COMPARATIVE STATISTICAL MICROARRAY ANALYSIS OF YEAST DATA UNDER HEAT SHOCK STRESS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

DUYGU VAROL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

JULY 2014

Approval of the thesis:

COMPARATIVE STATISTICAL MICROARRAY ANALYSIS OF YEAST DATA UNDER HEAT SHOCK STRESS

submitted by **DUYGU VAROL** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İnci Batmaz _____
Head of Department, **Statistics**

Assoc. Prof. Dr. Vilda Purutçuoğlu _____
Supervisor, **Statistics Department, METU**

Assoc. Prof. Dr. Remziye Yılmaz _____
Co-supervisor, **MBB R&D Center, METU**

Examining Committee Members:

Prof. Dr. Füsun İnci Eyidoğan
Institute of Educational Sciences, Başkent University _____

Assoc. Prof. Dr. Vilda Purutçuoğlu
Supervisor, Statistics Department, METU _____

Assoc. Prof. Dr. Remziye Yılmaz
Co-Supervisor, MBB R&D Center, METU _____

Prof. Dr. Gerhard-Wilhelm Weber
Institute of Applied Mathematics, METU _____

Assoc. Prof. Dr. Özlem İlk Dağ
Statistics Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: DUYGU VAROL

Signature :

ABSTRACT

COMPARATIVE STATISTICAL MICROARRAY ANALYSIS OF YEAST DATA UNDER HEAT SHOCK STRESS

Varol, Duygu

M.S., Department of Statistics

Supervisor : Assoc. Prof. Dr. Vilda Purutçuoğlu

Co-Supervisor : Assoc. Prof. Dr. Remziye Yılmaz

July 2014, 92 pages

The microarray technology is one of the widely used experimental methods in biological and biochemical sciences. By this innovation, a number of genes can be analyzed simultaneously by means of statistical methods. Hereby in this study we analyze a new one-channel microarray dataset that is measured to investigate the changes in heat shock stress of yeast. The data that are generated in the Molecular Biology and Biotechnology R-D Center at the Middle East Technical University has not been evaluated yet in different researches. Hence in this study we perform detailed comparative analyses of these measurements and critically assessed the biological findings. For this purpose, in the thesis, we implement the normalization, the detection of differentially expressed genes, multiple testing under different error rates, clustering and the search of gene annotation as well as pathway analyses by comparing the most well-known approaches in each step. Finally, the biological results are evaluated to get new knowledge about the yeast under changes in temperature.

Keywords: One-channel Microarray Analysis, Comparative Normalization, Gene Clustering, Yeast Data, Heat-Shock Pathway.

ÖZ

ISI ŞOKU STRESİ ALTINDAKİ EKMEK MAYASI TEK - KANALLI MİKRODİZİN VERİSİNİN KARŞILAŞTIRMALI İSTATİSTİKSEL ANALİZİ

Varol, Duygu
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi : Doç. Dr. Vilda Purutçuoğlu
Ortak Tez Yöneticisi : Doç. Dr. Remziye Yılmaz

Temmuz 2014 , 92 sayfa

Mikrodizin teknolojisi, biyoloji ve biyokimya bilimlerinde en çok kullanılan deneysel metotlardan birisidir. Bu yenilik sayesinde birçok gen istatistiksel yöntemler aracılığıyla eş zamanlı olarak analiz edilebilmektedir. Dolayısıyla bu çalışmada ekmek mayasının sıcaklık şoku stresi altındaki değişimini incelemek için ölçülen yeni bir tek-kanallı mikrodizin veri setini analiz etmekteyiz. Orta Doğu Teknik Üniversitesine ait Moleküler Biyoloji ve Biyoteknoloji AR-GE Merkezinde toplanan bu veri, henüz hiçbir farklı araştırmada değerlendirilmemiştir. Dolayısıyla bu çalışmada, bu ölçümlerin detaylı, karşılaştırmalı analizini uygulamakta ve biyolojik bulguları sorgulayıcı biçimde değerlendirmekteyiz. Bu amaçla normalizasyon, farklı ekspresyonlu genlerin bulunması, değişik hata oranları altında çoklu test, sınıflandırma ve gen açıklaması araştırması ve de yolak analizi araştırması, her bir adımdaki en iyi bilinen yaklaşımlarla karşılaştırma yaparak uygulanmaktadır. Son olarak biyolojik sonuçlar sıcaklık değişimi altında ekmek mayası

hakkında yeni bilgiler bulmak amacıyla değerlendirilmektedir.

Anahtar Kelimeler: Tek-kanallı Mikrodizin Analizi, Karşılaştırmalı Normalizasyon, Gen Sınıflaması, (Ekmek) Maya Verisi, Isı-Şoku Yolağı.

To my family

ACKNOWLEDGMENTS

I would like to thank to my supervisor Assoc. Prof. Dr. Vilda Purutçuoğlu for her guidance, patience, sincereness, encouragement and supervision throughout this thesis study. It has been a great chance for me to work with her. I am also deeply indebted to my co-supervisor Assoc. Prof. Dr. Remziye Yılmaz with whom we carried out the biological part of my thesis. She was very helpful during thesis period and her innovative ideas impressed me a lot.

I would like to thank the Middle East Technical University, Molecular Biology and Biotechnology R&D Center as they provide us to use the microarray yeast data in this study.

I also compassionately express my special thanks to Ezgi Ayyıldız, Çiğdem Güngör, Gonca Mert and Serap Görgü for their motivation and nice friendship. I would like to extend my thanks to my beloved homemate Seher Gök for her support and friendship.

I would like to send my ultimate appreciation to my parents and my dear sister Damla for their endless patience, encouragement, support and love.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
a.s.w	Average Silhouette Width
C	Control Group
cDNA	Complementary Deoxyribonucleic Acid
CMARS	Conic Multivariate Adaptive Regression Splines
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic Acid
EMBL	European Molecular Biology Laboratory
FC	Fold-Change
FDR	False Discovery Rate
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FGED	Functional Genomics Data
FWER	Familywise Error Rate
GC-RMA	Robust Microarray Analysis Based on GC Content
GO	Gene Ontology
HIPAM	Hierarchical Partitioning Around Medoids
HOPACH	Hierarchical Ordered Partitioning and Collapsing Hybrid
HSP	Heat Shock Protein
ID	Identity
LS	Least Square
MA	Moving Average
MARS	Multivariate Adaptive Regression Splines
MAS5.0	Microarray Suite Software
MBB	Molecular Biology and Biotechnology
MBEI	Model Based Gene Expression Index
MeSH	Medical Subject Headings

MGED	Microarray Gene Expression Data
MIAME	Minimum Information About a Microarray Experiment
MM	Mismatches
mRNA	Messenger Ribonucleic Acid
PAM	Partitioning Around Medoids
PAMSAM	Partitioning Around Medoids With Sammon Mapping
PCA	Principle Component Analysis
pFDR	Positive False Discovery Rate
pFNR	Positive False Nondiscovery Rate
PM	Perfect Matches
RefSeq	The Reference Sequence
RMA	Robust Microarray Analysis
RMARS	Robustification of Multivariate Adaptive Regression Splines
RNA	Ribonucleic Acid
ROC	Receiving Operating Characteristic
SAM	Sammon Mapping
SGD	Saccharomyces Genome Database
TIGR CMR	The Institute for Genomic Research Comprehensive Microbial Resource
TIGR GI	The Institute for Genomic Research Gene Indices
TN	True Negative
TP	True Positive
T1	Treatment 1 Heat Shock
T2	Treatment 2 Heat Change
YFGdb	Yeast Functional Genomics Database

CHAPTER 1

INTRODUCTION

1.1 Microarray

The genetic material of an organism is called as genome. Although each cell of an organism includes identical genom, the active genes are valid in the cells of the organism. For this reason, the having knowledge of the genes that are active has a fundamental and vital role in biological science. Microarray analysis is a powerful implement used for obtaining knowledge about the molecular sources of biological problems for the past decades. It is used to understand whether the gene is active. Although the initial idea of a microarray experiment dates back to earlier, Mark Schena and his colleagues developed the microarrays at Stanford University in early 1990s and the result of the first microarray analysis experiment obtained from the yeast data by Affymetrix technology was represented in 1994 (Schena, 2003). The advantage of the microarray technology is that thousands of genes can be examined in one microarray at the same time. The most studied microarray which belongs to one of the human, mouse, rat or yeast organism are the mass-produced. Also the companies can produce private microarray for special researches on request and Affymetrix is the most well-known microarray producer (Wit and McClure, 2004). In Figure 1.1, a picture of the Affymetrix microarray is presented for illustration.

Hereby the microarray is a small chip which is produced from nylon membrane, silicon or glass. There are thousands of probes on the microarray surface in rows and columns in order as shown simply in Figure 1.2. Each probe includes spotted strands of the Deoxyribonucleic Acid (DNA). The strands of DNA on

the probe are identical and refer a gene. That is, each probe represents one gene. The location of each gene on the array surface is known and recorded in advance (Schena, 2003; Stekel, 2003).

On the other hand to gather the measurements from genes, the following procedure is used. Since the messenger Ribonucleic Acid (mRNA) molecule transfers the required genetic code from DNA to the ribosome organelle for the new protein synthesis, in the process of a typical microarray experiment, initially the mRNA molecules obtained from the cell at issue are labeled with an enzyme which produces the complementary DNA (cDNA). Thus fluorescent nucleotides are attached to cDNA. When cDNA binds to its complement on the convenient probe, its fluorescent tag is stood and colors. Then, the microarray slide is scanned and the fluorescent intensities of probes are measured (Schena, 2003).

There are two types of microarrays if we classify them in terms of their structures. These are the one-channel and two-channel microarrays. In the one-channel microarrays, each condition is implemented in every single array. On the other side in the two-channel microarrays, two conditions are studied simultaneously in a single array and the dye effect is utilized to separate the observed signals under different conditions.

Figure 1.1: Affymetrix GeneChip.

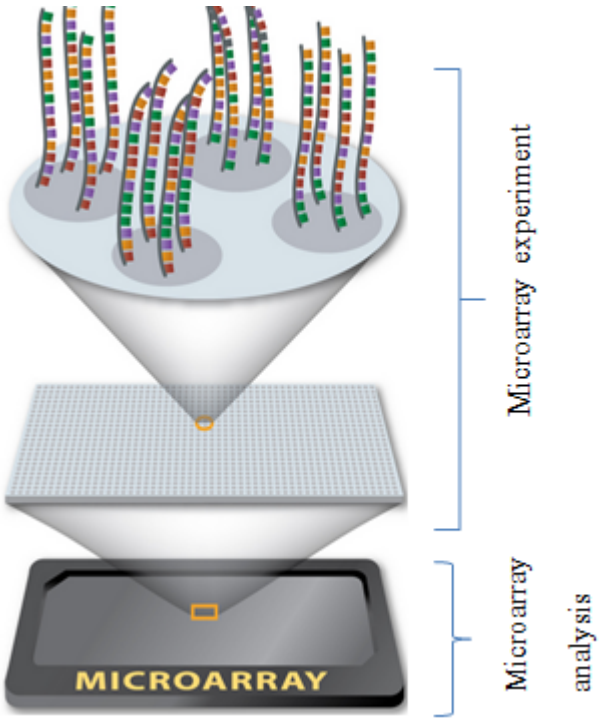

Figure 1.2: Structure of the DNA microarray (Learn.Genetics, 2014).

1.2 Aim of the Study

The aim of the study is to investigate the yeast data on hand in detailed microarray studies in order to find biologically significant results. Accordingly we detect differentially expressed genes under different conditions and then classify these active genes according to their features. In this process, all required statistical methods are performed comparatively in each step and the best performed ones are used in further analyses. Hence, as the first step in the analysis different background normalization algorithms such as MAS5.0, RMA, MBEI and GC-RMA are applied and the algorithm which gives the most accurate findings is chosen for the normalization procedure. In order to determine differentially expressed genes, two different methods are executed. Then the fold-change method for varied cut-off values is implemented and the results are recorded. Later the hypothesis testing, as the other method, is fulfilled by using different multiple testing procedures and the outcomes are saved. Finally, the list of genes recorded

as differentially expressed is investigated in terms of their biological validation and the cluster analysis is performed to classify these genes according to their similarities. In short, the microarray analysis of the yeast data under certain stress is performed and the biologically significant results are represented.

CHAPTER 2

MICROARRAY DATA ANALYSIS

The microarray datasets produced from different users need to be utilizable and comparable among scientist and software developers. Hereby the Microarray Gene Expression Data (MGED) Society is a cooperation which is founded for this goal (Moreau et al., 2003). Indeed it carries on as the Functional Genomics Data (FGED) Society and it is not only interested in microarray data today. On the other hand there is a proposal called the Minimum Information About a Microarray Experiment (MIAME) suggested by MGED in order to develop a standard for microarray experiments from their experimental designs to their normalizations. By means of this proposal, the data are saved with required sufficient information for the interpretation and comparison (Brazma et al., 2001). There are different data sources for gene sequences which are also used in the microarray analysis. The major international databases for the gene sequence hold all published gene sequences. The GenBank of America, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) can be supposed as primary ones. The other important databases can be referred as UniGene, TIGR GI, RefSeq for the secondary gene sequence databases and Ensemble, TIGR CMR, SGD for the genomic databases (Stekel, 2003). But the data used in this study are generated at the Middle East Technical University, Molecular Biology and Biotechnology R&D Center and it is already uploaded in Yeast Functional Genomics Database (YFGdb) for public users.

In order to analysis microarray data there exist different software packages. For example, Microsoft Excel can be used for this analysis. Apart from Excel, there are other advanced statistical softwares too such as R and Matlab. Here the first

one is mostly preferable as it is freely downloadable. Moreover, there are some softwares which are written especially for analyzing microarray data and respect the MIAME procedure. These softwares can be exemplified such as GeneSpring, J-Express and Expression Profiler. GeneSpring enables the users to make statistical analyses and biological understanding (Stekel, 2003). Hereby in different steps of this study, Agilent GeneSpring 12.1 and R 2.15.2 are performed.

2.1 Microarray Data

A microarray chip includes different types of genes which have special functions. Although probes have almost all investigated genes, there are also some genes which are utilized for the control such as *housekeeping* and *spike* genes (Schena, 2003; Wit and McClure, 2004; Stekel, 2003). The housekeeping genes are used to separate the cell function of tissues as they are common in every cell. Hence they are applied for normalization as their expressions are almost the same in different types of tissues. Another control probes are called as the spike genes. These genes are used to control the sample preparation and hybridization steps. As their expression levels are known, it is checked whether they have expected intensity levels. The control list of these genes for our yeast data can be seen in Table 3.1. In this study all of these genes are considered in the analysis.

2.2 Normalization

The microarray experiments investigate differential expressions of genes by using the transcribed mRNA. In the transcription step of DNA while the gene expression occurs, the DNA's genetic code is transmitted via RNA. In this process the amount of RNA is recorded to learn about the expressed genes. In other words, the microarray experiment aims to understand whether a gene is active under certain conditions by measuring the amount of transcribed RNA (Schena, 2003). In a microarray experiment, there are two sources of errors, namely, random error and systematic artifacts (Ewens, 2005). The random error is inhered in the data, whereas, the systematic error does not. Hereby the systematic error can

arise from different sources such as nonspecific hybridization, background signal or dye effect used in labelling the genes. Before any statistical analyses, the normalization techniques are applied to obtain as possible as pure data cleared from the systematic bias. By this way the erroneous signals added in the measurements during the sample preparation, array fabrication or hybridization can be excluded. Thereby the artificial bias in the data decreases.
In the traditional statistical point of view, the causes of systematic bias can be described in a regression model. However, this perspective is not functional as it is complicated and computationally demanding (Wit and McClure, 2004). Hereby the alternative approach is to model each normalization step separately. There are different levels of normalization in microarray experiments and when one tries to eliminate systematic artifacts, an ordered procedure is suggested to avoid the larger bias. The proposal order can be listed as follows.

1. Spatial correction,
2. Background correction,
3. Dye-effect correction (if it is a two-channel microarray analysis),
4. Within replicate rescaling,
5. Across-conditions rescaling.

The Affymetrix GeneChip is the most familiar one-channel microarray as shown in Figure 1.1. Each probe on the Affymetrix GeneChip has a piece of a certain gene and this gene-piece has a 25-base pair long that is also called oligonucleotide. In the Affymetrix GeneChip, there are 11 to 20 probes pairs per array. Moreover each probe pair consists of two components, namely, perfect matches and mismatches. It is assumed that the former captures the true transcribed intensity level and the latter collects the noisy signals nested in true signals. In terms of the structure of these two components, the mismatches are generated by only changing the 13th base of the perfect matches in the 25-base long oligonucleotide (Gentleman et al., 2005). The details for measuring the true intensities under such probe pairs are represented in Subsection 2.2.2 (Background Normalization). On the other hand, in the following parts, we initially

describe each normalization step and represent the most well-known methods in every step with their mathematical descriptions. Then we choose the most appropriate methods within each step to analyze a new dataset that is measured to evaluate the heat shock in yeast genes.

2.2.1 Spatial Normalization

The spatial normalization technique is used to remove the systematic bias arising from the location of probes on the array or the production phase of arrays. There are different approaches for this normalization process (Wit and McClure, 2004). The spatial smoothing is the most common approach among alternatives. In this method, according to the distinct part of the real data, a local linear trend surface is obtained. This smooth surface M is substracted from the actual data S. By this way the spatial bias is put away from the data and the location smoothed surface is found via

$$S_m(x, y) = S(x, y) - M(x, y), \tag{2.1}$$

where x and y are row and column coordinates of probes, respectively (Wit and McClure, 2004). Considering the regions of arrays and the variability of them, another smoothing is necessary for the data. Thus in the second step, the location scale parameter is calculated for each array and a standardization procedure is applied by dividing the located smoothed surface S_m into this parameter S_c via

$$S_{ms}(x, y) = \frac{S(x, y) - M(x, y)}{S_c(x, y)}. \tag{2.2}$$

After this correction, the data do not have the first and the second order spatial bias. If it is worried that these corrections change the scale of original data, Wit and McClure (2004) suggest to perform the median by the following expression.

$$S_{ms}^{original}(x, y) = S_{ms}(x, y) \times \text{median}S_c(x, y) + \text{median}S(x, y). \tag{2.3}$$

In this normalization step, all calculations are implemented under the log scale.

