Statistical Analysis of Microarrays to Determine Genetic Changes in Aging Yeast

A. Wise
J. Hardin, Thesis Advisor
Pomona College

April 1, 2005

Contents

Abstract

The application of statistical tools to microarray analysis can help to advance genetic research. The microarrays of interest in this study measure gene expression levels in seventeen samples from four generations of yeast. The goal of this paper is to help analyze the changes in gene expression levels in aging yeast by the identification of specific genes that are sensitive to aging through the application of various statistical techniques. The two tools described here are analysis of variance (ANOVA) and prediction analysis for microarrays (PAM). ANOVA models the variability and effect of different factors, such as generation. PAM classifies microarray samples based on the gene expression levels and allows us to measure the strength of groups based on gene expression across age categories. Both techniques identify specific genes which may be of interest for biologists to study further.

Chapter 1

Introduction

Microarray technology allows for the expression levels of thousands of genes in a cell to be measured simultaneously. This technology provides great potential in the fields of biology and medicine as the analysis of the data obtained from the microarray experiments provide insight into the roles of specific genes. This paper focuses on the analysis of data from a yeast DNA microarray experiment. In order for meaningful results to be obtained, numerous analytical techniques were applied to this data. Microarrays provide an exciting means through which to explore different statistical techniques.

The biological question that motivates the research of these data is how yeast genes change over time. Therefore, our primary interest in the analysis of yeast gene expression is to further uncover the quantitative relationship between the gene expression levels, both individually and as a whole, and the generation (age) of the yeast cell. The analysis will be performed using two different techniques, analysis of variance (ANOVA) and prediction analysis for microarrays (PAM). Through the use of these two statistical techniques, we will uncover genetic patterns in aging yeast. Through the similarities and differences in the results from and mathematics behind the various techniques, relationships between the techniques and their results can be inferred.

Analysis of variance (ANOVA) is a method that studies the variation in the data. With ANOVA we can create a linear model of the gene expression using generation and additional factors, including dye, which exist in our data. The resulting analysis identifies the specific genes for which generation had a significant effect given that we accounted for other forms of variability such as array and dye color.

Another method for analyzing the differences between generations is prediction analysis of microarrays (PAM). PAM is a clustering tool that provides insight into the robustness of different groupings of the generations of yeast [7]. PAM isolates and identifies specific genes using a threshold value and creates a model to predict the generation for a sample array.

Chapter 2

Background Biology and Data

The yeast data was collected by Professor Hoopes of the biology department at Pomona College. Yeast are single cellular organisms, so for a given yeast all the genetic information is isolated to one cell. To study the genetic effects of aging in yeast, yeast were obtained from generations 1.5, 4, 8, and 12. The genes from these cells were then analyzed through 17 array experiments (see Table 2.1). Generation 1.5 is a mixture of cells from the first and second generation yeast cells, which is a consequence of the challenge for biologists to isolate cells from the first generation.

For each of the microarray experiments, 'spotted arrays' were used. On each glass slide (array) there is a single spot which measures the expression levels for a specific gene. The gene expression levels for 6528 'spots' across these four generations of yeast were measured. Of these 'spots', only 6310 are genes of interest, the remaining 218 are either control genes or spots on the array that were empty of genetic information. The measurements of all the expression values were modified with a normalization program used by the biologists before releasing the data. Additional normalization techniques performed on this data will be described in section 3.1.

To measure the expression level of the genes, for each slide two mRNA samples are reverse transcribed into cDNA samples for a specific cell and compared. The first cDNA sample is from the generation of interest (1.5, 4, 8 or 12) and is labeled using a fluorescent dye of one color (often red); the second sample is from the base generation (1.5) and is labeled using a different fluorescent dye (often green). Note that the red and the green dyes will emit different levels of intensities for the same gene expression level. To control for biases in dye intensity due to the differences in the dye color, for certain arrays the labels for the red and the green dyes were switched (see Table 2.1). Throughout the equations in this paper, when we refer to the red(R) and the green(G) dye intensities, we mean those representing the interest and base generation respectively.

On each array, for each spot, the intensity ratio between the generation of interest and the base generation reflects the change in the gene activity for the generation of interest as compared to the baseline generation [8]. For example, consider array 8 in which the genetic information from generation 1.5 was labeled with green dye and the genetic information

Array Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Generation	1.5	1.5	1.5	1.5	1.5	4	4	4	4	8	8	8	8	12	12	12	12
Base Dye Color	R	G	G	G	G	R	G	G	R	G	G	R	R	G	G	R	R

Table 2.1: The generation and base dye color of each of the 17 arrays. If the base dye color is red (R), then the expression level of the generation of interest was measured by the green(G) dye intensity.

from generation 4 was dyed red. For a given spot, the measurement of red dye intensity reflects the amount of that gene present in fourth generation yeast cells. Similarly, for a given gene the measurement of the green dye intensity reflects the amount of that gene in generation 1.5 yeast cells. T a particular spot, if the red dye intensity is greater than the green dye intensity, the expression level for the fourth generation is greater than the 1.5 generation for the gene on that spot. In this example the ratio of the red to the green dye (or generation 4 to generation 1.5) is greater than one and so the gene measured at this spot has become more active in the fourth generation as the yeast has aged.

Chapter 3

Initial Methodology

3.1 Normalization

The microarray procedure is subject to biases both within arrays, for example, the location of the gene on the array and dye intensity for each spot, and between the arrays, for example, the specific slide a cell is on. These biases affect the measured gene expression of this data. Before the data are analyzed, the values obtained from each array should be adjusted to remove biases. We can apply a normalization technique that uses loess smoothing and scaling to minimize systematic location variations in the gene expression levels. Once the spatial location and dye biases have been accounted for, expression levels across the slides and the biological differences between the genes and samples can be better measured [8].

