Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DE ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Christina DIONG

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M. Sc.(Mathematics)

GRADE - DEGREE

Department of Mathematics

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

A Method for Linking Microarray Data to Database Information

A. Dabrowski

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

M. Andrade

S. Mills

D. Sankoff

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

A METHOD FOR LINKING MICROARRAY DATA TO DATABASE INFORMATION

By
Christina Diong, B.Sc.
October 2004

A Thesis
submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of
Master of Science in Mathematics[1]

[1]The M.Sc. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Your file Votre référence
ISBN: 0-494-01460-1
Our file Notre référence
ISBN: 0-494-01460-1

Abstract

Microarray technology is a new method of examining the whole genome expression profile. There are over 2,000 manuscripts in total published on microarray data analysis. There are also many database resources for genomic information. Although the human genome is largely known, the degree to which each gene is expressed is not known. Numerous authors have addressed this problem using microarray or database information. Here, we develop an approach that links the gene expression microarray data with the available genomic database information spatially. This provide and alternative method to clustering tools in examining differential gene expression.

Microarray data were pre-processed and normalized so that the data are independent of the technology. Out of 9,600 genes from the original microarray slide, there are 8474 normalized genes for subsequent linkage with the genomic database. The choice of genomic database is `Gene Ontology` (`GO`). The Genomic database information was represented in a 2-dimensional map using Correspondence Analysis (CA). The normalized microarray data were then linked with the genomic database information using response surface methodology.

Acknowledgements

I would like to thank my supervisor, Prof. Andre Dabrowski, for his help and advice. I would also like to thank Dr. Diaz-Mitoma and Dr. Ikhuri Alvarez-Maya of the Children's Hospital of Eastern Ontario for providing us with concrete microarray data sets and applications.

I would like to thank the Ottawa-Carleton Institute of Mathematics and Statistics for funding throughout my research.

Dedication

To my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 What is DNA?

Deoxyribonucleic acid, better known as DNA, is a molecule that encodes genetic information in the nucleus of cells and viruses. It determines the structure, function and behavior of the cell (see [1]). DNA is commonly recognized as two paired chains of chemical bases. The chemical bases in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). These bases are divided into two classes which are the purines (A and G) and pyrimidines (C and T). Two bases form a base pair. Adenine (A) specifically binds to thymine (T) and cytosine (C) binds to guanine (G). These chemical relations are called the Watson-Crick rules. When a base is attached to a sugar, it is called a nucleocide. If a phosphate group attaches to this nucleocide then it becomes a nucleotide. The nucleotide is the basic repeat unit of a DNA strand. Two strands are called complementary if for any base on one strand, the other strand contains the complement to this base. Two complementary single-stranded DNA chains form a stable double helix as shown in Figure 1.

1.2 What is Gene Expression?

Gene expression is a cellular process by which the genetic information flows from a DNA transcript to messenger RNA by an enzyme called RNA polymerase, and

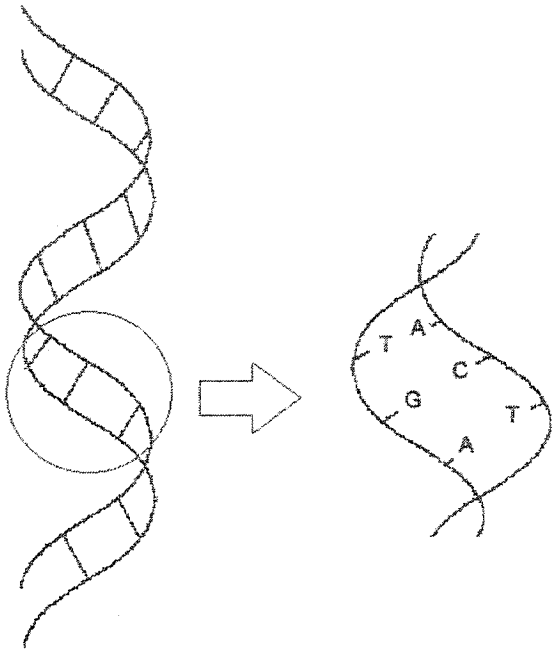


Figure 1: A Double Stranded DNA

which is then translated using a different chemical language (mRNA) to a protein (see Figure 2). Proteins are chains of amino acid molecules. There are 20 amino acids that can be combined to build proteins. If a lot of mRNA (and then protein) is produced, the gene is said to be highly expressed.

Genes make up only a subset of the entire amount of DNA in a cell. Every cell of an individual organism contains the same DNA, carrying the same information. However,a kidney cell is obviously different from a muscle cell for example. This occurs because not all the genes are expressed in the same way in all cells. The differentiation between cells is given by different patterns of gene activations which in turn control the production of proteins.

Today, genomes of many model organisms have been sequenced and in April, 2003, The Human Genome Project announced completion of the DNA reference sequence of *Homo Sapiens*. Although the sequences are completed, the degree to which each gene

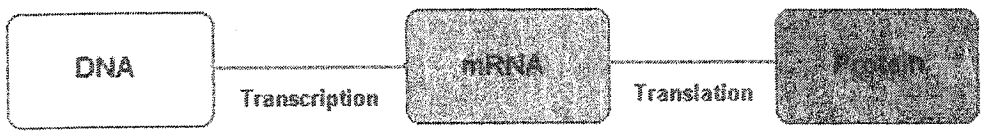


Figure 2: The Flow Of Genetic Information

is expressed in a cell is not known. There are several methods available to look at the gene expression levels for only a few genes at a time. But with over 30,000 genes in a genome, these methods would require decades to address a complete genome. The microarray is one of the methods of examining the whole genome expression profile. It allows the interrogation of thousands of genes at the same time. Microarray is a new scientific word derived from the Greek word *mikro* (small) and the Old French word *arayer* (arranged). It is able to take a snapshot of a whole gene expression pattern in a given sample and compare various samples with each other.

1.3 Microarray Experiment

There are many different types of microarray experiment. One of the most common types is the Nucleic Acid Microarray which has two methods, cDNA Microarrays and Oligonucleotide Microarrays. Another is the *In situ* synthesis in which a different technology used. One of the most common commercial technologies is the Affymetrix Microarray. See [3] for more detailed information on the different types of microarray experiment. We will look at cDNA microarrays in this thesis.

The fundamental basis of microarrays is the process of hybridization (see Figure 3). Hybridization is the process of joining two complimentary strands of DNA or one of DNA and RNA to form a double-stranded molecule.

A DNA array, also known as a microarray slide, is a substrate (nylon, glass or plastic membrane) where single stranded DNA (ssDNA) with various sequences are

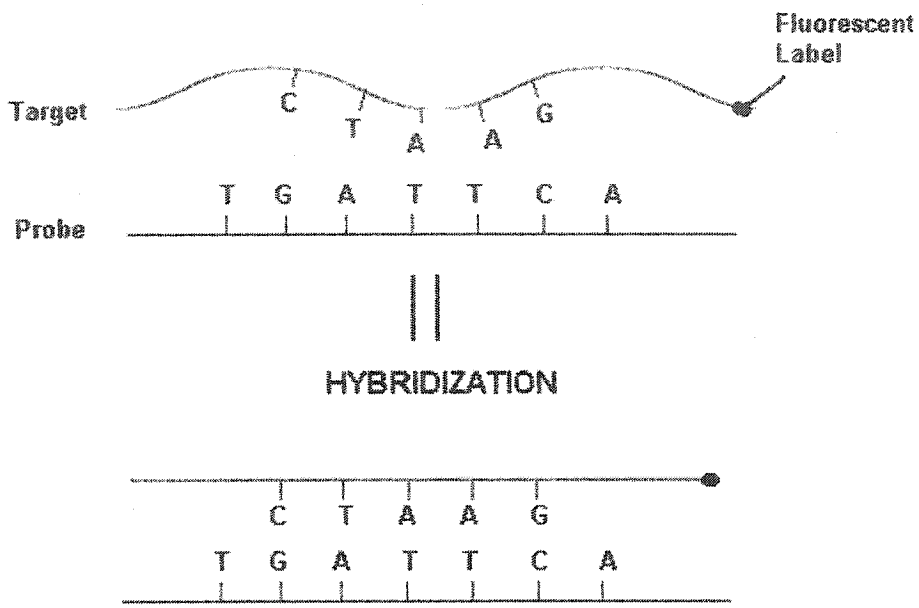


Figure 3: Hybridization. The target ssDNA hybridized to the probe.

printed on the surface of the substrate in a localized features that are arranged in a grid-like pattern as shown in Figure 4. The single stranded DNA printed on the microarray slide is called a **probe**.

The microarray experiment is done by washing the probe with a solution containing another single stranded DNAs from a particular biological sample under study. These sets of DNA are called the **target**. The idea is that the ssDNA in the sample solution which contain two or more channels, the targets, will hybridize to those complementary sequences on the surface of the array, the probe. The target is labelled with a fluorescent dye, radioactive elements, or another method so the hybridization spot can be detected and quantified easily. When using fluorescent dye, the targets are usually labelled using red-fluorescent dye, Cyanine 5 (Cy5) and green-fluorescent

dye, Cyanine 3 (Cy3) for each sample and control. Also see [2] for more information about microarray experiments.

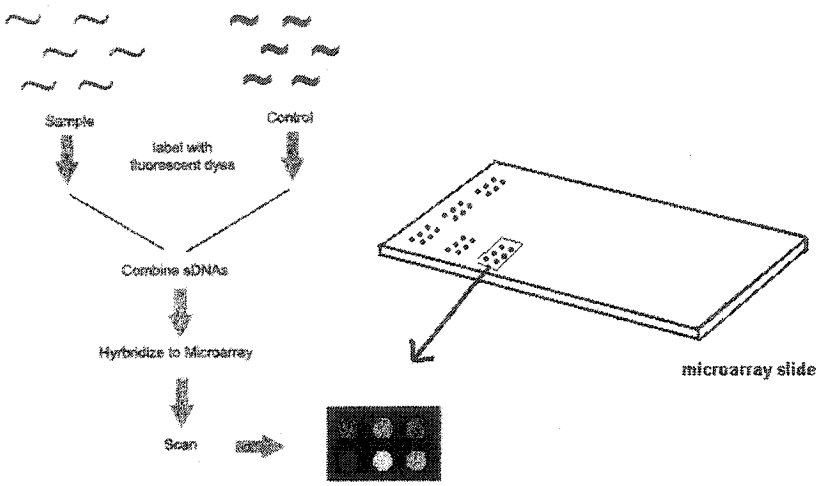


Figure 4: The Microarray Experiment

After hybridization, the intensity of each spot will be measured. The intensities, which are also the average differences between matches and mismatches, are related to the amount of mRNA present in the biological sample and in turn, with the amount of protein produced by the gene corresponding to the given feature. Therefore the main goal in a microarray experiment is to measure the intensity of the spots and quantify the gene expression values based on those intensities. Differential gene expression refers to difference in gene expression across different samples.

1.4 Genomic Database

The root word "genome" is universally defined as the total DNA content of a haploid cell or half the DNA content of a diploid cell. Genomics includes sequencing DNA and collecting genome variations within a population as well as studying the

transcriptional control of genes. For this thesis, a genomic database provides the accumulated biological knowledge on gene function. This may include experimental data, but more usually includes qualitative information on gene function and form drawn from a number of sources. The idea is to collect all the available information from diverse sources into one database location, and to use this database to identify gene properties. There are many different formats and different vocabularies to describe the gene annotations and the principal goal is to find patterns or structures valid across many genes. One way of meeting the challenges is to provide an ontology. This will be discussed in Chapter 3.

1.5 Approach of Thesis

The approach of this thesis is to link the genomic database which contains existing biological knowledge about a gene with the gene expression level as measured by the intensity values from a current microarray data.

Since microarray experiments are subject to considerable experimental and biological variation, we hope to use the prior knowledge of gene function embedded in the ontology database to augment the information from the experiment and so permit a more stable analysis of the data. This approach will be a competitor to the popular clustering method currently used to examine microarray data.

In Chapter 2, we will present image processing and the variables obtained in a microarray data. We will also present methods for microarray data pre-processing and normalization in this chapter. Chapter 3 presents brief information on genomic databases. We will describe the type of database information used here in the thesis and methods for data representation of the genomic database. Chapter 4 presents methods to link the genomic database with the gene expression level from the microarray data. The final chapter presents the conclusions and recommended future work of our same field of study.

Chapter 2

Microarray Data

2.1 The Microarray Data

A typical two-channel or two-color microarray experiment involves two samples such as a tumor sample and a healthy tissue sample. In this thesis, we will use the genetic expression of Peripheral Blood Mononuclear Cells (PBMC) from HIV positive humans compared with genetic expression of PBMC from HIV positive monkeys to illustrate the methods developed here. We thank Dr. Diaz-Mitoma and Dr. Ikuri Alvarez-Maya from the Children's Hospital of Eastern Ontario for supplying this data. We do not claim biological results here, we only illustrate a statistical approach to the data. This data set comprised three experiments. Note that each experiment was divided into two parts where the first part, part(a), is the first half of the 19,000 genes and the second part, part(b), is the second half of the 19,000 genes. We will examine only the first part here.

The available data consist of three slides/experiments where each slide carries two channels, Cy5 and Cy3. The list below shows how the human and monkey samples were assigned to each channel on each slide.

1. Experiment 1: Cy3 Human; Cy5 Monkey;

   Channel 1 = Cy5, Channel 2 = Cy3

2. Experiment 2: Cy3 Monkey; Cy5 Human;

Channel 1 = Cy5, Channel 2 = Cy3

3. Experiment 3: Cy3 Monkey; Cy5 Human;

   Channel 1 = Cy3, Channel 2 = Cy5

As described in Section 1.3, for each experiment, the labelled cDNA were mixed in equal proportions and were hybridized to the probes on the glass slides. The slides are scanned to produce digital images. For each array, the scanning is done in two phases. First, the array is illuminated with the laser light that excites the fluorescent dye corresponding to one channel, for instance in Experiment 1, the red channel corresponds to the monkey sample. An image is captured for this wavelength. Subsequently, the array is illuminated with a laser light having a frequency that excites the fluorescent dye of the green channel corresponding to the human sample. Another image is captured. The red and green intensity of each spot in each image will be proportional to the amount of matching mRNA present in the human and monkey sample. Figure 5 shows an example of a two sample microarray slide.

Let us consider, in Experiment 1, a certain gene spot that is expressed abundantly in the human sample and weakly in the monkey sample (eg spot A in Figure 6). The spot corresponding to this gene will yield an intense spot for the green channel due to abundant mRNA labelled with Cy3 coming from the human sample (left in Figure 6). The same spot will be dark for the red channel since there is little mRNA from this gene in the monkey tissue (right in Figure 6). Superposing the two images will produce a green spot. A gene expressed in the monkey sample and not expressed in the human sample will produce a red spot and a gene expressed in both samples will provide equal amounts of red and green and the spot will appear as yellow. A gene not expressed in either sample will provide a black spot.

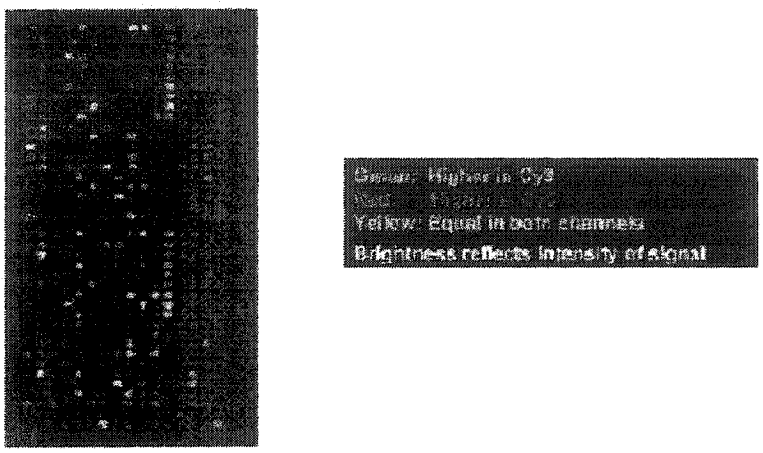


Figure 5: cDNA Microarray Fluorescent Colors for Human and Monkey Data.

The images of the arrays are measured using computer programs. A simple computer program could accomplish the image processing task by superimposing an array of circles with the defined dimensions and spacing on the given image. The pixels falling inside the circles would be considered signal and those outside would be background, here shown in Figure 7. Once the spots have been found, an image segmentation step is necessary in order to decide which pixels from the spot should be considered for the calculation of the signal, which pixels are from the background and which pixels are just noise or artifacts and should be eliminated. This is a challenging issue that is not addressed in this thesis. We will assume that this step has been properly executed.

The PBMC data was obtained using the QuantArray(R) Microarray Analysis Software [4]. The variables obtained for each channel from QuantArray(R) include intensity, background, intensity standard deviation, background standard deviation, diameter, area, footprint, circulation, spot uniformity, background uniformity, signal to noise ratio and confidence. The variables that are of interest in genomic expressions

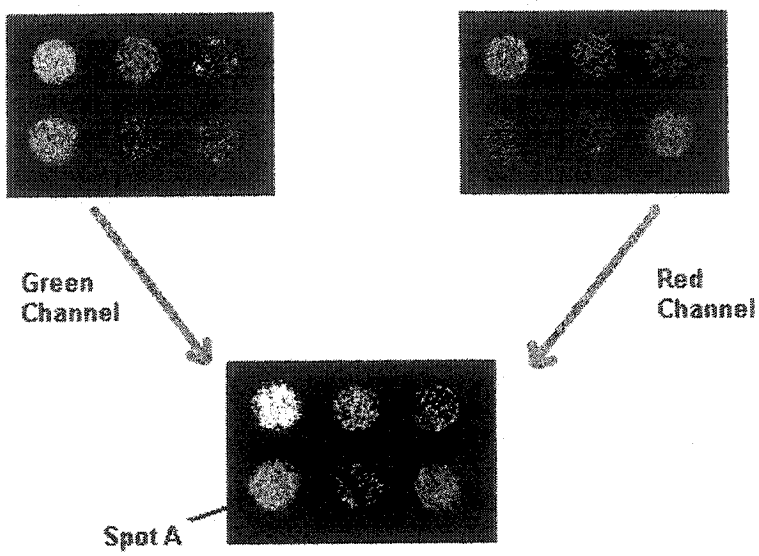


Figure 6: A synthetic image obtained by overlapping two channels. The colors of the images are rendered here in a gray-scale image.

are the channel intensities and backgrounds. In Table 1 is presented the first 2 observations of a sample QuantArray(R) Microarray data set.

Number	Array Row	Array Column	Row
1	1	1	1
2	1	1	1
Number	Column	Name	X Location
1	1	R11726	1540
2	2	R11726	1730
Number	Y Location	ch1 Intensity	ch1 Background
1	2890	559.810791	178.947372
2	2910	536.679016	165.10527
Number	ch1 Intensity Std Dev	ch1 Background Std Dev	ch1 Diameter
1	279.336487	192.149292	155.945114
2	440.182068	235.929993	152.435654
Number	ch1 Area	ch1 Footprint	ch1 Circularity
1	7400	31.489767	0.841288
2	8100	4.033427	0.808437
Number	ch1 Spot Uniformity	ch1 Bkg. Uniformity	ch1 Signal Noise Ratio
1	0.99115	0.994698	2.913416
2	0.987022	0.991875	2.274738
Number	ch1 Confidence	ch2 Intensity	ch2 Background
1	1	1175.157104	376.236847
2	1	1264.101685	359.236847
Number	ch2 Intensity Std Dev	ch2 Background Std Dev	ch2 Diameter
1	773.5448	464.52182	148.62944
2	533.319763	357.530609	138.197678
Number	ch2 Area	ch2 Footprint	ch2 Circularity
1	7000	31.489767	0.805993
2	5900	4.033427	0.75827
Number	ch2 Spot Uniformity	ch2 Bkg. Uniformity	ch2 Signal Noise Ratio
1	0.975418	0.984398	2.529821
2	0.986153	0.991051	3.535646
Number	ch2 Confidence	Ignore Filter	
1	1	1	
2	1	1	

Table 1: QuantArray(R) Microarray Data Sets for two spots

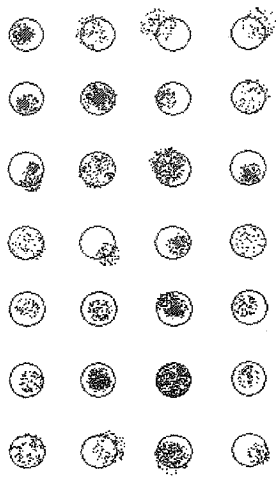


Figure 7: Illustration of image processing challenges.

Microarray data are very noisy. Even if an experiment is performed twice with the same materials and preparations under exactly the same conditions, it is likely that, after the scanning and image processing steps, many genes will probably be characterized by different quantification values (see [6]). Therefore, microarray data needs to be pre-processed and normalized to account for any systematic differences across data sets before being subject to interpretation.

