AUTOMATED BIOLOGICAL DATA ACQUISITION AND INTEGRATION USING MACHINE LEARNING TECHNIQUES

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

BY

LEVENT ÇARKACIOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

FEBRUARY 2009

Approval of the thesis:

AUTOMATED BIOLOGICAL DATA ACQUISITION AND INTEGRATION USING MACHINE LEARNING TECHNIQUES

submitted by **LEVENT ÇARKACIOĞLU** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering, Middle East Technical University** by,

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

Prof. Dr. Müslüm Bozyigit _____
Head of Department, **Computer Engineering**

Prof. Dr. Volkan Atalay _____
Supervisor, **Computer Engineering, METU**

Examining Committee Members:

Prof. Dr. Faruk Polat _____
Computer Engineering, METU

Prof. Dr. Volkan Atalay _____
Computer Engineering, METU

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

Assist. Prof. Dr. Tolga Can _____
Computer Engineering, METU

Assist. Prof. Dr. Özlen Konu _____
Molecular Biology and Genetics, Bilkent University

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 : Levent Çarkacıoğlu

Signature :

ABSTRACT

AUTOMATED BIOLOGICAL DATA ACQUISITION AND INTEGRATION USING MACHINE LEARNING TECHNIQUES

Çarkacıoğlu, Levent

Ph.D., Department of Computer Engineering

Supervisor: Prof. Dr. Volkan Atalay

February 2009, 104 pages

Since the initial genome sequencing projects along with the recent advances on technology, molecular biology and large scale transcriptome analysis result in data accumulation at a large scale. These data have been provided in different platforms and come from different laboratories therefore, there is a need for compilation and comprehensive analysis. In this thesis, we addressed the automatization of biological data acquisition and integration from these non-uniform data using machine learning techniques. We focused on two different mining studies in the scope of this thesis. In the first study, we worked on characterizing expression patterns of housekeeping genes. We described methodologies to compare measures of housekeeping genes with non-housekeeping genes. In the second study, we proposed a novel framework, bi-k-bi clustering, for finding association rules of gene pairs that can easily operate on large scale and multiple heterogeneous data sets. Results in both studies showed consistency and relatedness with the available literature. Furthermore, our results provided some novel insights waiting to be experimented by the biologists.

Keywords: bioinformatics, housekeeping genes, biclustering, association pattern discovery, gene expression profiles

ÖZ

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANILARAK OTOMATİKLEŞMİŞ BİYOLOJİK VERİ KAYNAŞIM VE KAZANCI

Çarkacıoğlu, Levent

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Volkan Atalay

Şubat 2009, 104 sayfa

İlk genom düzenleme projelerinden bu yana teknoloji, moleküler biyoloji ve büyük ölçekli transkriptom analizindeki gelişmeler çok sayıda veri birikmesine sebep olmuştur. Bu veriler farklı platformlar tarafından sağlanıp farklı labrotuvarlardan geldiklerinden dolayı derlenmiş ve ayrıntılı bir analize ihtiyaç duyulmaktadır. Bu tezde farklı veri tabanları üzerinde makine öğrenmesi teknikleri kullanarak, otomatikleşmiş biyolojik veri kaynaşım ve kazancını inceledik. Bu tez kapsamında iki farklı madencilik çalışmasına odaklandık. Birinci çalışmada ev idaresi genlerinin ifade desenlerinin nitelendirilmesi üzerinde çalıştık. Ev idaresi olan ve ev idaresi olmayan genlere ait ölçüleri karşılaştırmak için metodolojiler tanımladık. İkinci çalışmada ise, gen çiftleri için işbirliği kuralları bulmak amacı ile büyük ölçekli ve çoklu hetorejen veri kümelerinde kolayca işleyebilen yeni bir çerçeveyi, bi-k-bi kümeleme çerçevesini, önerdik. Her iki çalışmada elde ettiğimiz sonuçlar mevcut literatür ile alaka ve uyum gösterdi. Bunun yanı sıra, elde ettiğimiz sonuçlar biyologlar tarafından deneylendirilmeyi bekleyen bazı yeni kavramlar da sağladı.

Anahtar Kelimeler: bioinformatik, ev idaresi genleri, bikümeleme, işbirliği buluşu, gen ifade profilleri

ACKNOWLEDGMENTS

First of all I would like to thank very much to my supervisor Prof. Dr. Volkan Atalay for his helpful, positive and wonderful guidance during my study. I also can never thank enough to his dearest wife, Dr. Rengül Çetin-Atalay, for her wonderful comments, great patience on my never-ending pages of e-mails and suggestions about the biological aspects of this study.

I would like to thank Dr. Tolga Can for his friendly attitude and his perfect guidance on the computational aspects, Dr. Özlen Konu for her suggestions about the biological and statistical aspects of this study.

I am very glad for having a Ph.D. degree in Computer Engineering but even more glad for having this degree in the Department of Computer Engineering of **METU**. I would like to thank to every single person in this department who helps to keep this department as a big supportive family.

During the 6.5 years of this study, I worked as a full time computer engineer in Information Technologies Division of the Central Bank Of The Republic Of Turkey. I would like to thank my company and my colleagues for their support.

Finally, I'd like to thank my dearest family. I'm always grateful to my mother and father for their never-ending support. I would like to thank my brother, Dr. Abdurrahman Çarkacıoğlu, for sharing his academic views during my study but, even more, thank very much for his great brotherhood during my whole life.

To my family

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
ACKNOWLEDGMENTS vi
TABLE OF CONTENTS viii
LIST OF FIGURES x
LIST OF TABLES xi
LIST OF ABBREVIATIONS xiii
CHAPTER

1 INTRODUCTION 1
1.1 Gene Expression Profiling Using Microarray Technology 2
1.2 Problem Definition 3
1.3 Contributions 5
1.4 Organization 7

2 CONSTRUCTION OF THE LARGE SCALE MICROARRAY DATABASE 8
2.1 Gene Expression Data Construction 9
2.2 Metadata Construction 10
2.3 Normalization and Ranking Methods 12
2.3.1 Global Mean Normalization 12
2.3.2 Linear and Percentile Ranking 13

3 IDENTIFICATION OF REFERENCE GENE SETS 16
3.1 Related Work 17
3.2 Methods and Materials 19
3.2.1 Materials 19
3.2.2 Methods 19

3.3 Results . . . 22
3.3.1 Experimental Results . . . 26
3.4 Discussion . . . 27

4 BI-K-BI CLUSTERING 29
4.1 Related Work . . . 30
4.2 Methods and Materials . . . 32
4.2.1 Materials . . . 32
4.2.2 Methods . . . 32
4.2.3 Finding Gene Pairs Having Similar Expression Profiles . . . 32
4.2.4 Bi-k-Bi Clustering Algorithm . . . 36
4.2.5 Illustrative Example . . . 40
4.3 Results . . . 42
4.4 Discussion . . . 48

5 CONCLUSION 50
REFERENCES . . . 53
APPENDICES

A DATABASE DETAILS 61
A.1 Implementation Details . . . 61
A.2 Database Schema . . . 62

B RESULTS OF KOLMOGOROV-SIMIRNOV TESTS 65

C REFERENCE GENE SET OVER ALL SETS 68

D STABILITY VALUES FOR 17 HOUSEKEEPING GENES 90

E BI-K-BI CLUSTERING GENE-PAIR RULE EXAMPLE 91
VITA . . . 104

LIST OF FIGURES

FIGURES

Figure 1.1 DNA structure. 1
Figure 1.2 A sample Affymetrix microarray chip. 2
Figure 1.3 Microarray experiment flow (adapted from [68]). 4
Figure 1.4 Scanned microarray digital image (adapted from [68]). 5

Figure 2.1 Sample GDS and GPL files (adapted from [9]). 9
Figure 2.2 GDS data extraction flow. 10
Figure 2.3 Metadata construction flow. 11

Figure 3.1 Housekeeping vs. random non-housekeeping gene graphs (C_v *thresholding*). 23
Figure 3.2 Housekeeping vs. random non-housekeeping gene graphs (*PO thresholding*). 24
Figure 3.3 ROC Curve. 25

Figure 4.1 The overview of the three step methodology for mining gene expression data from multiple data sets. 33
Figure 4.2 Workflow of the bi-k-bi clustering algorithm. 37

Figure E.1 Rank values in the microarray samples of the genes in the rule . . . 103

LIST OF TABLES

TABLES

Table 2.1 Data set distribution according to Mesh Ontology nodes used in analysis 15

Table 4.1 Average percentile rank values of the genes for the illustrative example. 41

Table 4.2 Similarity measure values (P_S, C_v, ρ, ρ_w) of the gene pairs in the illustrative example where t_{Cv}=0.2 42

Table 4.3 NCBI-GEO Data sets vs. gene pairs in the illustrative example where $t_{\rho w} \geq 1.5$. 43

Table 4.4 Clusters for the illustrative example with support v_1=2. 43

Table 4.5 K-means clustering for the assignment of gene expression levels in the illustrative example: High and Low. 44

Table 4.6 Expression level assigned gene pairs vs. NCBI-GEO Microarray Samples for the illustrative example. 45

Table 4.7 Rules for the illustrative example with support v_1=2 and v_2=3. . . 45

Table 4.8 NCBI GEO datasets used in working sets. 49

Table B.1 Kolmogorov-Simirnov tests for pair wise distribution of non-housekeeping genes (C_v *thresholding*) 66

Table B.2 Kolmogorov-Simirnov tests for pair wise distribution of non-housekeeping genes (C_v *thresholding*) 66

Table B.3 Kolmogorov-Simirnov tests for pair wise distribution of housekeeping and non-housekeeping genes (*PO thresholding*)] 67

Table B.4 Kolmogorov-Simirnov tests for pair wise distribution of non-housekeeping genes (*PO thresholding*) 67

Table C.1 Reference gene set over all sets (C_v=0.12 and Sensitivity=0.5) 68

Table D.1 Stability values for 17 housekeeping genes 90

Table E.1 The most observed gene-pair rule in the breast cancer datasets . . . 91

Table E.2 GEO microarray samples used in the rule 93

Table E.3 GEO microarray samples not used in the rule 99

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
FN	False Negative
FP	False Positive
GDS	GEO Data Set
GEO	Gene Expression Omnibus
GSM	GEO Sample Microarray
GPL	GEO Platform
NCBI	National Center for Biotechnology Information
mRNA	Messanger Ribonucleic Acid
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcription Polymerase Chain Reaction
ROC	Receiver Operator Characteristic
SENS	Sensitivity
SPEC	Specificity
TN	True Negative
TP	True Positive

CHAPTER 1

INTRODUCTION

The cell is the structural and functional unit of all known living organisms. Cells provide structure for the body, convert food to energy, and carry out specialized functions. One of the most important functions of the cell is to replicate itself. During replication, the hereditary information is also carried to its offsprings. The hereditary information of the cell is coded in a structure called Deoxyribonucleic Acid (DNA), as shown in Figure 1.1.

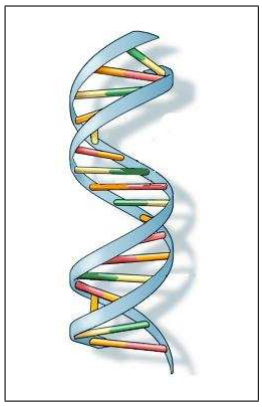

Figure 1.1: DNA structure.

The basic physical and functional unit of heredity in the DNA is called gene. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. Recent studies estimated that humans have 20,000 to 25,000 genes.

Many functions in the cell are performed by large and complex molecules called proteins. In the cell body, proteins are produced by structures, named as ribosome. The information needed to make proteins is coded in genes. The information transfer

from gene to ribosome in the cell takes place in two steps. In the first step, named as transcription, the information in the gene is transferred to the messenger Ribonucleic Acid (mRNA). In the second step, named as translation, the mRNA carrying the information interacts with the ribosome. These two steps together, transcription and translation, are named as gene expression.

The study of gene expression is a very important and complex process that allows scientists to understand the responses of a cell to the changes in its environment.

1.1 Gene Expression Profiling Using Microarray Technology

Due to the advances in technology, it has been possible to deposit very small volumes of many objects into a very small area. A microarray is an arrayed series of thousands of microscopic spots, printed on a solid substance such as glass, plastic or silicon biochip, as shown in Figure 1.2. These arrayed chips allow scientists to analyze expressions of thousands of genes together.

Figure 1.2: A sample Affymetrix microarray chip.

The analysis of gene expressions using microarray chips consists of two steps. In the first step, a biological experiment is performed. The flow of a microarray experiment is given in Figure 1.3. RNA is first isolated from different tissues, developmental stages or samples subjected to appropriate treatments. RNA is then labeled and hybridized to the arrays using an experimental strategy that allows expression to be assayed and compared between appropriate sample pairs. Common strategies include the use of a single label

and independent arrays for each sample, or a single array with distinguishable fluorescent dye labels for the individual RNAs [42, 51]. Two different mRNA populations, labeled with different fluorescent dye (Cy5 red, Cye3 Green) and are excited by a laser. Each fluorescent dye excites at a different wavelength, which is captured using a photo detector attached to a filter tuned to the particular fluorescent label.

In the second step, experimental results of the first step are collected. A microarray scanner is used to detect the fluorescent labels on each spot of the microarray. The more fluorescent a spot is, the more copies of that gene were present in the RNA sample. The microscope and camera work together to generate digital images of the array during this scan. A sample image is shown in Figure 1.4. A special program is then applied for quantization of the digital image. There exist many commercial and free software packages for microarray image quantization. Although there are minor differences between these software packages, most of them give high-quality, reproducible measures of hybridization intensities [51]. They typically apply many image processing techniques on the image and generate expression data of each spot in the array by subtracting background data from the spot.

Microarray technology allows researchers simultaneously monitor expression levels of thousands of genes in a single experiment. These experiments are valuable tools in the understanding of genes, biological networks, and cellular states. Studies of microarray experiments have targeted many important goals, such as finding differentially expressed genes, defining pathways, drug targeting, clustering.

Measurement of the expression values of thousands of genes in a microarray experiment is called, gene expression profiling. Gene expression profiling is based on the fact that during the transcription step of gene expression, only a fraction of the genes are expressed.

1.2 Problem Definition

In general, a microarray experiment is performed for a particular type of research. The excess in the amount of microarray experiments with the research resulted in significant data accumulation. Since these experiments have been provided in different platforms and come from different laboratories, it is necessary to compile and analyze comprehensively. In recent years, applications of data mining techniques for gene expression profiles

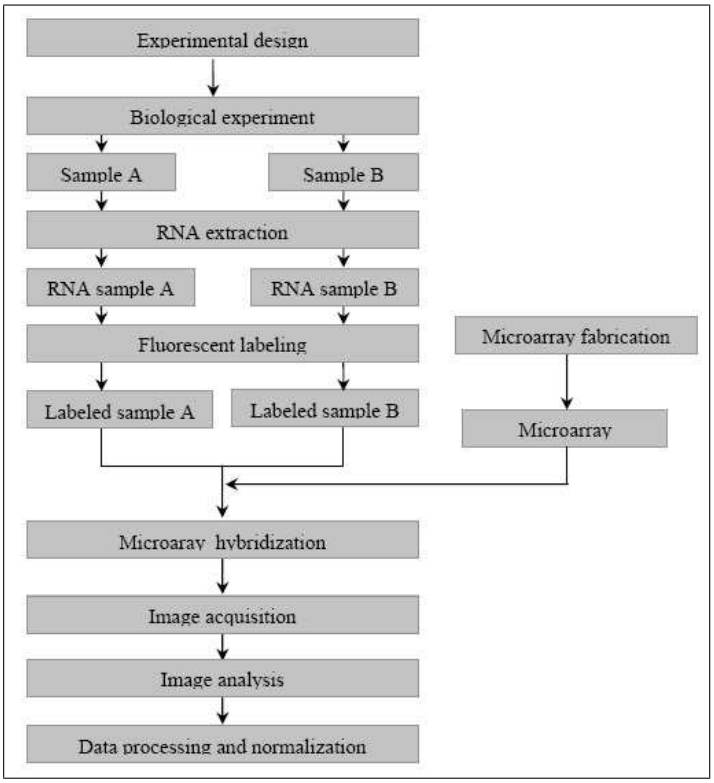

Figure 1.3: Microarray experiment flow (adapted from [68]).

using microarray experiments have became a matter of interest. In the very early stage of microarray technology few experiments and spreadsheets were enough for analysis. However, due to increase in gene expression data sets, spreadsheets have been less adequate tools. Therefore, application of machine learning techniques on gene expression data sets has became more popular. At first, supervised and unsupervised machine learning techniques have been applied on gene expression data. However, these methods have several shortcomings. Therefore, in the recent years, biclustering, association pattern discovery and pattern based clustering methods have also been studied. Most of the mining studies on gene expression profile analysis in the literature target single gene expression data set and they cannot handle large amount of gene expression data that exists in public databases in reasonable amount of time and space.

The basic problem is the automation of biological data acquisition and integration of non-uniform microarray experiments submitted by different experimenters using machine learning techniques. Specifically, the problem is to mine patterns of gene expression profiles using multiple heterogeneous microarray data sets that can easily work on a

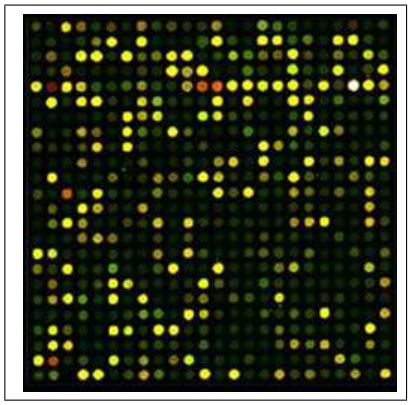

Figure 1.4: Scanned microarray digital image (adapted from [68]).

desktop computer with limited computational resources.

1.3 Contributions

In this thesis, we first constructed a large scale database consisting of microarray data along with the associated metadata.

For this purpose, we first downloaded all available public *Homo sapiens* microarray data sets and stored expression values for the spots of the microarray experiments in the database. We then queried, constructed and stored the associated metadata for these data sets. In order to make experiments comparable, we applied normalization methods on this heterogeneous large scale database. Global mean normalization and ranking methods has been applied on the database. After the construction of this large scale database, we applied data mining techniques on the database. We mainly focused on two different mining problems in the scope of this thesis. In the first problem, we worked on characterizing expression patterns of housekeeping genes. Genes involved in cellular maintenance functions and generally assumed to have expression levels unaffected by experimental conditions are called housekeeping genes. We defined a set of measures and described a methodology to compare measures of housekeeping genes with non-housekeeping genes. In the first step of the methodology, we defined a measure for the rank change of a gene within a set. In the second step, we defined a measure for the ratio of the sets that a gene has rank change under a given threshold. However, the ratio measure by itself does not give statistically significant results. Therefore, we defined a measure for the percentage of occurrence of a gene in the third step. In the last step, we

computed and analyzed the ratio and percentage of occurrence values of housekeeping and non-housekeeping genes for different threshold values.

Recent studies showed that several widely used housekeeping genes might have altered expression under different experimental conditions. In this problem, we tested whether a reduced set of genes is invariably expressed across different experimental conditions so that they can be used as reference genes. Our contributions in this problem can be summarized as follows.

- We defined a scaling process. By scaling all expression values into a comparable platform, we were able to work on multiple data sets. We applied our scaling process on approximately 142 million microarray spots from 9090 microarray samples grouped into 381 data sets. However, previous studies in this area generally worked on a few homogenous or specially curated data sets.

- Results have proven the claim that expressions of housekeeping genes are less variable across different experiment sets when compared with the expressions of randomly selected gene sets.

- We showed that cell specific reference gene sets are less variable than housekeeping genes in terms of gene expression profiles. However, all previous studies focused to find a general reference gene set.

In the second problem, we proposed a novel framework, called bi-k-bi clustering, for finding association rules of gene pairs. The bi-k-bi clustering framework can easily operate on large scale and multiple heterogeneous data sets.

The framework consists of three steps. In the first step, we applied normalization in order to scale all microarray samples in our database into a comparable platform. In the second step, we defined a function to find gene pairs with similar expression profiles from the first step. In the last step, we applied the bi-k-bi clustering algorithm to the gene pairs that have similar expression profiles and construct rules consisting of gene pairs and associated samples. The bi-k-bi clustering algorithm in the third step consists of three phases. In the first phase, a coarse analysis is performed and the working set has been reduced by focusing on gene pairs having similar expression profiles with their associated data sets. In second phase, a label for each gene in each experiment is assigned using the k-means clustering algorithm. In the last phase, a detailed analysis is done on the

labeled gene pairs with their associated microarray samples. The framework is tested on all available data sets and more specifically five different functionally concise groups of data sets independently.

Our contributions in this problem can be summarized as follows.

- We defined a similarity measure to find gene pairs having similar expression profiles.
- We defined a coarse to fine approach to work on large scale data in reasonable amount of time and space. At the coarse stage, we eliminated gene pairs having less similar expression profiles by using our similarity measure. At the fine stage, we performed detailed analysis on the gene pairs having similar expression profiles.
- By the use of dynamic thresholding on expression profiles, we alleviated the disadvantages of crude thresholding on expression data.
- We extended a maximum frequent itemset algorithm to a biclustering algorithm because available biclustering algorithms have high space complexity.

1.4 Organization

Organization of this manuscript is as follows. The next chapter describes the details for the construction of the large scale microarray database used in our studies. Third chapter presents the problem of characterizing expression patterns of housekeeping genes. Fourth chapter describes our novel bi-k-bi clustering framework. Chapter 5 concludes and gives some future directions on the thesis.

CHAPTER 2

CONSTRUCTION OF THE LARGE SCALE MICROARRAY DATABASE

The excess in the amount of microarray experiments requires the collection and service of these data under a standard format. NCBI Gene Expression Omnibus (GEO) project is a public repository for microarray experimental data submissions that consists of thousands of microarray experiments [9]. GEO is a curated, online resource for gene expression data browsing, query and retrieval.

A *microarray sample* is a biological experiment performed on a single microarray chip to monitor expression levels of thousands of genes. Microarray samples experimented for specific studies are grouped into data sets. Each dataset is associated with a platform file; and association of spots between datasets and platforms are established using common spot identifiers. GEO provides gene expression data in terms of microarray samples (GSM) grouped in to datasets (GDS) with their associated platforms (GPL). Datasets and platforms are submitted by different experimenters and they reside as separate files in the GEO repository. A sample GDS file and its associated GPL file in the GEO repository is given in Figure 2.1.

The remainder of this chapter is organized as follows. In the first section, we give detailed information about the construction of gene expression data in our database. In the next section, we describe the metadata construction for the data sets in our database. In the last section, we describe normalization and ranking methods applied on the expression values in our database. Database schema and implementation details of the database are given in Appendix A.

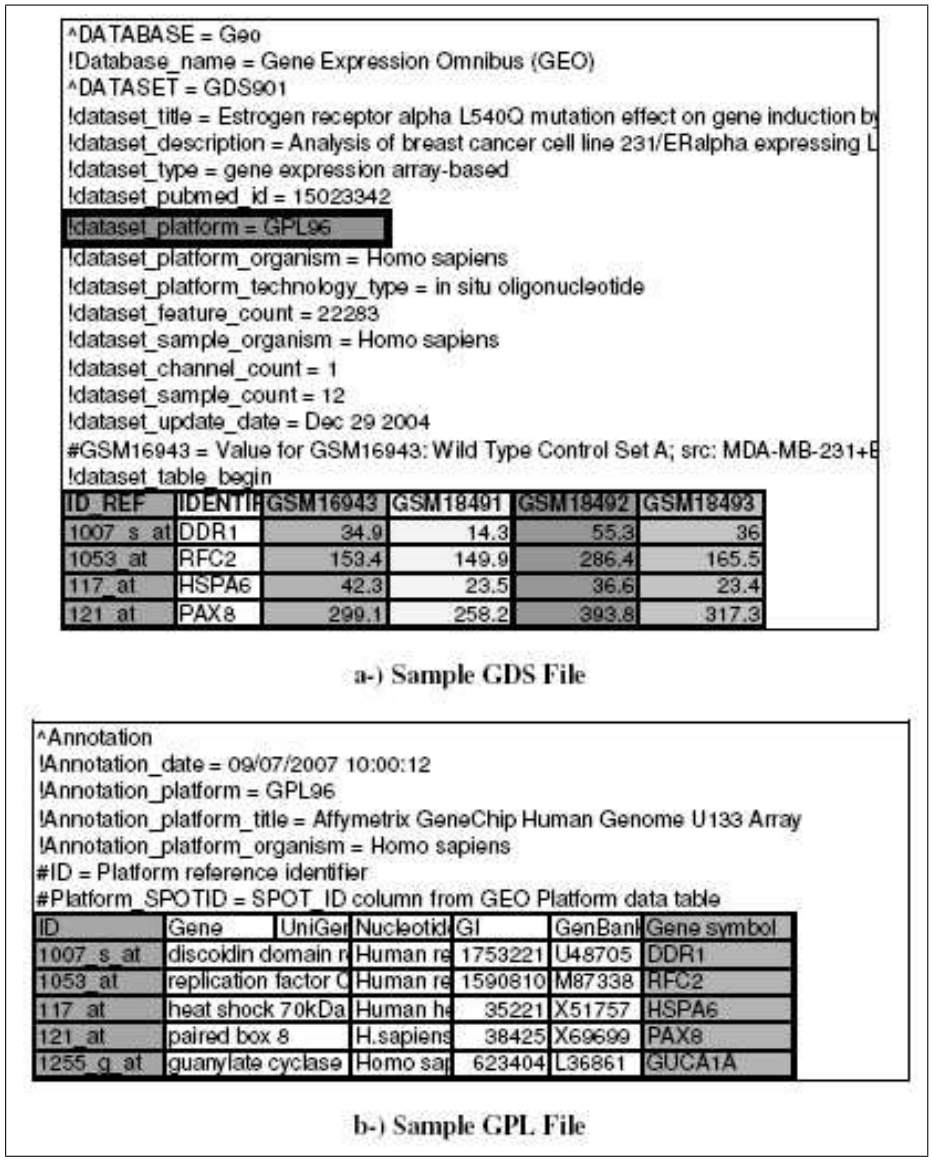

Figure 2.1: Sample GDS and GPL files (adapted from [9]).