3.1.1 MA-plot

MA-plots are a helpful way to observe the biases described previoulsy and assist in the normalization of the data [8]. To create the plot, the data must first be transformed into values representing ratios(M) and overall intensity(A). We want to modify our data so that our ratios of interest(M) are not dependent on overall intensity(A). The adjustment will help remove the effects of the differences between the red and green dye. For each of the 17 arrays, the red(R) and the green(G) intensity values have been measured for each gene. Therefore, for each gene on each array we set $M_{ij} = log_2(\frac{R_{ij}}{G_{ij}})$ and $A_{ij} = \frac{1}{2}log_2(R_{ij} \cdot G_{ij})$, where $i = 1, 2, \ldots, 17$ is the array number and $j = 1, 2, \ldots, 6528$ is the gene. Note that for samples with dye swawps, $M_{ij} = log_2(\frac{G_{ij}}{R_{ij}})$. Hence, M_{ij} represents the log of the ratio of the dye intensities while A_{ij} represents the average log intensity for a specific gene on a specific array. Because M_{ij} is the log ratio, $M_{ij} = 0$ reflects no change in gene intensity over the generation. In the calculation of A_{ij}, the $(\frac{1}{2})$ term is so A_{ij} will have the same log range as $log_2(R_{ij})$ and $log_2(G_{ij})$,while log-base two is used because it is easier to express ratios in powers of two (twice as big, four time as big,...). Also note that our original data is not lost in these transformations as we can back solve for R_{ij} and G_{ij} from M_{ij} and A_{ij} and obtain equations 3.2 from equations 3.1.

$$M_{ij} = log_2(\frac{R_{ij}}{G_{ij}}) \quad \text{and} \quad A_{ij} = \frac{1}{2} log_2(R_{ij} \cdot G_{ij}) \tag{3.1}$$

$$\Leftrightarrow$$

$$R_{ij} = (2^{2A_{ij}+M_{ij}})^{\frac{1}{2}} \quad \text{and} \quad G_{ij} = (2^{2A_{ij}-M_{ij}})^{\frac{1}{2}} \tag{3.2}$$

After transforming the yeast data, for each array i a scatter plot of the j M_{ij} values is plotted against the corresponding A_{ij} values. In total there are 17 plots with 6528 spots, assuming that no data points are missing. See figure 3.1(a) for an example MA-plot of array 12.

The normalization technique applied to the data expressed in the MA-plots requires multiple steps.

3.1.2 Within print-tip group normalization

When the cDNA array is printed, the spots are broken up into 24 groups based on their location. Each group contains 272 genes (6528/24) and uses a different print tip. The 24 print tips work simultaneously on their groups, each moving in the same order [1]. One method for removing unwanted variability in the data is to perform within print-tip normalization. Within print-tip group normalization removes the intensity dependence of M and the unwanted variability within each print-tip group that results from the location of the print-tip group on the array and the amount of dye the group received, by assuming that all of the genes should be centered at $M = 0$. If the data values were all without the potential errors caused by the aforementioned sources of variability, or "true", then the intensity log ratios M in an MA-plot should be symmetrically distributed around some horizontal line, indicating no ratio dependence on intensity. The line will be centered at $M = 0$ as we assume that most genes do not change from generation to generation, so the the R to G ratio is 1 [8]. Furthermore, genes with "true" non-zero log ratios (genes which have a change in expression level as the yeast aged) should be randomly distributed both throughout the slide and the print-tip groups so that the distribution of these genes is the same between the 24 print-tip groups. Hence, the intensity ratios within each print-tip group should be centered around $M = 0$.

Within print-tip group normalization uses a loess smoothing technique. The specific smoother used on our data was derived by Cleveland and is often referred to as Cleveland's smoother or a locally weighted running-line smoother [2]. The idea behind this smoother is that a regression line will not accurately fit the data over the entire range of X because the data are not linear, however over small intervals of X the data model can be approximated with a regression line. These regression lines are calculated by locally weighting the data at given points, and then connecting these lines to create a smoothing spline. Due to the local weighting, this smoother is robust as it will not be affected by a small percentage of differentially expressed genes that appear as outliers in the MA-plot [8].

The specifics of Cleveland's technique will be described with respect to the calculation of the smoothing spline for a given print-tip group on a given array. We begin with some

notation: let the subscripts i, k and l denote the i^{th} array, k^{th} print-tip group and l^{th} gene in a print-tip group, where i is as described section 3.1.1, $j = 1, 2, \ldots, 24$ and $k = 1, 2, \ldots, 272$. So M_{ikl} denotes the l^{th} log ratio in the k^{th} print-tip group on the i^{th} array, and similarly for A_{ikl} with log intensity. For a specific print-tip group from a given array, we are interested in estimating the expected value of M_{ikl} that corresponds to a specific A_{ikl} value.

First we need to choose in advance the value of f, the proportion of points from a specific print-tip group on a given array used to estimate M_{ik}. f is often referred to as the *span*. If f is too close to zero, then the lines will be too ragged, and if f is too close to 1, a weighted least squares line will result. We used a standard f value of 0.4, and hence the $272 \times 0.4 (\sim 109)$ nearest neighbors. Next, let $A^* = A_{ikl}$ for some l, and we measure the distance of A^* from each of the observed A_{ikl} values so that:

$$\delta_{ijk} = |A_{ikl} - A^*| \tag{3.3}$$

The 109^{th} A_{ikl} values that correspond to smallest 109 δ_{ikl} are the nearest neighbors to A^*. Set the maximum of these 109 δ_{ikl} equal to $\delta_{ik(max)}$. Let $Q_{ikl} = \frac{\delta_{ikl}}{\delta_{ik(max)}}$. The weight, w_{ikl}, of each gene is then calculated as follows:

$$w_{ikl} = \begin{cases} (1 - Q_{ikl}^3)^3 \text{ if } 0 \leq Q_{ikl} < 1 \\ 0 \text{ otherwise} \end{cases} \tag{3.4}$$

Hence, a very close A_{ikl} to A^*, will have a small δ_{ikl} resulting in a small Q_{ikl} and a large weight w_{ikl}. Whereas an A_{ikl} value that is far from A^* will either have a large weight or no weight at all.

Next, weighted least squares are used to predict $\hat{M}_{ikl}$ that corresponds to A^*. Choose b_0 and b_1 that minimize

$$\sum_{l=1}^{272} w_{ikl}(M_{ikl} - b_0 - b_1 A_{ikl})^2$$

are used in our equation. An estimate of $\hat{M}_{ik}$ is then:

$$\hat{M}_{ik.} = b_0 + b_1 A^* \tag{3.5}$$

Note how $\hat{M}_{ikl}$ changes as A^* is altered. We find A^* for all values k, and solve for $\hat{M}_{ikl}$ to get the estimated expected value of M_{ikl}. The smoothing spline, which we will denote $c_{ik}(A)$, is then obtained by the lines connecting the points $(A_{ikl}, \hat{M}_{ikl})$. The loess smoother is applied separately to each of the 24 print-tip groups in each array so that all of the $c_{ik}(A)$ are calculated. For array 12, the 24 loess fits to the MA-plot can be viewed in figure 3.1(b).

We are assuming that for each print-tip group, $M_{ik} = 0$. Therefore, for each gene on each array from the k^{th} print-tip group, a new normalized value for $M_{ikl(new)}$ is calculated by subtracting $c_{ik}(A_{ikl})$from the old M_{ikl} value as shown below

$$M_{ikl(new)} = M_{ikl} - c_{ik}(A_{ikl}) \tag{3.6}$$

Which normalizes each gene within its print-tip group. See the normalized data and the smoothing spline calculated from the new data in figure 3.1(b).