2.2 Data Pre-processing and Normalization

Microarrays data sets need to be pre-processed and normalized so that the data are independent of the particular experiment and technology used. There are many types of microarray data normalization in the literature and there were no standard methods used by all researchers. Normalization method varies with the type of experiment used. Here, data pre-processing and normalization were done following the steps in Figure 8:

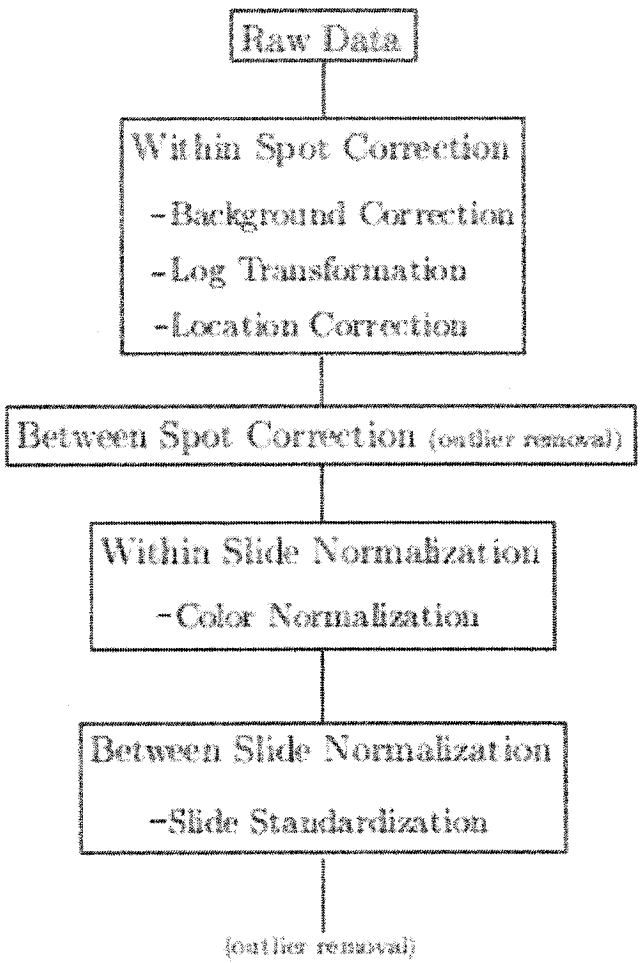


Figure 8: Flow Chart of Data Pre-processing

2.2.1 Within Spot Correction

Background Correction

One of the most commonly used data cleaning processes is background correction. It is calculated as follows:

Background Corrected Intensity Value = Intensity Value - Background Value (1)

The background corrected intensity values represent the contribution of hybridization of labelled target to the probe minus any non-specific hybridization, as well as the natural fluorescence of the probe itself. The signal intensity values should be higher than the background value. However, when the background value is higher than the intensity value, the result would be a negative number which would not be meaningful.

The unusual high background is taken to represent a local problem with the array such as dust and scratches, and so the intensity value is regarded as unreliable. Hence it could be eliminated (see [10]). There are studies suggesting that background intensities must be considered for the normalization of microarray data (see [11],[12]). Many other studies used background corrected values in the analysis of microarray data. Later in this section, we compared both the background corrected and non-background corrected value and concluded that we will proceed with the background corrected value.

Log Transformation

The logarithmic (log) function has been used to pre-process microarray data from the very beginning because of the long-tailed empirical distribution of responses. If one wants to compare the intensities (raw intensities or background corrected intensity values) of an infected sample with a control sample, the log transformation eliminates disproportion between the two relative changes. Log transformation provides values that are more easily interpretable.

In an example shown in Chapter 12, Section 2.1 of [6], if one consider two genes that have background corrected intensity values of 1000 in the control sample and the subsequent measurement of the same two genes in the infected sample are 100 and 10,000 respectively. The absolute difference between the control and infected samples of the two genes are:

$$10000 - 1000 = 9000 >> 1000 - 100 = 900$$

However, from the biological point of view the phenomenon is the same, namely both genes registered a 10-fold change. The only difference between the genes is that the 10-fold change was an increase for one gene and a decrease for the other gene. Applying log transformation, the values transformed into:

$$log_{10}(100) = 2$$

$$log_{10}(1000) = 3$$

and

$$log_{10}(10000) = 4$$

This time, the genes are shown to be vary by:

$$2 - 3 = -1$$

and

$$4 - 3 = 1$$

reflecting the fact that the phenomena affecting the two genes are the same only that they happen in different directions. Another very strong reason for log transformation is related to the shape of the distribution of the values. Figure 9 illustrates the effect of log transformation on the intensity values. Note that intensity values range over a very large interval, from zero to tens of thousands. Hence, the top panel of Figure 9 shows the skewed distribution having a very long tail towards high intensity values. The bottom panel shows the distribution of the same values after log transformation.

Logarithms to base 2 were used here after background correction. The reason is that later in the analysis, the ratio of Cy5 and Cy3 intensities will be used. The ratio is transformed into a difference between the logs of the intensities. Therefore, 2-fold up-regulated genes correspond to a log_2 ratio of +1, and 2-fold down-regulated genes correspond to a log_2 ratio of -1. Genes that are not differentially expressed have a log_2 ratio of 0.

Location Correction

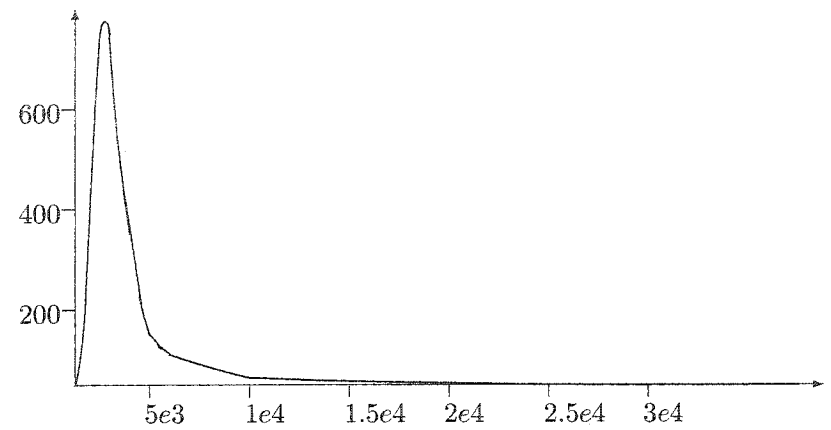
As mentioned before, microarray data are very noisy and are subject to many biases. One of the biases is a spatial bias where the intensity depends on the spatial position of each spot on the microarray (see [13]). This can be seen on a geographical location plot where each plot represents the *logged* background corrected value for the corresponding spot in a microarray slide. Figure 10 shows the geographical plots of *logged* background corrected values for each channel of Experiment 1, part (a). The top panel shows the plot for red channel and the bottom panel shows the plot for green channel. Note from the plots, there is a valley on the left hand side of each slide. It clearly shows that the intensities were not randomly placed on the slide for both channels. These were corrected after location correction.

We introduced location correction for spatial bias along the slide. The steps for location correction are the following, separately for each slide:

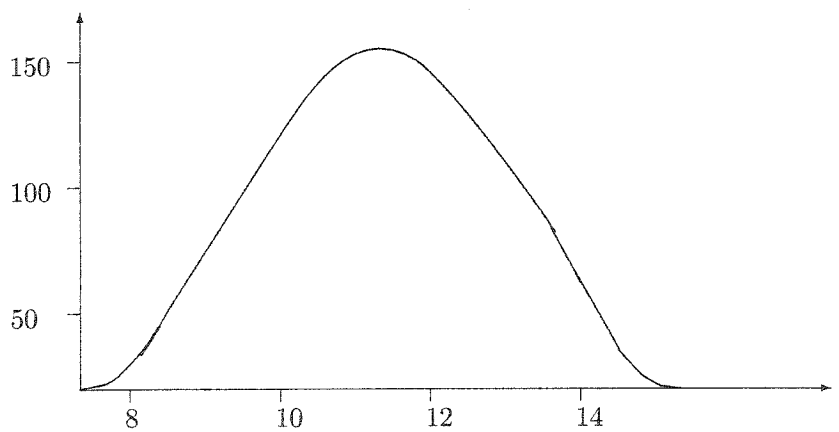
1. Plot the *logged* background corrected value for each spot using X and Y location of the slide as that spot geographic location.

2. Fit a response surface to the intensity data to estimate spatial effects.

3. Compute the residuals from the curve fit, and retain those as data for subsequent analysis.



Distribution of raw intensity values



Distribution of log intensity values

Figure 9: The effect of log transformation to the distribution of intensity values

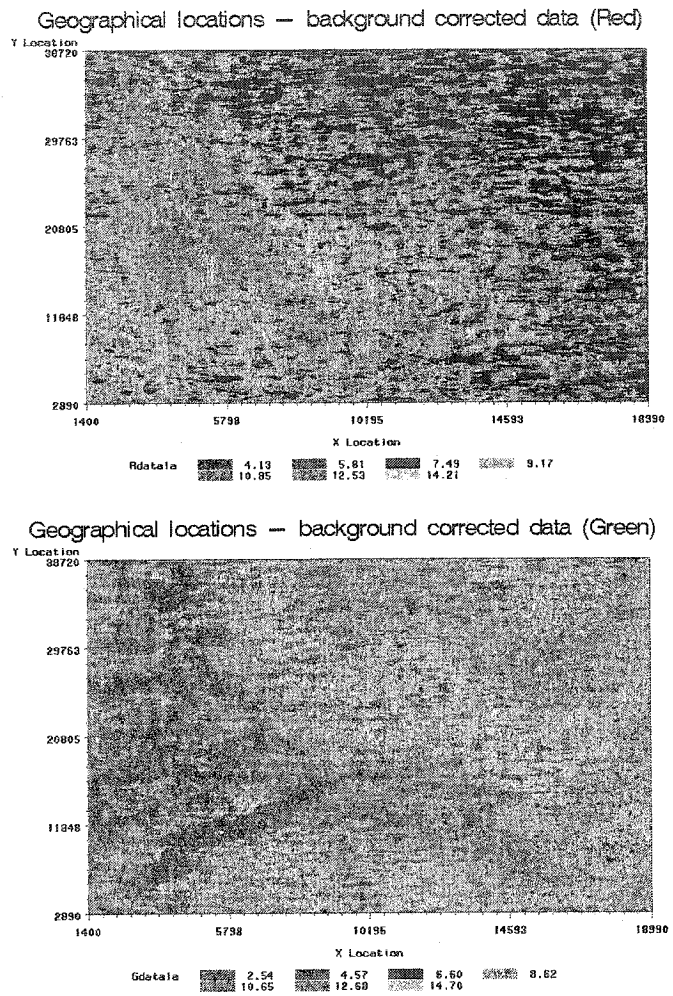


Figure 10: Geographical location of background corrected data for Experiment 1(a). Top panel represents red channel and bottom panel represents green channel.

Response Surface Methodology (RSM)

RSM is a set of techniques that encompasses setting up a series of experiments (designing a set of experiments) that will yield adequate and reliable measurements of the response of interest. It determines a mathematical model that best fits the data collected from the design chosen by conducting appropriate tests of hypothesis concerning the model's parameters (see [14]).

The response variable, y is the measured quantity whose value is assumed to be affected by changing the levels of the factors, x_1, x_2, ..., x_k, that is,

$$y = \phi(x_1, x_2, ..., x_k) + \epsilon \tag{2}$$

If we denote the expected response by $E(y) = \phi(x_1, x_2, ..., x_k) = \eta$, then the surface represented by

$$\eta = \phi(x_1, x_2, ..., x_k) \tag{3}$$

is called a **response surface**.

The response surface is represented graphically by η plotted versus the levels of x_1 and x_2 in the x_1, x_2 plane. In the contour plot, lines of constant response are drawn in the x_1, x_2 plane. Each contour corresponds to a particular height of the response surface.

In RSM problems, the relationship between the responses and the independent variables, factors, are unknown. Therefore, the first step in RSM is to find a suitable approximation for the true functional relationship between y and the set of independent variables and usually a low-order polynomial is employed. If the response is modelled by a linear function of the independent variables, then the approximation function is the first-order model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{4}$$

If there is a curvature in the system, then a polynomial of higher degree will be used, such as the second-order model

$$y = \beta_0 + \Sigma\beta_i x_i + \Sigma\beta_{ii} x_i^2 + ... + \Sigma\Sigma\beta_{ij} x_i x_j + \epsilon \quad (5)$$

The method of least squares is used to estimate the parameters in the approximating polynomials. The response surface analysis is then performed using the fitted surface. Designs for fitting the response surfaces are called response surface designs.

This procedure was done using PROC RSREG from statistical software SAS 9.1, a product that is the registered trademark of SAS Institute Inc. [15]. The Response Surface SAS OUTPUT for the red channel of Experiment 1 is shown below. From the output, we see that the linear and quadratic fit are significant and the lack of fit test (p-value 0.3086) is not significant. This fit is the same for both red and green channels of all the three experiments.

The RSREG Procedure

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
X_Location	10195	8795.000000
Y_Location	20805	17915

Response Surface for Variable Rdata1a: Rdata1a

Response Mean	8.952044
Root MSE	1.143467
R-Square	0.0562
Coefficient of Variation	12.7733

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	2	1283.463079	0.0485	490.80	<.0001
Quadratic	2	197.527930	0.0075	75.54	<.0001
Crossproduct	1	7.179256	0.0003	5.49	0.0191
Total Model	5	1488.170265	0.0562	227.63	<.0001

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	19119	25000	1.307603	2.71	0.3086
Pure Error	2	0.965137	0.482569		
Total Error	19121	25001	1.307517		

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F	Label
X_Location	3	892.432362	297.477454	227.51	<.0001	X Location
Y_Location	3	601.137811	200.379270	153.25	<.0001	Y Location

The responses for location correction here are the *logged* background corrected value. Each channel was modelled separately. The independent variables are the X and Y location of the microarray slides. After being fitted for the response surface curve, the residuals for the curve were computed. These residuals will be used for subsequent analysis; the estimated surface was discarded. The geographical plots after location correction are shown in Figure 11. Note that the prominent streak in the lower left of the bottom graph was not removed by the RSM process. The streak is a polynomial curve out of the range of the rest of the responses on the slide. The estimated response surface was fitted globally. Hence, the streak was not removed in the residuals plot.

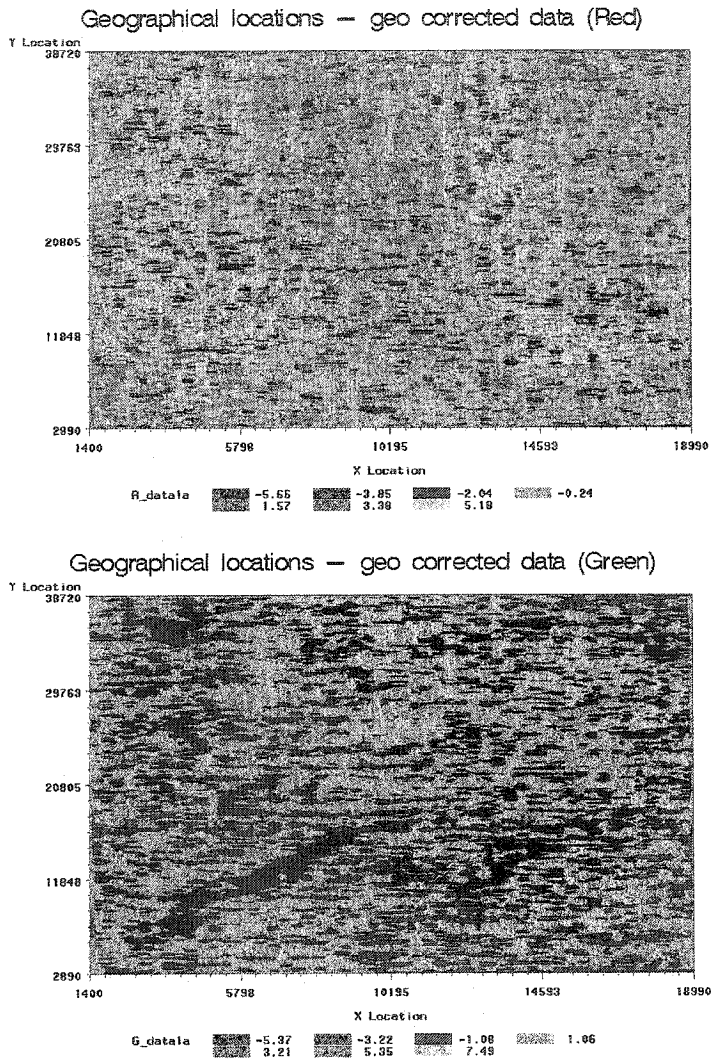


Figure 11: Geographical location of background corrected and location corrected data for Experiment 1(a). Top panel represents red channel and bottom panel represents green channel. Note that the prominent streak in the lower left of the bottom panel was not removed by the RSM process.

Comparison Between Background Corrected Data and Non-background Corrected Data

The background correction method is commonly used in microarray data analysis. But we have questions as to whether to use background correction and location correction method or just the location correction method applied to the raw intensity value. For example, background correction may result a negative intensities inappropriate for log transformation. A comparison of these methods was done in parallel:

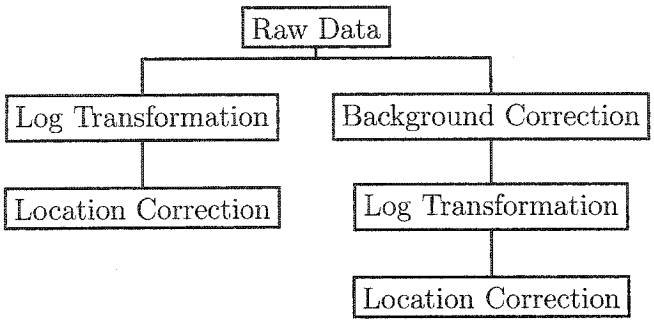


Figure 12: Parallel Comparison of Background Correction with Non-background Correction

The analysis for comparison was divided into two paths, where along the first path, we applied a log transformation to raw intensity values followed by location correction. The other path was background correction, then log transformation followed by location correction. The intensity values for both paths were plotted on a scatter plot of path 1 vs. path 2 for each red and green channel. Figure 13 shows the scatter plots for these comparisons of Experiment 1(a).

Note that for background corrected data sets, negative values were eliminated from the analysis. Hence for this comparison, only the remaining data points will be compared. From the plots, the straight line indicates that both data sets are close. There are several outliers in the plots. After examining the outliers, these are due to a very large background intensity value but not larger than the raw intensity value; the difference is positive and hence they were kept in the data sets. The background corrected value will be used here for subsequent analysis since both paths were close

and many other microarray data analyses in the literature used background corrected data.

2.2.2 Between Spot Correction

Each PBMC microarray data sets from a single slide has multiple observations per gene. The multiple observations are in pairs for each gene spot. In our sample PBMC data, the paired observations are located next to each other on the slide. Some genes have more than one pair in different locations. When we observe the intensity values for the paired observations, some are very different from one another. These are likely due to the basic noise of microarray experiments.

Note that earlier in the background correction section, negative background corrected values were eliminated from the data sets. If, in one pair of genes, one has a positive background corrected value and the other has a negative background corrected value, both genes will be eliminated from the data sets. One could retain this single spots, but the design becomes unbalanced and we would not pursue that direction here. We now do an analysis to compare all the replicated observations and to eliminate observations that are too far apart. The steps to compare these observations are the following:

1. Transform the observations in a Cartesian product match merge format.
2. Plot scatter plots of each Cartesian paired observations.
3. Eliminate outliers of the scatter plot.

Figure 14 shows the Cartesian transformation for the paired genes. After this transformation, scatter plots for each pair were plotted. The plots are shown in Figure 15. Outliers of the plot will be eliminated from the data sets. The process for data elimination are as follows:

1. Compute the differences of the pair (sub.) and sum of the pair.
2. Plot sub. vs. sum.
3. Compute absolute value of the sub.
4. Compute square root of |sub|.
5. Plot histogram of the square root value.
6. Eliminate data points from the extreme tails of the histogram.

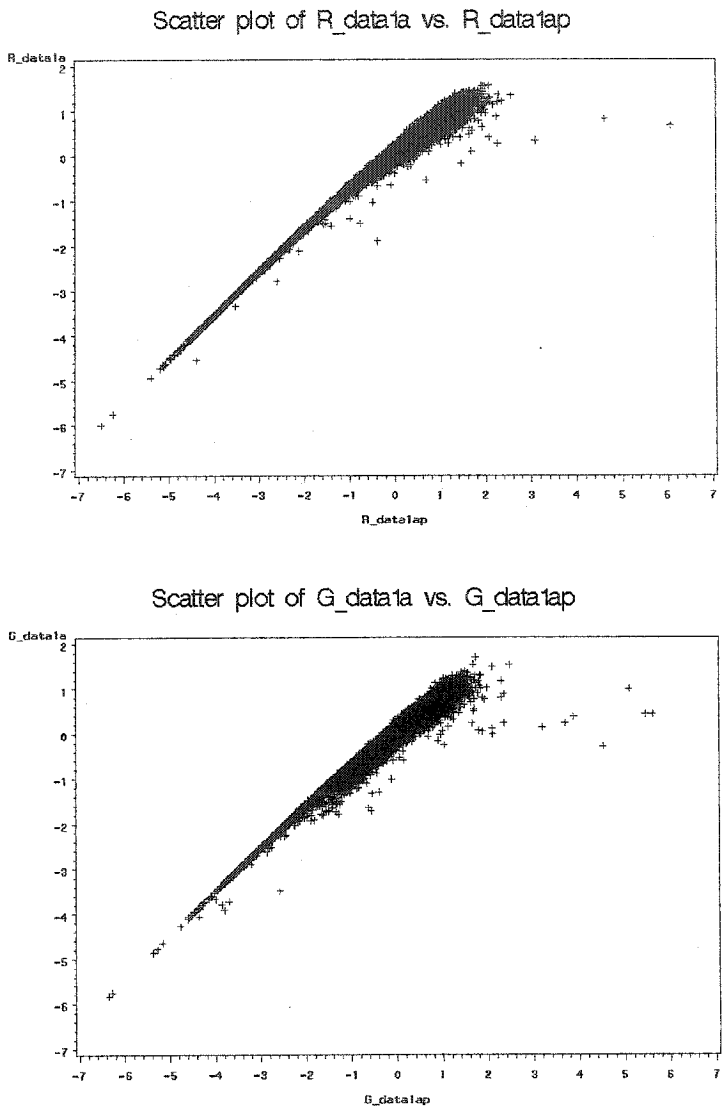


Figure 13: Scatter plot for comparing background corrected data (horizontal axis) with non-background corrected data (vertical axis) of Experiment 1(a). R_data1a and G_data1a represents the red and green channel of non-background corrected data and $R\_data1ap$ and $G\_data1ap$ represents the red and green channel of background corrected data.