2.1 Gene Expression Data Construction

We have downloaded all available public *Homo sapiens* microarray data sets with their associated platform files from NCBI-GEO (August, 30, 2006) and stored these files in our server. Then, we developed an application to process these files and store the processed expression values in our database. The execution flow for this process is given in Figure 2.2.

Since, NCBI-GEO contains microarray experiments submitted by different experiments all over the world, there can occasionally be misconfigured data sets or misleading

spot identifiers. We marked these sets as suspicious and ignored these data sets during processing.

NCBI-GEO gene expression data may occasionally contain spots with missing expression values due to noise or unobserved signal in microarray images. There are two common strategies employed by previous work to cope with missing spot values: the first strategy is to treat the missing spot's expression value as zero, while the second strategy is to ignore the spot. We used both strategies to construct two different data sets. The first data set, in which missing spot values were treated as zero, contains 155 million microarray spots from 9090 microarray experiments grouped into 381 GEO data sets. The second data set, in which missing spots were ignored, approximately contains 142 million microarray spots from 9090 microarray experiments grouped into 381 GEO data sets.

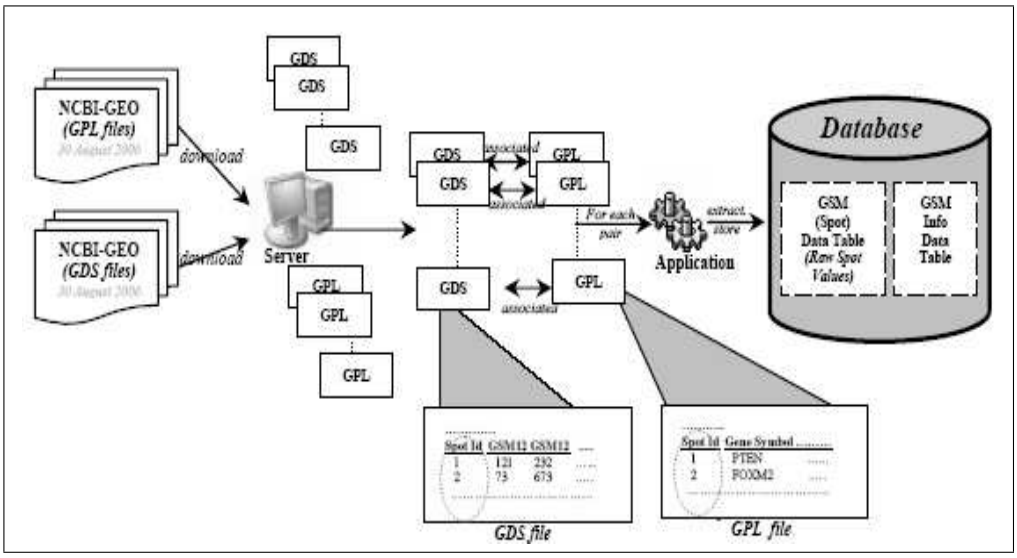

Figure 2.2: GDS data extraction flow.

2.2 Metadata Construction

In general, a set of microarray experiments is conducted for a particular type of research and referenced in medical publications. Such related experiments are grouped into GEO data sets and each data set is associated with its respective publications. Most of these

publications are categorized in a hierarchical structure using the controlled vocabulary provided by the Medical Subject Headings (Mesh Headings) ontology. In order to store the metadata for the data sets in our database we developed three applications. In the first application, we downloaded the Mesh Ontology nodes from NCBI-Mesh Ontology database and stored these nodes in our database. In the second application, we queried, extracted, and then stored mesh headings of the associated publications from NCBI-PubMed, NCBI-MESH Heading databases for the data sets in our database. Finally, in the last application, we grouped data sets in our database into a hierarchical tree structure with their associated mesh headings by using the Mesh Ontology. The execution flow for metadata construction is given in Figure 2.3.

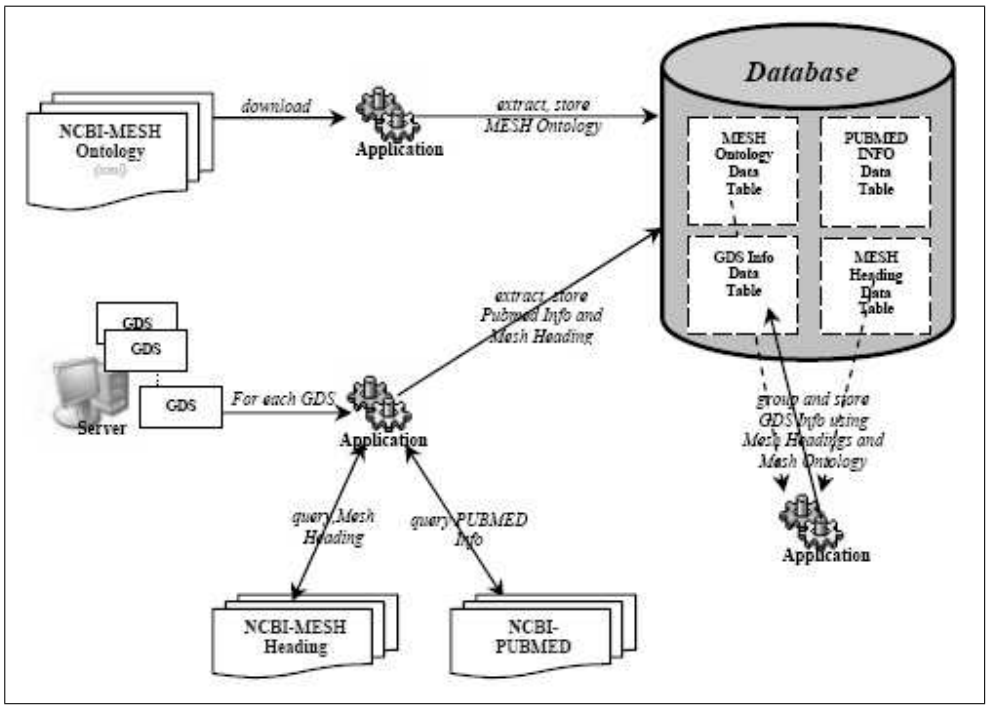

Figure 2.3: Metadata construction flow.

Of the 381 GEO data sets in our database, 341 are associated with 272 different medical publications and 264 of these publications are associated with 5754 Mesh headings. The grouping of data sets according to the selected Mesh Ontology nodes that we used in our analysis is given in Table 2.1.

2.3 Normalization and Ranking Methods

Data normalization is a process by which multiple sources of data are made comparable, quantitatively. Yet, there are many different platforms from which microarray data are extracted. Furthermore, these data sets include differential number of probe sets or clones printed, some which belong to the same transcript. The platforms also differ in their use of normalization, background correction and summarization protocols, which are necessary for preprocessing of the microarray data. Various normalization and/or data integration methodologies have been proposed [66]. Among these global normalization and normalization based on invariant probe sets have been widely used. Different scaling and normalization methods by examining their performance in terms of rank-intensity plots (RIPs, intensities plotted against ranks) have been studied [29, 36, 46]. The results in the previous studies indicated that gene expression data exhibit similar RIPs. Based on this finding, ranking genes might represent a unifying method for integration and comparison of multiple data sets without extensive manipulation of the original data distribution.

In order to analyze gene expression values across several microarray samples, we defined and applied a two step methodology on each sample in our database:

1. Global mean normalization over all spots in a microarray sample.

2. Linear and percentile ranking over the normalized spots within the same sample.

2.3.1 Global Mean Normalization

We preferred global mean normalization since it is one of the least intrusive methods available and does not alter the distribution of microarray data. Accordingly, a log-based sample mean value is subtracted from each spot's log based expression value, for each microarray sample in the database. Let n be the number of spots in a microarray sample and E=$\{ch_1, ch_2, ..., ch_n\}$ be the set of channel values of the spots within the microarray sample. The function for global mean normalization of a spot channel value, $f_{GBN}(\mathrm{ch}_x)$, is given in Equation 2.1.

$$f_{GBN}(ch_x) = \log(ch_x) - \log\left(\frac{\sum_{i=1}^{n} ch_i}{n}\right) \tag{2.1}$$

2.3.2 Linear and Percentile Ranking

Cross-experiment normalization can be achieved by using ranking methods. We applied linear and percentile ranking methods independently on the normalized channel values within each experiment in each GEO data set. This additional process provides a rank for each gene within an experiment. The rank measure is comparable across experiments and allows us to analyze the behavior of a gene on a larger scale.

Let n be the number of spots in a microarray sample and E= $\{ch_1, ch_2, ..., ch_n\}$ be the set of channel values of the spots within the microarray sample. Functions we have used for linear, $\text{Rank}_{Linear}(\text{ch}_x)$, and percentile ranking, $\text{Rank}_{Percentile}(\text{ch}_x)$, are given in Equation 2.2 and Equation 2.5 respectively. It is important to note that, each spot is ranked among the other spots only within the sample it belongs to.

$$Rank_{Linear}(ch_x) = \frac{f_{GBM}(ch_x) - Min(f_{GBM}(E))}{Max(f_{GBM}(E)) - Min(f_{GBM}(E))} \times 100 \tag{2.2}$$

$$Q_b(ch_x) = |f_{GBM}(ch_i)|, \quad \forall \quad ch_i \in E \quad \text{where} \quad f_{GBM}(ch_i) < f_{GBM}(ch_x) \tag{2.3}$$

$$Q_a(ch_x) = |f_{GBM}(ch_i)|, \quad \forall \quad ch_i \in E \quad \text{where} \quad f_{GBM}(ch_i) = f_{GBM}(ch_x) \tag{2.4}$$

$$Rank_{Percentile}(ch_x) = \frac{\sum_{i=1}^{n} Q_b(ch_x) + 0.5 \times Q_a(ch_x)}{n} \times 100 \tag{2.5}$$

The | | operator in the Equation 2.2 and Equation 2.5 indicates the cardinal function.

Each spot in a microarray sample is associated with at least one gene symbol. Some genes may also have multiple probes within the same sample. In order to work with a single rank value for each gene probe in the sample, we calculate, for each gene, an average change of its rank within the sample. Let GS =$\{G_1, G_2, ..., G_k\}$ be the set of k genes and S =$\{e_1, e_2, ..., e_n\}$ be a set of n microarray samples. For each gene G_i in a sample e_i, a single rank value, $\text{r}(\text{G}_i, \text{e}_i)$ is computed. For a gene that occurs multiple probes in a sample, if the variation across rank values of the spots of that gene probe does not exceed a threshold then the average gene probe rank value is used. Otherwise all the probes for that gene are ignored for this sample. We plotted the average rank change changes of genes that have multiple probe representations. We then, analyzed

these changes and experimentally noticed that more than 20% change in the rank values of the multiple probes of a gene, is most probably due to noise in the experiment and we concluded that all the probes of that gene can be ignored. Given a set $\mathbf{S}$, that has n samples, we have at most n rank values for a gene G_i.

Table 2.1: Data set distribution according to Mesh Ontology nodes used in analysis

Mesh Tree Number	**Mesh Tree Name**	**Number of Sets**
ALL	ALL	381
A10	Tissues	77
A10.272	Epithelium	16
A10.690	Muscles	46
A11	Cells	223
A11.118	Blood Cells	47
A11.148	Bone Marrow Cells	14
A11.251	Cells, Cultured	157
A11.284	Cellular Structures	40
A11.329	Connective Tissue Cells	34
A11.436	Epithelial Cells	48
A11.627	Myeloid Cells	17
A11.733	Phagocytes	14
A11.872	Stem Cells	22
A15	Hemic and Immune Systems	74
A15.145	Blood	51
A15.378	Hematopoietic System	14
A15.382	Immune System	60
C04	Neoplasms	108
C04.557	Neoplasms by Histologic Type	68
C04.588	Neoplasms by Site	68
C04.697	Neoplastic Processes	16

CHAPTER 3

IDENTIFICATION OF REFERENCE GENE SETS

Recent advances in large scale transcriptome analysis resulted in significant data accumulation. Since these data have been provided in different platforms and come from different laboratories, there is a need for compilation and comprehensive analysis. To analyze a gene of interest across several array experiments from different sources, the gene's expression data must be scaled in a comparable platform. Various normalization methods are available to scale microarray data within the same experiment, yet it becomes problematic to compare arrays of different sources without using references.

Previous studies have provided gene lists involved in cellular maintenance functions; thus these genes are called housekeeping genes that are generally assumed to have expression levels unaffected by experimental conditions [23, 30]. However, recent studies indicated that several widely used housekeeping genes might have altered expression under different experimental conditions [60, 63]. Therefore, it is essential to confirm the reliability of available housekeeping gene sets as well as to determine whether there are other invariably expressed gene set(s) under large number of experimental conditions.

In this study, we aimed to test whether a reduced set of clones/probes is invariably expressed across various biological phenomena so that they can be used as reference genes. We have selected a set of housekeeping genes by Eisenberg et al. (2003) and Hsiao et al. (2001). We analyzed the expression patterns of these selected genes in NCBI-GEO *Homo sapiens* datasets in our database.

The remainder of this chapter is organized as follows. In the first section, we give brief information about the related work done in this area. In the second section, we present

detailed information about methods and materials used in the study. Computational and experimental results are presented in the third section. Discussions of the study are provided in the last section.

3.1 Related Work

During the last decade, there has been remarkable progress in the identification of transcriptome blueprint of human cells through small or large-scale quantitative gene expression studies. Since the gene expression is major the determinant of the protein abundance in the cell, transcriptome analysis experiments have been widely applied to reveal the molecular mechanisms of disease conditions. Depending on the cell fate expression levels of genes varies. Previously it was assumed that there are certain genes that care and manage cellular activities in different tissues and named as housekeeping genes. But recent studies on identifying genes, which are ubiquitously and constantly expressed, demonstrated that there are not such genes for all cell types. Housekeeping genes were traditionally defined as ubiquitous genes (i.e., widely and persistently expressed in all tissues) that perform necessary functions in normal cellular physiology. Accordingly, Warrington et al. (2000) identified 535 transcripts that can qualify for the above description. Similarly, Hsiao et al. (2001) reported a set of 451 maintenance genes 358 of which were common with those reported by Warrington et al. (2000). Many of the housekeeping genes were found to be highly expressed; moreover, they belonged to functional groups involved in cellular metabolism, protein biosynthesis and cell signaling [30]. A cluster analysis based on these 451 housekeeping gene expression data could accurately cluster related tissue samples together suggesting that many of them also were differentially expressed among these 19 tissues. Later in another study, 575 human genes were identified as constituvely expressed in all tissues [23]. This study also pointed out that the gene structure of housekeeping genes are compact in terms of their organization in genome. Lercher et al. (2002) have showed that ubiquitously expressed genes cluster in their distribution throughout the genome base on SAGE (Serial Analysis of Gene Expression) datasets. This finding also explained why highly expressed genes form tight genomic expression modules since many of the ubiquitous genes were highly expressed. More recently, Su et al. (2002) compiled expression of multiple independent tissues from human and mouse to find that 6% of all genes studied were expressed ubiquitously yet

differentially.

On the other hand, recent expression technologies require the determination of a set of ubiquitously yet constantly expressed gene set(s) across biological phenomena (e.g., developmental stage, response to stress, pathological conditions) for accurate normalization of microarray as well as real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) datasets which are used for amplifying defined pieces of a RNA molecules. For example, Warrington et al. (2000) have found only a subset of genes (i.e., 47 out of 535) as being not differentially expressed across multiple tissues. Hsiao et al. (2001) also reported a smaller subset out of 451 as being relatively constantly expressed. Therefore, it is crucial to determine whether such a gene sets exist among known housekeeping gene sets using available microarray data that represent multiple normal, experimental and pathological conditions. Furthermore, it is important to determine whether condition-specific subsets of genes could be extracted as reference sets for normalization studies.

Several attempts have been made to compare microarray data from different platforms and species. Studying experiment sets individually and comparison of these individual results to obtain a meta-statistics also have been performed [39]. One recent study, Pan et al. (2006) used a graph-theoretical approach in which individual microarray datasets were analyzed for co-expression links within datasets and then consistent co-expression links among experiments were extracted. Yoon et al. (2006) have applied a normalization method that incorporates gene-specific mean and standard deviation information from various GEO microarray datasets. They implemented within array standardization and gene-specific multi-array normalization to integrate multiple microarray datasets. Their method allowed for redistribution of the sample mean of a selected gene across all available biological cases. Integration of multiple microarray datasets for functional protein network detection has been recently proposed [31]. In this study, 40 different datasets were obtained using dual channel and single channel methods from GEO to extract normalized correlations between gene pairs [31].

Recent studies have focused on finding appropriate genes for normalization in many cancers such as colon carcinoma, breast carcinomas, renal cell carcinoma and bladder cancer [6, 34, 47, 47]. A recent study showed that ACTB, GAPDH and TBP were differentially expressed in eight pathological stages of hepatitis C virus induced hepatocellular carcinoma (HCC), and thus are inappropriate for normalization [64]. Another study investigated the expression levels of candidate genes in hepatitis virus B related hepato-

cellular carcinoma (HCC), comparing malignant and non-malignant samples [25]. This study also showed that GAPDH and ACTB are regulated during HCC carcinogenesis. On the contrary, a study aiming to validate reference genes in Caco-2 cells during differentiation, identified GAPDH and ACTB as most stable genes among six candidates [50]. Reference gene expression was also studied in neutrophils, reticulocytes, and fibroblasts [69, 53, 55].

These individual studies on reference gene selection clearly show that even the reference genes that have been used for normalization for long time are subject to differential regulation under specific treatments or between different cell lines or tissues. This finding discredits important gene expression studies conducted so far that showed expression changes in different conditions or different cells after normalizing the expression values with endogenous control genes that are themselves prone to regulation.

High throughput analysis encompassing different tissues and cell lines is necessary to determine a pool of reference genes that can be used for normalization when analyzing gene expression among or within different tissues or cell lines [33].

The use of two or more housekeeping genes for normalization has been suggested to aid in reducing the effect of small differential expressions of housekeeping genes and making the decision of which reference gene to use easier [21, 34].

3.2 Methods and Materials

3.2.1 Materials

In this study, we used the curated gene expression data in our database. The curation details of the expression data in our database is described in Chapter 2.

In order to analyze the gene expression behavior of genes that may be used as reference genes, we have selected a set of experimentally determined housekeeping genes by Eisenberg et al. (2003) that contains 566 genes.

3.2.2 Methods

Expression data for a certain gene is analyzed with respect to its variance across experiments. Therefore, we first compute, for each gene, an average change of its rank in a GEO data set. The idea is that housekeeping genes should exhibit relatively constant expression levels and their average rank change should be lower than that of non-housekeeping

genes.

Let G=$\{G_1,...,G_m\}$ be the set of m genes and S = $\{e_1,...,e_n\}$ be a GEO data set of n microarray experiments. For gene G_i in experiment e_j, a single rank value, r(G_i,e_j), is computed. The details for this rank computation of the spots in the database are described in Chapter 2.

Given a set **S** with n experiments, we have at most n rank values for a gene. The average amount of change in the rank of gene G_i, in set **S** is then computed by the Coefficient of Variation, C_v(S,G_i), of r(G_i,e_j) in **S**, as given in Equation 3.3.

$$\mu\left(S_i, G_i\right) = \frac{\sum_{j=1}^{N} r\left(G_i, e_j\right)}{N} \quad \forall e_j \in S \tag{3.1}$$

$$\sigma\left(S, G_i\right) = \sqrt{\frac{\sum_{j=1}^{N}\left(r\left(G_i, e_j\right) - \mu\left(S, G_i\right)\right)^2}{N}} \quad \forall e_j \in S \tag{3.2}$$

$$C_v\left(S, G_i\right) = \frac{\sigma\left(S, G_i\right)}{\mu\left(S, G_i\right)} \tag{3.3}$$

In this analysis, we compute a C_v(S,G_i) for gene G_i, in each GEO data set **S**. However, not every gene is observed in all GEO data sets. Therefore most genes have C_v(S,G_i) computed only for a subset of the GEO data sets.

The next step is to analyze the C_v(S,G_i) values, i.e., the average amount of change in the rank of a gene in set **S**, by thresholding on this change value. For gene G_i and threshold, t, we compute the ratio, Ratio$_t$(G_i), of experiment sets in which a gene has C_v(S,G_i) values lower than t, as shown in Equation 3.6. This ratio is computed by considering only the GEO data sets in which the gene is observed. In this analysis, housekeeping genes are expected to have higher Ratio$_t$(G_i) values, since they should generally exhibit low C_v(S,G_i) values in many GEO data sets. For different values of t, i.e., 0.5, 0.1, 0.05, and 0.01, we compared housekeeping genes to five different randomly selected sets of non-housekeeping genes. The sets of non-housekeeping genes are selected such that they have the same mean rank distribution as that of the housekeeping gene set.

$$R_1 = \{S | C_v\left(S, G_i\right) \leq t\} \tag{3.4}$$

$$R_2 = \{S | C_v\left(S, G_i\right) \geq 0\} \tag{3.5}$$

$$Ratio_t\left(G_i\right) = \frac{|R_1|}{|R_2|} \tag{3.6}$$

The | | operator in the Equation 3.6 indicates the cardinal function.

Let f_{PO}(r) be the function that gives the number of genes with Ratio$_t$ greater than a given r and occur in at least in PO% of the datasets, as shown in Equation 3.7. We plotted and compared the graph of f_{PO}(r) for housekeeping genes and randomly selected non-housekeeping genes. The plotted graphs are presented in the Section 3.3.

$$f_{PO}\left(r\right) = \{G_i | Ratio_t\left(G_i\right) \geq r\} \tag{3.7}$$

The Ratio$_t$(G$_i$) value, by itself, does not provide a measure that can be used to identify statistically significant reference genes. For example, a gene that is observed in a single GEO data set and have a C$_v$(S,G$_i$) value smaller than the threshold will achieve a perfect ratio. To cope with such cases, we performed another analysis by filtering the genes based on the percentage of the number of sets that each gene is observed, which we call percentage of occurrence, PO. For different values of PO, i.e.: 75%, 50%, 25% and 3% with different C$_v$(S,G$_i$) thresholds, i.e.: 0.5, 0.1, 0.05 and 0.01, sorted Ratio$_t$(G$_i$) graphs comparing housekeeping and five different randomly selected sets of non-housekeeping genes are analyzed in Section 3.3.

One of our major goals in this study is to define a reference gene set that can be used for global normalization of microarray experiments. The C$_v$(S,G$_i$) measure can be utilized to build a classifier that can be used to predict novel reference genes. We built a classifier with a fixed $PO = 75\%$ and a fixed Ratio$_t = 0.90$. We analyzed the accuracy of our classifier in predicting the experimental housekeeping genes. The stringency of our classifier is adjusted by varying the C$_v$(S,G$_i$) threshold t. We plot the receiver operating characteristic of our classifier at different stringency levels. At a given threshold, t, our classifier identifies a set of candidate reference genes. The experimental housekeeping genes are regarded as true positives (TP) and the non-housekeeping genes are regarded as false positives (FP) in this candidate reference gene set. For different C$_v$ values, we plot the graph of sensitivity and specificity measures, as given in Equation 3.8 and in Equation 3.9.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.9}$$

A good classifier is supposed to capture most of the known housekeeping genes while providing a relatively small number of false positives. However, the false positives are candidate reference genes that can be used for normalization purposes as well and they are investigated further in Section 3.3.

3.3 Results

Recent studies have focused on housekeeping genes on a single or a few number of microrarray experiment sets. One of our main goals in this study is to check the consistency of these housekeeping genes across large scale experiment sets. For this purpose, we ranked each gene within an experiment and defined measures over the rank values. We then described methodologies to compare these measures of housekeeping genes with those of randomly selected non-housekeeping genes. We paid attention for the randomly selected non-housekeeping genes to have the same mean rank distribution as of those housekeeping genes. Details for experiment sets, housekeeping gene set and further implementation details are presented in Section 3.2.

In order to compare housekeeping genes with randomly selected non-housekeeping genes, we plotted the graph of their sorted $\text{Ratio}_t(\text{G}_i)$ (which is the ratio of number of sets a gene G_i has C_v below a threshold value t to the number of sets in which gene G_i occurs) using thresholding on C_v values and filtering on the percentage of occurrence values (PO). For C_v thresholds, t=0.5, 0.1, 0.05 and 0.01 and percentage of occurrences, PO=75%, 50%, 25% and 5% we analyzed genes ordered by both linear ranking and percentile ranking. When t=0.5, analyses performed using both the percentile- and linearly-ranked GEO data sets have shown that nearly all genes (housekeeping or not) exhibited high Ratio_t values for all PO values. At t=0.1 and at t=0.05 as well, housekeeping genes had significantly higher Ratio_t values than those of random gene sets for all PO values. Yet, percentile ranked genes resulted in higher Ratio_t values than those of linearly-ranked genes. Finally, for $t = 0.01$, Ratio_t difference between the housekeeping and non-housekeeping genes was relatively less in either the percentile- or linearly-ranked GEO data sets; but again housekeeping genes behaved slightly better than the random sets in all data tables for all PO values. For the sake of simplicity, we present the graphs

of Ratio$_t$ for PO=50% for C$_v$ thresholds a-) 0.5, b-) 0.1, c-) 0.05 and d-) 0.01 in Figure 3.1 while in Figure 3.2, graphs for C$_v$ threshold 0.05 of different *PO* thresholds, a-) 75%, b-) 50%, c-) 25% and d-) 5% are shown.