2.2.2 Background Normalization

The main aim in the microarray experiment is to obtain the true signal from the optical measurement which includes the background signal. There are some correction methods to remove this artificial signal on the true intensities. These methods can be separated into the two parts, namely, deterministic and probabilistic approaches. Among deterministic approaches the first method considers that the background signal is postulated as an additive effect, i.e., the observed signal S consists of the true signal T and the background signal B such that

$$S = B + T. \tag{2.4}$$

Therefore, the background signal cannot be quantified, but the background value near the spot can be measured. In order to eliminate this spurious sign, B can be substracted from the observed value to get the true signal $\hat{T}$ (Eisen, 1999) by

$$\hat{T} = S - B. \tag{2.5}$$

On the other hand Wit and McClure (2004) suggest to use the value of the empty arrays' signal instead of the ones near the spot value. It is thought that the lowest signal should be zero and if not, it can be concluded as the systematic artifact. In this deterministic method, the mean or median is calculated for all empty probes. Then this value is substracted from all spots and negative values are recorded as zero like the following expression:

$$\hat{y}_i = \max\{y_i - \mu_e, 0\}. \tag{2.6}$$

Here, y_i is the original probe intensity, μ_e is the average intensity of all empty probes, $\hat{y}_i$ is the corrected probe intensity and i is the index of probes. On the other side for the probabilistic approaches, although there are a number of methods, here we consider only the most well-known ones in the literature. These are MAS 5.0, RMA, dChip (or MBEI) and GC-RMA techniques. Below we present the mathematical details of each of these methods.

2.2.2.1 MAS 5.0

MAS 5.0 (Microarray Suite Software) is one of the commonly used gene expression indices that are used for oligonucleotides (Hubbel et al., 2002). In this background normalization method, it is considered that MM probes arise only from stray signals which are the result of binding on the array surface instead of probes and the effect of MM probes on PM probes is calculated in an additive way such that the true signal T_{ij} is obtained from the given equation.

$$\log T_{ij} = \log(\mathrm{PM}_{ij} - S_{ij}),$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The indices i and j represent n genes and m probes, respectively. As a component of the true signal, S_{ij} indicates the stray signal while PM_{ij} presents the perfect match. MM_{ij} is checked out in two different way dependent on the amount of S_{ij}. If $PM_{ij} > MM_{ij}$, then the amount of the stray signal accounts for $S_{ij} = MM_{ij}$. If $PM_{ij} < MM_{ij}$, the amount of the stray signal is found via $S_{ij} = logPM_{ij}$ - SB_i^+. Here, SB_i^+ shows a specific background signal and provides a robust estimate for each gene expression. SB for gene i and probe j is calculated as follows.

$$SB_{ij} = T_{bi}[\log \mathrm{PM}_{ij} - \log \mathrm{MM}_{ij}]. \tag{2.7}$$

In Equation (2.7), T_{bi} points out the one-step Tukey biweight estimator of location and each data point takes weight according to the distance from the median as calculated below.

$$T_{bi} = \frac{x_j - \widetilde{\mu}_i^m}{\widetilde{\sigma}_i^m},$$

where the data point $x_j = \log PM_{ij}$ - $\log MM_{ij}$, $\widetilde{\mu}_i^m$ is the median of the ith gene in the mth probe and $\widetilde{\sigma}_i^m$ is the median absolute deviation. If u_{ij} is obtained by dividing each x_j via its by median absolute deviation, the weight for each data point is computed according to u_{ij}'s value as follows:

$$w(u) = (1 - u^2)^2, \text{for } 0 \leq \mid u \mid \leq 1$$

and zero in other cases. By this way the true signal is found by

$$T_{bi} = \frac{\sum_{j=1}^{n} w(u)x_j}{\sum_{j=1}^{n} w(u_j)}.$$

As SB_i^+ should be positive by guarantying that $PM_{ij} > MM_{ij}$, it is checked with a threshold value γ. If $SB_i^+ < \gamma$, the given correction below is applied in place of SB_i^+ as follows (Affymetrix, 2002).

$$SB_i^+ = \frac{\gamma}{1 + 0.1(\gamma - SB_i^+)}.$$

On conclusion although this method can compute PM and MM while $PM > MM$, it can include add-hoc adjustment if $MM > PM$ to guarantee the positively estimated true signals. In the end this adjustment can cause bias in the estimation (Purutçuoğlu and Wit, 2007).

2.2.2.2 RMA

The RMA (Robust Microarray Analysis) background normalization method is based on only PM intensities (Irizarry et al., 2003). It is considered that PM includes the true signal S and the background signal b which arises from the non-specific hybridization. Hence the true signal is obtained from PM intensities via the conditional expectation as follows:

$$S_{aij}^* = E(S_{aij}|S_{aij} + b_{aij}) = E(S_{aij}|PM_{aij}),$$

where $i = 1, \ldots, n$, $j = 1, \ldots, m$ and $a = 1, \ldots, k$ stand for the ith gene, the jth probe and the ath array, respectively, while true signals are distributed exponential, background signals are normal with mean $E(b_{aij}) = \beta_a$. In this method before evaluating gene expression values, first of all, the quantile normalization is applied across arrays to make their distributions identical. Then these normalized values are transformed to the logarithmic scale in order to describe an additive model as below:

$$\log_2(S_{aij}^{**}) = \mu_{ai} + \alpha_{aj} + \varepsilon_{aij}$$

in which μ_{ai} represents the gene expression level, α_{aj} indicates the probe effect and ε_{aij} shows a normally distributed random error term with mean zero.
On the other hand when RMA results are compared with the ones of MAS 5.0 and MBEI, it has been shown that RMA has smaller standard deviation, particularly, for genes which have low intensity values. Moreover under different concentration levels its fold-change results have found as better. Also, if ROC (Receiving operating characteristic) curves is utilized to assign active genes, the normalized values with RMA are more sensitive. In spite of these advantages, the RMA normalization is not an effective way if there are many numbers of arrays in the analysis. Because it uses the least squares (LS) method in the estimation of true signals as MBEI. Thus while the number of arrays increases, the number of conditions in the LS estimator becomes very restrictive to find unique solutions from associated normal equations (Irizarry et al., 2003; Purutçuoğlu et al., 2011).

2.2.2.3 MBEI (dChip)

The MBEI (Model Based Gene Expression Index) method proposes a multiplicative model for observed gene expression values for each probe pair (Li and Wong, 2001). In this method it is suggested that the probe intensity and model expression index θ_a have a linear relation for a gene in the ath array. This relation is different for each probe pair and follows an increasing manner. If an observed signal is high regarding all the other signals in probes, related model expression index θ_a has also higher rate. In addition, it is agreed that PM intensities react faster than MM intensities. Then the model based on PM and MM intensities is as follows.

$$MM_{aj} = v_j + \theta_a \alpha_j + \varepsilon_{aj}^m$$

and

$$PM_{aj} = v_j + \theta_a \alpha_j + \theta_a \phi_j + \varepsilon_{aj}^p,$$

where v_j is the constant term which arises from the non-specific hybridization, α_j indicates an increasing rate for MM values and ϕ_j represents another increasing rate for PM intensities for the jth probe pair of the ath array. Moreover ε_{aj}^m and ε_{aj}^p stand for random errors of MM and PM, respectively. Then the final state of model for the true signal in the jth probe and the ath array is shown as below.

$$Y_{aj} = PM_{aj} - MM_{aj} = \theta_a \phi_j + \varepsilon_{aj},$$

in which ε_{aj} is the random error term and has mean zero and variance σ^2.
As in the estimation of model parameters, the square estimation method is implemented, MBEI is not good for the large number of arrays. Moreover different from MAS 5.0, MBEI performs on the original scale. Furthermore, if MBEI and MAS 5.0 are compared with respect to the accuracy measure, MBEI gives better results (Li and Wong, 2001; Purutçuoğlu et al., 2011).

2.2.2.4 GC-RMA

The GC-RMA (Robust Microarray Analysis based on GC content) is an extended method of RMA (Gentleman et al., 2005). The GC-RMA method supposes that MM probe intensities include some information about the true signal as well as the background signal unlike RMA. This method is the first method suggesting this idea. Hereby the PM values appear in the form of the summation of the optical noise O and the non-specific hybridization which arises from misbinding of some part of target genes and the true signal S. These noisy signals from the non-specific hybridization are caused by the irrelevant sequence, rather than the expected complementary sequence. On the other hand the MM values involve background signal and some proportional true signal whose fraction is indicated via ϕ in Equation (2.8) and (2.9). Thereby PM and MM values are shown in the model as follows:

$$PM = O_{PM} + N_{PM} + S, \tag{2.8}$$

$$MM = O_{MM} + N_{MM} + \phi S, \tag{2.9}$$

where the optical noise O and the non-specific hybridization N values are independent functions of the probe affinity meaning that the summation of the base effect depend on the probe position. On the other side in inference of model parameters GC-RMA implements the maximum likelihood approach with empirical Bayesian method. Whereas since this model has large number of parameters caused by gene and probe specific terms, it uses a simplied version of the model to decrease the complexity in the calculation. Accordingly, in the estimation it sets ϕ to zero, resulting in almost no difference in the practical model.
Moreover regarding its performance with other methods, GC-RMA gives better result than RMA in terms of accuracy like MAS 5.0 while RMA performs better according to precision which is the inverse of covariances (Purutçuoğlu et al., 2011).
The other methods in the literatures are mainly extensions of this method. Because from various analyses it has been show that PM and MM values are correlated implying that MM intensities also include part of the true signals (Purutçuoğlu and Wit, 2007; Purutçuoğlu, 2012). However as this is an active research area, none of the current extended methods gives always outputs better than others based on the model selection criteria in this context (Cope et al., 2003). Therefore, in this study we merely focus on those underlying methods that can be considered as the most well-known approaches in the background normalization of oligonucleotides.

2.2.3 Dye-Effect Normalization

In microarray studies the data are represented as an image at the end of some processes. As this image is obtained via an optical scanner, the dye is used to provide visible genes in the analyses. In two-channel microarray studies Cy3 and Cy5 dyes are utilized to separate treatment and control group (Schena, 2003; Wit and McClure, 2004; Stekel, 2003). But as these dyes have different sensitivities, they need to be corrected to equate their effects, resulting in no artificial variation in the hybridization process. Hence the dye normalization is used to discard this source of the systematic bias in the measurements. There are different methods to get rid of the dye effect.

First method is called the dye-swap approach (Ye, 2008). In this technique, after control and treatment intensity levels are measured with dyes Cy3 and Cy5 for each probe, the dyes are swapped for these two groups in another two-channel array and the intensity values are taken again. Then, the average intensity values are calculated for each probe by using these two measurements' sets. The dye-swap method has some disadvantages. As discussed in the array effect in microarray studies, we cannot be certain whether the dye effect is eliminated under such exchange. In addition, if there are many numbers of arrays in an experiment, the dye-swap method cannot be an effective way (Wit and McClure, 2004; Stekel, 2003).

The other method includes the smoothing of the data. Before the smoothness, the data are transformed into the logarithmic scale and its scatter plot is drawn.This plot is called the Moving average (MA) scatter plot. In a MA scatter plot, M values represent log ratios of two groups' intensities and are stated on the vertical axis while A values indicate the mean of two groups' intensities in log scale and lie on the horizontal axis. Then the following transformation in Equation (2.10) and (2.11) is performed:

$$m_i = \log(R_i) - \log(G_i), \tag{2.10}$$

and

$$a_i = \frac{1}{2}(\log(R_i) + \log(G_i)), \tag{2.11}$$

where i stands for the probe indice and R (red) and G (green) present two groups with Cy5 and Cy3 dyes, respectively. Later the invariant genes are detected via the MA plot. The invariant genes are known as genes whose behaviors do not change under both dyes' types. Finally a smoothing function is fitted with a proper smoothing parameter. For smoothing, the nonparametric *loess* method can be used. The loess method implies the locally weighted linear regression and is used to estimate the regression surface. According to this smoothed regression line, $\tilde{f}$, the normalized MA values are obtained as follows:

$$\widetilde{m_i} = m_i - \tilde{f}(a_i), \tag{2.12}$$

$$\widetilde{a_i} = a_i. \tag{2.13}$$

By using the normalized values in Equations (2.12) and (2.13), the normalized red and green values given in the two-channel microarray are calculated via Equation (2.10) and (2.11) as noted below (Wit and McClure, 2004).

$$\log \widetilde{R_i} = \widetilde{a_i} + \widetilde{m_i}/2 \tag{2.14}$$

and

$$\log \widetilde{G_i} = \widetilde{a_i} - \widetilde{m_i}/2. \tag{2.15}$$

2.2.4 Normalization within conditions

The previously discussed normalization methods may not be sufficient to obtain the true signal. Because some systematic bias can occur during the sample preparation or scanning each probe. Hence, to compare different arrays, the normalization within and across condition is utilized.
It is expected that the replicates under the same condition should result in same findings, whereas, as mentioned beforehand, it is not the case in general. A quantile normalization approach is suggested to overcome this problem (Bolstad et al., 2003). Briefly in this method, all replicates are tried to put into the same scale and a common mean of distributions for all replicates is calculated in order to apply it as a new scale.

2.2.4.1 Quantile Normalization

In microarray studies the aim of the normalization is to remove the artificial variation in the data. While there are different methods for the normalization of a single array, it is also essential that any number of arrays can be comparable. At this stage, arrays to compare should be cleaned out potential systematic bias which can arise from hybridization or scanning process. To make it possible, arrays are transformed in such a way that the intensity values from different arrays are in the same scale. Hereby the quantile normalization method is used to combine the gene expression values in such a manner that they will have the same location and scale parameter. If an intensity value is represented by x_{ij} ,

where $i = 1, \ldots, p$ for probes and $j = 1, \ldots, n$ for arrays, then m_j and s_j are assigned as the locale and scale parameter, respectively. Thereby the normalized intensities are presented via

$$x^*_{ij} = \frac{x_{ij} - m_j}{s_j}$$

and the normalized values are back transformed to the original scale as given below.

$$x^{**}_{ij} = m + s x^*_{ij}. \tag{2.16}$$

Now, the intensity values x^{**}_{ij} found from Equation (2.16) are in the same scale. In the calculation of m and s for the location and the scale parameter, respectively, the mean and the standard deviation can be used or the median and the median absolute deviation can be preferred to get more robust results (Wit and McClure, 2004).

Here we explain the procedure of the quantile normalization method via a numeric example. It is supposed that there are intensity values of four genes from three arrays as in Table 2.1. First of all, each column is sorted in ascending

Table 2.1: Intensity values of four genes from three arrays.

	Array1	**Array2**	**Array3**
Gene1	2	7	4
Gene2	5	6	5
Gene3	4	3	6
Gene4	8	5	3

order and their rank values are recorded as in Table 2.2. Then, the mean for

Table 2.2: Ascending ordered rank values for the intensities in Table 2.1.

	Array1	**Array2**	**Array3**
Gene1	1	4	2
Gene2	3	3	3
Gene3	2	1	4
Gene4	4	2	1

each gene over rows is calculated as shown in Table 2.3 and these values take

the place of original values according to their rank values at the beginning as presented in Table 2.4.

Table 2.3: Sorted genes over arrays

	Array1	**Array2**	**Array3**	**Mean of ranks per gene**
Gene1	2	3	3	2.67
Gene2	4	5	4	4.33
Gene3	5	6	5	5.33
Gene4	8	7	6	7

Table 2.4: Quantile normalized values presented in Table 2.1

	Array1	**Array2**	**Array3**
Gene1	2.67	7	4.33
Gene2	5.33	5.33	5.33
Gene3	4.33	2.67	7
Gene4	7	4.33	2.67

Once the underlying normalization techniques are performed, the normalized data are checked one more time to detect the effect of the normalization and the quality of the normalized measurements. In the following part we briefly explain this step via the principal component analysis. Then we start the statistical analysis of new biological data from their normalization and then the detection of their differentially expressed genes with further analysis and finally the clustering at selected genes to see biologically related species.

2.3 Quality Control

In microarray studies major interest is a multidimensional dataset. As explained previously, before analyzing the data statistically, certain normalization techniques are applied to remove the systematic bias in the observations. After the normalization procedure, the visualization methods are performed in order to realize problematic measurements or to sight whether biological replicates and different experimental conditions are grouped together as expected. Hereby the

dimension reduction methods are performed to decrease the size of the multi-dimensional data so that the data can be visualized easily. There are several methods to reduce the dimension. The most widely implemented one among alternatives is the principle component analysis (PCA) (Stekel, 2003; Lee, 2004; Wit and McClure, 2004). In the PCA approach, each axis represents a linear combination of original axes that indicate original dimensions (Johnson, 2007; Everitt, 2011). The new axes are called the principal components. These axes are orthogonal to each other like the original ones. Thereby the first principal component is one of the new axes and contains the largest variation among the components. Similarly, the second and the third principal components have the largest variations in descending order after the first principal components. In theory, the principal components are the eigenvectors which are derived from the original data. The covariance matrix or correlation matrix is used to obtain eigenvectors. For a given data matrix, if n random variables are discussed and are represented via a random vector $X' = [X_1, X_2, ..., X_n]$, and if its covariance matrix $\sum$ has the eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$, the principal components which are the uncorrelated linear combinations are derived as follows.

$$\begin{aligned} Y_1 &= a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n, \\ Y_2 &= a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n, \\ &\vdots \\ Y_n &= a_n'X = a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n. \end{aligned} \tag{2.17}$$

As it is mentioned beforehand, the variance of the first principal component $Var(Y_1)$ is the maximum one and each variance $Var(Y_i)$ is maximized providing $a_i'a_i = 1$. Also as the linear combinations are uncorrelated, their pairwise covariances are zero as shown by the expression in Equation (2.18) and (2.19)

$$Var(Y_i) = a_i' \sum a_i, \tag{2.18}$$

$$Cov(Y_i, Y_k) = a_i' \sum a_k = 0 \text{ for } k < i, \tag{2.19}$$

where $i, k = 1, \ldots, n$ and $a_1, a_2, \ldots, a_n$ are the related eigenvectors.

2.4 Differential Expression

In microarray studies, it is tried to find which genes are active under the investigated condition. In other words, how the expression levels of a gene from the treatment group change regarding the control group. There are different methods for detecting differential expression levels of genes such as frequentist or classical approaches and the Bayesian hypothesis testing (Ye, 2008). In this study we are merely interested in frequentist or classical approaches due to their advantages in computational time.
In the hypothesis testing method it is examined whether the difference between population means for a gene under distinct condition is zero. A typical hypothesis testing has four elements. These are a hypothesis, a test statistic, an error rate and a decision rule.

2.4.1 Hypothesis

The hypotheses are used for the *gene discovery* . The tested hypothesis is named as the *null hypothesis*, H_0, and represents that the gene under consideration is not differentially expressed. That is, there is no difference between control and treatment group for this gene. The null hypothesis is tested against the *alternative hypothesis*, H_1, saying that the gene is differentially expressed. The alternative hypothesis can be constructed within the context of the gene to observe whether it is more or less active. Hereby when H_{g0} is rejected, it means that the associated gene is differentially expressed, or active or significant. The result of the hypothesis test is obtained as a *p-value* and evaluated in terms of a certain cut-off value. If the p-value is less than the cut-off, H_0 is rejected in favour of H_1, meaning that the expression level of the gene is changed as indicated in the alternative hypothesis.
Hereby the hypotheses constituted for a gene g are represented via

$$\begin{aligned} H_{\mathrm{g0}} &: \text{gene g is not differentially expressed} \\ H_{\mathrm{g1}} &: \text{gene g is differentially expressed} \end{aligned}$$

If the genes are tested simultaneously, the multiple testing problem can arise. If the mutually exclusive hypotheses are tested for each gene according to a fixed p-value, a very small rate is obtained for the assessment. In this case, H_0 can be rejected falsely, although it should be accepted. In order to solve this problem different types of error rates are defined for the hypotheses and the most convenient one can be chosen according to the single gene test or the simultaneous testing procedure of genes (Lee, 2004).

