MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Development of Improved Cancer Classification Method by Integrating Metadata in Microarray Data Analysis

By

Md. Ahsan Habib

Supervised by

Prof. Dr. M. A. Mottalib
Head, CSE Department
Islamic University of Technology (IUT)

Computer Science and Engineering (CSE) Department
Islamic University of Technology (IUT)
Organisation of Islamic Cooperation (OIC)
Gazipur, Bangladesh

September 2012

Development of Improved Cancer Classification Method by Integrating Metadata in Microarray Data Analysis

By
Md. Ahsan Habib
St. ID # 084602

Supervised by

Prof. Dr. M. A. Mottalib
Head, CSE Department
Islamic University of Technology (IUT)

A thesis submitted to Computer Science and Engineering Department in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering

Computer Science and Engineering (CSE) Department
Islamic University of Technology (IUT)
Organisation of Islamic Cooperation (OIC)
Gazipur, Bangladesh

September 2012

RECOMMENDATION OF THE BOARD OF EXAMINERS

The thesis titled "**Development of Improved Cancer Classification Method by Integrating Metadata in Microarray Data Analysis.**" submitted by Md. Ahsan Habib (Student No. 084602) of academic year 2011-2012 has been found as satisfactory and accepted as partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering on September 2012.

1. ---------- Chairman

Prof. Dr. M. A. Mottalib **(Supervisor)**
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
Board Bazar, Gazipur-1704, Bangladesh

2. ---------- Member

Dr. Muhammad Mahbub Alam
Associate Professor
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
Board Bazar, Gazipur-1704, Bangladesh

3. ---------- Member

Dr. Md. Kamrul Hasan
Assistant Professor
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
Board Bazar, Gazipur-1704, Bangladesh

4. ---------- External Member

Prof. Dr. Chowdhury Mofizur Rahman
Pro. Vice-Chancellor
United International University (UIU)
Dhanmondi, Dhaka, Bangladesh

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

DECLARATION OF CANDIDATE

This is to certify that the work presented in this thesis is the outcome of the analysis and investigation carried out by the candidate under the supervision of **Prof. Dr. M. A. Mottalib** in Computer Science and Engineering (CSE) Department, IUT, Gazipur, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references are given.

_____

Md. Ahsan Habib

Student No: 084602

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

DEDICATION

My beloved parents - Dr. Md. Aman Ullah & Tahmina Khatum

My parents in law - Md. Nazrul Islam & Zohura Begum

My family - wife Naima Islam & son Nayeem Ul Islam and

My sisters - Rajia, Tania, Sonia

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

ACKNOWLEDGEMENTS

I must sense grateful to the Almighty Allah to complete the dissertation. At the outset, I would like to express gratitude to my supervisor ***Professor Dr. M. A. Mottalib***, who has supported my plans to continue this work as an M.Sc. Engineering research in bioinformatics field. In addition, his continuous support, valuable advices, encouragements, patience and enthusiasm guided me for completion of the research successfully. Moreover, I would like to express my deep appreciation to ***Hawlader Abdullah Al Mamun*** of the Department who have influenced me to carry out the thesis in bioinformatics field. I also wish to take opportunity to articulate my sincerest and heartiest thanks to ***Dr. Muhammad Mahbub Alam, Tareque Mohmud Chowdhury, Dr. Md. Kamrul Hasan, Dr. Md. Hasanul Kabir, Md. Ali-Al-Mamun, Hasan Mahmud, Md. Sakhawat Hossen, Md. Safiur Rahman Mahdi, Md. Abid Hasan, Faisal Ahmed, Shaikh Jeeshan Kabeer*** and all other teachers of CSE Department for their inspiration and co-operation.

I would like to acknowledge to ***S. M. Mahbubur Rashid*** of Dhaka University and ***Abdullah-Al-Emran*** of Mawlana Bhashani Science & Technology University (MBSTU) to have their patience and time to share modern and updated knowledge in Genetics Engineering and Molecular Biology. I am highly grateful to the authority and my colleagues of MBSTU to cooperate and grant leave for completing the M.Sc program.

I would like to thank honorable examiners of the thesis committee for valuable comments and criticism that helped to improve the manuscript.

I thank all staffs of CSE Department, friends of IUT and family members for their unconditional support and encouragement. The work would never been completed without the consistent support of them. I wish to express my gratitude to IUT for providing an excellent environment for research.

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

ABSTRACT

Microarray experiments provide a high throughput to measure expressions of thousands of genes simultaneously. A systematic and computational analysis of this vast amount of data provides understanding and insight into many aspects of biological processes like Single Nucleotide Polymorphism (SNP), genetic disorders, cancer identification. Expression analysis of DNA (Deoxyribonucleic Acid) of microarray experiments is far from straightforward from statistical point of view. Dimensionality problem of microarray data, identifying significant and informative genes or DNA sequences are the prime challenges for cancer classification. Therefore, the aim of the thesis in cancer classification is to integrate metadata for optimal subset of genes that are useful for expert and embedded system design for different types of cancer classification.

In this thesis, different filtering and classification algorithms will be compared on different set of data to conclude which will be efficient and effective method for cancer classification from microarray data. For the improvement of cancer classification performance and biological validation of optimal subset of genes, metadata ranking will be evaluated and integrated for cancer classification. The method will achieve better performance based on gene-independent covariance, trustworthy gene feature ranking and metadata ranking factors. The performance of the proposed method will be evaluated by different filtering to overcome the dimensionality problem, minimum number of genes used in classification techniques on publicly available benchmark dataset of ALL, brain, breast, kidney, lung, prostate for intra-cancer and inter-cancer classification.

TABLE OF CONTENTS

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

LIST OF FIGURES

LIST OF TABLES

ABBYY PDF Transformer 2.0
Click here to buy
www.ABBYY.com

1. CHAPTER 1

INTRODUCTION

1.1 Background

DNA microarray technology, introduced in 1995, allows the measurement of thousands of gene expression values at a time, providing insight into the global gene expression patterns of cells being studied [1-3]. Despite the need for further technological developments with microarray [4], the approach remains powerful for studying the myriad of transcription-related pathways involved in cellular growth, differentiation, and transformation in various organisms and genome-wide ideal approaches to molecular cancer classification.

The field of cancer classification with microarray data leads to a lot of research activity in recent years. Owing to this effort, microarray data analysis and gene expression profiling are being used as more efficient techniques in clinical practice. In cancer research, microarrays are often used to support exact phenotyping in early stages of the disease, which potentially allows for tailored treatment and better cure rates. But, existing classification techniques suffer from high dimensionality problem of microarray data.

A major challenge with microarray research in cancer classification is how to find informative genes that can be used for effective discriminating variables in relation to different conditions, such as classifying healthy and diseased tissue samples. The amount of relevant genes is typically small, as “the majority of the active cellular mRNA is not affected by the biological differences” [5]. In the articles [6-7], the approach to this problem of classification in high dimension, encountered with microarray data, is to first apply dimension reduction techniques. After dimension reduction, standard classification/prediction tools such as Linear discernment analysis or logistic discrimination (LD) can reduce subspaces. The information retained plays an important role in the subsequent prediction. Clustering and classification are extensively studied problems in statistics and machine learning domain. Many algorithms, such as decision tree, linear discriminant analysis, neural network, and the Bayesian network have been

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

proposed and widely applied in practical problems. Recently years, researchers have paid attention to cancer clustering and classification using gene expression data.

Gene classification methods can be generally classified into two major groups: filter and wrapper methods [8]. Filter method examines the intrinsic characteristics of genes as the measuring criterion, while wrapper method evaluates genes based on the performance of an induction algorithm usually involving a classifier. Filter method is more popular than wrapper method in gene selection area, because it can generally achieve satisfactory performance with much less computational cost. But these methods cannot find optimal number of genes for classification that may be used for implementation of gene based embedded medical diagnosis system for cancer patients.

1.2 The Significance of the Problem

The monitoring of gene activity (or expression) from thousands of genes in parallel at the same time could lead to identify different types of diseases. Although the microarray (MA) technologies are very powerful, their complexity and selecting genes make the observed data very difficult to integrate across experiments. In addition, existing gene selection, classification and clustering algorithms do not consider whether the genes are biologically validated or not. But it is necessary to integrate publicly available datasets and research papers for extracting genomic relevant information and validating biological hypotheses or inferring some significance of genes for further research.

To diagnosis cancer or critical diseases, the exploitation of different tests is still hampered due to lack of appropriate knowledge of a pathologist. Though microarray technology has provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment, the vast amount of raw gene expression data leads to statistical and analytical challenges including the classification of the dataset of patients into correct cancer classes. For this reason, it is to identify the differentially expressed genes responsible for different cancers that would be useful for embedded system design to identify and predict cancer class.

1.3 Problem Statement

Modeling an efficient method for cancer classification and selecting optimal number of significant genes from microarray data suffers for the scarcity of real life and real time samples. Besides, the biological validation of the genes through experiment is costly and time consuming. While very effective methods for binary classification (i.e. classification into two classes) are known, the methods do not necessarily perform as well in the multi-class case.

Microarray data suffer for high directionality problems of gene feature for efficient cancer classifications which lead space and time complexity in microarray data analysis. Moreover, it is also noticed that selecting a set of proper genes can significantly reduce the inconsistency of microarray data experiment. Obviously, it will be more interesting to find out a set of genes that enable a biologically validated good classification performance over different subsets of patients in the expert embedded system.

Many challenges in microarray need to be addressed before new knowledge about gene expression can be revealed. Some of the problems are: a) Bias and confounding Problem: which occurred during study, design phase of microarray and can lead to erroneous conclusion. Technical factors, such as differences in physical, batch of reagents used and various levels of skill in technician could possibly cause bias. Confounding on the other hand, take place when another factors distorts the true relationship between the study variables of interest. b) Cross-platform comparisons of gene expression studies are difficult to conduct when microarrays were constructed using different standards. Thus, the results may not be reproduced. To deal with this problem, it is to develop to improve reproducibility, sensitivity and robustness in gene expression analysis. c) Microarray data is high dimensional data characterized by thousand of genes in few sample sizes, which cause significant problems such as irrelevant and noise genes, complexity in constructing classifiers, and multiple missing gene expression values due to improper scanning. Moreover, most of studies that applied microarray data are suffered from data over fitting which requires additional validation. d) Mislabeled data or questioned tissues result by experts also another types of drawback that could decrease the accuracy of experimental results and led to imprecise conclusion about gene expression patterns. e) Biological

relevancy result is another integral criterion that should be taken into account in analyzing microarray data rather than only focusing on accuracy of cancer classification. Although there is no doubt gaining high accuracy classification results are important in microarray data analysis, but revealing the biological information during the process of cancer classification is also essential. For instance determination of genes that are under expressed or over expressed in cancerous cells could assists domain experts in designing and planning more appropriate treatments for cancer patients. Therefore, most of domain experts are interested in classifiers that not only produce high classification accuracy but also reveal important biological information and identify minimal number of gene sets.

1.4 Thesis Objectives

The current era of bioinformatics research through microarray data analysis trends to identify risk factors of diseases, to adapt long-term countermeasures, chronic diseases or cancers where 4Ps (Predictive, Personalized, Preemptive and Participatory) of medicine rule is considered. In near future, treatments must be tailored in order to take into account the characteristics of individual patients. Shifting the focus of medicine from the current doctor centric, curative paradigm to preventing diseases will require the active involvement of patients. With the advent of personalized medicine, biomarkers, including genetic markers, will be tested for each patient in order to diagnose specific forms of diseases, predict disease progression and patient outcome, and propose the best therapeutic options. The study mainly focuses on the reduction of high dimensional microarray data for informative gene selection. The optimal set of genes will be will be used for improved cancer classification technique.

The thesis proposes an improved cancer classification method based on the integration of metadata in microarray data analysis as well as addresses the following criteria:

- To work on microarray gene feature selection and ranking with the help of metadata useful for cancer classification.
- To find out the optimal number of gene features from microarray data and to verify the model by checking the success and error rate for the prediction of cancers.

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

- To develop an expert and embedded system as an advanced bioinformatics tool to diagnosis intra-cancers and inter cancers.
- To conclude which classification methods are useful for microarray data analysis using metadata of microarray databases.

1.5 Thesis Contributions

The main contributions of the thesis are summarized follow:

- Metadata based gene feature ranking technique has been achieved to classify different types of cancers by integrating pubmed data.
- An efficient feature reduction technique has been obtained for high dimensional microarray data.
- A comparison with different classification techniques has made with optimal features to show which method is appropriate for microarray data.
- To sketch the relationships of the selected genes with different type of cancers and discover the best gene responsible for cancer .

1.6 Thesis Outline

The thesis is organised in six chapters. ***Chapter 1*** presents an introduction to the thesis outlining the problem and objectives. ***Chapter 2*** includes the preliminaries on Bioinformatics. ***Chapter 3*** reviews the related literature on gene feature selection and microarray data processing for classification methods. ***Chapter 4*** discusses the detailed implementation of Cancer Classification Method by integrating Metadata. ***Chapter 5*** expands experimental setup and performance evaluation of proposed method. Finally, ***Chapter 6*** draws conclusion with a note of future scope of research in this area.

ABBYY PDF Transformer 2.0
Click here to buy
www.ABBYY.com

2. CHAPTER 2

PRELIMINARIES

This chapter provides a basic understanding of Bioinformatics and microarray experiments for readers who are not familiar with molecular biology. Further reading is encouraged as an understanding of the underlying biology in the thesis. For a more detailed description of genes and genetic analysis see [9-14]. Readers with a good knowledge of biology and microarray technology are advised to skip directly to *Chapter 3*.

2.1 Terminologies

Bioinformatics: Bioinformatics is the application of computer science and information technology to the field of biology and medicine. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, data mining, image processing, modeling and simulation, signal processing, discrete mathematics, control and system theory, circuit theory, and statistics. Bioinformatics generates new knowledge as well as the computational tools to create that knowledge.

Cell: A *cell* is the minimal unit of life. There are a multitude of specific chemical transformations that not only provide the energy needed by a cell, but also coordinate all of the events and activities within that cell [9].

Central Dogma of Molecular Biology: The basics of molecular biology has been summarized in a concept called the Central Dogma of Molecular Biology [10]. DNA molecules contain biological information coded in an alphabet of four letters, A (Adenosine), T (Thymine), C (Cytosine), G (Guanine). The succession of these letters is referred as a sequence of DNA that constitutes the complete genetic information defining the structure and function of an organism. Proteins can be viewed as effectors of the genetic information contained in DNA coding sequences. They are formed using the genetic code of the DNA to convert the information contained in the 4-letter alphabet into a new alphabet of 20 amino acids. Despite an apparent simplicity of this translation

procedure, the conversion of the DNA-based information requires two steps in eucariotyc cells since the genetic material in the nucleus is physically separated from the site of protein synthesis in the cytoplasm of the cell. Transcription constitutes the intermediate step, where a DNA segment that constitutes a gene is read and transcribed into a single stranded molecule of RNA (the 4 letter alphabet remains with the replacement of Thymine molecules by Uracyle molecules). RNAs that contain information to be translated into proteins are called messenger RNAs, since they constitute the physical vector that carry the genetic information from the nucleus to the cytoplasm where it is translated into proteins via molecules called ribosomes.

DNA (Deoxyribonucleic Acid): DNA is a very stable molecule that forms the "blueprint" [11] of an organism. Deoxyribonucleic Acid is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The DNA segments carrying this genetic information are called genes. Likewise, other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. Along with RNA and proteins, DNA is one of the three major macromolecules that are essential for all known forms of life.