3.1.3 Scaled print-tip normalization

The within print-tip group normalization removed the print-tip and some of the spatial bias as well as the intensity dependence. However, to more thoroughly remove the spatial bias we must scale across all of the print-tip groups, which is accomplished through scaling the print-tip group data. Scaling data that has undergone within print-tip group normalization is called scaled print-tip normalization. The scaling method assumes that the log ratios from each of the k^{th} print-tip group follows a distribution with mean zero and variance s_{ik}^2. Furthermore, as mentioned before, we assume that the print-tip groups would have the same distributions on each of the i^{th} arrays. If we let$s_{ik}^2 = a_{ik}^2\sigma_i^2$, where σ_i^2 is the variance of the log ratios if there is no error within the i^{th} array and a_{ik}^2 is the scale factor for the k^{th} print-tip group, then to scale the data we want all of the s_{ik}^2/a_{ik}^2 so that all print-tip groups on array i have variance σ_i^2.

To estimate a_{ik} the median absolute deviation (MAD) is used. MAD is a robust alternative method to the maximum likelihood estimate for a_{ij},which is affected by outliers [8]. MAD is defined as follows:

$$MAD_{ik} = median_l(|M_{ikl} - \text{median}_{il}(M_{ikl})|) \tag{3.7}$$

where M_{ikl} denotes the l^{th} log ratio in the k^{th} print-tip group on the i^{th} array that has been normalized as described in section 3.1.2.

Next we estimate a_{ik} with $\hat{a}_{ik}$, where MAD is described in equation 3.7.

$$\hat{a}_i = (MAD_{ik})/[(\prod_{k=1}^{24} MAD_{ik})^{1/24}]$$

Once all the $\hat{a}_{ik}$ have been estimated, the print-tip groups on all of the arrays can be scaled. See the normalized data in figure 3.1(c).

3.1.4 Across array normalization

While scaled print-tip normalization has removed the spatial bias and intensity dependence that exist within the slide, the difference in variability that exists between the 17 arrays has not been accounted for. For instance, room temperature when an experiment is performed may affect the dyes and causes variability across the slides. Therefore, the method used to scale the print-tip groups, see section 3.1.3, is applied to all of the arrays to scale the variance between them. Now, we assume that the data from the i^{th} array follows a distribution with mean zero and variance $s_i^2 = \sigma_i^2$, as calculated for each array in section 3.1.3. $\sigma_i^2 = a_i^2\sigma^2$, where σ^2 is the variance of the "true" log ratios if there is no error with or across arrays and a_i^2 is the scale factor for the i^{th} array. We can estimate a_i^2 as described in equation 3.8, and then scale across our arrays.

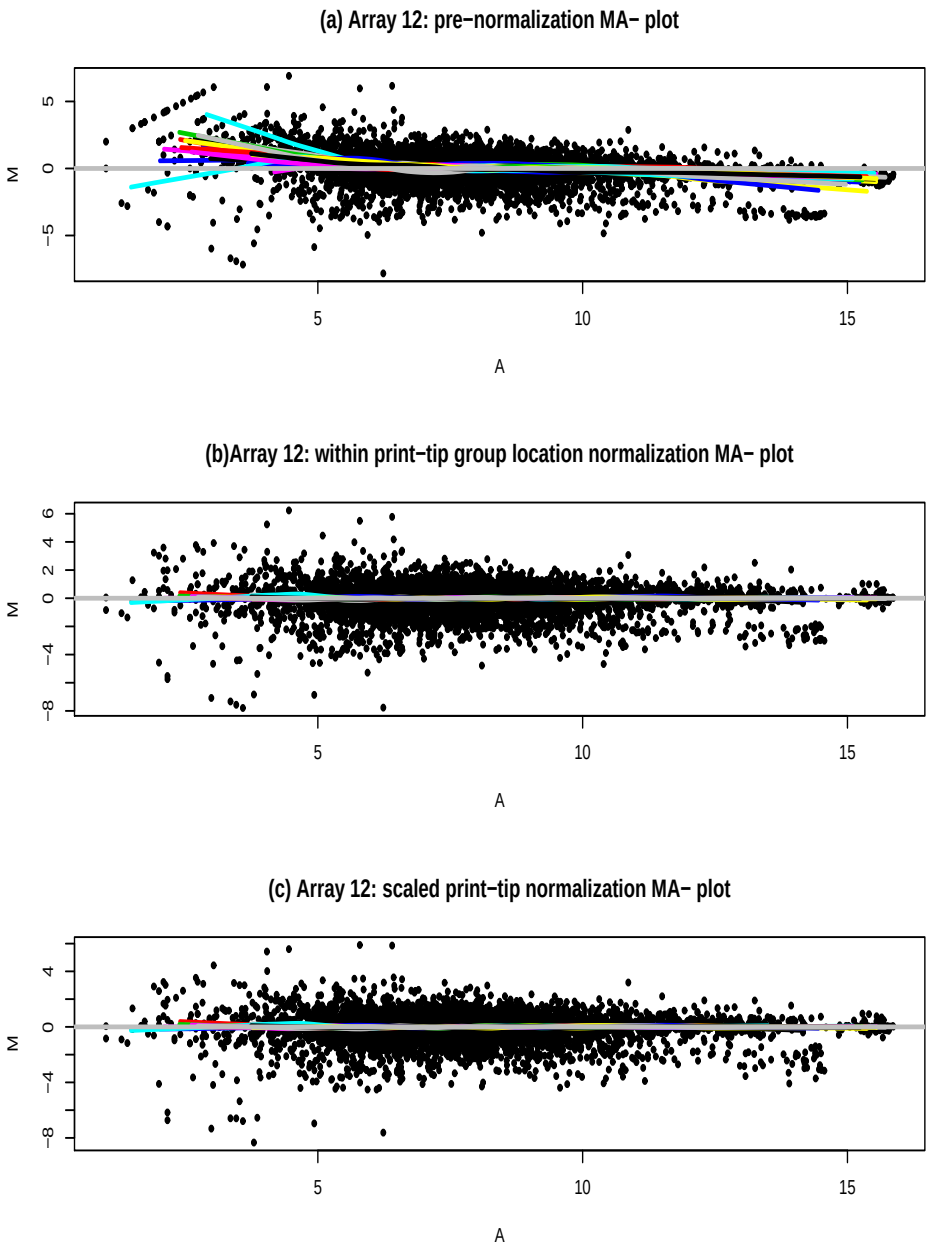

Figure 3.1: MA-plots of array 12, the different lines represent the loess smoothing splines for each print tip group. a)The plot and splines before normalization. b)The plot and splines after within print-tip group location normalization. c)The plot and splines after the data shown in (b) has been scaled across print-tip group, known as scaled print-tip normalization.