2.4.2 Test statistic

The test statistic is a kind of the summary calculated from the data and used for deciding which hypothesis should be supported either null hypothesis or its alternative. Various test statistics can be performed according to different data and conditions. Among alternatives the t-statistic is the most preferable. In this test, the difference of means is divided by the standard deviation to obtain the test statistics (Ross, 2010).

2.4.3 Error rates

In hypothesis testing if only one genes is discussed, types of two errors can be considered. These are false positive and false negative errors. The *false positive* error occurs if the null hypothesis is wrongly rejected. On the other hand, if it is wrongly accepted, the error is called as the *false negative*. Between these two errors, when the probability of one of them increases, the other decreases. Accordingly if the significance level raises, the probability of the false positive also enhances while the false negative decreases. Thereby, the *power of a test* is defined as the difference of the false negative rate (FNR) minus one increases.
On the other side if the multiple testing procedure is followed to test n genes simultaneously, different types of errors, as presented in Table 2.5 are used. One of the error rates is called the false positive rates indicated via

$$FPR = E[FP/n_0]. \tag{2.20}$$

Table 2.5: Correct and incorrect decisions of hypotheses.

Genes	**Fail to reject H_0**	**Reject H_0**	**Total number of sample size**
Inactive genes	TN	FP	n_0
Active genes	FN	TP	$n - n_0$
Total number of sample size	$n - n_r$	n_r	n

The familywise error rate (FWER) is defined as the probability that at least one gene among all inactive genes is declared as differentially expressed eventhough it is not differentially expressed. Hereby it is denoted by

$$FWER = P(FP > 0). \tag{2.21}$$

Finally the false discovery rate (FDR) shows the expected number of not differentially expressed genes among the genes that are claimed as differentially expressed and it is presented via (Draghici, 2012):

$$FDR = E[FP/n_r]. \tag{2.22}$$

2.4.4 Decision rules

The decision rule is a method that is used to decide whether the null hypothesis should be rejected. The widely used decision rule is the p-value approach. The p-value represents the probability of the occurrence of the possible maximum test-statistic value which provides the trueness of the null hypothesis.

2.5 Hypothesis Testing

In microarray analysis, the p-value calculated according to the test statistic is used for testing the null hypothesis and determining whether a gene is active. As there are different test statistics, different p-values can be obtained from the same hypothesis. Thus, it is crucial that the most convenient test statistic is chosen. There are certain methods developed under different conditions and

the proper one can be preferred to get the most suitable test statistics for the calculated gene expression level. Below we present the most well-known ones.

2.5.1 *t*-statistic

The t-statistic is commonly used to compare the two different means. It is able to be used if the observations are normally distributed and independent from each other. While the expression levels of a gene under two different conditions are compared, there are two distinct cases according to the validity of the assumption of equal variances. If the assumption of equal variances is held for two groups, the Student-t distribution is suitable. In the calculations, firstly, a pooled sample variance is calculated for two conditions as shown in Equation (2.23) (Draghici, 2012). Then using this pooled sample variance, the t-statistic presented in Equation (2.24) is obtained for the hypothesis testing.

$$s_{gp}^2 = \frac{(n_{g1}-1)s_{g1}^2 + (n_{g2}-1)s_{g2}^2}{n_{g1}+n_{g2}-2}, \tag{2.23}$$

where

$$t'_g = \frac{\bar{x}_{g1.} - \bar{x}_{g2.}}{\sqrt{s_{gp}^2(\frac{1}{n_{g1}} + \frac{1}{n_{g2}})}}. \tag{2.24}$$

In Equation (2.23) and (2.24), n_{g1} and n_{g2} represent the sample size of the observation under the control and treatment groups, respectively, for gene g. Accordingly s_{g1}^2 and s_{g2}^2 are the sample variance of the measurement for control and treatment groups, in order, under the gth gene. Finally $\bar{x}_{g1.}$ and $\bar{x}_{g2.}$ show the associated mean value for the same genes.
The calculated test statistic t'_g has the Student-t distribution with $n_{g1}+n_{g2}-2$ degrees of freedom. Accordingly, the p-value for gene g is obtained via

$$t'_g \sim t_{n_{g1}+n_{g2}-2}$$

and

$$p-value = 2 \times P(t_{n_{g1}+n_{g2}-2} \geq |t'_g|).$$

If the assumption of equal variances for two groups is not verified, the test statistic is calculated as follows (Stekel, 2003).

$$t'_g = \frac{\bar{x}_{g1} - \bar{x}_{g2}}{\sqrt{\left(\frac{s^2_{g1}}{n_{g1}} + \frac{s^2_{g2}}{n_{g2}}\right)}} \tag{2.25}$$

where s^2_{g1} and s^2_{g2} are the sample variances for two groups. The calculated test statistic is approximately t-distributed with a special degree of freedom that depends on the squared standard errors of the two groups and is represented via

$$t'_g \quad \sim \quad t_v$$

while

$$v \quad = \quad \frac{(z_1 + z_2)^2}{z_1^2/(n_1 - 1) + z_2^2/(n_2 - 1)}.$$

Here z_i symbolizes the statement s_i^2/n_i for each group. Thereby the p-value for gene g is denoted by

$$p - value = 2 \times P(t_v \geq |t'_g|).$$

2.5.2 Wilcoxon rank sum statistic

The Wilcoxon-Mann-Whitney statistic, also known as Mann-Whitney or Mann-Whitney-U test, is a nonparametric test statistic (Gibbons, 2003; Draghici, 2012). In this testing procedure if the data are not normally distributed but they have approximately similar shape, we can compare the means of a gene under two conditions. In this method observations from two groups are initially pooled and considered as one sample. Then the data are sorted from smallest to highest. The test statistic is calculated by summing up the ranks of observations that come from one of the groups. The general form of the test statistic, for example for the first group, is represented via

$$t'_g = \sum_{j=1}^{n_{g1}+n_{g2}} rank(x_{g1j}).$$

In the calculation of t'_g, since the two groups are thought as one sample, the number of observation becomes $n_{g1} + n_{g2}$. If all the observations from the first

group are smaller than the observations from the second group, the test statistic is computed via $\sum_{j=1}^{n_{g1}} j$. When all the observations from the first group are larger than the observations from the second group, then the test statistic is found from $\sum_{j=n_{g2}+1}^{n_{g1}+n_{g2}} j$. Thereby the test statistic is distributed as

$$T'_g \sim Wilcoxon(n_{g1}, n_{g2}).$$

In the final stage, the p-value can be obtained from the reading table of the Wilcoxon distribution, via

$$p - value = 2 \times \min\{P(T'_g \leq t'_g), P(T'_g \geq t'_g)\}.$$

2.5.3 Global error likelihood ratio test statistic

The global error likelihood ratio test statistic can be performed under the assumption that the errors of each group have a constant variance for all genes. The likelihood represents the probability to obtain the data which are on hand accepting that a specific model is true. Basically the test statistic θ_g is interested in the ratio of the likelihoods under the null and alternative hypothesis for gene g as shown the following expression (Bain, 1992).

$$\theta_g = \frac{\text{likelihood for active gene g}}{\text{likelihood for inactive gene g}}$$

If a large ratio is obtained, it is a supporting evidence of an active gene g . However, as the θ_g for each gene is evaluated simultaneously and as this test assumes that the data have both an additive and a multiplicative variance at lower and higher intensity levels, respectively, the calculation can be complicated. The data that have both an additive and a multiplicative variance can be shown as below.

$$x_{g1j} = \mu_{g1} + \mu_{g1}\delta_{g1j} + \gamma_{g1j} \tag{2.26}$$

and

$$x_{g2j} = \mu_{g2} + \mu_{g2}\delta_{g2j} + \gamma_{g2j} \tag{2.27}$$

in which x_{g1j} and x_{g2j} are the gene expression levels of the jth replicate of two groups, μ_{g1} and μ_{g2} are the true mean expression of gene g in two conditions.

Moreover δ_{g1j} and δ_{g2j} present the multiplicative error terms while γ_{g1j} and γ_{g2j} indicate the additive error terms. The error terms come from the bivariate distribution with zero mean and variances $\sigma^2_{\delta_1}$, $\sigma^2_{\delta_2}$, $\sigma^2_{\gamma_1}$, and $\sigma^2_{\gamma_2}$ with correlations ρ_δ and ρ_γ (Wit and McClure, 2004).
The parameters $\phi = (\sigma^2_{\delta_1}, \sigma^2_{\delta_2}, \rho_\delta, \sigma^2_{\gamma_1}, \sigma^2_{\gamma_2}, \rho_\gamma)$ and $\mu = \{(\mu_{g1,g2}) : g = 1, ..., n\}$ are calculated from the maximum-likelihood method such that

$$\begin{aligned} L(\phi, \mu) &= \prod_{g=1}^{n} L_g(\phi, \mu_{g1}, \mu_{g2}) \\ &= \prod_{g=1}^{n} \prod_{j=1}^{J} p(x_{g1j}, x_{g2j} | \phi, \mu_{g1}, \mu_{g2}). \end{aligned}$$

Then, in order to get the test statistic, the following maximization is applied under the null hypothesis.

$$\theta_g = -2 \ln \left(\frac{\max_{\mu_g} L_g(\phi, \mu_g, \mu_g)}{\max_{\mu_{g1,g2}} L_g(\phi, \mu_{g1}, \mu_{g2})} \right). \tag{2.28}$$

If Equation (2.28) results in 1, it means that the null hypothesis is true. Here, Equation (2.28) is distributed as χ^2 with 1 degree of freedom under the null hypothesis. Hence to obtain a p-value, the inverse of the χ^2 distribution is subtracted from one.

2.6 Decision rules for multiple testing

As it is explained before, if more than one genes are tested simultaneously, the multiple testing problem can be occurred. In order to overcome this challenge, different decision rules are suggested. One of these decision rules can be chosen according to the purpose of the researcher. In all the decision rules, the ordered p-value denoted by $P_{(i)}$ and implying the ith smallest p-value for each pairwise hypothesis is utilized.

2.6.1 Decision criteria

In the analysis of microarray studies different decision criteria are used to specify the rejection or acceptance of the null hypothesis. Below we list these criteria with their mathematical description.

2.6.1.1 Familywise error rate

If one is interested in the probability that at least one gene is asserted as active eventhough it is inactive, the familywise error rate (FWER) is convenient to use. There are two FWER procedures. The expression of FWER is shown in Equation (2.21). These are Bonferroni and Hochberg methods (Draghici, 2012) whose mathematical details are presented in Subsection 2.6.2.

2.6.1.2 False discovery rate

In the false discovery rate (FDR) method, the average proportion of not differentially expressed genes among all genes which are declared as differentially expressed is investigated. Equation (2.22) indicates the mathematical equality of this value (Draghici, 2012).

2.6.1.3 False nondiscovery rate

The false nondiscovery rate (FNR) is the expectation proportion of false negatives among all the accepted null hypothesis under the condition that the probability of rejecting all null hypotheses is zero. It is formulated as below.

$$FNR = E\{\frac{FN}{n - n_r} \mid (n - n_r) > 0\} P\{(n - n_r) > 0\}. \tag{2.29}$$

Here n is the total sample size and n_r presents the total sample size rejecting the null hypothesis. Finally $E(\cdot)$ denotes the expectation of the given value (McLachlan, 2004).

2.6.1.4 Positive false discovery rate

The positive false discovery rate (pFDR) is used when it is interested in the FDR under the condition that at least one hypothesis is rejected (McLachlan, 2004).

$$pFDR = E(FP/n_r | n_r > 0). \tag{2.30}$$

Similar to the previous equation, n_r refers to the total sample size rejecting the null hypothesis.

2.6.1.5 Positive false nondiscovery rate

According to FNR, the positive false nondiscovery rate (pFNR) is computed as below (McLachlan, 2004).

$$pFNR = E(FN/(n - n_r)|(n - n_r) > 0), \tag{2.31}$$

while n is the total sample size and n_r denotes the sample size for rejecting the null hypothesis as used previously.

2.6.2 Multiple testing procedure

Once the decision criteria are chosen regarding the purpose of the study, they are used to compute the test statistics for multiple comparison of genes. Below we describe these testing procedures in details.

2.6.2.1 Bonferroni FWER method

In this method, the cut-off level α is divided by the total number of genes n and this new value is used as the significance level. Hereby if a p-value for a gene is less than this new α, the gene is recorded as active. By this way the method is guarantees that FWER does not exceed α. But as it is very strict to control FWER, it decreases the power of the test (Lee, 2004).

2.6.2.2 Hochberg FWER method

In this method, the largest p-value, P_n, is compared with α. If $P_n \leq \alpha$, it is accepted that other p-values are also less than α. Thus, all null hypotheses which are tested for one gene are rejected and all of the genes are declared as active. Otherwise, the second largest p-value, P_{n-1}, is compared with $\alpha/2$. If

P_{n-1} is less than $\alpha/2$, then the related hypotheses are rejected. This procedure goes on sequentially by this way until the comparison via the smallest p-value. Here each ordered p-value is formulated by

$$P_{(g)} \leq \frac{\alpha}{n-g+1}.$$

In this inequality the null hypothesis $H_{(g)}$ is evaluated for $g = 1, \ldots, k$ where k has the largest p-value for gene g (McLachlan, 2004).

2.6.2.3 Benjamini and Hochberg FDR method

If it is assumed that k has the largest p-value for gene g, the null hypotheses $H_{(g)}$ are evaluated according to the following formula.

$$P_{(g)} \leq \frac{g\alpha}{np_0},$$

where $g = 1, \ldots, k$. As the true fraction of not differential genes, denoted by p_0, is typically unknown in advance, p_0 is set to 1 in practice (Lee, 2004).

2.6.2.4 Permutation correction

The Westfall and Young (W-Y) correction follows a permutation step and Bonferroni step-down method. First, the p-values for each gene are calculated for the original dataset that includes control and treatment groups and later the Bonferroni multiple testing correction is utilized in order to adjust p-values. Then, the control and treatment groups are changed and assigned randomly as control and treatment groups. The new p-values for each gene are calculated for this permutation and corrected according to the Bonferroni method. This process is repeated for a few thousand of times and p-values are recorded. The final p-value for gene i is founded using the proportion given as below which is interested in the number of the permutation that has a higher test statistic than the original one (Draghici, 2012).

$$p_i = \frac{\text{number of permutations that } t_i \leq u_j^{(k)}}{\text{number of all permutations}}, \tag{2.32}$$

where $u_j^{(k)}$ is the adjusted value for the kth permutation and t_i is the computed Student-t test statistic.

2.7 Clustering

Statistical learning can be examined under the two titles, namely supervised and unsupervised learning. In these methods, the observations are divided into different groups in such a way that similar observations according to their features are assigned in the same group. While the groups are previously stated in the supervised learning, there is no prior knowledge about what are the groups in the unsupervised learning. In this study since the knowledge about the optimal number of clusters is not available, we perform the unsupervised methods whose mathematical details are presented as below.

2.7.1 *k*-Means Clustering

The k-means clustering is a non-hierarchical method splits the data into the k separate groups. The significant point of this method is that the number of groups k is specified at the beginning and each observation belongs to only one of these k groups.

The best clustering can be achieved if the within-cluster variation is minimized. In order to get it, the observations in the same cluster should be too close to each other while they are quite away from the other observations in the remained clusters. Indeed this constraint is related with the similarity or dissimilarity of observations. There are different procedures in order to decide on whether the observations are similar. The *Euclidean distance* is the most frequently used dissimilarity measure and is indicated as given below.

$$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2}, \tag{2.33}$$

where x_{ik} and x_{jk} are the kth variable value for the observations i and j with p-dimension.

As the minimum within-cluster variation has the same meaning with the minimization of the summation of the clusters' variation, the optimum clustering can be represented via (James, 2013)

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \epsilon C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}, \tag{2.34}$$

where $C_1, \ldots, C_K$ indicate clusters, i and i' stand for pairwise observations of stated clusters and j represents the features of observations. Now this definition becomes an optimization problem in such a way that an iterative algorithm can be performed to be able to reach the best clustering. After deciding the number of cluster, at first each observation is assigned to one of these clusters randomly and the cluster centroid of each cluster is calculated. Then each observation is put in a new cluster whose centroid is the closest according to the Euclidean distance. Later the centroid of each cluster is calculated again and each observation is assigned to the closest centroid anew. This iterative process is repeated until the time that the result does not change anymore. The important point is that as the result depends on the first assignation of the observations for clusters, the algorithm needs to run many times from the beginning by using different random initial cluster numbers for observations. Finally, the results obtained according to the optimization criteria of Equation (2.34) are compared and the best clustering is chosen.

2.7.2 Hierarchical Clustering

The hierarchical clustering is a method whose representation is based on the hierarchical structure, also called *dendrogram*. The dendrogram is a tree diagram and provides convenience for the interpretation. While performing the hierarchical clustering, we can perform the divisive (top-down) or agglomerative (bottom-up) techniques (Everitt, 2001). In the divisive clustering technique, the data are thought as one cluster and split into the two clusters. Then each of the current new cluster is divided again into the two clusters and this process is gone until each observation owns an individual cluster. On the other hand in the agglomerative clustering, every observation is accepted as one cluster and then the closest two observations are though pairwise and are put together in the same cluster. Hereby the algorithm is run sequentially in such a way that each pair of the closest two clusters is joined as a new cluster and this process is held up to obtain one cluster. The number of clusters can be determined by using horizontal cut on the dendrogram. Accordingly every separate branch of the diagram represents a cluster. But one should be careful while interpreting

the dendrogram. Because the observations or clusters cannot be evaluated as similar based on the affinity on the horizontal axis. On the contrary, two observations are accepted as similar if they are subjected to the same branch. In other words, the similarity is investigated on the vertical axis.
Hence the clustering method is defined with respect to the similarity or dissimilarity of the observations. There are different measures for the evaluation of similarity and dissimilarity. The correlation distance measure is one of the commonly used measures to compute the distance between two observations such as the Euclidean distance. While r_{xy} stands for the Pearson correlation coefficient of vectors x and y, the correlation distance is calculated as given below (Shay, 2003).

$$d_{corr}(x, y) = 1 - r_{xy}. \tag{2.35}$$

As it is seen in the hierarchical clustering, the calculation of the distance between two groups which have more than one observation is also needed. The different types of *linkage* is used to assess the dissimilarity between two groups. The most favorite linkage types are the single, average, complete and the centroid linkage.

- *Single linkage* method minimizes the dissimilarity in such a way that all possible pairwise dissimilarities between observations which come from two different clusters are calculated and the smallest one is preferred.
- *Average linkage* method gives the average dissimilarity in the sense that all possible pairwise dissimilarities between observations which come from two different clusters are computed and the mean of them is found.
- *Complete linkage* method maximizes the dissimilarity such that all possible pairwise dissimilarities between observations which are originated from two different clusters are calculated and the largest one is preferred.
- *Centroid linkage* method gives the distance between the centroid of clusters as the dissimilarity measure. The mean of the observations for each cluster is found to detect the centroid of each group.

