# GENE SELECTION BY 1-D DISCRETE WAVELET TRANSFORM

# FOR CLASSIFYING CANCER SAMPLES USING DNA MICROARRAY DATA

A Thesis

Presented to

The Graduate Faculty of The University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Adarsh Jose

May, 2009

GENE SELECTION BY 1-D DISCRETE WAVELET TRANSFORM

FOR CLASSIFYING CANCER SAMPLES USING DNA MICROARRAY DATA

Adarsh Jose

Thesis

Approved:

_____
Advisor
Dr. Dale H. Mugler

_____
Co-Advisor
Dr. Zhong-Hui Duan

_____
Committee Member
Dr. Daniel B. Sheffer

Accepted:

_____
Department Chair
Dr. Daniel B. Sheffer

_____
Dean of the College
Dr. George K. Haritos

_____
Dean of the Graduate School
Dr. George R. Newkome

_____
Date

ABSTRACT

Selecting a set of highly discriminant genes for biological samples is an important task for designing highly efficient classifiers using DNA microarray data. The wavelet transform is a very common tool in signal processing applications, but its potential in the analysis of microarray gene expression data is yet to be explored fully.

In this thesis, a simple wavelet based feature selection method is presented that assigns scores to genes for differentiating samples between two classes. The term 'gene expression signal' is used to refer to the gene expression levels across a set of pre-grouped samples. The expression signal is decomposed using several levels of the wavelet transform. The scoring method is based on the observation that the third level 1-D wavelet approximation of a gene expression signal captures the differential expression levels of the gene between two classes. The genes with the highest scores are selected to form a feature set to be used for sample classification. The method was implemented using MATLAB®. Experiments based on three real microarray gene expression datasets were carried out to examine the efficiency of the method. The classification performance of the method was compared to two standard filter based methods: the t-test and BSS/WSS methods using the 3-Nearest Neighbor Classifier. The results show that the wavelet-based method performs at least as well as the sum of squares and the wavelet based method in classifying cancer samples.

The results demonstrate that 1-D wavelet analysis can be a useful tool for studying gene expression patterns across pre-grouped samples.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my co-advisors Dr. Dale Mugler and Dr. Zong-Hui Duan. They have been a great source of knowledge and inspiration throughout the course of the thesis. Their unwavering support, timely suggestions and untiring advices helped to give a direction to my efforts and eventually take the theses to a successful completion.

Dr. Daniel Sheffer was very supportive and helpful as the third committee member. I am very thankful to him for his suggestions during the thesis meetings. They were of immense help and benefit during the course of the thesis work.

I would also like to extend my gratitude to the rest of the faculty of the Biomedical Engineering Department, Dr. Igor Tsukerman from the Department of Electrical Engineering and Dr. Amy Milsted from the Department of Biology for making my life at the University of Akron a wonderful experience.

Finally, I would like to thank my family, and friends who have stood by me at various phases during the last two and a half years. Their prayers and moral support was extremely crucial in making this thesis a success.