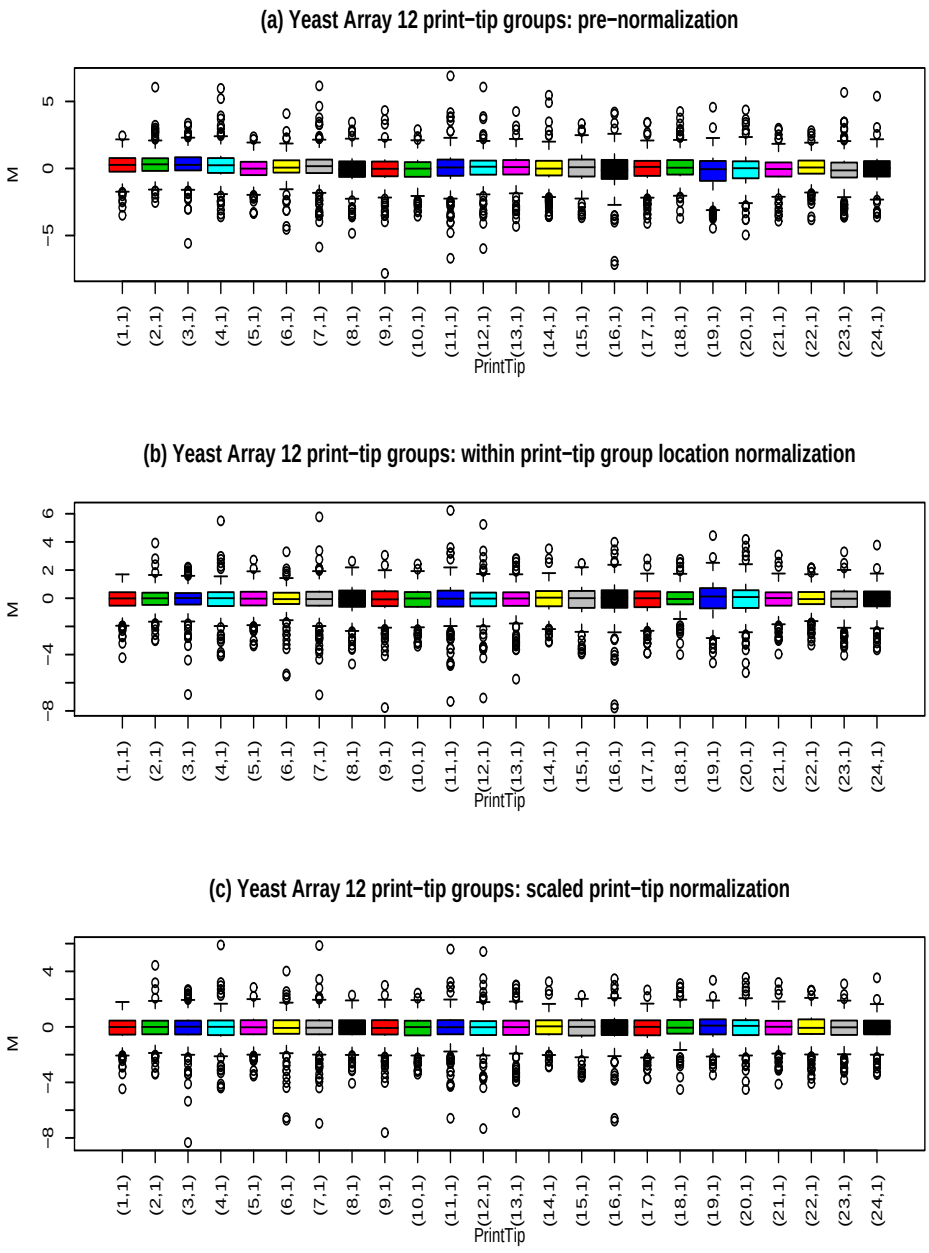

Figure 3.2: Box plots of the 24 print-tip groups from array 12. a)Plots of the print-tip groups before normalization. b)Plots of the print-tip groups after within print-tip group location normalization. c)Plots of the print-tip groups after the data shown in (b) has been scaled across print-tip group, known as scaled print-tip normalization. Notice both the centering and the spread of the box plots.

