i1}^2 & \cdots & x_{i1}^p \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{in_i} & x_{in_i}^2 & \cdots & x_{in_i}^p \end{bmatrix}. \quad (4.8)$$

Then, the conditional probability of y_i give x_i as $N(y_i \mid \mathbf{X_i}\beta, \sigma^2\mathbf{I})$. The polynomial

regression mixture model of K clusters is defined to be:

$$\begin{aligned} p(y_i \mid x_i, \boldsymbol{\theta}) &= \sum_{k=1}^{K} \alpha_k p_k(y_i \mid x_i, \theta_k) \\ &= \sum_{k=1}^{K} \alpha_k N(y_i \mid \mathbf{X_i}\beta_k, \sigma_k^2 \mathbf{I}), \end{aligned} \tag{4.9}$$

where α_k is the mixing probability in kth cluster, p_k is the conditional probability of a Gaussian distribution with mean $\mathbf{X_i}\beta_k$ and covariance matrix σ^2_kI. The log-likelihood function N observations becomes

$$\log p(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k p_k(y_i \mid x_i, \theta_k). \tag{4.10}$$

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

$$L_c = \sum_{i=1}^{N} \log \alpha_{z_i} N(y_i \mid \mathbf{X_i}\beta_{z_i}, \sigma_{z_i}^2 \mathbf{I}). \tag{4.11}$$

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

$$\begin{aligned} w_{ik} = p(z_i = k \mid y_i, x_i) &\propto \alpha_k p_k(y_i \mid x_i) \\ &= \alpha_k N(y_i \mid \mathbf{X}_i \beta_k, \sigma_k^2 \mathbf{I}). \end{aligned} \tag{4.12}$$

And the conditional expectation Q is:

$$Q = E[L_c \mid y_i, x_i] = \sum_{i=1}^{N} \sum_{k=1}^{N} w_{ik} \log \alpha_k N(y_i \mid X_i \beta_k, \sigma_k^2 \mathbf{I}). \tag{4.13}$$

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

$$\hat{\beta}_k = \left[\sum_{i=1}^{N} w_{ik} \mathbf{X}_\mathbf{i}^{'} \mathbf{X}_\mathbf{i} \right]^{-1} \sum_{i=1}^{N} w_{ik} X_i^{'} y_i, \quad (4.14)$$

$$\hat{\sigma}_k^2 = \frac{1}{\sum_{i=1}^{N} w_{ik}} \sum_{i=1}^{N} w_{ik} \left\| y_i - \mathbf{X}_\mathbf{i} \beta_k \right\|^2, \quad (4.15)$$

and

$$\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{N} w_{ik}. \quad (4.16)$$

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 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 expression 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 4.6.

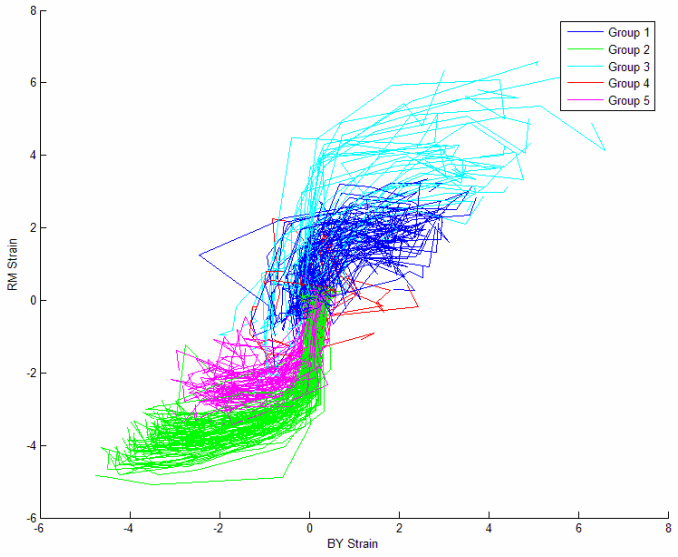


Figure 4.6: The typical results of two dimensional expression curves in experiment 1 for five groups are plotted.

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:

$$BIC = -2\log(L_{ML}) + K_a \log N, \quad (4.17)$$

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 for cluster sizes from 2 to 10 in Figure 4.7. As the BIC curve is decreasing when the cluster size is increasing in Figure 4.7, 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 Chapter 6.

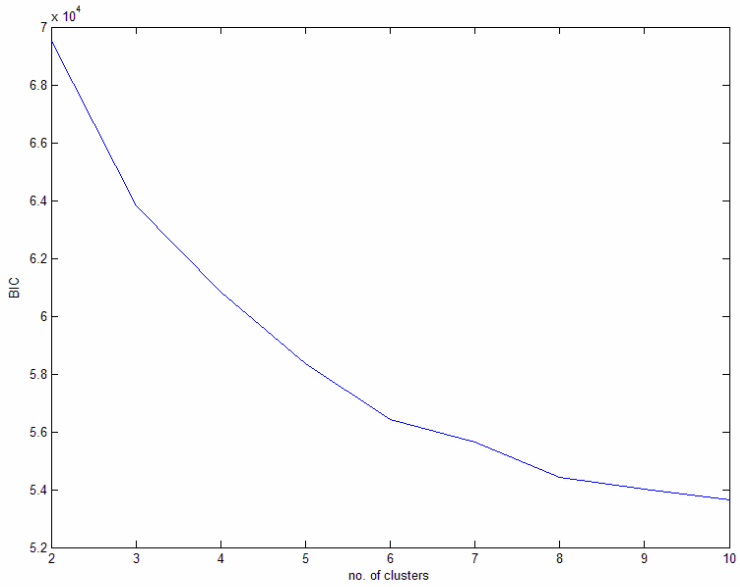


Figure 4.7: Model Selection by BIC is shown for curve clustering.