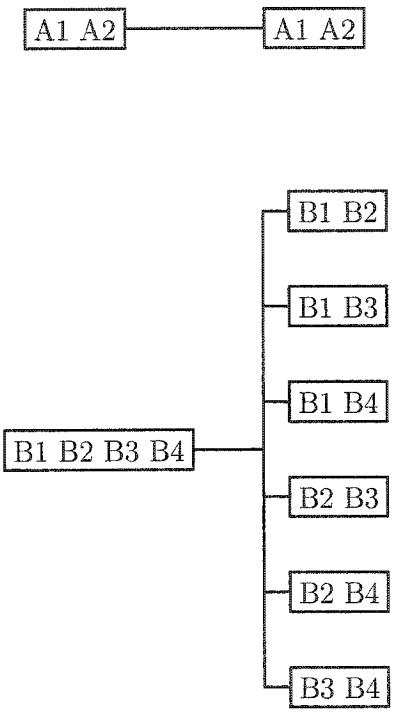


Figure 14: Cartesian Transformation of the set {A1, A2}, and of the set {B1, B2, B3, B4}.

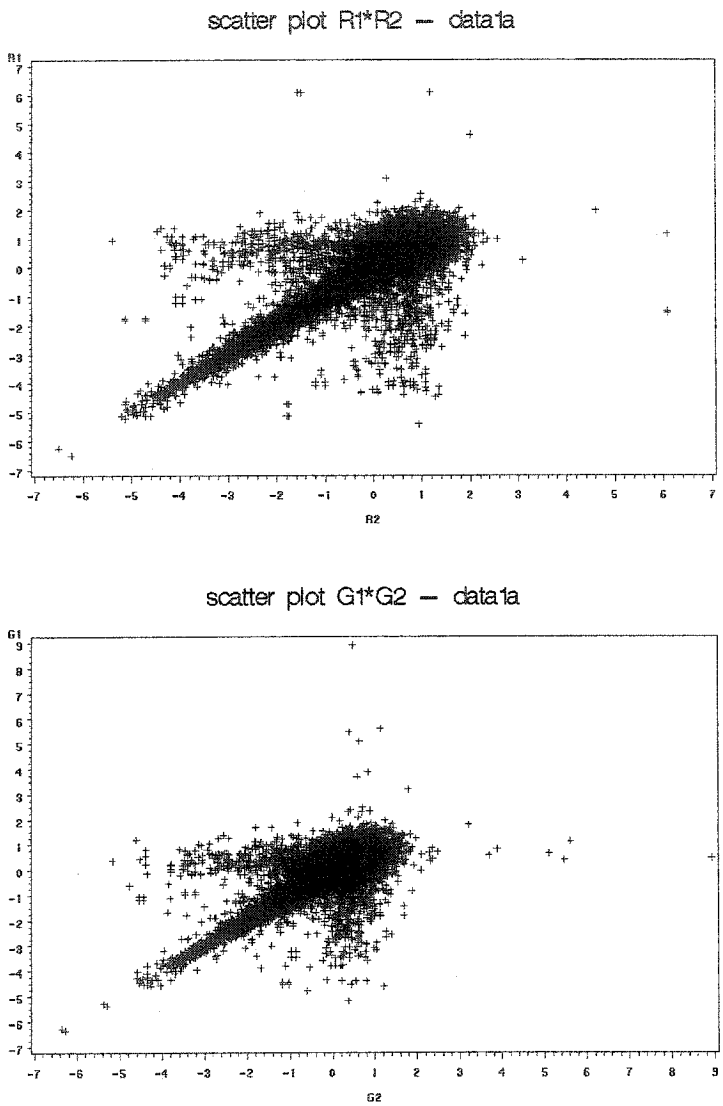


Figure 15: Scatter plot of Cartesian paired genes for Experiment 1(a). Top panel is the plot for red channel and bottom panel is the plot for green channel. Note the plumes of mismatched points off the diagonal.

Figure 16 shows the plot of subtraction between the pairs vs. the sum of the pairs for red channel, experiment 1(a). Figure 17 shows the histogram of the square root of subtracted value. Here, for Experiment 1(a), red channel value, data sets that are larger than 1.17 will be eliminated, removing about 2% of the data points. The scatter plot after outlier removal is shown in Figure 18. This remaining observations will be used for further analysis.

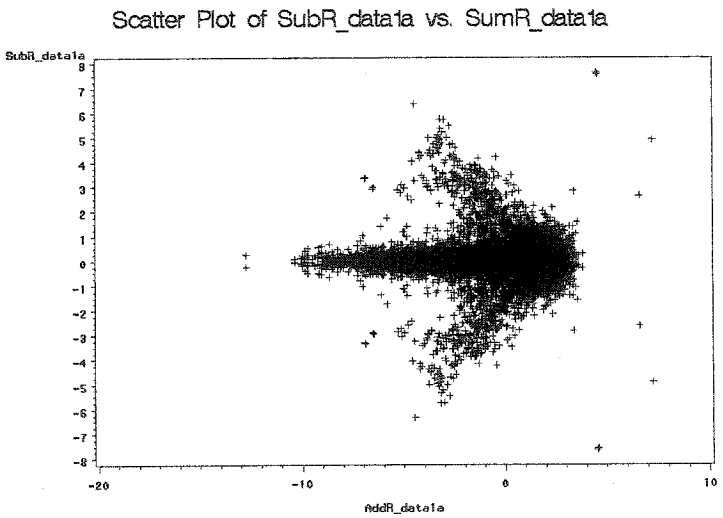


Figure 16: Scatter plot of sub. vs. sum for Cartesian paired genes of red channel for Experiment 1(a)

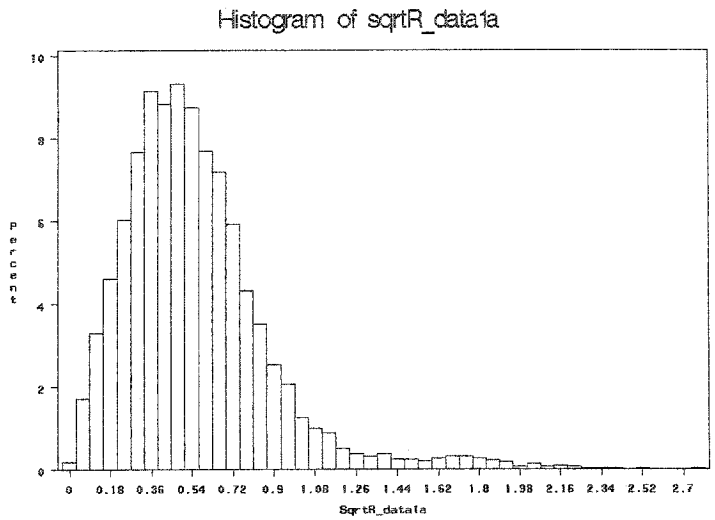


Figure 17: Histogram of square root subtracted value of red channel for Experiment 1(a)

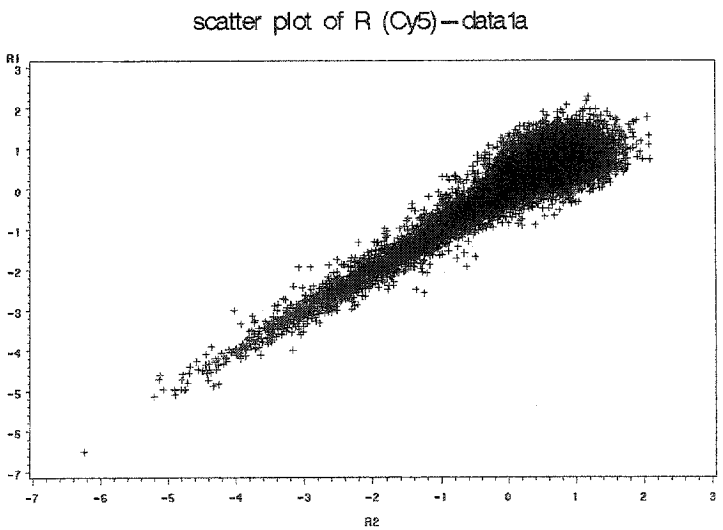


Figure 18: Scatter plot of paired genes for red channel, Experiment 1(a) after between spot outlier removal.

2.2.3 Within Slide Normalization

Within slide normalization is done separately for each slide, using only the red and green intensities for this slide. We have already corrected errors within spots, between spots and also corrected for location biases. These approaches adjust overall problems but do not address dye non-linearity.

We recall that the two samples (human and monkey) were labelled with two different fluorescent dyes in two separate chemical reactions, and their intensity was measured with two different lasers operating at two different wavelengths. In addition, the features on the array are distributed on different parts of the surface of the array. We need to ensure that there is no dye bias or error introduced by the experimental method and the Cy3 and Cy5 intensities are able to be compared on an equal footing. Here, the normalization process is followed by color normalization. There are many methods to correct color distortion, such as 'curve fitting and correction', 'piece-wise linear normalization' and 'LOWESS/LOESS normalization' (see [6]). We will only discuss the LOWESS/LOESS normalization method which will be used as one of our steps in the microarray data cleaning process. Also see [16] for various types of normalization processes of microarray data.

LOWESS/LOESS Normalization

The LOWESS transformation, also known as LOESS, stands for Locally Weighted Polynomial Regression ([17], [18]). The function fitted by LOWESS is a polynomial of the form:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... \tag{6}$$

From this section onwards, dye color Cy3 will be represented by G (green) and Cy5 will be represented by R (red). The starting point for LOWESS normalization is to plot a ratio-intensity plot (R-I plot) of M vs. A.

$$M = log_2 \frac{R}{G} \tag{7}$$

$$A = log_2\sqrt{RG} \tag{8}$$

It has been noted that the log_2(ratio) values often have a systematic dependence on intensity, most often observed as a deviation from 0 for low intensity spots (see [8]). Hence, the $log_2\frac{R}{G}$ ratio should be 0, independent of intensity and it is expected to see no differential expression and consequently all $log_2(ratio)$ measures should be 0 on average. Figure 19 shows the R-I plots for all the part (a) data from Experiments 1(a), 2(a) and 3(a). The plots clearly show a strong non-linear dye distortion.

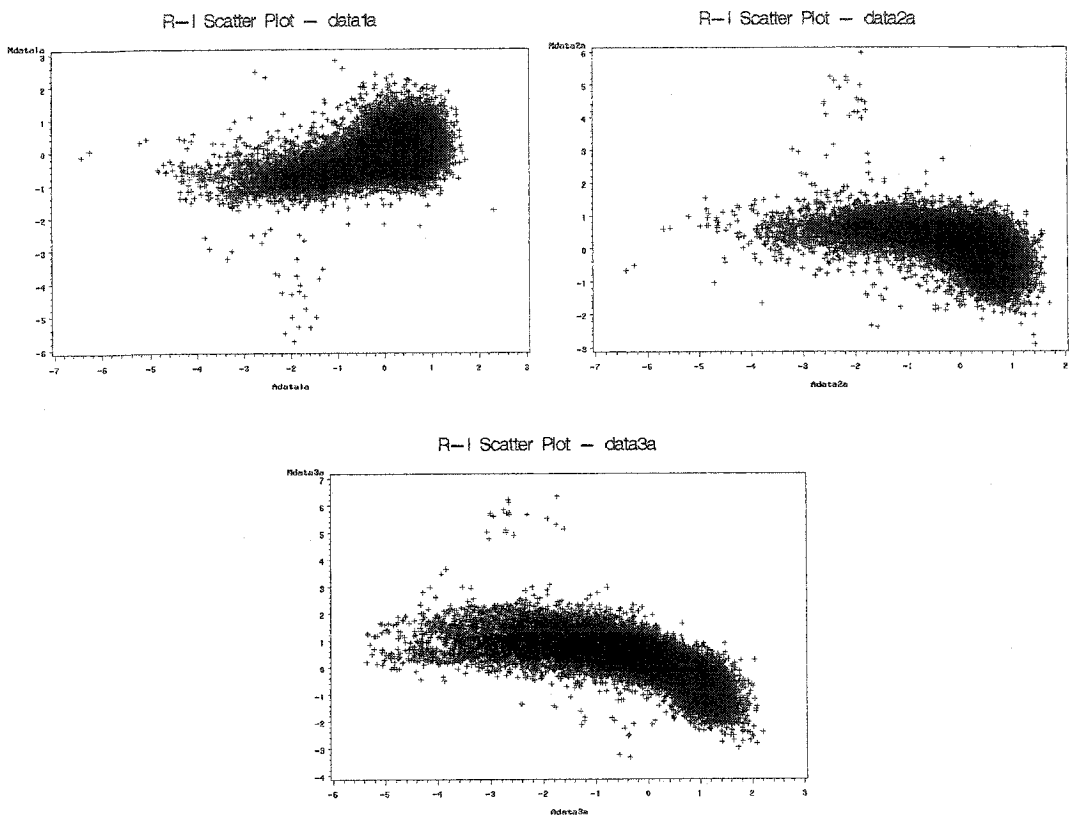


Figure 19: R-I plots for the 3 slides. Notice a non-linear dye distortion for all the slides. Top left represents plot for Experiment 1(a), top right represents plot for Experiment 2(a) and bottom represents plot for Experiment 3(a).

LOWESS detects deviations from the expected behavior and corrects them by performing a local weighted linear regression for each data point in the R-I plot. We used `PROC LOESS` from `SAS 9.1` to find the best smoothing parameter of the fitted curve. For the LOWESS method in `SAS 9.1`, weighted least squares is used to fit the curve of the predictors at the centers of neighborhoods. The fraction of the data, called the smoothing parameter, in each local neighborhood controls the smoothness of the fitted curve.

There are many weight functions that can be applied, but one of the most common is the tri-cube weight function (see [8]),

$$w(u) = 1 - (|u|^3)^3 \tag{9}$$

where u is the distance from a particular data point to those in its neighborhood.

The data sets are corrected by subtracting from each observation the loess fitted curve, $C_i(A)$. We retain the residuals.

$$log_2\frac{R}{G} \longrightarrow log_2\frac{R}{G} - C_i(A) \tag{10}$$

where $C_i(A)$ is the LOWESS fit to the M vs. A plot.

We chose a range of smoothing parameter from 0.1, 0.2, 0.3 to 0.4 and examined which smoothing parameter fits best to the LOWESS curve. Each smoothing parameter has different fitted value. Figure 20 shows the R-I plots for all the four smoothing parameters of Experiment 1(a) before LOWESS fit and Figure 21 shows R-I plots after LOWESS fit.

After examining the plots, we observed that a good fit is obtained with smoothing parameter value 0.2. To aid the interpretation of these scatter plots, we also examine the R-I plots after LOWESS and found that the choice 0.2 is reasonable. With smoothing parameter value 0.1, there is gross over-fitting in the sense that the original data are exactly interpolated. The R-I plot for smoothing parameter 0.3 and 0.4 has a decreasing trend at the beginning of the range of the A variable. For Experiments 2 and 3, the same smoothing parameter was chosen.

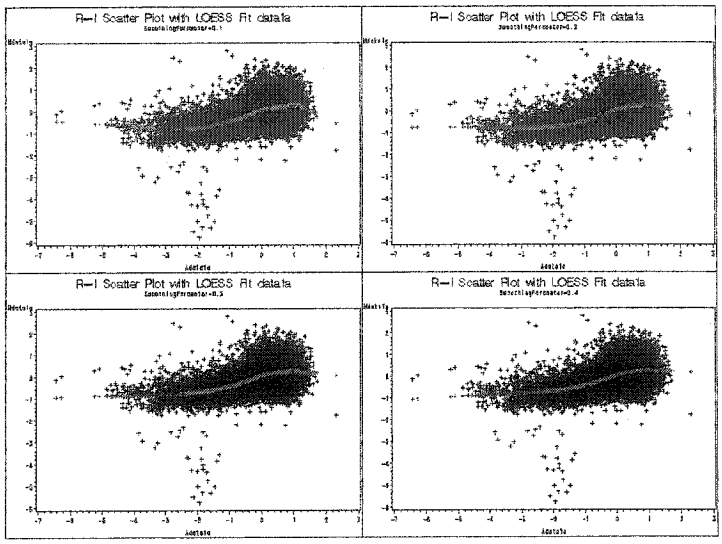


Figure 20: R-I plots for Experiment 1(a) before LOWESS. The four plots have four different smoothing parameter used to fit the curve for the plot.

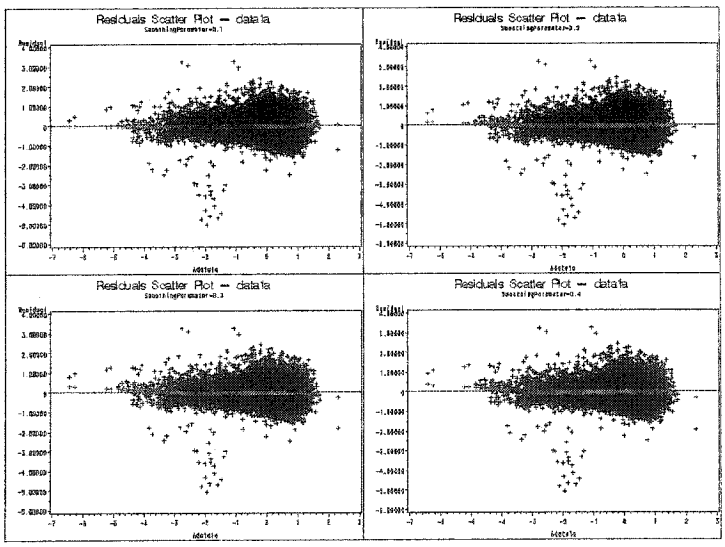


Figure 21: R-I plots after LOWESS. Observed that smoothing parameter 0.2 has better fit which lies on residuals zero.

The corrected M, $\widetilde{M}$ and A values were transformed back to Red, R and Green, G *logged* intensity values as follows:

$$log_2(R) = \frac{\widetilde{M} + 2A}{2} \tag{11}$$

$$log_2(G) = \frac{2A - \widetilde{M}}{2} \tag{12}$$

We recall that these data sets are replicated, therefore, mean values for each gene were computed.

$$log_2(X) = \sum_{j=1}^{n} \frac{log_2(X_j)}{n} \tag{13}$$

where,

$$X = R \text{ or } G \text{ and} j = 1, 2, ..., n.$$

2.2.4 Between Slide Normalization

As mentioned before, the purpose of color normalization is to eliminate the data artifacts introduced by the dyes. Dye bias is obviously seen in a cDNA microarray experiment. It is very rare to have dye intensity values equal on average and often the intensity values are higher for the green dye (see [8]). Usually, a flip dye experiment is used to control such phenomena (see [6]). The data sets used in this thesis as discussed in Section 2.1 above do form a flip dye experiment where the second slide used the reverse assignment of dies from the first slide.

Note that in the previous sections, data normalization was done separately for each slide. Here, we look at normalization methods that allow us to make comparisons between the different slides. Note that in a microarray experiment, each hybridization reaction may be slightly different, and so the overall intensities of different slides may be be different. In order to compare the samples hybridized to different slides on an equal footing, we will correct the variability of each slide. There is also a paper written on quality control of DNA array hybridization data where they compared intensity values of several hybridization experiments in another approach (see [9]).

There are three standard methods for data standardization so that the arrays or slides can be compared on an equal footing. These methods assume that the variations in the distributions between arrays are a result of experimental conditions and do not represent biological variability. (See [10]). If this assumption is not true, then these methods are not appropriate. The three methods in [10] are:

1. Scaling
2. Centering
3. Distribution normalization

The scaling method subtracts the mean log intensity over all of the data on the array from each individual log intensity measurement on the array. This ensures only that the means of all the distributions are equal. The centering method ensures that both the means and the standard deviations of all the distributions are equal. The distribution normalization involves centering and computing a new distribution with mean 0 and standard deviation 1.

The scaling method does not account for differences in range. We used the centering method, where data sets were centered to ensure the means and the standard deviations of all of the distributions are equal. This method is simpler than distribution normalization and it is the most commonly used method for data standardization. Each intensity value were subtracted with the mean intensity of each slide and divided by the standard deviation. Data standardization are perform separately for each R and G intensity.

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \tag{14}$$

where,

$$i = \text{slide 1, 2 or 3}$$

$$\mu = \text{mean intensity}$$

$$\sigma = \text{standard deviation}$$

The next step to correct for color distortion is to plot a comparison scatter plot. The comparisons are performed for Slide 1 and 2, Slide 1 and 3, and Slide 2 and 3 for both human and monkey samples separately.