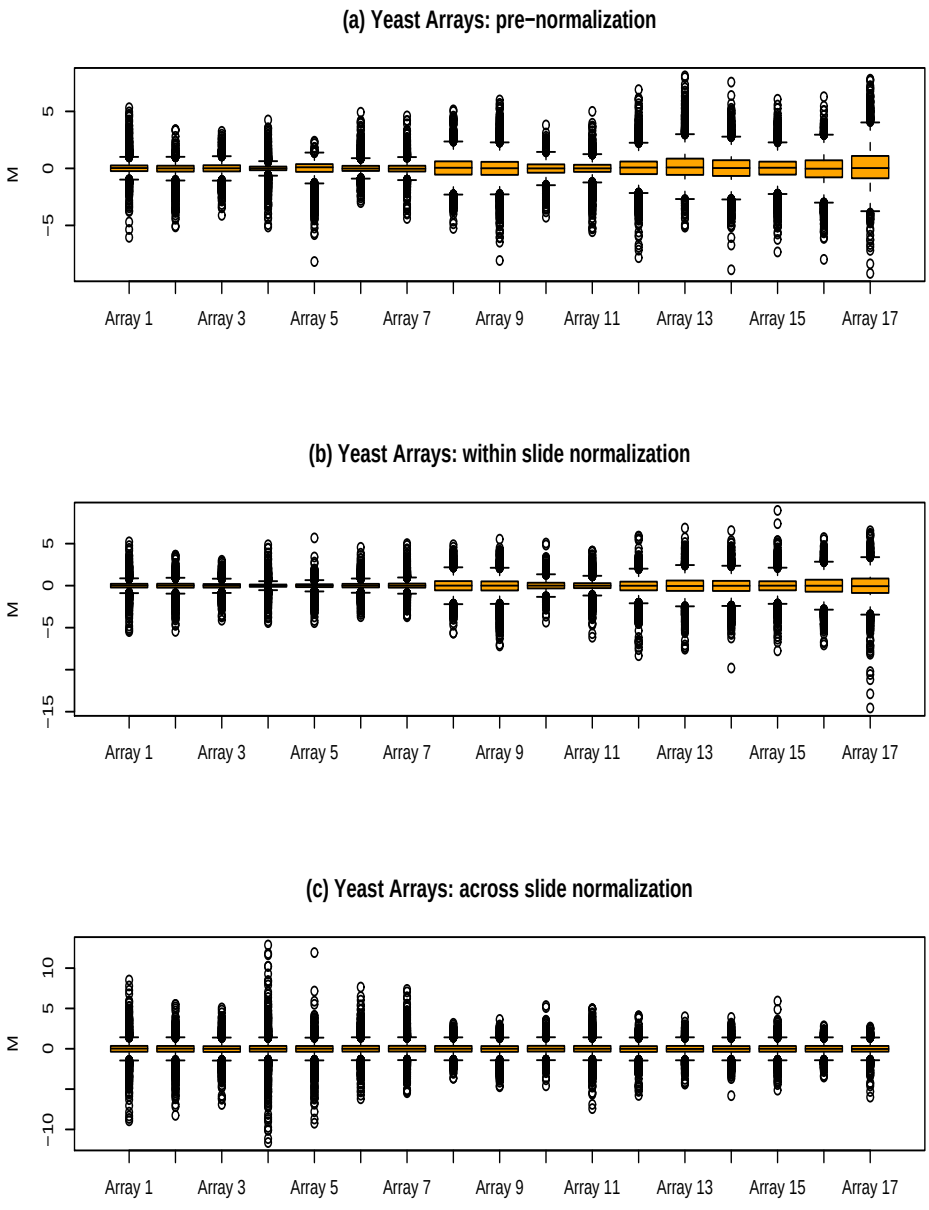

Figure 3.3: Box plots of M for all 17 arrays. a)The plots before normalization. b)The plots of the after scaled print-tip normalization. c)The plots after normalization has been applied across the arrays in addition to the normalization in b). Notice both the centering and the spread of the box plots.

3.1.5 Results of Normalization

The results of the of the different stages of normalization can be seen in figures 3.1, 3.2 and 3.3. The original pre-normalization smoothing splines for the different print tip groups are not always close to zero; whereas, after print-tip normalization, the smoothing splines are much closer to a horizontal line at zero, as depicted by the MA-plots of array 12 (which measures generations 1.5 and 8) in figures 3.1(a) and 3.1(b), respectively. As can be observed from comparing the boxplots of the print-tips groups of array 12 in figure 3.2(a) to 3.2(b), the print-tip normalization centers the data around zero. However, from figure 3.2(b) we can also see that the variability between the print-tip groups was not the same. Scaled print-tip normalization remedied this problem as shown in figures 3.2(c) and 3.1(c). Furthermore, as depicted in figure 3.3(b) as compared to figure 3.3(a), the loess smoothing technique consistently normalizes the data within each of the arrays. These results reflect that within a given array, the average log ratio values, M, across the log intensity values, A, are no longer dependent on A. They also reflect the removal of unwanted data biases due to differences in location on array, and print tip group between all of the genes.

After print-tip normalization there still exists biases between the arrays as depicted by the variability between the arrays in figure 3.3(b). The across array normalization adjusted for this variability, see figure 3.3(c). The addition of this normalization reduces the biases between the arrays. We can't remove all the variability across A because the print-tip runs all along A. The lower A will naturally have more variability because it has smaller raw values which reflect low gene activity and dye intensity at spot. These smaller amounts will be harder to measure and small imprecision measurements will have more of an effect on the ratio value. In addition, there still remains some bias from the differences in dye color, this will be accounted for in the ANOVA application in Chapter 4. Although not all of the bias that exists due to the microarray procedure can be removed, the normalization technique has improved the quality of the data and allows for the natural variability of the data to be better observed.

After normalization, as the biologists are interested in studying only the yeast's genes, the empty and control genes are removed from the data set to isolate the genes of interest.

3.2 Imputing Missing Data

Once the data have been normalized, the missing values can be estimated. The technique we used is one available in R by the PAM package that uses a "k-nearest neighbor" imputation method [3]. Once the data have been imputed using PAM, all techniques can be applied to a full data set.

To understand the "k-nearest neighbor" imputation method, consider all of the M_{ij} data to be placed in a 6310×17 matrix, where the columns represent the array and the rows represent the genes of interest. To begin, we chose the chose k=10, so we will use the 10 nearest neighbors to estimate the missing values. For each row with missing data, if there are too few data points in that row (we used the number of arrays with missing values

for that gene to be > 15),we discard that entire row (gene) as there is not enough data to calculate the 10 nearest neighbors. 301 genes are discarded, resulting in a 6009×17 matrix.

Using the new matrix, for each row with any missing data, the 10 nearest neighbors are calculated using a Euclidean metric based on the columns with pre-existing data. The 10 genes who have coordinate gene values closest to the pre-existing coordinate values of our gene of interest are used as neighbors. If any of the gene rows (not including the one to be filled) are missing some of the coordinates used to calculate the distance, then the average of the distance from the non-missing coordinates is used. Once the 10 neighbors are determined, the missing value of the gene of interest in column (array) i,where $i = 1, 2, ..., 17$, is imputed from the average of the non-missing values from column i in these 10 neighbors. This procedure is preformed for every missing data point, so all of the entries in the matrix are filled.

Unless all of the data across all of the generations for a given gene is present, PAM cannot use any of the data from that gene. There are 3610 genes that have no missing data across all 17 arrays. Without imputation almost half of the genes and the data they contain will be lost. With imputation there are 6009 genes which have a complete data set. Imputing data has a positive affect in that it allows us to work with more data points, and we don't lose existing data from a gene if there are any missing data for that gene across the arrays. However, we are not fully confident in the accuracy of this method and whether the imputed data should be used or not. Therefore, throughout this analysis, the imputed data will be used to explain the techniques and the results, although the non-missing data are also run simultaneously.

Chapter 4

ANOVA

To study the effect that generation has on the intensity of a gene, analysis of variance (ANOVA) will be used. ANOVA is an analytical tool that models sources of variability from different factors. ANOVA analyzes the variability of the factors from the model in the data, compares the variability within these factors to the variability between these factors, and computes the significance of the effects on the model.