In this study the correlation distance with the complete linkage approach are performed as they give better results in microarray studies (Gibbons, 2002).

Because as expected among alternatives, the correlation distance enables us to detect the functionally related genes. And the complete linkage is mostly preferred when we expect approximately spherical clusters. Here the Ward method can be an alternative linkage. But it is sensitive to outliers which may be also seen generally in microarray studies (Johnson, 2007).

2.7.3 Hierarchical PAM

The hierarchical partitioning around medoids (HIPAM) is a clustering method proposed by Wit and McClure (2004) and consists in the PAM algorithm by Kaufman and Rousseeuw (2005).

The partitioning around medoids (PAM) algorithm, also known as the k-medoids technique, is a clustering method which is similar to the k-means clustering as they divide the data into the k clusters. In order to cluster the data into k groups, k representative points are determined for each class. While representative point of each group is a calculated value from observations in that class and named as centroid in the k-means clustering, the representative point of each group is chosen itself out of the data and called as the *medoid* in the PAM method. After selecting k objects from the data as medoids for the representation of clusters, the pairwise distance between each observation and each medoid is calculated by using a distance measure such as the Euclidean or Pearson correlation coefficient. Then each observation is assigned to one of these clusters with whom their distances are minimum. The mean of these minimum distances of observations to medoids is assigned as the average dissimilarity. Different average dissimilarity values are obtained depending on the choice of the medoids. The best option which gives the minimum average dissimilarity among all possibilities is decided within this scope.

On the other hand the PAM algorithm has certain advantages in the sense that it is more consistent than the k-means since the centers of clusters are decided among the observations. Moreover it is faster while studying with high-dimension data such as microarray data. However, even if good clustering results can be obtained for the different choice of k, it is required that k must be held in advance and these results are not comparable. Furthermore since it is not a

hierarchical method, it cannot detect any hierarchical structure if the original data have this feature (Wit and McClure, 2004).

Hereby the HIPAM is suggested to overcome this drawback. HIPAM is a hierarchical divisive method in which it stops the down at one point according to a validation rule so that a good clustering can be obtained for the validation rule, the *average silhouette width* (a.s.w) is proposed by Kaufman and Rousseeuw (2005) in this algorithm. The silhouette width, $s(i)$, and the average silhouette width, $a.s.w$; can be shown as below.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2.36}$$

where $a(i)$ represents the average dissimilarity of the ith observation over all other observations in its cluster and $b(i)$ indicates the minimum of the average dissimilarities which are calculated for each cluster separately except the cluster i as the average dissimilarity of the ith observation over all other observations in its cluster. Accordingly,

$$a.s.w = \frac{\sum_{i=1}^{n} s(i)}{n} = \bar{s}, \tag{2.37}$$

in which n is the total number of observations in the data. The high value of $\bar{s}$ for an observation and related cluster at issue indicates a better clustering. This value is obtained when a.s.w is maximized and the clustering that gives this maximum a.s.w is taken as the global validation of HIPAM. However, the global a.s.w gives several clusters in large data. As it is crucial for investigating the data in detail and finding more homogeneous subgroup, the local HIPAM can be preferred. In the local HIPAM, it is looked for whether there are homogeneous subgroups in the generated clusters. In order to decide where splitting is stopped, the homogeneity of the major node and the average homogeneity of sub-nodes are compared. The splitting is gone as long as the homogeneity of the major node is less than the average homogeneity of sub-nodes. In other words, as the higher a.s.w means a lower homogeneity, it is concluded that the split is required if the a.s.w of major node is higher than the average a.s.w of the sub-nodes. The

usage of the mean average silhouette width as a comparison rule for the splitting is originated from the HOPACH algorithm (Pollard and van der Laan, 2002). Finally for the comparison with a.s.w of the major node, different limiting values can be chosen such as the pre-defined constant or certain amount of the mean a.s.w of sub-nodes (Wit and McClure, 2004).

2.7.4 PAMSAM

The partitioning around medoids with the Sammon mapping algorithm (PAMSAM) method integrates two different algorithms in the sense that the PAM algorithm (Kaufman and Rousseeuw, 2005) is performed for clustering while the Sammon mapping, i.e. SAM, method (Wit and McClure, 2004) is used for the visualization of clustering results. As it is explained previously, the PAM method is a kind of the k-means algorithm. On the other hand, the Sammon mapping is a multidimensional scaling method and offers a lower-dimensional representation. It is performed more quickly than other dimension reduction methods since it uses the distance matrix, rather than the original data matrix. Hereby after performing the PAMSAM algorithm, ones obtains a two-dimensional graph of sub-clusters that present their components in a two-dimensional graphs.

On the other hand apart from these major methods, if required, two-way clustering can be performed by combining different clustering methods as well. For instance after obtaining groups via the k-means clustering algorithm if it is expected that there is a hierarchical structure in the analyzed data, the hierarchical clustering can be applied for each of these groups separately (Erkan, 2011).

CHAPTER 3

MATERIAL AND METHOD

The microarrays are used to decide whether the average expression levels of genes change under different conditions. Hence, the amount of RNA that binds to its complement located on the probes and included on the array surface are measured and evaluated (Frazee et al., 2014). For these purposes, the experiment is initially designed and then its numerical assessment is performed. In this study we merely deal with the analysis part of the microarray whose experimental part is already done in the selected yeast data. Figure 3.1 shortly presents these steps based on the underlying two parts and Figure 1.2 shows them in a simple picture.

3.1 Materials

All cells are exposed to rapid environmental changes and they must response these stress in order to survive. Temperature is one of the most important stress for a cell. The heat shock response is a vital mechanism which enables the cells protection against the heat stress. In response to heat stress, the expression levels of certain genes change. The yeast data used in this study are a commercial baker's yeast from Pakmaya (Turkey). The yeast has an important product in the food industry, resulting in the economy. The environmental conditions have quite influence on the yeast. One of the most important factors which affect this food is the temperature. Hereby the purpose of the microarray experiment performed in this study is to investigate the effect of the heat stress on the yeast.

Accordingly in this thesis, twelve microarray datasets which are obtained from

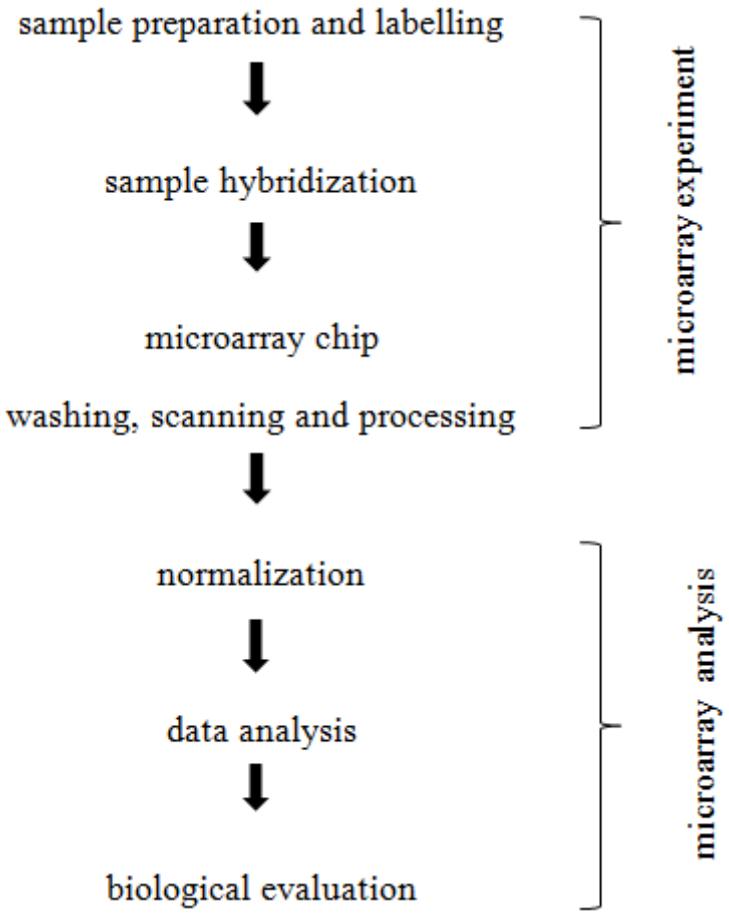

Figure 3.1: Steps of a microarray study

the Affymetrix technology are applied. The selected arrays are the GeneChip Yeast Genome 2.0 Arrays that include probe sets for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. These are the commonly used species for the yeast. This array has nearly 5,744 probe sets for 5,841 of the 5,845 genes existed in *Saccharomyces cerevisiae* and 5,021 probe sets for all 5,031 genes existed in *Schizosaccharomyces pombe*. But in here, the studied species are the *Saccharomyces cerevisiae* which has the first completely sequenced genome in the literature (Goffeau et al., 1996). The array content for *Saccharomyces cerevisiae* was obtained from the public data source GenBank (Affymetrix, 2012).

Table 3.1: Affymetrix GeneChip Yeast Genome 2.0 Array.

Number of probe sets, *S.cerevisiae*	5,744
Number of probe sets, *S.pombe*	5,021
Number of transcripts, *S.cerevisiae*	5,841
Number of transcripts, *S.pombe*	5,031
Number of arrays in set	One
Array format	169
Feature size	11m
Oligonucleotide probe length	25-mer
Probe pairs per sequence	11
Hybridization controls	*bioB, bioC, bioD,* from *Escherichia coli* and and *cre* from P1 bacteriophage
Poly-A controls	*dap, lys, phe, thr, trp* from *Bacillus subtilis*
Housekeeping/control genes	GAPDH, actin, *EAF5*, *SRB4, TFIID, RIP1* *URA3* and *WBP1*
Detection sensitivity	1:100,000

3.1.1 Experimental Model

The data used in this study are generated at the Middle East Technical University, Molecular Biology and Biotechnology R&D Center (Yilmaz et al., 2012). The dataset includes the gene expressions of *Saccharomyces cerevisiae* under certain stress in the sense that the effect of the heat shock and the heat change are investigated as distinct stress levels. Accordingly in the data, there are one control group and two treatment groups assigned for the heat shock and the heat change, respectively. Moreover two biological replicates are used for each treatment group. In the heat shock, cells are incubated at 25 °C for six hours and then at 37 °C for one hour. In the heat change, cells are incubated at 37 °C for six hours and then at 25 °C for one hour. On the other hand the cells situated in the control groups are incubated at 30 °C along the experiment as it is represented in Figure 3.2. Later the gene expression for each group is measured after six hours and one hour in order to investigate the activations under distinct times. Hereby twelve one-channel microarrays are recorded in this dataset

whose brief description and physical feature are also presented in Table 3.1. In this study data are considered as three groups, namely control, treatment 1 and treatment 2, among columns as it is represented in Figure 3.2 and average values are used for each gene over four microarrays included in each condition.

3.2 Method

In this chapter we present the real life microarray data and their statistical analyses by performing the normalization of the measurements, checking quality control of the normalization, detecting the differentially expressed genes and the significance of genes via the multiple testing procedure as well as by implementing clustering of the biologically interesting genes.
All analysis in this chapter are carried out by the Agilent GeneSpring 12.1 software (Agilent, 2012) and the R 2.15.2 programme (Maechler et al., 2014; Lucas, 2014; Witten et al., 2013; Wit et al., 2012). Accordingly the normalization, quality control and the detection of significant genes are run in the GeneSpring software. On the other hand the test for the homogeneity of variance and the cluster analyses are implemented via the R programme.

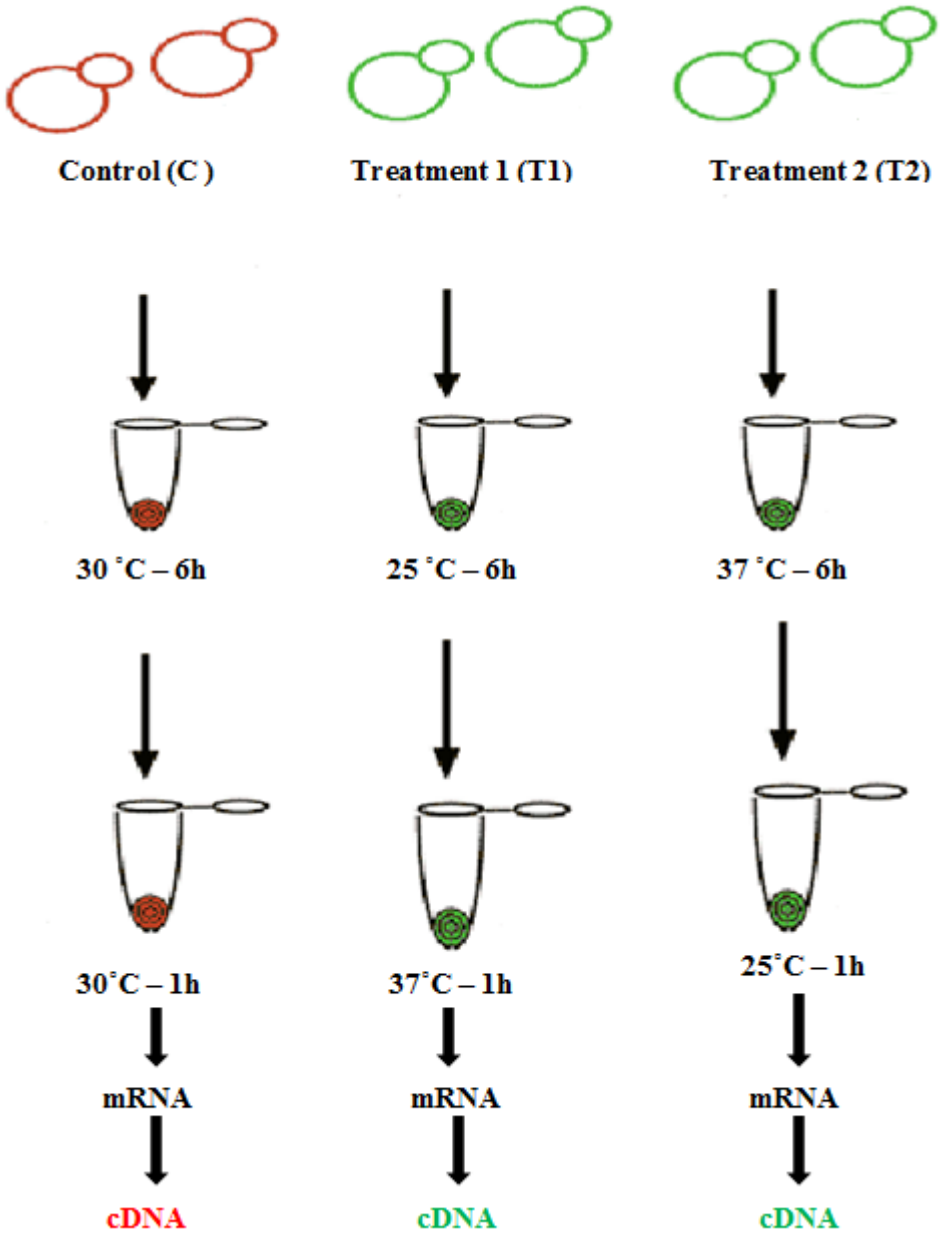

Figure 3.2: Experimental design

CHAPTER 4

RESULTS

4.1 Normalization of Data and Quality Control

As it is explained in the previous chapter, before statistical analyses, the normalization methods of the data are performed. One of the steps in this procedure is the background normalization as described in Chapter 2. Hereby with its alternative approaches the most preferred background normalization methods are considered as MAS5.0, RMA, MBEI and GC-RMA. The GeneSpring software offers some summarization algorithms for Affymetrix expression data. These algorithms include the background correction, normalization and probe summarization steps. These five techniques are applied to raw data automatically under the background correction step and after the quantile normalization as default. After the normalization their figures from principal component analysis (PCA) are compared to implement the quality control step. Here we choose the best performed method regarding their PCA results. Accordingly after the normalization, we expect that the control and treatment groups with their replicates can be shown separately. Figures 4.1, 4.2, 4.3, 4.4 and 4.5 display PCA plots of the raw and the normalized data by using MAS5.0, RMA, MBEI and GC-RMA, respectively. In the PCA figures, different shapes indicate conditions while different colors stand for biological replicates of each sample. The squares, triangles and circles stand for control group (C), heat shock (T1) and heat change (T2), respectively. On the other hand colors blue, red, grey, brown, purple and green indicate the two measurements which are taken after six hours and then one hour later as explained in the experimental design of C, T1 and T2

samples, respectively. For each method the percentage of the first three principal component, namely, the percentage of the variation mostly captured by the linear combinations are presented in Table 4.1. As it is seen clearly in the PCA plots, the best result is obtained via the RMA method. Because in Figure 4.3 the samples stand for the closest outputs with its biological replicate and they are separated from the rest.

Moreover the boxplots for the raw data and the data normalized with the RMA method can be seen in Figures 4.6 and 4.7, in order. Here since RMA already uses the spatially normalized and the quantile normalized data from its calculations, the measurements after the RMA method are, indeed, normalized based on all steps for the typical oligonucleotides.

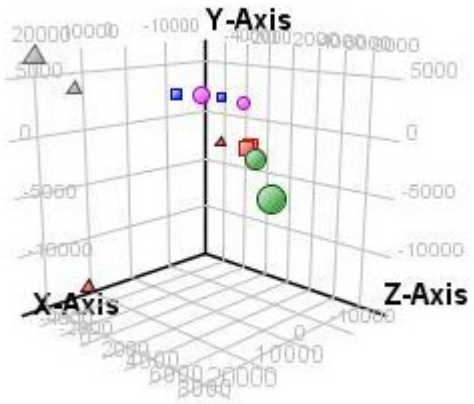

Figure 4.1: PCA of raw data.

Table 4.1: Percentage of principal components.

	X-Axis	**Y-Axis**	**Z-Axis**
Raw	34.59	13.65	12.09
MAS5.0	21.53	17.47	15.07
RMA	22.63	18.04	13.46
MBEI	24.37	18.43	14.97
GC-RMA	33.44	17.47	11.75

On conclusion from the comparison of different normalization methods, we select the RMA method as it gives the best result. After RMA is applied, the nor-

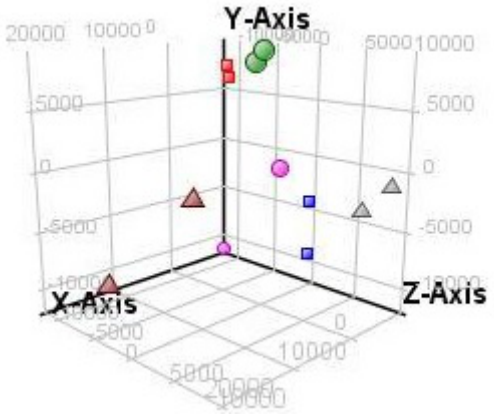

Figure 4.2: PCA for the MAS5.0 method.