5. Regression Models with Time Shifts

The analysis of variance (ANOVA) has been applied for microarray data in literature (Kerr, 2000, Kerr, 2002, Kerr, 2002 Chi and Churchill, 2003, Dudoit, 2003, Cui and Churchill, 2003, Taesung, 2003). 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:

$$\begin{aligned}\log(Ratio(time)) &= \mu + \mu_{strain} + \mu_{time} + \mu_{exp} + \mu_{gene} + \mu_{time*gene} \\ &+ \gamma\ glucose(time + time\_shift) + error.\end{aligned} \quad (5.1)$$

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:

$$\begin{aligned}\log(Ratio(time)) = \mu + \mu_{time} + \mu_{exp} + \mu_{gene} + \mu_{time*gene} \\ + \gamma\ glucose(time + time\_shift) + error.\end{aligned} \quad (5.2)$$

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 null hypothesis of H0: beta = 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 droppes. 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 droppes. 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:

$$F_{Glucose} = \frac{SS_{Glucose}/1}{SS_{Error}/df_{Error}}. \quad (5.3)$$

The degree of freedom for the sum of squares of Glucose is equal to 1 since the Glucose term is treated as a one-dimensional independent 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 $\mu_{time*gene}$ are all equal to zeros); and H_1, the alternative hypothesis that gene expressions do vary by times (the time-gene interaction terms of $\mu_{time*gene}$ are not all equal to zeros.). The F statistics become.

$$F_{Time \cdot Gene} = \frac{SS_{Time*Gene}/df_{Time*Gene}}{SS_{Error}/df_{Error}}, \quad (5.4)$$

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

6. Results

In this chapter, the results by two different kinds of clustering methods are compared. Firstly, the PSE is considered. The results are plotted and tabulated below.

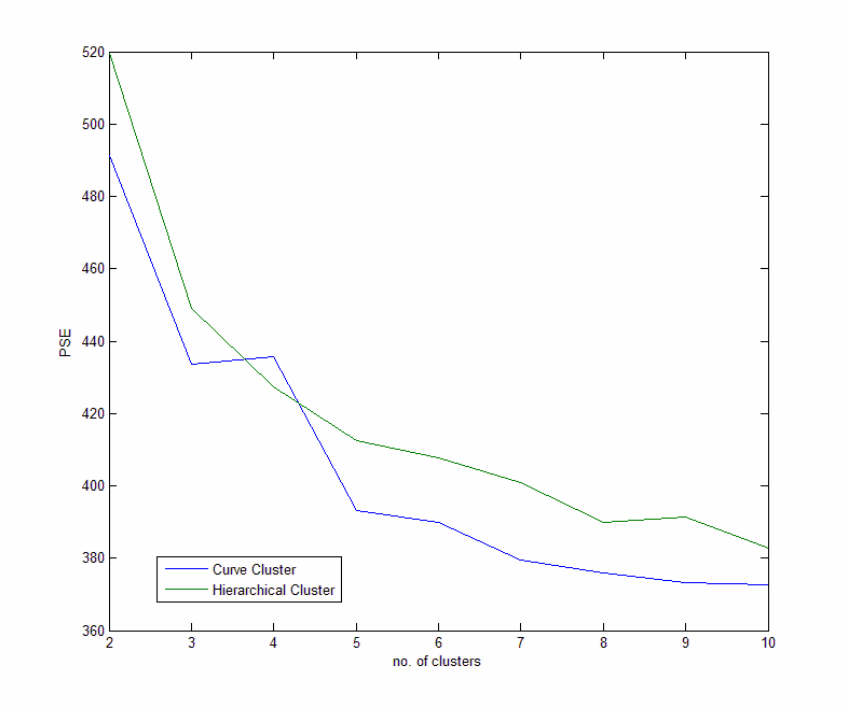
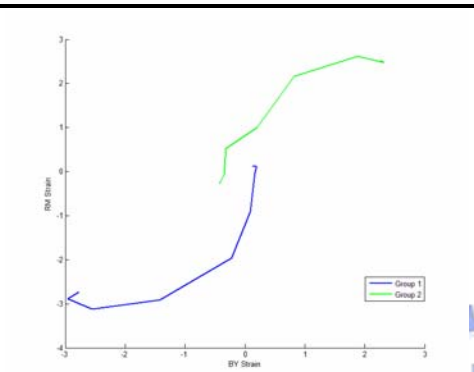
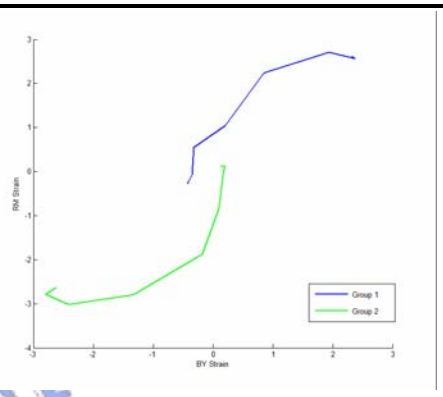
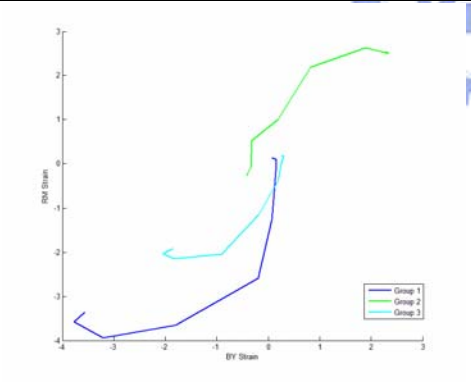
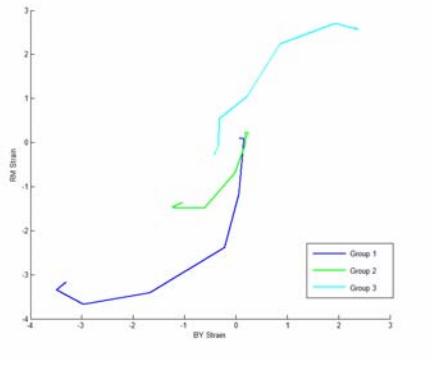