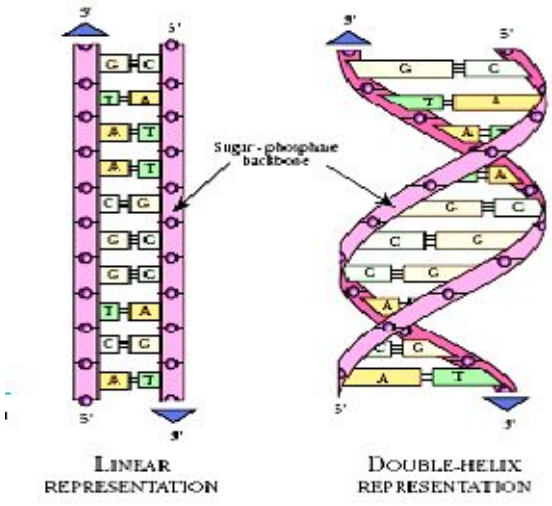

Figure 2.1 : Representations of double strand DNA

DNA is the central data repository of the cell. It is compound of two parallel strands. Each strand consists of four different types of molecules, which are called nucleotides. The four types of nucleotides are marked as: A (Adenine), C (Cytosine), G (Guanine) and T

(Thymine). Thus, each strand is a text composed from 4 letters. Nucleotides tend to bond in pairs. T nucleotide bonds with A nucleotide while C nucleotide bonds with G. The double-helix of the DNA is constructed of two complementary strands.

RNA (Ribonucleic Acid): Ribonucleic acid is part of a group of molecules known as the nucleic acids, which are one of the four major macromolecules (along with lipids, carbohydrates and proteins) essential for all known forms of life. Like DNA, RNA[11] is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase, a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. All cellular organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

Gene: Genes are the units of the DNA sequence that control the identifiable hereditary traits of an organism. A *gene*[9] can be defined as a segment of DNA that specifies a functional RNA. It is a name given to some stretches of DNA and RNA that code for a polypeptide or for an RNA chain that has a function in the organism. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring, although some organelles (e.g. mitochondria) are self-replicating and are not coded for by the organism's DNA. All organisms have many genes corresponding to various biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type or increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life.

Genome: In modern molecular biology and genetics, the genome [11] is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA/RNA. The human (Homo sapiens) genome is stored on 23 chromosome pairs in the cell nucleus and in the small mitochondrial DNA. The human genome occupies a total of just over three billion DNA base pairs. The haploid human genome contains just over 20,000 protein-coding genes. In fact, only about 1.5% of the genome codes for proteins,

while the rest consists of non-coding RNA genes, regulatory sequences, introns, and noncoding DNA (once known as junk DNA). The genome defines the genetic construction of an organism or cell, or the *genotype.* The *phenotype,* on the other hand, is the total set of characteristics displayed by an organism under a particular set of environmental factors. The outward appearance of an organism (phenotype) may or may not directly reflect the genes that are present (genotype).

Proteins: Proteins [11] are biochemical compounds consisting of one or more polypeptides typically folded into a globular or fibrous form, facilitating a biological function. Proteins can be viewed as effectors of the genetic information contained in DNA coding sequences.

Cancer: Cancer [13] known medically as a malignant neoplasm, is a broad group of various diseases, all involving unregulated cell growth. In cancer, cells divide and grow uncontrollably, forming malignant tumors, and invade nearby parts of the body. The cancer may also spread to more distant parts of the body through the lymphatic system or bloodstream. Determining what causes cancer is complex. Many things are known to increase the risk of cancer. Approximately five to ten percent of cancers are entirely hereditary.

2.2 DNA Microarray

DNA microarray (also known as DNA chip or biochip) [1-3, 14] is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Microarray technologies open exciting perspectives for gene expression profiling studies. Comparisons of the levels of RNA molecules can be used to decipher the thousands of processes going on simultaneously in living organisms where healthy and diseased cells can yield vital information on the causes of diseases.

Microarrays offer an efficient method [3, 14] of gathering data that can be used to determine the expression pattern of thousands of genes. The mRNA expression pattern from different tissues in normal and diseases states could reveal which genes and environmental conditions can lead to disease. The experimental steps of typical microarray began with extraction of mRNA from a tissues sample or probe. The mRNA is then labeled with fluorescent nucleotides, eventually yielding fluorescent (typically red) cDNA. The sample later is incubated with similarly processed cDNA reference (typically green). The labeled probe and reference are then mixed and applied to the surface of DNA microarrays, allowing fluorescent sequences in the probereference mix to attach to the cDNA adherent to the glass slide. The attraction of labeled cDNA from the probe and reference for a particular spot on microarray depends on the extent to which the sequences in the mix (probe-reference) complement the DNA affixed to the slide.

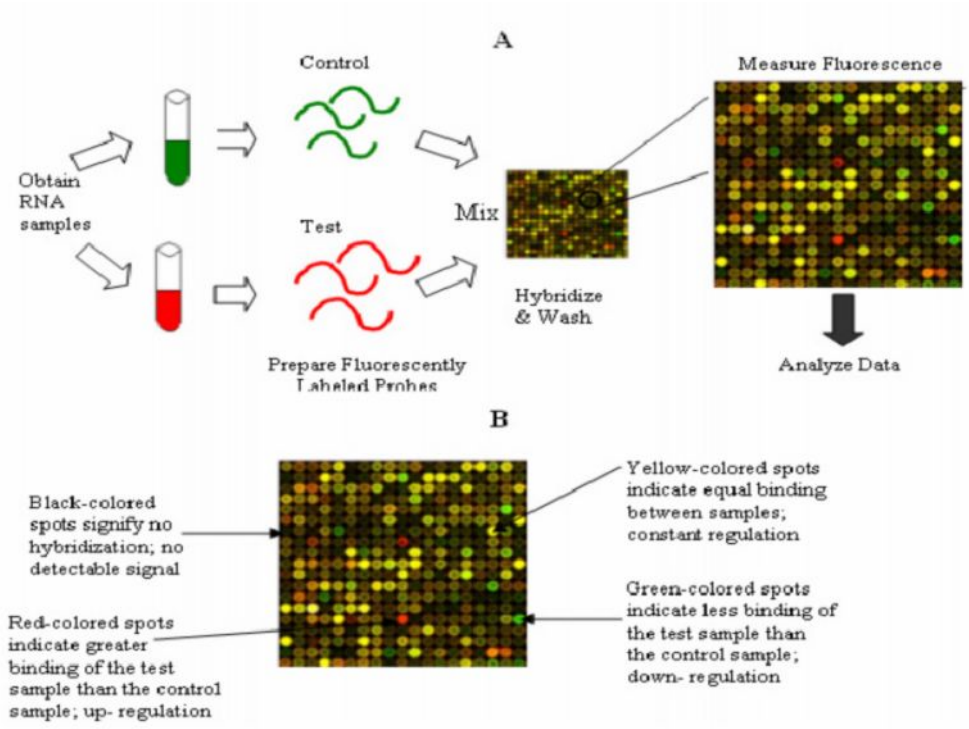

Figure 2.2 : Microarray experiment

A perfect compliment, in which a nucleotide sequence on a strand of cDNA exactly matches a DNA sequence affixed to the slide, is known as hybridization. Hybridization is the key element in microarray technology. The populated microarray is then excited by a

laser and the consequential fluorescent at each spot in the microarray is measured. If neither the probe nor the reference samples hybridize with the gene spotted on the slide, the spot will appear in the black color. However, if hybridization is predominantly with the probe, the spot will be in red (Cy5). Conversely, if hybridization is primarily between the reference and DNA affixed to the slide, the spot will fluoresce green (Cy3). The spot can also incandescent yellow, when cDNA from probe and reference samples hybridize equally at a given spot, indicating that they share the same number of complementary nucleotides in particular spot. Using image processing software, the red-to-green fluorescence will be digitized and providing the ratio values output indicating the expression of genes. The process of microarray experiment is illustrated in Figure 2.2.

Typically, microarrays are scanned with a 3M pixel size, generating a .DAT file of >40 Mb. Affymetrix software is used to define position of oligonucleotides and calculate signal intensities of individual oligonucleotides. This converts the .DAT file to a .CEL file that still contains probe-level signal information but has now averaged the individual pixels for a given probe [14].

Finally, the gene expression data set can be noted by the following matrix $M\{w_{ij} \mid 1 \le i \le n, 1 \le j \le m\}$,where the rows (G { $g_1,\ldots,g_n$ }) from the expression patterns of genes, the columns (S { s1,…,sm }) from the expression profiles of samples, and w_{ij} is the measured expression level of gene i in sample j . Thus, M is defined as:

$$M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \leftarrow g_i, i = 1, \cdots, n$$

$$\uparrow$$

$$s_j, j = 1, \cdots, m.$$

2.2.1 Microarray Platforms

Physically, microarrays are rectangular matrices on which DNA molecules, called probes, are pre-affixed in row and column intersections. There are three platforms for Microarrays [3] shown below in the Table 2.1.

Table 2.1 : Different microarray platforms

Probe	Arraying Technique	Microarray Platform
cDNA	Robotic spotting	Spotted cDNA Microarrays
Oligonucleotides (Affymetrix)	Robotic spotting	Spotted Oligonucleotide Microarrays
	In situ synthesis	In situ Oligonucleotide Microarrays

2.2.2 Microarray structures

Each type of microarray has its own unique analysis features. A major commercial source of oligonucleotide arrays (GeneChip, Affymetrix) has a design distinct from spotted arrays.

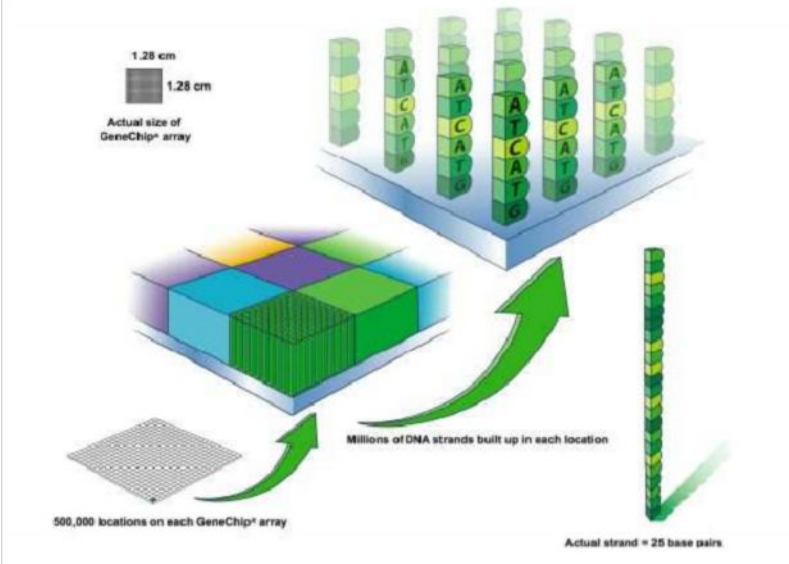

Figure 2.3 : Microarray structure

A GeneChip probe array consists of a number of *cells* (square-shaped areas on the array) and each contains many copies of a unique probe. Probes are tiled in probe pairs consisting of a perfect match (PM) and a mismatch (MM).

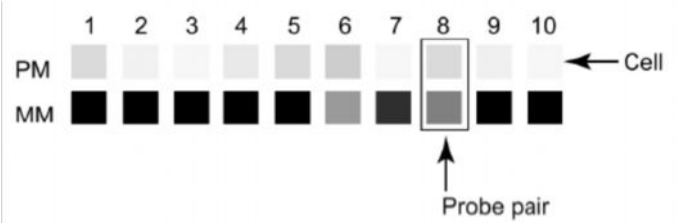

Figure 2.4 : Probe structure in microarray

The sequence of the PM and MM are the same, except for a base substitution in the middle of the MM probe sequence. A probe set includes a series of probe pairs and represents an expressed transcript

2.3 Metadata

The terms annotation and meta-data interchangeably. Both terms refer to the large body of information that is available, primarily through databases, on different aspects of the biological systems under study. This includes sequence information, catalogs of gene names and symbols, structural information, and virtually any relevant publication. We use the term meta-data, which means data about data, as well as the term annotation, because in many of the analyses existing metadata are used to annotate analytic resources or results. Public databases of microarray gene expression data have been quickly growing as the use of high-throughput techniques has become routine in genome-wide studies. Major repositories [15-17] of microarray data, e.g. Gene Expression Omnibus, Array Express or Stanford Microarray Database, are exceptionally rich mines of genomic information and exploiting their content, through meta-analysis, represents an unprecedented opportunity to improve the interpretation and validation of expression studies. Meta-analysis of large microarray expression datasets allows researchers to confirm biological hypotheses, formulated from results of a study, in a relatively inexpensive way, i.e. using data independently obtained in another laboratory, without the need of novel experiments. Meta-analysis also offers the opportunity of re-analyzing formerly available data, in combination with new samples and state-of-the-art computational methods, thus increasing the reliability and robustness of results. Finally, meta-analysis enhances the capabilities of bioinformatics methods to obtain precise estimates of gene expression

differentials and to assess the heterogeneity of overall estimates. Biological validation and the integration of annotation information in data analysis Methods are introduced in this thesis which can improve the interpretation and consistency of results of current biological validation methods.

In microarray-based experiment, different gene feature identifiers are associated with a reasonably large set of biological data for chip specific probe labels. Each of the chip-specific data packages is maintained by database that maps from the probes on the array to Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] pathways for metadata integration. Then by the use of annotate functions simple HTTP queries is sent to web service providers for the assay based meta-data for queried gene features so that navigating the hierarchy, determining parents and children of selected gene are tagged.

2.3.1 KEGG and cMAP

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] provides a data resource that is primarily concentrated on pathways. KEGG associates each pathway with a number. On the other hand, The cancer Molecular Analysis Project (cMAP) provides software and data for the comprehensive exploration of data relevant to cancer. cMAP provides pathway data in a format that is amenable to computational manipulation.

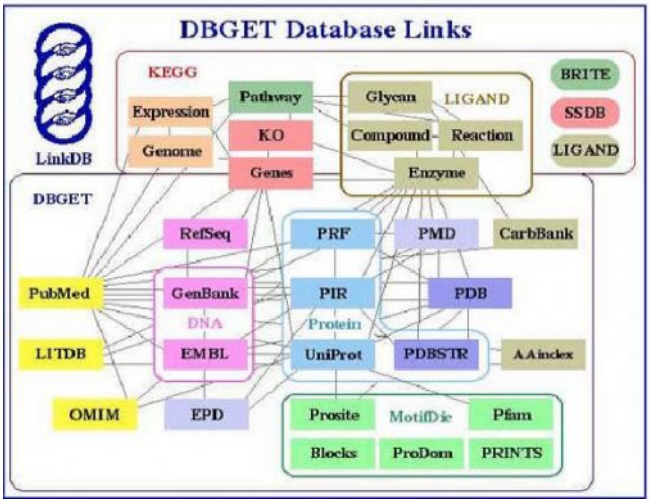

Figure 2.5 : Kyoto Encyclopedia of Genes and Genomes

An ultimate goal of bioinformatics is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity, such as molecular interaction networks involving various cellular processes and phenotypes (morphological, physiological, and behavioral aspects) of entire organisms from genomic information. KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information).