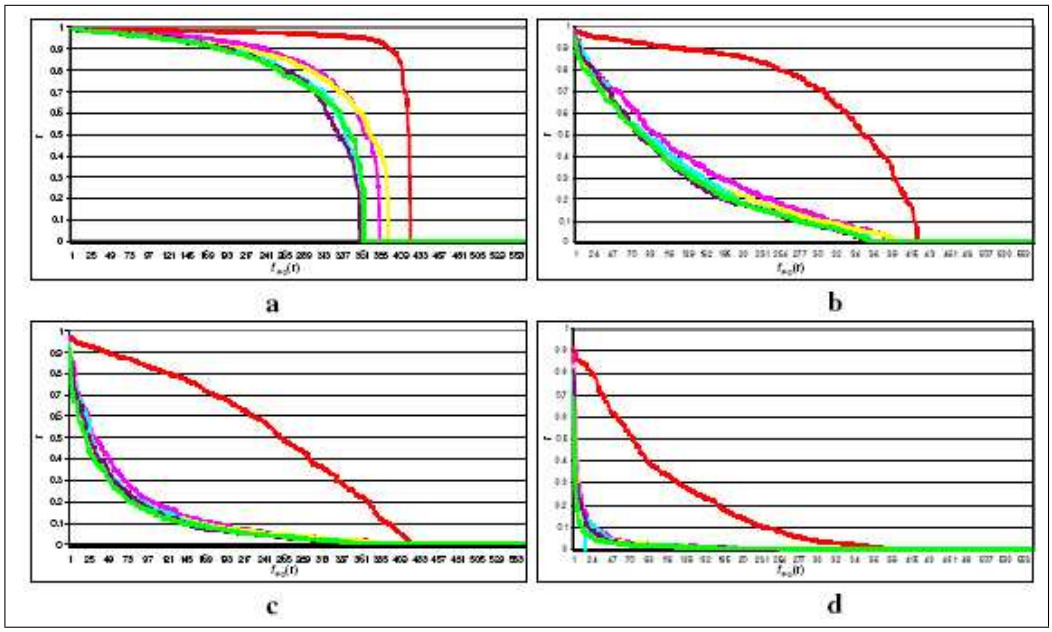

Figure 3.1: *Graph of ratio of the number of sets in which gene has coefficient of variation less than a-) 0.5, b-) 0.1, c-) 0.05, d-) 0.01 to the number of sets in which the gene is observed. Gene is observed at least 50% of the total sets. The x axis indicates the number of genes having a ratio value greater than ratio value at the corresponding y axis. Curve with red color presents housekeeping genes while curves with other colors shows 5 random sets of genes excluding the housekeeping genes. Random sets of genes have the same mean rank distribution as of those housekeeping genes.*

As shown in Table B.1 and Table B.3, for different C$_v$ values and *PO* values, we applied Kolmogorov-Simirnov tests on graph data and statistically proved that housekeeping genes behave significantly different than randomly selected non-housekeeping gene sets. Bonferroni adjusted Kolmogorov-Simirnov tests were used to show randomly selected non-housekeeping gene sets behave similarly as shown in Table B.2 and Table B.4.

From the results, the claim that housekeeping gene expression is less variable across

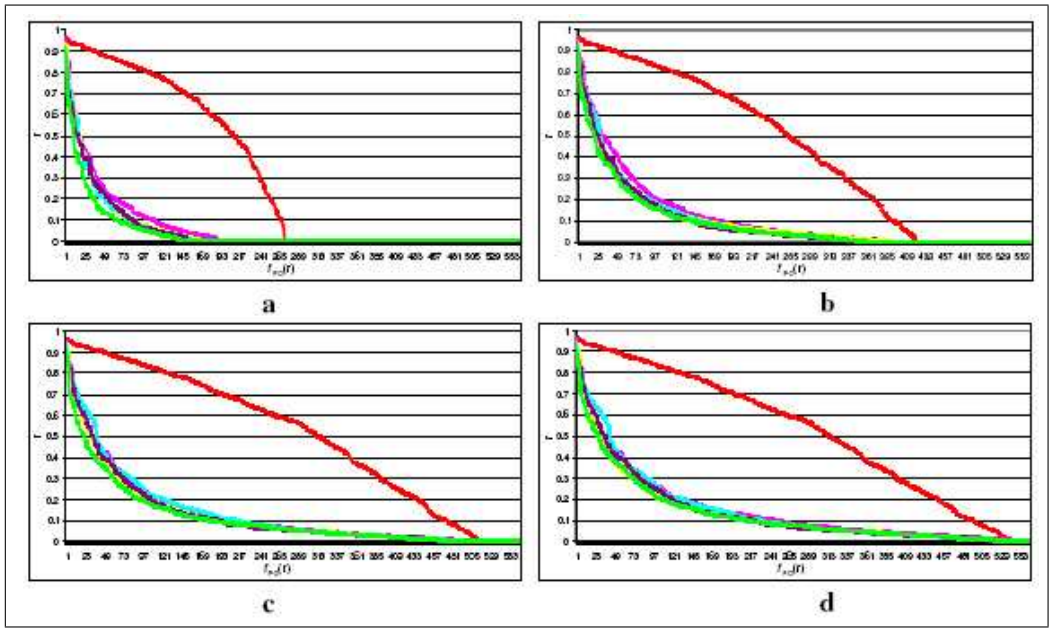

Figure 3.2: *Graph of ratio of the number of sets in which gene has coefficient of variation less than 0.05 to the number of sets in which the gene is observed. Gene is observed at least a-) 75%, b-) 50%, c-) 25% and d-) 5% of the total sets. The x axis indicated the number of genes having a ratio value greater than ratio value at the corresponding y axis. Curve with red color presents housekeeping genes while curves with other colors shows 5 random sets of genes excluding the housekeeping genes. Random sets of genes have the same mean rank distribution as of those housekeeping genes.*

different experiment sets when compared with randomly selected gene sets is also consistent in large scale analysis.

Another primary goal in this study is to define a set of genes that can be used for global normalization of microarray experiments. In order to define such a reference gene set we defined some measures and methodologies. We applied thresholding on the C_v measures and filtering on *PO* values. We defined a reference gene as a gene that has a C_v value less than the threshold in 90% of its occurrences and has a *PO* value greater than 75%. In this methodology the selection of C_v threshold value is important. In order to find a good C_v threshold value, we analyzed the sensitivity and specificity of this reference selection for different C_v threshold values. For different C_v values, we plot the graph of sensitivity and specificity measures (Receiver Operating Curve, ROC). Further details for these measures and methodology can be found in the Section 3.2. The ROC

graph is given in Figure 3.3.

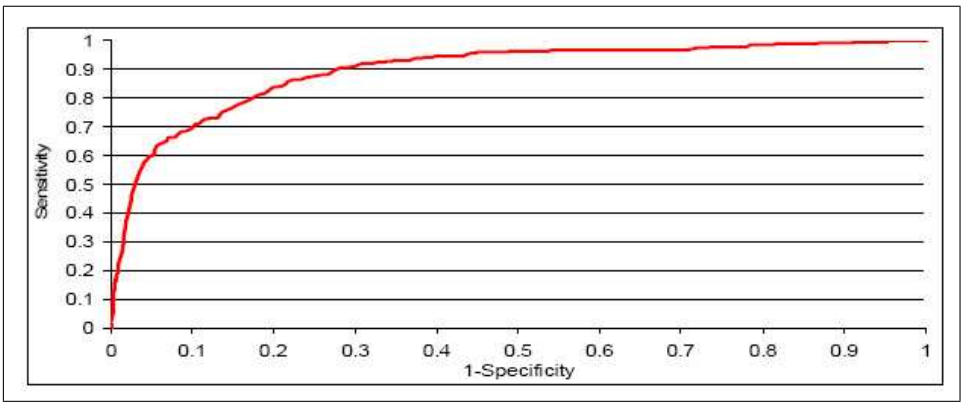

Figure 3.3: *ROC Curve for coefficient of variations (C_v) ranging from 0.01 to 10 for genes observed at least 75% of the total sets and having ratio value ($Ratio_t$) greater than 0.9. $Ratio_t$ is the number of sets in which the gene has C_v less than the threshold over the total number of occurrences of that gene. The x axis indicated (1-specificity) for the corresponding sensitivity in the y axis. Each point in the graph corresponds to a C_v threshold.*

Results of our computational experiments can be reached through the web site (http://www.i-cancer.org/HKgenes) for the use of biological experimentalists. The global reference gene set we have found, which are computed by considering all available microarray experiments in the database, is given in Appendix C and is also accessible from the web site.

In addition, the initial results demonstrated that rather than focusing on a general reference gene set, cell specific reference gene sets gave better results. Finding a reference gene for the analysis of an experiment is a common problem of the experimentalists. Different cells under different conditions may require using different reference gene sets. Keeping these problems at the center of our research, we aimed to provide reference gene sets for different cell types. Therefore, cell origin specific twenty reference gene sets, according to the Mesh Ontology nodes given in Section 2, were generated based on the Mesh Headings of the published microarray experimental data. The list of reference set according to Mesh Headings are also accessible from the web site.

3.3.1 Experimental Results

Expression levels of 17 housekeeping genes (AARS, ACTB, CFL1, EEF2, GAPDH, GSTO1, H2AFZ, HBXIP, RPL30, RPL41, RPL7, RPN2, RPS10, RPS17, RPS3A, SOD1, TPT1) were biologically experimented and measured in 16 different cell lines consisting of 8 Hepatocellular Carcinoma cell lines (HepG2, Focus, Mahlavu, Hep3B, Hep3B-TR, Huh7, SkHep1, PLC), 5 Breast Carcinoma cell lines (MDA-MB453, HCC1937, BT20, T470, CAMA I) and 3 Colon Carcinoma cell lines (HCT116, HT29, SW620).

NormFinder and geNorm are softwares used in defining gene expression stability. geNorm is a pair wise comparison-based model that calculates a gene expression stability measure M for each gene based on the average pair wise variation between all tested genes. The genes are ranked according to their expression stability through stepwise exclusion of the gene with the highest M value indicating least stable expression. Hence, a low stability M value indicates a gene with stable expression. NormFinder is a model-based approach that estimates the variation between sample subgroups, such as hepatocellular carcinogenesis (HCC), Breast and Colon Carcinoma cell lines, as well as the overall expression variation of the tested genes.

We used geNorm and Normfinder softwares to rank stability values of these 17 candidate reference genes. The results are given in Table D.1 The ranking of the housekeeping genes is almost similar with both softwares. RPS10, RPL41 and RPL30, RPS3A were the most stable genes among all 16 cell lines according to geNorm and NormFinder respectively. Two commonly used genes for normalization, ACTB and GAPDH, were less stable than the ribosomal genes and ranked lower in the stability rank list of both softwares. Stability within Hepatocellular Carcinoma, Breast Carcinoma and Colon Carcinoma cell lines were also calculated separately. The ribosomal genes RPL30, RPL41, RPL7, RPS10 and RPS3A were stable within and among different cell lines.

Comparison of the actual expression values by analysis of variance (ANOVA) identified RPL30 as the most stable gene with lowest variance within and between the HCC, Breast and Colon Carcinoma cell line groups, followed by RPL41, RPS10 and CFL1. The variance within each of these three groups was low for RPS3A, RPS17, RPL7, ACTB, H2AFZ, and HBXIP reference gene expression but the variance between the three groups were high, indicating that these reference genes are suitable for normalization when cell lines from one of the indicated group are being compared, but not when comparing cell

lines from different sources like HCC, Breast and Colon Carcinoma. GSTO1, TPT1, RPN2, SOD1 showed the highest variability in geNorm, NormFinder and ANOVA analysis, and hence were not accurate reference genes for normalization.

While the commonly used reference gene GAPDH was one of the stable genes in hepatocellular carcinogenesis (HCC) cell lines, it was among the least stable genes in Breast and Colon Carcinoma cell lines. It had a high variance in terms of stability value and actual expression value between the three types of carcinoma cell lines investigated. This suggested that GAPDH can be used as a reference gene when dealing with HCC cell lines, but not Breast and Colon Carcinoma cell lines. ACTB was more stable than GAPDH in Breast and Colon Carcinoma cell lines and had a stability value similar to GAPDH in HCC cell lines.

Recent studies also showed that ACTB, GAPDH and TBP were differentially expressed in eight pathological stages of hepatitis C virus induced, and thus are inappropriate for normalization [64, 69]. Further, GAPDH and ACTB are regulated during HCC [25]. Similarly, our results showed that when comparing expression levels in HCC, Breast Carcinoma or Colon Carcinoma, other reference genes, especially RPL30, RPL41, RPL7, RPS10 and RPS3A should be used for normalization rather than ACTB and GAPDH in order to reach more reliable expression data.

3.4 Discussion

Recent advances in large scale transcriptome analysis result in data accumulation at a large scale. Transcriptome data must be comparable although they are generated from different experiments performed under various biological conditions. We downloaded all *Homo sapiens* microarray experiments from NCBI-GEO and generate a curated large gene expression database. Previous studies have provided lists gene lists involved in cellular maintenance functions; thus these genes are called housekeeping genes and are generally assumed to have their expression levels unaffected by experimental conditions. We characterized expression patterns of the published set of housekeeping genes across the large number of microarray experiments in this database. We have selected the sets of housekeeping genes by Eisenberg et al. (2003) and Hsiao et al. (2001) in our analysis.

We described methodologies to compare these measures of housekeeping genes with those of randomly selected non-housekeeping genes using the gene expression data in our

curated database. Accordingly, the present study supports the claim that housekeeping gene expression is less variable across different experiment sets when compared with randomly selected gene sets. We further compared housekeeping genes published by Hsiao et al. (2001) with randomly selected non-housekeeping genes independently. The results showed similar behaviour compared with the results with housekeeping list published by Eisenberg et al. (2003).

Since cells perform different activities in the body, gene pairs involved in cellular functions varies among tissues. Therefore, except from the ribosomal protein genes that are used to supply energy to the cell, it is not possible to find a general set of genes that are involved and expressed in all cellular functions in all cell types. In addition, along with the recent avilable literature our initial results demonstrated that rather than focusing on a general reference gene set, cell specific reference gene sets gave better results. Therefore, cell origin specific twenty reference gene sets were generated based on the Mesh Headings of the published microarray experimental data.

Further, we performed biological experiments with some of the reference genes in our results which are not published in the available literature. These experiments confirmed 14 novel genes that can be as reference genes within various cell types.

Future studies will include determination of different subsets of genes that could be used as house-keeping gene sets for a particular biological condition (e.g., cancer); these gene sets can be useful in microarray data normalization as well as real-time RT-PCR confirmation studies used for amplifying defined pieces of RNA molecules.

CHAPTER 4

BI-K-BI CLUSTERING

Due to the increase in gene expression profile data sets in recent years, application of data mining techniques on these data became a matter of interest. Various methods have been proposed for mining gene expression profiles. However most of these methods have several limitations.

The proposed methods generally work on a single gene expression data set and cannot handle large number of gene expression data sets in public databases in reasonable amount of time and space. As a solution we propose a two level biclustering approach that works at the data set and experiment (i.e., condition) levels and discovers similar behaving gene pairs in multiple data sets. Our approach does not produce a subset of genes in a subset of conditions; rather, we report *pairs of genes* that behave similarly in a subset of conditions. This property is a setback compared to existing methods; however, it allows for mining gene expression patterns on a larger scale on a desktop computer with limited computational resources.

In this study, we combined the ideas from biclustering and association pattern discovery approaches. Our framework mainly consisted of a preprocessing phase followed by three phases. In the preprocessing phase, we scaled gene expression values of all data sets into a comparable platform and computed an all-to-all similarity measure over the whole set for finding gene pairs with similar expression profiles. In the first phase of the bi-k-bi clustering, we applied biclustering on gene pairs in order to eliminate unrelated gene pairs and data sets. In the second phase, we assigned labels to each gene in each sample of the working data set to indicate expression levels (high-expressed or low-expressed). Finally, we generated rules of gene pairs associated with samples by applying biclustering on the working set. Our method outputted clusters of labeled gene pairs with their

associated microarray samples. We aimed to help biologists to discover significant relationships among gene pairs and work easily on these relationships by concentrating on their associated microarray samples.

We applied our framework on all available NCBI GEO Homo sapiens data sets and more specifically five different functionally concise groups of NCBI GEO data sets independently (i.e.: Breast Cancer, Normal Human Tissue, Obesity, Liver and Colon). We searched for the existence of resulting gene pairs in protein-protein interaction databases to recover those pairs that act in concert at the transcriptional as well as post-transcriptional levels.

The remainder of this chapter is organized as follows. In the first section, we give brief information about the related work done in this area. The detailed information about methods and materials are given in the second section. In the third section, we discuss the test results of our bi-k-bi clustering framework. Discussions of the framework are provided in the last section.

4.1 Related Work

In the early stage of microarray technology few experiments and spreadsheets were enough in analysis. As microarray samples get larger and larger, spreadsheets will become less and less of an adequate tool for doing analysis and data mining techniques should find more and more use in analyzing expression data in the samples.

During this early decade, both supervised and unsupervised data mining methods are applied to gene expression data. Supervised methods use predefined sample groups and try to assign any new sample to a proper group [28]. Unsupervised methods concentrate on the idea that genes with similar expression profiles might share common mechanisms and functions, thus can be grouped [28]. Support Vector Machines and Neural Networks are examples studies of supervised techniques [14, 61]. Principal component analysis, singular value decomposition, k-means clustering, hierarchical clustering and self organizing maps are sample studies of unsupervised techniques [1, 5, 58, 22, 57]. Both supervised and unsupervised techniques discover co-regulated genes over the full set. These methods also allow a gene or a condition to occur in more than one cluster/pattern.

In recent years, biclustering and association pattern discovery (APD) (a.k.a association rule mining (ARM)) methods have been proposed to discover genes with similar

expression profiles in a subset of conditions (samples) [13, 19, 35]. These methods also allow a gene or a condition to occur in more than one cluster/pattern.

Cheng and Church (2000) introduced the concept of biclustering for gene expression data and proposed a greedy approach based on a uniformity criterion. Their method used randomly generated values to replace missing values in the data set. This approach is likely to reduce the quality of the discovered biclusters. To address this problem, a probabilistic algorithm (FLOC) that handles missing values and discovers higher quality, overlapping biclusters has been proposed [65].

Ben-Dor et al. (2002) introduced the order preserving submatrix (OPSM) model for biclustering that focuses on the coherence of relative order of conditions rather than the coherence of expression values. The OPSM model is improved by allowing some flexibility among conditions in an order equivalent group [44]. This new model is called as, OP-Cluster, and is able to tolerate the effect of noise that reduces the efficiency of the stricter OPSM model [44].

Pattern based clustering problem is related to the biclustering problem. The p-Clustering algorithm is able to discover genes that exhibit similar expression patterns in a subset of conditions [62]. Pei et al. (2003) improved the p-Clustering algorithm by proposing a more efficient and scalable method, MaPle, to find maximal pattern-based clusters. All of the biclustering methods mentioned above discover clusters of genes that behave similarly in a subset of experimental conditions. However, the computational resources required by existing biclustering methods do not allow for analysis of gene expression data on a large scale.

Similar to biclustering methods association pattern discovery (APD) methods can be used to discover genes with similar expression profiles in a subset of conditions. APD methods were first used to discover associations among subsets of items from large transaction databases [4]. APD methods detect sets of elements that frequently co-occur in a database and establish relationships between them of the form of $X \rightarrow Y$, meaning that, when X occurs it is likely that Y also occurs [4].

Various methods have been proposed on associations and relationships among subsets of genes (e.g.: X= {Condition_1 (sample_1), Condition_2 (sample_2) }, Y= {gene A $\uparrow$, gene B $\downarrow$, gene C $\uparrow$}, which means, in Condition_1 and Condition_2 when gene A is up regulated, gene B is down regulated and gene C is also up regulated) [11, 20, 26, 35, 59].

Most of the studies with APD methods work on homogeneously curated one or two

data sets experimented for a specific study. APD methods that work on binary data use thresholding on the expression values of genes [26]. It is clear that a crude discretization such as using thresholding lead to certain loss of information. As a solution, a quantitative association rule mining approach has been proposed [26]. However, microarray experiments are generally not sufficiently robust and noise-resistant. It is not easy to decide whether or not an association of a small quantity is noise. Therefore, quantitative associations among genes do not generally give valuable information to a biologist. Another disadvantage of existing APD methods is that these methods try to find out rules over the whole set of genes. However, focusing on association rules among genes having similar expression profiles reduces the working data set by eliminating uncorrelated data patterns and is therefore expected to give more accurate results in a more efficient manner.

4.2 Methods and Materials

We follow a three step methodology in our study. In the first step we apply normalization in order to scale all microarray samples in our database into a comparable platform. In the second step, we define a function to find gene pairs with similar expression profiles from the first step. Finally, we apply the bi-k-bi clustering algorithm to the gene pairs that have similar expression profiles and construct rules consisting of gene pairs and associated samples. The overview of the proposed methodology is given in Figure 4.1 and explained in detail below.

4.2.1 Materials

In this study, we used the curated gene expression data in our database. The curation details of the expression data in our database are described in Chapter 2.

4.2.2 Methods

4.2.3 Finding Gene Pairs Having Similar Expression Profiles

In order to find gene pairs having similar expression profiles we follow a three phase procedure. In the first phase, we define a similarity measure for finding similar rank behaving gene pairs and calculate this measure for each gene pair and each data set. In

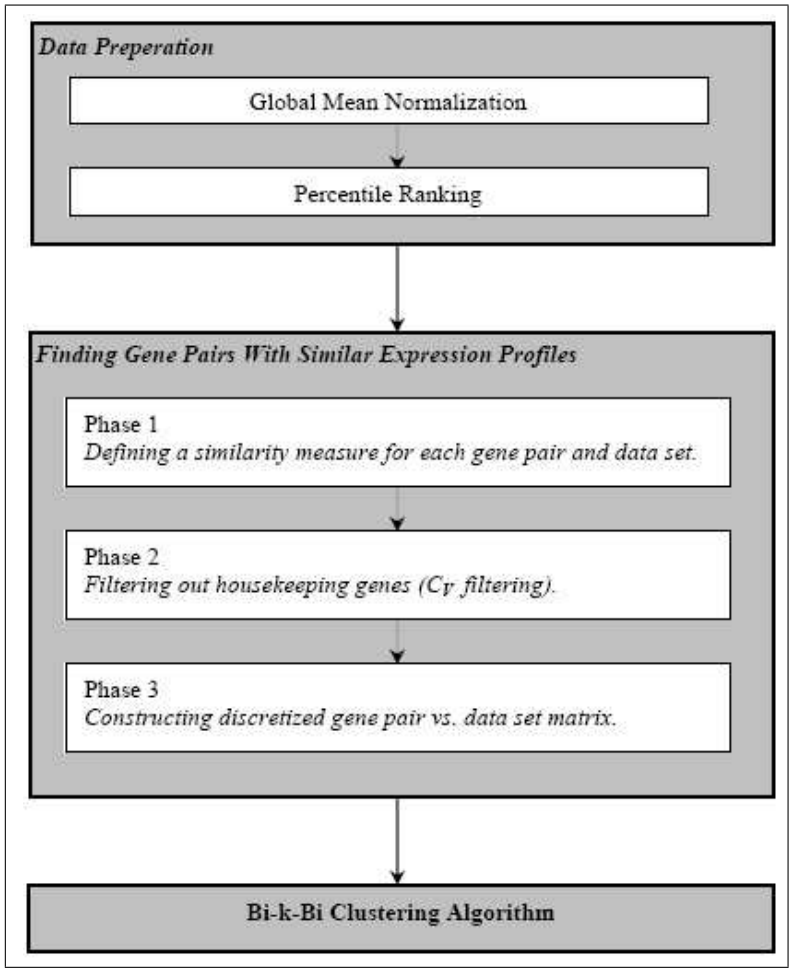

Figure 4.1: The overview of the three step methodology for mining gene expression data from multiple data sets.

the next phase, we apply a filter to eliminate gene pairs that consists of housekeeping genes. Finally, in the last phase, we construct a matrix from the values calculated in the first phase and apply thresholding to discretize this matrix.

Phase 1

Three commonly used similarity measures for finding similar behaving gene pairs are the Euclidean Distance, Pearson's Correlation Coefficient, and Spearman Rank Correlation Coefficient (ρ) [7]. Euclidean Distance and Pearson's Correlation Coefficient methods are sensitive to magnitude and shape [37]. Whereas, Pearson's Correlation Coefficient assumes approximate Gaussian distribution of the points and therefore it is not robust for non-Gaussian distributions [32, 37]. On the other hand, Spearman Rank Correlation

Coefficient does not require Gaussian distribution and it is more robust against outliers. Therefore, we decided to use Spearman Rank Correlation Coefficient in our analysis.

Spearman Rank Correlation Coefficient has the drawback of data loss due to consequence of ranking [32]. Since we look for gene pairs having similar patterns over scaled expression values, the drawback of ranking has a negligible effect over our analysis.

Spearman Rank Correlation Coefficient is also sensitive to missing values. One of the approaches for missing values is to ignore the missing value pairs during correlation computation. Since not every gene is observed in all microarray samples, missing expression value pairs are ignored during computation. In order to cope with the negative effect of the "*ignore*" approach; we used Weighted Spearman Rank Correlation Coefficient, ρ_w, in our analysis.

Let S=$\{S_1,S_2,...,S_m\}$ be m data sets in the database where S_i =$\{e_{i1},e_{i2},...,e_{in}\}$ is a set having n microarray samples where $n \geq 3$ and GS=$\{G_1,\ G_2,...,G_k\}$ be the set of k distinct genes in the database. Also let P_S be the percentage of non-ignored value pairs during Spearman Rank Correlation Coefficient calculation. We define a log based non linear weight function, $f_w(P_S)$, as given in Equation (4.1). Since, $P_S \in$ [0,100] then, by using Equation (4.1), $f_w(P_S) \in$ [1,2].

$$f_w(P_S) = \begin{cases} \log(P_S) & \text{for} \quad P_S > 10 \\ \log(10) & \text{for} \quad P_S \leq 10 \end{cases} \tag{4.1}$$

Let r(G_i,e_k) be the average percentile ranked expression value of gene G_i and r(G_j,e_k) be the average percentile ranked expression value of gene G_j in the microarray sample e_k of the data set S_a. We then define the Weighted Spearman Rank Correlation Coefficient, $\rho_w(G_{i,j},S_a)$, as given in Equation (4.3).

$$R(G_{i,j}, S_a) = \{\forall e_k \in S_a | (r(e_k, G_i), r(e_k, G_j))\} \tag{4.2}$$

$$\rho_w(G_{i,j}, S_a) = \rho(R(G_{i,j}, S_a)) \quad \times \quad f_w(P_{Sa}) \tag{4.3}$$

Phase 2

Gene pairs that are not modulated in microarray samples are expected to (and does) have high ρ_w values. In this study we mainly focus on genes whose expression profiles are

similarly affected among samples. Therefore, genes whose expressions are not modulated in any of the experimental conditions, (i.e., housekeeping genes) does not have much impact on the association rules of genes which we aim to find.