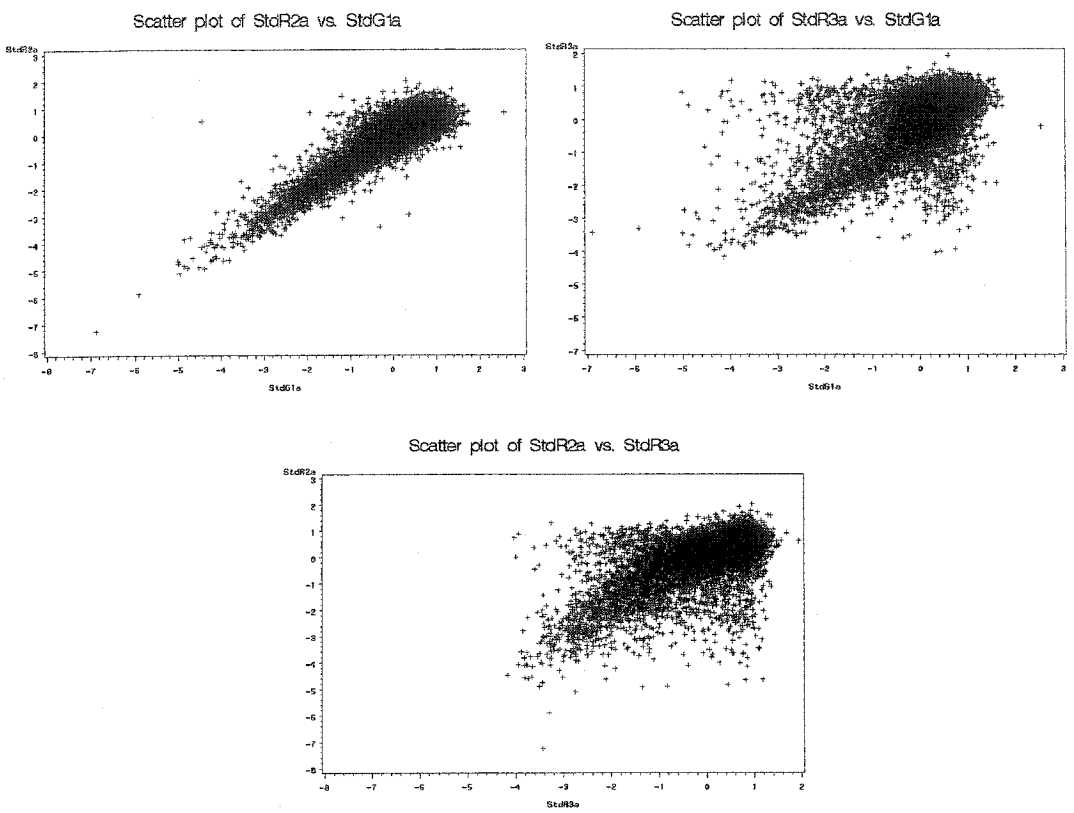


Figure 22: Comparison scatter plots for **human** sample. All the three plots appear along the diagonal. Notice any comparison involving Slide 3 has many outliers.

The scatter plots for human samples are shown in Figure 22 and scatter plots for monkey samples are shown in Figure 23. The top left of each figure represents plot for comparison between Slide 1 and 2, top right represents comparison between Slide 1 and 3 and the bottom panel represents comparison between Slide 2 and 3. Here, the scatter plots appear along the diagonal because at this point, data sets were

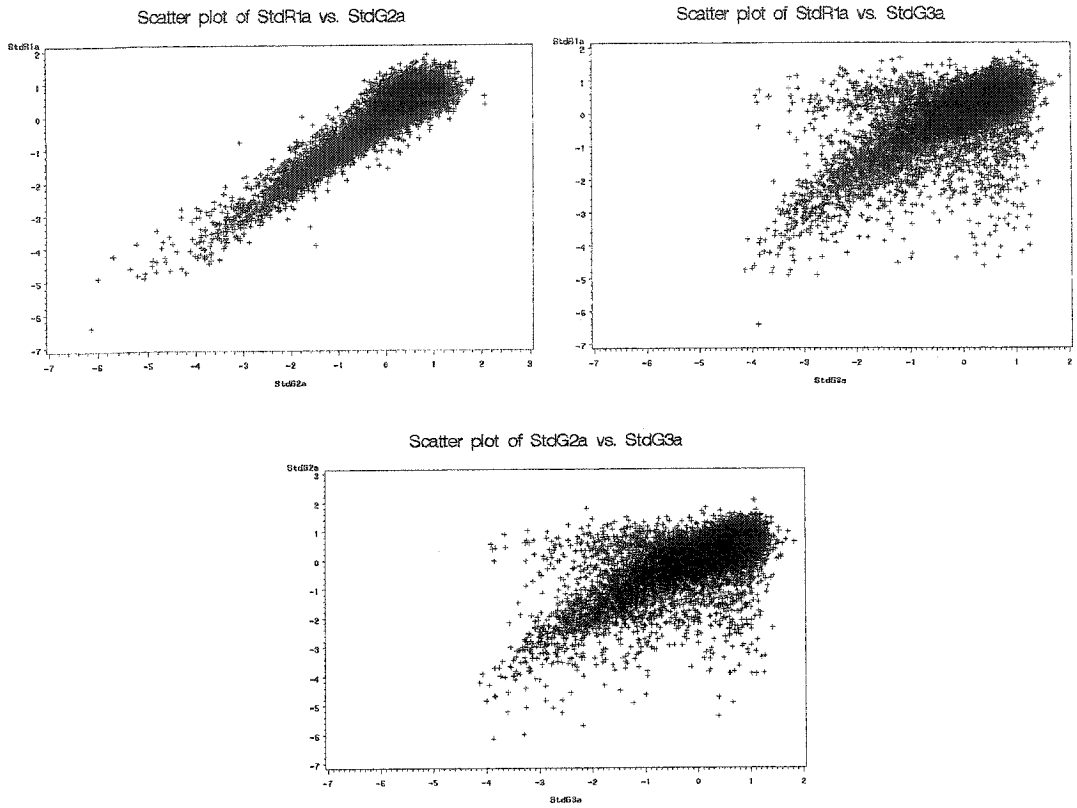


Figure 23: Comparison scatter plots for **monkey** sample. All the three plots appear along the diagonal. Notice any comparison involving Slide 3 has many outliers. Consequently Slide 3 shall be discarded.

normalized and corrected for color distortions. Most points in these scatter plot are **expected** to appear along the diagonal because most of the genes of an organism are expected to be unchanged. Therefore any plots depart from the diagonal line is due either to the inherent random noise or to the dyes.

Data were normalized by eliminating outliers on the scatter plots. Methods for eliminating outliers are similar with Section 2.2.2 as follows:

1. Compute the difference of the pair (sub.) and sum of the pair
2. Plot of sub. vs. sum
3. Compute absolute value of the subtraction value
4. Compute square root of the subtraction value
5. Plot histogram of the square root value
6. Eliminate data sets from the end-tailed of histogram

We noticed that most points on the scatter plot that compares slide 1 and slide 2 lie along the diagonal and there are very few outliers. But scatter plots that involved Slide 3 had many outliers. If we eliminate every outlier involving Slide 3, we will lose a lot of observations. We do not know if the third slide that switched channel is not comparable with Slide 1 and 2. Therefore, we split the analysis into two paths where for the first path, we removed Slide 3 and only do a between slide normalization for Slide 1 and Slide 2. The second path, we took the average of Slide 1 and Slide 2 and compared it with Slide 3. The outliers for this comparison were not eliminated from the data sets. It will be kept as it is for the average of Slide 1 and Slide 2. For Slide 3, it will be classed as missing data.

Path 1 : Between Slide 1 and 2

The outlier removal methods were discussed in the Sections 2.2.2. Figure 24 shows the scatter plots of Slide 1 vs. Slide 2 for each human and monkey samples after outlier removal. After all the normalization procedures, there are 8474 genes for experiment part (a). Before data cleaning, there were 9600 pairs of genes from experiment part (a).

Path 2: Between Average of Slide 1 and 2 with Slide 3

For this path, the average of Slide 1 and Slide 2 will be compared with Slide 3. As mentioned, all the outliers from Slide 3 will be treated as missing data. They were not deleted from the data sets. After detecting outliers, there were 7799 balance data and were plot on the scatter plots in Figure 25.

In either case, after normalization process, the *logged* ratio of the two channels will be computed, $log_2\frac{R}{G}$ or $log_2\frac{G}{R}$. We used the $\frac{Human}{Monkey}$ ratio. Note that the intensity values were *logged* from the first step of data pre-processing, hence the calculation for these *logged* ratios are log_2Human - log_2Monkey.

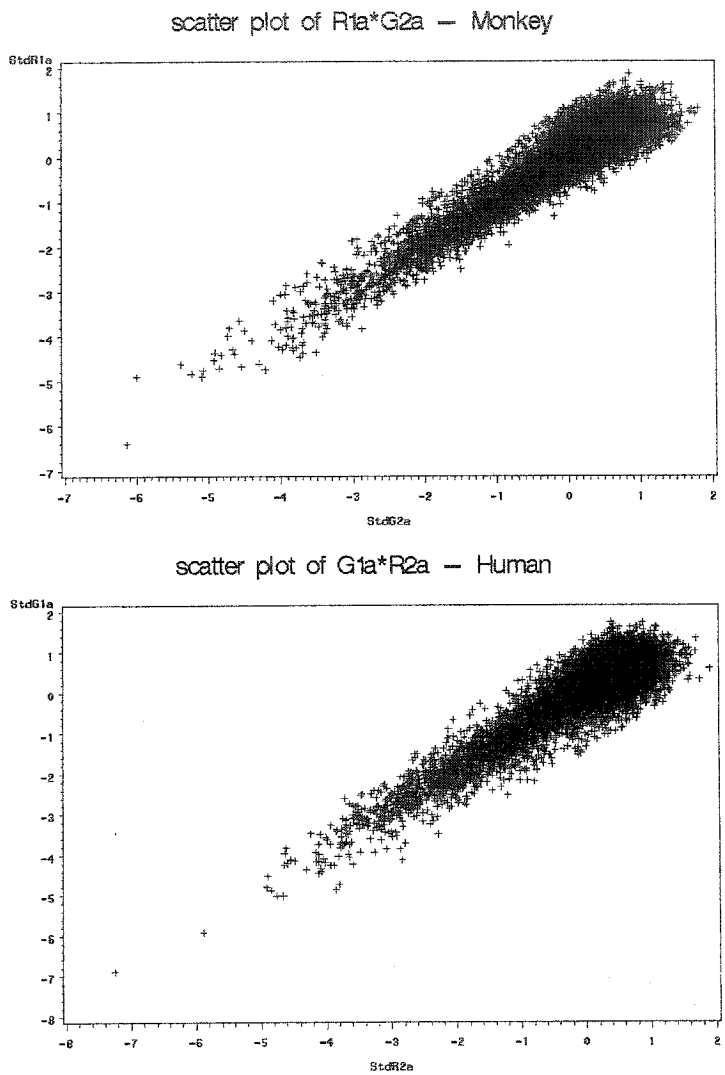


Figure 24: Scatter plot after between slides outlier removal - Slide 1 vs. Slide 2. Top panel is the plot for monkey sample and bottom panel is the plot for human sample.

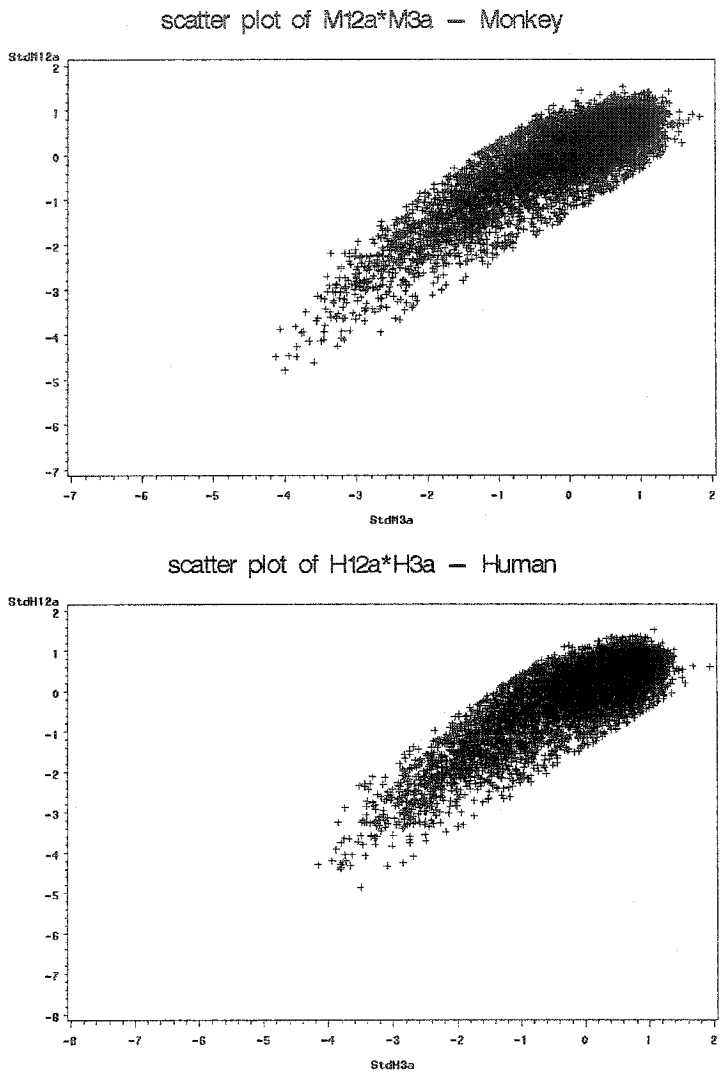


Figure 25: Scatter plot after between slides outlier removal - average Slide 1 and Slide 2 vs. Slide 3. Top panel is the plot for monkey sample and bottom panel is the plot for human sample.

Chapter 3

Gene Ontology

3.1 Data Acquisition and Preparation

As mentioned in Section 1.4, The Gene Ontology (GO) describes the known attributes of a gene. It is a set of controlled vocabularies used to describe biological features within a specific domain of biological knowledge (see [26], [23]). The GO Consortium, [26] decided that three terms were needed to describe different aspects of every protein (see [24]). The independent GOs are **biological process** (BP), **cellular component** (CC) and **molecular function** (MF). They are all attributes of a gene, a gene product or a gene product group. A gene product can have more than one molecular function, be used in one or more biological process and may be associated with one or more cellular components.

A BP is defined as a biological objective to which the gene product contributes. A process is the result of one or more ordered assemblies of MF. The MF is defined as the biochemical activity of a gene product. The MF only describes what is done without specifying where or when the actually activity takes place. The CC refers to the place in the cell where the gene product is active. It must be noted that not all terms are applicable to all organisms [5].

The GO information is a tabular database independent of any other and it is not populated with gene products of any organism. This database uses GO terms to annotate objects such as genes or gene products. There are many types of genetic

data available on the internet from several public databases. In this thesis, we used the database information from the Stanford Microarray Database (see [25]).

We obtained the genomic database information for the list of genes in our PBMC microarray data. Each gene spot of our microarray data has various types of gene names in different context. To be more consistent, we used a gene name of the type *Unigene Name* in the form of '*Xaaaaa*', where '*X*' is a letter and '*aaaaa*' are the corresponding numbers.

The genomic database information were obtained from the SOURCE resource. SOURCE is a functional genome resource for human, mouse and rat genes. The URL is as follows:

$$http://genome-www5.stanford.edu/cgi-bin/source/sourceBatchSearch$$

Following are the steps to download the GO database information from the web-site mentioned above:

1. Enter the list of gene names, in a text file, obtained from PBMC microarray data into "input file".
2. Select type of input as "Unigene Name".
3. Select organism as "Homo sapiens".
4. Check only the "Gene Ontology Annotation (full)" in the "Choose field(s) of extraction" section.
5. Uncheck every box in "Error Conditions" section.
6. Click "submit".

A sample of the GO database information is shown in Table 2. Here, we will use only the frequencies of BP, CC and MF for each gene. This choice of database information were selected for convenience purposes. There is no biological significance to this choice of database information and we do not expect it to be biologically useful.

Note that not every gene has GO database information. From the gene list from slide part (a), only 3172 out of 9600 genes have GO database information. We used

SAS 9.1 to count the number of BP, CC and MF in each gene and present the GO database information in the form of a contingency table. Table 3 shows the first 5 observations of the GO database contingency table for a few of the genes appearing in the microarray data sets. This database will now be represented geographically with Correspondence Analysis using SAS 9.1.

Gene	GO Entry
R11726	molecular function|glucose transporter activity|TAS|GO:0005355|GOA|10671487
	molecular function|transporter activity|IEA|GO:0005215|GOA|na
	biological process|glucose transport|TAS|GO:0015758|GOA|10671487
	biological process|carbohydrate metabolism|TAS|GO:0005975|GOA|10671487
	cellular component|integral to membrane|IEA|GO:0016021|GOA|na
	cellular component|integral to plasma membrane|TAS|GO:0005887|GOA|10671487
	biological process|carbohydrate transport|IEA|GO:0008643|GOA|na
	molecular function|sugar porter activity|IEA|GO:0005351|GOA|na
R37412	
R11793	molecular function|calcium-release channel activity|TAS|GO:0015278|GOA|9030597
	molecular function|receptor activity|NR|GO:0004872|GOA|na
	biological process|cation transport|IEA|GO:0006812|GOA|na
	biological process|calcium ion transport|TAS|GO:0006816|GOA|7511586
	biological process|muscle contraction|TAS|GO:0006936|GOA|9030597
	cellular component|smooth endoplasmic reticulum|TAS|GO:0005790|GOA|2298749
	cellular component|integral to plasma membrane|TAS|GO:0005887|GOA|2298749
R12521	molecular function|trypsin activity|IEA|GO:0004295|GOA|na
	molecular function|sugar binding|IEA|GO:0005529|GOA|na
	molecular function|peptidase activity|IEA|GO:0008233|GOA|na
	molecular function|calcium ion binding|IEA|GO:0005509|GOA|na
	biological process|proteolysis and peptidolysis|IEA|GO:0006508|GOA|na
	molecular function|chymotrypsin activity|IEA|GO:0004263|GOA|na
	biological process|heterophilic cell adhesion|IEA|GO:0007157|GOA|na
R15106	

Table 2: First 5 observations of GO database obtained from Stanford Microarray Database. Missing value indicates no GO database information.

Gene	Gene Ontology		
Acc	BP	CC	MF
H00136	2	0	1
H00168	0	0	1
H00222	1	0	5
H00239	1	1	3
H00293	5	2	1

Table 3: First 5 lines on the contingency table constructed from Experiment 1(a) and the corresponding GO information on BP, CC and MF.

3.2 Correspondence Analysis

The genomic information of Table 3 used is nonmetric data and it is a cross-tabulation of two categorical variables (Gene and GO). Therefore, Correspondence Analysis (CA) can be used to provide a multivariate representation of the contingency table. CA is a technique that facilitates both dimensional reduction and perceptual mapping of objects relative to these attributes, see [22]. This technique transforms the nonmetric data to a metric level and performs dimensional reduction and perceptual mapping. Note that CA was also used in analysis for microarray data (see [19], [20], [21]).

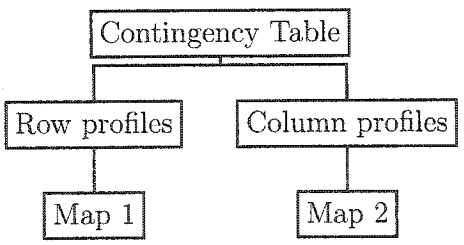


Figure 26: Analytical Process of Correspondence Analysis

A sample table shown in Table 4 will be used to describe the steps of the CA method. The analysis starts by transforming the frequencies in the contingency table

Genes	BP	CC	MF	Totals
gene1	a_1	b_1	c_1	T_1
gene2	a_2	b_2	c_2	T_2
gene3	a_3	b_3	c_3	T_3
.....	....	....	....	....
Totals	a_T	b_T	c_T	T

Table 4: Contingency Table of Genes and GOs

into proportions by rows shown in the top panel of Table 5 and proportions by columns shown in the bottom panel of Table 5. Note that the variables were mapped separately for each row and columns of the contingency table. The maps will later be merged into one 2-dimensional map as shown in Figure 27.

To illustrate, the row profiles for *gene1* were calculated in the following way: $\frac{a_1}{T_1} = a_{r1}$, $\frac{b_1}{T_1} = b_{r1}$ and $\frac{c_1}{T_1} = c_{r1}$. The table also shows the average row profile which is the profile of the marginal distribution or the column variable (GO), which is the following: $\frac{a_T}{T} = n_a$, $\frac{b_T}{T} = n_b$ and $\frac{c_T}{T} = n_c$. The last column in the top panel of Table 5 shows the row masses or the marginal profile. These are composed of the relative frequency distributions of the sums of the rows (marginal distribution), which is the following: $\frac{T_1}{T} = m_1$, $\frac{T_1}{T} = m_2$ and $\frac{T_3}{T} = m_3$.

The column profiles for *gene1* were calculated in the following way: $\frac{a_1}{a_T} = a_{c1}$, $\frac{b_1}{b_T} = b_{c1}$ and $\frac{c_1}{c_T} = c_{c1}$. Note that the row masses equal the average column profile and the column masses equal the average row profile.

For *Map 1*, the map of gene variables, the average row profile is the weighted average of the row profiles. This point is the *centroid* and it is placed at the origin on the principal axes. If a profile is very different from the average profile, then the point will lie far from the origin, whereas profiles that are close to the average will be represented by points close to the centroid. The distance corresponds to the chi squared separation.

Chi squared distance, d, represents the (squared) metric in CA, and is calculated separately for each row and column profiles. The distances between the different points may be calculated by the formula

Row Profiles

Genes	**BP**	**CC**	**MF**	Total	Row Masses
*gene*1	a_{r1}	b_{r1}	c_{r1}	1.00	m_1
*gene*2	a_{r2}	b_{r2}	c_{r2}	1.00	m_2
*gene*3	a_{r3}	b_{r3}	c_{r3}	1.00	m_3
.....	....	....	....	....	
Ave. row profile	n_a	n_b	n_c		

Column Profiles

Genes	**BP**	**CC**	**MF**	Ave. Column Profile
*gene*1	a_{c1}	b_{c1}	c_{c1}	m_1
*gene*2	a_{c2}	b_{c2}	c_{c2}	m_2
*gene*3	a_{c3}	b_{c3}	c_{c3}	m_3
.....	....	....	....	....
Column masses	n_a	n_b	n_c	

Table 5: Profiles and Masses for Table 4

$$d(i,i') = \sqrt{\sum_j \frac{(a_{ij} - a_{i'j})^2}{a_{.j}}} \tag{15}$$

where $d(i,i')$ is the "chi square" distance between the points i and i', a_{ij} are elements in the row profile, and $a_{.j}$ are elements in the average row profile.