2.4 Microarray Data Repositories

Pooling experimental data requires the standard annotation of the experiments. It also requires interoperability among data repositories supported by standard services and workflows. Interoperable data repositories constitute an enabling resource for meta-analysis. Public datasets have been created in response to the growing demand for publicly available repositories for high-throughput gene expression data. Such public repositories represent an important resource for the biological research community as they provide unrestricted access to microarray data published by other researchers. As such, they complement local in-house gene expression databases by providing reference data for comparative studies. Among them, the Gene Expression Omnibus (GEO) repository developed by the National Center for Biotechnology Information (NCBI) is publicly accessible on the NCBI website at http://www.ncbi.nlm.nih.gov/geo [15]. GEO archives and helps disseminate microarray and other forms of high-throughput data generated by the scientific community. GEO data can be viewed from the perspective of the experiment or the gene. The experiment- centric view presents the entire study, while the gene-centric view displays quantitative gene expression measurements for one given gene across a dataset, with links to gene annotations. Other efforts to archive experiments and make them accessible to the whole community include the Stanford Microarray Database (SMD) [16] and the ArrayExpress database of microarray [17], developed by the

European Bioinformatics Institute. All these repositories promote standard exchange formats such as MAGE-TAB. Moreover, data submitted to these repositories are required to have a common set of core elements. As many other resources in this domain, including local experimental databases, data sets in public repositories are compliant with the standards that define a minimum information about a microarray experiment.

2.5 Affymetrix Microarrays Databases

The microarray experiments carried out in our study employed the Affymetrix GeneChip system. Affymetrix probes are designed using publicly available information. The sequences, from which the probe sets were derived, were selected from GenBank, dbEST, and RefSeq. The sequence clusters were created from the UniGene database and then refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database. Sequences from these databases were collected and clustered into groups of similar sequences.

The probes are manufactured on the chip [19-23] using photolithography (a process of using light to control the manufacture of multiple layers of material), which is adapted from the computer chip industry. Each GeneChip contains approximately 1,000,000 features. Each probe is spotted as a pair, one being a perfect match (PM), and the other with a mismatch (MM) at the centre. These probe pairs allow the quantization and subtraction of signals caused by non-specific cross-hybridization. The differences in hybridization signals between the partners, as well as their intensity ratios, serve as indicators of specific target abundance. Each gene or transcript is represented on the GeneChip probe pairs. The probe sets are given different suffixes to describe their uniqueness and or their ability to bind different genes or splice variants.

- "_at" describes probes set that are unique to one gene
- "_a_at" describes probe sets that recognize multiple transcripts from the same gene

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

- "_s_at" describes probe sets with common probes among multiple transcripts from separate genes. The _s_at probe sets can represent shorter forms of alternatively polyadenylated transcripts, common regions in the 3′ ends of multiple alternative splice forms, or highly similar transcripts. Approximately 90% of the _s_at probe sets represent splice variants. Some transcripts will also be represented by unique _at probe sets.
- "_x_at" designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridisation are dropped in order to design the _x_at probe sets. These probe sets share some probes identically with two or more sequences and therefore, these probe sets may cross-hybridise in an unpredictable manner.

This chapter focuses on the preliminary concepts of molecular biology and microarray data analysis. It is explained how DNA microarray data is prepared and the structure of microarray data.

ABBYY PDF Transformer 2.0
Click here to buy
www.ABBYY.com

3. CHAPTER 3

REVIEW OF RELATED LITERATURE

3.1 Introduction

Classification based on microarray data faces with many challenges. The main challenge is the overwhelming number of genes compared to the number of available training samples, and many genes are not relevant to the distinction of samples. These irrelevant genes have negative effect on the accuracy of the classifier and increase data acquisition cost as well as learning time. Moreover, different combination of genes may provide similar classification accuracy. Another challenge is that DNA array data contain technical and biological noises. So, development of a reliable classifier based on gene expression levels is getting more attention.

In machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly-identified observations is available. The corresponding unsupervised procedure is known as clustering and involves grouping data into categories based on some measure of inherent similarity. Each classification technique suffers from high dimensionality of data. As microarray benchmark data of repository contains huge data produced from different sources in the world, the data need to preprocess for cancer classification. The following steps such as normalization [24-27], probe level expression measurement [29-32], and gene filtering for dimensionality reduction, high level analysis and biological relationship are essential for better classification for microarray data. The steps are listed in the following table 3.1. The performance of classification depends on the appropriate set of microarray data. There are several new techniques has been made in microarray data analysis in the last decades.

This chapter contains a review of works of classifier design, affymetrix microarray data processing, dimensionality reduction, gene feature filtering, selection methods, statistical analysis, different classification techniques and performance analysis metrics.

Table 3.1 : Summary of microarray data analysis

Analysis Stage	Description	Examples of Methods
Normalization	Equalizes overall signal across arrays to be compared, ensures linearity of response across abundance classes	Whole chip [26] Quartile [27-28]
Probe reduction	Combines signals from multiple probes or probe pairs to define "expression level". Identifies genes with invalid or hyper-variable expression levels.	Weighted average (MAS 4) [29] Tukey bi-weight (MAS 5) [30] Model-based (MBEI) [31] Log scale linear additive (RMA) [32]
Statistical comparative factors	Compares expression of a gene across two or more arrays to determine significant changes in expression.	t-test, S-score [33] rank order (MAS 5) [30] Permutation (SAM) [36-37]
Classification and Clustering	Identifies significant correlations in expression data across experiments/conditions.	Classification [38-40] hierarchical clustering [39] k-means clustering [39] principle components analysis[53] etc.
Biological overlay	Identify functions for given genes, clusters of genes; hypothesis generation.	Multiple database access (Source)[41] Gene Ontology rankings (GenMAPP, MAPPFinder, DAVID/EASE)[42-45] PubMed correlations (PubGene)[49]

3.2 Classifier Design

In classification theory, it is required to assign an object to different classes. The assignment is based on attributes of the objects which are different between the members of different classes but similar between members of same class. These attributes are called features. A classifier takes features as input data and labels an object with a class label. The outcome of the classifier is always not correct. The reliability of the classifier is determined by the probability of error it makes. In a binary classifier, where there are only two classes there are two types of errors, termed as false positive and false negative. The probability that an object is classified as class 1 while it belongs to class 0 is referred to as false positive and the probability that an object is classified as class 0, while it belongs to class 1 is referred to as false negative. The goal of classifier design algorithm is to come up with classifiers with low probability of errors.

There are several well known classification algorithms such as Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Nearest mean classifier, Support vector machines (SVM), Neural Networks etc., all of which try to estimate optimal Bayesian classifier. The performance of each classifier depends on the problem conditions. Therefore, it is used all the available data when designing a classifier. But, when it is to train a classifier from a data base, the behavior of probability of error versus number of features is different. Initially, with the increase in the number of features the probability of error decrease, but after a while it starts increasing with any added feature. This problem is referred to as the Curse of Dimensionality or Peaking Phenomenon. This phenomenon is depicted in figure 3.1. This figure is a plot between number of features and probability of error. The curve $E[\varepsilon_{d,n}]$ represents the behavior of probability of error with increase in number of features obtained and the curve ε_d represents the behavior of probability of error obtained from classification theory. Therefore, in supervised classification, there are an optimal number of features that should be identified from the original set to design a classifier. This is called feature selection.

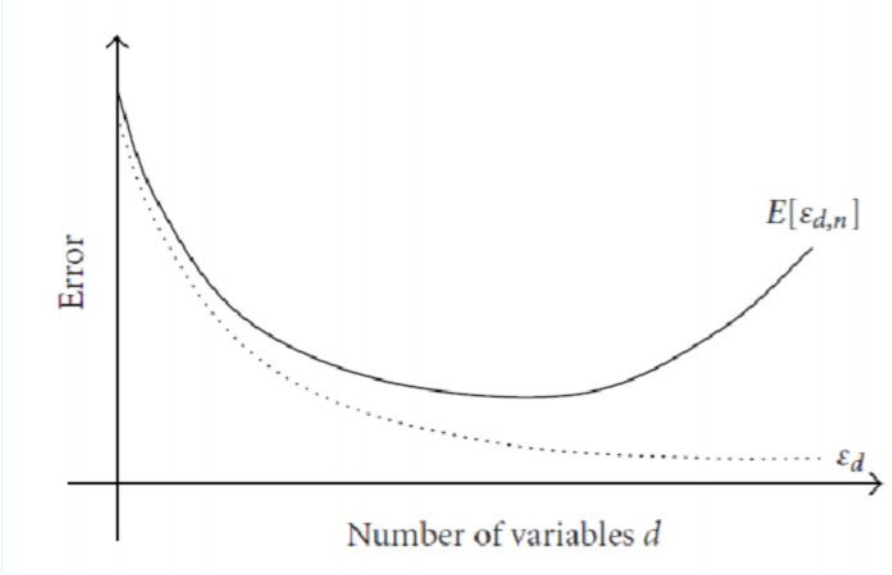

Figure 3.1 : Curse of Dimensionality or Peaking Phenomenon

In this section it is reviewed various classification rules, feature selection methods and error estimation methods used for classifier training in this thesis. The following classification rules are discussed: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Nearest Mean Classifier (NMC). The Feature selection methods described are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Bidirectional Selection (BDS).

3.2.1 Classifier Design

A classifier design involves 3 steps:

- Selecting a set of features that can best differentiate between the classes.
- Finding a classification rule or models that classifies based on the values of the features and decides the class of the object.
- Evaluate the performance of the designed classifier using error estimation methods.

Classification rules can be divided into two categories; parametric and non-parametric. Parametric classifiers assume that the underlying distribution of data is fully described by a minimum number of parameters like mean, variance and covariance. Examples of parametric models are Linear Discriminant Analysis (LDA), Quadratic discriminant Analysis (QDA). Non parametric models are mathematical procedures for hypothesis testing which make no assumption of the probability

distribution of the variable being assessed. Examples of non parametric model are k-nearest-neighbor rule, Kernel rules, neural networks etc. In general the performance of both types depends heavily on the sample size but for small sample sizes the parametric classifiers tend to perform better.

3.3 Classification

Data classification [40-47] is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. A tuple, X, is represented by an n-dimensional attribute vector, $X = (x_1, x_2, : : : , x_n)$, depicting n measurements made on the tuple from n database attributes, respectively, $A_1, A_2, : : : , A_n$. Each tuple, X, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are selected from the database under analysis. In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects. Because the class label of each training tuple is provided, this step is also known as supervised learning (i.e., the learning of the classifier is "supervised" in that it is told to which class each training tuple belongs). It contrasts with unsupervised learning (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

The features are gene expression coefficients and patterns correspond to patients. If it is limited to two-class classification problems, then two classes may be considered with the symbols (+) and (-). The training patterns are used to build a decision function (or discriminant function) D(x), that is a scalar function of an input pattern x. New patterns are classified according to the sign of the decision function:

$$D(x) > 0 \Rightarrow x \in \text{class}(+)$$
$$D(x) < 0 \Rightarrow x \in \text{class}(-)$$
$$D(x) = 0, \text{decision boundary}$$

A data set is said to be “linearly separable” if a linear discriminant function can separate it without error.

Classification and prediction [40-49] are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. While classification predicts categorical labels (classes), prediction models continuous-valued functions. Predictive accuracy, computational speed, robustness, scalability, and interpretability are five criteria for the evaluation of classification and prediction methods.

ID3, C4.5, and CART are greedy algorithms [44] for the induction of decision trees. Each algorithm uses an attribute selection measure to select the attribute tested for each nonleaf node in the tree. Pruning algorithms attempt to improve accuracy by removing tree branches reflecting noise in the data. Early decision tree algorithms typically assume that the data are memory resident—a limitation to genetic mining on large dataset.

Naïve Bayesian classification and Bayesian belief networks are based on Bayes, theorem of posterior probability. Unlike naïve Bayesian classification (which assumes class conditional independence), Bayesian belief networks allow class conditional independencies to be defined between subsets of variables. A rule-based classifier uses a set of IF-THEN rules for classification. Rules can be extracted from a decision tree. Rules may also be generated directly from training data using sequential covering algorithms and associative classification algorithms.

Backpropagation is a neural network algorithm for classification that employs a method of gradient descent. It searches for a set of weights that can model the data so as to minimize the mean squared distance between the network’s class prediction and the actual class label of data tuples. Rules may be extracted from trained neural networks in order to help improve the interpretability of the learned network. A Support Vector Machine (SVM) [45] is an algorithm for the classification of both linear and nonlinear data. It transforms the original data in a higher dimension, from where it can find a hyperplane for separation of the data using essential training tuples called support vectors.

Decision tree classifiers, Bayesian classifiers, classification by backpropagation, support vector machines, and classification based on association are all examples of eager learners in that they use training data to construct a generalization model and in this way are ready for classifying new sample. This contrasts with lazy learners or instance based methods of classification, such as nearest-neighbor classifiers and case-based reasoning classifiers, which store all of the training data in pattern space and wait until presented with a test tuple before performing generalization. Hence, lazy learners require efficient indexing techniques.

In genetic algorithms, populations of rules "evolve" via operations of crossover and mutation until all rules within a population satisfy a specified threshold or fitness function. Linear, nonlinear, and generalized linear models of regression can be used for prediction. Many nonlinear problems can be converted to linear problems by performing transformations on the predictor variables. Unlike decision trees, regression trees and model trees are used for prediction. In regression trees, each leaf stores a continuous valued prediction. In model trees, each leaf holds a regression model. Stratified k-fold cross-validation is a recommended method for accuracy estimation.

3.4 Microarray Data Processing

The information of microarray probe represented by pixels of 640 x 640 with mean and standard deviation in CEL file that is available in standard microarray data repository. Data preprocessing and normalization are common tasks prior to any study in order to clean up data of any biases or systematic errors due to array technology and measurement instruments. The following steps are required for microarray data processing:

• ***Background correction:*** A first step consists in performing a background correction to laminate bias due to uneven level of noise across the slide.
• ***Spot quality assessment***: Then one can filter the spots according to their quality. In the case of Stanford spotted arrays, this is assessed by a quantity called spot correlation. The higher the value, the better. Spots poorly scanned are thus not taken into account.

• ***Set lower or percentage cut-offs***: Another simple method is to set cutoffs to exclude from analysis any genes for which the expression values do not meet the requirements, threshold or percentage of valid data.

• ***Log2 transform:*** This transformation is widely used since it allows a similar treatment for genes that are up or down-regulated. Logarithms treat numbers and their reciprocal symmetrically. An interesting feature of this transform is to suppress minor distortions and make the distribution of expression values ”more normal”. It removes also outliers due to very low level of expression of controls compared to cases.

• ***Normalization:*** This transformation adjusts intensities so that quantities across the matrix can be compared more accurately. It tends to minimize the effects of systematic variation in measurements. There is no unique way to normalize data. This research topic is currently very active and many methods can be found in the literature [52]). Among them, global normalization by centering the distribution according to the mean or the median is useful.