Previous studies by Hsiao et al. (2001) and Eisenberg et al. (2003) have provided gene lists involved in cellular maintenance functions; thus these genes are called housekeeping genes that are generally assumed to have expression levels unaffected by experimental conditions. However, recent studies indicated that several widely used housekeeping genes might have altered expression under different experimental conditions [63, 60]. Therefore, in order to eliminate genes that are not effected in microarray samples, we use the Coefficient of Variation, C_v, as a filter.

Let $C_v(G_i,S_a)$ be the Coefficient of Variation of gene G_i and $C_v(G_j,S_a)$ be the Coefficient of Variation of gene G_j calculated using the percentile ranked expression values of microarray samples of the data set S_a, as defined in Equation (4.4). Also let $\rho_w(G_{i,j},S_a)$ be the Weighted Spearman Rank Correlation Coefficient for $G_{i,j}$ among the expression values of G_i and G_j in the data set S_a, calculated by using Equation (4.3). We then apply a filter on $\rho_w(G_{i,j},S_a)$ using the C_v values as defined in Equation (4.5).

$$C_v(G_i, S_a) = C_v(\{r(G_i, e_k)\}) \quad \forall e_k \in S_a \tag{4.4}$$

$$\rho_w(G_{i,j}, S_a) = \begin{cases} 0 & \text{if } C_v(G_i, S_a) \leq t_{Cv} \text{ and } C_v(G_j, S_a) \leq t_{Cv} \\ \rho_w(G_{i,j}, S_a) & \text{otherwise} \end{cases} \tag{4.5}$$

Phase 3

For each gene pair, $G_{i,j}$ =(G_i, G_j) where G_i,G_j $\in$ GS, and each data set S, we calculate Weighted Spearman Rank Correlation Coefficient using Equation(4.5).

Given a query gene G_i, this calculation forms out a $k \times m$ matrix, $\mathbf{A}$, with k rows and m columns. This matrix is defined by its set of rows, X=$\{G_{i1}, G_{i2}, G_{i3}, \ldots, G_{ik}\}$ for the given gene and its set of columns, Y=$\{S_1, S_2, \ldots, S_m\}$.

A cell in this matrix, a_{MN}, is a real value representing the Weighted Spearman Correlation Coefficient of M^{th} gene pair for the N^{th} working set as defined in Equation (4.6).

$$a_{MN} = \rho_w(G_{i,j}, S_N) \quad \text{where} \quad M = G_{i,j} \quad \text{and} \quad N = S_N \tag{4.6}$$

Since $\rho \in$ [-1,+1] and $f_w(P) \in$ [1,2] then $\rho_w \in$ [-2,+2]. Setting the threshold, $t_{\rho w}$=1.5, for ρ_w makes $\rho \leq 0.75$; which is a reasonable threshold for two random variables that show similar behavior.

Weighted Spearman Rank Correlation gives a scale on the strength of the similarity of two random variables. Therefore, in order to decide whether two genes behave similarly, we apply thresholding and discretize the matrix to mark gene pairs versus sets showing similar behavior. A cell, a_{MN}, in the matrix is then defined in Equation (4.7).

$$a_{MN} = \begin{cases} 1 & \rho_w(G_{i,j}, S_N) \geq t_{\rho w} \quad \text{where} \quad M = G_{i,j} \quad \text{and} \quad N = S_N \\ 0 & \rho_w(G_{i,j}, S_N) < t_{\rho w} \quad \text{where} \quad M = G_{i,j} \quad \text{and} \quad N = S_N \end{cases} \tag{4.7}$$

There exist approximately 20,200 distinct genes among the 371 NCBI-GEO data sets complied in our database. Therefore, the matrix we used in our analysis has more than 150,000,000,000 members. By the use of the symmetry property of Spearman Rank Correlation Coefficient, thresholding, filtering weight values during ρ_w calculation and using the fact that microarray samples generally do not include all genes, we had 12,000,000,000 ρ_w calculations to construct this matrix.

4.2.4 Bi-k-Bi Clustering Algorithm

The bi-k-bi clustering algorithm was applied in three steps (Figure 4.2 and Algorithm 1) which operates on the discretized matrix, **A**, defined in Section 4.2.3.

In the first step we perform a coarse analysis in order to reduce the working set by focusing on gene pairs with their associated data sets in which they have similar expression profiles. The second step is the assignment of labels for genes in microarray samples by k-means clustering over all available gene expression data in our database. Finally, a detailed analysis of labeled gene pairs associated with microarray samples, so called rules, is performed. As it can be deduced, the first and last step should be done in order, however, the second step, where we assign a label for each gene in each microarray sample, can be performed in any order.

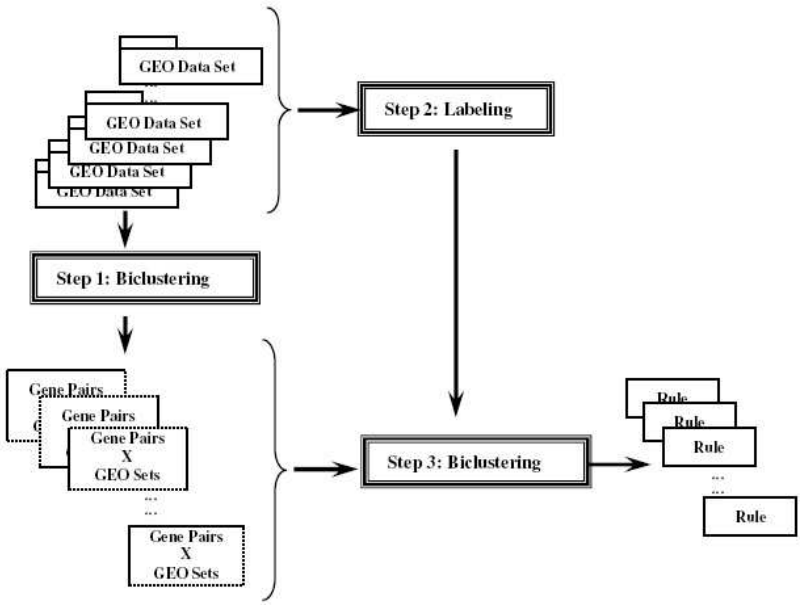

Figure 4.2: Workflow of the bi-k-bi clustering algorithm.

Algorithm 1 Bi-k-Bi Clustering Algorithm

1: Find clusters of similar behaving gene pairs versus sets using the **biclustering** algorithm, as described in Section 4.2.4.

2: Associate a label to each gene in each experiment using **k-means** clustering algorithm, as described in Section 4.2.4.

3: **loop**

4: For each cluster of similar behaving gene pairs

5: Construct (sample,labeled gene pairs) using the function, ClusterLabel(C), defined in Equation (4.11).

6: **end loop**

7: Find clusters of labeled gene pairs versus experiments, called as rules, using **biclustering** algorithm, as described in Section 4.2.4.

Finding Clusters of Gene Pairs That Have Similar Expression Profiles

After the discretization process of the matrix from the finding similar gene expression pair step (Figure 4.1), we first focus on identification of both groups of gene and set pairs in this matrix. It is important to keep in mind that we should find the groups of gene pairs and sets having 1's in the matrix which represent similarly behaved gene pairs after bi-k-bi clustering algorithm. Biclustering algorithms are used for this purpose and tools available in the literature [32, 19, 62, 70]. Initially in this study we used a freely available Biclustering Analysis Toolbox (BiCAT) [8]. However due to the large size of our matrixes we faced memory problems. For this reason, we designed and implemented a time and memory efficient biclustering algorithm.

Association Pattern Discovery (APD) methods have been applied on gene expression data in order to find out groups of co-regulated gene patterns [20, 18, 26, 28]. APD originates from market basket analysis and aim to find interesting relationships hidden in large data sets. Such relationships can be represented as frequent itemsets and association rules. APD methods are inherited from the area of frequent itemsets and association rule mining. However, studies done on this area have commonly been focused on finding gene patterns as association rules [20].

Maximum Frequent Itemset (MFI) in transactional databases is the problem of mining maximum itemsets from the transactional database. Thus, in a given set of items and transaction set, MFI algorithms find out maximum sets of items occur for a given support. For example, for support v, items that occur at least in v transactions are reported. It is important to note that MFI reports the maximum itemsets and does not report subsets. There have been many studies for mining frequent itemsets in the literature [2, 3, 27].

MAFIA is one of the MFI algorithms which performs best when mining long itemsets and it outperforms other algorithms on dense data [15]. The algorithm applies space pruning techniques and adaptive compression that makes optimal use of memory and running time. A free implementation with source code of MAFIA is publicly available (http://himalaya-tools.sourceforge.net /Mafia/doxygen-Mafia/index.html).

MAFIA, like most MFI algorithms, outputs the list of frequent itemsets for a given support. Applying a post processing on this output by adding the associated transactions of the itemsets, we can have biclusters of itemsets and transactions. (i.e., biclusters having number of transactions greater than the support, v). The post processing has at

most $O(n^2)$ time complexity, which does not have much effect on the running time of the original MAFIA.

According to this approach, we modified the MAFIA algorithm in order to output associated transactions with the itemsets. We use this modified algorithm as our biclustering algorithm. We then represent transactions as data sets and itemsets as gene pairs having similar expression profile using the discretized matrix defined in Section 4.2.3.

Let $G_{i,j} \in GS$ and $S_k \in S$ and v_1 be the minimum number of sets in the clustered results; applying our biclustering algorithm gives clusters as in Equation (4.8).

$$Cluster_c = [\{G_{i,j}\}, \{S_k\}] \text{ where sizeof}(S) \geq v_1 \tag{4.8}$$

Labeling Gene Pairs in Experiments

In order to use association rules among the biclustered gene pairs, gene pairs having similar expression profiles (i.e., biclusters constructed in the previous subsection) should be labeled. In this step, we assign a label for each gene pair with the corresponding experiment in which it occurs.

In order to label gene pairs in our database, a preprocessing step should be applied. In this preprocessing step, rank values of the genes in the experiments are discretized. High rank values are labeled with **High** ("*High-expressed*"), low rank values with **Low** ("*Low-expressed*").

It is clear that a crude discretization such as using thresholding on rank data lead to certain loss of information [26]. In order to alleviate this loss of information we decided to use clustering on the rank values for labeling. Since **k-means clustering** is a fast and efficient clustering algorithm, we apply k-means clustering on all rank values and assign a label for each gene in each experiment using these clusters. When the number of clusters (i.e., k=2) are known k-means clustering is also advantageous in both time and space complexity.

Let $S = \{e_1, e_2, ..., e_n\}$ be a set of n microarray experiments, $G_{i,j}$ be a gene pair in S and $e_k \in S$. Also let $l_{i,k}$ be the label assigned to gene G_i in e_k and $l_{j,k}$ be the label for gene G_j within the same experiment e_k (i.e., $l_{i,k}$, $l_{j,k} \in$ {High, Low}). We define the gene pair experiment labeling function, ExpLabel($G_{i,j}$,e_k), for the gene pair $G_{i,j}$ within the experiment e_k as in Equation (4.9).

$$ExpLabel(G_{i,j}, e_k) = l_{i,j,k} \quad \text{where} \quad l_{i,j,k} = [G_i(l_{i,k}), G_j(l_{j,k})] \tag{4.9}$$

For a given set, we define the set labeling function, SetLabel($G_{i,j}$,S), as in Equation (4.10).

$$SetLabel(G_{i,j}, S) = \{ExpLabel(G_{i,j}, e_k)\} \quad \forall e_k \in S \tag{4.10}$$

Let C be a bicluster, we define the cluster labeling function, ClusterLabel(C), as in Equation (4.11).

$$ClusterLabel(C) = \{SetLabel(G_{i,j}, S_m)\} \quad \forall S_m, \forall G_{i,j} \in C \tag{4.11}$$

Finally, we apply the function, ClusterLabel(C), in Equation (4.11) to all biclusters found in Equation (4.8) and construct the data to be used in the third step of our bi-k-bi clustering algorithm.

Extracting Rules

In the third step of our bi-k-bi clustering algorithm we aim to find association rules as clusters of labeled gene pairs versus experiments. Thus, we apply biclustering algorithm on the clusters of gene pairs with similar expression profiles obtained as a result of k-means clustering. We use the same biclustering algorithm described in Section 4.2.4 and find out sets of clusters, which we call them as rules, for a given support v_2 (i.e., number of experiments in the rule is greater than v_2).

Let e_k be an experiment and $G_{i,j}$ be a gene pair in or database. Further, let v_1 be the support for the minimum number of data sets, gene pair $G_{i,j}$ have similar expression profiles and v_2 be the support for the minimum number of experiments in a rule. This final step of the algorithm outputs sets of rules as of the form given in Equation (4.12).

$$Rule = [\{ExpLabel(G_{i,j}, e_k)\}, \{e_k\}] \text{ for support } v_1 \text{ and } v_2 \tag{4.12}$$

4.2.5 Illustrative Example

As an illustrative example, consider the randomly selected sample genes $Gene_1$, $Gene_2$, $Gene_3$ and NCBI GEO data sets GDS1, GDS2 and GDS3 from our database. The average percentile rank values of these sample genes are given in Table 4.1.

Table 4.1: Average percentile rank values of the genes for the illustrative example.

NCBI-GEO GDS	**NCBI-GEO GSM**	**Average Percentile Rank**		
		Gene$_1$	**Gene$_2$**	**Gene$_3$**
GDS1	GSM11	36	51	45
	GSM12	51	60	65
	GSM13	38	57	-
	GSM14	54	-	-
	GSM15	-	-	55
	GSM16	55	-	44
GDS2	GSM21	62	90	35
	GSM22	59	86	21
	GSM23	44	84	42
GDS3	GSM31	87	87	60
	GSM32	94	93	46

First, we find gene pairs with similar expression profiles as described in Section 4.2.3. For this purpose, we define thresholds t_{Cv}=0.2 and $t_{\rho w}$=1.5. We then compute similarity measures (P_S, C_v, ρ, ρ_w) for each gene pair. The computed values for this illustrative example are given in Table 4.2. By using these computed values and thresholds, we prepare the input data for the bi-k-bi clustering framework as given in Table 4.3.

We apply the bi-k-bi clustering framework on the discretized matrix generated from the gene pairs with similar expression profiles and data sets.

In the first phase of the framework, we apply biclustering and generate clusters as described in Section 4.2.4. The resulting cluster for this illustrative example with support v_1=2 is given in Table 4.4.

In the second phase of the framework, we apply k-means to assign **High**/**Low** labels to the genes and generate (sample, labeled gene pairs) for the clusters as described in Section 4.2.4. List of labels for the genes in the microarray samples and the cluster for this illustrative example are given in Table 4.5 and Table 4.6 respectively.

In the last phase of our framework, we apply biclustering and generate association rules as described in Section 4.2.4. The resulting rule for this illustrative example with

Table 4.2: Similarity measure values (P_S, C_v, ρ, ρ_w) of the gene pairs in the illustrative example where t_{Cv}=0.2

Gene pair	GDS 1	GDS 2	GDS 3
$Gene_1$ - $Gene_2$	P_S=50 $C_v(Gene_1)$=0.19 $C_v(Gene_2)$=0.08 ρ=1 ρ_w =1.69	P_S=100 $C_v(Gene_1)$=0.17 $C_v(Gene_2)$=0.03 ρ=1 ρ_w =2	P_S=100 $C_v(Gene_1)$=0.05 $C_v(Gene_2)$=0.04 ρ=1 ρ_w =2
$Gene_1$ - $Gene_3$	P_S=50 $C_v(Gene_1)$=0.19 $C_v(Gene_3)$=0.18 ρ=-1 ρ_w =-0.84	P_S=100 $C_v(Gene_1)$=0.17 $C_v(Gene_3)$=0.32 ρ=0.5 ρ_w =0	P_S=100 $C_v(Gene_1)$=0.05 $C_v(Gene_3)$=0.18 ρ=-1 ρ_w =-2
$Gene_2$ - $Gene_3$	P_S=33 $C_v(Gene_2)$=0.08 $C_v(Gene_3)$=0.18 ρ=1 ρ_w =1.51	P_S=100 $C_v(Gene_2)$=0.03 $C_v(Gene_3)$=0.32 ρ=-0.5 ρ_w =0	P_S=100 $C_v(Gene_2)$=0.04 $C_v(Gene_3)$=0.18 ρ=-1 ρ_w =-2

support v_2=3 is given in Table 4.7.

4.3 Results

The proposed bi-k-bi clustering framework was applied on all available NCBI GEO *Homo sapiens* data sets (i.e. 9090 microarray samples grouped into 372 data sets). Our findings indicated that majority of the gene-pairs belonged to categories with known housekeeping gene functions, such as ribosomal protein genes and metabolic pathway genes. Housekeeping genes generally are assumed to have expression levels unaffected by experimental conditions thus are expected to exhibit relatively stable expression at high levels. Coefficient of Variation parameter of the bi-k-bi clustering approach allowed us to determine the most stably expressed genes in the database [17]. By applying an experimentally defined C_v filter threshold on the C_v, i.e., 0.1, we filtered out the likely housekeeping genes

Table 4.3: NCBI-GEO Data sets vs. gene pairs in the illustrative example where $t_{\rho w} \geq 1.5$.

NCBI-GEO GDS	**Gene pairs**
GDS1	[$Gene_1$,$Gene_2$],[$Gene_2$,$Gene_3$]
GDS2	[$Gene_1$,$Gene_2$]
GDS3	[$Gene_1$,$Gene_2$],[$Gene_1$,$Gene_3$], [$Gene_2$,$Gene_3$]

Table 4.4: Clusters for the illustrative example with support v_1=2.

Cluster Name	**Cluster Members**
c_1 (2x2)	{[$Gene_1$,$Gene_2$],[$Gene_2$,$Gene_3$]} x {GDS1,GDS3}

thereby leaving gene-pairs potentially involved in cellular signaling processes as well as those function among tissues and pathological states in highly divergent manners.

However, the results after C_v filtering may still contain so called housekeeping genes that pass the C_v threshold since recent studies have indicated that several widely used housekeeping genes have altered expression under experimental conditions [60, 63]. One of the important aspects of the present study is its ability to associate a set of gene-pair rules with subsets of the available microarray data. Therefore, we applied our framework on five different groups of data sets: *i*) Breast Cancer, *ii*) Normal Human Tissue, *iii*) Obesity, *iv*) Liver and *v*) Colon and extracted rules. For each group we constructed a working set consisting of NCBI GEO microarray data sets [9]. We manually curated working sets by using the free text titles and descriptions supplied by experimenters in NCBI GEO data sets. Microarray sample and data set distribution of each working set is as follows: *i*) Breast Cancer: 188 microarray samples grouped into 10 data sets; *ii*) Normal Human: 120 microarray samples grouped into 5 data sets; *iii*) Obesity: 275 microarray samples grouped into 8 data sets; *iv*) Liver: 35 microarray samples grouped into 2 data sets; and *v*) Colon: 132 microarray samples grouped into 9 data sets. The corresponding NCBI GEO data sets with their titles for each working set were provided in Table 4.8.

Table 4.5: K-means clustering for the assignment of gene expression levels in the illustrative example: High and Low.

NCBI-GEO GDS	**NCBI-GEO**	**Assigned Labels**		
	GSM	**Gene$_1$**	**Gene$_2$**	**Gene$_3$**
GDS1	GSM11	LOW	HIGH	LOW
	GSM12	HIGH	HIGH	HIGH
	GSM13	LOW	HIGH	-
	GSM14	HIGH	-	-
	GSM15	-	-	HIGH
	GSM16	HIGH	-	LOW
GDS2	GSM21	HIGH	HIGH	LOW
	GSM22	HIGH	HIGH	LOW
	GSM23	LOW	HIGH	LOW
GDS3	GSM31	HIGH	HIGH	HIGH
	GSM32	HIGH	HIGH	LOW

We applied the bi-k-bi clustering framework on each working set independently. C_v filter threshold, t_{Cv}, was set as 0.1 to specifically focus on gene-pairs with variable expression. Results for each subset can be accessed online (http://www.i-cancer.org/∼levent /rules). A search engine that enables users to query genes within the result sets can also be accessed through the web site (http://www.i-cancer.org/∼levent/query).

The most observed gene-pair rule in the breast cancer subsets, FOXM1-TPX2, existed in 80% of the microarray samples, given in Table E.1, is analyzed. Microarray samples valid and not valid for rule are listed in Table E.2 and Table E.3 respectively. Further, the graph of rank values in the microarray samples of the genes in this rule is also given in Figure E.1. Interestingly, a few of the experiment sets were included or excluded as a whole for the rule determination. In particular, FOXM1-TPX1 rule was restricted to all samples of GDS817, GDS820, GDS901 while this rule was not observed in any of the samples of the experiment sets GDS901 and GDS1250. On the other hand, different samples of GDS2250 were used for determining the FOXM1-TPX2 rule; for example, normal breast tissue expression together with six non-basal-like and one basal-like tumor

Table 4.6: Expression level assigned gene pairs vs. NCBI-GEO Microarray Samples for the illustrative example.

Labeled gene pairs	**NCBI-GEO GSM**
$Gene_1$(HIGH),$Gene_2$(HIGH)	GDS1_GSM12, GDS3_GSM31, GDS3_GSM32
$Gene_1$(HIGH),$Gene_2$(LOW)	-
$Gene_1$(LOW),$Gene_2$(HIGH)	GDS1_GSM11, GDS1_GSM13,
$Gene_1$(LOW),$Gene_2$(LOW)	-
$Gene_2$(HIGH),$Gene_3$(HIGH)	GDS1_GSM12, GDS3_GSM31
$Gene_2$(HIGH),$Gene_3$(LOW)	GDS1_GSM16, GDS3_GSM32,
$Gene_2$(LOW),$Gene_3$(HIGH)	-
$Gene_2$(LOW),$Gene_3$(LOW)	-

Table 4.7: Rules for the illustrative example with support v_1=2 and v_2=3.

Rule Name	**Rule Members**
$Rule_1$ (2x3)	{[$Gene_1$(HIGH),$Gene_2$(HIGH)],[$Gene_2$(HIGH),$Gene_3$(HIGH)] } x { GDS1_GSM12, GDS3_GSM31, GDS3_GSM32 }

samples did not contribute to the rule [52]. These results might indicate that normal cells do not exhibit FOXM1-TPX2 rule unlike subsets of breast tumors, i.e., basal-like. Previous studies support our findings such that FOXM1 and TPX2 were up regulated in breast tumor studies [54].

For the tissue expression pair-rules, the two most commonly observed composite rules were [CHD1 (HIGH) - PSD4 (HIGH)] [CHD1 (HIGH) - ZNF384 (HIGH)] and [MBD6 (HIGH) - PSD4 (HIGH)] [POLDIP2 (HIGH) - SFXN4 (HIGH)]. The first rule was determined based on the tissue samples from GDS422 and GDS423. Similarly, [MBD6 (HIGH) - PSD4 (HIGH)] [POLDIP2 (HIGH) - SFXN4 (HIGH)] rule was supported only by the GDS423 and GDS424 experiment sets. These findings suggested that representative sequences of the same gene might be non-equivalent among different platforms influencing the degree of the association due to either presence/absence of alternative splicing events, multiple hits in the transcriptome, and/or dinucleotide content of the probesets. Since

[CHD1 (HIGH) - PSD4 (HIGH)] [CHD1 (HIGH) - ZNF384 (HIGH)] rule was commonly observed in all of the tissues studied (i.e., bone marrow, liver, heart, spleen, lung, kidney, skeletal muscle, thymus, brain, spinal cord, prostate, and pancreas), this triplet might represent an alternative reference gene set for non-diseased tissue normalization studies. Moreover, CDH1, PSD4, and ZNF384 or MBD6 and PSD4, or POLDIP2 and SFXN4 genes were not shown to be co-expressed previously in the literature thus represent novel links in cellular signaling pathways.

The most commonly observed pair-rule in comparing the adipocyte expression profiling was [CRIP2 (HIGH) - RGS5 (HIGH)]; however this rule was not able to separate lean versus obese type adipocytes. Furthermore, [CRIP2 (HIGH) - RGS5 (HIGH)] rule did not hold in none of the preadipocyte/stromal vascular cells (GDS1480) or lean/obese skeletal muscle tissue (GDS268) or some of the lean/obese adipocytes (i.e., some of GDS1493, all of GDS1496, all of GDS1497, some of GDS1498). The same issue observed for the tissue datasets explained in the previous paragraph was present for the lean/obese samples (20 non-obese. BMI 25+/-3 kg/m^2, and 19 obese. BMI 55+/-8 kg/m^2, non-diabetic Pima Indians; [40]). GDS1493, GDS1495, GDS1496, GDS1497, and GDS1498 belonged to five different platforms (HG-U95A-E) and although were used for the same set of adipocytes. Nevertheless, each platform might contain different probesets for the same gene thus might not always support the [CRIP2 (HIGH) - RGS5 (HIGH)] or other rules.

In liver datasets, [CD38 (LOW) - PENK (LOW)] [C2ORF27 (LOW) - HELZ (HIGH)] [GYS1 (HIGH) - SNX9 (LOW)] rule was observed most commonly across almost all liver experiment sets. The rule displayed low-high, high-low, and low-low expression pairs suggesting positive and negative regulation on these gene-pairs and warrants further experimental confirmation in the context of liver cancer. Analyzing the pair-rules in and experiment specific-manner rather than globally provided rules with functional importance. For example, composite rule [AKR1B10 (HIGH) - NQO1 (HIGH)] [TBC1D9B (HIGH) - UBE2G1 (HIGH)] [HADHB (HIGH) - VNN1 (HIGH)] [CYP4F2 (HIGH) - DIO1 (HIGH)] [CD38 (LOW) - PENK (LOW)] [C2ORF27 (LOW) - HELZ (HIGH)] [GYS1 (HIGH) - SNX9 (LOW)] was able to separate the long-term high dose treatment (i.e., 3 samples out of 4 150mg/kg/day for 15 days and 4 samples out of 5 400mg/kg/day for 15 days) effects of peroxisome proliferator-activated receptor-a agonist ciprofibrate, which caused hepatocellular carcinoma, from the vehicle control (5 out of 5) and low dose treatment (3mg/kg/day and 30mg/kg/day for 15 days) on primate liver samples

(GDS1442) [16]. Therefore, this rule might represent a novel co-expressed gene set that is dose-dependently regulated by peroxisome proliferator -activated receptor- during liver carcinogenesis.