These distances will be used to plot in a 2-dimensional space. To find the axis, the weighted sum of squared distances (z^2) from the points to the axis will be used. The weights are the row masses (m). Thus, the intention is to minimize $\sum mz^2$. The problem is solved by means of principal component analysis, and the result of this analysis provides a number of useful descriptive statistics in addition to the graphical display. A similar procedure was used for column profiles to perform *Map 2*.

This procedure is quite tedious to attempt by hand; we used `PROC CORRESP` from `SAS 9.1` to map the gene and `GO` from the contingency table.

3.3 CA Applied to GO database

The maximum number of dimensions (or axes) is the minimum of the number of rows and columns, minus one. Our contingency table has 3 columns and 3172 rows. Hence, we will produce a 2-dimensional CA map. The 2-dimensional map is shown in Figure 27. Note that genes that have similar GO are located close to one another. The Correspondence Analysis SAS Output is shown below:

Correspondence Analysis - plangoa

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	11 22 33 44 55
0.42252	0.17852	2668.35	53.27	53.27	**********
0.39571	0.15658	2340.44	46.73	100.00	*********
Total	0.33510	5008.79	100.00		

Degrees of Freedom = 6342

Column Coordinates

	Dim1	Dim2
bp	-0.1598	-1.2932
cc	1.7256	0.5600
mf	-0.8651	0.8802

Summary Statistics for the Column Points

	Quality	Mass	Inertia
bp	1.0000	0.3706	0.2947
cc	1.0000	0.2330	0.4038

mf	1.0000	0.3963	0.3015

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
bp	0.0095	0.6199
cc	0.6939	0.0731
mf	0.2966	0.3070

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
bp	0	2	2
cc	1	0	1
mf	2	2	2

Squared Cosines for the Column Points

	Dim1	Dim2
bp	0.0171	0.9829
cc	0.9154	0.0846
mf	0.5241	0.4759

The total chi-square statistic, which is a measure of the association between the rows and columns is 5008.79. The chi squares and inertia are explained equally for both the dimensions. About 53.27% explain *Dimension 1* and 46.73% explain *Dimension 2*. This indicates that the association between the row and column categories is essentially two dimensional.

It is always interesting to try to interpret the dimensions of the CA. From the SAS OUTPUT, we see that we can think of *Dimension 1* as $2*CC-MF$, and of *Dimension 2* as $\frac{1}{2}(CC+MF)-BP$. Thus, high *Dimension 1* indicates a gene influencing several locations through relatively few mechanisms. High *Dimension 2* shows numerous

influences on fewer biological process.

There are 3172 out of 9600 genes on this map. As mentioned in Section 2.2.4, the microarray data were split into two paths. Recall the first path discarded slide 3. The normalized data sets have 8474 observations. For these 8474 observations, there are 2883 genes with GO database information available to link with the microarray gene expression data.

For Path 2, there are 9600 unbalanced and 7799 balanced normalized data sets. As mentioned in Section 2.2.4, the outliers for Path 2 were not eliminated, it was kept as missing data for Slide 3. Here, there are 2656 genes with GO database information available to link with microarray gene expression data. These will be performed in the next chapter.

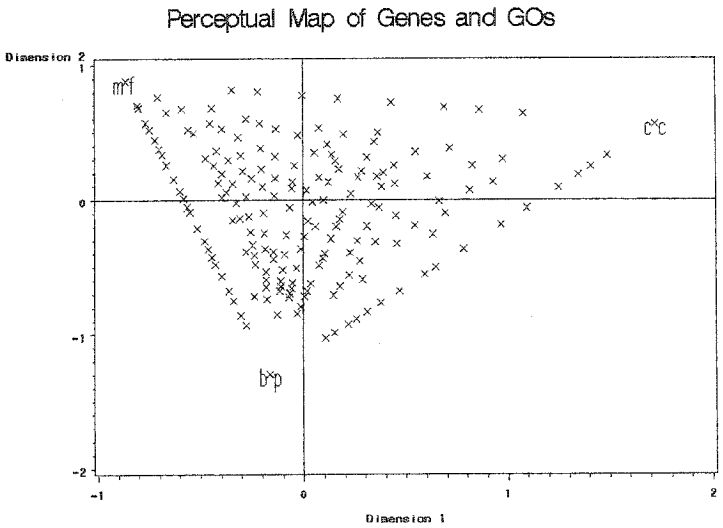


Figure 27: Perceptual Map of Genes and GO

Chapter 4

Spatial Linkage

As mentioned in Chapter 3, there are 3172 genes in our microarray experiment that have information on Gene Ontology. These genes were located on a 2-dimensional map (shown in Figure 27) using Correspondence Analysis applied to the Gene Ontologies. Also recall that the data cleaning process of the experimental data were split into 2 parallel path, where for the first path, we removed slide 3 and did a between slide correction of slide 1 and slide 2. We will name the first path as Path 1 for convenience. For the second path, we kept slide 3 and compared it with the average of slide 1 and slide 2. We will name this section as Path 2. The genomic database information was linked to the normalized microarray data by plotting a third dimension on the 2-dimensional perceptual CA map.

4.1 Path 1

For Path 1, there were 2883 genes from the microarray data that has GO information. The normalized intensity value for these 2883 genes was plotted on each respecting gene spot on the CA map. Missing values will not be plotted on the CA map. Figure 28 shows the 3-dimensional plot for the Path 1 normalized microarray data of Experiment 1(a). The top panel shows the plots in three dimensions and the bottom panel shows the plot for dimension R, ratio vs. *Dimension 1*. For Experiment 2(a), the 3-dimensional plot were shown in Figure 29.

We then apply the response surface method to the surfaces of the 3-dimensional plot to produce fitted model of the intensity ratios using the axes of the GO perceptual plot as predictors. We used PROC RSREG from SAS 9.1 to find the fitted response surface model for the surface. The fitted surfaces for both Experiment 1(a) and 2(a) are shown in Figure 30. Note that the surfaces are similar for both experiments. Figure 31 shows the contour plots of the two experiments after response surface analysis.

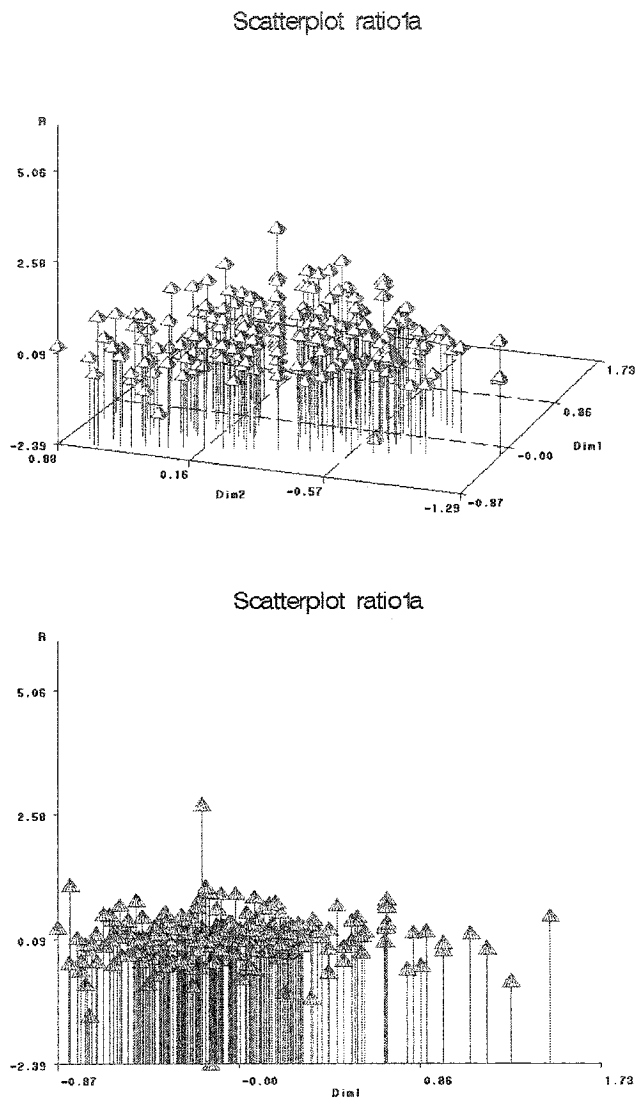


Figure 28: 3-dimensional scatter plot for Experiment 1(a). Top panel shows the plot in three dimensions and bottom panel shows the plot for R, ratio vs. *Dimension 1.*

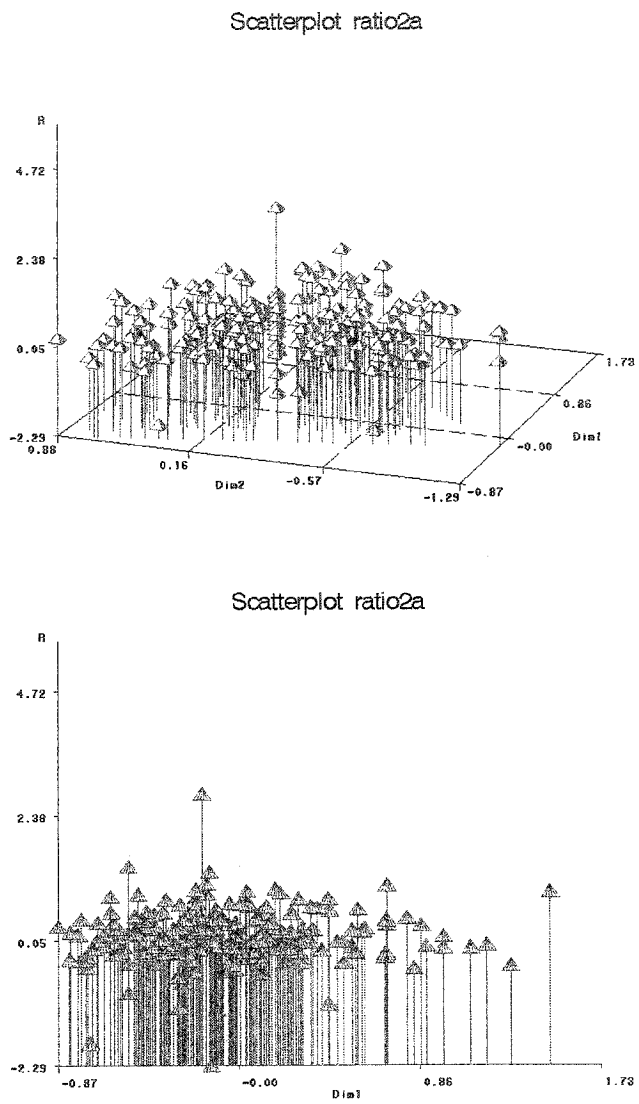


Figure 29: 3-dimensional scatter plot for Experiment 2(a). Top panel shows the plot in three dimensions and bottom panel shows the plot for R, ratio vs. *Dimension 1*.

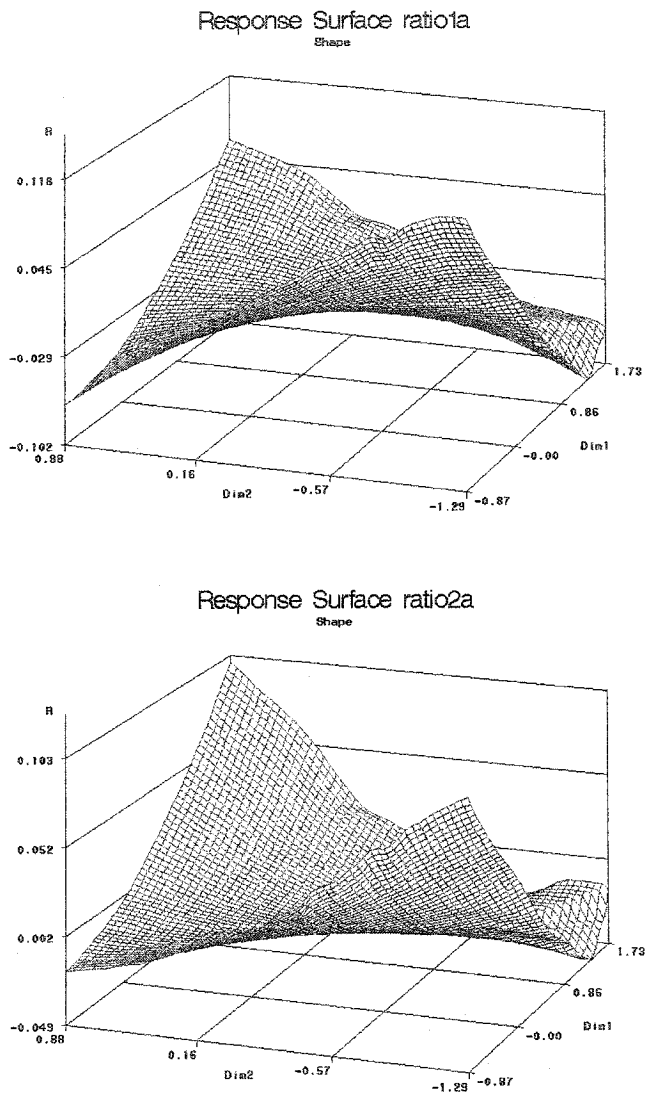


Figure 30: Response Surface fitted surface for Experiment 1(a) and 2(a). Top panel - Experiment 1(a). Bottom panel - Experiment 2(a).

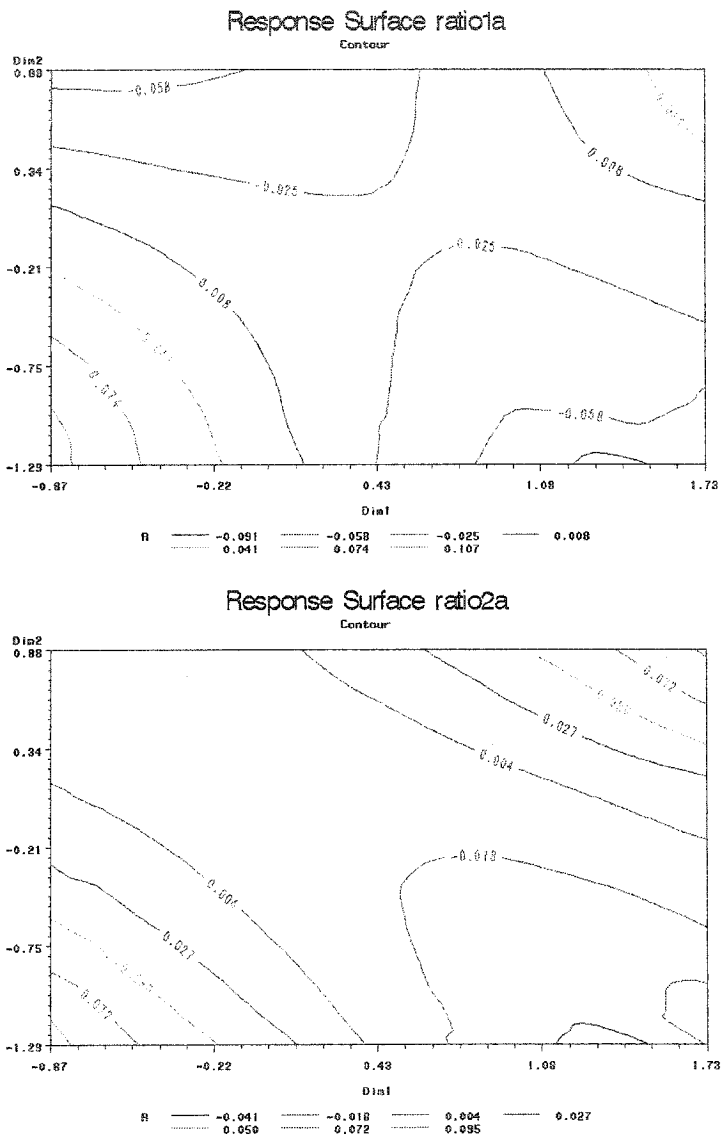


Figure 31: Contour plot of Experiment 1(a) and 2(a). Top panel - Experiment 1(a). Bottom panel - Experiment 2(a).

4.2 Path 2

For Path 2, there were 2656 genes from the microarray data that have GO database information. Figure 32 shows the 3-dimensional plot for the normalized microarray data of the average of Experiment 1(a) and 2(a). For convenience, we will name the average experiment as Experiment 1-2(a). The top panel shows the plot in three dimensions and the bottom panel shows the plot for R, ratio vs. *Dimension 1*. For Experiment 3(a), the 3-dimensional plots were shown in Figure 33. Figure 35 shows the contour plot of the experiments after response surfaces.

The fitted response surface plot for both of the experiments are shown in Figure 34. Note that the surfaces are different. We reconfirm that Slide 3 has poor overall comparability to the other two slides and we will eliminate Slide 3 from our microarray gene expression data. We will only use the microarray data from Slide 1 and Slide 2.

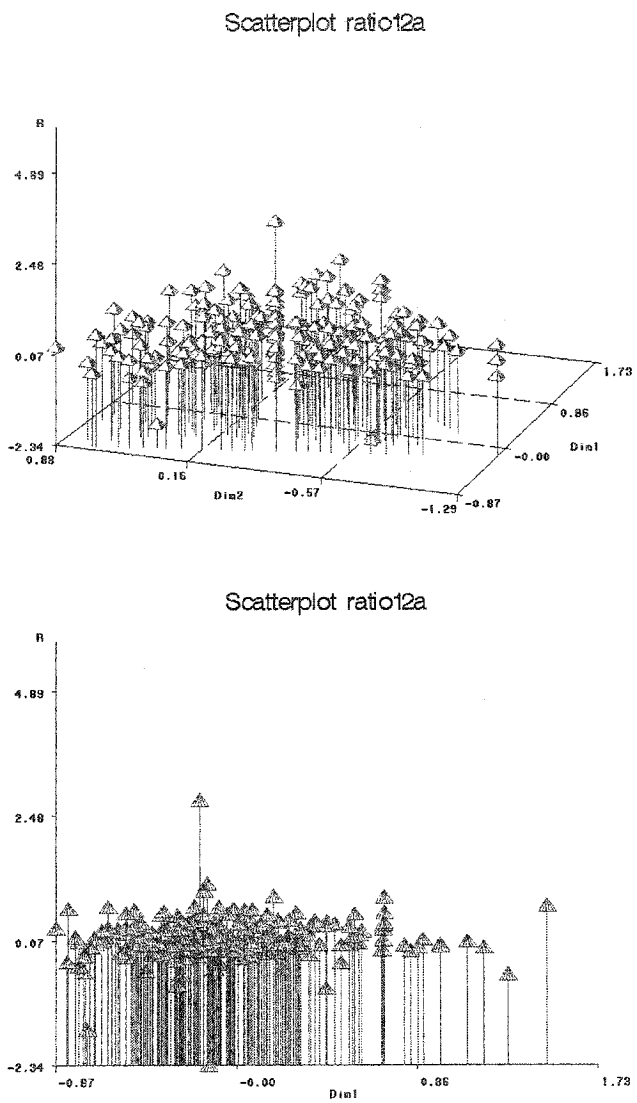


Figure 32: 3-dimensional scatter plot for Experiment 1-2(a). Top panel shows the plot in three dimensions and bottom panel shows the plot for R, ratio vs. *Dimension 1*.

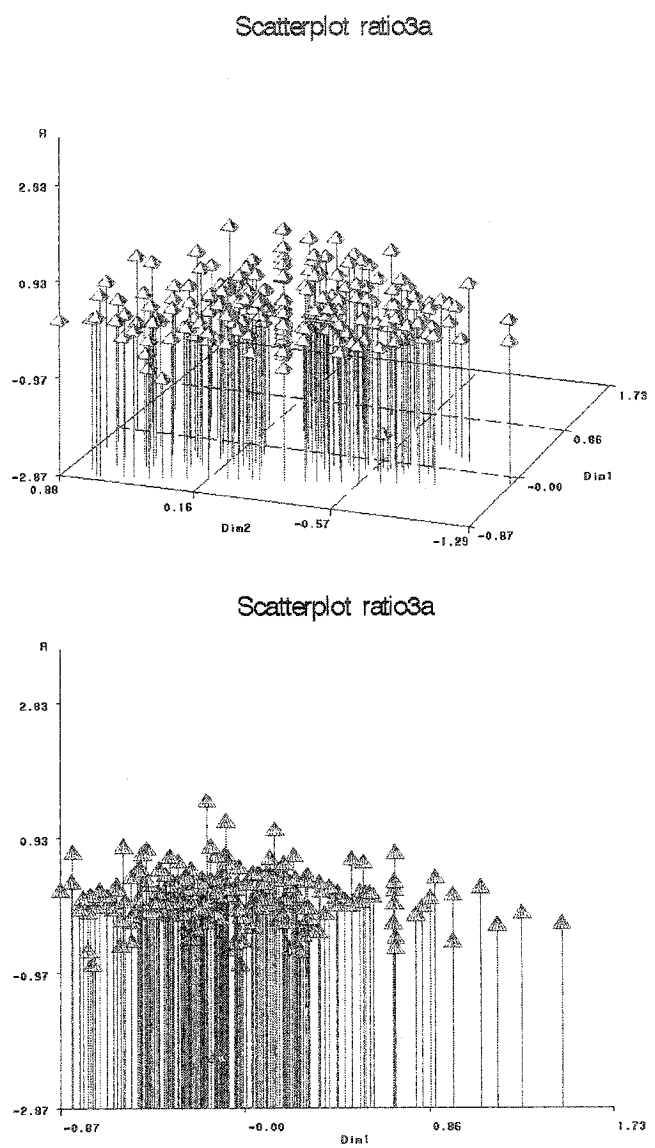


Figure 33: 3-dimensional scatter plot for Experiment 3(a). Top panel shows the plot in three dimensions and bottom panel shows the plot for R, ratio vs. *Dimension 1*.