3.5 Low Level Analysis

Low-level analysis and assessment of quality control are a crucial aspect of microarray experiments. Many approaches have been proposed for ensuring adequate quality control in microarray experiments. The issues are scaling factor, background noise, percent of genes etc. Following signal acquisition and calculation of individual oligonucleotide intensities, the first step in microarray analysis, is normalization of signal intensities [22]. Although normalizing across all probes on an array (whole chip) [26-27] is often adequate, there can be significant problems with non-linearity, particularly with high-abundance genes. Once normalized, individual oligonucleotide probe pairs are usually “reduced” to a single number representing the expression level for the given gene. Multiple algorithms [28-32] for deriving this expression intensity have been developed. Affymetrix originally devised a “trimmed mean” method for determining an “average difference” value (MAS 4). But this was prone to large fluctuations with lower abundance genes, even producing “negative” values. A more recent version, MAS 5.0 uses a statistical expression algorithm to calculate the signal on the oligonucleotide array. This analysis produces a “change p-value” and a “change call”. Another probe-based method, model based expression index (MBEI), was developed by Li and

Wong for oligonucleotide array analysis. They demonstrated a new method, referred to as the log scale robust multi-array analysis (RMA) [32].

3.6 Dimensionality Reduction and Feature selection

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation has a dimensionality that corresponds to the intrinsic dimensionality of the data. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data. Dimensionality reduction is important in many domains, since it facilitates classification, visualization, and compression of high-dimensional data, by mitigating the curse of dimensionality and other undesired properties of high-dimensional spaces. It is investigated the linear techniques PCA and nonlinear filtering for cancer classification.

3.6.1 Consistency Driven Feature Selection

Almuallim and Dieterich [33,54] describe an algorithm originally designed for boolean domains called FOCUS. FOCUS exhaustively searches the space of feature subsets until it finds the minimum combination of features that divides the training data into pure classes (that is, where every combination of feature values is associated with a single class). This is referred to as the "min-features bias". Following feature selection, the final feature subset is passed to ID3 which constructs a decision tree.

3.6.2 Feature Selection Through Discretization

Setiono and Liu [34,55] note that discretization has the potential to perform feature selection among numeric features. If a numeric feature can justifiably be discretized to a single value, then it can safely be removed from the data. The combined discretization and feature selection algorithm Chi2 uses a chi-square statistic $\boldsymbol{x}^2$ to perform discretization.

3.6.3 Principal Component Analysis (PCA)

PCA is one method used to reduce the number of features used to represent data. The benefits of this dimensionality reduction include providing a simpler representation of the data, reduction in memory, and faster classification. PCA [35, 53] uses a linear transformation to obtain a simplified data set retaining the characteristics of the original data set.

Suppose that the data to be reduced consist of tuples or data vectors described by n attributes or dimensions. Principal components analysis, or PCA (also called the Karhunen-Loeve, or K-L, method), searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$. The original data are thus projected onto a much smaller space, resulting in dimensionality reduction. Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA "combines" the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set. PCA often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result. The basic procedure is as follows:

a) The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.

b) PCA computes k orthonormal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components. The input data are a linear combination of the principal components.

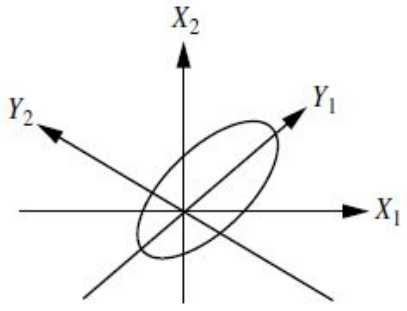

Figure 3.2 : Principal component analysis

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

c) The principal components are sorted in order of decreasing "significance" or strength. The principal components essentially serve as a new set of axes for the data, providing important information about variance. That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on. For example, Figure 3.1 shows the first two principal components, Y1 and Y2, for the given set of data originally mapped to the axes X1 and X2. This information helps identify groups or patterns within the data.

d) Because the components are sorted according to decreasing order of "significance," the size of the data can be reduced by eliminating the weaker components, that is, those with low variance. Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

PCA is computationally inexpensive, can be applied to ordered and unordered attributes, and can handle sparse data and skewed data. But it suffers from multidimensional data like microarray data[34]. Principal components may be used as inputs to multiple regression and cluster analysis.

3.6.4 Information Gain (IG) for Decision Tree

Information Gain (IG) can be used to measure the information content in a feature [52], and is commonly used for decision tree induction. Maximizing IG is equivalent to minimizing:

$$\sum_{i=1}^{V} \frac{n_i}{N} \sum_{j=1}^{K} \frac{n_i}{N} \left(-\frac{n_{ij}}{n_i} \log_2 \frac{n_{ij}}{n_i}\right),$$

where K is the number of classes, V is the number of values of the attribute, N is the total number of examples, n_i is the number of examples having the ith value of the attribute and n_{ij} is the number of examples in the latter group belonging to the jth class.

3.6.5 Gene Feature Reduction

Cancer classification using gene expression data is a nontrivial task due to the very nature of the gene expression data. The expression data has very high dimensionality [54], usually in the order of

thousands to tens of thousands of genes. The situation is more complicated with the number of sample sizes, usually below hundred. The high dimensionality of the features and the low population size usually cause over-fitting of the classifier. A term "curse of dimensionality" is coined to refer to this situation. Computational expenses also impose important limitations. Another key issue is, due to not all genes being related to the cancer, it is difficult to extract biologically meaningful genes.

The techniques for dimensionality reduction of gene can be divided into transformation and selection based reduction. Transformation based reduction such as Principal Component Analysis (PCA) transforms the original features of a dataset with a typically reduced number of uncorrelated ones, termed principal component. In contrast, selection reduction techniques attempt to determine a minimal feature subset from a problem domain while retaining the meaning of the original feature sets. Thus, selection based reduction techniques have become the main preference in many bioinformatics applications, especially microarray data analysis since it offers the advantage of interpretability by a domain expert.

Feature selection [55] is the process of systematically reducing the dimensionality of a dataset to an optimal subset of attributes for classification purposes. Problem of feature selection is hence, an important issue in cancer classification. It has been shown that, in many applications feature selection process improves a classifier's prediction capability.

The objectives of feature selection techniques are many, the major ones are: i.) To avoid over fitting and improve model performance, for example selecting highly informative genes could enhance the accuracy of classification model. ii.) To provide faster and more cost-effective models, and iii.) To gain a deeper insight into the underlying processes that generated the data.

Feature selection techniques have many benefits, it also introduces extra complexity level which requires thoughtful experiment design to address the challenging tasks, yet provide fruitful results. In the context of classification, feature selection techniques can be organized into three categories,

depending on how they combine the feature selection search with the construction of the classification model: filter method, wrapper method and embedded method.

Filter method rank each feature according to some univariate metric and only the highest ranking features are used while the remaining low ranking features are eliminated. This method also relies on general characteristics of the training data to select some features without involving any learning algorithm. Therefore, the results of filter model will not affecting any classification algorithm. Moreover, filter methods also provide very easy way to calculate and can simply scale to large-scale microarray datasets since it only have a short running time. Univariate filter methods such as Bayesian Network Information Gain (IG) and Signal-to-Ratio(SNR) and Euclidean Distance have been extensively used in microarray data to identify informative genes.

3.7 Text and Literature Mining

Text and literature mining is emerging as a promising area for data mining in biology. One important representation of text and documents is the so-called bag-of-words (BOW) representation, where each word in the text represents one variable, and its value consists of the frequency of the specific word in the text. It goes without saying that such a representation of the text may lead to very high dimensional datasets, pointing out the need for feature selection techniques.

Although the application of feature selection techniques is common in the field of text classification, the application in the biomedical domain is still in its infancy. Some examples of FS techniques in the biomedical domain include the work, who use the Kullback–Leibler divergence as a univariate filter method to find discriminating words in a medical annotation task, the work of Eom and Zhang, who use symmetrical uncertainty (an entropy-based filter method) for identifying relevant features for protein interaction discovery, and the work, which discusses the use of feature selection for a document classification task.

It can be expected that, for tasks such as biomedical document clustering and classification, the large number of feature selection techniques that were already developed in the text mining community will be of practical use for researchers in biomedical literature mining [47-48].

Frequent pattern growth (FP-growth) is a method of mining frequent itemsets without candidate generation. It constructs a highly compact data structure (an FP-tree) to compress the original transaction database. Rather than employing the generateand test strategy of Apriori-like methods, it focuses on frequent pattern (fragment) growth, which avoids costly candidate generation, resulting in greater efficiency. An interesting method in this attempt is called frequent-pattern growth, or simply FP-growth, which adopts a divide-and-conquer strategy as follows. First, it compresses the gene databases representing frequent items into a frequent-pattern tree, or FP-tree, which retains the itemset association information. It then divides the compressed database into a set of conditional databases. Each associated with one frequent item or "pattern fragment," and mines each such database separately.

3.8 Meta-analysis

One advantage of integrating large numbers of microarray studies and compiling them in a data-warehouse is that it makes it possible to compare the results of different studies and to determine which methods are robust and produce consistent results across a range of studies. There are, however, many problems associated with the comparison of gene expression profiles across disparate microarray data sets. In studies performed in 2004 to 2011 by several teams. Different technologies seemed to show good agreement within and across labs using the same RNA samples. The variability between two labs using the same technology was higher than that between two technologies within the same lab. Moreover, the source of RNA samples can make a difference in microarray data. Several methods have been developed to address these variability issues in multiple, independent data sets generated on various platforms.

3.8.1 Challenges of Integrative Analysis of Expression Data

In recent years, different strategies are to combine results from independent. The choice of the most effective meta-analysis technique depends on the type of and on the objective of the study. Meta-analysis strategies can be divided into two broad classes: data integration and data combination. Statistical techniques as vote counting, p-value or rank combination and effect size estimation have been used for meta-analyses based on data integration. Instead, data combination encompasses the

direct comparison of different studies, is applicable only when expression profiles have been obtained using the same array technology (e.g. Affymetrix).

Despites numerous efforts, mining and analyzing publicly available microarray data still represents a bioinformatics challenge and the lack of appropriate tools able to overcome critical issues, as annotation, cross-platform comparison and handling of metadata, is still hampering the potentialities of large-scale meta-analyses. Performing a meta-analysis of independent microarray studies requires to carefully handle the heterogeneity of array designs, which complicates cross-platform integration, and of sample descriptions, which impact the correct characterization of specimens.

This chapter reviews and summarizes the analysis stages and methods for microarrays from data preprocessing. Here it is explained low level analysis, dimensionality reduction techniques, classifications and clustering techniques, text and literature mining, challenges of integrative analysis of expression data.

ABBYY PDF Transformer 2.0
Click here to buy
www.ABBYY.com

4. CHAPTER 4

PROPOSED CANCER CLASSIFICATION METHOD BY INTEGRATING METADATA

4.1 Cancer Classification

People all over the world are suffering from cancers and the patients of cancer are increasing day by day. cancer is affected due to abnormal change in DNA sequence of a cell. So, identification of cancer and classify to appropriate type are the great challenge in medical science. In the past few decades, "technology explosion" has created an immense impact on both biomedical research and clinical medicine. Tremendous strides were made with the aid of numerous new technologies such as recombinant DNA methods, DNA sequencing, magnetic resonance imaging (MRI), polymerase chain reaction (PCR), monoclonal antibodies, and so forth. Despite these, major hurdles remain. In the field of cancer medicine, limited successes are still overshadowed by the tremendous morbidity and mortality incurred by this devastating disease. It has become increasingly important to integrate new technologies into both cancer research and clinical practice if we hope to win the battle against cancer.

Although some methods used to distinguish and classify human malignancies rely on a variety of clinical, molecular and morphological parameters, precise cancer diagnosis remains a challenging task. Existing diagnostic classes are often heterogeneous and include diseases with different clinical courses, therapeutic response and metastatic potential. Genome-wide gene expression measurements can give an insight into genetic pathways and gene networks. They can point to new molecular markers that can be widely used in clinical diagnosis, and lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more reliable classification. In this chapter, an improved cancer classification method is proposed by integrating metadata. After microarray data preprocessing, gene features will be ranked by the integration of covariance and metadata ranking. Then subset of gene feature will be considered by supervised attribute selection which will be used for classification model preparation. The proposed method for cancer classification by integrating metadata is shown in the following figure 4.1.

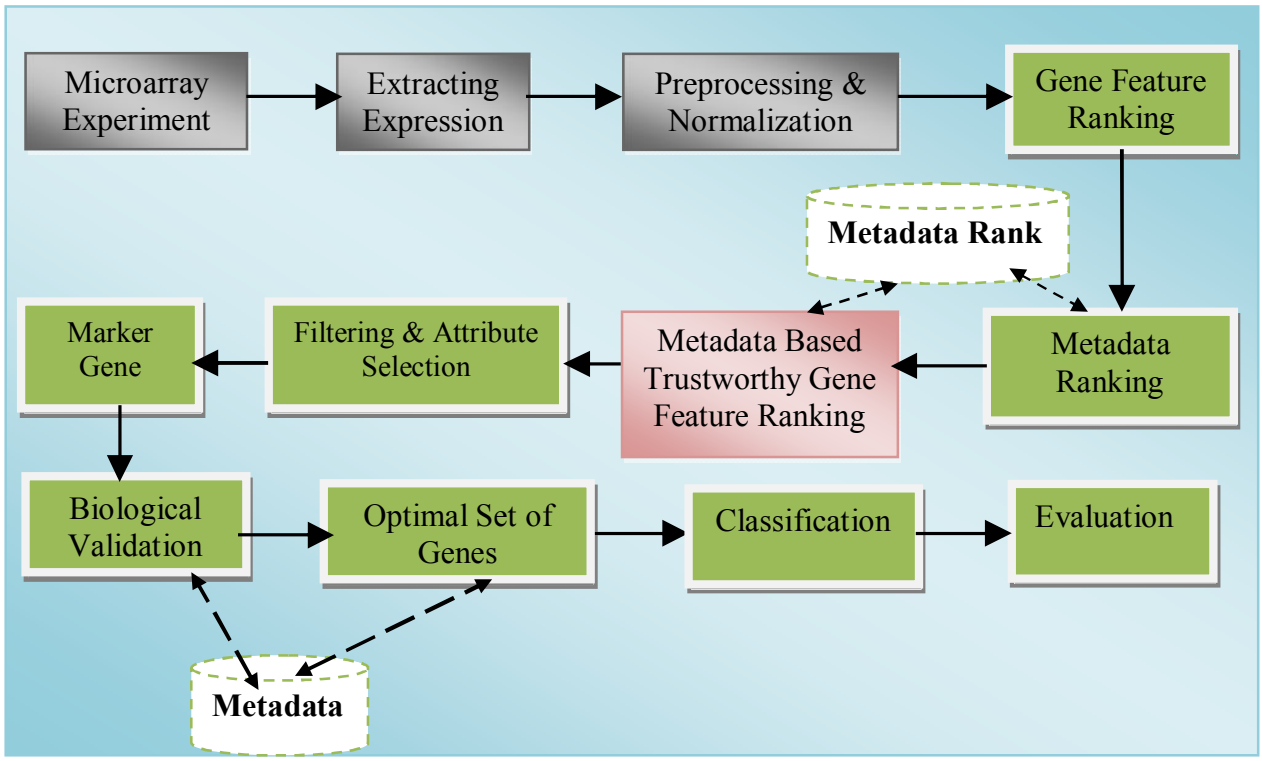

Figure 4.1 : Metadata based cancer classification for microarray data

4.2 Microarray Data Preprocessing

The Microarray data is preprocessed according to the steps described in the section 3.4. In this case normalization is one of the important steps because different types of cancer dataset are used for the model. Preprocessing of the data is an important step prior to classification and clustering. The large number of genes present in a microarray experiment may be excessive for application of some algorithms with limited resources. As many of the algorithms are based on Euclidean distance between samples, the first step should consist of normalization to avoid samples with the larger dynamic range to take over the process. A good review of normalization of microarray data can be found in [52]. A second step is the removal of all genes that show low variation across samples, which may affect negatively the clustering process.