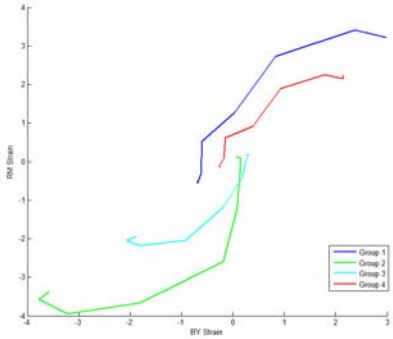
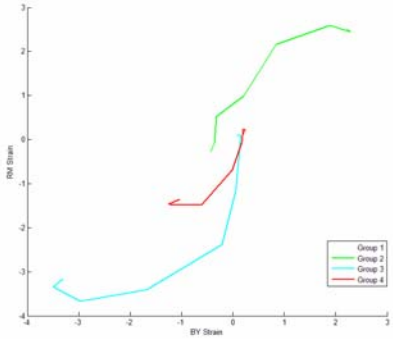
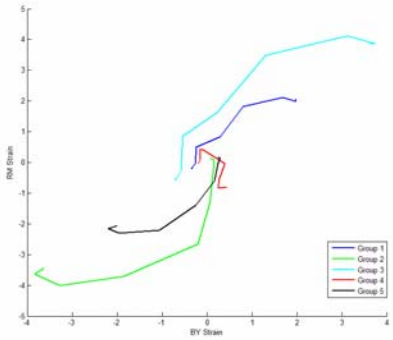
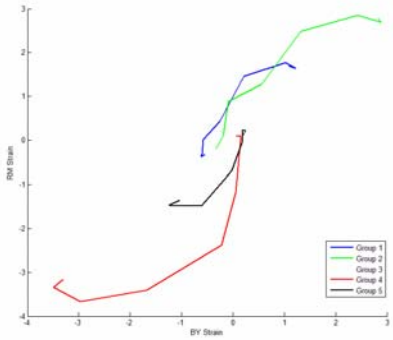
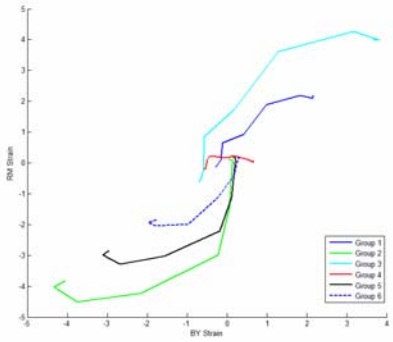
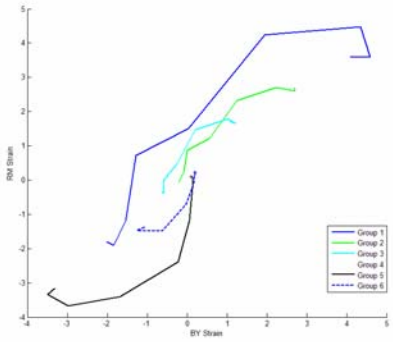
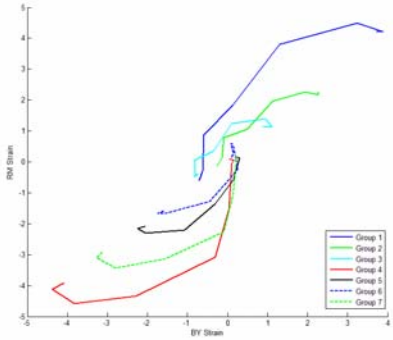
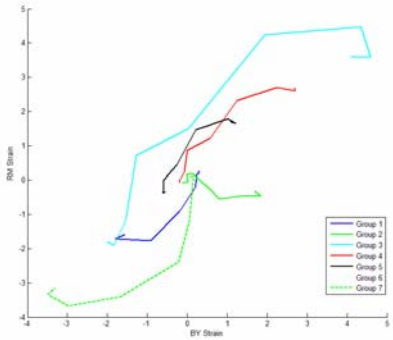
Figure 6.1: PSE comparisons of different number of clusters are shown for two different cludtering methods.

No. of Groups	2	3	4	5	6	7	8	9	10
H.Clust	519.55	449.1	427.49	412.41	407.79	400.81	389.94	391.29	382.67
C.Clust	491.43	433.64	435.6	393.16	389.91	379.63	375.9	373.21	372.52

Table 6.1: the detail values of PSE

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 6.2.