Colon disease datasets investigated in the present study were more divergent in nature including colon cancer cell lines (metastatic/non-metastatic, treated/untreated) as well colon biopsy samples (normal, chrons disease, colitis) and primary tumors (with/without recurrence). The most commonly observed gene-pair rule, [MRPL12 (HIGH) - RP2 (LOW)] was not able to separate colon cancer samples from other colon diseases nor from cell line experiments. On the other hand, [GPR64 (LOW) - RRAD (LOW)] rule seem to cluster almost all colon cancer cell lines together although it fails to distinguish between colon cancer recurrence or colon diseases.

GEO data sets consists of biological experiments realized for a specific purpose. For example, a GEO data set may contain microarray samples done on normal and cancer tissues together where normal tissue samples are often used as control samples by the experimenters. Microarray samples in rules generated by our framework consisted of only control or target samples of data sets included. Only small number of rules included mixtures of control and target samples of the same data set while misplaced samples in these rules represented generally duplicates of original samples. Since most rules were determined by the contribution of the majority of the members of a particular microarray experiment set this suggested that experimental conditions together with biological differences also played a role in the observed co-expression patterns of the genes.

A literature survey performed on the genes in the rules generated by our framework resulted in confirmation of expressional regulation of cancer-related genes. FOXM1, TPX2, HIST1H1C, HIST1H2BK, IFIT3, RSAD2, ISG15, HIST1H2BE, DBF4, STIL, KNTC2, MELK, IFIT2, DLG7, BUB1B, HCP5, OASL genes were the most commonly occurring up-regulated genes in rules found using the Breast Cancer Working Set. In previous breast cancer studies, some of these genes have been reported as being up-regulated including FOXM1, TPX2, HIST1H1C, HIST1H2BK, KNTC2, MELK and BUB1B [45, 54, 54, 10, 38, 43, 24].

In the light of all these information, we can say that rules found using the proposed bi-k-bi clustering framework show consistency and relatedness with the literature.

4.4 Discussion

As gene expression profile data sets became more available, application of data mining techniques on these data give valuable hints about gene pattern associations. Several methods have been proposed for mining gene expression profiles. However most of these methods have several limitations.

In this paper, we proposed a novel framework, bi-k-bi clustering, for finding association rules of gene pairs that can easily operate on large scale data. Our framework outputs rules consisting of labeled gene pairs with their associated microarray samples.

One of the most important aspects of this study is its ability to deal with large scale and multiple heterogeneous data sets. By the use of dynamic thresholding on expression profiles, we also alleviate the disadvantages of crude thresholding on expression data.

Available biclustering algorithms require space complexity on large amount of data. For this purpose, we also modified an existing MFI algorithm for biclustering. The proposed methods generally work on a single gene expression data set and cannot handle large number of gene expression data sets in public databases in reasonable amount of time and space. In our experiments, the available biclustering methods we tested were not able to produce an output for the Breast Cancer Data Set with about 20,190 genes and 188 conditions.

In order to test our framework, we applied it on all available NCBI GEO *Homo sapiens* data sets and more specifically five different functionally concise groups of NCBI GEO data sets independently (i.e.: Breast Cancer, Normal Human Tissue, Obesity, Liver and Colon). Gene-pair rules and their association with a given sample set exhibited concordance with the literature. Furthermore, our results provided novel insights into the co-regulated gene pairs among a compendium of tissues as well as diverse conditions of human cancers.

Table 4.8: NCBI GEO datasets used in working sets.

Working Set	**NCBI GEO Dataset**
Breast Cancer	**GDS360**:BREAST CANCER AND DOCETAXEL TREATMENT **GDS817**:BREAST CANCER CELL EXPRESSION PROFILES (HG-U95A) **GDS820**:BREAST CANCER CELL EXPRESSION PROFILES (HG-U133A) **GDS823**:BREAST CANCER CELL EXPRESSION PROFILES (HG-U133B) **GDS881**:BREAST CANCER AND SELECTIVE ESTROGEN RECEPTOR MODULATORS **GDS901**:ESTROGEN RECEPTOR ALPHA L540Q MUTATION EFFECT ON GENE INDUCTION BY ESTRADIOL: TIME COURSE **GDS1250**:ATYPICAL DUCTAL HYPERPLASIA AND BREAST CANCER **GDS1326**:BREAST CANCER CELLS REEXPRESSING ESTROGEN RECEPTOR ALPHA RESPONSE TO 17BETA-ESTRADIOL **GDS1329**:MOLECULAR APOCRINE BREAST TUMORS **GDS2250**:BASAL-LIKE BREAST CANCER TUMORS
Normal Human Tissue	**GDS422**:NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95A) **GDS423**:NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95B) **GDS424**:NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95C) **GDS425**:NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95D) **GDS426**:NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95E)
Obesity	**GDS268**:OBESITY AND FATTY ACID OXIDATIONC **GDS1480**:OBESITY: PREADIPOCYTE EXPRESSION PROFILE (HG-U133A) **GDS1481**:OBESITY: PREADIPOCYTE EXPRESSION PROFILE (HG-U133B) **GDS1493**:OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95A) **GDS1495**:OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95B)) **GDS1496**:OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95C) **GDS1497**:OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95D) **GDS1498**:OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95E)
Liver	**GDS1373**:PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR SUBTYPE ACTIVATION EFFECT ON LIVER CELL **GDS1442**:PPARI AGONIST CIPROFIBRATE EFFECT ON LIVER
Colon	**GDS559**:INFLAMMATORY BOWEL DISEASE (HG-U133A) **GDS560**:INFLAMMATORY BOWEL DISEASE (HG-U133B) **GDS709**:ENTEROCYTE DIFFERENTIATION TIME COURSE **GDS756**:COLON CANCER PROGRESSION **GDS1263**:DUKES B COLON CANCER RECURRENCE **GDS1330**:CROHN DISEASE AND ULCERATIVE COLITIS COMPARISON **GDS1386**:COLORECTAL CARCINOMA SUBTYPE WITH MICROSATELLITE INSTABILITY (HG-U133A) **GDS1387**:COLORECTAL CARCINOMA SUBTYPE WITH MICROSATELLITE INSTABILITY (HG-U133B) **GDS1942**:TRANSGENIC KRAPPEL-LIKE FACTOR 4 INDUCTION: TIME COURSE

CHAPTER 5

CONCLUSION

Due to the advances in technology, it has been possible to deposit very small volumes of many objects in to a very small area. A microarray is an arrayed series of thousands of microscopic spots, printed on a solid substance. Microarray technology allows researchers simultaneously monitor expression levels of thousands of genes in a single experiment.

The excess in the amount of microarray experiments requires the collection and service of these data under a standard format. NCBI Gene Expression Omnibus (GEO) project is a public repository for microarray sample submissions from all over the world [9]. Since these data have been provided in different platforms and come from different laboratories, there is a need for compilation and comprehensive analysis.

As gene expression profile data sets became more available, application of data mining techniques on these data give valuable hints about gene pattern associations. In this thesis, our primary motivation is to address the automation of biological data acquisition and integration from these non-uniform microarray experiments submitted by different experiments using machine learning techniques. Several methods have been proposed for mining gene expression profiles. However most of these methods have several shortcomings. In the scope of this thesis, we studied methods and techniques to alleviate these shortcomings. We focused on mining problems that can easily work on large scale, multiple heterogeneous data sets on a desktop computer with limited computational resources.

At first, we download all *Homo sapiens* microarray experiments from NCBI-GEO and constructed a large scale database from these experimental data along with its associated metadata. After the construction of this curated database, we considered two different mining problems.

In the first problem, we aimed to characterize expression patterns of a published set of housekeeping genes across large number of heterogeneous microarray experiments in our curated database. However, most of the studies available in the literature generally worked on a few homogenous or specially curated data sets. In order to work with multiple heterogeneous data sets we defined and applied a scaling process on the constructed database. We then described methodologies to compare measures of housekeeping genes with those of randomly selected non-housekeeping genes. Our results have supported the claim that housekeeping gene expression is less variable across different experiment sets when compared with randomly selected gene sets. In addition, all previous studies in characterization expression patterns of housekeeping genes focused to find a general reference gene set. However, our initial results demonstrated that rather than focusing on a general reference gene set, cell specific reference gene sets gave better results. We further generated cell origin specific twenty reference gene sets based on the Mesh Headings of the published microarray experimental data. Future directions for this problem can be summarized as follows.

- Reference gene set results can be analyzed and biologically experimented by the scientists.
- Microarray database, described in Chapter 2, can be periodically updated and enlarged with the inclusion of new microarray experiments from NCBI and other available microarray repositories.
- The methodologies and measures can be executed for custom data sets submitted by the users.
- Different subsets of genes can be determined and can be used as housekeeping gene sets for a particular biological condition (e.g., cancer); these gene sets can be useful in microarray data normalization as well as real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) confirmation studies which are performed to amplify defined pieces of RNA molecules.

In the second problem, we proposed a novel framework, bi-k-bi clustering, for finding association rules of gene pairs that can easily operate on large scale and multiple heterogeneous data sets. Our framework proposed a two level biclustering approach that works at the data set and experiment (i.e., condition) levels and discovers similar behaving

gene pairs in multiple data sets. Our approach did not produce a subset of genes in a subset of conditions; rather, we reported *pairs of genes* that behave similarly in a subset of conditions. This property was a setback compared to existing methods; however, it allowed for mining gene expression patterns on a larger scale on a desktop computer with limited computational resources. Our motivation in this problem is to propose a framework that help biologists discover significant gene pair relations and reason about them using their associated microarray samples. We further targeted this framework to work on a desktop computer in a reasonable time and space.

We applied the framework on all available GEO Homo sapiens data sets as well as on five different groups of GEO data sets independently (i.e.: Breast Cancer, Normal Human Tissue, Obesity, Liver and Colon) to test our framework. Gene-pair rules and their association with a given sample set were in accord with the literature. Furthermore, our results provided novel insights into the co-regulated gene pairs among a compendium of tissues as well as diverse conditions of human cancers. Future directions for this problem can be summarized as follows.

- Resulting rules from the tested data set groups can be biologically experimented and analyzed by the scientists to discover novel gene pair relations.
- Rules for other user specified data set groups can be generated.
- The bi-k-bi clustering framework can be extended to work with custom data sets submitted by the users. This extension will help scientists to concentrate and discover novel rules from their own data sets.
- Instead of finding rules of gene pairs the bi-k-bi clustering framework can be extended for finding rules with all associated genes.

REFERENCES

[1] Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing*, 5, 2000.

[2] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 108–1180. ACM, 2000.

[3] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. *J. Parallel Distrib. Comput.*, 61(3):350–371, 2001.

[4] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U S A*, 97(18):10101–10106, 2000.

[6] C. L. Andersen, J. L. Jensen, and T. F. Ørntoft. Normalization of real-time quantitative reverse transcription-pcr data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research*, 64(15):5245–5250, 2004.

[7] R. Balasubramaniyan, E. Hüllermeier, N. Weskamp, and J. Kämper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–1077, 2005.

[8] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler. Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283, 2006.

[9] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue), January 2007.

[10] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[11] C. Becquet, S. Blachon, B. Jeudy, J. F. Boulicaut, and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. *Genome Biology*, 3(12), 2002.

[12] A. Ben-Dor, R. Chor, B.and Karp, and Z. Yakhini. Discovering local structure in gene expression data: The order-preserving submatrix problem. *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 49–57, 2002.

[13] D. Berrar, W. Dubitzky, M. Granzow, and R. Eils. Analysis of gene expression and drug activity data by knowledge-based association mining. *In Proceedings of Critical Assessment of Microarray Data Analysis Techniques (CAMDA'01)*, pages 25–28, 2001.

[14] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using suport vector machines. *Proc. Natl. Acad. Science*, 97:262–267, 2000.

[15] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu. Mafia: A maximal frequent itemset algorithm for transactional databases. *ICDE '01: Proceedings of the 17th International Conference on Data Engineering*, pages 443–452, 2001.

[16] N. F. Cariello, E. H. Romach, H. M. Colton, H. Ni, L. Yoon, J. G. Falls, W. Casey, D. Creech, S. P. Anderson, G. R. Benavides, D. J. Hoivik, R. Brown, and R. T. Miller. Gene expression profiling of the ppar-alpha agonist ciprofibrate in the cynomolgus monkey liver. *Toxicol Science*, 88(1):250–264, 2005.

[17] L. Carkacioglu, T. Can, O. Konu, V. Atalay, and R. Cetin-Atalay. Expression pattern analysis of housekeeping genes across large number of microarray experiments. In *5th European Conference on Computational Biology (ECCB)*, 2006.

[18] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo, and A. Pascual-Montano. Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, 7(1):54, 2006.

[19] Y. Cheng and G. M. Church. Biclustering of expression data. *Proc Int. Conf. Intell. Syst. Mol. Biol.*, 8:93–103, 2000.

[20] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.

[21] JB. de Kok, RW. Roelofs, BA. Giesendorf, JL. Pennings, ET. Waas, T. Feuth, DW. Swinkels, and PN. Span. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab Invest*, 85:154–159, 2005.

[22] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A*, 25(15):14863–8, 1998.

[23] E. Eisenberg and E. Y. Levanon. Human housekeeping genes are compact. *Trends in Genetics*, 19:362, 2003.

[24] J. Fridlyand, A. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B. Ljung, A. Jain, J. McLennan, J. Ziegler, K. Chin, S. Devries, H. Feiler, J. Gray, F. Waldman, D. Pinkel, and D. Albertson. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6(1):96, 2006.

[25] Q. Gao, XY. Wang, J. Fan, Qiu SJ., J. Zhou, YH. Shi, YS. Xiao, Y. Xu, XW. Huang, and J. Sun. Selection of reference genes for real-time pcr in human hepatocellular carcinoma tissues. *Journal of Cancer Research and Clinical Oncology*, 134(9):979–986, 2008.

[26] E. Georgii, L. Richter, U. Rückert, and S. Kramer. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 21(2):123–129, 2005.

[27] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 163–170. IEEE Computer Society, 2001.

[28] A. Gyenesei, U. Wagner, S. Barkow-Oesterreicher, E. Stolte, and R. Schlapbach. Mining co-regulated gene profiles for the detection of functional associations in gene expression data. *Bioinformatics*, 23(15):1927–1935, 2007.

[29] A. O. Hero. Gene selection and ranking with microarray data. In *Proc. Seventh EURASIP/IEEE International Symposium on Signal Processing and its Applications*, Paris, 2003.

[30] L. Hsiao, F. Dangond, T. Yoshide, R. Hong, R. V. Jensen, J. Misra, W. Dillon, F. L. Lee, K. E. Clark, P. Haverty, Z. Weng, G. L Mutter, M. P. Frosh, M. E. Macdonald, E. L. Mildord, C. P. Crum, R. Bueno, R. E. Pratt, M. Mahadevappa, J. A. Warrington, G. Stephanopolous, and S. R. Gullans. A compendium of gene expression in normal human tissues. *Physiol Genomics*, 7(2):97–104, 2001.

[31] C. Huttenhower, M. Hibbs, C. Myers, and O. G. Troyanskaya. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22(23):2890–2897, 2006.

[32] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[33] P. Jin, Y. Zhao, Y. Ngalame, MC. Panelli, D. Nagorsen, V. Monsurro, K. Smith, N. Hu, H. Su, PR. Taylor, FM. Marincola, and E. Wang. Selection and validation of endogenous reference genes using a high throughput approach. *BMC Genomics*, 5:55, 2004.

[34] M. Jung, A. Ramankulov, J. Roigas, M. Johannsen, M. Ringsdorf, G. Kristiansen, and K. Jung.) in search of suitable reference genes for gene expression studies of human renal cell carcinoma by realtime pcr. *BMC Mol Biol*, 47(8), 2007.

[35] P. Kotala, P. Zhou, S. Mudivarthy, W. Perrizo, and E. Deckard. Gene expression profiling of dna microarray data using peano count trees. *In Proceedings of the First Virtual Conference on Genomics and Bioinformatics*, pages 15–16, 2001.

[36] T. C. Kroll and S. Wölfl. Ranking: a closer look on globalization methods for normalization of gene expression arrays. *Bioinformatics*, 30(11):50, 2002.

[37] K. Kyungpil, Z. Shibo, J. Keni, C. Li, L. In-Beum, F. J. Lewis, and H. Haiyan. Measuring similarities between gene expression profiles through new data transformations. *BMC Bioinformatics*, 8:29+, 2007.

[38] M. Lacroix. Significance, detection and markers of disseminated breast cancer cells. *Endocr. Relat. Cancer*, 13(4):1033–67, 2006.

[39] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14:1085–1094, 2004.

[40] Y. H. Lee, S. Nair, E. Rousseau, P. A. Tataranni, C. Bogardus, and P. A. Permana. Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese pima indians: increased expression of inflammation-related genes. *Diabetologia*, 48(9):1776–83, 2005.

[41] M. J. Lercher, A. O. Urrutia, and L. D. Hurst. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, 31(2):180–183, 2002.

[42] Y. F. Leung and D. Cavalieri. Fundamentals of cdna microarray data analysis. *TRENDS in Genetics*, 19(11):649–659, 2003.

[43] M. Lin, J. Park, T. Nishidate, Y. Nakamura, and T. Katagiri. Involvement of maternal embryonic leucine zipper kinase (melk) in mammary carcinogenesis through interaction with bcl-g, a pro-apoptotic member of the bcl-2 family. *Breast Cancer Res.*, 9(1):R17, 2007.

[44] J. Liu, J. Yang, and W. Wang. Biclustering in gene expression data by tendency. *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pages 182–193, 2004.

[45] P. A. Madureira, R. Varshochi, D. Constantinidou, R. E. Francis, R. C. Coombes, K. Yao, and W. Lam. The forkhead box m1 protein regulates the transcription of the estrogen receptor alpha in breast cancer cells. *J. Bio. Chem.*, 281(35):25167–25176, 2006.

[46] S. N. Mukherjee, S. J. Roberts, P. Sykacek, and S. J. Gurr. Gene ranking using bootstrapped p-values. *SIGKDD Explorations*, 5(2):16–22, 2003.

[47] F. Ohl, M. Jung, A. Radonic, M. Sachs, SA. Loening, and K. Jung. Identification and validation of suitable endogenous reference genes for gene expression studies of human bladder cancer. *J. Urol*, 175(5):1915–1920, 2006.

[48] F. Pan, K. Kamath, K. Zhang, S. Pulapura, A. Achar, J. Nunez-Iglesias, Y. Huang, X. Yan, J. Han, H. Hu, M. Xu, J. Hu, and X. J. Zhou. Integrative array analyzer: a software package for analysis of cross-platform and cross-species microarray data. *Bioinformatics*, 22(13):1665–1667, 2006.

[49] J. Pei, X. Zhang, M. Cho, H. Wang, and P. S. Yu. Maple: a fast algorithm for maximal pattern-based clustering. *ICDM 2003. Third IEEE International Conference on Data Mining*, pages 259–266, 2003.

[50] C. Piana, M. Wirth, S. Gerbes, H. Viernstein, F. Gabor, and S. Toegel. Validation of reference genes for qpcr studies on caco-2 cell differentiation. *Eur J Pharm Biopharm.*, 2008.

[51] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32 Suppl:496–501, December 2002.

[52] A. L. Richardson, Z. C. Wang, A. De Nicolo, X. Lu, M. Brown, A. Miron, X. Liao, J. D. Iglehart, D. M. Livingston, and S. Ganesan. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, 9(2):121–32, 2006.

[53] N. Silver, S. Best, J. Jiang, and SL. Thein. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time pcr. *BMC Mol Biology*, 33(7), 2007.

[54] C. Sotiriou, P. Wirapati, S. Loi, S. Harris, A.and Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.*, 98(4):262–72, 2006.

[55] G. Spinsanti, C. Panti, D. Bucalossi, L. Marsili, S. Casini, F. Frati, and MC. Fossi. Selection of reliable reference genes for qrt-pcr studies on cetacean fibroblast cultures exposed to ocs, pbdes, and 17beta-estradiol. *Aquat Toxicol.*, 3(87):178–186, 2008.

[56] A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A*, 99(7):4465–4470, 2002.

[57] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewanand, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U S A*, 96(6):12907–12, 1999.

[58] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–5, 1999.

[59] A. Tuzhilin and G. Adomavicius. Handling very large numbers of association rules in the analysis of microarray data. *In Proceedings of the Eighth ACM SIGKDD International Conference on Data Mining and Knowledge Discovery*, pages 396–404, 2002.

[60] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman. Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7), 2002.

[61] J. Vohradsky. Neural network model of gene expression. *Proc. Natl. Acad. Science*, 15(3):846–54, 2001.

[62] H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. *SIGMOD Conference*, 2002.

[63] J. A. Warrington, M. Nair, A. Mahadevappa, and M. Tsyganskaya. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics*, 2:143–147, 2000.

[64] S. Waxman and E. Wurmbach. De-regulation of common housekeeping genes in hepatocellular carcinoma. *BMC Genomics*, pages 243+, 2007.

[65] J. Yang, H. Wang, W. Wang, and P. Yu. Enhanced biclustering on expression data. *BIBE '03: Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*, pages 321–327, 2003.

[66] Y. Yang, S. Dudoit, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cdna microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):15, 2002.

[67] S. Yoon, Y. Yang, J. Choi, V. Peng, and J. Seong. Large scale data mining approach for gene-specific standardization of microarray gene expression data. *Bioinformatics*, 22(23):2898–2904, 2006.

[68] F. L. Yuk and C. Duccio. Fundamentals of cdna microarray data analysis. *Trends Genetics*, 19:649–659, 2003.

[69] X. Zhang, L. Ding, and AJ. Sandfordm. Selection of reference genes for gene expression studies in human neutrophils by real-time pcr. *BMC Mol Biology*, 6(1), 2005.

[70] L. Zhao and M. J. Zaki. Microcluster: Efficient deterministic biclustering of microarray data. *IEEE Intelligent Systems*, 20(6):40–49, 2005.

APPENDIX A

DATABASE DETAILS

A.1 Implementation Details

MySQL database (*Version 4*) running on Linux operating system is used to store microarray data. Schema for this database is given in Section A.2. In order to insert microarray spot values along with the associated metadata to this database, we developed a number of applications running on the Linux operating system. All these applications are developed with the C programming language and compiled with the GNU C compiler (*gcc*). The brief information of these applications are given as follows.

- **Raw Data Extraction**: Stores raw channel data of the microarray samples. Further, this application stores the GDS and GSM metadata in the database. It first unzips the GDS and its associated GPL file. Then it extracts gene symbol and channel value of the gene symbol for the corresponding spot using GDS and GPL files.

- **Normalization**: Computes and stores the global mean normalized values of the spots in the database.

- **Ranking**: Computes and stores linear and percentile rank values of the spots in the database.

- **Mesh Heading metadata construction**: Connects to NCBI web site and downloads the Mesh Heading information of the data sets in the database (*Some GDS files that may not contain a PubMed Id. For these data sets, publication details are manually queried from known journal and conference web sites*).

- **Rank per gene computation**: Computes the average percentile rank value of each gene symbol in each GSM. In order to ignore the gene symbol, the application also checks whether the rank change of a gene symbol within the GSM exceeds the threshold (i.e.: 20) or not.

- **C_v computation**: Computes the Coefficient of Variation (C_v) for each gene among all the data sets in the database.

- **Mesh Ontology construction**: Downloads the Mesh Heading Ontology (Mesh Tree) from NCBI web site and stores the ontology in the database.

A.2 Database Schema

```
/*
   Spot value of each GSM is stored in this table
*/
CREATE TABLE RANK_GEO_ALL_GSM_EI (
         ROW_ID              INT NOT NULL AUTO_INCREMENT,
         DATASETNAME         VARCHAR(30),
         NAME                VARCHAR(30),
         ID                  VARCHAR(30),
         GENESYMBOL          VARCHAR(30),
         LOG2_CHANNEL1_VALUE DOUBLE,            /* Log Value of the Channel */
         MEAN_CENTERED_LOG2_CH1_VALUE DOUBLE,   /* Global Mean Centered Value */
         RANK_MEAN_CENTERED_CH1_VALUE SMALLINT, /* Linear Rank Value */
         PERC_MEAN_CENTERED_CH1_VALUE SMALLINT, /* Percentile Rank Value */
         PRIMARY KEY (ROW_ID)
)TYPE=INNODB;


/*
   Information about GSMs are stored in this table
*/
CREATE TABLE RANK_GEO_ALL_GSM_INFO_EI (
         ROW_ID             INT NOT NULL AUTO_INCREMENT,
         GSMNAME            VARCHAR(50) NOT NULL,
         DATASETNAME        VARCHAR(50) NOT NULL,
         FEATURECOUNT       INT,
         CHANNELCOUNT       SMALLINT,
         AVG_LOG2_CHANNEL1  DOUBLE,  /* Average LOG value of spots in the GSM */
         PRIMARY KEY (ROW_ID,GSMNAME,DATASETNAME)
)TYPE=INNODB;
```

```
/*
   Associated PubMed Id for the GDSs are stored in this table
*/
CREATE TABLE RANK_GEO_ALL_PUBMED_GDS_EI  (
         ROW_ID           INT NOT NULL AUTO_INCREMENT,
         GDSNAME          VARCHAR(30) NOT NULL ,
         PUBMEDID         INT NOT NULL,  /* PubMed ID */
         PRIMARY KEY (ROW_ID,GDSNAME,PUBMEDID)
)TYPE=INNODB;


/*
   GDS file processed are stored in this table
*/
CREATE TABLE FILES_READ_GDS  (
         ROW_ID           INT NOT NULL AUTO_INCREMENT,
         FILENAME         VARCHAR(100),
         SIZE             INT,
         PRIMARY KEY (ROW_ID)
)TYPE=INNODB;


/*
   Mesh Heading Information for PubMedIds in the RANK_GEO_ALL_PUBMED_GDS_EI
   table are stored in this table
*/
CREATE TABLE PUBMED_MESH  (
         ROW_ID          INT  NOT NULL AUTO_INCREMENT,
         PUBMEDID        INT NOT NULL,
         MESHHEAD        VARCHAR(100) NOT NULL,
         QUALIFIER       VARCHAR(255) NOT NULL,
         PRIMARY KEY (ROW_ID)
)TYPE=INNODB;


/*
   Average Percentile Rank Value of genes in each GSM.
*/
CREATE TABLE PERCRANK_PER_GENE_EI (
        ROW_ID           INT NOT NULL AUTO_INCREMENT,
        GDSNAME          CHAR(10) NOT NULL,
        GSMNAME          CHAR(10) NOT NULL,
        GENESYMBOL       CHAR(50) NOT NULL,
        AVG_PERC_RANK    DOUBLE,  /* Average Percentile Rank Value  */
        PRIMARY KEY (ROW_ID,GDSNAME,GSMNAME,GENESYMBOL)
)TYPE=INNODB;
```

```
/*
   CV values of genes are stored in this table
*/
CREATE TABLE GEO_PERC_CV_EI (
         ROW_ID          INT NOT NULL AUTO_INCREMENT,
         GENESYMBOL      CHAR(50) NOT NULL,
         CV              DOUBLE,
         PRIMARY KEY (ROW_ID,GENESYMBOL)
)TYPE=INNODB;


/*
   Housekeeping Gene symbols listed by Eisenberg,et.al.
*/
CREATE TABLE HKSET2 (
        ROW_ID        INT NOT NULL AUTO_INCREMENT,
        NUCLEOTIDENAME VARCHAR(50),
        GENESYMBOL    VARCHAR(50),
        PRIMARY KEY (ROW_ID)
)TYPE=INNODB;

/*
   Mesh Ontology is stored in this table
*/
CREATE TABLE MESH_TREENUMBERS  (
         ROW_ID            INT  NOT NULL AUTO_INCREMENT,
         MESHHEAD          VARCHAR(100) NOT NULL,
         TREENO            VARCHAR(255) NOT NULL,
         PRIMARY KEY (ROW_ID)
)TYPE=INNODB;
```

APPENDIX B

RESULTS OF KOLMOGOROV-SIMIRNOV TESTS

Tables for the two sample Kolmogorov-Simirnov Test (two-tailed test) results of the hypothesis comparing pair wise Ratio_t distributions of housekeeping genes with random non-housekeeping genes and distributions of random genes among each other are given in Table B.1, Table B.2 and Table B.3, Table B.4 repectively.