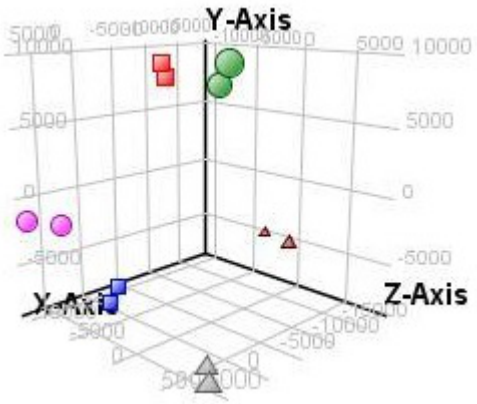

Figure 4.3: PCA for the RMA method.

malized data are analyzed in order to determine differentially expressed genes among the control and treatment groups.

4.2 Detection of Active Genes via Fold-Change

In order to detect active genes whose expression level changes between any two groups, the fold-change can be used. The fold-change calculates the ratio of intensity values for a gene from two selected groups. Here different cut-off values can be chosen for the fold-change. The fold-change computes the absolute

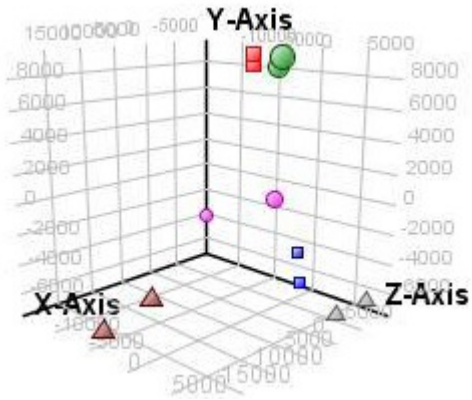

Figure 4.4: PCA for the MBEI(dChip) method.

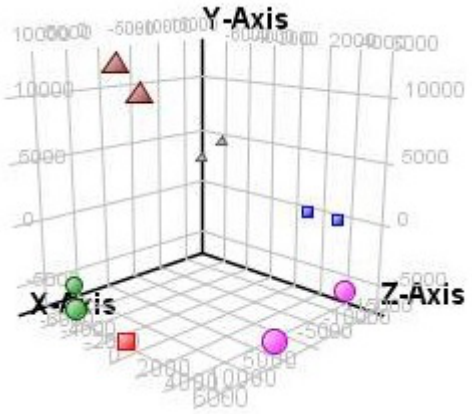

Figure 4.5: PCA for the GC-RMA method.

ratio and gives results for the chosen cut-off. The active genes according to the fold-change results can be up-regulated or down-regulated genes in such a way that it can be computed for two conditions as below.

$$\text{Fold-Change} = \frac{\text{Gene expression under condition1}}{\text{Gene expression under condition2}}.$$

For pairs of conditions such as T1/C, T2/C and T1/T2, the fold-changes are performed and the numbers of differentially expressed genes out of 10928 genes for 2 and 3 fold-changes are shown in Tables 4.2 and 4.3.

Moreover the fold-change analysis is performed to investigate the differential expression levels of the heat shock protein (HSP) genes as they are a model

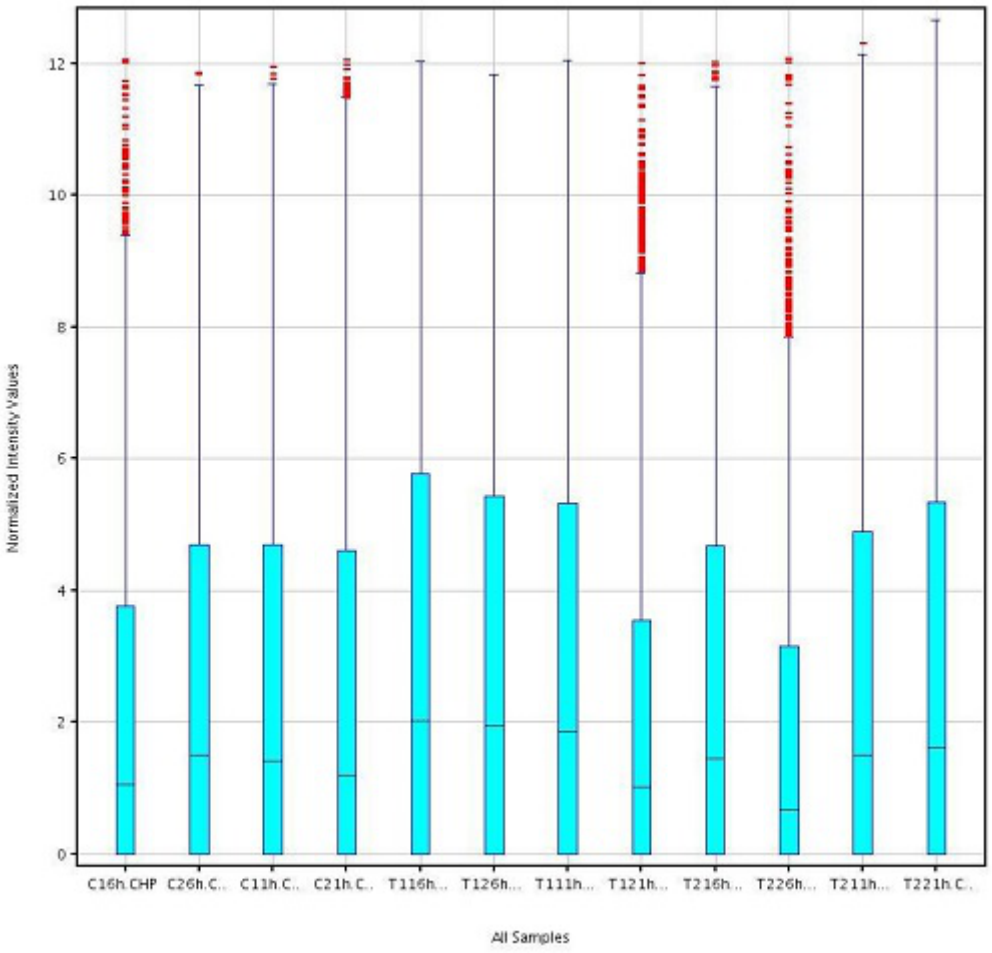

Figure 4.6: Boxplot of the raw (unnormalized) data in log2 scale.

group under the heat stress. As the cut-off value, 1.1 is chosen in order to observe all HSP genes included in yeast *Saccharomyces cerevisiae*. Thus, in this study it is investigated how the expression levels of HSP genes change under the experimental conditions. The fold-change results are presented in Table 4.4 and grouped in Table 4.5. The negative values of results indicate the down-regulation while positive ones represent the up-regulation. As it can be seen in Table 4.5, HSP genes under our experimental conditions response in different levels. Least active genes that have smaller fold-change are located in the first column. The genes in the second column are more differentially expressed than the HSP genes in the first column. HSP12, HSP42 and HSP78 have the highest expression level when they are evaluated in terms of absolute values. They are

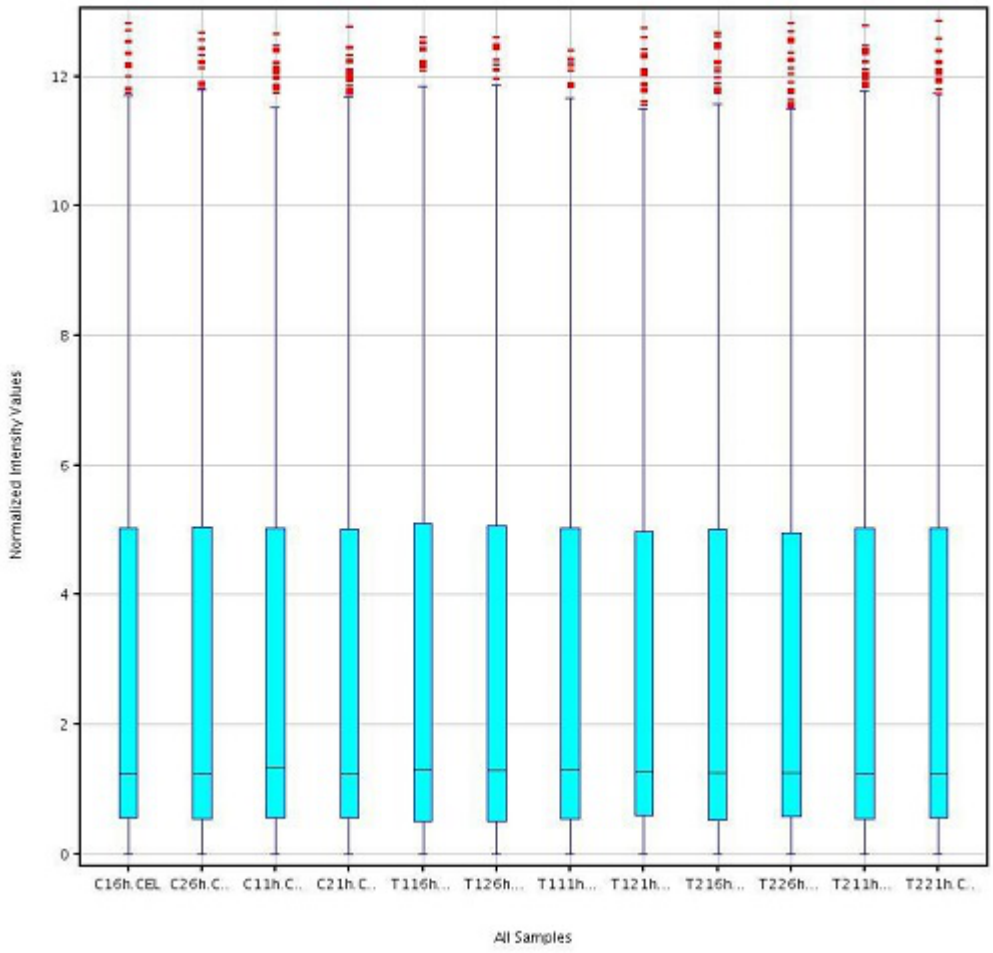

Figure 4.7: Boxplot of the RMA normalized data on the log2 scale.

down-regulated in T1/T2 condition, that is, the expression levels of these genes decrease while the temperature increases.

As stated beforehand, the fold-change is one of the common approaches in order to detect differentially expressed genes. Here different cut-off values can be chosen according to the data and to the question of interest while investigating the active genes. In this study 2 and 3 fold-changes are calculated and their associated results are represented. However ANOVA method is preferred for the further analysis as it gives more moderate results for the aim of this thesis in terms of the number of differentially expressed genes.

Table 4.2: Results of 2 fold-changes.

	T1/C	T2/C	T1/T2
Number of up-regulated genes	348	34	527
Number of down-regulated genes	232	48	279

Table 4.3: Results of 3 fold-changes.

	T1/C	T2/C	T1/T2
Number of up-regulated genes	111	16	119
Number of down-regulated genes	57	24	60

4.3 Detection of Active Genes via ANOVA

An alternative approach for the detection of the differentially expressed genes can be the ANOVA (Analysis of Variance) approach. In this analysis by choosing one of the multiple testing procedures explained in the first chapter, the null hypotheses that consider not differentially expressed genes are tested simultaneously. Before using ANOVA, the homogeneity of variances among conditions needs to be tested. In this study we checked and validated it via the levene test (Kutner et al., 2005). The test result indicates the homogeneity under the p-value 0.894 and 0.354 based on both mean and median of the signals, respectively. Then the numbers of differentially expressed genes for different error rates are compared for various computed $p-$values. Hereby the numbers of active genes out of 10928 genes under distinct significance level are presented in Table 4.6.

After performing the fold-change and ANOVA analyses, the shared genes for these two methods are recorded in Tables 4.7 and 4.8 and the full list of genes' probe set IDs are given in Appendix A. In Tables A.1 and A.3, the column for T2/C is not included as there is no common up-regulated gene under this condition.

Once the active genes are detected, they are used in the pathway analysis to investigate functionally related genes. The changes in expression levels of HSP

Table 4.4: FC results of HSP genes.

	T1/C	**T2/C**	**T1/T2**
hsp10	-1.21	1.03	-1.25
HSP10	-1.16	1.45	-1.67
HSP12	-2.06	1.02	-2.09
HSP31	-1.71	1.12	-1.89
HSP42	-1.39	1.61	-2.23
HSP60	-1.25	1.41	-1.76
HSP78	-1.42	1.54	-2.19
HSP82	1.13	1.99	-1.75
hsp90	1.04	-1.09	1.13
HSP104	-1.10	1.80	-1.99
HSP150	1.12	1.13	-1.01

genes are also investigated for the ANOVA results, but none of the HSP genes is detected as active under the taken experimental conditions of this study.

4.4 Pathway Analysis

A biological pathway represents the chemical reactions and interactions occurred while a specific biological action goes on in a cell. There are different types of biological pathways. GeneSpring software enables us to search various databases, namely, the organism specific Interaction Databases, BridgeDb databases and HomoloGene annotations while performing the pathway analysis.
After determining the differentially expressed genes, this analysis is implemented in order to understand the role of active genes in biological processes and confirm the genes that are included in different pathways, especially, in the heat shock pathway, under the experimental conditions of this study. It is decided that the active genes obtained via the Benjamini-Hochberg multiple testing procedure under the significance level α=0.05 and given in Tables 4.9, 4.10 and 4.11 can be used for the further calculations. However, the pathway analysis is also implemented for the differentially expressed genes obtained from the fold-change analysis and the results are represented in Appendix B. The pathway analysis results for differentially expressed genes obtained via the Benjamini-Hochberg

Table 4.5: FC interval of HSP genes.

	1<|FC|<1.5	**1.5≤|FC|<2**	**|FC|≥2**
T1/C	hsp10 HSP10 HSP42 HSP60 HSP78 HSP82 hsp90 HSP104 HSP150	HSP31	
T2/C	hsp10 HSP10 HSP12 HSP31 HSP60 hsp90 HSP150	HSP42 HSP78 HSP82 HSP104	
T1/T2	hsp10 hsp90 HSP150	HSP12 HSP31 HSP60 HSP82 HSP104	HSP12 HSP42 HSP78

multiple testing procedure for α=0.05 are displayed in Figure 4.8. As it is seen in the Figure 4.8, some of our differentially expressed genes are included in certain pathways. We are especially interested in the heat-shock pathway which is shown in Figures 4.9 and 4.10.The results indicate that there are 124 genes in this pathway and the expression levels of four of them change under the heat-shock stress performed in our experiment. The genes CDC28, SLT2, RAD53 are up-regulated under the conditions T1/C, T1/T2 and down-regulated under the condition T2/C while CTT1 is down-regulated under the conditions T1/C, T1/T2 and up-regulated under the condition T2/C. If it is zoomed in the heat-shock pathway as shown in Figure 4.9, these four genes can be seen closely like in Figure 4.11, separately. Hereby their shapes and connections have a biological meaning as they are illustrated in Figure 4.12. In order to have more biological information about genes, we detect them from the Gene

Table 4.6: Numbers of differentially expressed genes.

Multiple Testing Procedure	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Bonferroni FWER	1	4	5
Bonferroni Holm FWER	1	4	5
Benjamini and Hochberg FDR	4	90	290

Table 4.7: Overlap of Benjamini Hochberg FDR (α=0.05) and FC2.

	T1/C	T2/C	T1/T2
Number of up-regulated genes	33	0	41
Number of down-regulated genes	14	3	9

Ontology (GO) database. GO is a bioinformatics tool and enables us to get knowledge about genes and genes' product attributes across species as well as databases with vocabularies. Here the ontologies are collected under the three titles, named as biological process, molecular function and cellular component. By visiting the web site of Affymetrix, GO results can be queried for certain genes via probe set ID (i.e., 1778851_at) or gene symbol (i.e., CDC28). As a result, the GO findings for active genes included in the heat-shock pathway can be seen in Figures 4.13, 4.14, 4.15 and 4.16.

4.5 Clustering

After finding differentially expressed genes and investigate whether they are included in the heat shock pathway under the heat stress, different clustering methods are performed to detect which genes are similar and grouped together according to certain features.

The dendrogram of differentially expressed genes based on the Benjamini and Hochberg FDR multiple testing procedure under the significance level $\alpha = 0.05$ and the correlation distance can be seen in Figure 4.17. It displays a hierarchical structure. The probe set IDs of genes included in the clusters obtained when the branches of dendrogram are splitted into the six clusters are represented

Table 4.8: Overlap of Benjamini Hochberg FDR (α=0.05) and FC3

	T1/C	T2/C	T1/T2
Number of up-regulated genes	5	0	10
Number of down-regulated genes	5	1	4

Pathway	p-value(RMA_2p_nb)	Matched Entities(RMA...	Pathway Enti
123-58 pathway analysi...	3.2484126E-11	12	74
Transcription Regulators	4.805545E-11	12	89
Direct Interactions	1.2344551E-6	25	1058
Direct Interactions	1.5116175E-10	19	220
Direct Interactions	0.9957361	1	627
Direct Interactions	0.14624327	2	82
Direct Interactions	1.5116175E-10	19	220
MeSH heat shock pathway	0.020108648	4	124

Name	DB	DB ID	[C]	[T1]	[T2]
CDC28	Ensembl	YBR160W	5.094	6.361	4.902
SLT2	Ensembl	YHR030C	5.922	6.604	5.726
CTT1	Ensembl	YGR088W	9.761	7.191	9.738
RAD53	Ensembl	YPL153C	1.923	3.497	1.287

Figure 4.8: Results of the pathway analysis based on the Benjamini-Hochberg FDR multiple testing procedure.

in Tables 4.12 and 4.13. If it is pointed out the genes included in the heat shock pathway, it is appeared that genes CDC28, SLT2 and RAD53 are in the first cluster while CTT1 is in the third cluster. Accordingly when their GO results are investigated to determine the common features and differences of these four genes, it is occurred that there is no significant difference among these genes in terms of the cellular component. In point of the molecular function, CDC28, RAD53 and SLT2 are effective in the protein kinase activity and the kinase activity commonly while CTT1 is different from them and features in the catalase activity. In regard to the biological process, CDC28, RAD53 and SLT2 are effective in the protein phosphorylation while CTT1 features in the response to the oxidative stress.