Curve Clustering	Hierarchical Clustering
 Numbers of cluster size = 2	 Numbers of cluster size = 2
 Numbers of cluster size = 3	 Numbers of cluster size = 3

 Numbers of cluster size = 4	 Numbers of cluster size = 4
 Numbers of cluster size = 5	 Numbers of cluster size = 5
 Numbers of cluster size = 6	 Numbers of cluster size = 6
	

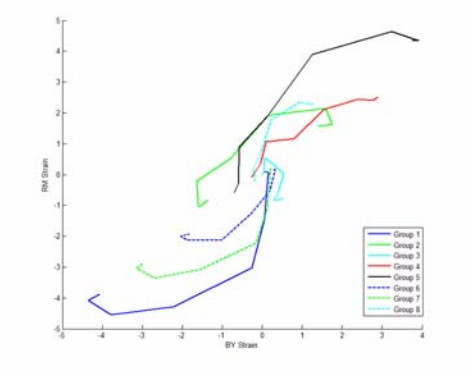
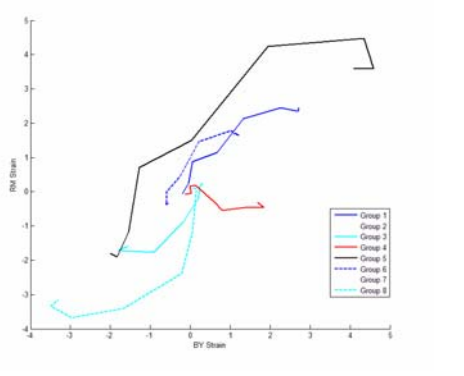
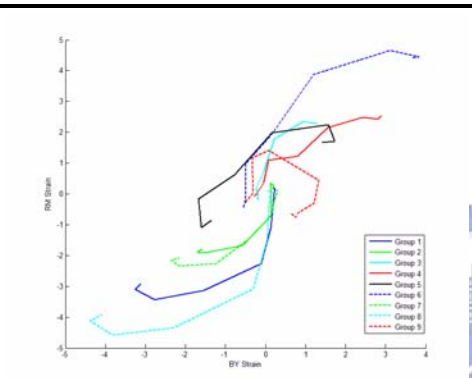
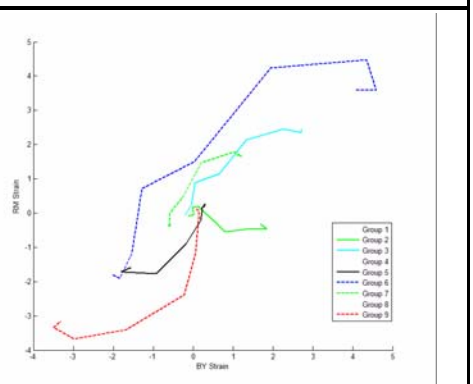
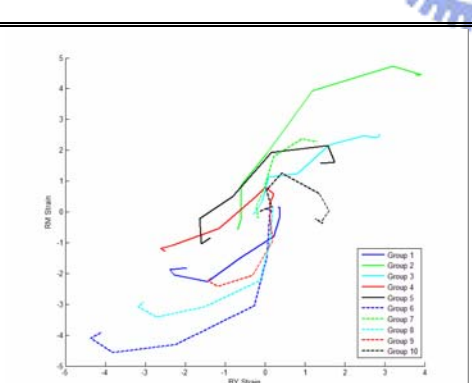
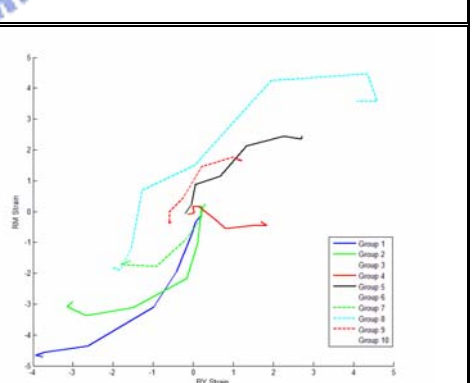
Numbers of cluster size = 7	Numbers of cluster size = 7
	
Numbers of cluster size = 8	Numbers of cluster size = 8
	
Numbers of cluster size = 9	Numbers of cluster size = 9
	
Numbers of cluster size = 10	Numbers of cluster size = 10

Figure 6.2: Mean curves of every group are shown for two clustering methods with different clustering sizes.

From the above results for two clustering method, there 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 and the known genes in these five clusters are listed in Table 6.2. Expression profiles in group 1, 2, 3 and 5 show similar time trends and positive correlations in two strains. However, consistent genes in group 4 show different time trends and patterns that will be explored below.

Group	Average expression profiles of three experiments for consistency genes in two strains	Known genes in this group:

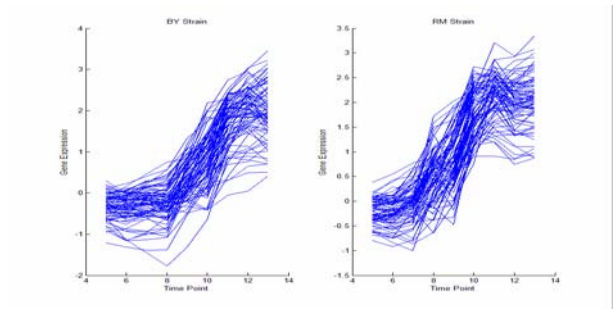
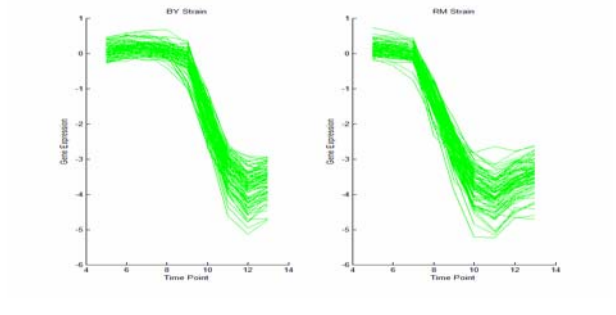
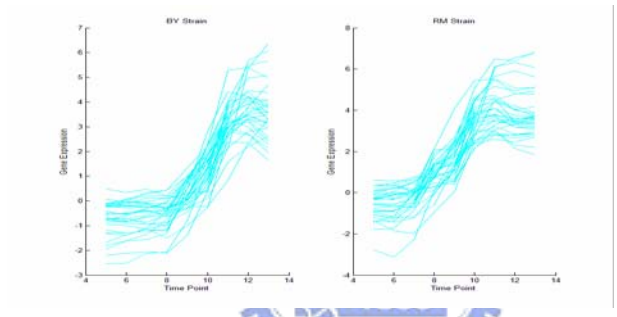
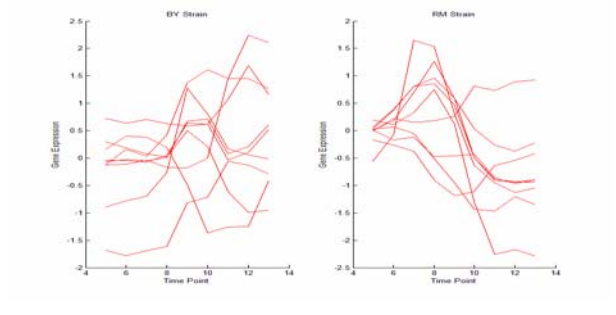
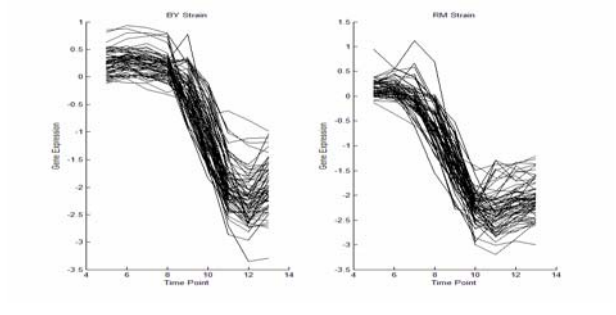
1		ybl045c ykl026c ylr258w
2		yhl015w yil053w ylr029c
3		ygr043c ygl191w ynl117w
4		none of all
5		ymr290c ylr180w

Table 6.2: The clustering results by curve clustering are shown when the number of cluster size is five.

The results of time shifts determined by regression models in these five clusters are listed in Table 6.3. From Table 6.3, gene expressions appear to vary later than glucose consumption do in most groups, except for group 4. Genes in group 4 are interesting because there are negative correlations between gene expressions in BY and RM strains as shown in the mean curves in Figure 6.2 when the cluster size is five. The regression results show that the gene expression profiles in group 4 are inhomogeneous. The time profiles of consistent genes for two strains in group 4 are further investigated in Figure 6.3. From Figure 6.3, 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 4 in Table 6.3 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.

Group	Results of regression models with the most significant effects of glucose assoication among three time shifts are listed

1

Use Model (5.1):

Tests of Between-Subjects Effects

Dependent Variable: Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4697.270[a]	732	6.417	24.260	.000
Intercept	276.954	1	276.954	1047.047	.000
Strain	10.631	1	10.631	40.192	.000
EXP	31.094	2	15.547	58.777	.000
TIME	32.776	8	4.097	15.489	.000
GeneID	247.401	80	3.093	11.691	.000
TIME * GeneID	305.768	640	.478	1.806	.000
GlucoseTSN1	109.097	1	109.097	412.448	.000
Error	963.080	3641	.265		
Total	8476.517	4374			
Corrected Total	5660.350	4373			

a. R Squared = .830 (Adjusted R Squared = .796)

Group Time shift = -1

2

Use Model (5.1):

Tests of Between-Subjects Effects

Dependent Variable: Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14340.742[a]	786	18.245	97.051	.000
Intercept	2111.956	1	2111.956	11234.038	.000
Strain	.589	1	.589	3.132	.077
EXP	74.000	2	37.000	196.813	.000
TIME	369.563	8	46.195	245.725	.000
GeneID	252.918	86	2.941	15.643	.000
TIME * GeneID	169.660	688	.247	1.312	.000
GlucoseTSN1	1044.116	1	1044.116	5553.923	.000
Error	735.253	3911	.188		
Total	28855.247	4698			
Corrected Total	15075.995	4697			

a. R Squared = .951 (Adjusted R Squared = .941)

GroupTime shift = -1

3

Use Model (5.1):

Tests of Between-Subjects Effects

Dependent Variable: Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7447.260[a]	300	24.824	43.301	.000
Intercept	510.573	1	510.573	890.590	.000
Strain	12.108	1	12.108	21.120	.000
EXP	51.409	2	25.704	44.836	.000
TIME	70.787	8	8.848	15.434	.000
GeneID	354.168	32	11.068	19.305	.000
TIME * GeneID	552.746	256	2.159	3.766	.000
GlucoseTSN1	239.956	1	239.956	418.553	.000
Error	849.053	1481	.573		
Total	12023.959	1782			
Corrected Total	8296.314	1781			

a. R Squared = .898 (Adjusted R Squared = .877)