4.1 Methodology

As we are interested in how the gene ratio changes over time, we model the M values. There are two sources of variability we feel obliged to deal with in our data: dye color (D) and generation (G). Note that we are not modeling the array factor as we are assuming that our normalization technique has adequately accounted for this variability, and we do not have enough degrees of freedom. Let M_{dgj} be the be the log ratio reading for j^{th} gene of the d^{th} dye color used for the base geneation and g^{th} generation group of which there are n_{dg} values. Note that $j = 1, 2, \ldots, 6009$, d=1, 2 (corresponding to red and green groups, respectively) and $g = 1, 2, 3, 4$ (corresponding to generations 1.5, 4, 8 and 12 respectively) and n_{dg} is the number or arrays which are from both the d^{th} and g^{th} groups. For instance, as shown in table 2.1, for $d = 2$ and $g = 2$, n_{dg} is 2 as there are 2 arrays from generation 4 (the second group) that had a green base dye (the second dye color). For each gene there are 17 M values used as one ratio measurement is taken from each of 17 arrays. Our ANOVA model on M_{dgj} is as follows:

$$M_{dgj} = \mu.. + D_{dj} + G_{gj} + (DG)_{dgj} + \epsilon_{dgj} \tag{4.1}$$

The terms D_{dj} and G_{gj} represent the main effects and account for the overall differences in dyes and generations for one gene, respectively. The term DG_{dgj} accounts for the interaction effect between the dye colors and generations. This term can be understood as measuring, for example, if the red dye amplifies higher at generation 12 then to any other generation. ϵ_{dgj} represents the error measurement and is assumed independent and normally distributed with mean 0 and variance σ^2. μ represents the underlying mean of

a given gene. To calculate the effects, and their significance, of the factors in the model above we solve for the sources of variation between the factors and within the factors. To solve for the variability of a specific factor we first calculate the sum of squares for the main effects, SSD and SSG, the interaction effect $SSDG$ and the error, SSR (where R stands for residuals). For a two-factor ANOVA model with interaction terms, the sums of squares are defined below:

$$SSG_j = \sum_{g=1}^{4}\sum_{d=1}^{2}\sum_{h=1}^{n_{dg}}(\bar{M}_{.g.j} - \bar{M}_{...j})^2 \tag{4.2}$$

$$SSD_j = \sum_{d=1}^{2}\sum_{g=1}^{4}\sum_{h=1}^{n_{dg}}(\bar{M}_{d..j} - \bar{M}_{...j})^2 \tag{4.3}$$

$$SSDG_j = \sum_{d=1}^{2}\sum_{g=1}^{4}\sum_{h=1}^{n_{dg}}(\bar{M}_{dg.j} - \bar{M}_{d..j} - \hat{M}_{.g.j} + \hat{M}_{...j})^2 \tag{4.4}$$

$$SSR_j = \sum_{d=1}^{2}\sum_{g=1}^{4}\sum_{h=1}^{n_{dg}}(M_{dghj} - \hat{M}_{dg.j})^2 \tag{4.5}$$

The mean squares (MS) for each factor is the sum of squares divided by the degrees of freedom, ν, of that factor. For the main effects, the degrees of freedom are the number of groups minus one. Therefore, for the dye color and generation factors, $\nu_D = 1$ and $\nu_G = 3$, respectively. For the interaction effect, the degrees of freedom are $\nu_{DG} = \nu_D \times \nu_G$=3. The degree of freedom for the error is the total number of samples, 17, minus the sum of the other degrees of freedom minus 1, so $\nu_r = 9$. The mean squares are calculated for all of the factors. The calculation for MSR_j for a given gene is shown below:

$$MSR_j = \frac{SSR_j}{9}$$

To measure the effect of a given factor, we are interested in how the variability within that factor compares to the variability across all factors, which is calculated by the ratio of mean squares of a given effect to the mean squares of the residual. A higher ratio reflects a greater variability across the factor, so that the effect of that factor will be greater. The calculation of this ratio for generation for gene j is shown below:

$$\frac{MRG_j}{MSR_j} = \frac{SSG_j/v_G}{SSR_j/v_R} \tag{4.6}$$

If we assume that our data is approximately normally distributed, we know the sums of squares each have an approximate χ^2 distribution with parameter, ν, the degrees of freedom of the variable. In addition, the ratio of 2 independent χ^2 random variables divided by their respective degrees of freedom has an F-distribution with the 2 degrees of freedom as its two parameters. Therefore, as demonstrated in equation 4.6, the distribution of the mean

Source of Variation	df	Sum of Squares	Mean Square	F-statistic	p-value
Generation	3	13.4248	4.4761	12.1069	0.001640
Dye Color	1	1.7261	1.7261	4.6688	0.05899758
Interaction	3	3.2077	1.0692	2.8920	0.094563
Residual	9	3.3274	0.3697		

Table 4.1: ANOVA output for gene HSP26. All three effects are significant at an $\alpha = 0.1$ level, and generation has a significant effect on the model at an $\alpha = 0.01$

p-value	Gene Name
4.38E-06	GLY1
1.53E-05	ENT3
2.80E-05	YGL053W
5.17E-05	NMD4
7.97E-05	YDL173W
8.30E-05	ZDS2
8.32E-05	ADE5,7
9.44E-05	SMY1

Table 4.2: The p-values of the 8 genes on which generation had the most significant effect. All of the effects are significant at an $\alpha = 0.0001$ level.

squares ratio is approximated by the F-distribution. Therefore, we use the F-distribution to calculate the probability, or p-value, of seeing a particular mean square ratio (which we will can an F-statistic) or more extreme based on chance for each of the main and interaction effects on each gene. We use R to compute all values for each of the genes.

4.2 Results

For each gene an ANOVA table was created as illustrated in table 4.1. The p-values associated with generation were collected and ordered for all of the genes. P-values identify the specific genes for which generation had a significant effect given that dye and generation-dye interaction were accounted for. These p-values measure the probability of seeing our data or more extreme if the generation effect measured were in fact due to chance rather than an actual effect. This is referred to as a type I error, and α is the probability of a type I arrising from chance. A threshold α value should be chosen beforehand to determine the size of the p-values ($p \leq \alpha$) for which we will reject that an effect is due to chance. The genes for which generation had the most significant effect in the model are given in table 4.2. These genes, as well as other genes for which the p-value $\leq \alpha$ should be studied more closely as they are more likely to provide insight into the effects generation (or aging) has on gene activity.

Chapter 5

PAM

Prediction analysis for microarrays(PAM) is a clustering technique used for class, such as generation, prediction. PAM uses gene expression data to calculate the shrunken centroid for each class and then predicts which class an unknown sample would fall into based on the nearest shrunken centroid. Through this process, PAM also identifies the specific genes that most determine the centroid. PAM also performs soft-thresholding, a technique that eliminates excess noise from genes that do not vary across the classes.