In Table B.1 and Table B.3, coefficient of variation thresholds are a-) 0.5, b-) 0.1, c-) 0.05, d-) 0.01 and genes are observed at least 50% of the total sets. In Table B.2 and Table B.4, coefficient of variation threshold is 0.05 and genes are observed at least a-) 75%, b-) 50%, c-) 25% and d-) 5% of the total sets. The hypothesis used in tests is:

- H_0: Two distributions are not different.
- H_a: Two distributions are different.

In Table B.1 and Table B.3, the computed p-values are lower than the significance level (alpha=0.05) in all cases. Therefore, H_0 is rejected with the risk it is true is lower than 0.01%.

In Table B.2 and Table B.4, H_0 is rejected in cases where the computed p-values are lower than the significance level (alpha=0.05). Test results shown in bold face have p-values higher than the significance level and H_0 is accepted in these cases. However, when Bonferroni adjustment is done on the p-values, H_0 is rejected. The risk to reject H_0 while it is true is at most lower than 3%.

Table B.1: Kolmogorov-Simirnov tests for pair wise distribution of non-housekeeping genes (C_v *thresholding*)

Two sample Kolmogorov -	a			b			c			d		
Simirnov test / Two-tailed test	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha
Housekeeping Set vs. Random Set1	*0.443*	*<0.0001*	*0.05*	*0.505*	*<0.0001*	*0.05*	*0.511*	*<0.0001*	*0.05*	*0.458*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set2	*0.454*	*<0.0001*	*0.05*	*0.484*	*<0.0001*	*0.05*	*0.493*	*<0.0001*	*0.05*	*0.463*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set3	*0.390*	*<0.0001*	*0.05*	*0.452*	*<0.0001*	*0.05*	*0.436*	*<0.0001*	*0.05*	*0.417*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set4	*0.417*	*<0.0001*	*0.05*	*0.481*	*<0.0001*	*0.05*	*0.479*	*<0.0001*	*0.05*	*0.424*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set5	*0.431*	*<0.0001*	*0.05*	*0.461*	*<0.0001*	*0.05*	*0.465*	*<0.0001*	*0.05*	*0.429*	*<0.0001*	*0.05*

Table B.2: Kolmogorov-Simirnov tests for pair wise distribution of non-housekeeping genes (C_v *thresholding*)

Two sample Kolmogorov -	a			b			c			d		
Simirnov test / Two-tailed test	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha
Random Set1 vs. Random Set2	*0.049*	*0.357*	*0.05*	*0.034*	*0.769*	*0.05*	*0.041*	*0.571*	*0.05*	*0.030*	*0.761*	*0.05*
Random Set2 vs. Random Set3	**0.127**	**0.000**	**0.05**	**0.104**	**0.003**	**0.05**	**0.102**	**0.004**	**0.05**	*0.055*	*0.197*	*0.05*
Random Set3 vs. Random Set4	*0.088*	*0.019*	*0.05*	*0.071*	*0.091*	*0.05*	*0.102*	*0.004*	*0.05*	*0.025*	*0.919*	*0.05*
Random Set4 vs. Random Set5	*0.034*	*0.776*	*0.05*	*0.037*	*0.681*	*0.05*	*0.051*	*0.329*	*0.05*	*0.032*	*0.726*	*0.05*

Table B.3: Kolmogorov-Simirnov tests for pair wise distribution of housekeeping and non-housekeeping genes (*PO thresholding*)]

Two sample Kolmogorov -	a			b			c			d		
Simirnov test / Two-tailed test	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha
Housekeeping Set vs. Random Set1	*0.385*	*<0.0001*	*0.05*	*0.511*	*<0.0001*	*0.05*	*0.587*	*<0.0001*	*0.05*	*0.555*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set2	*0.367*	*<0.0001*	*0.05*	*0.493*	*<0.0001*	*0.05*	*0.576*	*<0.0001*	*0.05*	*0.511*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set3	*0.352*	*<0.0001*	*0.05*	*0.436*	*<0.0001*	*0.05*	*0.518*	*<0.0001*	*0.05*	*0.500*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set4	*0.362*	*<0.0001*	*0.05*	*0.479*	*<0.0001*	*0.05*	*0.564*	*<0.0001*	*0.05*	*0.519*	*<0.0001*	*0.05*
Housekeeping Set vs. Random Set5	*0.348*	*<0.0001*	*0.05*	*0.465*	*<0.0001*	*0.05*	*0.544*	*<0.0001*	*0.05*	*0.512*	*<0.0001*	*0.05*

Table B.4: Kolmogorov-Simirnov tests for pair wise distribution of non-housekeeping genes (*PO thresholding*)

Two sample Kolmogorov -	a			b			c			d		
Simirnov test / Two-tailed test	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha	D	p-value	alpha
Random Set1 vs. Random Set2	*0.042*	*0.243*	*0.05*	*0.041*	*0.571*	*0.05*	*0.042*	*0.606*	*0.05*	**0.083**	**0.039**	**0.05**
Random Set2 vs. Random Set3	*0.172*	*0.172*	*0.05*	**0.102**	**0.004**	**0.05**	**0.005**	**0.005**	**0.05**	*0.053*	*0.397*	*0.05*
Random Set3 vs. Random Set4	*0.076*	*0.076*	*0.05*	*0.074*	*0.067*	*0.05*	**0.022**	**0.022**	**0.05**	*0.057*	*0.319*	*0.05*
Random Set4 vs. Random Set5	*0.032*	*0.529*	*0.05*	*0.051*	*0.329*	*0.05*	*0.352*	*0.352*	*0.05*	*0.053*	*0.393*	*0.05*

APPENDIX C

REFERENCE GENE SET OVER ALL SETS

Table C.1: Reference gene set over all sets (C_v=0.12 and Sensitivity=0.5)

Gene Symbol	**Name**	**Ratio$_t$**	**Percentile Rank**			**Is in HK Set**
			Mean	**Median**	**Std. Dev.**	
AARS	(alanyl-tRNA synthetase)	0.926	85.659	91	16.005	No
ABCF1	(ATP-binding cassette; sub-family F (GCN20); member 1)	0.941	80.98	86	17.601	No
ACADVL	(acyl-Coenzyme A dehydrogenase; very long chain)	0.913	87.286	91	12.952	No
ACTB	(actin; beta)	0.949	94.965	99	15.704	Yes
ACTG1	(actin; gamma 1)	0.971	95.342	99	14.66	Yes
ADAR	(adenosine deaminase; RNA-specific)	0.929	90.092	94	13.87	Yes
ADIPOR2	(adiponectin receptor 2)	0.939	84.809	87	11.649	No
ADRM1	(adhesion regulating molecule 1)	0.909	83.378	88	13.797	No

Table C.1: Cont'd

AGPAT1	(1-acylglycerol-3-phosphate O-acyltransferase 1 (lysophosphatidic acid acyltransferase; alpha))	0.926	84.729	88	12.879	Yes
AHSA1	(AHA1; activator of heat shock 90kDa protein ATPase homolog 1 (yeast))	0.935	85.409	89	13.125	Yes
AKR1A1	(aldo-keto reductase family 1; member A1 (aldehyde reductase))	0.916	85.17	89	14.943	Yes
ALDH9A1	(aldehyde dehydrogenase 9 family; member A1)	0.902	83.285	87	14.571	No
ALDOA	(aldolase A; fructose-bisphosphate)	0.926	92.929	98	15.779	Yes
ANP32B	(acidic (leucine-rich) nuclear phosphoprotein 32 family; member B)	0.931	91.073	95.5	15.328	Yes
AP2S1	(adaptor-related protein complex 2; sigma 1 subunit)	0.935	88.409	91.667	12.359	Yes
APEX1	(APEX nuclease (multifunctional DNA repair enzyme) 1)	0.913	87.861	92	14.107	No
ARF1	(ADP-ribosylation factor 1)	0.935	88.726	95	19.684	Yes
ARF3	(ADP-ribosylation factor 3)	0.947	85.986	92	19.61	Yes
ARF4	(ADP-ribosylation factor 4)	0.937	89.188	93	13.693	Yes
ARF5	(ADP-ribosylation factor 5)	0.92	85.273	89	13.575	Yes
ARPC2	(actin related protein 2/3 complex; subunit 2; 34kDa)	0.952	92.016	96	15.303	Yes
ARPC3	(actin related protein 2/3 complex; subunit 3; 21kDa)	0.908	85.798	91	15.839	Yes
ATF4	(activating transcription factor 4 (tax-responsive enhancer element B67))	0.952	92.415	96	14.716	Yes
ATP5A1	(ATP synthase; H+ transporting; mitochondrial F1 complex; alpha subunit 1; cardiac muscle)	0.924	92.503	97	14.54	Yes

Table C.1: Cont'd

ATP5G1	(ATP synthase; H+ transporting; mitochondrial F0 complex; subunit C1 (subunit 9))	0.94	86.73	90	12.854	Yes
ATP5I	(ATP synthase; H+ transporting; mitochondrial F0 complex; subunit E)	0.969	92.411	94.5	8.863	Yes
ATP6V0B	(ATPase; H+ transporting; lysosomal 21kDa; V0 subunit b)	0.947	88.247	94	18.572	Yes
ATP6V0C	(ATPase; H+ transporting; lysosomal 16kDa; V0 subunit c)	0.945	90.741	94.5	13.877	Yes
ATP6V1E1	(ATPase; H+ transporting; lysosomal 31kDa; V1 subunit E1)	0.939	86.196	90	15.279	Yes
ATP6V1F	(ATPase; H+ transporting; lysosomal 14kDa; V1 subunit F)	0.958	90.552	93	11.547	Yes
ATP6V1G1	(ATPase; H+ transporting; lysosomal 13kDa; V1 subunit G1)	0.919	85.787	91	13.868	Yes
B2M	(beta-2-microglobulin)	0.975	96.399	99	10.644	Yes
BANF1	(barrier to autointegration factor 1)	0.927	89.011	94	13.538	Yes
BCAP31	(B-cell receptor-associated protein 31)	0.943	89.347	94	13.947	Yes
BECN1	(beclin 1 (coiled-coil; myosin-like BCL2 interacting protein))	0.926	83.88	85.5	12.35	Yes
BRD2	(bromodomain containing 2)	0.94	84.245	86.333	9.798	No
C11ORF58	(chromosome 11 open reading frame 58)	0.923	86.13	91	13.773	No

Table C.1: Cont'd

C14ORF2	(chromosome 14 open reading frame 2)	0.919	85.303	91	14.969	Yes
C19ORF10	(chromosome 19 open reading frame 10)	0.917	85.007	88	12.172	No
C7ORF24	(chromosome 7 open reading frame 24)	0.925	83.841	87	13.342	No
CALM2	(calmodulin 2 (phosphorylase kinase; delta))	0.926	93.761	98	15.126	Yes
CANX	(calnexin)	0.954	90.163	95	17.561	Yes
CAP1	(CAP; adenylate cyclase-associated protein 1 (yeast))	0.958	91.646	95.5	13.317	No
CAPNS1	(calpain; small subunit 1)	0.92	88.136	96	21.604	Yes
CASC3	(cancer susceptibility candidate 3)	0.931	82.904	87	14.356	Yes
CCT3	(chaperonin containing TCP1; subunit 3 (gamma))	0.936	89.843	94	13.739	Yes
CCT4	(chaperonin containing TCP1; subunit 4 (delta))	0.929	86.215	95	17.527	No
CCT7	(chaperonin containing TCP1; subunit 7 (eta))	0.905	86.725	93	17.469	Yes
CD81	(CD81 molecule)	0.94	90.939	97	16.058	Yes
CDIPT	(CDP-diacylglycerol–inositol 3-phosphatidyltransferase (phosphatidylinositol synthase))	0.957	88.245	92	13.971	No
CDK2AP1	(CDK2-associated protein 1)	0.909	87.181	93	15.485	No
CFL1	(cofilin 1 (non-muscle))	0.974	94.945	99	15.778	Yes

Table C.1: Cont'd

CHMP2A	(chromatin modifying protein 2A)	0.929	85.71	90	13.396	No
CIB1	(calcium and integrin binding 1 (calmyrin))	0.934	83.899	88	13.439	No
CIRBP	(cold inducible RNA binding protein)	0.948	89.756	93.5	12.696	No
CKAP1	(tubulin folding cofactor B)	0.922	86.926	90.333	13.088	Yes
CLIC1	(chloride intracellular channel 1)	0.901	90.014	97	16.581	No
CLSTN1	(calsyntenin 1)	0.902	84.539	89	16.557	Yes
CNIH	(cornichon homolog (Drosophila))	0.918	86.314	91	14.125	No
COBRA1	(cofactor of BRCA1)	0.904	80.428	83	14.498	Yes
COPS6	(COP9 constitutive photomorphogenic homolog subunit 6 (Arabidopsis))	0.971	86.604	88	11.116	Yes
COX6B1	(cytochrome c oxidase subunit Vib polypeptide 1 (ubiquitous))	0.956	92.793	96	12.601	Yes
COX6C	(cytochrome c oxidase subunit VIc)	0.937	91.712	97	15.523	No
COX7A2	(cytochrome c oxidase subunit VIIa polypeptide 2 (liver))	0.961	93.552	97	14.196	Yes
COX7A2L	(cytochrome c oxidase subunit VIIa polypeptide 2 like)	0.924	88.045	93	16.537	Yes
COX7B	(cytochrome c oxidase subunit VIIb)	0.914	88.414	94	15.18	No
COX7C	(cytochrome c oxidase subunit VIIc)	0.965	92.564	95.667	12.348	Yes
COX8A	(cytochrome c oxidase subunit 8A (ubiquitous))	0.958	93.52	97	13.918	Yes

Table C.1: Cont'd

CPSF4	(cleavage and polyadenylation specific factor 4; 30kDa)	0.918	78.131	80	11.377	No
CS	(citrate synthase)	0.948	90.335	94	12.26	No
CSNK2B	(casein kinase 2; beta polypeptide)	0.922	88.671	94	16.042	Yes
CYB5R3	(cytochrome b5 reductase 3)	0.942	91.971	95	13.185	No
CYC1	(cytochrome c-1)	0.907	87.155	91	13.966	Yes
DAD1	(defender against cell death 1)	0.935	88.7	94	17.207	Yes
DCTN2	(dynactin 2 (p50))	0.93	83.167	86	14.337	No
DCTN3	(dynactin 3 (p22))	0.946	84.6	88	12.578	No
DDB1	(damage-specific DNA binding protein 1; 127kDa)	0.946	88.244	92	12.297	No
DDX48	(DEAD (Asp-Glu-Ala-Asp) box polypeptide 48)	0.944	87.382	91	12.456	No
DDX5	(DEAD (Asp-Glu-Ala-Asp) box polypeptide 5)	0.929	90.558	96	18.331	No
DNAJA1	(DnaJ (Hsp40) homolog; subfamily A; member 1)	0.93	87.993	91	12.277	No
DRG1	(developmentally regulated GTP binding protein 1)	0.93	86.623	90	12.45	No
DULLARD	(dullard homolog (Xenopus laevis))	0.95	85	89	16.587	Yes
DYNLL1	(dynein; light chain; LC8-type 1)	0.953	92.886	97	13.473	No
DYNLT1	(dynein; light chain; Tctex-type 1)	0.924	88.241	92	13.196	No

Table C.1: Cont'd

ECHS1	(enoyl Coenzyme A hydratase; short chain; 1; mitochondrial)	0.911	85.447	92	17.3	No
EEF1A1	(eukaryotic translation elongation factor 1 alpha 1)	0.93	94.242	99	15.134	No
EEF1B2	(eukaryotic translation elongation factor 1 beta 2)	0.959	94.59	98	12.635	No
EEF2	(eukaryotic translation elongation factor 2)	0.979	95.237	98	13.723	No
EI24	(etoposide induced 2.4 mRNA)	0.916	80.339	83.5	14.205	No
EIF2B2	(eukaryotic translation initiation factor 2B; subunit 2 beta; 39kDa)	0.927	77.183	80	12.372	No
EIF3S12	(eukaryotic translation initiation factor 3; subunit 12)	0.934	90.279	95	13.889	No
EIF3S2	(eukaryotic translation initiation factor 3; subunit 2 beta; 36kDa)	0.949	88.284	92	11.938	Yes
EIF3S3	(eukaryotic translation initiation factor 3; subunit 3 gamma; 40kDa)	0.937	91.133	96	15.845	No
EIF3S4	(eukaryotic translation initiation factor 3; subunit 4 delta; 44kDa)	0.929	89.707	94	12.958	Yes
EIF3S5	(eukaryotic translation initiation factor 3; subunit 5 epsilon; 47kDa)	0.941	90.756	96	18.856	Yes
EIF3S6	(eukaryotic translation initiation factor 3; subunit 6 48kDa)	0.904	91.67	96	14.61	No
EIF3S7	(eukaryotic translation initiation factor 3; subunit 7 zeta; 66/67kDa)	0.935	89.167	96	19.99	Yes
EIF4A2	(eukaryotic translation initiation factor 4A; isoform 2)	0.923	91.637	98	18.16	Yes
EIF4B	(eukaryotic translation initiation factor 4B)	0.931	90.168	94.5	14.436	No
ERH	(enhancer of rudimentary homolog (Drosophila))	0.935	88.616	95	18.268	Yes

Table C.1: Cont'd

ERP29	(endoplasmic reticulum protein 29)	0.913	88.348	94	15.547	No
ETF1	(eukaryotic translation termination factor 1)	0.915	84.25	88	13.268	No
ETHE1	(ethylmalonic encephalopathy 1)	0.904	79.975	84	15.04	No
FAM32A	(family with sequence similarity 32; member A)	0.944	82.476	86	12.399	No
FAU	(Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived); ribosomal protein S30)	0.962	93.723	98	18.297	Yes
FIBP	(fibroblast growth factor (acidic) intracellular binding protein)	0.913	82.547	87	13.303	No
FTH1	(ferritin; heavy polypeptide 1)	0.946	93.066	97	13.45	Yes
GABARAPL2	(GABA(A) receptor-associated protein-like 2)	0.959	91.951	95	11.259	Yes
GANAB	(glucosidase; alpha; neutral AB)	0.912	87.137	91	12.511	Yes
GARS	(glycyl-tRNA synthetase)	0.906	87.343	92	14.299	No
GDI1	(GDP dissociation inhibitor 1)	0.937	86.932	92	16.809	Yes
GDI2	(GDP dissociation inhibitor 2)	0.943	88.961	94	17.403	Yes
GLO1	(glyoxalase I)	0.925	88.122	94	18.466	No
GNB2	(guanine nucleotide binding protein (G protein); beta polypeptide 2)	0.907	85.698	89	12.478	Yes
GNG5	(guanine nucleotide binding protein (G protein); gamma 5)	0.935	89.914	95	13.639	No
GOT2	(glutamic-oxaloacetic transaminase 2; mitochondrial (aspartate aminotransferase 2))	0.923	85.95	90	14.019	Yes
GPAA1	(glycosylphosphatidylinositol anchor attachment protein 1 homolog (yeast))	0.926	85.022	88.333	12.437	Yes

Table C.1: Cont'd

GPX1	(glutathione peroxidase 1)	0.923	91.148	96	15.33	No
GPX4	(glutathione peroxidase 4 (phospholipid hydroperoxidase))	0.939	91.121	95	12.617	Yes
GSTO1	(glutathione S-transferase omega 1)	0.937	90.035	94	13.538	No
H2AFZ	(H2A histone family; member Z)	0.943	90.656	94	11.714	No
H3F3A	(H3 histone; family 3A)	0.965	95.522	98	11.995	Yes
HADH2	(hydroxysteroid (17-beta) dehydrogenase 10)	0.917	83.389	88	15.573	No
HADHB	(hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein); beta subunit)	0.946	90.819	94	11.808	Yes
HBXIP	(hepatitis B virus x interacting protein)	0.953	89.04	92	12.786	No
HDGF	(hepatoma-derived growth factor (high-mobility group protein 1-like))	0.959	90.753	93.5	10.982	Yes
HINT1	(histidine triad nucleotide binding protein 1)	0.941	93.823	97	11.16	Yes
HLA-A	(major histocompatibility complex; class I; A)	0.931	93.954	99	14.997	No
HLA-G	(HLA-G histocompatibility antigen; class I; G)	0.915	87.686	94.75	19.633	Yes
HMGB1	(high-mobility group box 1)	0.923	85.77	94.667	17.465	Yes
HMGN1	(high-mobility group nucleosome binding domain 1)	0.94	90.374	96	15.911	No
HMGN2	(high-mobility group nucleosomal binding domain 2)	0.943	94.842	98	13.005	No
HMGN4	(high mobility group nucleosomal binding domain 4)	0.904	81.249	84	12.548	No

Table C.1: Cont'd

HNRPC	(heterogeneous nuclear ribonucleoprotein C (C1/C2))	0.919	85.057	92	21.648	No
HNRPK	(heterogeneous nuclear ribonucleoprotein K)	0.944	91.231	96.5	17.852	Yes
HSBP1	(heat shock factor binding protein 1)	0.926	85.166	90	16.606	Yes
HSP90AA1	(heat shock protein 90kDa alpha (cytosolic); class A member 1)	0.958	93.571	97.75	15.844	No
HYOU1	(hypoxia up-regulated 1)	0.936	87.481	92	13.435	Yes
IER2	(immediate early response 2)	0.933	90.927	95	12.127	Yes
IKBKG	(inhibitor of kappa light polypeptide gene enhancer in B-cells; kinase gamma)	0.902	77.387	79	12.734	No
ILF2	(interleukin enhancer binding factor 2; 45kDa)	0.92	84.184	90	18.837	Yes
ITGB4BP	(integrin beta 4 binding protein)	0.949	86.768	91	12.938	No
KARS	(lysyl-tRNA synthetase)	0.932	89.835	95	17.833	Yes
KDELR2	(KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2)	0.911	84.622	89	13.635	No
KRT10	(keratin 10 (epidermolytic hyperkeratosis; keratosis palmaris et plantaris))	0.904	83.352	88	16.998	No
LAPTM4A	(lysosomal-associated protein transmembrane 4 alpha)	0.952	92.447	96	11.915	No
LASP1	(LIM and SH3 protein 1)	0.927	87.622	94	17.666	Yes
LDHA	(lactate dehydrogenase A)	0.949	93.581	98	15.961	Yes

Table C.1: Cont'd

LSM3	(LSM3 homolog; U6 small nuclear RNA associated (S. cerevisiae))	0.901	83.618	87	13.608	No
LTA4H	(leukotriene A4 hydrolase)	0.939	88.004	91	13.455	No
LYPLA2	(lysophospholipase II)	0.945	85.945	88	13.139	Yes
MARCKSL1	(MARCKS-like 1)	0.902	85.545	91	16.767	No
MDH1	(malate dehydrogenase 1; NAD (soluble))	0.912	89.83	96	16.442	Yes
MGRN1	(mahogunin; ring finger 1)	0.927	80.149	84	15.134	No
MIF	(SMT3 suppressor of mif two 3 homolog 4 (S. cerevisiae))	0.903	90.522	98	19.517	Yes
MLF2	(myeloid leukemia factor 2)	0.908	83.598	88	13.388	Yes
MRCL3	(myosin regulatory light chain MRCL3)	0.902	87.378	94	18.344	No
MRPL49	(mitochondrial ribosomal protein L49)	0.916	83.409	88	13.706	No
MYST2	(MYST histone acetyltransferase 2)	0.912	74.891	78	16.66	Yes
NACA	(nascent-polypeptide-associated complex alpha polypeptide)	0.968	95.573	98.5	12.09	Yes
NARS	(asparaginyl-tRNA synthetase)	0.937	89.69	95	18.024	No
NCL	(nucleolin)	0.961	93.296	97	13.34	Yes
NCOA4	(nuclear receptor coactivator 4)	0.932	89.811	95	15.989	No
NDUFA1	(NADH dehydrogenase (ubiquinone) 1 alpha subcomplex; 1; 7.5kDa)	0.943	93.347	97	13.056	Yes
NDUFAB1	(NADH dehydrogenase (ubiquinone) 1; alpha/beta subcomplex; 1; 8kDa)	0.936	90.585	95	13.374	No