Group Time shift = -1

4

Use Model (5.1):

Tests of Between-Subjects Effects

Dependent Variable: Ratio

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	209.151[a]	84	2.490	3.789	.000
Intercept	32.683	1	32.683	49.735	.000
Strain	46.515	1	46.515	70.783	.000
EXP	21.048	2	10.524	16.015	.000
TIME	51.164	8	6.395	9.732	.000
GeneID	53.823	8	6.728	10.238	.000
TIME * GeneID	82.979	64	1.297	1.973	.000
GlucoseTS1	37.527	1	37.527	57.106	.000
Error	263.516	401	.657		
Total	474.115	486			
Corrected Total	472.667	485			

a. R Squared = .442 (Adjusted R Squared = .326)

Group Time shift = 1

Use Model (5.2):

BY **RM**

Glucose terms are not significant for time shifts of -5, -4, -3, -2, -1, 0, 1 hours!

Tests of Between-Subjects Effects

Dependent Variable: Ratio

Source	ype III Sur of Squares	df	Iean Squar	F	Sig.
Corrected M	174.158[a]	83	2.098	8.077	.000
Intercept	1.318	1	1.318	5.072	.026
EXP	13.338	2	6.669	25.670	.000
TIME	45.341	8	5.668	21.816	.000
GeneID	40.309	8	5.039	19.395	.000
TIME * Ge	61.877	64	.967	3.722	.000
GlucoseTS1	2.760	1	2.760	10.625	.001
Error	41.306	159	.260		
Total	229.299	243			
Corrected T	215.465	242			

a.R Squared = .808 (Adjusted R Squared = .708)

Group Time shift = 1

5 **Use Model (5.1):**

Tests of Between-Subjects Effects

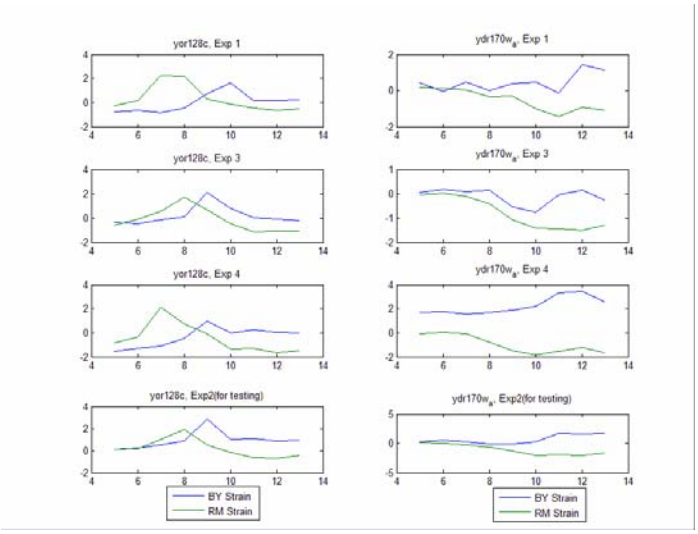
Dependent Variable: Ratio

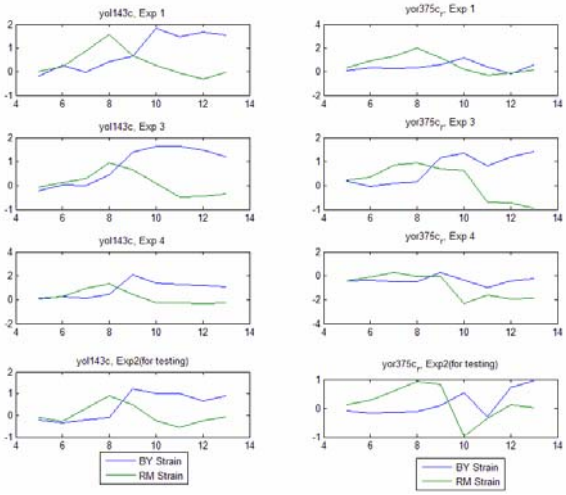
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4150.884[a]	597	6.953	43.012	.000
Intercept	444.144	1	444.144	2747.554	.000
Strain	2.416	1	2.416	14.947	.000
EXP	40.540	2	20.270	125.393	.000
TIME	66.297	8	8.287	51.265	.000
GeneID	104.833	65	1.613	9.977	.000
TIME * GeneID	128.851	520	.248	1.533	.000
GlucoseTSN1	211.632	1	211.632	1309.192	.000
Error	479.456	2966	.162		
Total	7778.110	3564			
Corrected Total	4630.340	3563			

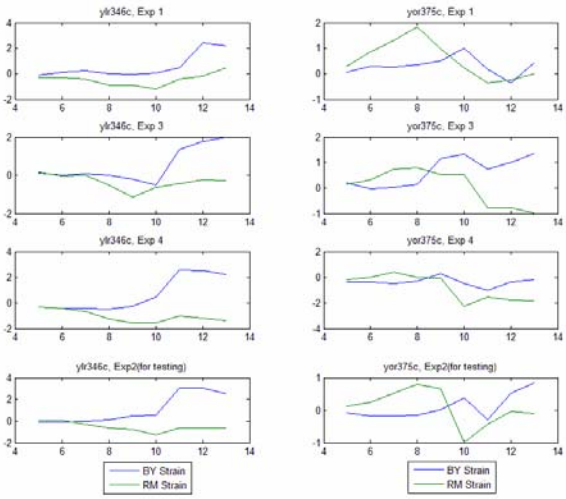
a. R Squared = .896 (Adjusted R Squared = .876)