As it is explained in Chapter 2, HIPAM is a hierarchical divisive method in which it stops the down at one point according to a validation rule. If HIPAM

Figure 4.9: Representation of the Heat-Shock pathway with 124 genes.

is performed for the differentially expressed genes based on the Benjamini and Hochberg FDR multiple testing procedure under the significance level $\alpha = 0.05$ and the correlation distance, it splits the active genes into the three clusters. The probe set IDs of active genes included in these three clusters can be seen in Tables 4.14, 4.15 and 4.16. If it is looked for the active genes included in the heat shock pathway, it is shown that genes CDC28 and CTT1 are in the first cluster while RAD53 and SLT2 are in the third cluster. CDC28 and CTT1 have only one common feature in terms of the cellular component which is also the only one effect of CTT1. They have a part in the cytoplasm. On the other hand RAD53 and SLT2 have many common features. Especially in point of the molecular function, almost all of the effect which are the nucleotide binding, protein kinase activity, protein serine/threonine kinase activity, protein binding, ATP binding, kinase activity, transferase activity and transferring phosphorus-containing groups of these two genes are the same. In regard to the biological

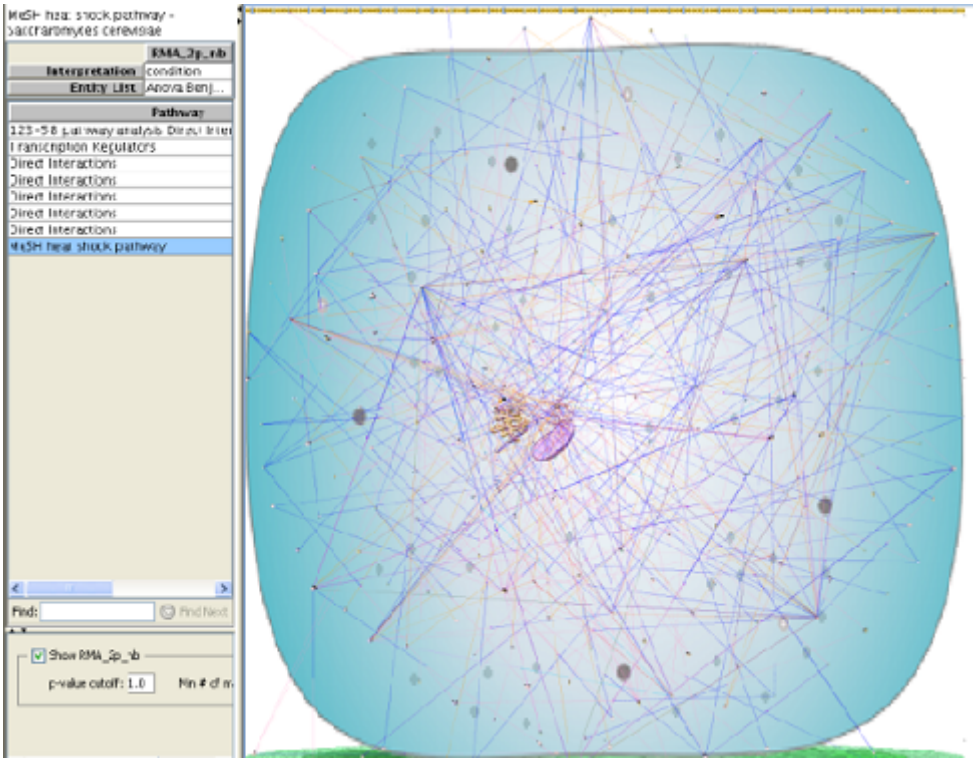

Figure 4.10: Pathway view in the GO database with 124 genes.

process, the protein phosphorylation and the phosphorylation are the common features of RAD53 and SLT2. In the sense of the cellular component, they are located in the nucleus.

As distinct from the supervised learning, the number of groups are not known in advance in the unsupervised learning. In order to decide the number of clusters, the R programme presents the `cutree` function. This function considers the distance over y-axis and uses a chosen percentile point to cut the tree obtained from the hierarchical clustering. Whereas there is a trade-off between the cut-off point and the number of clusters. Hereby the optimal result can be evaluated by performing different choices of cut-off percentiles (Erkan, 2011).

As it is performed whole genome analysis and interested in the global stress response, the hierarchical structure is not typically expected as long as a special background knowledge supports this constraint. Only a generalized hierarchical structure can be discussed if it is interested in the hierarchical interactions of a

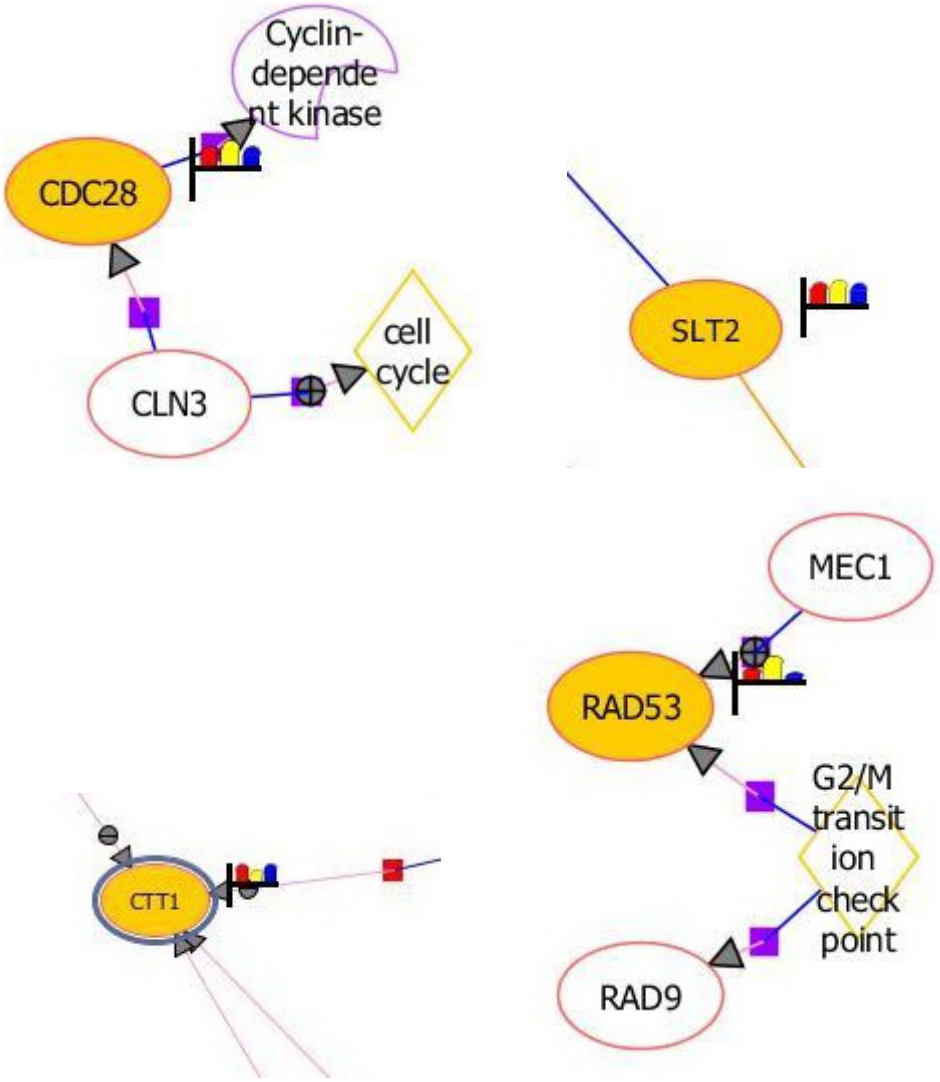

Figure 4.11: Neighborhood view of CDC28, SLT2, RAD53 and CTT1 genes in the Heat-Shock pathway.

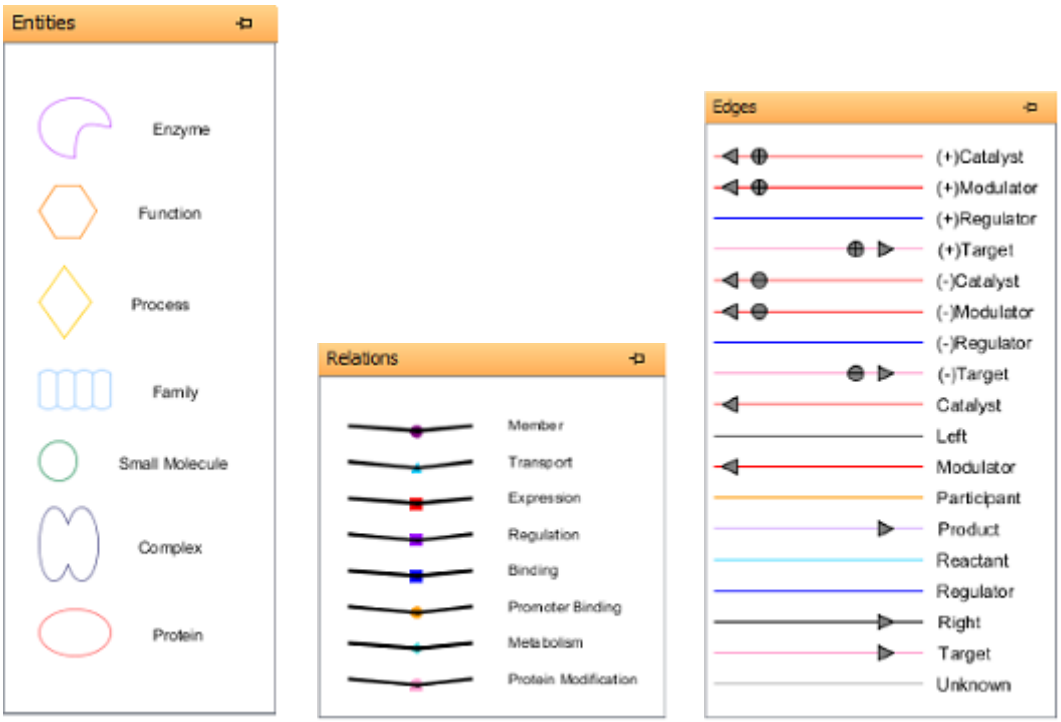

Figure 4.12: Meaning of legends for entities, relations and edges in the Heat-Shock pathway shown in Figure 4.9-4.11

specific gene (Yu, 2006). For this reason, here we focus on the non-hierarchical clustering methods for the biological validation. As non-hierarchical clustering methods, k-means and PAMSAM are applied the differentially expressed genes. These two non-hierarchical unsupervised methods are performed by using different values for the number of clusters k and the results are investigated by considering how the genes included in the heat shock pathway are separated. The results for k is equal to 5 and 9 are represented in Appendices C and D. After this point, it is decided to proceed with the PAMSAM results for biological validation because it is more robust than its alternatives as it is based on the average dissimilarities, instead of the sum of squared dissimilarities (Kaufman, 2005). However, as PAMSAM is an unsupervised learning and k is not known in advance, it is challenging to determine the optimal k. In order to overcome this problem, in this study, we prefer another method, called consensus clustering, that can enable us to decide on the optimal number of clusters (Monti, 2003). The consensus clustering is implemented via `ConsensusClusterPlus` function in R (Wilkerson et al., 2013). In the analysis optimal k is observed as one as

seen the clustering barcharts under different k values. As it is demonstrated in Figure 4.19, a smooth display can not be obtained in none of these graphes as one group always dominates until $k = 8$. Therefore it is decided that $k = 9$ can be appropriate in terms of biological the consideration. The behavior of genes in each cluster is represented in Figure 4.18. If it is looked for the active genes included in the heat shock pathway, it is shown that genes CDC28 and SLT2 are in the third cluster while RAD53 and CTT1 are in the separate clusters. CDC28 and SLT2 have many common features. If they are investigated in point of the molecular function, it is seen that they are almost participated in the common reactions which are the nucleotide binding, protein kinase activity, protein serine/threonine kinase activity, protein binding, ATP binding, kinase activity, transferase activity and transferring phosphorus-containing groups. In terms of the biological process, the protein phosphorylation and phosphorylation are the common features of CDC28 and SLT2. In the cellular component term, they are involved in the nucleus and cytoplasm.

Table 4.9: Active genes for Benjamini Hochberg FDR ($\alpha = 0.05$).

Probe Set ID	Gene Symbol	T1/C	T2/C	T1/T2
1777576_at	MTW1	up	up	up
1769591_at	NUP60	up	up	up
1773412_at	RFA1	up	down	up
1771711_at	SEN34	up	down	up
1776843_at	MCM2	up	down	up
1773826_at	ALK2	up	down	up
1778825_s_at	YBL005W-B	down	down	down
1775439_at	ETR1	down	up	down
1780080_at	HSL7	up	down	up
1777367_at	SLI15	up	down	up
1778851_at	**CDC28**	**up**	**down**	**up**
1775317_at	MRPL11	down	up	down
1774694_at	SYO1	up	down	up
1769387_at	ARP2	down	up	down
1779020_at	RPC11	up	down	up
1770307_at	PST1	up	down	up
1780104_at	DNF2	up	down	up
1775856_at	TRM1	up	up	up
1772392_at	GCD6	up	up	up
1771051_at	LYS4	up	down	up
1772049_s_at	YBL005W-B	down	down	down
1773300_at	SNU13	up	down	up
1776921_at	RAD3	up	up	up
1775387_at	GYP8	down	up	down
1780054_at	ECO1	up	up	up
1778624_at	NIF3	down	up	down
1776707_at	MCM6	up	down	up
1778806_at	MSB2	up	down	up
1769955_at	**CTT1**	**down**	**up**	**down**
1778966_at	CLC1	down	up	down
1772612_at	BUB1	up	up	up
1770410_at	YGR210C	up	up	up
1769783_at	SCW4	down	down	down
1773125_at	YGR283C	up	down	up
1778188_at	OTU2	up	down	up
1772139_at	**SLT2**	**up**	**down**	**up**
1776211_at	SET1	up	down	up
1775179_at	YHR127W	up	down	up

Table 4.10: Active genes for Benjamini Hochberg FDR ($\alpha = 0.05$).

Probe Set ID	Gene Symbol	T1/C	T2/C	T1/T2
1777403_at	IMP3	up	down	up
1770277_at	MNL1	up	down	up
1772095_at	MNI1	up	down	up
1778327_at	URM1	up	up	up
1769659_at	YPS6	down	down	down
1777962_at	GON7	up	down	up
1774138_at	ARP3	down	down	down
1772477_at	YJR124C	up	down	up
1775611_at	NMD5	up	up	up
1775663_at	ECM17	down	down	down
1777630_at	URA1	up	down	up
1769665_at	SLD2	up	down	up
1778809_at	GFA1	up	down	up
1773832_at	UTP30	up	up	up
1779773_at	CMS1	up	up	up
1773903_at	XYL2	down	up	down
1771781_at	DPH5	up	up	up
1770532_at	MET17	down	up	down
1770954_at	YLR363W-A	up	down	up
1774347_at	IKI3	up	down	up
1779745_at	TSR2	up	down	up
1771355_at	NBP1	up	up	up
1775683_at	NDI1	down	up	down
1772101_at	OGG1	up	down	up
1780031_at	DFG5	up	down	up
1769419_at	RKR1	up	up	up
1774223_at	HAS1	up	down	up
1780240_at	RFA2	up	down	up
1771976_at	STB1	up	down	up
1773745_at	SLA2	up	down	up
1770300_at	CNM67	up	down	up
1773099_at	KRE33	up	down	up
1776908_at	TOM22	down	up	down
1774653_at	AIM37	down	up	down
1778146_at	OCA2	down	down	up
1776905_at	SMM1	up	down	up
1774361_at	COQ2	down	up	down

Table 4.11: Active genes for Benjamini Hochberg FDR ($\alpha = 0.05$).

Probe Set ID	Gene Symbol	T1/C	T2/C	T1/T2
1774175_at	ESF2	up	up	up
1777980_at	DCP1	up	down	up
1779198_at	MPD2	down	up	down
1773490_at	REX4	up	down	up
1771804_at	TSR4	up	up	up
1769834_at	HST3	down	down	up
1779007_at	ARP8	down	up	down
1776391_at	MSB1	up	down	up
1775357_at	YOR283W	up	down	up
1771416_at	YOR389W	down	down	down
1770864_at	**RAD53**	**up**	**down**	**up**
1779955_at	ATG21	down	down	down
1771463_at	RPA135	up	down	up
1771501_at	DSS4	up	down	up
1778715_at	SPBC32H8.04c	down	down	up

biological process	molecular function	cellular component
meiotic DNA double-strand break processing	nucleotide binding	astral microtubule
chromatin remodeling	RNA polymerase II core binding	cyclin-dependent protein kinase holoenzyme complex
7-methylguanosine mRNA capping	protein kinase activity	nucleus
protein phosphorylation	protein serine/threonine kinase activity	cytoplasm
cell cycle	cyclin-dependent protein serine/threonine kinase activity	endoplasmic reticulum
mitosis	protein binding	spindle pole body
synaptonemal complex assembly	ATP binding	ribosome
regulation of budding cell apical bud growth	kinase activity	cellular bud neck
regulation of double-strand break repair via homologous recombination	transferase activity	
regulation of filamentous growth	transferase activity, transferring phosphorus-containing groups	
positive regulation of nuclear cell cycle DNA replication	histone binding	
positive regulation of spindle pole body separation		
positive regulation of triglyceride catabolic process		
vesicle-mediated transport		
phosphorylation		
regulation of protein localization		
negative regulation of sister chromatid cohesion		
negative regulation of transcription, DNA-dependent		
positive regulation of transcription, DNA-dependent		
negative regulation of mitotic cell cycle		
positive regulation of mitotic cell cycle		
positive regulation of transcription from RNA polymerase II promoter		
cell division		
positive regulation of meiotic cell cycle		
negative regulation of meiotic cell cycle		
phosphorylation of RNA polymerase II C-terminal domain		
negative regulation of double-strand break repair via nonhomologous end joining		

Figure 4.13: Biological knowledge for the gene CDC28 with the probe set ID 1778851_at in the Affymetrix website.

biological process	molecular function	cellular component
MAPK cascade	nucleotide binding	nucleus
barrier septum assembly	protein kinase activity	cytoplasm
response to acid	protein serine/threonine kinase activity	cellular bud tip
protein phosphorylation	MAP kinase activity	
signal transduction	protein binding	
regulation of cell size	ATP binding	
fungal-type cell wall biogenesis	kinase activity	
phosphorylation	transferase activity	
peroxisome degradation	transferase activity, transferring phosphorus-containing groups	
endoplasmic reticulum unfolded protein response		
UFP-specific transcription factor mRNA processing involved in endoplasmic reticulum unfolded protein response		
regulation of transcription factor import into nucleus		
regulation of fungal-type cell wall organization		

Figure 4.14: Biological knowledge for the gene SLT2 with the probe set ID 1772139_at in the Affymetrix website.

biological process	molecular function	cellular component
response to reactive oxygen species	catalase activity	
response to oxidative stress	peroxidase activity	
hydrogen peroxide catabolic process	oxidoreductase activity	cytoplasm
oxidation-reduction process	heme binding	
	metal ion binding	

Figure 4.15: Biological knowledge for the gene CTT1 with the probe set ID 1769955_at in the Affymetrix website.

biological process	molecular function	cellular component
DNA damage checkpoint	nucleotide binding	nucleus
nucleobase-containing compound metabolic process	DNA replication origin binding	cytosol
DNA replication initiation	protein kinase activity	
DNA repair	protein serine/threonine kinase activity	
protein phosphorylation	protein serine/threonine/tyrosine kinase activity	
response to DNA damage stimulus	protein tyrosine kinase activity	
cell cycle	protein binding	
protein localization	ATP binding	
deoxyribonucleoside triphosphate biosynthetic process	kinase activity	
phosphorylation	transferase activity	
	transferase activity, transferring phosphorus-containing groups	

Figure 4.16: Biological knowledge for the gene RAD53 with the probe set ID 1770864_at in the Affymetrix website.