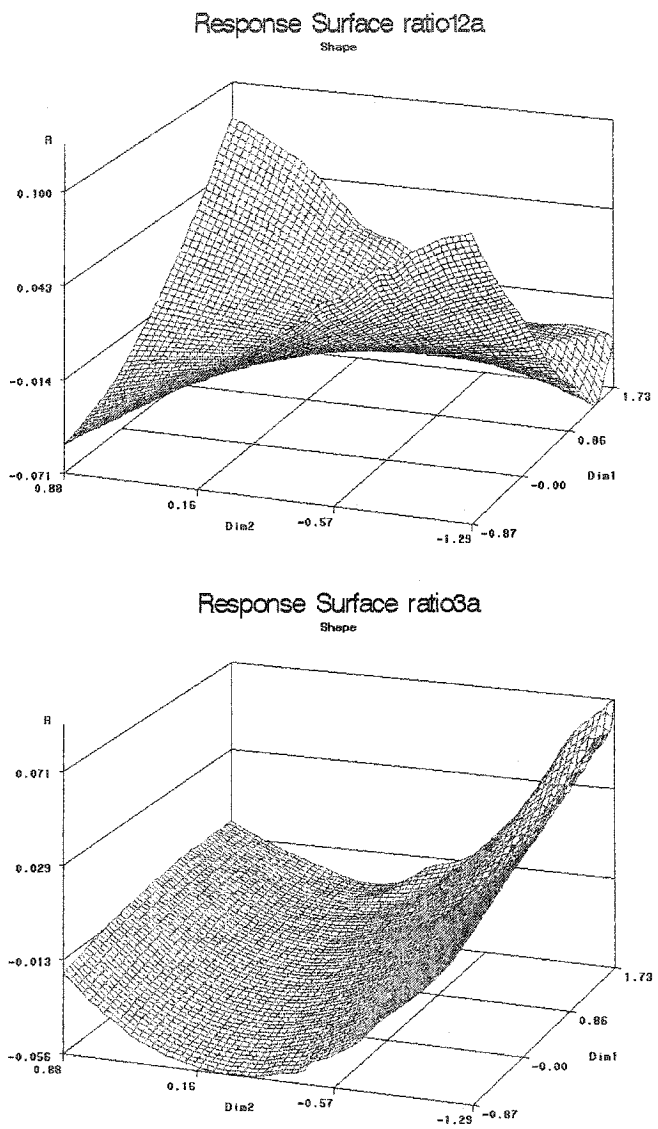


Figure 34: Response Surface fitted surface for Experiment 1-2(a) and 3(a). Top panel - Experiment 1-2(a). Bottom panel - Experiment 3(a).

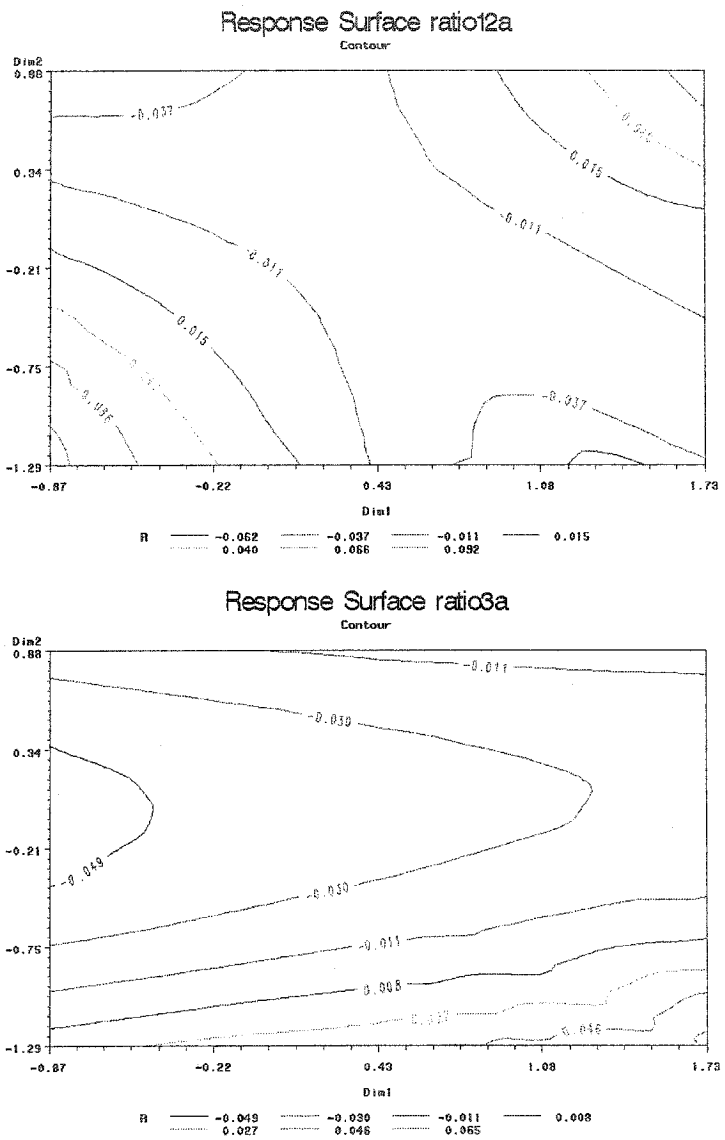


Figure 35: Contour plots of Experiment 1-2(a) and 3(a). Top panel - Experiment 1-2(a). Bottom panel - Experiment 3(a).

As a final step, we will take the average ratios of Slide 1 and Slide 2 and plot a new 3-dimensional representation of the spatial link between microarray data and genomic database. The 3-dimensional plots are shown in Figure 36. The response surfaces fitted plots and contour plots are shown in Figure 37.

A plot of the actual normalized intensity values vs. the response surface predicted normalized intensity values are shown in Figure 38. This plot shows that the response surface fitted to the microarray data are not useful. Indeed the inter-quantile range of the actual observed values is 0.374 ($Q3 - Q1 = 0.183 - (-0.191)$), whereas the full range of the predicted values is 0.101 ($max - min = 0.058 - (-0.043)$).

This is expected because the choice of our genomic databases are not biologically significant. Hence our response surface model is not a good fit. Nonetheless, we have shown that the methods exist to provide a graphical representation of microarray data as a function of archived database information.

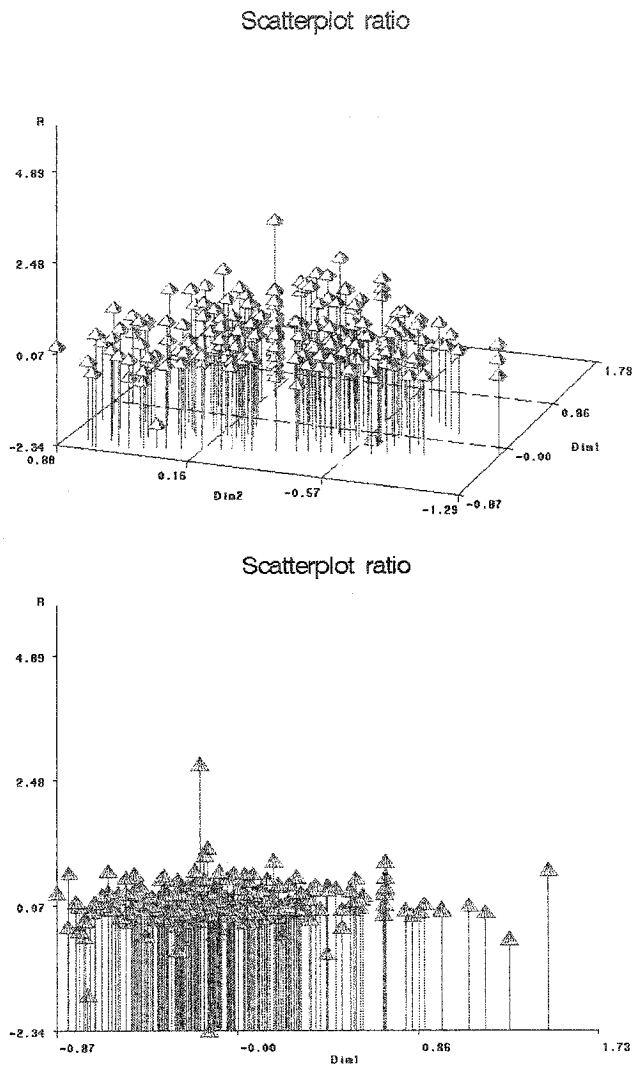


Figure 36: 3-dimensional scatter plot. Top panel shows the plot in three dimensions and bottom panel shows the plot for R, ratio vs. *Dimension 1*.

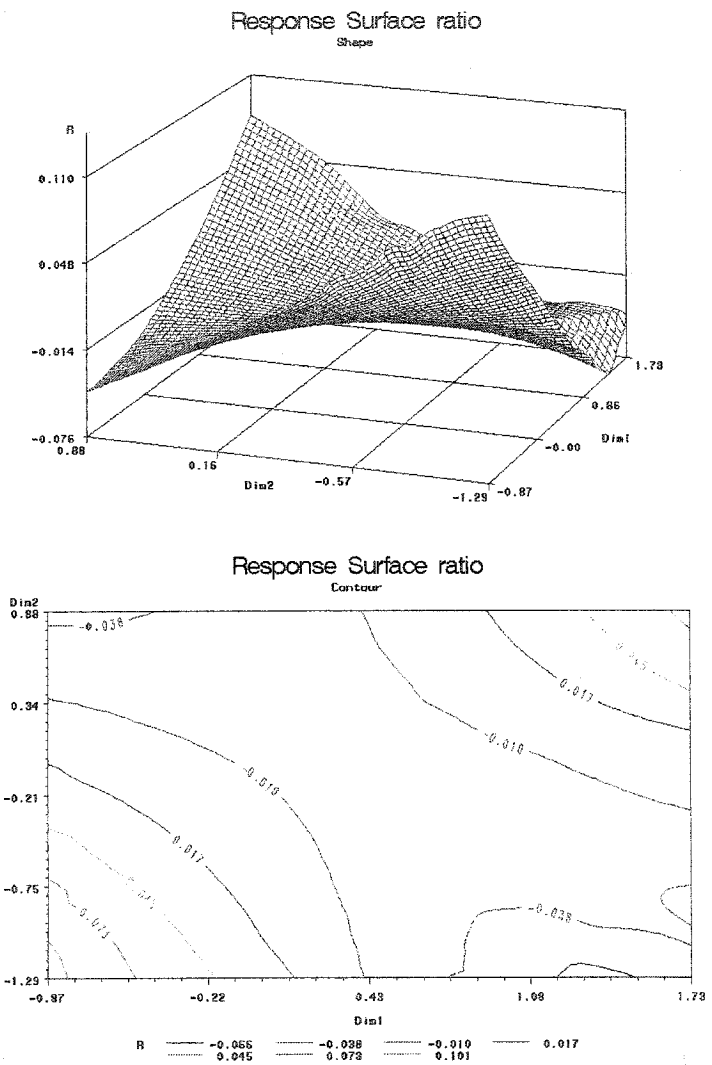


Figure 37: Response Surfaces predicted surface and contour

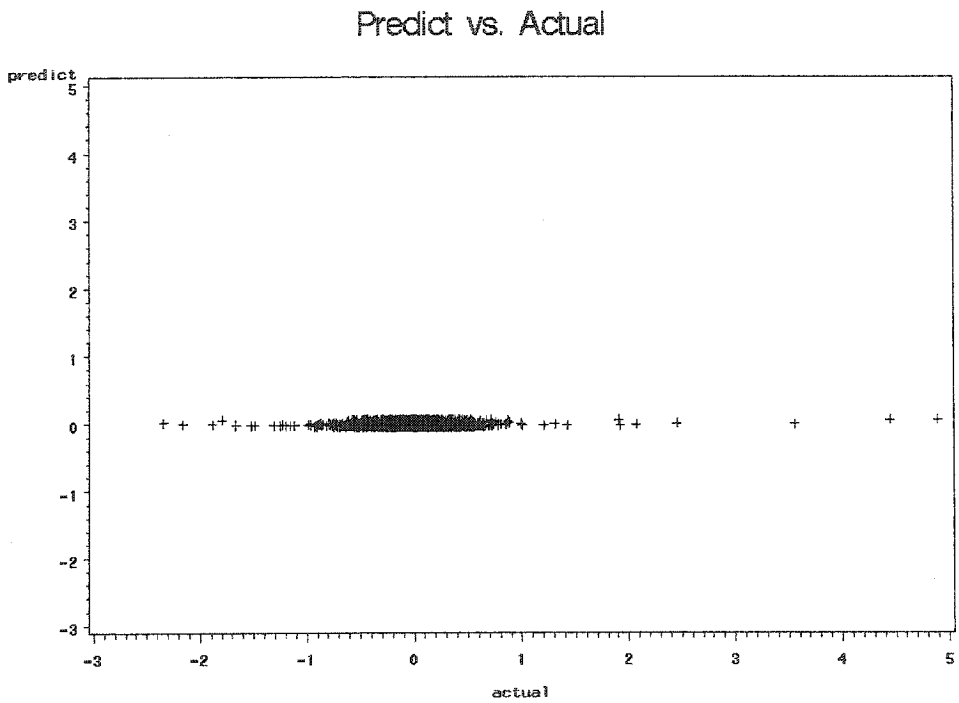


Figure 38: Plot of predicted value vs. actual value. Note that the predicted values are basically the mean of the actual observations.

Chapter 5

Conclusions and Future Work

We recommend that future work should look into the availability of biologically significant genomic databases. There are tremendous amounts of resources publicly available on-line. The National Center for Biotechnology Information (NCBI) provides sequence, protein, structure and genome databases. Other resources are UniGene and LocusLink. There are also graphically represented databases of cellular processes in the Kyoto Encyclopedia of Genes and Genomes (KEGG), GenMAPP, ArrayExpress and SAGEmap. A paper by Stephanopoulos et. al [27], describes mapping physiological states from microarray expression measurements. There are also various database designs described in [28]. With a better genomic database, we expect that the response surface model obtained from the CA genomic map could be a better fit. In particular we anticipate using more than three columns in Table 3.

There is considerable scope as well to improve the methods used here. For example, using spatial statistics methods such as Kriging to relate microarray data to GO coordinates may incorporate statistical dependence among proximate observations in a way that RSM cannot accomplish. A step that is not as complex as Kriging but more adaptable to abrupt surfaces is to use a two-dimensional LOWESS surface both to fit data to the GO map and to correct for spatial effects. LOWESS could also be used instead of RSM in the location correction step of the microarray data normalization. This would have helped in eliminating the streak in Figure 11.

Should a non-trivial model of gene expression be found in terms of GO database,

it then becomes possible to both predict gene expression from GO characteristics, and to estimate GO characteristics from gene expression. This would be an exciting development if it succeeded.

Chapter 6

Appendices

6.1 The Source Code

6.1.1 Data Preparation

Correspondence Analysis for GO Data Sets

```
%let dir=C:\Thesis\Analysis\;
%let macros=SASwork\macros2\;

options SASAUTOS="&dir&macros"; options MAUTOSOURCE;

/* PBMC DATA IMPORTATION FOR CORRESPONDENCE ANALYSIS */

/* Import Gene Ontology Datasets */
%importation2(type=EXCEL2000,path=&dir,file=PBMC_GOa.xls,out=goa)

/* DATA PREPARATION */
%preparation(data=goa,out=goa2)

/* CORRESPONDANCE ANALYSIS */
%correspondance(data=goa2,out=coorgoa,plan=plangoa)
```

Importation of Microarray Data

```
/* MICROARRAY DATA IMPORTATION FOR NORMALIZATION*/
```

```
/* import intensities data from excel file */
%importation2(type=EXCEL2000,path=&dir,file=data1a.xls,out=intens1a)
%importation2(type=EXCEL2000,path=&dir,file=data2a.xls,out=intens2a)
%importation2(type=EXCEL2000,path=&dir,file=data3a.xls,out=intens3a)
```

6.1.2 Data Pre-Processing and Normalization

Within Spot Correction

```
/*INTENSITY PREPARATION - WITHIN SPOT CORRECTION*/

/* change the value of substract to 'false' if do not want to do
background subtraction and do log transformation for intensity
values */
%prepareintens(data=intens1a,out=intens1acl,var=data1a,substract=true)
%prepareintens(data=intens2a,out=intens2acl,var=data2a,substract=true)
%prepareintens(data=intens3a,out=intens3acl,var=data3a,substract=true)

/* DO location correction using response surfaces method */
%rsreg(data=intens1acl, var=data1a, out1=Rpredata1a, out2=Gpredata1a)
%rsreg(data=intens2acl, var=data2a, out1=Rpredata2a, out2=Gpredata2a)
%rsreg(data=intens3acl, var=data3a, out1=Rpredata3a, out2=Gpredata3a)

/*PREPARE location corrected data*/
%dataset(data1=Rpredata1a, data2=Gpredata1a, data3=intens1acl, var1=data1a,
        var2=_data1a, out1=Rdata1a, out2=Gdata1a, out3=data1a, del_neg=no)
%dataset(data1=Rpredata2a, data2=Gpredata2a, data3=intens2acl, var1=data2a,
        var2=_data2a, out1=Rdata2a, out2=Gdata2a, out3=data2a, del_neg=no)
%dataset(data1=Rpredata3a, data2=Gpredata3a, data3=intens3acl, var1=data3a,
        var2=_data3a, out1=Rdata3a, out2=Gdata3a, out3=data3a, del_neg=no)
```

Between Spot Correction

```
/*BETWEEN SPOT CORRECTION*/

/*Prepare rep A & B in Cartesian Product match merge: PLOT of
replicates A vs. B for each color*/
%ab2(set=data1a, var=data1a, out=xdata1a)
%ab2(set=data2a, var=data2a, out=xdata2a)
%ab2(set=data3a, var=data3a, out=xdata3a)
```

```
/*prepare data for eliminating outliers*/
%prepareab2(data=xdata1a, var=data1a)
%prepareab2(data=xdata2a, var=data2a)
%prepareab2(data=xdata3a, var=data3a)

/*scatter plots and histogram for eliminating outliers*/
%plot(data=xdata1a, var=R_data1a)
%plot(data=xdata1a, var=G_data1a)
%plot(data=xdata2a, var=R_data2a)
%plot(data=xdata2a, var=G_data2a)
%plot(data=xdata3a, var=R_data3a)
%plot(data=xdata3a, var=G_data3a)

/*eliminate outliers*/
%delete1a(data=data1acl, set=xdata1a, var=data1a)
%delete2a(data=data2acl, set=xdata2a, var=data2a)
%delete3a(data=data3acl, set=xdata3a, var=data3a)

/*scatter plot after eliminating outliers*/
%plot1(data=data1acl, var=data1a)
%plot1(data=data2acl, var=data2a)
%plot1(data=data3acl, var=data3a)

/*Compress back data sets*/
%compress(data1=intens1acl, data2=data1acl, out=dat1a)
%compress(data1=intens2acl, data2=data2acl, out=dat2a)
%compress(data1=intens3acl, data2=data3acl, out=dat3a)
```

Within Slide Correction

```
/*INTENSITY PREPARATION - WITHIN SLIDE CORRECTION*/

/*join the replicates into one variable and calculate M and A*/
%prepareint2(data=int1a, set=dat1a, var=data1a)
%prepareint2(data=int2a, set=dat2a, var=data2a)
%prepareint2(data=int3a, set=dat3a, var=data3a)

/*R_I Scatter Plot to determine color distortion*/
%RIPlot(data=int1a, var=data1a)
%RIPlot(data=int2a, var=data2a)
%RIPlot(data=int3a, var=data3a)

/*LOESS Normalization*/
%LOESS(data=int1a, var=data1a)
%LOESS(data=int2a, var=data2a)
%LOESS(data=int3a, var=data3a)
```

```
%LOESSDATA(data1=normint1a, data2=int1a, var=data1a, pred=norm1a, resid=residata1a, out1=normint1ax,
           out2=loess1a, SmoothingParameter=0.2)
%LOESSDATA(data1=normint2a, data2=int2a, var=data2a, pred=norm2a, resid=residata2a, out1=normint2ax,
           out2=loess2a, SmoothingParameter=0.2)
%LOESSDATA(data1=normint3a, data2=int3a, var=data3a, pred=norm3a, resid=residata3a, out1=normint3ax,
           out2=loess3a, SmoothingParameter=0.2)

/*transform data for response surfaces*/
%transform(data=norm_1a, set=loess1a, part=1a, var=data1a)
%transform(data=norm_2a, set=loess2a, part=2a, var=data2a)
%transform(data=norm_3a, set=loess3a, part=3a, var=data3a)

/*MEAN REPLICATES OF THE NORMALIZED INTENSITIES*/
%meanorm(data=norm_1a, out=norm1a, part=1a)
%meanorm(data=norm_2a, out=norm2a, part=2a)
%meanorm(data=norm_3a, out=norm3a, part=3a)

/*Standardize the corrected intensity values to do a between
slides correction. Use Macro %means and  %standard*/
/* MERGE normalized data of the 3 experiments into one table for
scatter plots */

%means(data=norm1a, var=1a)
%means(data=norm2a, var=2a)
%means(data=norm3a, var=3a)

/*Standardize the different experiment to merge later*/
%standard(data=norm1a, Rmean=0.0026465, Gmean=0.0035815, Rstd=0.9594938, Gstd=0.9503553, var=1a)
%standard(data=norm2a, Rmean=0.0150846, Gmean=0.0146289, Rstd=0.9426471, Gstd=0.9574666, var=2a)
%standard(data=norm3a, Rmean=-0.0025578, Gmean=-0.0039856, Rstd=1.3350992, Gstd=1.3474196, var=3a)

/* MERGE normalized data of the 3 experiments into one table for
scatter plots */
%mergeintens(intens1=norm1a,intens2=norm2a,intens3=norm3a,part=a,out=norma)

/*scatter plots after eliminating outliers, location and color
normalization*/
%scatter(data=norma, var1=1a, var2=2a)
%scatter(data=norma, var1=1a, var2=3a)
%scatter(data=norma, var1=2a, var2=1a)
%scatter(data=norma, var1=2a, var2=3a)
%scatter(data=norma, var1=3a, var2=1a)
%scatter(data=norma, var1=3a, var2=2a)

/* Found from the scatter plots above, intensities for slide 3 are
uncomparable with other slides. Hence, the analysis will split
```

```
into 3 methods:

1) discard slide 3 and continue with slide 1 and 2

2) keep slide 3, take the average of slide 1 and 2 - outliers for
slide 3 will be kept as missing data and keep intensities of slide
1 and 2 as usual */

quit;
```