Table C.1: Cont'd

NDUFB3	(NADH dehydrogenase (ubiquinone) 1 beta subcomplex; 3; 12kDa)	0.902	84.484	89	15.979	No
NDUFB5	(NADH dehydrogenase (ubiquinone) 1 beta subcomplex; 5; 16kDa)	0.946	87.597	91	12.4	No
NDUFS2	(NADH dehydrogenase (ubiquinone) Fe-S protein 2; 49kDa (NADH-coenzyme Q reductase))	0.943	85.939	90	13.479	No
NDUFS3	(NADH dehydrogenase (ubiquinone) Fe-S protein 3; 30kDa (NADH-coenzyme Q reductase))	0.932	87.626	92	14.107	No
NDUFS5	(NADH dehydrogenase (ubiquinone) Fe-S protein 5; 15kDa (NADH-coenzyme Q reductase))	0.963	92.672	96	11.779	Yes
NDUFS6	(NADH dehydrogenase (ubiquinone) Fe-S protein 6; 13kDa (NADH-coenzyme Q reductase))	0.935	86.411	90	13.479	No
NFE2L1	(nuclear factor (erythroid-derived 2)-like 1)	0.913	82.222	88	17.618	No
NIFUN	(IscU iron-sulfur cluster scaffold homolog (E. coli))	0.918	89.848	94	13.826	No
NME2	(non-metastatic cells 2; protein (NM23B) expressed in)	0.957	93.534	97	13.017	Yes
NONO	(non-POU domain containing; octamer-binding)	0.949	89.066	95	19.616	Yes
NPC2	(Niemann-Pick disease; type C2)	0.949	89.689	95	15.517	No
NXF1	(nuclear RNA export factor 1)	0.919	82.726	87	14.365	Yes
OAZ1	(ornithine decarboxylase antizyme 1)	0.967	93.532	98.5	18.639	No
ODC1	(ornithine decarboxylase 1)	0.929	90.499	95	12.36	Yes
OS9	(amplified in osteosarcoma)	0.937	88.556	92	13.232	No
PABPC1	(poly(A) binding protein; cytoplasmic 1)	0.933	92.888	98	16.416	Yes

Table C.1: Cont'd

PABPC4	(poly(A) binding protein; cytoplasmic 4 (inducible form))	0.92	86.835	91	13.878	No
PARK7	(Parkinson disease (autosomal recessive; early onset) 7)	0.958	92.69	97	17.688	Yes
PCBP1	(poly(rC) binding protein 1)	0.936	90.51	95	15.644	No
PCMT1	(protein-L-isoaspartate (D-aspartate) O-methyltransferase)	0.907	81.287	85.333	15.529	No
PEBP1	(phosphatidylethanolamine binding protein 1)	0.914	86.217	91.5	16.607	No
PFN1	(profilin 1)	0.937	92.514	97	14.633	Yes
PGAM1	(phosphoglycerate mutase 1 (brain))	0.958	92.488	97	13.858	No
PLOD3	(procollagen-lysine; 2-oxoglutarate 5-dioxygenase 3)	0.933	83.31	86	12.652	No
POLR2G	(polymerase (RNA) II (DNA directed) polypeptide G)	0.933	86.32	91	14.069	No
POLR2H	(polymerase (RNA) II (DNA directed) polypeptide H)	0.929	83.104	86	11.704	No
PPP1CC	(protein phosphatase 1; catalytic subunit; gamma isoform)	0.945	89.628	94	14.205	No
PPP1R11	(protein phosphatase 1; regulatory (inhibitor) subunit 11)	0.937	85.419	89	12.402	Yes
PPP6C	(protein phosphatase 6; catalytic subunit)	0.909	81.746	86	14.138	No
PPT1	(palmitoyl-protein thioesterase 1 (ceroid-lipofuscinosis; neuronal 1; infantile))	0.911	85.404	90	15.788	No
PRDX1	(peroxiredoxin 1)	0.949	93.859	98	13.256	Yes
PRPF8	(PRP8 pre-mRNA processing factor 8 homolog (S. cerevisiae))	0.909	86.049	92	19.079	Yes

Table C.1: Cont'd

PSAP	(prosaposin (variant Gaucher disease and variant metachromatic leukodystrophy))	0.944	92.733	97	14.567	No
PSMA1	(proteasome (prosome; macropain) subunit; alpha type; 1)	0.901	86.2	94	15.625	No
PSMA4	(proteasome (prosome; macropain) subunit; alpha type; 4)	0.927	88.458	92	12.358	No
PSMA6	(proteasome (prosome; macropain) subunit; alpha type; 6)	0.935	91.702	96	14.149	No
PSMB3	(proteasome (prosome; macropain) subunit; beta type; 3)	0.946	90.727	94	12.212	No
PSMB4	(proteasome (prosome; macropain) subunit; beta type; 4)	0.918	89.58	94	14.69	Yes
PSMB6	(proteasome (prosome; macropain) subunit; beta type; 6)	0.943	88.775	93	12.731	No
PSMB7	(proteasome (prosome; macropain) subunit; beta type; 7)	0.939	87.1	91	12.578	Yes
PSMC1	(proteasome (prosome; macropain) 26S subunit; ATPase; 1)	0.968	92.289	95	10.364	No
PSMC2	(proteasome (prosome; macropain) 26S subunit; ATPase; 2)	0.917	84.637	88	12.182	No
PSMC5	(proteasome (prosome; macropain) 26S subunit; ATPase; 5)	0.939	87.029	92	15.827	No
PSMD1	(proteasome (prosome; macropain) 26S subunit; non-ATPase; 1)	0.916	84.531	89	14.58	No
PSMD2	(proteasome (prosome; macropain) 26S subunit; non-ATPase; 2)	0.95	89.762	93	12.741	No
PSMD6	(proteasome (prosome; macropain) 26S subunit; non-ATPase; 6)	0.913	85.486	91	15.861	No

Table C.1: Cont'd

PSMD7	(proteasome (prosome; macropain) 26S subunit; non-ATPase; 7 (Mov34 homolog))	0.926	83.364	86	12.004	No
PSME1	(proteasome (prosome; macropain) activator subunit 1 (PA28 alpha))	0.959	89.749	94	12.799	No
PSME2	(proteasome (prosome; macropain) activator subunit 2 (PA28 beta))	0.945	89.391	93	11.962	Yes
PTDSS1	(phosphatidylserine synthase 1)	0.952	89.62	93	12.802	Yes
PTP4A2	(protein tyrosine phosphatase type IVA; member 2)	0.92	88.766	92.5	14.637	No
RAB8A	(RAB8A; member RAS oncogene family)	0.906	84.054	88	13.952	Yes
RABAC1	(Rab acceptor 1 (prenylated))	0.933	88.487	93	13.785	Yes
RAC1	(ras-related C3 botulinum toxin substrate 1 (rho family; small GTP binding protein Rac1))	0.93	91.921	95.5	13.491	Yes
RAD23A	(RAD23 homolog A (S. cerevisiae))	0.904	85.385	88	12.322	Yes
RAN	(RAN; member RAS oncogene family)	0.93	91.681	95	12.311	Yes
RBMX	(RNA binding motif protein; X-linked)	0.907	84.843	90	15.351	No
RHOA	(ras homolog gene family; member A)	0.915	90.192	97.5	21.91	Yes
RHOG	(ras homolog gene family; member G (rho G))	0.911	83.634	88	14.911	No
RING1	(ring finger protein 1)	0.933	81.467	84.5	13.505	Yes
RNPS1	(RNA binding protein S1; serine-rich domain)	0.93	86.552	91.5	18.598	Yes
RPL10A	(ribosomal protein L10a)	0.949	92.091	98	19.777	Yes

Table C.1: Cont'd

RPL11	(ribosomal protein L11)	0.947	93.402	99	18.579	Yes
RPL12	(ribosomal protein L12)	0.949	95.298	99	14.252	No
RPL13	(ribosomal protein L13)	0.965	92.765	98.75	19.627	Yes
RPL13A	(ribosomal protein L13a)	0.965	96.004	99	13.176	Yes
RPL15	(ribosomal protein L15)	0.904	92.633	98	16.211	Yes
RPL19	(ribosomal protein L19)	0.939	93.226	99	19.572	Yes
RPL21	(ribosomal protein L21)	0.961	94.158	99	18.146	No
RPL22	(ribosomal protein L22)	0.967	94.784	97.2	11.95	No
RPL23A	(ribosomal protein L23a)	0.952	94.018	99	15.782	No
RPL24	(ribosomal protein L24)	0.933	92.901	99	20.163	No
RPL27	(ribosomal protein L27)	0.973	94.484	99	17.652	Yes
RPL28	(ribosomal protein L28)	0.916	88.625	99	23.502	No
RPL29	(ribosomal protein L29)	0.958	95.353	98.5	12.452	Yes
RPL3	(ribosomal protein L3)	0.973	95.852	99	13.387	Yes
RPL30	(ribosomal protein L30)	0.962	94.383	99	18.104	No
RPL32	(ribosomal protein L32)	0.941	95.06	99	14.992	Yes
RPL34	(ribosomal protein L34)	0.948	92.766	99	20.247	Yes
RPL35	(ribosomal protein L35)	0.958	93.498	98	17.986	Yes
RPL36A	(ribosomal protein L36a)	0.955	95.333	99	12.362	No
RPL36AL	(ribosomal protein L36a-like)	0.931	92.618	96	13.206	Yes
RPL37	(ribosomal protein L37)	0.927	94.077	99	16.235	Yes
RPL38	(ribosomal protein L38)	0.97	93.957	97.333	11.816	Yes
RPL4	(ribosomal protein L4)	0.933	93.318	98.333	17.631	No
RPL41	(ribosomal protein L41)	0.933	93.786	99	18.188	No
RPL6	(ribosomal protein L6)	0.958	93.461	98	18.981	No
RPL7	(ribosomal protein L7)	0.972	96.105	99	12.414	No

Table C.1: Cont'd

RPL8	(ribosomal protein L8)	0.955	95.497	99	13.617	Yes
RPL9	(ribosomal protein L9)	0.968	93.767	99	18.444	No
RPLP0	(ribosomal protein; large; P0)	0.947	95.257	98.8	13.53	No
RPN2	(ribophorin II)	0.965	91.518	95	12.506	No
RPS10	(ribosomal protein S10)	0.976	96.491	99	11.661	Yes
RPS11	(ribosomal protein S11)	0.962	92.232	99	19.844	Yes
RPS13	(ribosomal protein S13)	0.969	96.238	99	12.376	Yes
RPS15	(ribosomal protein S15)	0.948	93.719	99	16.815	Yes
RPS16	(ribosomal protein S16)	0.967	95.918	99	12.538	Yes
RPS17	(ribosomal protein S17)	0.971	96.31	99	11.97	No
RPS18	(ribosomal protein S18)	0.955	95.939	99	13.673	Yes
RPS23	(ribosomal protein S23)	0.944	94.469	99	14.299	No
RPS24	(ribosomal protein S24)	0.95	93.272	99	19.099	Yes
RPS25	(ribosomal protein S25)	0.959	95.031	98	12.315	Yes
RPS26	(ribosomal protein S26)	0.901	85.651	96	18.834	No
RPS27	(ribosomal protein S27 (metal-lopanstimulin 1))	0.908	93.194	99	16.497	No
RPS27A	(ribosomal protein S27a)	0.964	93.965	98	17.508	Yes
RPS3	(ribosomal protein S3)	0.949	94.633	99	14.723	No
RPS3A	(ribosomal protein S3A)	0.962	96.092	99	12.972	No
RPS4X	(ribosomal protein S4; X-linked)	0.959	96.3	99	11.823	No
RPS5	(ribosomal protein S5)	0.95	93.301	99	18.893	Yes
RPS6	(ribosomal protein S6)	0.96	95.172	99	13.901	No
RPS7	(ribosomal protein S7)	0.955	94.847	98	13.464	No
RPS9	(ribosomal protein S9)	0.949	94.785	98	13.198	Yes
RRAGA	(Ras-related GTP binding A)	0.913	86.945	92	13.936	Yes
RUSC1	(RUN and SH3 domain containing 1)	0.925	79.342	82	12.926	No
SCAMP3	(secretory carrier membrane protein 3)	0.911	82.887	87	13.77	Yes

Table C.1: Cont'd

SDHA	(succinate dehydrogenase complex; subunit A; flavoprotein (Fp))	0.937	87.189	91	12.408	Yes
SDHC	(succinate dehydrogenase complex; subunit C; integral membrane protein; 15kDa)	0.922	81.433	85	12.87	No
SEC11L1	(SEC11 homolog A (S. cerevisiae))	0.943	88.658	93	13.734	No
SEC61G	(Sec61 gamma subunit)	0.902	87.012	92	13.907	Yes
SEPHS2	(selenophosphate synthetase 2)	0.901	78.587	80	12.537	No
SEPT2	(septin 2)	0.923	90.706	94.5	12.574	No
SEPW1	(selenoprotein W; 1)	0.924	88.251	92	13.179	No
SET	(SET translocation (myeloid leukemia-associated))	0.935	90.429	94.714	13.451	No
SFRS2	(splicing factor; arginine/serine-rich 2)	0.932	88.056	92.333	14.093	Yes
SFRS9	(splicing factor; arginine/serine-rich 9)	0.935	87.776	93.5	17.839	Yes
SHFM1	(split hand/foot malformation (ectrodactyly) type 1)	0.903	83.318	88	16.157	No
SIAHBP1	(fuse-binding protein-interacting repressor)	0.938	87.498	91	13.504	Yes
SLC25A3	(solute carrier family 25 (mitochondrial carrier; phosphate carrier); member 3)	0.959	93.23	98	17.605	Yes
SLC25A6	(solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator); member 6)	0.95	93.475	97	12.484	No

Table C.1: Cont'd

SNRPB	(small nuclear ribonucleoprotein polypeptides B and B1)	0.922	88.557	94	15.825	Yes
SNRPD2	(small nuclear ribonucleoprotein D2 polypeptide 16.5kDa)	0.952	93.075	96	12.388	Yes
SNRPG	(small nuclear ribonucleoprotein polypeptide G)	0.927	87.325	93	16.384	Yes
SNX3	(sorting nexin 3)	0.944	92.295	95	11.479	Yes
SOD1	(superoxide dismutase 1; soluble (amyotrophic lateral sclerosis 1 (adult)))	0.962	93.632	97	11.849	Yes
SPAG7	(sperm associated antigen 7)	0.956	85.947	90	16.78	Yes
SRP14	(signal recognition particle 14kDa (homologous Alu RNA binding protein))	0.961	92.734	97	18.239	Yes
SRP19	(signal recognition particle 19kDa)	0.913	82.874	87	15.954	No
SRP9	(signal recognition particle 9kDa)	0.921	90.706	95	13.766	No
SSR2	(signal sequence receptor; beta (translocon-associated protein beta))	0.959	90.43	95	14.483	Yes
SSR4	(signal sequence receptor; delta (translocon-associated protein delta))	0.946	92.481	95	11.267	No
ST13	(suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein))	0.909	80.986	87	16.01	No
SUMO3	(SMT3 suppressor of mif two 3 homolog 3 (S. cerevisiae))	0.905	85.098	91	17.719	Yes
SUPT5H	(suppressor of Ty 5 homolog (S. cerevisiae))	0.908	79.375	82	12.256	No
TAF10	(TAF10 RNA polymerase II; TATA box binding protein (TBP)-associated factor; 30kDa)	0.944	87.72	93	17.603	No
TALDO1	(transaldolase 1)	0.927	89.618	93	13.983	Yes

Table C.1: Cont'd

TAX1BP1	(Tax1 (human T-cell leukemia virus type I) binding protein 1)	0.913	87.383	90.5	12.606	No
TEGT	(testis enhanced gene transcript (BAX inhibitor 1))	0.946	91.11	95.5	15.957	Yes
TERF2IP	(telomeric repeat binding factor 2; interacting protein)	0.929	83.561	87	12.822	Yes
TMED2	(transmembrane emp24 domain trafficking protein 2)	0.913	86.102	93.667	21.14	No
TMEM147	(transmembrane protein 147)	0.949	89.124	93	12.748	No
TMEM66	(transmembrane protein 66)	0.944	89.777	93	12.26	No
TMSB10	(thymosin; beta 10)	0.951	95.04	98	13.06	Yes
TOMM20	(translocase of outer mitochondrial membrane 20 homolog (yeast))	0.908	87.775	92	14.269	No
TOMM34	(translocase of outer mitochondrial membrane 34)	0.924	80.269	82	10.876	No
TPT1	(tumor protein; translationally-controlled 1)	0.968	96.054	99	13.058	No
TRIM28	(tripartite motif-containing 28)	0.914	87.701	93	15.719	Yes
TUBB	(tubulin; beta)	0.964	92.963	96	10.651	Yes
TUBB2C	(tubulin; beta 2C)	0.953	93.172	97	11.478	No

Table C.1: Cont'd

TXNRD1	(thioredoxin reductase 1)	0.92	84.603	89	15.908	No
UBB	(ubiquitin B)	0.967	93.684	97	12.144	Yes
UBC	(ubiquitin C)	0.974	96.436	98.75	11.247	Yes
UBE1	(ubiquitin-activating enzyme E1 (A1S9T and BN75 temperature sensitivity complementing))	0.918	89.44	95	17.086	Yes
UBE2D3	(ubiquitin-conjugating enzyme E2D 3 (UBC4/5 homolog; yeast))	0.916	87.378	91.5	12.451	No
UQCRFS1	(ubiquinol-cytochrome c reductase; Rieske iron-sulfur polypeptide 1)	0.936	91.877	96	12.759	Yes
UQCRH	(ubiquinol-cytochrome c reductase hinge protein)	0.95	92.651	96	13.053	Yes
UQCRQ	(ubiquinol-cytochrome c reductase; complex III subunit VII; 9.5kDa)	0.949	93.693	97	12.807	No
USP11	(ubiquitin specific peptidase 11)	0.911	85.366	90	15.103	Yes
VCP	(valosin-containing protein)	0.94	85.904	88	12.01	No
VDAC2	(voltage-dependent anion channel 2)	0.959	91.582	96	12.5	No
VPS72	(vacuolar protein sorting 72 (S. cerevisiae))	0.929	81.995	86	13.959	No
WBSCR1	(Williams-Beuren syndrome chromosome region 1)	0.944	90.961	95	14.32	No
XBP1	(X-box binding protein 1)	0.916	85.836	90	14.23	Yes
XPO6	(exportin 6)	0.955	85.116	89	14.396	No
YWHAB	(tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein; beta polypeptide)	0.958	91.346	94.667	12.071	Yes

Table C.1: Cont'd

YWHAQ	(tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein; theta polypeptide)	0.958	92.683	96	11.7	Yes
ZNF289	(zinc finger protein 289; ID1 regulated)	0.929	82.425	86	11.827	No
ZNF384	(zinc finger protein 384)	0.908	80.674	83	11.565	Yes

APPENDIX D

STABILITY VALUES FOR 17 HOUSEKEEPING GENES

Table D.1: Stability values for 17 housekeeping genes

Gene Symbol	**Stability Value**							
	Norm Finder				**geNorm**			
	Over All	HCC	Breast	Colon	Over All	HCC	Breast	Colon
RPS10	0,215	0,165	0,355	0,100	0,000	0,000	0,000	0,000
RPL41	0,224	0,250	0,246	0,094	0,000	0,000	0,000	0,000
RPL7	0,217	0,164	0,245	0,171	0,268	0,220	0,253	0,276
RPS3A	0,191	0,198	0,095	0,057	0,339	0,296	0,378	0,128
RPL30	0,174	0,089	0,223	0,188	0,365	0,310	0,310	0,247
HBXIP	0,288	0,246	0,277	0,243	0,438	0,449	0,165	0,201
CFL1	0,257	0,246	0,157	0,298	0,474	0,420	0,347	0,456
RPS17	0,332	0,164	0,419	0,324	0,508	0,270	0,499	0,309
EEF2	0,368	0,319	0,255	0,598	0,552	0,525	0,428	0,651
ACTB	0,408	0,364	0,204	0,276	0,597	0,503	0,466	0,410
GAPDH	0,331	0,211	0,523	0,331	0,635	0,384	0,780	0,523
H2AFZ	0,418	0,284	0,548	0,210	0,675	0,482	0,823	0,170
AARS	0,513	0,507	0,618	0,241	0,721	0,581	0,661	0,490
TPT1	0,539	0,613	0,397	0,326	0,764	0,717	0,548	0,348
SOD1	0,531	0,635	0,525	0,396	0,802	0,652	0,720	0,592
RPN2	0,617	0,666	0,247	0,437	0,854	0,786	0,592	0,557
GSTO1	0,868	0,997	0,985	0,366	0,942	0,905	0,932	0,377

APPENDIX E

BI-K-BI CLUSTERING GENE-PAIR RULE EXAMPLE

Table E.1: The most observed gene-pair rule in the breast cancer datasets where p=70% and q=80%

Gene Pairs :	[FOXM1 (HIGH) - TPX2 (HIGH)]
Experiments : (*GDS_ GSM)*	[GDS1326_GSM41355], [GDS1326_GSM41356], [GDS1326_GSM41357], [GDS1326_GSM41352], [GDS1326_GSM41353], [GDS1326_GSM41354], [GDS1326_GSM41349], [GDS1326_GSM41350], [GDS1326_GSM41351], [GDS1329_GSM26883], [GDS1329_GSM26886], [GDS1329_GSM26887], [GDS1329_GSM26903], [GDS1329_GSM26910], [GDS1329_GSM26871], [GDS1329_GSM26880], [GDS1329_GSM26882], [GDS1329_GSM26884], [GDS1329_GSM26888], [GDS1329_GSM26889], [GDS1329_GSM26892], [GDS1329_GSM26893], [GDS1329_GSM26895], [GDS1329_GSM26898], [GDS1329_GSM26900], [GDS1329_GSM26902], [GDS1329_GSM26905], [GDS1329_GSM26906], [GDS1329_GSM26908],[GDS1329_GSM26912], [GDS1329_GSM26804], [GDS1329_GSM26867], [GDS1329_GSM26868], [GDS1329_GSM26870], [GDS1329_GSM26873], [GDS1329_GSM26875], [GDS1329_GSM26876], [GDS1329_GSM26879], [GDS1329_GSM26881], [GDS1329_GSM26890], [GDS1329_GSM26891], [GDS1329_GSM26894], [GDS1329_GSM26896], [GDS1329_GSM26897], [GDS1329_GSM26899], [GDS1329_GSM26901], [GDS1329_GSM26904], [GDS1329_GSM26907], [GDS1329_GSM26909], [GDS1329_GSM26911], [GDS1329_GSM26914],

Table E.1: Cont'd

	[GDS2250_GSM85494], [GDS2250_GSM85495], [GDS2250_GSM85496], [GDS2250_GSM85497], [GDS2250_GSM85498], [GDS2250_GSM85499], [GDS2250_GSM85500], [GDS2250_GSM85504], [GDS2250_GSM85505], [GDS2250_GSM85506], [GDS2250_GSM85509], [GDS2250_GSM85510], [GDS2250_GSM85511], [GDS2250_GSM85512], [GDS2250_GSM85491], [GDS2250_GSM85492], [GDS2250_GSM85473], [GDS2250_GSM85474], [GDS2250_GSM85475], [GDS2250_GSM85476], [GDS2250_GSM85477], [GDS2250_GSM85478], [GDS2250_GSM85479], [GDS2250_GSM85481], [GDS2250_GSM85482], [GDS2250_GSM85483], [GDS2250_GSM85484], [GDS2250_GSM85485], [GDS2250_GSM85486], [GDS2250_GSM85487], [GDS2250_GSM85488], [GDS2250_GSM85489], [GDS2250_GSM85490], [GDS360_GSM4901], [GDS360_GSM4902], [GDS360_GSM4904], [GDS360_GSM4905], [GDS360_GSM4906], [GDS360_GSM4910], [GDS360_GSM4913], [GDS360_GSM4916], [GDS360_GSM4918], [GDS360_GSM4922], [GDS360_GSM4924], [GDS360_GSM4903] , [GDS360_GSM4907], [GDS360_GSM4908], [GDS360_GSM4914], [GDS360_GSM4917], [GDS360_GSM4919], [GDS360_GSM4920], [GDS360_GSM4921], [GDS360_GSM4923], [GDS817_GSM21240], [GDS817_GSM21241], [GDS817_GSM21236], [GDS817_GSM21237] , [GDS817_GSM21238], [GDS817_GSM21239], [GDS820_GSM21246], [GDS820_GSM21247], [GDS820_GSM21242], [GDS820_GSM21243], [GDS820_GSM21244], [GDS820_GSM21245], [GDS881_GSM13097], [GDS881_GSM13098], [GDS881_GSM13099], [GDS881_GSM13138], [GDS881_GSM13139], [GDS881_GSM13140], [GDS881_GSM15900], [GDS881_GSM15901], [GDS881_GSM15902], [GDS881_GSM15903], [GDS881_GSM15904], [GDS881_GSM15905], [GDS881_GSM15906], [GDS881_GSM15907], [GDS881_GSM15908], [GDS881_GSM15909], [GDS881_GSM15910]

Table E.2: List of GSMs among Breast Cancer datasets where the rule [FOXM1 (HIGH) - TPX2 (HIGH)] is valid.