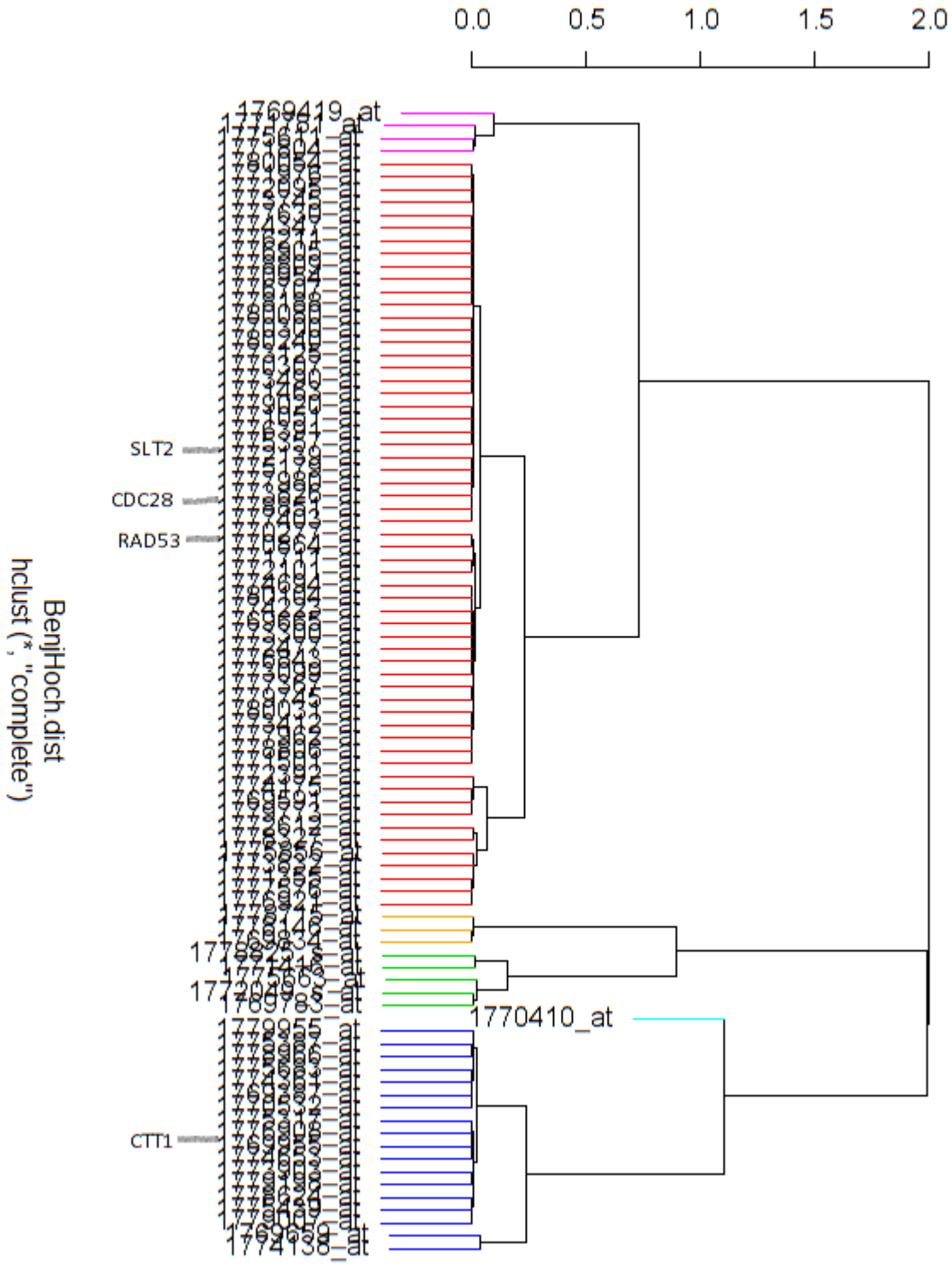

Figure 4.17: Dendrogram of differentially expressed genes based on the Benjamini and Hochberg FDR multiple testing procedure under the significance level $\alpha = 0.05$ and the correlation distance.

Table 4.12: Entities, i.e. probe set ID's, in each cluster shown in Figure 4.17.

Cluster 1		
1777576_at	1778806_at	1779745_at
1769591_at	1772612_at	1771355_at
1773412_at	1773125_at	1772101_at
1771711_at	1778188_at	1780031_at
1776843_at	**1772139_at**	1774223_at
1773826_at	1776211_at	1780240_at
1780080_at	1775179_at	1771976_at
1777367_at	1777403_at	1773745_at
1778851_at	1770277_at	1770300_at
1774694_at	1772095_at	1773099_at
1779020_at	1778327_at	1776905_at
1770307_at	1777962_at	1774175_at
1780104_at	1772477_at	1777980_at
1775856_at	1777630_at	1773490_at
1772392_at	1769665_at	1776391_at
1771051_at	1778809_at	1775357_at
1773300_at	1773832_at	**1770864_at**
1776921_at	1779773_at	1771463_at
1780054_at	1770954_at	1771501_at
1776707_at	1774347_at	

Table 4.13: Entities, i.e. probe set ID's, in each cluster shown in Figure 4.17.

Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1778825_s_at	1775439_at	1770410_at	1775611_at	1778146_at
1772049_s_at	1775317_at		1771781_at	1769834_at
1769783_at	1769387_at		1769419_at	1778715_at
1775663_at	1775387_at		1771804_at	
1771416_at	1778624_at			
	1769955_at			
	1778966_at			
	1769659_at			
	1774138_at			
	1773903_at			
	1770532_at			
	1775683_at			
	1776908_at			
	1774653_at			
	1774361_at			
	1779198_at			
	1779007_at			
	1779955_at			

Table 4.14: Entities, i.e. probe set ID's, in the cluster 1 for HIPAM.

Cluster 1				
1777576_at	1769591_at	1773412_at	1771711_at	1776843_at
1775439_at	1780080_at	**1778851_at**	1775317_at	1774694_at
1770307_at	1780104_at	1772392_at	1772049_s_at	1773300_at
1776921_at	1780054_at	1778624_at	1776707_at	1778806_at
1769955_at	1778188_at	1776211_at	1775179_at	1777403_at
1770277_at	1772095_at	1778327_at	1773832_at	1773903_at
1774347_at	1779745_at	1775683_at	1772101_at	1771976_at
1773745_at	1773099_at	1776908_at	1778146_at	1777980_at
1779198_at	1771804_at	1771416_at		

Table 4.15: Entities, i.e. probe set ID's, in the cluster 2 for HIPAM.

Cluster 2				
1778825_s_at	1777367_at	1769387_at	1775856_at	1771051_at
1775387_at	1778966_at	1770410_at	1777962_at	1772477_at
1775611_at	1777630_at	1779773_at	1771781_at	1780031_at
1780240_at	1774653_at	1776905_at	1774175_at	1769834_at
1776391_at	1775357_at	1779955_at	1771463_at	

Table 4.16: Entities, i.e. probe set ID's, in the cluster 3 for HIPAM.

Cluster 3				
1773826_at	1779020_at	1772612_at	1769783_at	1773125_at
1769659_at	1774138_at	1775663_at	1769665_at	1778809_at
1770954_at	1771355_at	1769419_at	1774223_at	1770300_at
1773490_at	1779007_at	**1770864_at**	1771501_at	1778715_at
1772139_at	1770532_at	1774361_at		

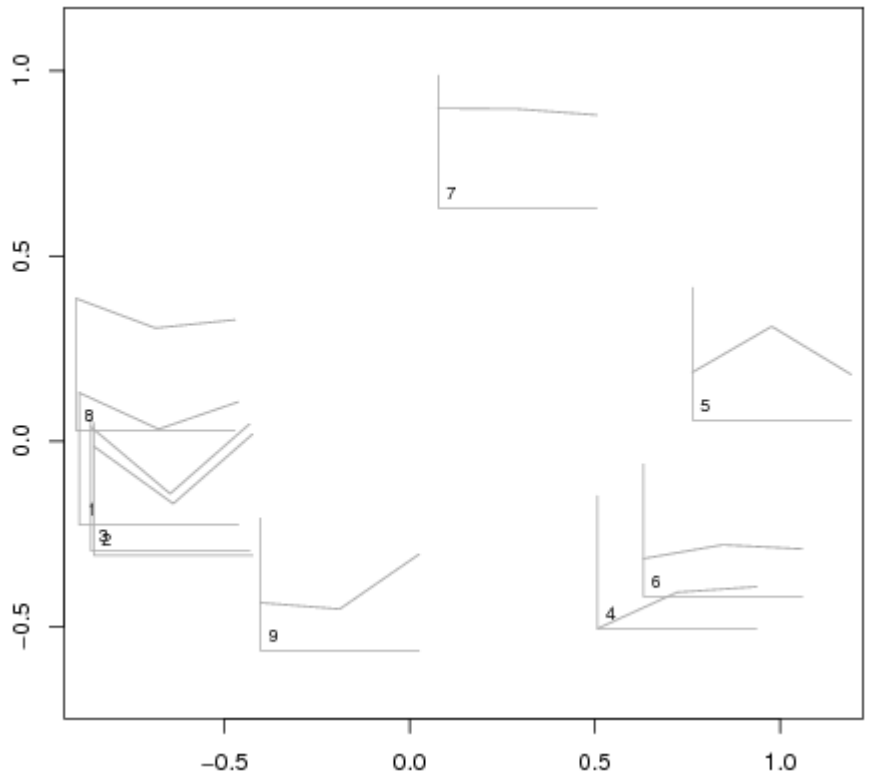

Figure 4.18: Behavior of genes under the PAMSAM for k=9.

Table 4.17: Entities, i.e. probe set ID's, in the cluster 1 under the PAMSAM method for k=9.

Cluster 1				
1777576_at	1769591_at	1775856_at	1772392_at	1776921_at
1779773_at	1771355_at	1774175_at	1773832_at	

Table 4.18: Entities, i.e. probe set ID's, in the cluster 2 under the PAMSAM method for k=9.

Cluster 2				
1773412_at	1771711_at	1776843_at	1780080_at	1777367_at
1770307_at	1780104_at	1773300_at	1778806_at	1773125_at
1777962_at	1772477_at	1769665_at	1779745_at	1772101_at
1774223_at	1780240_at	1773099_at	1773490_at	**1770864_at**
1771501_at	1774694_at	1770277_at	1780031_at	1771463_at

Table 4.19: Entities, i.e. probe set ID's, in the cluster 3 under the PAMSAM method for k=9.

Cluster 3				
1773826_at	**1778851_at**	1779020_at	1771051_at	1780054_at
1772612_at	1778188_at	**1772139_at**	1776211_at	1775179_at
1772095_at	1778327_at	1777630_at	1778809_at	1770954_at
1771976_at	1773745_at	1770300_at	1776905_at	1777980_at
1775357_at	1776707_at	1777403_at	1774347_at	1776391_at

Table 4.20: Entities, i.e. probe set ID's, in the cluster 4 under the PAMSAM method for k=9.

Cluster 4	
1778825_s_at	1771416_at

Table 4.21: Entities, i.e. probe set ID's, in the cluster 5 under the PAMSAM method for k=9.

Cluster 5				
1775439_at	1775317_at	1769387_at	1775387_at	1778624_at
1778966_at	1774138_at	1773903_at	1770532_at	1775683_at
1774653_at	1774361_at	1779198_at	1779007_at	1779955_at
1769955_at	1776908_at			

Table 4.22: Entities, i.e. probe set ID's, in the cluster 6 under the PAMSAM method for k=9.

Cluster 6			
1772049_s_at	1769783_at	1769659_at	1775663_at

Table 4.23: Entities, i.e. probe set ID's, in the cluster 7 under the PAMSAM method for k=9.

Cluster 7
1770410_at

Table 4.24: Entities, i.e. probe set ID's, in the cluster 8 under the PAMSAM method for k=9.

Cluster 8			
1775611_at	1771781_at	1769419_at	1771804_at

Table 4.25: Entities, i.e. probe set ID's, in the cluster 9 under the PAMSAM method for k=9.

Cluster 9		
1778146_at	1769834_at	1778715_at

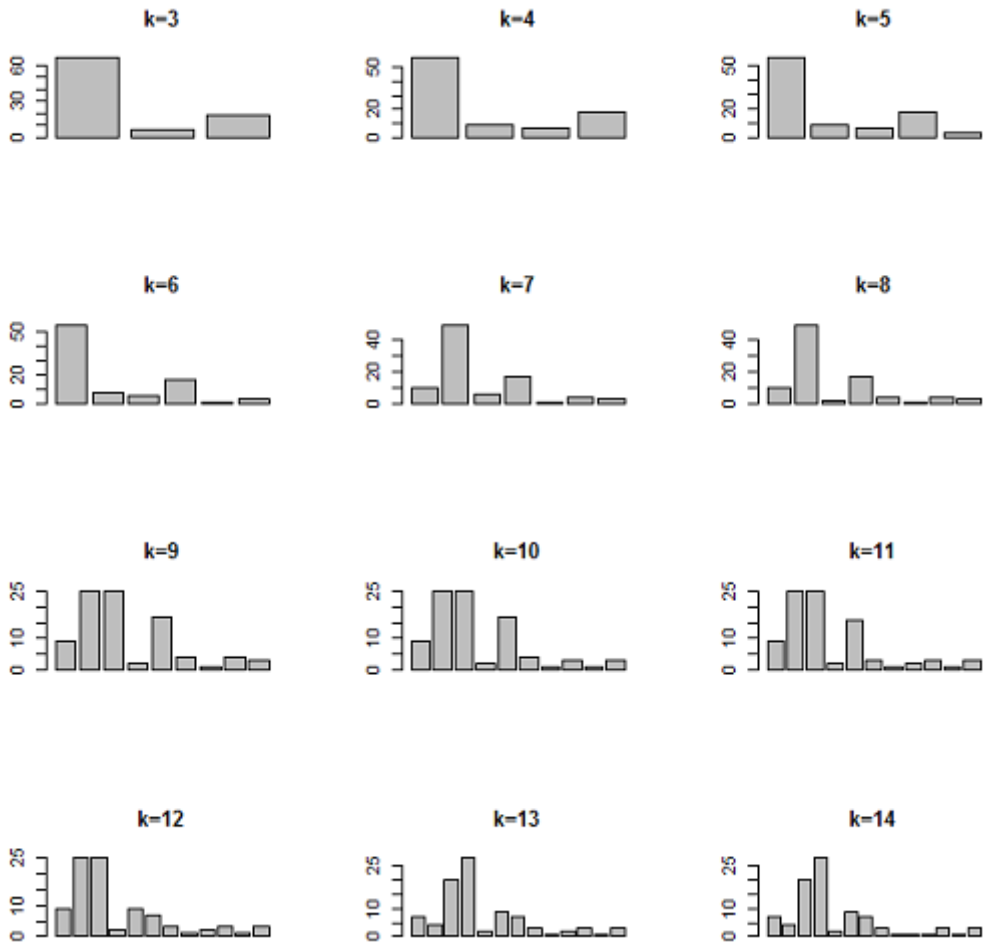

Figure 4.19: Distributions of the total number genes for different k values under the PAMSAM method.

CHAPTER 5

CONCLUSION

The aim of the study is to investigate the microarray yeast data on hand in detailed in order to find biologically significant results. For this purpose all required statistical methods are performed comparatively in each step and the best performed ones are used in the further analyses. Then the results of analyses are evaluated for their biological validations. Below we list the performed calculations for this study.

- As the first step of analyses, different background normalization methods are implemented to the raw data and the results are compared by utilizing their PCA graphs. It is decided to choose RMA since it produces the best result in terms of the grouping of the arrays.
- Next, the fold-changes for various cut-off values and ANOVA under different multiple testing procedures and significance levels analyses are performed to detect differentially expressed genes under distinct conditions. Then the further calculation is continued with the results of the Benjamini and Hochberg FDR multiple testing procedure under the significance level $\alpha = 0.05$. This analysis determines 90 active genes out of 10928 genes.
- Later, the pathway analysis is applied in order to understand the role of active genes in the biological process and to confirm the genes that are included in different pathways, especially, in the heat shock pathway, under the heat stress. The findings show that CDC28, SLT2, RAD53 and CTT1 genes included in the heat shock pathway present up or down regulation under the given experimental conditions.

- Finally, the cluster analysis is performed to classify these 90 genes according to their similarities based on certain features. For the clustering, different methods are implemented and the PAMSAM method is preferred for the final biological validation as it is a non-hierarchical method and typically gives more stable results than its alternative. As PAMSAM is a non-hierarchical method and the number of clusters k is not known in advance, i.e., the method is an unsupervised learning approach, the optimal k is detected via the consensus analysis as well.

On conclusion the yeast data used in this study is the first-time statistically analyzed and the comparative methods are performed in order to get the best results. The findings are investigated in terms of the biological validation and significant results are achieved. As explained in section about the experimental model and represented in Figure 3.2, the data consist of twelve microarrays and have intensity values obtained from the measurement for each time period after six hours and then one hour again. Here the data are considered under three conditions, namely, control, treatment 1 and treatment 2 for the presented heat effects.

For the future study, we intend to analyze the yeast data on hand in more details by constructing the conditions to answer other biologically interesting questions. For instance the significant results among rows or among their two measurements under each condition as shown in Figure 3.2 can be other biologically interesting questions. Thereby we believe that more specific knowledge can be obtained in terms of biological validation when these analyses are performed under different perspectives.

Furthermore the clustering analyses are used for the fold-change data. Hereby the genes which are significant can be also analyzed separately. Moreover, there are alternative approaches that provide different algorithms to assess the cluster stability (Volkovich et al., 2008; Barzily et al., 2009). These methods can be performed in order to determine optimal number of clusters k. Finally we consider to extend the analysis by adding the modelling of the significant genes belonging to the heat shock pathway. Here different approaches can be used. For instance we can perform MARS (Friedman, 1991), CMARS (Yerlikaya-Özkurt

et al., 2012), and RMARS (Özmen et al., 2013) which are the non-parametric and deterministic techniques which can smartly handle the non-linearity of the model and the dependence structure between measurements. Alternatively we can apply another deterministic approach called the Gaussian graphical model (Whittaker, 1990), where the interactions between the genes can be reconstructed under the conditional independence of the genes in the system under the normality assumption. Another method for modelling these genes can be the gene-environment network which uses the numerical solutions such as the Euler approximation of different level of Heun approximations in order to detect the interactions between genes in the quasi system (Thomas, 2010; Defterli et al., 2012; Defterli, 2011).

REFERENCES

[1] Affymetrix (2002). *Statistical algorithms description document.* Affymetrix, 1-28.

[2] Affymetrix (2012). *Data Sheet, GeneChip Yeast Genome 2.0 Array.* Affymetrix.

[3] Agilent Technologies (2012). *Agilent GeneSpring User Manual.*

[4] Bain, L. J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics.* Duxbury Classic Series.

[5] Barzily, Z., Volkovich, Z., Akteke-Öztürk, B., Weber, G.-W. (2009). *On a Minimal Spanning Tree Approach in the Cluster Validation Problem.* Informatica, 20(2), 187-202.