Group Time Shift = -1

Table 6.3: Results of regression models with the most significant effects of glucose assoication among three time shifts are listed for five groups.


yor128c, Exp 1
ydr170w$_a$, Exp 1
yor128c, Exp 3
ydr170w$_a$, Exp 3
yor128c, Exp 4
ydr170w$_a$, Exp 4
yor128c, Exp2(for testing)
ydr170w$_a$, Exp2(for testing)
BY Strain
RM Strain


yol143c, Exp 1
yor375c$_r$, Exp 1
yol143c, Exp 3
yor375c$_r$, Exp 3
yol143c, Exp 4
yor375c$_r$, Exp 4
yol143c, Exp2(for testing)
yor375c$_r$, Exp2(for testing)
BY Strain
RM Strain


ylr346c, Exp 1
yor375c, Exp 1
ylr346c, Exp 3
yor375c, Exp 3
ylr346c, Exp 4
yor375c, Exp 4
ylr346c, Exp2(for testing)
yor375c, Exp2(for testing)
BY Strain
RM Strain

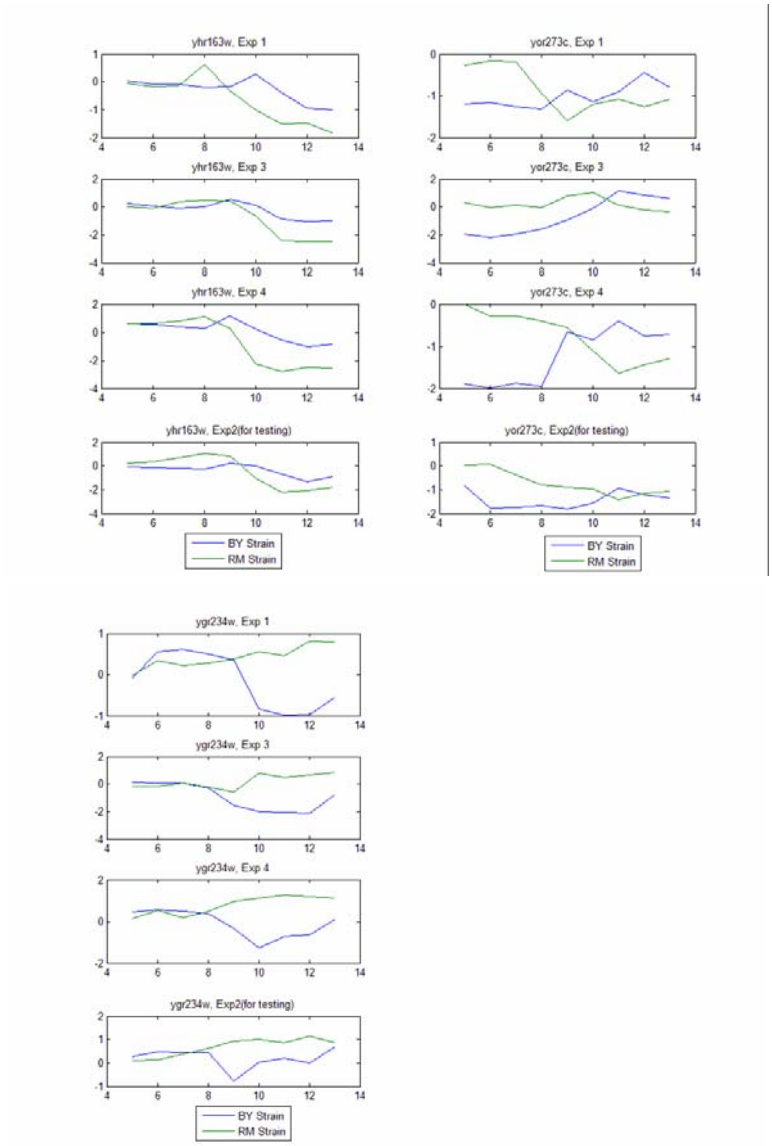


Figure 6.3: Time profiles of genes for two strains in group 4 are shown when the number of cluster size equal to five

The lists of consistent genes in these five groups are reported in Table 6.3. The clustering consistency for all and known genes can be further evaluated by Table 6.5 and 6.6. From Table 6.5 and 6.6, the probabilities of consistent genes in three experiments of the training set among all and known genes are over 56% and 42%

respectively. 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.

Group	Consistent genes included in the group					
1	yil087c	ymr181c	ygl187c	ypl135w	ypl154c	ydl110c
	ykr076w	ymr271c	yhr138c	ypl222w	yhl021c	ykl016c
	ylr270w	yor120w	ylr395c	ydr018c	yll009c	yll020c
	ylr356w	yor100c	yir039c	ydl067c	yml131w	yor136w
	ygl188c	ypr193c	ylr038c	yjl161w	ynl037c	ypl201c
	yil111w	ydl222c	ymr081c	ykl142w	yol083w	ygr174c
	yjl163c	ydl124w	yol077w_a	ylr295c	yor289w	yjl144w
	ymr251w_a	ydl021w	ydr343c	ydl181w	yor285w	ylr080w
	yol152w	ydr530c	ykr049c	yml120c	ypl271w	ypl134c
	ybl045c	ykl026c	ynl237w	yol084w	ypl078c	ydl168w
	ydr513w	ylr164w	ypl123c	ypr006c	ypr149w	ygr194c
	ydr322c_a	ylr258w	yor317w	ypr002w	ypl154c	yjl164c
	yel060c	ycl064c	yer015w	ypl186c	yhl021c	ylr294c
	yhl032c	ydr377w	yml081c_a	ypl087w	yll009c	yor374w