5.1 Methodology

5.1.1 Centroid shrinkage

The PAM procedure is thoroughly explained in the PAM *Users guide and manual*, [4]. As we are interested in studying the generation of yeast, our data will be broken into four classes,g, as descibed in section 4.1. Let x_{ij} be the expression for the j^{th} gene from the i^{th} array. In addition, let C_g be the indices of the samples in class g of which there are n_g. The j^{th} component of the centroid for generation class a is the mean expression value in class g for gene j and can be written as:

$$\bar{x}_j g = \sum_{i \in C_g} x_{ij}/n_g \tag{5.1}$$

The j^{th} component of the overall centroid across the generations can similarly be expressed as the mean expression value of gene j across all generation. A t-test-like statistic for gene j that compares its expression in generation g to the other generations is denoted by d_{jg}.

$$d_{jg} = \frac{\bar{x}_{jg} - \bar{x}_i}{m_g \cdot s_j} \tag{5.2}$$

where $m_g = (1/n_g - 1/n)^{1/2}$, which scales the standard error of the denominator, and s_j is the pooled within-class standard deviation.

To shrink the class centroids towards the overall centroids, PAM uses soft-thresholding. For a given threshold value Δ, the absolute value of each d_{jg} is reduced by Δ to create a new d'_{jg} such that:

$$d'_{jg} = \begin{cases} \text{sign}(d_{jg})(|d_{jg}| - \Delta) \text{ if } |d_{jg}| - \Delta > 1 \\ 0 \text{ if } |d_{jg}| - \Delta \leq 0 \end{cases} \tag{5.3}$$

Combining equations 5.2 and 5.3 we can calculate a new $\bar{x}'_{jg}$ for each gene as shown:

$$\bar{x}'_{jg} = \bar{x}_j + m_g s_j d'_{jg} \tag{5.4}$$

Therefore, if d'_{jg} is shrunk to zero and the centroid component for gene j is $\bar{x}_j$, then gene j will no longer have an effect on the centroid values for the g^{th} generation. If for all generations, d'_{jg} is zero, then gene j is $\bar{x}_j$ for all generations and no longer has an effect on the centroid values for any generation and can be eliminated from the entire classification model. Therefore, for a gene to remain in the model when there is a threshold value, the statistic $T_j = \max_k d_{jg}$ must be greater than zero. This technique will shrink the number of genes used in the class prediction as well as ignore the insubstantial deviations of a gene from the centroid.

A threshold value should be chosen based on the effect it has on the prediction and distribution of classes in the data.

5.1.2 Class prediction

Once a threshold value has been chosen, test samples are classified by their closest shrunken centroid. Consider a sample array from the data, and label the vector of the gene expression levels x^* such that $x^* = (x^*_1, x^*_2, \ldots, x^8_{6009})$.The discriminant score for generation class g is $\delta_g(x^*)$ and it defined as follows:

$$\delta_g(x^*) = \sum_{j=1}^{6009} \frac{x^*_j - \bar{x}'_{jg})^2}{s_j^2} - 2\log\pi_a \tag{5.5}$$

The first term of equation 5.5 is the standardized squared distance of x^* to the shrunken centroid of the a^{th} generation; the second term is a correction term which uses the prior probability π_g, where π_g is the overall proportion of generation class g in the population. Once $\delta_g(x^*)$ has been calculated for all 4 generations, the classification rule applies the minimum of these four values, $\delta_{g(min)}(x^*) = \min_g \delta_g(x^*)$ and sets the generation which corresponds to the $\delta_{a(min)}(x^*)$ equal to $C(x^*)$; the sample, x^*, is then assigned to class $C(x^*)$.

Furthermore the $\delta_a(x^*)$ can be used to help estimate the class probabilities.

$$\hat{p}_a(x^*) = \frac{e^{-\frac{1}{2}\delta_a(x^*)}}{\sum_{l=1}^{4} e^{-\frac{1}{2}\delta_{a(min)}(x^*)}} \tag{5.6}$$

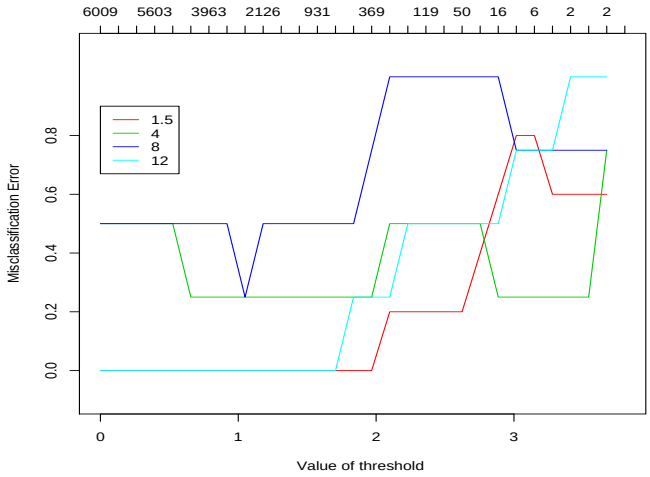

Figure 5.1: Plots of the misclassification error over the various thresholds which correspond to different numbers of genes used in grouping the yeast sample according to its generation. The label across the top axis gives the number of genes used in the classification.

5.2 Results

We can illustrate the results we obtain from PAM using imputed data with a misclassification graph, see figure 5.1. This graph shows four lines, one for each generation. The Y-axis represents the misclassification error, or the proportion of times that the model classified a sample known to be from the generation plotted as another generation. The X-axis shows how the misclassification error of each generation changes as the threshold is raised and the number of genes which remain in the model decreases. As seen from the plots, only generation 1.5 and 12 are accurately calculated at low thresholds, generations 4 and 8 are misclassified over 1/4 of the time no matter what threshold we use. We note similar results when we observe misclassification error at a single threshold through the use of a confusion table. For our example we use a threshold of 1.8, as it appears to be the largest threshold at which the genes still accurately classify some of the generations, see table 5.1. After the threshold increases past 1.8, especially once there are fewer than 369 genes, all four generations are misclassified with uncomfortably high error rates. This plot is rather disheartening as it reflects that hundreds of genes are needed to even come close to accurately classifying a generation. There are not a handful of identifiable genes that are strong indicators of a sample's true generation. This result supports that the centroids and the data should be further studied to better understand, and possibly remedy such bad results. Also, in general it is difficult to accurately describe the differences across 4 groups when our total sample size is so small, only 17.

generation	predicted				class
observed	1.5	4	8	12	error rate
1.5	5	0	0	0	0
4	0	3	0	1	0.25
8	0	0	2	2	.50
12	0	0	0	4	0
Overall error rate = 0.169					

Table 5.1: The confusion matrix for the data at a threshold of 1.8. Fewer than 900 genes made this threshold cut. The values represent the expected generation on the top, and the observed generation on the left. The overall error rate is 0.169.