[6] Bolstad, B. M., Irizarry, R. A., Astrand, M., Speed, T. P. (2003). *A comparison of normalization methods for high density oligonucleotide array data based on bias and variance.* Bioinformatics, 19(2), 185-193.

[7] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). *Minimum information about a microarray experiment (MIAME) - toward standards for microarray data.* Nature Genetics, 29, 365 - 371, doi:10.1038/ng1201.

[8] Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. and Speed, T. P. (2004). *A benchmark for Affymetrix GeneChip expression measures.* Bioinformatics, 20(3), 323-331.

[9] Defterli, Ö. (2011). *Modern mathematical methods in modeling and dynamics of regulatory systems of gene-environment networks.* Ph.D Thesis, Mathematics Department, Middle East Technical University, Turkey.

[10] Defterli, Ö., Purutçuoğlu, V., Weber, G.-W. (2012). *Advanced mathematical and statistical tools in the dynamic modelling and simulation of gene-environment networks.* Preprint: 1-22. Chapter in: Modeling, Optimization, Dynamics and Bioeconomy. Editor: D. Zilberman and A. Pinto. Springer-Verlag.

[11] Draghici, S. (2012). *Statistics and Data Analysis for Microarrays Using R and Bioconductor.* A Chapman and Hall Book.

[12] Eisen, M. (1999). *ScanAlyze User Manual.* Stanford University, Stanford, http://rana.lbl.gov/manuals/ScanAlyzeDoc.pdf. (last accessed on 16/7/2014)

[13] Erkan, (2011). *Mixed Effects Models for Time Series Gene Expressin Data.* Ph.D Thesis, Middle East Technical University, Turkey.

[14] Everitt, B. S., Landau, S. and Leese, M. (2001). *Cluster Analysis.* London, Arnold.

[15] Everitt, B. and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R.* Springer.

[16] Ewens, W. and Grant, G. (2005). *Statistical Methods in Bioinformatics: An Introduction.* Springer.

[17] Frazee, A. C., Sabunciyan, S., Hansen, K. H., Irizarry, R. A., Leek, J. T. (2014). *Differential expression analysis of RNA-seq data at single-base resolution.* Biostatistics, 1-14, doi:10.1093/biostatistics/kxt053.

[18] Friedman, J. H. (1991). *Multivariate Adaptive Regression Splines.* The Annals of Statistics, 91(1), 1-67.

[19] Gentleman, R., Irizarry, R. A., Carey, V. J., Dudoit, S., Huber, W. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Springer.

[20] Gibbons, F. D. and Roth, F. P. (2002). *Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation.* Cold Spring Harbor Laboratory Press, 12, 1574-1581.

[21] Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference.* Marcel Dekker, Inc.

[22] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G. (1996). *Life with 6000 genes.* Science, 274(5287):546,563-567.

[23] Hubbel, E., Liu, W. and Mei, R. (2002). *Robust estimators for expression analysis.* Bioinformatics, 18(12), 1585-1592.

[24] Irizarry, R. A., Hobbs, B., Collin, F. and Speed, T. P. (2003). *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 4(2), 249-264.

[25] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York, Springer.

[26] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statisticsl Analysis.* Upper Saddle River, N.J., Pearson Prentice Hall.

[27] Kaufman, L., Rousseeuw, P. J. (2005). *Finding Groups in Data, An Introduction to Cluster Analysis.* John Wiley and Sons.

[28] Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models.* McGraw-Hill.

[29] Learn.Genetics. (2014). *Generated from the figure at learn.genetics.utah.edu/content/labs/microarray/.* (last accessed on 16/7/2014)

[30] Lee, M.-L. T. (2004). *Analysis of Microarray Gene Expression Data.* Kluwer Academic Publishers.

[31] Li, C. and Wong, W.H. (2001). *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.* Proc Natl Acad Sci USA, 98, 31-36.

[32] Lucas, A. (2014). *Package 'amap'. R package version 0.8-12.* mulcyber.toulouse.inra.fr/projects/amap/. (last accessed on 16/7/2014)

[33] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2014). *cluster: Cluster Analysis Basics and Extensions. R package version 1.15.2.*

[34] McLachlan, G. J. (2004). *Analyzing Microarray Gene Expression Data.* Wiley.

[35] Monti, S., Tamayo, P., Mesirov, J., Golub, T. (2003). *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.* Machine Learning, 52, 91-118.

[36] Moreau, Y., Aerts, S., De Moor, B., De Strooper, B. and Dabrowski, M. (2003). *Comparison and meta-analysis of microarray data: from the bench to the computer desk.* Trends in Genetics, 19 (10), 570-577.

[37] Özmen, A., Weber, G.-W. (2013). *RMARS: Robustification of Multivariate Adaptive Regression Spline, and an Application in Finance.* Institute of Applied Mathematics, Middle East Technical University, Turkey.

[38] Pollard, K. and van der Laan, M. (2002). *Resampling-based methods for identification of significant subsets of genes in expression data.* Working Paper 121, University of California, Berkeley Division of Biostatistics Working Paper Series.

[39] Purutçuoğlu, V. and Wit, E. (2007). *Fgx: a frequentist gene expression index for affymetrix arrays.* Biostatistics, 8(2), 433-437.

[40] Purutçuoğlu, V., Kayış, E. and Weber, G.-W. (2011). *Survey of background normalizations for Affymetrix arrays and a case study.* 199-219. Chapter in: Advances in Intelligent Modelling and Simulation: Simulation Tools and Applications, Springer-Verlag, Berlin Heidelberg.

[41] Purutçuoğlu, V. (2012). *Robust gene expression index.* Mathematical Problems in Engineering, doi: 10.1155/2011/182758, 1-17.

[42] Quackenbush, J. (2002). *Review: Microarray data normalization and transformation.* Nature Genetics, 32, 496-501, doi:10.1038/ng1032.

[43] Ross, S. M. (2010). *Introductory Statistics.* Academic Press, Elsevier.

[44] Schena, M. (2003). *Microarray Analysis.* John Wiley and Sons, Hoboken, NJ.

[45] Shay, E. (2003). *Microarray Cluster Analysis and Applications Review.* Institute of Evolution, University of Haifa.

[46] Stekel, D. (2003). *Microarray Bioinformatics.* Cambridge University.

[47] Thomas, D. (2010). *Geneenvironment-wide association studies: emerging approaches.* Nature Reviews Genetics, 11, 259-272.

[48] Volkovich, Z., Barzily, Z., Morozensky, L. (2008). *A statistical model of cluster stability.* Elsevier, 41(7), 2174-2188.

[49] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* New York, John Wiley and Sons.

[50] Wilkerson, M. and Waltman, P. (2013). *ConsensusClusterPlus: ConsensusClusterPlus. R package version 1.16.0.*

[51] Wit, E. and McClure, J. (2012). *Package 'smida'. R package version 1.0.* www.math.rug.nl/ ernst/book/smida.html. (last accessed on 16/7/2014)

[52] Wit, E. and McClure, J. (2004). *Statistics for Microarrays Design, Analysis, and Inference.* John Wiley and Sons Ltd.

[53] Witten, D. and Tibshirani, R. (2013). *Package 'sparcl'. R package version 1.0.3.* http://cran.r-project.org/web/packages/sparcl/index.html. (last accessed on 16/7/2014)

[54] Ye, S. Q. (2008). *Bioinformatics A Practical Approach.* Chapman and Hall/CRC Mathematical and Computational Biology Series.

[55] Yerlikaya-Özkurt, F., Batmaz, İ., Weber, G.-W. (2012). *A Review of Conic Multivariate Adaptive Regression Splines (CMARS): A Powerful Tool for Predictive Data Minig.* Institute of Applied Mathematics, Middle East Technical University, Turkey.

[56] Yılmaz, R., Akça, O., Baloğlu, M. C., Öz, M. T., Öktem, H. A., Yücel, M. (2012). *Optimization of yeast (Saccharomyces cerevisiae) RNA isolation method for real-time quantitative PCR and microarray analysis.* African Journal of Biotechnology, 11(5), 1046-1053.

[57] Yu, H., Gerstein, M. (2006). *Genomic analysis of the hierarchical structure of regulatory networks.* PNAS, 103(40), 14724-14731.

APPENDIX A

LIST OF DIFFERENTIALLY EXPRESSED GENES

Table A.1: Probe set ID's of up-regulated genes in Table 4.7.

T1/C	T1/T2	T1/C	T1/T2
1777576_at	1777576_at	1773832_at	1772095_at
1773412_at	1773412_at	1771781_at	1778327_at
1771711_at	1771711_at	1770954_at	1777962_at
1773826_at	1776843_at	1772101_at	1772477_at
1777367_at	1773826_at	1780240_at	1777630_at
1778851_at	1777367_at	1771976_at	1769665_at
1774694_at	1778851_at	1773099_at	1771781_at
1779020_at	1774694_at	1776905_at	1770954_at
1775856_at	1779020_at	1774175_at	1779745_at
1772392_at	1780104_at	1771804_at	1772101_at
1771051_at	1775856_at	1770864_at	1780031_at
1773300_at	1771051_at	1771463_at	1774223_at
1780054_at	1773300_at		1780240_at
1776707_at	1780054_at		1771976_at
1772612_at	1776707_at		1770300_at
1775179_at	1778806_at		1773099_at
1770277_at	1772612_at		1776905_at
1772095_at	1773125_at		1774175_at
1778327_at	1775179_at		1770864_at
1777630_at	1777403_at		1771463_at
1769665_at	1770277_at		

Table A.2: Probe set ID's of down-regulated genes in Table 4.7.

T1/C	T2/C	T1/T2
1778825_s_at	1778825_s_at	1775439_at
1775439_at	1772049_s_at	1772049_s_at
1772049_s_at	1769834_at	1769955_at
1769955_at		1769659_at
1769783_at		1773903_at
1769659_at		1770532_at
1775663_at		1775683_at
1773903_at		1774653_at
1770532_at		1774361_at
1775683_at		
1774653_at		
1774361_at		
1769834_at		
1771416_at		

Table A.3: Probe set ID's of up-regulated genes in Table 4.8.

T1/C	T1/T2
1771711_at	1771711_at
1775856_at	1773826_at
1771781_at	1777367_at
1780240_at	1773300_at
1774175_at	1770277_at
	1769665_at
	1772101_at
	1780240_at
	1773099_at
	1770864_at

Table A.4: Probe set ID's of down-regulated genes in Table 4.8.

T1/C	T2/C	T1/T2
1778825_s_at	1769834_at	1772049_s_at
1772049_s_at		1769955_at
1769955_at		1769659_at
1769659_at		1773903_at
1771416_at		

APPENDIX B

FOLD-CHANGE RESULTS FROM GO DATABASE

Pathway	p-value(R...	Matched ...	Pathway ...
123-58 pathway analysis Direct Interacti...	2.124202...	43	74
Transcription Regulators	1.876593...	46	89
Direct Interactions	2.614895...	310	1058
Direct Interactions	0.0	130	220
Direct Interactions	2.762342...	128	627
Direct Interactions	1.111535...	29	82
Direct Interactions	0.0	130	220
MeSH heat shock pathway	0.006762...	20	124

Name	DB	DB ID
YDJ1	Ensembl	YNL064C
CNS1	Ensembl	YBR155W
HSC82	Ensembl	YMR186W
SSB1	Ensembl	YDL229W
KAR2	Ensembl	YJL034W
CDC28	Ensembl	YBR160W
TAH1	Ensembl	YCR060W
SSE1	Ensembl	YPL106C
SSZ1	Ensembl	YHR064C
ADR1	Ensembl	YDR216W
HSP78	Ensembl	YDR258C
SSA4	Ensembl	YER103W
SRO9	Ensembl	YCL037C
CYC1	Ensembl	YJR048W
HSP30	Ensembl	YCR021C
PRC1	Ensembl	YMR297W
FES1	Ensembl	YBR101C
CTT1	Ensembl	YGR088W
RAD53	Ensembl	YPL153C
KAR1	Ensembl	YNL188W

Figure B.1: Results of the pathway analysis for 2 fold-change.

Pathway	p-value(RMA_...	Matched Entiti...	Pathway
123-58 pathway analysis Direct Interactions	1.6801294E-10	29	74
Transcription Regulators	1.8069435E-10	29	89
Direct Interactions	0.0	100	1058
Direct Interactions	1.3810697E-10	79	220
Direct Interactions	0.5320859	11	627
Direct Interactions	9.807662E-6	9	82
Direct Interactions	1.3810697E-10	79	220
MeSH heat shock pathway	0.36297235	3	124

Name	DB	DB ID
SSA4	Ensembl	YER103W
CTT1	Ensembl	YGR088W
RAD53	Ensembl	YPL153C

Figure B.2: Results of the pathway analysis for 3 fold-change.

APPENDIX C

RESULTS OF THE K-MEANS CLUSTERING

Table C.1: Entities, i.e. probe set ID's, in the cluster 1 under the k-means method for k=5.

Cluster 1			
1771711_at	1773826_at	1777367_at	1773300_at
1776707_at	1770277_at	1769665_at	1772101_at
1780240_at	1771976_at	1773099_at	**1770864_at**

Table C.2: Entities, i.e. probe set ID's, in the cluster 2 under the k-means method for k=5.

Cluster 2				
1773412_at	1776843_at	1780080_at	**1778851_at**	1774694_at
1770307_at	1780104_at	1771051_at	1778806_at	1773125_at
1772139_at	1776211_at	1775179_at	1777403_at	1772095_at
1772477_at	1777630_at	1778809_at	1770954_at	1774347_at
1780031_at	1774223_at	1773745_at	1770300_at	1776905_at
1773490_at	1776391_at	1775357_at	1771463_at	1771501_at
1779020_at	1778188_at	1777962_at	1779745_at	1777980_at

Table C.3: Entities, i.e. probe set ID's, in the cluster 3 under the k-means method for k=5.

Cluster 3				
1778825_s_at	1772049_s_at	1769783_at	1769659_at	1774138_at
1775663_at	1778146_at	1769834_at	1771416_at	1779955_at
		1778715_at		

Table C.4: Entities, i.e. probe set ID's, in the cluster 4 under the k-means method for k=5.

Cluster 4				
1775439_at	1775317_at	1769387_at	1775387_at	1778624_at
1778966_at	1773903_at	1770532_at	1775683_at	1776908_at
1774361_at	1779198_at	1779007_at	**1769955_at**	1774653_at

Table C.5: Entities, i.e. probe set ID's, in the cluster 5 under the k-means method for k=5.

Cluster 5				
1777576_at	1769591_at	1775856_at	1772392_at	1776921_at
1772612_at	1770410_at	1778327_at	1775611_at	1773832_at
1771781_at	1771355_at	1769419_at	1774175_at	1771804_at
1780054_at	1779773_at			

Table C.6: Entities, i.e. probe set ID's, in the cluster 1 under the k-means method for k=9.

Cluster 1		
1772049_s_at	**1769955_at**	1769659_at

Table C.7: Entities, i.e. probe set ID's, in the cluster 2 under the k-means method for k=9.

Cluster 2		
1778825_s_at	1769783_at	1774138_at
1775663_at	1771416_at	1779955_at

Table C.8: Entities, i.e. probe set ID's, in the cluster 3 under the k-means method for k=9.

Cluster 3				
1776843_at	1780080_at	1770307_at	1780104_at	1778806_at
1773125_at	1778188_at	**1772139_at**	1776211_at	1777403_at
1772095_at	1777962_at	1772477_at	1778809_at	1774347_at
1773745_at	1770300_at	1777980_at	1773490_at	1776391_at
1775357_at	1771501_at			

Table C.9: Entities, i.e. probe set ID's, in the cluster 4 under the k-means method for k=9.

Cluster 4			
1777576_at	1775856_at	1780054_at	1772612_at
1778327_at	1771781_at	1774175_at	

Table C.10: Entities, i.e. probe set ID's, in the cluster 5 under the k-means method for k=9.

Cluster 5				
1775439_at	1775317_at	1769387_at	1775387_at	1778624_at
1778966_at	1773903_at	1770532_at	1775683_at	1776908_at
1774653_at	1774361_at	1779198_at	1779007_at	

Table C.11: Entities, i.e. probe set ID's, in the cluster 6 under the k-means method for k=9.

Cluster 6				
1769591_at	1772392_at	1776921_at	1770410_at	1775611_at
1773832_at	1779773_at	1771355_at	1769419_at	1771804_at

Table C.12: Entities, i.e. probe set ID's, in the cluster 7 under the k-means method for k=9.

Cluster 7				
1771711_at	1773300_at	1772101_at	1780240_at	**1770864_at**

Table C.13: Entities, i.e. probe set ID's, in the cluster 8 under the k-means method for k=9.

Cluster 8				
1773412_at	1773826_at	1777367_at	**1778851_at**	1774694_at
1779020_at	1771051_at	1776707_at	1775179_at	1770277_at
1777630_at	1769665_at	1770954_at	1779745_at	1780031_at
1774223_at	1771976_at	1773099_at	1776905_at	1771463_at

Table C.14: Entities, i.e. probe set ID's, in the cluster 9 under the k-means method for k=9.

Cluster 9		
1778146_at	1769834_at	1778715_at

APPENDIX D

RESULTS OF THE PAMSAM CLUSTERING

Table D.1: Entities, i.e. probe set ID's, in the cluster 1 under the PAMSAM method for k=5.

Cluster 1				
1777576_at	1773412_at	1771711_at	1776843_at	1773826_at
1780080_at	1777367_at	**1778851_at**	1774694_at	1779020_at
1770307_at	1780104_at	1775856_at	1771051_at	1773300_at
1776921_at	1780054_at	1776707_at	1778806_at	1772612_at
1773125_at	1778188_at	**1772139_at**	1776211_at	1775179_at
1777403_at	1770277_at	1772095_at	1778327_at	1777962_at
1772477_at	1777630_at	1769665_at	1778809_at	1773832_at
1770954_at	1774347_at	1779745_at	1771355_at	1772101_at
1780031_at	1774223_at	1780240_at	1771976_at	1773745_at
1770300_at	1773099_at	1776905_at	1777980_at	1773490_at
1776391_at	1775357_at	**1770864_at**	1771463_at	1771501_at

Table D.2: Entities, i.e. probe set ID's, in the cluster 2 under the PAMSAM method for k=5.

Cluster 2			
1769591_at	1772392_at	1775611_at	1779773_at
1769419_at	1774175_at	1771804_at	1771781_at

Table D.3: Entities, i.e. probe set ID's, in the cluster 3 under the PAMSAM method for k=5.

Cluster 3		
1778825_s_at	1772049_s_at	1769783_at
1769659_at	1775663_at	1771416_at

Table D.4: Entities, i.e. probe set ID's, in the cluster 4 under the PAMSAM method for k=5.

Cluster 4				
1775439_at	1775317_at	1769387_at	1775387_at	1778624_at
1769955_at	1778966_at	1770410_at	1774138_at	1773903_at
1770532_at	1775683_at	1776908_at	1774653_at	1774361_at
1779198_at	1779007_at	1779955_at		

Table D.5: Entities, i.e. probe set ID's, in the cluster 5 under the PAMSAM method for k=5.

Cluster 5		
1778146_at	1769834_at	1778715_at