Between Slide Correction

Path 1

```
%let dir=C:\Thesis\Analysis\;
%let macros=SASwork\macros\;

options SASAUTOS="&dir&macros"; options MAUTOSOURCE;

/*data preparation for choosing outliers - datasets for slide 3
will be discarded at this point*/
%prepare(data=nora, set=norma, var1=1a, var2=2a, part=a)

/*scatter plots and histogram for eliminating outliers*/
%plot(data=nora, var=Ha)
%plot(data=nora, var=Ma)

/*eliminate outliers*/
%deleteaa(data=CDa, set=nora, var=a)

/*scatter plot after eliminating outliers*/
%plot2(data=CDa, var1=1a, var2=2a)

/*calculate ratio intensity for response surfaces with database
information*/
%ratio(data=ratio1a, set=CDa, var1=G1a, var2=R1a)
%ratio(data=ratio2a, set=CDa, var1=R2a, var2=G2a)

/* PLOT SURFACES FOR NORMALIZED EXPERIMENTS */
%surfaces(plan=plangoa, surfacei=surf1ai, surfaceii=surf1aii, surfaceiii=surf1aiii, var=ratio1a)
%surfaces(plan=plangoa, surfacei=surf2ai, surfaceii=surf2aii, surfaceiii=surf2aiii, var=ratio2a)

quit;
```

Path 2

```
%let dir=C:\Thesis\Analysis\;
%let macros=SASwork\macros\;

options SASAUTOS="&dir&macros"; options MAUTOSOURCE;

/*take average of slide 1 and 2*/
%average(data=norra, set=norma, StdR1=StdR1a, StdG1=StdG1a, StdR2=StdR2a, StdG2=StdG2a,
            StdR3=StdR3a, StdG3=StdG3a, StdR12=StdR12a, StdG12=StdG12a,
            StdH12=StdH12a, StdM12=StdM12a, StdH3=StdH3a, StdM3=StdM3a)

/*prepare data sets for plots for eliminating outliers*/
%prepare2(data=norra, set=norra, var1=12a, var2=3a, part=a)

/*scatter plots and histogram for eliminating outliers*/
%plot(data=norra, var=Ha)
%plot(data=norra, var=Ma)

/*keep outliers for slide12 and set missing data for slide 3*/
%deletea33(data=CDRa, set=norra, var=a)

/*scatter plot after eliminating outliers*/
%plot22(data=CDRa, var1=12a, var2=3a)

/*calculate ratio intensity for response surfaces with database
information*/
%ratio(data=ratio12a, set=CDRa, var1=H12a, var2=M12a)
%ratio(data=ratio3a,  set=CDRa, var1=H3a,  var2=M3a)

/* PLOT SURFACES FOR NORMALIZED EXPERIMENTS */
%surfaces(plan=plangoa, surfacei=surf12ai, surfaceii=surf12aii, surfaceiii=surf12aiii, var=ratio12a)
%surfaces(plan=plangoa, surfacei=surf3ai, surfaceii=surf3aii, surfaceiii=surf3aiii, var=ratio3a)

quit;
```

6.1.3 SAS Code for Plot of Geographical Locations

```
%let dir=C:\Thesis\Analysis\;
%let macros=SASwork\macros\;

options SASAUTOS="&dir&macros"; options MAUTOSOURCE;

/*GEOGRAPHICAL PLOTS - each channel*/
%geo1(data1=intens1acl,data2=data1a,part=1a,var=data1a)
%geo1(data1=intens2acl,data2=data2a,part=2a,var=data2a)
%geo1(data1=intens3acl,data2=data3a,part=3a,var=data3a)
```

```
/*GEOGRAPHICAL PLOTS - Red channel*/
%geo2(data1=intens1acl,data2=data1a,data3=int1a,data4=Norm_1a,part=1a,var=data1a)
%geo2(data1=intens2acl,data2=data2a,data3=int2a,data4=Norm_2a,part=2a,var=data2a)
%geo2(data1=intens3acl,data2=data3a,data3=int3a,data4=Norm_3a,part=3a,var=data3a)
/*
%geo(data1=intens1bcl,data2=data1b,data3=int1b,data4=Norm_1b,part=1b,var=data1b)
%geo(data1=intens1bcl,data2=data2b,data3=int2b,data4=Norm_2b,part=2b,var=data2b)
%geo(data1=intens3bcl,data2=data3b,data3=int3b,data4=Norm_3b,part=3b,var=data3b)
*/

/*GEOGRAPHICAL PLOTS - ratio*/
%geo(data1=intens1acl,data2=data1a,data3=int1a,data4=Norm_1a,part=1a,var=data1a)
%geo(data1=intens2acl,data2=data2a,data3=int2a,data4=Norm_2a,part=2a,var=data2a)
%geo(data1=intens3acl,data2=data3a,data3=int3a,data4=Norm_3a,part=3a,var=data3a)
/*
%geo(data1=intens1bcl,data2=data1b,data3=int1b,data4=Norm_1b,part=1b,var=data1b)
%geo(data1=intens1bcl,data2=data2b,data3=int2b,data4=Norm_2b,part=2b,var=data2b)
%geo(data1=intens3bcl,data2=data3b,data3=int3b,data4=Norm_3b,part=3b,var=data3b)
*/

quit;
```

6.1.4 SAS Code to Compare Background Corrected Method with Non-Background Corrected Method

```
%let dir=C:\Christina\Analysis\;
%let macros=SASwork\macros\;

options SASAUTOS="&dir&macros"; options MAUTOSOURCE;

/* DATA IMPORTATION */

/* import intensities data from excel file */
%importation2(type=EXCEL2000,path=&dir,file=data1a.xls,out=intens1a)
%importation2(type=EXCEL2000,path=&dir,file=data2a.xls,out=intens2a)
%importation2(type=EXCEL2000,path=&dir,file=data3a.xls,out=intens3a)

/*INTENSITY PREPARATION - WITHIN SPOT CORRECTION*/

/* change the value of subtract to 'false' if you do not want to
do intensities - background and log the intensity values*/
%prepareintens(data=intens1a,out=intens1acl,var=data1a,substract=false)
%prepareintens(data=intens2a,out=intens2acl,var=data2a,substract=false)
%prepareintens(data=intens3a,out=intens3acl,var=data3a,substract=false)
```

```
%prepareintens(data=intens1a,out=intens1ac,var=data1a,substract=true)
%prepareintens(data=intens2a,out=intens2ac,var=data2a,substract=true)
%prepareintens(data=intens3a,out=intens3ac,var=data3a,substract=true)

/*DO location correction using response surfaces method - for raw
data*/
%rsreg(data=intens1acl, var=data1a, out1=Rpredata1a, out2=Gpredata1a)
%rsreg(data=intens2acl, var=data2a, out1=Rpredata2a, out2=Gpredata2a)
%rsreg(data=intens3acl, var=data3a, out1=Rpredata3a, out2=Gpredata3a)

/*DO location correction using response surfaces method - for
background corrected data*/
%rsreg(data=intens1ac, var=data1a, out1=Rpredata1ap, out2=Gpredata1ap)
%rsreg(data=intens2ac, var=data2a, out1=Rpredata2ap, out2=Gpredata2ap)
%rsreg(data=intens3ac, var=data3a, out1=Rpredata3ap, out2=Gpredata3ap)

/*PREPARE location corrected data - for raw data*/
%dataset(data1=Rpredata1a, data2=Gpredata1a, data3=intens1acl, var1=data1a, var2=_data1a,
        out1=Rdata1a, out2=Gdata1a, out3=data1a, del_neg=no)
%dataset(data1=Rpredata2a, data2=Gpredata2a, data3=intens2acl, var1=data2a, var2=_data2a,
        out1=Rdata2a, out2=Gdata2a, out3=data2a, del_neg=no)
%dataset(data1=Rpredata3a, data2=Gpredata3a, data3=intens3acl, var1=data3a, var2=_data3a,
        out1=Rdata3a, out2=Gdata3a, out3=data3a, del_neg=no)

/*PREPARE location corrected data - for background corrected
data*/
%dataset(data1=Rpredata1ap, data2=Gpredata1ap, data3=intens1ac, var1=data1a, var2=_data1ap,
        out1=Rdata1ap, out2=Gdata1ap, out3=data1ap, del_neg=no)
%dataset(data1=Rpredata2ap, data2=Gpredata2ap, data3=intens2ac, var1=data2a, var2=_data2ap,
        out1=Rdata2ap, out2=Gdata2ap, out3=data2ap, del_neg=no)
%dataset(data1=Rpredata3ap, data2=Gpredata3ap, data3=intens3ac, var1=data3a, var2=_data3ap,
        out1=Rdata3ap, out2=Gdata3ap, out3=data3ap, del_neg=no)

%All(data=All1a, set1=data1a, set2=data1ap)
%All(data=All2a, set1=data2a, set2=data2ap)
%All(data=All3a, set1=data3a, set2=data3ap)

%scatter1(data=All1a, var1=data1a, var2=data1ap)
%scatter1(data=All2a, var1=data2a, var2=data2ap)
%scatter1(data=All3a, var1=data3a, var2=data3ap)

/*Plot of Experiment 1 in a better scale*/
%scat1(data=All1a, var1=data1a, var2=data1ap)

quit;
```

6.1.5 SAS Code for Actual vs. Predicted Values

```
%let dir=C:\Thesis\Analysis\;
%let macros=SASwork\macros\;

options SASAUTOS="&dir&macros"; options MAUTOSOURCE;

/*PROGPRINb1_*/ /*take average of slide 1 and 2*/
%averatio(set=CDa)

/* PLOT SURFACES FOR NORMALIZED EXPERIMENTS */
%surfaces(plan=plangoa, surfacei=surfai, surfaceii=surfaii,
          surfaceiii=surfaiii, var=ratio)

%preact(data=preact, actual=surfai, predict=surfaii)

/*To get min, max, q1, q3 of actual and predicted value*/
%range(data=preact)

quit;
```

6.2 SAS Code for Macros

importation2

```
%MACRO importation(type=,path=, file= ,out= );

PROC IMPORT OUT= WORK.&out
        DATAFILE= "&path&file"
        DBMS=&type REPLACE;
     GETNAMES=YES;
RUN;

%MEND importation;

%MACRO importation2(type=, path=, file= ,out= );
/* do nothing if the dataset already exists */
%if %sysfunc(exist(&out)) %then
%put Data set &out does exist.;
%else
%importation(type=&type,path=&path, file=&file,out=&out)
%MEND importation2;
```

preparation

```
%MACRO preparation(data=, out=);

/* count the number of cc, mf, bp for each gene */ DATA &out;
    set &data;
    keep acc cc bp mf;
    bp=0;
    cc=0;
    mf=0;
    DO z=f2, f5, f8 ,f11, f14, f17, f20, f23, f26, f29, f32,
         f35, f38, f41, f44, f47, f50, f53, f56, f59, f62, f65;
        if z='mf' then mf=mf+1;
        if z='cc' then cc=cc+1;
        if z='bp' then bp=bp+1;
    end;
run;

proc sort;
    by acc;
run; /* delete repeats */ data &out;
    set &out end=final;
    by acc;
    if last.acc;
run;
%MEND preparation;
```

correspondance

```
%MACRO correspondance(data=, out=, plan=);

title1 "Correspondence analysis of categories - &plan"; /* for
HTML file output*/ *ods listing close; options nodate nonumber;
*ods html body="\corresp-&plan-.htm" path="&chemin&output";

/* correspondence analysis for categories*/ proc corresp
data=&data profile=row norow=print /*noprint*/ /*observed*/
/*cellchi2*/ /*rp cp*/
        outc=&out;
    var bp cc mf;
    id acc;
run;

*ods html close; *ods listing; /* saving the bidimensional graph
of genes */ Data &plan;
    set work.&out;
    keep Acc Dim1 Dim2;
    if _n_=1 then delete;
    if acc='cc' then delete;
    if acc='mf' then delete;
```

```
    if acc='bp' then delete;
run ;

Proc sort;
    by acc;
run;

/*%plotit(data=&out, datatype=corresp, plotvars=Dim1 Dim2)*/
/* ploting */ data travail ;
    set &out;
    keep text size xsys ysys x y;
    x=dim1;
    y=dim2;
    If _TYPE_='INERTIA' then delete;
    xsys='2';
    ysys='2';
    text=acc;
    if text NE 'bp' and text NE 'cc' and text NE 'mf' then
        size=0;
    if text = 'bp' or text = 'cc' or text = 'mf' then
        size=2;
    label y ='Dimension 2'
        x='Dimension 1';
run;

proc gplot data=travail;
    symbol1 V=X;
    plot y*x=1 / annotate=travail frame
        href=0 vref=0  ;
run;

proc datasets library=work;
    delete travail;
run;
%MEND correspondance;
```

prepareintens

```
%MACRO prepareintens(data=,out=,var=,substract=);

DATA &out;
    set &data;
    keep acc X_Location Y_Location R&var G&var;

    /* deleting junk genes */
    if acc='N/A2' or acc='N/A1' or acc='' or acc='g' then delete;

    %if &substract=true %then
    %do;
    %put Background substract for &data ;
    /* Subtract background */
    if R&var-Rb&var<=0 then delete;
    if G&var-Gb&var<=0 then delete;
    R&var=log(R&var-Rb&var)/log(2);
    G&var=log(G&var-Gb&var)/log(2);
    /*delete negative intensities*/
    %end;

    %if &substract=false %then
    %do;
    R&var=log(R&var)/log(2);
    G&var=log(G&var)/log(2);
    %end;
run;

PROC sort data=&out;
    by acc;
run;

%MEND prepareintens;
```

rsreg

```
%MACRO rsreg(data=, var=, out1=, out2=);

proc rsreg data=&data out=&out1;
    model R&var= X_Location Y_Location/predict lackfit;
run;

proc rsreg data=&data out=&out2;
    model G&var= X_Location Y_Location/predict lackfit;
run;

%MEND rsreg;
```

dataset

```
%MACRO dataset(data1=, data2=, data3=, var1=, var2=, out1=,
                out2=, out3=, del_neg=);

data &out1; set &data1;
    retain obs 0;
    obs+1;
    keep obs X_Location Y_Location R&var1 Rpre&var1;
    Rpre&var1=R&var1;
run;

data &out1; set &out1;
    drop R&var1;
run;

data &out2; set &data2;
    retain obs 0;
    obs+1;
    keep obs X_Location Y_Location G&var1 Gpre&var1;
    Gpre&var1=G&var1;
run;

data &out2; set &out2;
    drop G&var1;
run;

data &data3; set &data3;
    retain obs 0;
    obs+1;
run;

data &out3;
    merge &out1 &out2 &data3;
    by obs;
run;

data &out3; set &out3;
    keep obs acc X_Location Y_Location R&var2 G&var2;
    R&var2=Rpre&var1-R&var1;
    G&var2=Gpre&var1-G&var1;
run; /* data &out3; set &out3; keep obs acc X_Location Y_Location
R&var2 G&var2 logR&var2 logG&var2 M&var2 A&var2;

%if &del_neg=no %then
%do;
    if R&var2<=0 then logR&var2 = log(-R&var2)/log(2);
    else logR&var2 = log(R&var2)/log(2);
    if G&var2<=0 then logG&var2 = log(-G&var2)/log(2);
    else logG&var2 = log(G&var2)/log(2);
    if R&var2/G&var2<=0 then M&var2=log(-R&var2/G&var2)/log(2);
    else M&var2=log(R&var2/G&var2)/log(2);
    if R&var2*G&var2<=0 then A&var2=log(sqrt(-R&var2*G&var2))/log(2);
    else A&var2=log(sqrt(R&var2*G&var2))/log(2);
    if acc='' or acc='g' then delete;
%end;
```

```
%if &del_neg=yes %then
%do;
    if R&var2<=0 then delete;
    if G&var2<=0 then delete;
    logR&var2 = log(R&var2)/log(2);
    logG&var2 = log(G&var2)/log(2);
    M&var2=log(R&var2/G&var2)/log(2);
    A&var2=log(sqrt(R&var2*G&var2))/log(2);
    if acc='' or acc='g' then delete;
%end;

run: */ proc sort data=&out3;
    by acc;
run;

proc datasets library=work;
    delete &data1 &data2 &out1 &out2;
    run;
quit;
%MEND dataset;
```

ab2

```
%MACRO ab2(set=, var=, out=);

data temp1; set &set; R1=R_&var; G1=G_&var; keep acc X_Location
Y_Location R1 G1; run;

data temp2; set &set; R2=R_&var; G2=G_&var; X2=X_Location;
Y2=Y_Location; keep acc X2 Y2 R2 G2; run;

proc sql;
    create table work.&out as
    select * from work.temp1 inner join work.temp2
    on (temp1.acc = temp2.acc);
run;

data &out; set &out; if (R1=R2) and (G1=G2) and (X_Location=X2)
and (Y_Location=Y2) then delete; run;

proc sort data=&out; by acc; run;

proc gplot data=&out; title "scatter plot R1*R2 - &var"; plot
R1*R2; run; quit;

proc gplot data=&out; title "scatter plot G1*G2 - &var"; plot
G1*G2; run; quit;

proc datasets library=work; delete temp1 temp2; run; quit;

%MEND ab2;
```

prepareab2

```
%MACRO prepareab2(data=, var=);

data &data; set &data; keep acc X_Location Y_Location R1 G1 R2 G2
SubR_&var SubG_&var AddR_&var AddG_&var AbsR_&var AbsG_&var
SqrtR_&var SqrtG_&var;

SubR_&var = R1 - R2; SubG_&var = G1 - G2; AddR_&var = R1 + R2;
AddG_&var = G1 + G2; AbsR_&var = abs(SubR_&var);

AbsG_&var = abs(SubG_&var); SqrtR_&var= sqrt(AbsR_&var);
SqrtG_&var= sqrt(AbsG_&var); run;

%MEND prepareab2;
```

plot

```
%MACRO plot(data=, var=);

proc gplot data=&data; title "Scatter Plot of Sub&var vs.
Sum&var";
    plot Sub&var*Add&var;
run;

proc univariate data=&data noprint; title "Histogram of sub&var";
    var sub&var;
    histogram;
run;

proc univariate data=&data noprint; title "Histogram of sqrt&var";
    var Sqrt&var;
    histogram;
run; quit;

%MEND plot;
```

delete1a

```
%MACRO delete1a(data=, set=, var=);

 data &data; set &set;
    if sqrtR_&var >= 1.17 then delete;
    if sqrtG_&var >= 1.20 then delete;
 run;

%MEND delete1a;
```

delete2a

```
%MACRO delete2a(data=, set=, var=);

 data &data; set &set;
    if sqrtR_&var >= 1.20 then delete;
    if sqrtG_&var >= 1.14 then delete;
 run;

%MEND delete2a;
```

delete3a

```
%MACRO delete3a(data=, set=, var=);

 data &data; set &set;
    if sqrtR_&var >= 1.11 then delete;
    if sqrtG_&var >= 1.14 then delete;
 run;

%MEND delete3a;
```

plot1

```
%MACRO plot1(data=,var=);

data temp; set &data; if lag(acc)=acc then delete; run;

proc gplot data=temp; title "scatter plot of R (Cy5)-&var";

plot R1*R2; run;

proc gplot data=temp; title "scatter plot of G (Cy3)-&var";

plot G1*G2; run;
```

```
proc datasets library=work; delete temp; run; quit;

%MEND plot1;
```

compress

```
%MACRO compress(data1=, data2=, out=);

proc sort data=&data1; by X_Location Y_Location; run;

proc sort data=&data2; by X_Location Y_Location; run;

data temp; merge &data1 &data2; by X_location Y_Location; run;
proc sort data=temp;by acc; run; proc sort data=&data1; by acc;
run; proc sort data=&data2; by acc; run;

data &out; set temp; keep acc obs X_Location Y_Location R1 G1;

if R1=" " then delete; if lag(obs)=obs then delete; run;

proc datasets library=work; delete temp; run; quit;

%MEND compress;
```

prepareint2

```
%MACRO prepareint2(data=, set=, var=);

data &data; set &set; keep acc X_Location Y_Location R&var G&var
M&var A&var; R&var=R1; G&var=G1; M&var=R&var-G&var;
A&var=(R&var+G&var)/2; output; run;

proc datasets library=work; delete &set; run; quit;

%MEND prepareint2;
```

RIPlot

```
%MACRO RIplot(data=, var=);

title1 "R-I Scatter Plot - &var"; proc gplot data=&data;
    plot M&var*A&var;
run; quit;

%MEND RIplot;
```