2	ygr148c	yjl191w	yjr145c	yhr203c	ydr418w	ydr025w
	ylr344w	yml024w	ymr242c	yil069c	ygl030w	yfl034c_a
	yml063w	ypl081w	ymr230w	ypl143w	ygr162w	yer131w
	yol120c	ydr064w	yhl015w	ydl083c	yil052c	ykr094c
	yer102w	yjr094w_a	yjl190c	ydl075w	ylr048w	ymr143w
	yfr031c_a	ymr116c	ygl147c	yer117w	ymr098c	yel054c
	yhr021c	ylr367w	ygl135w	ylr101c	yor312c	yer056c_a
	yjr123w	ynl178w	ygl031c	ypl249c_a	ydr450w	ygr214w
	ylr075w	ynl162w	yil053w	ygl123w	yhr010w	ylr029c
	ylr388w	yor063w	ylr061w	ygr034w	yjl136c	ynl302c
	ynl096c	yol127w	yor234c	ygl103w	ymr121c	yor293w
	ymr142c	ypr132w	ypl198w	yjl189w	ynl069c	ypr102c
	ynl301c	yhl001w	ydr447c	yml026c	yol040c	
	yor096w	yil018w	yfr032c_a	ynl209w	yol121c	
	yor369c	yhr141c	ygr085c	ypl079w	ydl082w	
3	ymr105c	yer053c_a_r	ymr107w	yel039c	ynl160w	
	ydr178w	ylr327c	ygr183c	ymr175w	yfl030w	
	ykl217w	ynr002c	ylr366w	ynl117w	yer150w	
	ygl121c	ypr160w	ymr250w	q0080	yer053c_a	

		ygr043c	ygl191w	ynr034w_a	yil160c	ymr256c
		ykl148c	ynl134c	yol052c_a	yml054c	
		yhr001w_a	ylr149c	yer067w	ylr178c	
4				yor128c		
				ydr170w_a		
				yol143c		
				yor375c_r		
				ylr346c		
				yor375c		
				yhr163w		
				yor273c		
				ygr234w		
5	yor272w	ygl076c	ydr324c	yor254c	yol077c	yfl045c
	ygl029w	ygr272c	ydr496c	ypl131w	ypr069c	ykl153w
	ykl081w	yjl158c	yer055c	yhr052w	yhr170w	ynl132w
	ykr059w_r	yjl138c	ygl120c	yjl177w	ylr432w	ypl090c
	ymr075c_a	yfl045c_r	ykl006w	ylr287c_a	ypl043w	yor247w_r
	yor108w	ylr180w	yjr063w	ypr187w	yor340c	ypr190c
	ymr290c	yor344c	yol097c	yhr064c	ypl126w	

	ypl211w	ydr101c	ypl273w	yjr070c	ygr118w	
	yor310c	ykl056c	yhr007c	ylr121c	yil096c	
	ydr087c	ynl110c	ydl229w	ylr406c	yhr216w	
	yer110c	ypl160w	ylr167w	yml022w	ydl192w	

Table 6.4: Consistent genes are reported for five groups.

Max. no. of occurrence in one group among three experiments	3	2	2
No. of Genes	276 (56.56%)	203 (41.60%)	9 (1.84%)

Table 6.5: Degrees of clustering consistency for all genes are tabulated.

Max. no. of occurrence in one group among three experiments	3	2	1
No. of known	11	14	1

genes provided by Dr. Sung	(42.31%)	(53.85%)	(3.85%)

Table 6.6: Degrees of clustering consistency for known genes are tabulated.

7. Conclusion and Discussion

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 fifth cluster shows very interesting pattern of negative 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 before the time that glucose consumption drops.

These consistent genes show negative correlations in two strains are: yor128c, 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. These will be of interest to investigate in future studies.

Reference

1. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278(5338):680-6.
2. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 8;95(25):14863-8.
3. Gaffney, S. (2004). Probabilistic Curve-Aligned Clustering and Prediction with Mixture Models. Ph.D. Dissertation, Department of Computer Science, University of California, Irvine.
4. Gaffney, S. and Smyth, P. (2004). Joint Probabilistic Curve Clustering and Alignment. Advances in Neural Information Processing Systems NY: MIT Press.
5. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell. 2000 Dec;11(12):4241-57.
6. Kerr MK, Churchill GA.. Experimental design for gene expression microarrays. Biostatistics. 2001 Jun;2(2):183-201.
7. Kerr MK and Churchill GA. Statistical design and analysis of gene expression microarray. Genetical Research 2001b; 77:123-128.
8. Kerr MK, Leiter E, Picard L and Churchill GA. Analysis of a designed microarray experiment. Proceedings of the IEEE-Eurasip Nonlinear Signal and Inage Processing Workshop. June 3-6 2001.
9. Kerr MK. Design considerations for efficient and effective microarray studies.Biometrics. 2003 Dec; 59(4):822-8.
10. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol. 2000;7(6):819-37..
11. Park T, Yi SG, Lee S, Lee SY, Yoo DH, Ahn JI, Lee YS, Statistical tests for

identifying differentially expressed genes in time-course microarray experiments. Bioinformatics 2003 Vol. 19 no. 6 2003, pages 694–703

12. Schuller HJ. Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. Curr Genet. 2003 Jun;43(3):139-60.