In a microarray experiment, there are many sources of variation. Some types of variation, such as differences of gene expressions, may be highly informative as they may be of biological origin. Other types of variation, however, may be undesirable and can confound subsequent analysis,

leading to wrong conclusions. In particular, there are certain systematic sources of variation, usually owing to a particular microarray technology, that should be corrected prior to further analysis. The process of removing such systematic variability is called normalization. There may be a number of reasons for normalizing microarray data. For example, there may be a systematic difference in quantities of starting RNA, resulting in one sample being consistently overrepresented. There may also be differences in labeling or detection efficiencies between the fluorescent dyes (e.g., Cy3, Cy5), again leading to systematic over expression of one of the samples. Thus, in order to make meaningful biological comparisons, the measured intensities must be properly adjusted to counteract such systematic differences.

4.3 Gene Feature Ranking

Cancer classification from microarray data is based on the significant gene/gene features selected by analyzing the data. The application of different statistical methods must be handled with care as the microarray contains huge amount of features. To overcome this challenge, Gene Feature Ranking is used followed by proper normalizing. In the proposed method, metadata based Gene Feature Ranking is proposed as a feature selection method for microarray data. Initially, gene features are ranked based on their *covariance* measure which indicate more discriminative characteristics among the genes because these(genes) cannot be discriminated by expression measure. the genes are ranked based on their covariance measure. Genes with covariance measure having a threshold are selected. It is noteworthy to mention that most existing approaches [20-25] for reducing the number of genes are based on the measurement of expression using “spot quality” instead of “quality measures”[11]. Using the covariance measure, the proposed classification method is able to better classify the cancer dataset.

Feature ranking is computed by the covariance of a gene is measured by evaluating a value called *covariance measure* ***(C_i)*** on preprocessed microarray dataset by the following equation:

$$C_i = \frac{\| \text{cov}_{\alpha} - \text{cov}_{\beta} \|}{\frac{\text{cov}_{\alpha} + \text{cov}_{\beta}}{2}} \quad \text{..... ....} (1)$$

where cov_{α} and cov_{β} indicate the average covariance of the probe set (which represents gene). The threshold value should be high if the average of the covariance between the two sub-types is low and

the difference between the two covariance is high. The covariance measure **(C_g)** will be furthermore integrated with metadata ranking to calculate covariance measure and metadata based trustworthy rank of genes. The genes having values below the threshold are considered to be too noisy and are discarded.

4.4 Metadata Ranking

The terms annotation and meta-data interchangeably. This includes sequence information, catalogs of gene names and symbols, structural information, and virtually any relevant publication. The relationship between biological metadata integration and a data-analytic workflow is inherently complex[50]. Annotation may be used to perform gene feature reduction during statistical model building stage. The term meta-data(annotation) is used here as the gene is cited in PubMed archive.

Figure 4.2 : Accessing the list of metadata from online data repository

It is a service of the National Library of Medicine that provides a very rich resource of data and tools for working with papers published in journals that are related to medicine and health. Genes are linked to published papers. For this reason, a query to PubMed is requested to provide the number of citations for corresponding gene features that will be considered for metadata based ranking for trustworthy gene feature ranking. In this case, the number of outcome is considered as $\boldsymbol{n_i}$ for gene i. So, Gene wise metadata ranking is considered as M_i.

$$G_i = \sum n_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$M_i = G_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The value of M_i may be null for few of genes, so it is ignored in this worked. As the literature regarding the requested gene is dynamically changing day by day, the G_i measure in the metadata database of the integrated system will be updated periodically.

4.5 Metadata Integration

Gene features may be ranked based on their covariance measures [10] which indicate more discriminative characteristics among the genes. But the trustworthiness of the features for biological significance is important to maintain a balance between the discriminative qualities of the features and their trustworthiness for biological validation. For this reason, This research aims to integrate metadata for selecting biologically trustworthy genes.

To find the most significant genes, integration of metadata with gene features is required. After integration, the optimal number of genes are selected for classification. The block diagram of the overall process is shown in the figure 4.1.

Metadata based trustworthy feature ranking will be evaluated by the following equation:

$$T_i = C_i . M_i \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

The genes having higher values of T_i are more significant for classification. How many genes will be effective and efficient for classification model that will be examined for different conditions. The genes having lower T_i value are assumed to be too noisy or insignificant for classification and so are discarded. After integration it will be processed through a conventional feature selection algorithm.

4.6 Feature Selection

Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. In some cases, accuracy of classification can be

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

improved. Feature selection algorithms perform a search through the space of feature subsets and must address four basic issues affecting the nature of the search:

a) Starting point: Selecting a point in the feature subset space from which to begin the search can affect the direction of the search. One option is to begin with no features and successively add attributes. In this case, the search is said to proceed forward through the search space. Conversely, the search can begin with all features and successively remove them. In this case, the search proceeds backward through the search space. Another alternative is to begin somewhere in the middle and move outwards from this point.

b) Search organization: An exhaustive search of the feature subspace is prohibitive for all but a small initial number of features. With N initial features there exist 2^N possible subsets. Heuristic search strategies are more feasible than exhaustive ones and can give good results, although they do not guarantee finding the optimal subset.

c) Evaluation strategy: How feature subsets are evaluated is the single biggest differentiating factor among feature selection algorithms for machine learning. One paradigm the *filter* operates independent of any learning algorithm - undesirable features are filtered out of the data before learning begins. These algorithms use heuristics based on general characteristics of the data to evaluate the merit of feature subsets. Another thought argues that the bias of a particular induction algorithm should be taken into account when selecting features. This method, called the *wrapper,* uses an induction algorithm along with a statistical resampling technique such as cross-validation to estimate the final accuracy of feature subsets.

d) Stopping criterion: A feature selector must decide when to stop searching through the space of feature subsets. Depending on the evaluation strategy, a feature selector might stop adding or removing features when none of the alternatives improves upon the merit of a current feature subset. Alternatively, the algorithm might continue to revise the feature subset as long as the merit does not degrade. A further option could be to continue generating feature subsets until reaching the opposite end of the search space and then select the best.

All filter methods use heuristics based on general characteristics of the data rather than a learning algorithm to evaluate the merit of feature subsets. As a consequence, filter methods are generally

much faster than wrapper methods and so they are more practical for use on data of high dimensionality. The following criterion was used to generate gene lists (a) Fold change is the ratio of the mean of the experimental group to that of the baseline. It is a metric to define the gene's mRNA-expression level between two distinct experimental conditions. (b) The difference of Affymetrix expression units (gene expression values obtained after dChip processing) was also incorporated for finding differentially regulated genes. (c) The t-test assesses whether the means of two groups are statistically different from each other. Subsequently, the p-value is calculated from the t-test. The purpose of the t-test is to evaluate the null hypothesis that there is no difference between the means of two samples. The t-test is a parametric test which is used to analyze the mean and standard deviation of two or more groups of samples based on a number of underlying assumptions, including a normal distribution of the data within the test. Therefore, hypothesis testing facilitates the calculation of the probability of the observed value of the t-statistic occurring based on the assumption that the null hypothesis is true. For calculation of the probability, the data is assumed to be normally distributed. By convention, a p-value of ≤ 0.05 is usually considered sufficient to reject the null hypothesis, i.e. that there is a real difference between the means (≤ 0.01 would be considered strong evidence).

4.6.1 Correlation-based Feature Selection (CFS)

CFS is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features.

4.6.2 Marker Gene Selection

In the proposed classification method, correlation-based feature selection method is applied to the gene set selected by metadata based Trustworthy Gene Feature Ranking and all the gene in the dataset to compare the performance of classification. That means, one dataset is obtained after the

mtGFR method while other dataset is considered without mtGFR. Here it has been used a supervised attribute selection based the discriminative quality of a subset of attributes for individual predictive ability of each feature. From this, subsets of the list of genes that are highly correlated with the class are selected. During classification metadata rank is considered to optimize the number of genes. The classifiers will result a set of genes that are assessed as both discriminative and biologically trustworthy.

4.7 Significant Genes through Biological Validation

Every experiment provides static information on the expression level of a gene at a given time point, whereas the analysis of a row gives a picture of its expression pattern. Knowing the variation in expression across the time line provides interesting information on how a gene is regulated. Novel genes that might be linked to a particular disease can be found by comparing their patterns with those of genes that are known to be linked to that disease. One can also assign functional class properties of a newly discovered gene from known classes of genes that show similar patterns.

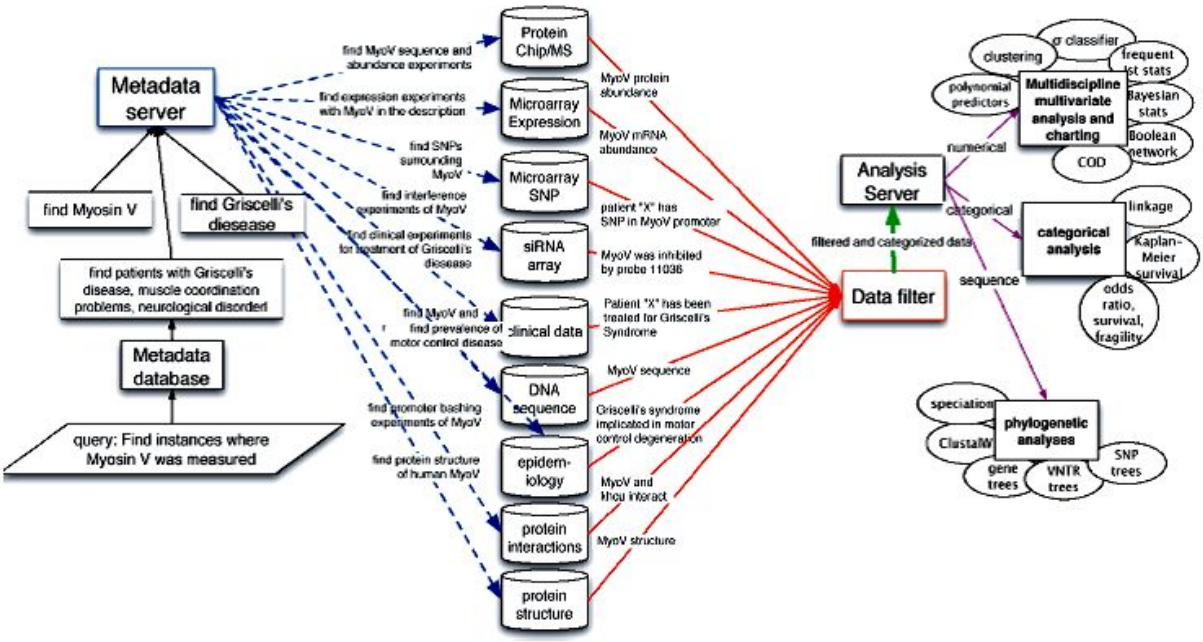

Figure 4.3 : A federated system of databases for systems biology

In figure 4.3 it is showed the left diagram (dotted lines) shows a putative query for a gene and associated disease information. The lines on the right show the reworked query results that provide much more consistency and usability for statistical analyses. The right side also highlights the data receipt modules that shuttle data into and out of analysis routines, which are split into 3 distinct groups—categorical (clinical or patient outcome, epidemiology, disease association), genetic and phylogenetic (multiple genome comparisons and evolution), and numerical data-mining Bayesian and frequency statistics, classification, neural networks, support vector machine (SVM, etc.). These analysis tools are used to interrogate as much of the data as possible in native format, group by group, and then to analyze all of the data that can be combined into one universal type (i.e., binned, or quantized values). Metadata based biological validation is considered by a scoring scheme through relationship analysis of gene with related protein, diseases, drug, cells, species, and other visualization connections. Optimal number of genes are selected from marker genes by the following equation called *Relation Score(g)*.

Relation Score (g) = S(Diseases) + S(Protein) + S(Drug) + S(Cells) + S(Species) + S(Others)...(6)

The genes with high relation score are considered for optimal set of gene for classification. The optimal set of genes are evaluated by recursive elimination of genes based on the score. The gene with lowest relation score is eliminated first up to threshold value.

The techniques and methods described in the previous sections certainly bring improvements to the data analysis, but in the end, the results of the analyses have to be translated back to the biology. Ultimately these biological facts will give confidence about results from the data analysis to the biologists which performed the experiments. This process can be described as biological validation and should be a vital part of the introduction of a new data analysis method, to prove the applicability of the method in practice.

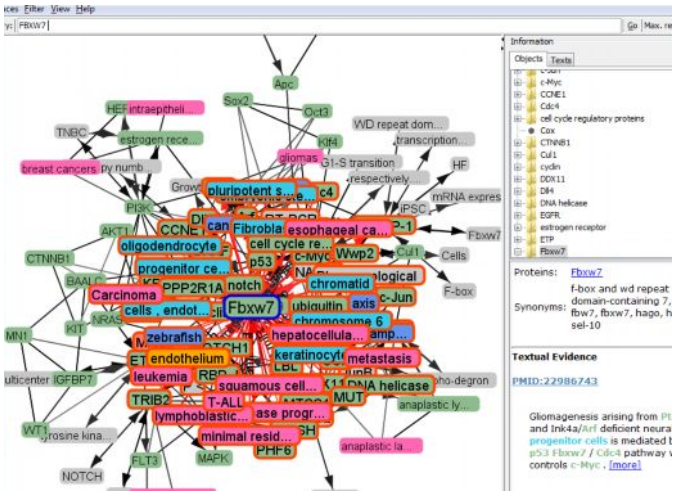

Figure 4.4 : Annotation for a gene using metadata integration

Biological validation can become a sort of vicious circle sometimes: when the results of the analysis agree with current knowledge about a system, one could ask what new insights can be gained from the experiments. On the other hand, one could begin to doubt the results from an experiment when the prior knowledge about a system is not directly reflected in the validation process. Therefore, it is useful to mention that biological validation can be considered as a test of the outcome of an experiment against current knowledge about a biological system at hand. The ultimate biological validation of the results from methodological analysis is of course the confirmation of newly generated hypotheses in the wet lab. The term biological validation is used in this thesis as the description of methodological results with the prior information which is available. Several methods are available which have the objective of biological validation. They range from reconstructing genetic networks to mining literature databases.

4.8 Classification and Evaluation

The genes selected by metadata based gene feature ranking are then used for cancer classification and the accuracy of the classification is evaluated. In the next chapter the simulation and performance analysis of different classification algorithms on different dataset are explained.

In this chapter the steps of classification using metadata integration is discussed. At first, the genes are ranked with their covariance measure, then the metadata is ranked with their weight function and

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

then these both measure is considered for metadata based trustworthy GFR is performed. After appropriate attribute selection, the classification is evaluated dynamically by the use of metadata.

ABBYY PDF Transformer 2.0
Click here to buy
www.ABBYY.com

5. CHAPTER 5

SIMULATION AND PERFORMANCE ANALYSIS

This chapter describes an evaluation of metadata based trustworthy Gene Feature Ranking (mtGFR) method for cancer classification using different types of datasets. With the help of mtGFR, the marker genes are known in advance and further a set of optimal set of genes are computed through biological validation process by using metadata. Then it is compared performance of different classification algorithms with and without feature selection by metadata based trustworthy gene feature ranking (mtGFR).