Chapter 6

Comparison of ANOVA and PAM

6.1 Methodology

To determine if ANOVA and PAM identify similar genes as important with respect to generation, we analyze the overlap between the significant ANOVA genes and the extreme PAM genes. Based on our PAM results, an important gene in this context is a gene which survives a high threshold, not necessarily one that helps accurately predict generation. Both our PAM and ANOVA techniques compared the variability of a given gene within a generation to the variability across generations. PAM performed this comparison to help shrink generation centroids by removing genes that did not identify a generation well, while ANOVA used it to measure the effect generation had on specific genes. The techniques were therefore similar, though not equivalent, as the test-statistic that determines if a gene is altogether eliminated from the PAM model is based on only one generation effect, whereas the ANOVA test-statistic accounts for the effects of all the generations both together and seperately. Still, due to the similarities, it would follow that PAM and ANOVA would identify many of the same genes as having a high threshold and significance level.

To test our hypothesis that PAM and ANOVA should identify similar genes, we calculated the amount of overlap in our PAM and ANOVA genes and compared it with the amount of overlap expected by chance if there was no similarity across methods. To calculate the amount we take the n most significant genes from the ANOVA results and the n genes from the PAM results with the highest threshold values, where $n = 1, 2, ..., 6009$. We then compare these two subsets and count the number of overlapping genes. To calculate the expected number of overlapping genes had the n genes from each subset been allocated at random, a hypergeometric distribution was used. This distribution is appropriate as there are a set number of trials, n, and each trial will be either a "success" or "failure" (overlap or not), and the trials are dependent in that once one element has been removed, the proportion of overlapping genes in the remaining population changes. Let X is a hypergeometric random variable with parameters n and N such that, n is the number of genes randomly chosen from a set of N genes twice and each time placed into first the PAM and then the ANOVA subsets. Then, X represents the number of the n genes that are in both

in both the PAM and ANOVA subsets. Furthermore, it can be shown that the expected value of a this hypergeomtric, and hence the expected value of the overlap is

$$E[X] = \frac{n^2}{N} \tag{6.1}$$

We can then compare this value to the observed value to determine how random our results were. The expected and observed overlap for all values of n were calculated, see figure 6.1.

6.2 Results

The result of the comparison between the expected and observed overlap in ANOVA and PAM genes can be seen in figure 6.1. The plot of the expected overlap, depicts the overlap we would expect to see had we randomly chosen n genes for both PAM and ANOVA. This plot also shows a 95 percent upper confidence bound for the expected number of overlaps, $E[X]$. Our observed data lies far outside of the 95 percent upper confidence interval. Therefore, we are confident that the observed overlap in the data is significantly greater than the overlap in a random sample. This result supports our hypothesis that there are similarities between the math behind each technique such that the resulting "important" genes have non-random overlap. However, in order for this overlap to be meaningful, it would be more reassuring if the values in PAM with a larger threshold did a more accurate job in modeling class prediction.

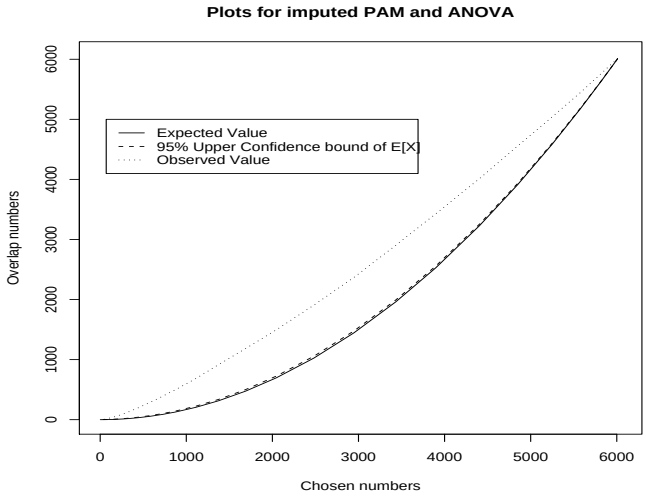

Figure 6.1: A plot of the number, n, of the most significant genes chosen from ANOVA against the number of those genes which were in the n most extreme genes from PAM. The lines shown represent the observed overlap from the data, the expected over if the data were random ($E[X]$), and the 95 percent upper confidence bound for $E[X]$.

Chapter 7

Discussion

The ANOVA and PAM techniques provide useful tools for analyzing the yeast data. Through ANOVA we were able to identify those genes on which generation had a significant effect. PAM allowed us to create a model for predicting the generation of a sample array based on centroid clustering. PAM also used a soft-thresholding technique to shrink centroids and remove genes consisting of only noise. Both the thresholding from PAM and the p-values from ANOVA used similar, though not identical techniques for measuring the variability of the genes across the generations. As shown, the number of genes that had overlap at high threshold values of PAM and low p-values of ANOVA was significant.

However, there are some unsettling results, such as those that PAM produced, as well as results which suggest that the normalization procedure be modified. For instance, raw data with low intensity values should possibly be removed as it is a source of variability that is hard to remove and probably not an accurate reflection of those "true" gene expression levels. The removal of low expression points may help improve the PAM data. The imputation method should also be better assessed to determine how "close" the closest k neighbors really are, and when during the initial methodology this method should be implemented (before or after normalization). To test how well the normalization procedure works, a different ANOVA model that accounts for more factors of variability, such as array and gene, should be run and the results analyzed to see how much normalization ANOVA had to perform after the initial normalization. Furthermore, it might be interesting to go into a more in depth analysis of different analytic tools by applying statistical analysis of microarrays (SAM) to the data and comparing all three results.

Bibliography

[1] Bengtsson, H., Calder, B., Mian, I.S., Callow, M., Rubin, E. and Speed, T.P. (2001) Identifying defferentially expressed genes in cDNA microarray experiments: making aging visible", *Sci Aging Knowledge Environ*, 12.

[2] Cleveland, W. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association,* 74, p. 829-836.

[3] Hatsie, T., Tibshirani, R., Naranimhan, B. and Chu, G. (2002) The pamr Package, users manual to *Pam: prediction analysis for microarrays*, Version 1.24.

[4] Hatsie, T., Tibshirani, R., Naranimhan, B. and Chu, G. PAM: 'Prediction Analysis of Microarrays'" *Users guide and manual*, available at http://www-stat.stanford.edu/ tibs/PAM/pam.pdf

[5] Kerr, K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2002) "Statistical Analysis of a Gene Expression Microarray Experiment with Replication",*Statistica Sinica.*

[6] Ramsey, F. and Schafer, D. (2002) On mixing properties of compact group extensions of hyperbolic systems. *The Statistical Sleuth: A Course in Methods of Data Analysis.*Duxbury, Pacific Grove, CA. Second Edition.

[7] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, 99, p. 6567-6572.

[8] Yang, Y., Dudoit, S., Luu, P., David, L.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation,*Nucleic Acid Research*, 30.