LOESS

```
%let opts=overlay;
%MACRO LOESS(data=, var=);

proc loess data=&data;
    title1"Loess Normalization of &var";
    model M&var=A&var/smooth=0.1 0.2 0.3 0.4 residual;
    ods output OutputStatistics=norm&data;
run;

/*proc print data=Results(obs=5);
     id obs;
   run;*/

goptions  nodisplay; proc gplot data=norm&data;
    title1 "R-I Scatter Plot with LOESS Fit &var";
    by SmoothingParameter;
    plot DepVar*A&var=1 Pred*A&var=2/ &opts name="fit" overlay;
run; quit;

goptions display; proc greplay nofs tc=sashelp.templt
template=l2r2;
   igout gseg;
   treplay 1:fit 2:fit2 3:fit1 4:fit3;
   treplay 1:fit4 2:fit6 3:fit5 4:fit7;
```

```
   treplay 1:fit8 2:fit10 3:fit9 4:fit11;

run; quit;


proc loess data=norm&data;
    title1"Scatter plots of residuals - &var";
    by SmoothingParameter;
    model Residual=A&var/smooth=0.2;
    ods output OutputStatistics=resid&data;
run;

    axis1 label = (angle=90 rotate=0)
    order = (-2.5 to 2.5 by 0.5);
    goptions nodisplay;

proc gplot data=resid&data;
    title1"Residuals Scatter Plot - &var";
    by SmoothingParameter;
    plot  DepVar*A&var Pred*A&var /
        &opts vref=0 lv=2 vm=1 name="resids" overlay;
run; quit;

goptions display; proc greplay nofs tc=sashelp.templt
template=l2r2;
   igout gseg;
   treplay 1:resids 2:resids2 3:resids1 4:resids3;
   treplay 1:resids4 2:resids6 3:resids5 4:resids7;
   treplay 1:resids8 2:resids10 3:resids9 4:resids11;
run; quit;


%MEND LOESS;
```

LOESSDATA

```
%MACRO LOESSDATA(data1=, data2=, var=, pred=, resid=, out1=, out2=,
                SmoothingParameter=);

data &out1;
    set &data1;
    keep A&var &pred &resid;
    %if &SmoothingParameter=0.2 %then
        %do;
        &pred=Pred;
        &resid=Residual;
        %end;
    %else delete;
run;

data &out2;
    set &data2;
    set &out1;
run;

proc sort; by acc; run;

%MEND LOESSDATA;
```

transform

```
%MACRO transform(data=, set=, part=, var=);

data &data; set &set;
    keep acc X_Location Y_Location R&var G&var
         R&part G&part resi&var A&var;
    R&part= (resi&var + 2*A&var)/2;
    G&part= (2*A&var - resi&var)/2;
    by acc;
run;
```

```
%MEND transform;
```

meanorm

```
/* MACRO location and color corrected values*/
%MACRO meanorm(data=,out=,part=);

/* taking the mean of replicates */ DATA &out;
    set &data end=final;
    by acc;
    keep acc Rdata&part Gdata&part Adata&part residata&part
         R&part G&part;

    Rsub&part + Rdata&part;
    Gsub&part + Gdata&part;
    Asub&part + Adata&part;
    resisub&part + residata&part;
    Rsb&part + R&part;
    Gsb&part + G&part;
    nb+1;
    If last.acc;
        Rdata&part=Rsub&part/nb;
        Gdata&part=Gsub&part/nb;
        Adata&part=Asub&part/nb;
        residata&part=resisub&part/nb;
        R&part=Rsb&part/nb;
        G&part=Gsb&part/nb;

        Rsub&part=0;
        Gsub&part=0;
        Asub&part=0;
        resisub&part=0;
        Rsb&part=0;
        Gsb&part=0;
        nb=0;
run;

%MEND meanorm;
```

means

```
%MACRO means(data=, var=);

proc means data=&data mean std;
    var R&var G&var;
run;

%MEND means;
```

standard

```
%MACRO standard(data=, Rmean=, Gmean=, Rstd=, Gstd=, var=);

data &data; set &data;
    keep acc StdR&var StdG&var;
    StdR&var = (R&var-&Rmean)/&Rstd;
    StdG&var = (G&var-&Gmean)/&Gstd;
run;

%MEND standard;
```

mergeintens

```
%MACRO mergeintens(intens1=,intens2=,intens3=,part=,out=);

Data &out;
    merge &intens1 (in=a) &intens2 (in=b) &intens3 (in=c);
    by acc;
    if a and b and c;
run;
```

```
/* creating dataset of unknow genes */ Data unknow&part;
    keep acc;
    merge &out (in=a) plango&part (in=b);
    by acc;
    if a and not b;
run;

%MEND mergeintens;
```

scatter

```
%MACRO scatter(data=, var1=, var2=);
goptions reset=global; proc gplot data=&data;
    title1 "Scatter plot of StdR&var1 vs. StdG&var2";
    plot StdR&var1*StdG&var2;
run; quit;

%MEND scatter;
```

prepare

```
%MACRO prepare(data=, set=, var1=, var2=, part=);

data &data; set &set;
    keep acc StdR&var1 StdG&var1 StdR&var2 StdG&var2 SubH&part
    SubM&part AddH&part AddM&part AbsH&part AbsM&part SqrtH&part
    SqrtM&part;
        SubH&part = StdG&var1 - StdR&var2;
        SubM&part = StdR&var1 - StdG&var2;
        AddH&part = StdG&var1 + StdR&var2;
        AddM&part = StdR&var1 + StdG&var2;
        AbsH&part = abs(SubH&part);
        AbsM&part = abs(SubM&part);
        SqrtH&part= sqrt(AbsH&part);
        SqrtM&part= sqrt(AbsM&part);
run;

%MEND prepare;
```

deleteaa

```
%MACRO deleteaa(data=, set=, var=);

data &data; set &set;
    if sqrtH&var >= 1.20 then delete;
    if sqrtM&var >= 1.08 then delete;
run;

%MEND deleteaa;
```

plot2

```
%MACRO plot2(data=,var1=, var2=);

proc gplot data=&data;
    title "scatter plot of G&var1*R&var2 - Human";
    plot StdG&var1*StdR&var2;
run;

proc gplot data=&data;
    title "scatter plot of R&var1*G&var2 - Monkey";
    plot StdR&var1*StdG&var2;
run; quit;

%MEND plot2;
```

ratio

```
%MACRO ratio(data=, set=, var1=, var2=);

data &data; set &set;
    keep acc &data;
```

```
    &data=Std&var1-Std&var2;
run;

%MEND ratio;
```

surfaces

```
%MACRO surfaces(plan=, surfacei=, surfaceii=, surfaceiii=, var=);

/* merge the dataset of R and the plan of genes */ Data &surfacei;
    keep acc dim1 dim2 R;
    merge &var (in=a) &plan (in=b);
    by acc;
    if a and b;
    R=&var;
run;

/* plot PLAN */ proc g3d data=&surfacei;
    title1 "Scatterplot &var";
    scatter Dim2*Dim1=R / caxis=black;
    scatter Dim2*Dim1=R / rotate=0 caxis=black tilt=0;
    scatter Dim2*Dim1=R / rotate=0 caxis=black tilt=90;
run;

/* RESPONSE SURFACES */
 title1 "Response Surface &var";

/* for HTML file output*/
 *ods listing close;
 options nodate nonumber;
*ods html body="rsreg-&var-.htm" path="&chemin&output";

/* using response surfaces */ proc rsreg data=&surfacei
out=&surfaceii /*noprint*/;
    model R=Dim1 Dim2 / predict;
    id acc;
run;

*ods html close; *ods listing;

proc g3d data=&surfaceii;
    title2 "Scatter";
    scatter Dim2*Dim1=R / caxis=black grid ;
run;

/* plot surface */ proc g3grid data=&surfaceii out=&surfaceiii;
    grid Dim2*Dim1=R / NAXIS1=60 NAXIS2=60 ;
run;

proc g3d data=&surfaceiii;
    title2 "Shape";
    plot Dim2*Dim1=R / caxis=black grid ;
run;

/* plot contour */ proc gcontour;
    title2 "Contour";
    plot Dim2*Dim1=R / caxis=black autolabel;
run;

%MEND surfaces;
```

average

```
%MACRO average(data=, set=, StdR1=, StdG1=, StdR2=, StdG2=, StdR3=,
               StdG3=, StdR12=, StdG12=, StdH12=, StdM12=, StdH3=,
               StdM3=);

data &data; set &set;
    keep acc &StdH12 &StdM12 &StdH3 &StdM3;
    &StdH12=(&StdG1+&StdR2)/2;
    &StdM12=(&StdR1+&StdG2)/2;
```

```
    &StdH3=&StdR3;
    &StdM3=&StdG3;
run; quit;

%MEND average;
```

prepare2

```
%MACRO prepare2(data=, set=, var1=, var2=, part=);

data &data; set &set;
    keep acc StdH&var1 StdM&var1 StdH&var2 StdM&var2 SubH&part
         SubM&part AddH&part AddM&part AbsH&part AbsM&part
         SqrtH&part SqrtM&part;
    SubH&part = StdH&var1 - StdH&var2;
    SubM&part = StdM&var1 - StdM&var2;
    AddH&part = StdH&var1 + StdH&var2;
    AddM&part = StdM&var1 + StdM&var2;
    AbsH&part = abs(SubH&part);
    AbsM&part = abs(SubM&part);
    SqrtH&part= sqrt(AbsH&part);
    SqrtM&part= sqrt(AbsM&part);
run;

%MEND prepare2;
```

deletea33

```
%MACRO deletea33(data=, set=, var=);

data &data; set &set;
    if sqrtH&var >= 1.20 then StdH3&var='';
    if sqrtM&var >= 1.16 then StdM3&var='';
run;

%MEND deletea33;
```

plot22

```
%MACRO plot22(data=,var1=, var2=);

proc gplot data=&data;
    title "scatter plot of H&var1*H&var2 - Human";
    plot StdH&var1*StdH&var2;
run;

proc gplot data=&data;
    title "scatter plot of M&var1*M&var2 - Monkey";
    plot StdM&var1*StdM&var2;
    run;
quit;

%MEND plot22;
```

geo

```
%MACRO geo(data1=, data2=, data3=, data4=, part=, var=);

 /*Geographical plots of log raw data*/
 /*calculation of log(R/G) for geo plot*/
data temp1; set &data1;
    keep acc X_Location Y_Location ratio;
    ratio=R&var-G&var;
run;

proc g3grid data=temp1 out=out1;
    grid Y_location*X_location=ratio
        /naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out1;
    title "Geographical locations - background corrected data";
```

```
    plot Y_location*X_location=ratio /pattern;
run;

 /*Geographical plots of location corrected (rsreg) data */
 /*calculation of log(R/G) for geo plot*/
data temp2; set &data2;
    keep acc X_Location Y_Location ratio;
    ratio=R_&var-G_&var;
run;

proc g3grid data=temp2 out=out2;
    grid Y_Location*X_Location=ratio
        /naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out2;
    title "Geographical locations - geo corrected data";
    plot Y_location*X_location=ratio / pattern;
run;

 /*Geographical plots after eliminating between spots outliers */
 /*calculation of log(R/G) for geo plot*/
data temp3; set &data3;
    keep acc X_Location Y_Location ratio;
    ratio=R&var-G&var;
run;

proc g3grid data=temp3 out=out3;
    grid Y_Location*X_Location=ratio
        /naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out3;
    title "Geographical locations - between spots corrected data";
    plot Y_location*X_location=ratio / pattern;
run;

 /*Geographical plots after Color Normalization (LOESS)*/
 /*calculation of log(R/G) for geo plot*/
data temp4; set &data4;
    keep acc X_Location Y_Location ratio;
    ratio=R&part-G&part;
run;

proc g3grid data=temp4 out=out4;
    grid Y_Location*X_Location=ratio
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out4;
    title "Geographical locations - LOESS corrected data";
    plot Y_location*X_location=ratio / pattern;
run;

proc datasets library=work;
    delete temp1 temp2 temp3 temp4 out1 out2 out3 out4;
run;

%MEND geo;
```

geo1

```
%MACRO geo1(data1=, data2=, part=, var=);

 /*Geographical plots of log raw data*/
proc g3grid data=&data1 out=out1;
    grid Y_location*X_location=R&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out1;
```

```
    title "Geographical locations
           -background corrected data (Red)";
    plot Y_location*X_location=R&var / pattern;
run;

    /*Geographical plots of log raw data*/
proc g3grid data=&data1 out=out1g;
    grid Y_location*X_location=G&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out1g;
    title "Geographical locations
           -background corrected data (Green)";
    plot Y_location*X_location=G&var / pattern;
run;

 /*Geographical plots of location corrected (rsreg) data */
proc g3grid data=&data2 out=out2;
    grid Y_Location*X_Location=R_&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out2;
    title "Geographical locations - geo corrected data (Red)";
    plot Y_location*X_location=R_&var / pattern;
run;

 /*Geographical plots of location corrected (rsreg) data */
proc g3grid data=&data2 out=out2g;
    grid Y_Location*X_Location=G_&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out2g;
    title "Geographical locations - geo corrected data (Green)";
    plot Y_location*X_location=G_&var / pattern;
run;

proc datasets library=work;
    delete out1 out1g out2 out2g;
run;

%MEND geo1;
```

geo2

```
%MACRO geo2(data1=, data2=, data3=, data4=, part=, var=);

 /*Geographical plots of log raw data*/
proc g3grid data=&data1 out=out1;
    grid Y_location*X_location=R&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out1;
    title "Geographical locations - background corrected data";
    plot Y_location*X_location=R&var / pattern;
run;

 /*Geographical plots of location corrected (rsreg) data */
proc g3grid data=&data2 out=out2;
    grid Y_Location*X_Location=R_&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out2;
    title "Geographical locations - geo corrected data";
    plot Y_location*X_location=R_&var / pattern;
run;
```

```
/*Geographical plots after eliminating between spots outliers*/
proc g3grid data=&data3 out=out3;
    grid Y_Location*X_Location=R&var
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out3;
    title "Geographical locations
           - between spots corrected data";
    plot Y_location*X_location=R&var / pattern;
run;

 /*Geographical plots after Color Normalization (LOESS)*/
proc g3grid data=&data4 out=out4;
    grid Y_Location*X_Location=R&part
        / naxis1=1000 naxis2=1000 join;
run;

proc  gcontour data=out4;
    title "Geographical locations - LOESS corrected data";
    plot Y_location*X_location=R&part / pattern;
run;

proc datasets library=work;
    delete out1 out2 out3 out4;
run;

%MEND geo2;
```

All

```
%MACRO All(data=, set1=, set2=);

Proc sort data=&set1 out=temp1;
    by X_Location Y_Location;
run;

Proc sort data=&set2 out=temp2;
    by X_Location Y_Location;
run;

data &data;
    merge temp1 temp2;
    keep X_Location Y_Location R_&set1 G_&set1 R_&set2 G_&set2;
    by X_Location Y_Location;
run; quit;

proc datasets library=work;
    delete temp1 temp2;
run;

%MEND All;
```

scatter1

```
%MACRO scatter1(data=, var1=, var2=);

goptions reset=global; proc gplot data=&data;
    title "Background corrected vs raw data";
    title1 "Scatter plot of R_&var1 vs. R_&var2";
    plot R_&var1*R_&var2;
run;

proc gplot data=&data;
    title "Background corrected vs raw data";
    title1 "Scatter plot of G_&var1 vs. G_&var2";
    plot G_&var1*G_&var2;
run;

quit;

%MEND scatter1;
```

scat1

```
%MACRO scat1(data=, var1=, var2=);

goptions reset=global; proc gplot data=&data;
    title "Background corrected vs raw data";
    title1 "Scatter plot of R_&var1 vs. R_&var2";
    plot R_&var1*R_&var2/   haxis=-7 to 7 by 1
                            vaxis=-7 to 2 by 1;
run;

proc gplot data=&data;
    title "Background corrected vs raw data";
    title1 "Scatter plot of G_&var1 vs. G_&var2";
    plot G_&var1*G_&var2/   haxis=-7 to 7 by 1
                            vaxis=-7 to 2 by 1;
run; quit;

%MEND scat1;
```

averatio

```
%MACRO averatio(set= );

data ratio; set &set;
    keep acc StdH StdM ratio;
    StdH=(StdG1a+StdR2a)/2;
    StdM=(StdR1a+StdG2a)/2;
    ratio=StdH-StdM;
run; quit;

%MEND averatio;
```

preact

```
%MACRO preact(data=, actual=, predict=);

data temp1; set &actual;
    keep acc actual;
    actual=R;
run;

data temp2; set &predict;
    keep acc predict;
    predict=R;
run;

data &data; merge temp1 temp2; by acc;
    keep acc actual predict;
run;

proc gplot data=&data;
    title "Predict vs. Actual";
    plot predict*actual;
run;

proc gplot data=&data;
    title "Predict vs. Actual";
plot predict*actual /   haxis = -3 to 5 by 1
                        vaxiz = -3 to 5 by 1;
run;

proc datasets library=work;
    delete temp1 temp2;
run; quit;

%MEND preact;
```

range

```
%MACRO range(data=);

proc univariate data=&data;
   var actual predict;
run;

%MEND range;
```

Bibliography

[1] Online Medical Dictionary: http://cancerweb.ncl.ac.uk/omd/

[2] A.Brazma, P.Hingamp, J.Quackenbush, G.Sherlock, P.Spellman, C.Stoeckert, J.Aach, W.Ansorge, C.A.Ball, H.C.Causton, T.Gaasterland, P.Glenisson, F.C.P.Holstege, I.F.Kim, V.Markowitz, J.C.Matese, H.Parkinson, A.Robinson, U.Sarkans, S.Schulze-Kremer, J.Stewart, R.Taylor, J.Vilo and M.Vingron. Minimum Information About a Microarray Experiment (MIAME) - Toward Standards For Microarray Data. Nature Genetics, Vol. 29: 365-371, Dec. 2001.

[3] Mark Schena. Microarray Analysis. John Wiley, New Jersey 2003.

[4] QuantArray Software, PerkinElmer Life Sciences Inc., Boston, MA, USA.

[5] M.Ashburner, C.A. Ball, J.A. Blake, D. Botstein, et al. Gene Ontology: tool for the unification of biology. Nature Genetics, Vol. 25: 25-29, 2000.

[6] Sorin Draghici. Data Analysis Toold For DNA Microarrays. CRC Press, London, UK 2003.

[7] J.Schuchhardt, D.Beule, E.Wolski, and H.Eickhoff. Normalization strategies for cDNA microarrays. Nucleic Acids Research, Vol. 28, No. 10: e47i-e47v, May 2000.

[8] Y.H. Yang, S. Dudoit, P. Luu and T.P. Speed. Normalization for cDNA Microarray Data. SPIE BiOS, San Jose, California, January 2001.

[9] T.Beiβbarth, K.Fellenberg, B.Brors, R.Arrbas-Prat, J.M.Boer, N.C.Hauser, M.Scheideler, J.D.Hoheisel, G.Schütz, A.Poustka and M.Vigron. Processing and Quality Control of DNA Array Hybridization Data. Bioinformatics, Vol. 16, No. 11: 1014-1022, 2000.

[10] Dov Stekel. Microarray Bioinformatics. Cambridge University Press, UK 2003.

[11] Jin Hyuk Kim, Dong Mi Shin and Yong Sung Lee. Effect of Local Background Intensities in the Normalization of cDNA Microarray Data With a Skewed Expression Profiles. Experimental and Molecul Medicine, Vol. 34, No. 3, 224-232, July 2002.

[12] R. Sasik, E.Calvo and J.Carbeil. Statistical Analysis of High-density Oligonucleotide Arrays: A Multiplicative Noise Model. Bioinformatics, Vol. 18, No. 12, 1633-1640, Dec. 2002.

[13] D.L.Wilson, M.J.Buckley, C.A.Helliwell and W.Wilson. New Normalization Methods for cDNA Microarray Data. Bioinformatics, Vol. 19, No. 11, 1325-1332, 2003.

[14] A.I.Khuri and J.A.Cornell. Response Surfaces: Designs And Analyses. New York: Marcel Dekker, 1996.

[15] SAS 9.1 2003, SAS Institute Inc., Cary, NC, USA.

[16] John Quackenbush. Computational Analysis of Microarray Data. Nature Genetics, Vol. 2: 418-427, June 2001.

[17] W.Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, Vol. 74: 829-836, 1979.

[18] W.Cleveland and S.Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association, Vol. 83: 596-610, 1983.

[19] H.Kishino and P.J.Waddel. Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data. Genome Informatics, Vol. 11: 83-95, 2000.

[20] K.Fellenberg, N.C.Hauser, B.Brors, A.Neutzner, J.D.Hoheisel and M.Vingron. Correspondence Analysis Applied to Microarrya Data. PNAS Genetics, Vol. 98, No. 19: 10781-10786, September 11, 2001.

[21] A.C.Culhane, G.Perrière, E.C.Considinem, T.G.Cotter and D.G.Higgins. Between-Group Analysis of Microarray Data. Bioinformatics, Vol. 18, No. 12:1600-1608, 2002.

[22] J.F.Hair, R.E.Anderson, R.L.Tatham, W.C.Black. Multivariate Data Analysis, 5th Edition. Prentice Hall, New Jersey, 1998.

[23] C.J.Stoeckert Jr., H.C.Causton and C.A.Ball. Microarray Databases: Standards and Ontologies. Nature Genetics Supplement, Vol. 32: 469-473, Dec. 2002.

[24] A.M.Campbell and L.J.Heyer. Discovering Genomics, Proteomics, and Bioinformatics. Benjamin Cummings, 2003.

[25] http://genome-www.stanford.edu/

[26] http://www.geneontology.org/

[27] G.Stephanopoulos, D.Hwang, W.A.Schmitt, J.Misra and G.Stephanopoulos. Mapping Physiological States From Microarray Expression Measurements. Bioinformatics, Vol. 18, No. 8: 1054-1063, 2002.

[28] D.R. Masys. Database Designs For Microarray Data. The Pharmacogenomics Journal, Vol. 1, No. 4: 232-233, 2001.