5.1 Experimental Setup and Dataset Description

The gene expression profiles of some specimens were generated in-house as a starting point for this study. To complement this analysis, public datasets were also downloaded from GEO (Gene Expression Omnibus) for comparison; as outlined. Several published datasets relating to breast cancer were downloaded from the GEO (Gene Expression Omnibus) (http://www.ncbi.nlm.nih.gov/geo/).

Table 5.1 : Cancer datasets of HG_U95Av2 platform used in the study

SN	Dataset	**Number of Classes**	**Number of Samples in the Dataset**
1.	ALL	2 (B-cell ALL and T-cell ALL)	128 (95 B-cell ALL and 33 T-cell ALL)
2.	Brain	1 (Tumor)	28 Tumor
3.	Breast	2 (Normal and tumor)	6 (1 normal and 5 tumor)
4.	Kidney	2 (Normal and tumor)	27 (21 normal and 6 tumor)
5.	Lung	2 (MPM and ADCA)	16 (8 MPM and 8 ADCA)
6.	Prostate	2 (Normal and tumor)	24 (12 Normal and 12 tumor)

Datasets from GEO carry a unique GEO ID and more information can be obtained by searching for the specified GEO number. For some experiments, gene expression values were available as raw data files, while for others they were available as processed data. These microarray datasets are: Acute Lymphoblastic leukemia cancer (ALL), Brain, Breast, Kidney, Lung, cancer and colon cancer. Table 5.1 and Figure 5.1 summarize the data sets.

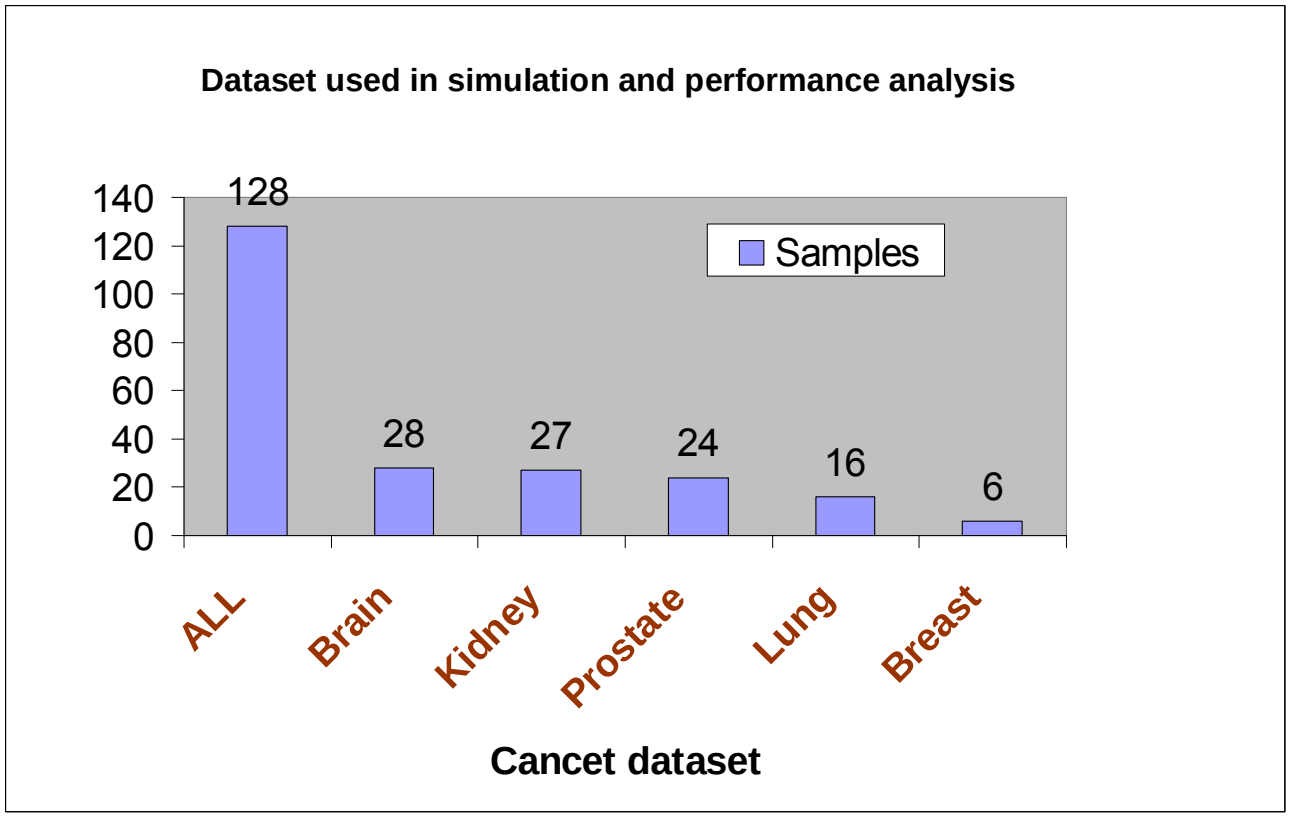

Figure 5.1 : Dataset used in simulation and performance analysis

5.2 Microarray Data Analysis

The dataset was retrieved in raw .CEL-format from the public repository Gene Expression Omnibus[16-20]. The CEL-files were subsequently processed using RMA on all the arrays of the patients for data simultaneously. Data was preprocessed and normalized using the robust multi-array average (RMA) expression measure. *RMA* is an expression measure obtained as a result of three pre-processing steps: background correction, quartile normalization and then a summarization based on a multi-array model fit robustly using the median polish algorithm.

The role of microarray experiments is often to test for regulation of tens of thousands of genes as an exploratory tool to derive candidate ranking lists of potentially regulated genes, which in subsequent steps will be biologically interpreted and validated by more precise techniques. The variation among

the dataset can me observed by intensity variation analysis (Figure:5.1), cluster dendogram (Figure: 5.2), heatmap graph (Figure: 5.3).

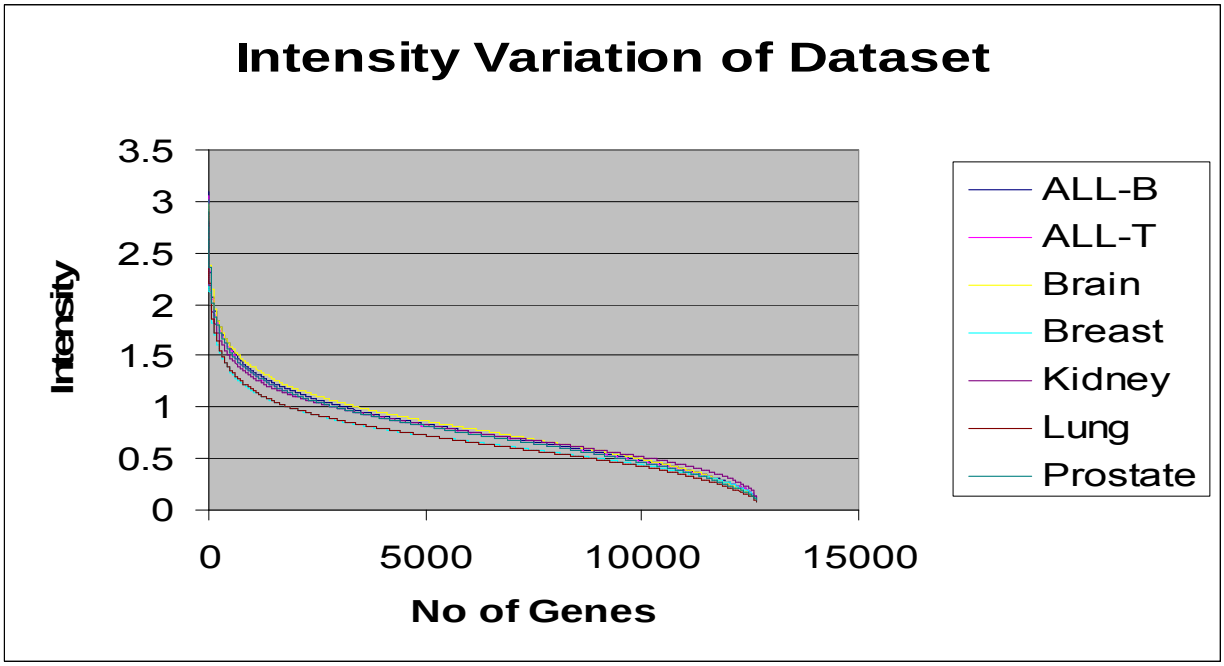

Figure 5.2 : Probe intensity of six dataset

From the probe intensity, it is observed that throughout the normalization process, the logarithmic intensity of the datasets are near to each other for further analysis. The variability of the dataset for classification is evaluated by cluster dendogram and heatmap that are very suitable for microarray data analysis. Those diagram are shown in figure 5.2 and figure 5.3 respectively.

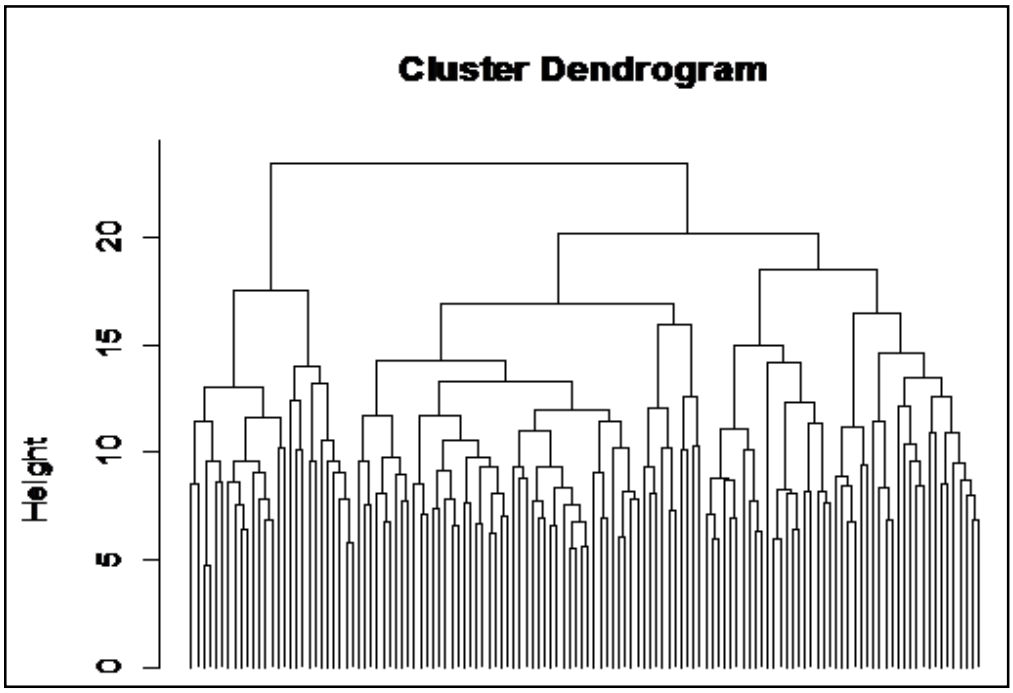

Figure 5.3 : Dendrogram analysis of six dataset

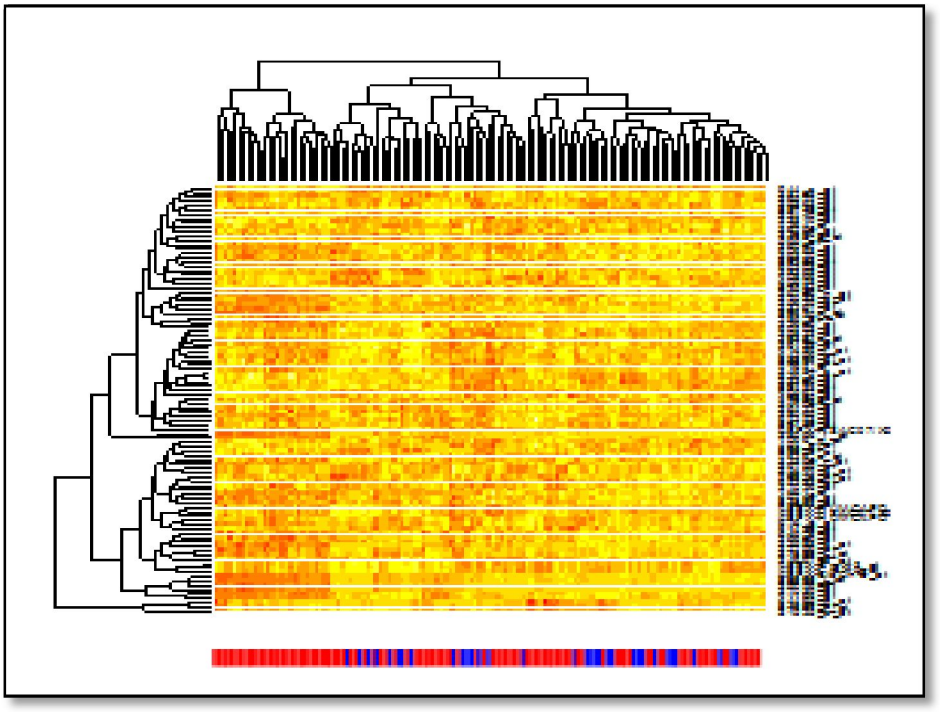

Figure 5.4 :Heatmap for ALL dataset

Analyzing the intensity variation analysis (Figure: 5.1), cluster dendogram (Figure: 5.2), heatmap graph (Figure: 5.3), it can be understood about the data variation of the experiment. In this thesis, our goal is to find out the biologically significant genes.

5.3 Gene Feature Ranking

Gene features are ranked based on their *covariance* measure which indicate more discriminative characteristics among the genes because these(genes) cannot be discriminated by expression intensity measure. the genes are ranked based on their covariance measure. Genes with higher covariance measure having a threshold are selected. Feature ranking is computed by the covariance of a gene is measured by evaluating a value called *covariance measure* ***(C_i)*** on preprocessed microarray dataset by the following equation:

$$C_i = \frac{\| \text{cov}_\alpha - \text{cov}_\beta \|}{\frac{\text{cov}_\alpha + \text{cov}_\beta}{2}}$$

where, cov_α and cov_β indicate the average covariance of the probe set.

Table 5.2 : Individual covariance of ALL-B and ALL-T with C_i

Gene Feature	CV of ALL-B	CV of ALL-T	C_i
33039_at	**0.145763**	**0.546125**	**1.157303**
38147_at	**0.290178**	**1.013103**	**1.109391**
36638_at	**0.782892**	**0.266824**	**0.983252**
39380_at	**1.316103**	**0.478088**	**0.934142**
36605_at	**1.307821**	**0.539424**	**0.831937**
32612_at	**0.870597**	**0.370645**	**0.805568**
1421_at	**0.223785**	**0.496206**	**0.756734**
39575_at	**0.153446**	**0.3357**	**0.745192**
31728_at	**1.173178**	**0.538884**	**0.740971**
36877_at	**0.527059**	**0.245436**	**0.729126**
39389_at	**1.091291**	**0.512287**	**0.72214**
33705_at	**0.599292**	**0.289341**	**0.697589**

Gene Feature	CV of ALL-B	CV of ALL-T	C_i
36502_at	0.565283	0.281968	0.668786
32168_s_at	0.702032	0.3527	0.662408
266_s_at	1.003146	0.507741	0.655781
34514_at	0.208497	0.411223	0.65425
33385_g_at	0.601202	0.305237	0.653029
33278_at	0.56762	0.289614	0.648613
……	……	……	……
……	……	……	……
40817_at	0.925961	0.926033	7.84E-05
38396_at	0.722368	0.722315	7.39E-05
37770_at	1.592563	1.592449	7.14E-05
38545_at	1.302629	1.302708	6.07E-05
38142_at	0.691654	0.691626	4.07E-05
37048_at	1.27455	1.2745	3.9E-05
1997_s_at	1.928329	1.928392	3.26E-05
AFFX-BioB-5_st	1.126051	1.126065	1.31E-05
39863_at	1.157021	1.157011	8.98E-06

5.4 Gene Filtering

During data preprocessing stage, it is to determine the filtering criteria to initially select the genes. The genes are selected through experiment and evaluated further. The particular decision function for filtering is assumed by analysing which genes are expressed for 50% and 60% patients.