GDS	GSM	Detail
GDS1326	GSM41355	VALUE FOR GSM41355: ADLACZ+VEHICLE JM4; SRC: MDA-MB-231
GDS1326	GSM41356	VALUE FOR GSM41356: ADLACZ+VEHICLE JMTM1; SRC: MDA-MB-231
GDS1326	GSM41357	VALUE FOR GSM41357: ADLACZ+VEHICLE JMTM16; SRC: MDA-MB-231
GDS1326	GSM41352	VALUE FOR GSM41352: ADLACZ+E2-8 JM6; SRC: MDA-MB-231
GDS1326	GSM41353	VALUE FOR GSM41353: ADLACZ+E2-8 JMTM2; SRC: MDA-MB-231
GDS1326	GSM41354	VALUE FOR GSM41354: ADLACZ+E2-8 JMTM17; SRC: MDA-MB-231
GDS1326	GSM41349	VALUE FOR GSM41349: ADERALPHA+VEHICLE JM1; SRC: MDA-MB-231
GDS1326	GSM41350	VALUE FOR GSM41350: ADERALPHA+VEHICLE JMTM4; SRC: MDA-MB-231
GDS1326	GSM41351	VALUE FOR GSM41351: ADERALPHA+VEHICLE JMTM18; SRC: MDA-MB-231
GDS1326	GSM41346	VALUE FOR GSM41346: ADERALPHA+E2-8 JM3; SRC: MDA-MB-231
GDS1326	GSM41347	VALUE FOR GSM41347: ADERALPHA+E2-8 JMTM5; SRC: MDA-MB-231
GDS1329	GSM26878	VALUE FOR GSM26878: PF14 ENPNT2N1G2; SRC: BIOPSY
GDS1329	GSM26883	VALUE FOR GSM26883: PF19 EPPUT4N0GU; SRC: BIOPSY
GDS1329	GSM26886	VALUE FOR GSM26886: PF22 ENPNT2N1G2; SRC: BIOPSY
GDS1329	GSM26887	VALUE FOR GSM26887: PF23 ENPNT2N0G2; SRC: BIOPSY
GDS1329	GSM26903	VALUE FOR GSM26903: PF39 EUPUT4N0GU; SRC: BIOPSY
GDS1329	GSM26910	VALUE FOR GSM26910: PF46 ENPNT4N1G3; SRC: BIOPSY
GDS1329	GSM26871	VALUE FOR GSM26871: PF06 ENPNT2N0G3; SRC: BIOPSY

Table E.2: Cont'd

GDS1329	GSM26880	VALUE FOR GSM26880: PF16 ENPNT2N0G3; SRC: BIOPSY
GDS1329	GSM26882	VALUE FOR GSM26882: PF18 ENPNT2N1G3; SRC: BIOPSY
GDS1329	GSM26884	VALUE FOR GSM26884: PF20 ENPNT3N1G2; SRC: BIOPSY
GDS1329	GSM26888	VALUE FOR GSM26888: PF24 ENPNTIN0G3; SRC: BIOPSY
GDS1329	GSM26889	VALUE FOR GSM26889: PF25 ENPNT3N2G2; SRC: BIOPSY
GDS1329	GSM26892	VALUE FOR GSM26892: PF28 ENPNT2N1G3; SRC: BIOPSY
GDS1329	GSM26893	VALUE FOR GSM26893: PF29 ENPNT3N1G3; SRC: BIOPSY
GDS1329	GSM26895	VALUE FOR GSM26895: PF31 ENPNT2N0G3; SRC: BIOPSY
GDS1329	GSM26898	VALUE FOR GSM26898: PF34 ENPNT3N1G3; SRC: BIOPSY
GDS1329	GSM26900	VALUE FOR GSM26900: PF36 ENPNT2N0G2; SRC: BIOPSY
GDS1329	GSM26902	VALUE FOR GSM26902: PF38 ENPNT2N1G3; SRC: BIOPSY
GDS1329	GSM26905	VALUE FOR GSM26905: PF41 ENPNT3N1G2; SRC: BIOPSY
GDS1329	GSM26906	VALUE FOR GSM26906: PF42 ENPNT2N2G3; SRC: BIOPSY
GDS1329	GSM26908	VALUE FOR GSM26908: PF44 ENPNT3N0G3; SRC: BIOPSY
GDS1329	GSM26912	VALUE FOR GSM26912: PF48 ENPNT2N0G3; SRC: BIOPSY
GDS1329	GSM26804	VALUE FOR GSM26804: PF01 EPPPT2N0G2; SRC: BIOPSY
GDS1329	GSM26867	VALUE FOR GSM26867: PF02 EPPPT4N1G2; SRC: BIOPSY
GDS1329	GSM26868	VALUE FOR GSM26868: PF03 EPPPT3N0GU; SRC: BIOPSY
GDS1329	GSM26870	VALUE FOR GSM26870: PF05 EPPNT2N1G1; SRC: BIOPSY
GDS1329	GSM26873	VALUE FOR GSM26873: PF09 EPPPT3N1G2; SRC: BIOPSY
GDS1329	GSM26875	VALUE FOR GSM26875: PF11 EPPPT3N1G2; SRC: BIOPSY
GDS1329	GSM26876	VALUE FOR GSM26876: PF12 EPPPTIN0G2; SRC: BIOPSY
GDS1329	GSM26879	VALUE FOR GSM26879: PF15 EPPNTIN1G3; SRC: BIOPSY
GDS1329	GSM26881	VALUE FOR GSM26881: PF17 EPPNT2N1G3; SRC: BIOPSY
GDS1329	GSM26890	VALUE FOR GSM26890: PF26 ENPNT3N0G3; SRC: BIOPSY
GDS1329	GSM26891	VALUE FOR GSM26891: PF27 EPPNT4N1G2; SRC: BIOPSY
GDS1329	GSM26894	VALUE FOR GSM26894: PF30 EPPPT2N0G3; SRC: BIOPSY
GDS1329	GSM26896	VALUE FOR GSM26896: PF32 ENPNT3N1G2; SRC: BIOPSY
GDS1329	GSM26897	VALUE FOR GSM26897: PF33 EPPNTIN0G2; SRC: BIOPSY
GDS1329	GSM26899	VALUE FOR GSM26899: PF35 EPPPT2N1G3; SRC: BIOPSY
GDS1329	GSM26901	VALUE FOR GSM26901: PF37 EPPPT3N1G2; SRC: BIOPSY

Table E.2: Cont'd

GDS1329	GSM26904	VALUE FOR GSM26904: PF40 EPPNT4N0G2; SRC: BIOPSY
GDS1329	GSM26907	VALUE FOR GSM26907: PF43 EPPPT2N1G2; SRC: BIOPSY
GDS1329	GSM26909	VALUE FOR GSM26909: PF45 EPPPT4N0G2; SRC: BIOPSY
GDS1329	GSM26911	VALUE FOR GSM26911: PF47 EPPPT3N1G3; SRC: BIOPSY
GDS1329	GSM26914	VALUE FOR GSM26914: PF50 EPPNT3N1G3; SRC: BIOPSY
GDS2250	GSM85494	VALUE FOR GSM85494: T183 U133P2; SRC: T183
GDS2250	GSM85495	VALUE FOR GSM85495: T117 U133P2; SRC: T117
GDS2250	GSM85496	VALUE FOR GSM85496: T161 U133P2; SRC: T161
GDS2250	GSM85497	VALUE FOR GSM85497: T30 U133P2; SRC: T30
GDS2250	GSM85498	VALUE FOR GSM85498: T84 U133P2; SRC: T84
GDS2250	GSM85499	VALUE FOR GSM85499: T115 U133P2; SRC: T115
GDS2250	GSM85500	VALUE FOR GSM85500: T44 U133P2; SRC: T44
GDS2250	GSM85504	VALUE FOR GSM85504: T175 U133P2; SRC: T175
GDS2250	GSM85505	VALUE FOR GSM85505: T178 U133P2; SRC: T178
GDS2250	GSM85506	VALUE FOR GSM85506: T41 U133P2; SRC: T41
GDS2250	GSM85509	VALUE FOR GSM85509: T74 U133P2; SRC: T74
GDS2250	GSM85510	VALUE FOR GSM85510: T162 U133P2; SRC: T162
GDS2250	GSM85511	VALUE FOR GSM85511: T145 U133P2; SRC: T145
GDS2250	GSM85512	VALUE FOR GSM85512: T119 U133P2; SRC: T119
GDS2250	GSM85491	VALUE FOR GSM85491: T151 U133P2; SRC: T151
GDS2250	GSM85492	VALUE FOR GSM85492: T152 U133P2; SRC: T152
GDS2250	GSM85473	VALUE FOR GSM85473: T118 U133P2; SRC: T118
GDS2250	GSM85474	VALUE FOR GSM85474: T134 U133P2; SRC: T134
GDS2250	GSM85475	VALUE FOR GSM85475: T140 U133P2; SRC: T140
GDS2250	GSM85476	VALUE FOR GSM85476: T141 U133P2; SRC: T141
GDS2250	GSM85477	VALUE FOR GSM85477: T146 U133P2; SRC: T146
GDS2250	GSM85478	VALUE FOR GSM85478: T147 U133P2; SRC: T147
GDS2250	GSM85479	VALUE FOR GSM85479: T149 U133P2; SRC: T149
GDS2250	GSM85481	VALUE FOR GSM85481: T21 U133P2; SRC: T21
GDS2250	GSM85482	VALUE FOR GSM85482: T56 U133P2; SRC: T56
GDS2250	GSM85483	VALUE FOR GSM85483: T116 U133P2; SRC: T116

Table E.2: Cont'd

GDS2250	GSM85484	VALUE FOR GSM85484: T144 U133P2; SRC: T144
GDS2250	GSM85485	VALUE FOR GSM85485: T129 U133P2; SRC: T129
GDS2250	GSM85486	VALUE FOR GSM85486: T143 U133P2; SRC: T143
GDS2250	GSM85487	VALUE FOR GSM85487: T38 U133P2; SRC: T38
GDS2250	GSM85488	VALUE FOR GSM85488: T123 U133P2; SRC: T123
GDS2250	GSM85489	VALUE FOR GSM85489: T137 U133P2; SRC: T137
GDS2250	GSM85490	VALUE FOR GSM85490: T130 U133P2; SRC: T130
GDS360	GSM4901	VALUE FOR GSM4901: 44; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4902	VALUE FOR GSM4902: 51; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4904	VALUE FOR GSM4904: 113; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4905	VALUE FOR GSM4905: 118; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4906	VALUE FOR GSM4906: 136; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4910	VALUE FOR GSM4910: 358; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4913	VALUE FOR GSM4913: 377; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4916	VALUE FOR GSM4916: 432; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4918	VALUE FOR GSM4918: 438; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4922	VALUE FOR GSM4922: 555; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4924	VALUE FOR GSM4924: 562; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4903	VALUE FOR GSM4903: 71; SRC: HUMAN BREAST CANCER CORE BIOPSY

Table E.2: Cont'd

GDS360	GSM4907	VALUE FOR GSM4907: 142; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4908	VALUE FOR GSM4908: 273; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4914	VALUE FOR GSM4914: 413; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4917	VALUE FOR GSM4917: 437; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4919	VALUE FOR GSM4919: 447; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4920	VALUE FOR GSM4920: 458; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4921	VALUE FOR GSM4921: 492; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4923	VALUE FOR GSM4923: 558; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS817	GSM21240	VALUE FOR GSM21240: NORMAL BREAST EPITHELIUM CONTROL REPLICATE 1 95A; SRC: HUMAN MAMMARY EPITHELIAL CELLS
GDS817	GSM21241	VALUE FOR GSM21241: NORMAL BREAST EPITHELIUM CONTROL REPLICATE 2 95A; SRC: HUMAN MAMMARY EPITHELIAL CELLS
GDS817	GSM21236	VALUE FOR GSM21236: BREAST CANCER CELLS MDA-MB-436 REPLICATE 1 95A; SRC: MDA-MB436 BREAST CANCER CELL LINE
GDS817	GSM21237	VALUE FOR GSM21237: BREAST CANCER CELLS MDA-MB-436 REPLICATE 2 95A; SRC: MDA-MB436 BREAST CANCER CELL LINE
GDS817	GSM21238	VALUE FOR GSM21238: BREAST CANCER CELLS HCC1954 REPLICATE 1 95A; SRC: HCC1954 BREAST CANCER CELL LINE

Table E.2: Cont'd

GDS817	GSM21239	VALUE FOR GSM21239: BREAST CANCER CELLS HCC1954 REPLICATE 2 95A; SRC: HCC1954 BREAST CANCER CELL LINE
GDS820	GSM21246	VALUE FOR GSM21246: NORMAL BREAST EPITHELIUM CONTROL REPLICATE 1 133A; SRC: HUMAN MAMMARY EPITHELIAL CELLS
GDS820	GSM21247	VALUE FOR GSM21247: NORMAL BREAST EPITHELIUM CONTROL REPLICATE 2 133A; SRC: HUMAN MAMMARY EPITHELIAL CELLS
GDS820	GSM21242	VALUE FOR GSM21242: BREAST CANCER CELLS MDA-MB-436 REPLICATE 1 133A; SRC: MDA-MB436 BREAST CANCER CELL LINE
GDS820	GSM21243	VALUE FOR GSM21243: BREAST CANCER CELLS MDA-MB-436 REPLICATE 2 133A; SRC: MDA-MB436 BREAST CANCER CELL LINE
GDS820	GSM21244	VALUE FOR GSM21244: BREAST CANCER CELLS HCC1954 REPLICATE 1 133A; SRC: HCC1954 BREAST CANCER CELL LINE
GDS820	GSM21245	VALUE FOR GSM21245: BREAST CANCER CELLS HCC1954 REPLICATE 2 133A; SRC: HCC1954 BREAST CANCER CELL LINE
GDS881	GSM13097	VALUE FOR GSM13097: CONTROL A; SRC: MCF-7
GDS881	GSM13098	VALUE FOR GSM13098: CONTROL B; SRC: MCF-7
GDS881	GSM13099	VALUE FOR GSM13099: E2 8H A; SRC: MCF-7
GDS881	GSM13138	VALUE FOR GSM13138: E2 8H B; SRC: MCF-7
GDS881	GSM13139	VALUE FOR GSM13139: E2 48H A; SRC: MCF-7
GDS881	GSM13140	VALUE FOR GSM13140: E2 48H B; SRC: MCF-7
GDS881	GSM15900	VALUE FOR GSM15900: E2+ICI 8H A; SRC: MCF-7
GDS881	GSM15901	VALUE FOR GSM15901: E2+ICI 8H B; SRC: MCF-7
GDS881	GSM15902	VALUE FOR GSM15902: E2+ICI 48H A; SRC: MCF-7
GDS881	GSM15903	VALUE FOR GSM15903: E2+ICI 48H B; SRC: MCF-7
GDS881	GSM15904	VALUE FOR GSM15904: E2+RAL 8H A; SRC: MCF-7
GDS881	GSM15905	VALUE FOR GSM15905: E2+RAL 8H B; SRC: MCF-7
GDS881	GSM15906	VALUE FOR GSM15906: E2+RAL 48H A; SRC: MCF-7

Table E.2: Cont'd

GDS881	GSM15907	VALUE FOR GSM15907: E2+RAL 48H B; SRC: MCF-7
GDS881	GSM15908	VALUE FOR GSM15908: E2+TOT 8H A; SRC: MCF-7
GDS881	GSM15909	VALUE FOR GSM15909: E2+TOT 8H B; SRC: MCF-7
GDS881	GSM15910	VALUE FOR GSM15910: E2+TOT 48H A; SRC: MCF-7

Table E.3: List of GSMs among Breast Cancer datasets where the rule [FOXM1 (HIGH) - TPX2 (HIGH)] is not valid.

GDS	**GSM**	**Detail**
GDS1250	GSM45657	VALUE FOR GSM45657: ADH1; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45658	VALUE FOR GSM45658: ADH2; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45659	VALUE FOR GSM45659: ADH3; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45660	VALUE FOR GSM45660: ADH4; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45661	VALUE FOR GSM45661: ADHC1; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45662	VALUE FOR GSM45662: ADHC2; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45663	VALUE FOR GSM45663: ADHC3; SRC: BREAST PRECANCEROUS TISSUE
GDS1250	GSM45664	VALUE FOR GSM45664: ADHC4; SRC: BREAST PRECANCEROUS TISSUE
GDS1326	**GSM41348**	**VALUE FOR GSM41348: ADERALPHA+E2-8 JMTM19; SRC: MDA-MB-231**
GDS1329	**GSM26869**	**VALUE FOR GSM26869: PF04 EPPPT2N1G1; SRC: BIOPSY**

Table E.3: Cont'd

GDS1329	GSM26872	VALUE FOR GSM26872: PF07 EPPPT2N0G2; SRC: BIOPSY
GDS1329	GSM26874	VALUE FOR GSM26874: PF10 EPPPT4N1G2; SRC: BIOPSY
GDS1329	GSM26877	VALUE FOR GSM26877: PF13 EPPPT2N1G1; SRC: BIOPSY
GDS1329	GSM26885	VALUE FOR GSM26885: PF21 EPPPT2N1G2; SRC: BIOPSY
GDS1329	GSM26913	VALUE FOR GSM26913: PF49 EPPPT2N1G1; SRC: BIOPSY
GDS2250	GSM85513	VALUE FOR GSM85513: NB42 U133P2; SRC: NB42
GDS2250	GSM85514	VALUE FOR GSM85514: NB58 U133P2; SRC: NB58
GDS2250	GSM85515	VALUE FOR GSM85515: NB60 U133P2; SRC: NB60
GDS2250	GSM85516	VALUE FOR GSM85516: NB64 U133P2; SRC: NB64
GDS2250	GSM85517	VALUE FOR GSM85517: NB69 U133P2; SRC: NB69
GDS2250	GSM85518	VALUE FOR GSM85518: NB83 U133P2; SRC: NB83
GDS2250	GSM85519	VALUE FOR GSM85519: NB87 U133P2; SRC: NB87
GDS2250	GSM85493	VALUE FOR GSM85493: T37 U133P2; SRC: T37
GDS2250	GSM85501	VALUE FOR GSM85501: T81 U133P2; SRC: T81
GDS2250	GSM85502	VALUE FOR GSM85502: T50 U133P2; SRC: T50
GDS2250	GSM85503	VALUE FOR GSM85503: T4 U133P2; SRC: T4
GDS2250	GSM85507	VALUE FOR GSM85507: T73 U133P2; SRC: T73
GDS2250	GSM85508	VALUE FOR GSM85508: T92 U133P2; SRC: T92
GDS2250	GSM85480	VALUE FOR GSM85480: T133 U133P2; SRC: T133
GDS360	GSM4909	VALUE FOR GSM4909: 356; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4911	VALUE FOR GSM4911: 359; SRC: HUMAN BREAST CANCER CORE BIOPSY
GDS360	GSM4912	VALUE FOR GSM4912: 370; SRC: HUMAN BREAST CANCER CORE BIOPSY

Table E.3: Cont'd

GDS360	**GSM4915**	**VALUE FOR GSM4915: 425; SRC: HUMAN BREAST CANCER CORE BIOPSY**
GDS823	GSM21252	VALUE FOR GSM21252: NORMAL BREAST EPITHELIUM CONTROL REPLICATE 1 133B; SRC: HUMAN MAMMARY EPITHELIAL CELLS
GDS823	GSM21253	VALUE FOR GSM21253: NORMAL BREAST EPITHELIUM CONTROL REPLICATE 2 133B; SRC: HUMAN MAMMARY EPITHELIAL CELLS
GDS823	GSM21248	VALUE FOR GSM21248: BREAST CANCER CELLS MDA-MB-436 REPLICATE 1 133B; SRC: MDA-MB436 BREAST CANCER CELL LINE
GDS823	GSM21249	VALUE FOR GSM21249: BREAST CANCER CELLS MDA-MB-436 REPLICATE 2 133B; SRC: MDA-MB436 BREAST CANCER CELL LINE
GDS823	GSM21250	VALUE FOR GSM21250: BREAST CANCER CELLS HCC1954 REPLICATE 1 133B; SRC: HCC1954 BREAST CANCER CELL LINE
GDS823	GSM21251	VALUE FOR GSM21251: BREAST CANCER CELLS HCC1954 REPLICATE 2 133B; SRC: HCC1954 BREAST CANCER CELL LINE
GDS881	GSM15911	VALUE FOR GSM15911: E2+TOT 48H B; SRC: MCF-7
GDS901	GSM16943	VALUE FOR GSM16943: WILD TYPE CONTROL SET A; SRC: MDA-MB-231+ERALPHA-WT
GDS901	GSM18491	VALUE FOR GSM18491: WT CONTROL SAMPLE B; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18492	VALUE FOR GSM18492: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING WILD-TYPE ERA 1 HR TREATMENT SAMPLE A; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18493	VALUE FOR GSM18493: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING WILD-TYPE ERA 1 HR TREATMENT SAMPLE B; SRC: MDA-MB-231ER+ BREAST CANCER CELLS

Table E.3: Cont'd

GDS901	GSM18494	VALUE FOR GSM18494: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING WILD-TYPE ERA 2 HR TREATMENT SAMPLE A; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18495	VALUE FOR GSM18495: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING WILD-TYPE ERA 2 HR TREATMENT SAMPLE B; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18496	VALUE FOR GSM18496: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING MUTANT ERA LQ CONTROL SAMPLE A; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18497	VALUE FOR GSM18497: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING MUTANT ERA LQ SAMPLE B; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18498	VALUE FOR GSM18498: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING MUTANT ERA LQ 1 HR TREATMENT OF 10 NM E2 SAMPLE A; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18499	VALUE FOR GSM18499: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING MUTANT ERA LQ 1 HR TREATMENT OF 10 NM E2 SAMPLE B; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18500	VALUE FOR GSM18500: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING MUTANT ERA LQ 2 HR TREATMENT OF 10 NM E2 SAMPLE A; SRC: MDA-MB-231ER+ BREAST CANCER CELLS
GDS901	GSM18501	VALUE FOR GSM18501: MDA-MB-231 BREAST CANCER CELLS STABLY EXPRESSING MUTANT ERA LQ 2 HR TREATMENT OF 10 NM E2 SAMPLE B; SRC: MDA-MB-231ER+ BREAST CANCER CELLS

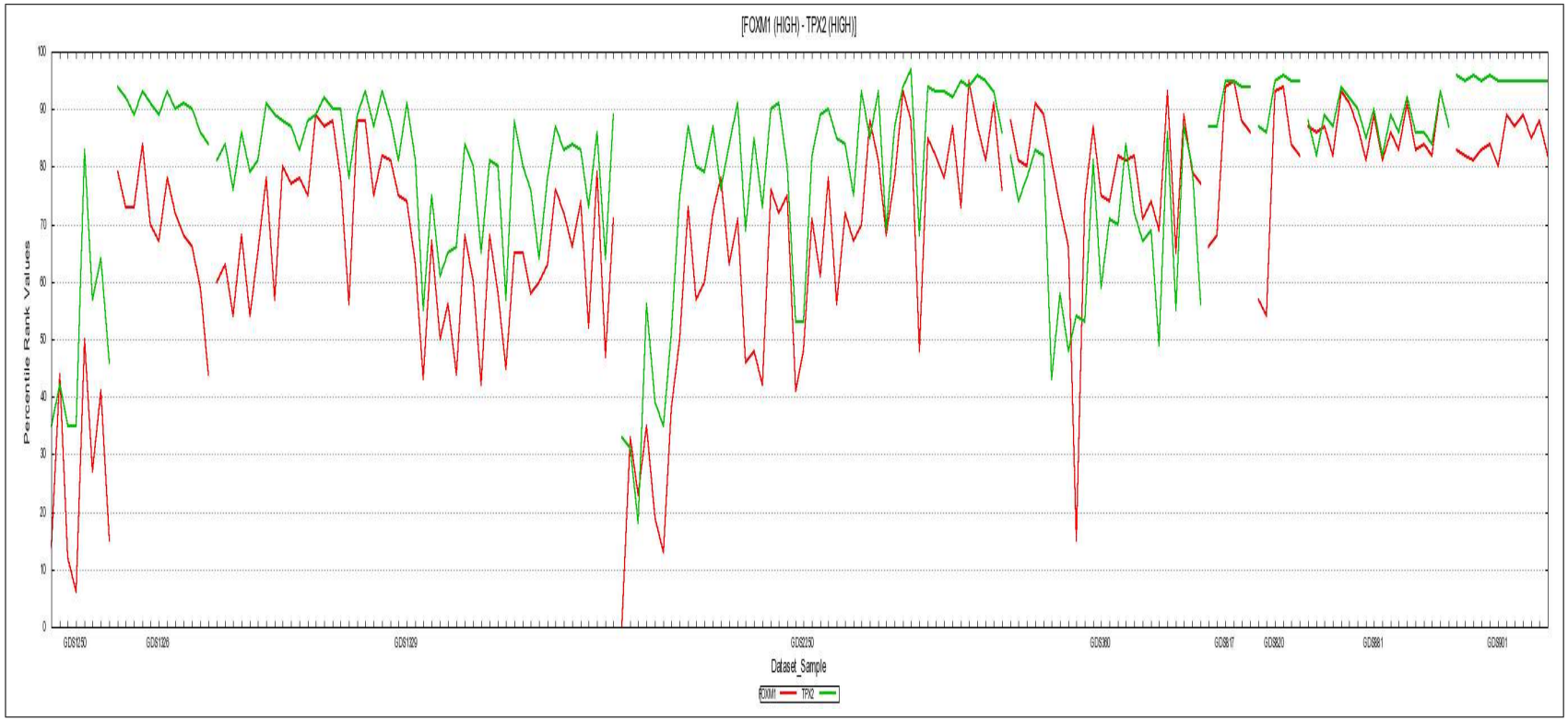

Figure E.1: *Graph of rank values of genes of the GSMs in the rule. Curve with green color presents gene with symbol TPX2 while curve with red color represent gene with gene symbol FOXM1. The y axis indicated the rank value of the gene at the corresponding GSM in the rule.*

VITA

Levent Çarkacıoğlu was born in 1977 in Afyon, Turkey. He graduated from Department of Computer Engineering, METU in 1999 and took his Masters degree form the same department in 2002 with his thesis "Developing ebXML Registry/Repository Mechanism". From 1999 to 2000, he worked as a software engineer in Defense Technologies and Engineering (STM A.Ş.). He is working as a computer engineer in the Information Technologies Division of Central Bank of The Republic of Turkey (TCMB) since 2000. His primary research interests include issues related with data mining, database management systems, large scale databases, bioinformatics and J2EE (Java) frameworks. Some of his selected publications include:

Journal:

- Carkacioglu, L., Cetin-Atalay, R., Konu, O., Atalay, V., and Can, T. Bi-k-Bi Clustering: Mining Large Scale Gene Expression Data Using Two-Level Biclustering, *Int. J. Data Mining and Bioinformatics*, 2009. (*Appear)*

Poster:

- Carkacioglu, L., Can, T., Konu, O., Atalay, V. and Cetin-Atalay, R. Expression Pattern Analysis of Housekeeping Genes Across Large Number of Microarray Experiments, *5th European Conference on Computational Biology (ECCB)*, Eilat, Israel, September, 2006.