Table 5.3 : Number of genes expressed for 50% and 60% samples

SN	Probe Intensity	IQR >	No of Gene expressed for 50% Sample	No of Gene expressed for 60% Sample
1.	200	0.8	858	703
2.	300	0.8	534	433
3.	400	0.8	384	304

SN	Probe Intensity	IQR >	No of Gene expressed for 50% Sample	No of Gene expressed for 60% Sample
4.	500	0.8	288	227
5.	600	0.8	221	173
6.	700	0.8	183	145
7.	800	0.8	197	119
8.	900	0.8	135	107
9.	1000	0.8	118	93

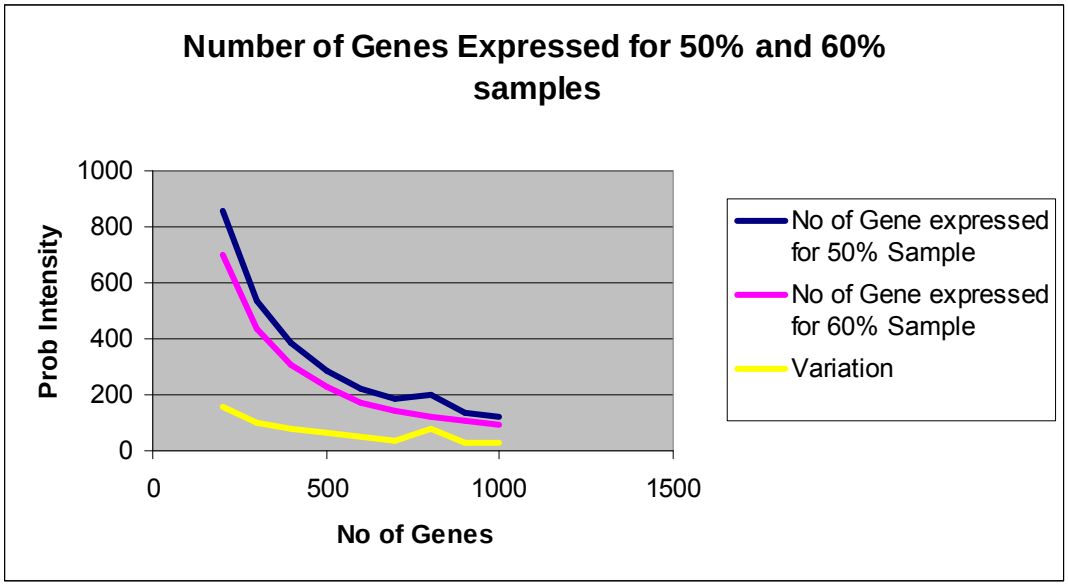

Figure 5.5 : Genes expressed for 50% and 60% samples based on intensity

The intensity of the samples are very near, so it is very difficult to distinguish the cancer classes based on intensity. But if covariance among the samples is considered then it may work for future classification techniques.

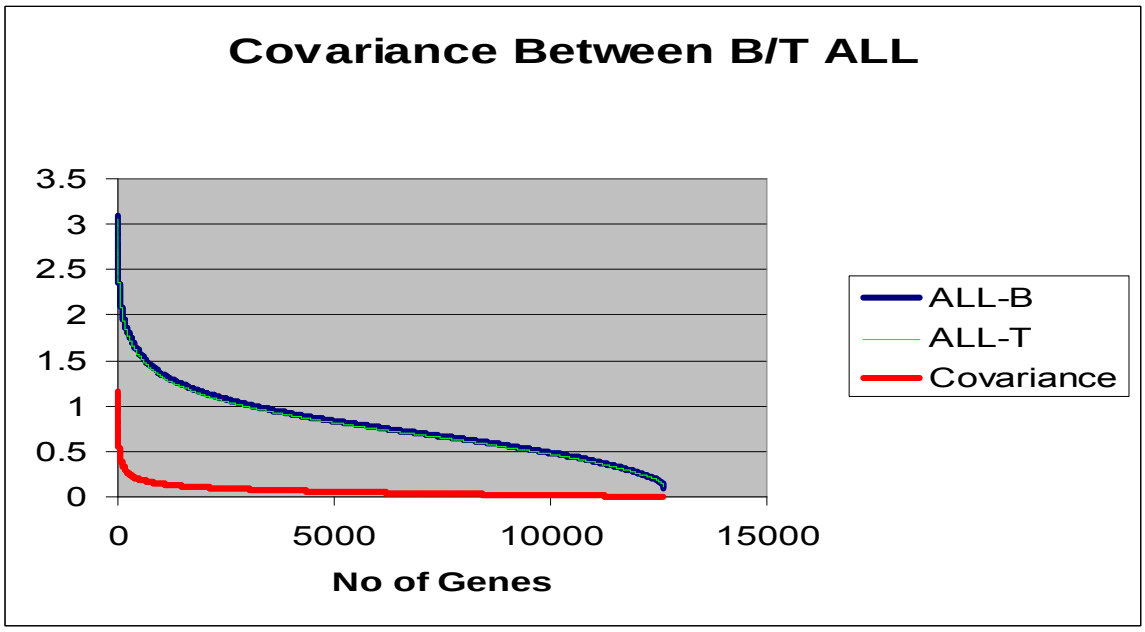

Figure 5.6 : Covariance of ALL subtypes

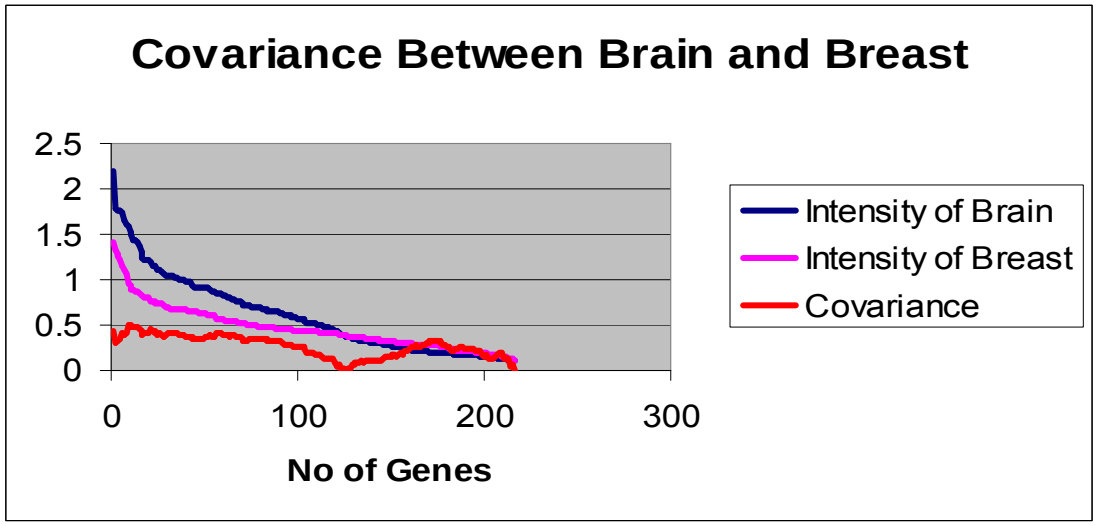

Figure 5.7 : Covariance between Brain and Breast dataset

So it is observed that the probe intensity expressed for 50% and 60% samples is very close. But the variation measure between them is significant. So in this research the following filtering criteria will be evaluated. Examined filter criteria in the research are as follows:

Table 5.4 : Examined filter criteria for 6 dataset

Filter Identifier	Probe Intensity	IQR >	% of Samples expressed
F1	300	0.8	50
F2	400	0.8	50
F3	500	0.8	50
F4	300	0.8	60
F5	400	0.8	60
F6	500	0.8	60

The filters are considered so that it may be analyzed what factor is the effect of expression over percentage of samples with the increasing the probe intensity.

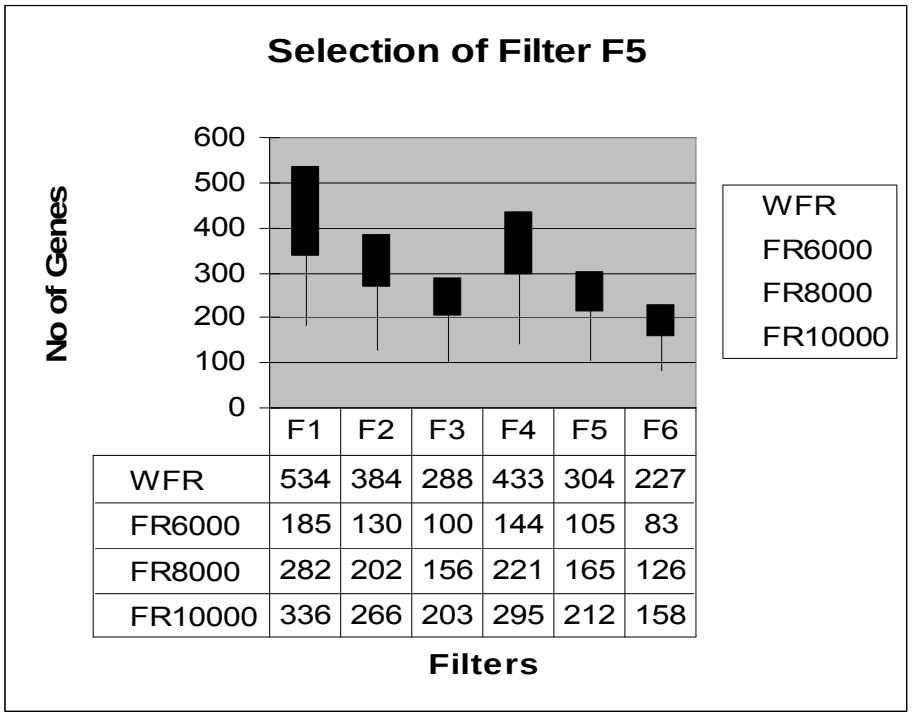

Figure 5.8 : Results for six datasets using F1-F6 filters

5.5 Selection of genes by PCA and Attribute Selection technique

To compare with attribute selection method for microarray data with PCA following table shows that attribute selection measure is better performed for the dataset. The table5.4 and 5.5 are given below.

Table 5.5 : Selection of genes by PCA and mtGFR on six dataset

Filter	% Sample Expressed	Intensity of Expression	IQR Range	Without GFR	PCA	PCA Using Gene Feature ranking					
						M6000	PCA	M8000	PCA	M10000	PCA
F1	0.5	300	0.8	534	65	185	49	282	56	336	60
F2	0.5	400	0.8	384	58	130	42	202	50	266	54
F3	0.5	500	0.8	288	51	100	35	156	43	203	47
F4	0.6	300	0.8	433	62	144	45	221	52	295	57
F5	0.6	400	0.8	304	53	105	37	165	46	212	49
F6	0.6	500	0.8	227	47	83	32	126	40	158	43

Table 5.6 : Selection of genes by attribute selection on six dataset

Filter	% Sample Expressed	Intensity of Expression	IQR Range	Without GFR	Attribute Selection	Attribute Selection Using Gene Feature ranking					
						M6000	Attribute Selection	M8000	Attribute Selection	M10000	Attribute Selection
F1	0.5	300	0.8	534	57	185	40	282	33	336	47
F2	0.5	400	0.8	384	45	130	39	202	28	266	42
F3	0.5	500	0.8	288	43	100	38	156	26	203	48
F4	0.6	300	0.8	433	51	144	31	221	32	295	42
F5	0.6	400	0.8	304	44	105	37	165	28	212	34
F6	0.6	500	0.8	227	44	83	37	126	40	158	44

5.6 Result for classification performance

Table 5.7 : Classification performance of F1-F6 filters

SN	Classification Algorithm	F1	F2	F3	F4	F5	F6	Average
1.	Bayes Net Classifiers	73	72	72	73	72	70	72
2.	Naive Bayes	68	67	69	67	68	68	68
3.	IB1	100	100	100	100	100	100	100
4.	KStar	100	100	100	100	100	100	100
5.	BF Tree	65	71	78	72	83	73	74
6.	FT	95	95	95	85	96	94	93
7.	J48 (C4.5)	95	95	95	85	95	95	93
8.	Average	85	86	87	83	88	86	86

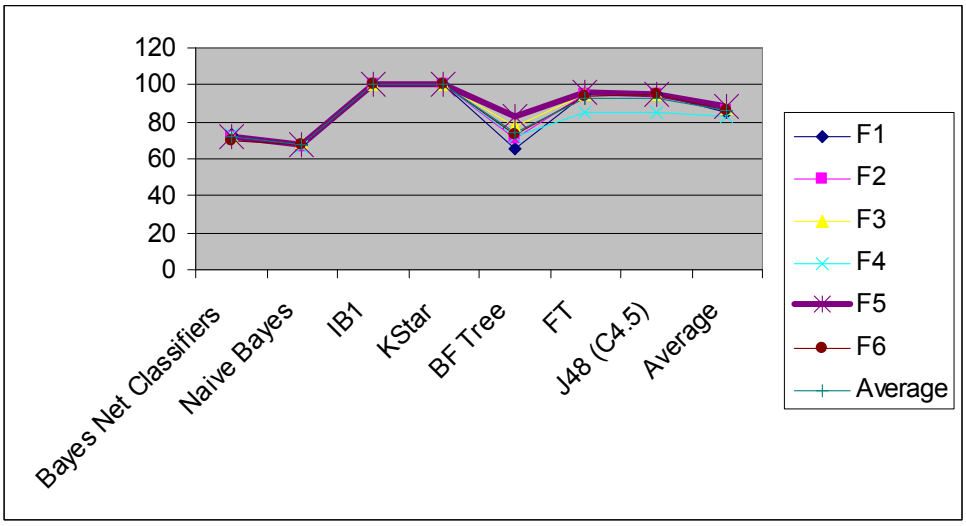

Figure 5.9 : Classification performance of filters

Table 5.8 : Time complexity of F1-F6 filters

SN	Classification Algorithm	F1	F2	F3	F4	F5	F6	Average
1.	Bayes Net Classifiers	0.023	0.015	0.015	0.020	0.020	0.010	0.017
2.	Naive Bayes	0.005	0.000	0.005	0.000	0.000	0.005	0.003
3.	IB1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4.	KStar	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5.	BF Tree	0.200	0.173	0.173	0.173	0.138	0.178	0.172
6.	FT	0.480	0.335	0.298	0.323	0.220	0.310	0.328
7.	J48 (C4.5)	0.045	0.035	0.033	0.033	0.060	0.045	0.042
8.	Average	0.108	0.080	0.075	0.078	0.063	0.078	0.080

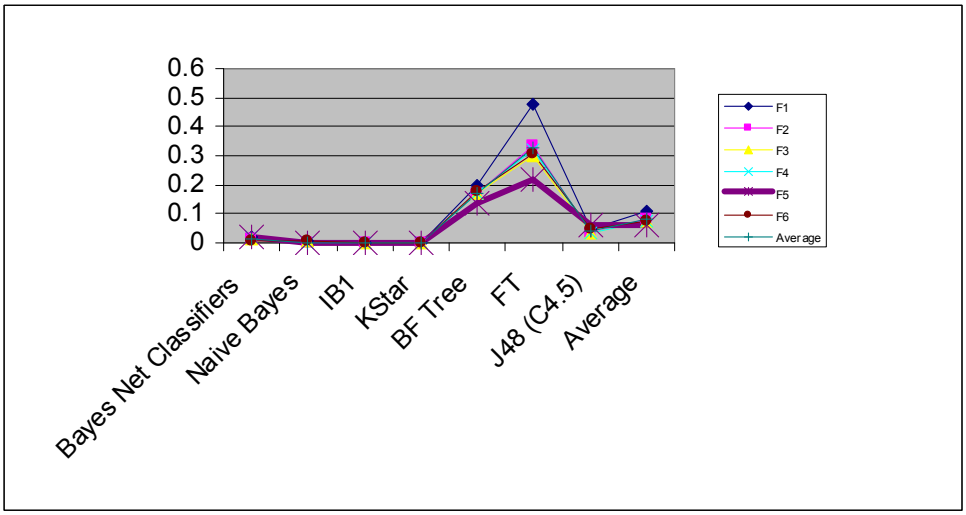

Figure 5.10 : Time complexity of F1-F6 filters

5.7 Experiment result for ALL Dataset

Table 5.9 : ALL selection of gene features for classification

SN	% Sample Expressed	Intensity of Expression	IQR Range	Without Gene Feature ranking	Attribute Selection	Attribute Selection Using Gene Feature ranking					
						M6000	Attribute Selection	M8000	Attribute Selection	M10000	Attribute Selection
F1	0.5	300	0.8	300	30	141	21	185	25	225	27
F2	0.5	400	0.8	186	25	91	16	115	20	138	22
F3	0.5	500	0.8	123	17	66	13	79	15	90	15
F4	0.6	300	0.8	229	26	104	18	135	20	164	23
F5	0.6	400	0.8	139	19	71	13	88	15	101	16
F6	0.6	500	0.8	99	15	50	13	60	14	68	14

Table 5.10 : Performance comparison of different classification algorithms

Name of Classification Algorithm	Accuracy (%)	
	Without metadata integration	Using metadata integration
BayesNet	99.22%	100%
Logistic-R	100%	100%
KStar	100%	100%
AdaboostM1	100%	100%
Multiboost AB	100%	100%
Decision Table	100%	100%
FT	100%	100%
Classification Via Regression	100%	99%
LWL	99.22%	99.23%
Attribute Selected Classifier	100%	99.22%
VFI	99.21%	99.21%
PART	100%	99.22%
BFTree	96.88%	99.20%
J48	100%	99.23%
Naïve Bayes with 55% splitting	100%	100%
Naïve Bayes	92.97%	96%

5.8 Significant Probe ID for ALL and Mixed cancer classification

Table 5.11 : Significant probe identified for ALL cancer classification

SN	Probe ID
1.	1202_g_at
2.	33238_at
3.	39318_at
4.	40116_at
5.	41215_s_at

Table 5.12: Significant probe identified for six dataset by F5

ProbeID	Gene_ID	MetaRanking
32321_at	3133	32
595_at	7128	31
41833_at	10899	31
649_s_at	7852	31
1612_s_at	3727	31
36224_g_at	6421	30
32378_at	5315	30
286_at	8337	29
286_at	723790	29
1836_at	10983	29
41185_f_at	6613	29
35830_at	23214	27
32052_at	3043	20
39110_at	1975	20
691_g_at	5034	18
31527_at	6187	17
39027_at	1327	16
38485_at	4717	14
38590_r_at	5757	12
41724_at	10134	9
32843_s_at	2091	7

ProbeID	Gene_ID	MetaRanking
37759_at	7805	6
1173_g_at	6303	4
911_s_at	805	4
911_s_at	285051	4
31444_s_at	305	4
31444_s_at	303	4
31444_s_at	302	4
34091_s_at	7431	4
503_at	5441	0

5.9 Biological validation through metadata analysis

Throughout the biological relation and metadata integration, an optimal set of genes are identified for the mixed data of six cancer data. Though, for the mixed data, some classification algorithm shows lower accuracy, but it shows better performance for single cancer dataset lke ALL. Now the gene will be integrated with metadata how a gene is interrelated with different types of proteins, diseases, drugs, biological pathways. In this case, it is used PubMed Alibaba interface to examine the biological validation for each gene (Figure 5.9).

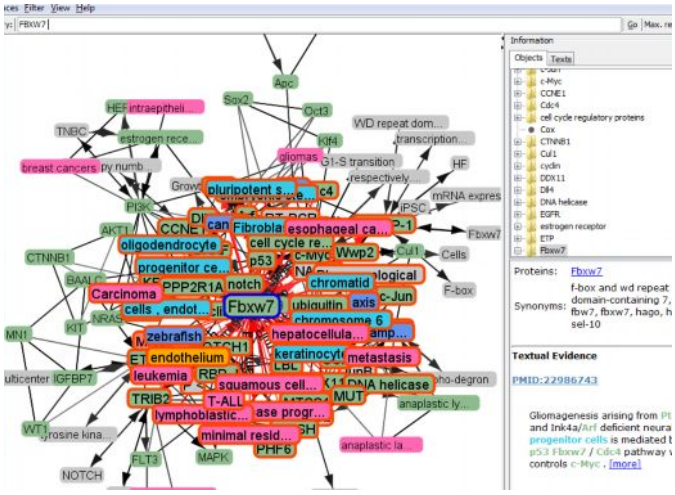

Figure 5.11: Biological relation of gene Fbxw7

The proposed method identifies the genes that are related to cancers. The sequence analysis of the genes can evaluated for which protein it codes. And if the protein is related with others, they also

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

cause cancers. Inter organism relation can also be evaluated by this method. In future, the performance of the classification for mixed data will be optimized by the help of metadata and an embedded smart system will be manufactured by the help of smart classifiers like metadata based trustworthy gene feature ranking method.

ABBYY PDF Transformer 2.0
Click here to buy
www.ABBYY.com

6. CHAPTER 6

CONCLUSION

6.1 Summary of Research

In this research, a novel approach called *Cancer Classification Method by Integrating Metadata in Microarray Data Analysis* is proposed using metadata based trustworthy Gene Feature Ranking (mtGFR). The method works with high trust values that is selected from a list of genes with high dimensionality and high metadata rank. By comparing to a suitable pre-defined threshold, only those genes that show sufficient *trustworthiness* to be considered for constructing the classification model are retained. Besides a biological relation score is proposed in this work to validate marker gens biologically. Through experimentation involving six dataset and two experiments, this model is useful for embedded system design for cancer identification in medical science in future.

6.2 Future Works

In future work, The improve of the accuracy of classification by incorporating domain specific knowledge including information obtained from relevant literature, gene symbol, nucleotide sequence, and protein databases that were maintained by National Center for Biotechnology Information (NCBI). One such database would be the Gene Ontology (GO) database where the categorization of gene product is done based on their associated biological processes, molecular functions, and cellular components. By using these, a smart classifier will be designed so that an embedded system can be developed to general purpose to diagnosis different types of cancers. In future, the model may be reevaluated with more set of cancer data. In this work the relation score is assumed to be straight forward for equal score for all type of relation. In future the score may be evaluated dynamically with the change of metadata measure and new invention of genes, diseases, protein, drug, cells and species.

The work is done on dataset extracted from HGU95av2 platform of microarray experiment. The other platforms should be integrated for improving the cancer classification in heterogeneous platform with more dataset. In that case the normalization process of the mixed dataset will be studied for better performance in classification.

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

In the future work, the null effect of metadata ranking will be generalized to integrate so that the biological significance of the gene whose known functions have not been discovered yet cannot bias the feature selection and classification process . If the gene with $M_i=0$ having higher covariance measure, then it may be an interesting genes for further biological research. How those genes should be integrated instead of nullifying the effect will be evaluated in future work.

REFERENCES

[1] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C.,Kobayashi, M., Horton, H., Brown, E.L., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1675–1680.

[2] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davis, R.W., 1996. Parallel human genome analysis:microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA 93, 10614–10619.

[3] Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

[4] Nguyen, D.V., Arpat, A.B., Wang, N., Carroll, R.J., 2002. DNA microarray experiments: biological and technological aspects. Biometrics 58, 701–717.

[5] Wolf, L., Shashua, A., et al. (2009). Selecting relevant genes with a spectral approach (No. CBCL Paper No.238). Cambridge, MA, USA: Massachusetts Institute of Technology.

[6] Hall MA, Correlation based feature selection for machine learning,, UW of Hamilton, NewZealand, 2012

[7] HIoskuldsson, A., 1988. PLS regression methods. J. Chemometr. 2, 211–228.

[8] Inza, I., Larranaga, P., et al. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence, 31(2), 91-103.

[9] Cooper, G.M. (2000). The Cell - A Molecular Approach, 2nd ed., Sinauer Associates Inc, Sunderland, Massachusetts

[10] Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., and Lewontin, R.C.. (1999). Modern Genetic Analysis, Freeman, New York.

[11] Lehninger, A.L., Nelson, D.L., and Cox, M.M. (2000) Principles of Biochemistry, Worth Publishing, 3rd ed.

[12] Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., and Gelbart, W.M. (2000). An Introduction for Genetic Analysis, Freeman, New York.

[13] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. (2002). Molecular Biology of the Cell, Garland Publishing, 4th ed.

[14] Dale, J.W. and von Schantz, M. (2002). From Genes to Genomes: Concepts and Applications of DNA Technology. John Wiley and Sons, Ltd, England.

[15] Microarray data repositories available at http://www.ncbi.nlm.nih.gov/geo

[16] Microarray data repositories available at http://smd.stanford.edu

[17] Microarray data repositories available at http://www.ebi.ac.uk/arrayexpress

[18] KEGG repository available at www.genome.jp/kegg/

[19] Draghici, S., Kulaeva, O., et al. (2003). Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics, 19(11), 1348-1359.

[20] Efron, B., Tibshirani, R., et al. (2001). Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association, 96, 1151-1160.

[21] Lee, K. E., Sha, N., et al. (2003). Gene selection: a Bayesian variable selection approach. Bioinformatics, 19(1), 90-97.

[22] Tibshirani, R. J. (2006). A simple method for assessing sample sizes in microarray experiments. BMC Bioinformatics, 7(106).

[23] Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS, 99(10), 6562-6566.

[24] Devijver, P. A., & Kittler, J. (1982). Pattern recognition: a statistical approach. London: Prentice-Hall Inc.

[25] Golub T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, (286):531–537, 1999.

[26] Yang YH, Nucleic Acids Res 3, e15, 2012

[27] Bolsted BM et al, Bioinformatics Quartiles, 2009

[28] Dudoit, S., Fridlyand, J., et al. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data: UC Berkeley.

[29] Affymetrix probe standard and expression measurement available at www.affymetrix.com/support/technical/manuals.affx, 1999

[30] Affymetrix probe standard and expression measurement available at www.affymetrix.com/support/technical/manuals.affx, 2001

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

[31] Calculation of Model-Based Expression Intensities (MBEI) standard available at http://biosun1.harvard.edu/complab/dchip/manual.htm

[32] Scheetz T, Microarray Analysis using RMA, 2005

[33] Ling M and Lee T, Statistical Analysis and consistent feature selection from Microarray Gene Expression Data, Klurwer Academic Publishers, New York, 2012

[34] Ding, C., & Peng, H. (2003). Minimum Redundancy Feature Selection for Gene Expression Data. Paper presented at the Proc. IEEE Computer Society Bioinformatics Conference (CSB 2003), Stanford, CA.

[35] Furey, T., Cristianini, N., et al. (2009). PCA and Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906-914.

[36] Jaeger, J., Sengupta, R., et al. (2003). Improved gene selection for classification of microarrays. Paper presented at the Pacific Symposium on Biocomputing, Kauai, Hawaii.

[37] Eisen, M. B., Spellman, P. T., et al. (1998). Cluster analysis and display of genomewide expression patterns. Paper presented at the Proc Natl Acad Sci USA.

[38] Mramor M et al. Visualization based cancer microarray data classification analysis, 2007

[39] Fan J et al, Clustering Analysis of DNA Microarray Data in Cancer Research, 2012

[40] Hasan A et al. Cancer Classification from Microarray Data using Gene Feature Ranking, International Journal of Data Mining and Emerging Technologies, Vol. 1 No.2, November, 2011, 54-60

[41] Bogdan D, Purvesh K, Arina D, and Sorin A, Predicting Novel Human Gene Ontology Annotations Using Semantic Analysis in Microarray Data, IEEE/ACM Transactions on computational Biology and Bioinformatics, Vil 70, No. 1, Jan-Mar 2012

[42] Cheng J., Sun S., Tracy A., Hubbell E., Morris J., Valmeekam V., Kimbrough A., Cline,M.S., Liu,G., Shigeta,R. et al. (2004) Gene feature selection for microarray data analysis. Bioinformatics, 2010, 1462–1463.

[43] Dahlquist K. D., Salomonis N., Vranizan K., Lawlor S.C. and Conklin B.R. (2002), A new technique for reduction gene features of microarray data.

[44] Mlecnik B., Scheideler M., Hackl H., Hartler J., Sanchez-Cabo F. and Trajanoski Z. (2005) Classification and Prediction of medical conditions for cancer. Nucleic Acids Res., 33, W633–W637.

[45] Al-Shahrour F., Diaz-Uriarte R. and Dopazo J. (2004) FatiGO: a tool for finding significant associations and classification of Gene Ontology terms with SVM. Bioinformatics, 20, 578–580.

[46] Chung H.J., Park C.H., Han M.R., Lee S., Ohn J.H., Kim J. and Kim J.H. (2005) ArrayXPath II: mapping and visualizing micro-array gene-expression data with high dimension. Nucleic Acids Res., 33, W621–W626.

[47] Pandey R., Guru R.K. and Mount D.W. (2004) Pathway miner: extracting gene association networks from molecular pathways for classifying and predicting the biological significance of gene expression microarray data. Bioinformatics, 20, 2156–2158.

[48] Goffard N. and Weiller G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. Nucleic Acids Res., 35, W176–W181.

[49] Gene expression and sequence data available at http://www.ncbi.nlm.nih.gov/pubmed

[50] Robert et al, Bioinformatics and Computational Biology Solution using R and Bioconductors, Springer 2010

[51] Jesmin, Rashid, Gene regulatory network reveals oxidative stress as the underlying molecular mechanism of type 2 diabetes and hypertension, 2012

[52] Keith Cheverst, Decision Tree construction using IG, 2009.

[53] K. Y. Yeung andW. L. Ruzzo, “Principal component analysis for clustering gene expression data,” Bioinformatics, vol. 17, no. 9, pp. 763–774, 2001.

[54] Chao S and Lihui C, Feature dimension reduction for microarray data analysis, 2004

[55] Hall MA, Correlation based feature selection for machine learning, New Zealand

[56] H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In Proceedings of the Ninth National Conference on Artificial Intelligence, pages 547–542. MIT Press, 1991.

[57] R. Setiono and H. Liu. Chi2: Feature selection and discretization of numeric attributes. In Proceedings of the Seventh IEEE Conference on Artificial Intelligence, 2012.