TABLE OF CONTENTS

Page

LIST OF TABLES……..……..….. ix

LIST OF FIGURES……..……..……...……..… x

CHAPTER

I. INTRODUCTION……...…….……..…..... 1

1.1. Overview……...…….. 1

1.2. DNA microarrays……..……..…… 2

1.3. Gene expression studies in cancer……...……..... 4

1.4. Classification using gene expression data……..... 4

1.5. Feature selection……..... 5

1.6. Feature extraction……..... 5

1.7. Wavelet transforms……..... 6

1.8. Hypotheses……..... 7

II. LITERATURE REVIEW……...…….. 8

2.1. Microarrays in biology……...……..... 8

2.2. Noise removal and normalization…….…..... 8

2.3. Microarray databases……..…… 9

2.4. Classification…….…… 9

2.5. Feature selection……..……..... 10

2.6. Wavelet transform in microarray studies…….…….. 12

III. DNA MICROARRAYS…….…….. 13

3.1. The cell……..…..... 13

3.2. Genes…….……. 13

3.3. Introduction……..……..... 14

3.4. Gene expression……..……. 14

3.5. The need for dna microarrays……..……..... 16

3.6. Monitoring gene expression……..……... 16

3.6.1. Hybridization……..…… 17

3.6.2. Microarray technology platforms……..….. 18

3.6.3. Selecting and synthesizing the probes…….…….. 19

3.6.4. Manufacturing the arrays…….…….. 20

3.6.5. Labeling the targets…….… 20

3.6.6. Scanning and analyzing the images obtained…….… 22

3.6.7. Microarray data……... 22

IV. FEATURE SELECTION IN MICROARRAY BASED CLASSIFICATION…..…...……. 23

4.1. Filter methods……..... 24

4.1.1. Parametric methods…….… 24

4.1.2. Non-parametric methods…….… 28

4.2. Multivariate methods for gene selection…….… 29

4.2.1. Correlation based feature selection……..... 30

4.2.2. Minimum redundancy maximum relevance method…… …….……... 30

4.2.3. Uncorrelated shrunken centroid based method……. 31

4.3. Wrapper methods……..... 31

V. WAVELET TRANSFORM……...……. 33

5.1. The continuous wavelet transform (cwt)… ……..……..... 35

5.2. The discrete wavelet transform (dwt) ……..……..... 36

5.2.1. Multi resolution analysis…… 37

5.2.2. Scaling and detail coefficients……..... 39

5.3. Fast wavelet transform …… 40

5.4. Smoothing using wavelets…… 42

VI. K-NEAREST NEIGHBOR CLASSIFIER…….... 44

VII. THE METHOD……..…… 46

7.1. Preprocessing…… 46

7.2. The gene selection method……... 47

7.3. Application……... 50

7.4. Experiments…….…… 50

7.4.1. Classification performance of the different methods……. 51

7.4.2. Shuffling test……... 52

7.4.3. Gene study…….…….. 52

VIII. RESULTS AND DISCUSSION……...……...……..... 53

8.1. Classification results……...…….…… 53

8.2. Gene study……...…… 60

8.2.1. Leukemia data……. 61

8.2.2. Lymphoma data…….…….. 61

8.2.3. Colon data……... 62

8.4. Some observations …….…….. 63

8.5. Limitations……..……..... 64

8.6. Significance of the Study...…….……. 64

8.7. Future works…….…..... 64

IX. CONCLUSIONS……... 65

BIBLIOGRAPHY……... 67

APPENDICES…….…. 74

APPENDIX A. STATISTICAL ANALYSIS FOR NULL HYPOTHESES……. 75

APPENDIX B. STATISTICAL ANALYSIS OF THE CLASSIFICATION RESULTS…..……. 79

APPENDIX C. GENE LISTS…… 90

LIST OF TABLES


| Table | | Page |
|---|---|---|
| 3.1 | The genetic code……..…….. | 15 |
| 4.1 | The summary of t-test and its modifications…….....……..... | 26 |
| 7.1 | Illustration of a typical microarray dataset…….. | 46 |
| 7.2 | A sample of the preprocessed and normalized data-points from the leukemia dataset..... | 47 |
| 7.3 | The Datasets studied with the two classes of samples and the Number of Samples in each Class…….....…….... | 50 |
| 7.4 | The Confusion matrix…….....…….....…….. | 51 |
| 8.1 | Division of Classes into Positive and Negative for estimating the Sensitivity and Specificity…….....…….....…….. | 58 |
| 8.2 | The top 10 genes from the three datasets identified by the Db-8 wavelet with their scores…….……..…….. | 61 |
| 8.3 | Number of genes in the list of top 100 genes, common to different methods……..…..…. | 62 |
| 8.4 | The genes from the top 100 list common for all the three gene selection method studied……..…….. | 63 |

LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| 1.1 | A 1,024 point signal decomposed to level 5 using the Sym 4 wavelet……..... | 6 |
| 3.1 | The three molecular genetics processes.… ……..……..... | 14 |
| 3.2 | Three labeling methods used for target labeling- Direct, Using Aminoallyl nucleotides and in vitro transcription based method…….. | 21 |
| 5.1 | Some commonly used wavelets……..... | 34 |
| 5.2 | Continuous wavelet transform of the sine wave with different frequencies using a Mexican hat wavelet……... | 35 |
| 5.3 | Multi Resolution Analysis of a highly noisy signal using db3 wavelet transform decomposed to the 3rd level…….. | 39 |
| 5.4 | The schematic of signal filtering and reconstruction……... | 41 |
| 5.5 | Wavelet smoothing operation……..…… | 42 |
| 5.6 | The scaling and wavelet functions of Db8 (Left) and Coif3 (Right) wavelets with their corresponding filter coefficients……... | 43 |
| 6.1 | Illustration of 3-NN classifiers…...…….. | 45 |
| 7.1 | The plot shows the gene scoring process for gene CST3 (Cystatin C)…..…….. | 49 |
| 7.2 | The plot illustrates the original expression signal of an informative gene LDHA (Lactate dehydrogenase A) (a) and its 3rd level approximation obtained using db-8 wavelet (b)……. | 50 |
| 8.1 | Mean Sensitivity for the different methods for the B-cell Lymphoma Dataset…….. | 53 |
| 8.2 | Mean Sensitivity for the different methods for the Leukemia Dataset…… | 54 |
| 8.3 | Mean Sensitivity for the different methods for the Colon Cancer Dataset…….. | 54 |
| 8.4 | Mean Specificity of the different methods for the B-cell Lymphoma Dataset…….. | 55 |
| 8.5 | Mean Specificity of the different methods for the Leukemia Dataset……. | 55 |
| 8.6 | Mean Specificity of the different methods for Colon Cancer Dataset…… | 56 |

8.7 Mean classification Accuracy of the different methods for the B-cell Lymphoma Dataset…….. 56

8.8 Mean classification Accuracy of the different methods for the Leukemia Dataset…….…… 57

8.9 Mean classification Accuracy of the different methods for Colon Cancer Dataset…… 57

# CHAPTER I

# INTRODUCTION

## 1.1. Overview

The most important challenge facing cancer biology is the identification of distinct tumor sub-types and development of specific therapies for these sub-types, which will maximize efficiency and minimize toxicity (Golub T, 1999).

Cancer is a class of diseases caused by the buildup of genetic and epigenetic changes, which results from alterations of sequences or expression of cancer-related genes like oncogenes, tumor suppressor genes, those involved in cell cycle control, adhesion, apatosis, DNA repair and angiogenesis (Squire, 2002). The traditional methods of cancer classification based on morphological appearance (shape and size) of the tumors have serious limitations, as tumors having very similar appearances can be of different sub-types and might respond completely differently to the same therapy.

DNA microarrays provide us a completely different perception of the different classes of cancer, one based on gene expression profile instead of morphological characteristics. For each sample, it gives a vector of expression levels of thousands of genes at the time of sample preparation. The gene expression profile, being a snap-shot of the functional state of the tissue being studied at the time of sample preparation, can characterize the different classes of cancer –normal from cancerous, one stage from the next, one sub-type from another – based on genomic characteristics which are the true causes of the changes rather than morphological features which are just "some of the effects" of the changes. Besides, it allows the application of systematic and unbiased pattern recognition techniques in place of the earlier methods, which depended on biological insights for classifying cancer (Golub T, 1999).

In pattern recognition, in order to design an unbiased and low variance classifier which performs well over the entire population, we need the sample size to be very large when compared to the number of variables. However, at least for now, the available sample size is very small compared to the number of variables involved in gene expression data. This raises several critical issues regarding design of classifiers using microarrays (Edward R. Dougherty, 2005) of which a method to somehow select a small subset of features or variables (genes) from the entire variable space which can classify accurately, not only the training set, but also the entire population.

The aim of this thesis is to explore the potential of using the 1-Dimensional Wavelet Transform as a feature selection (gene selection) tool. The idea is based on the hypothesis that a gene is informative for a given classification problem if it is differentially expressed in the different classes being studied. The method was tested on three publicly available Affymetrix datasets (Golub T 1999) (Shipp 2002) (Alon U. 1999).

The absolute value of the difference of mean between the classes of the signal recreated from the 3$^{rd}$ approximation of the 1-D Discrete wavelet transform is used as a score for ranking the genes, one at a time. The few highest ranked genes were used for classification and the feature sets tested using their classification accuracy.

Different wavelets were tried for the two-class problem. Db-3, Db-8 and Coiflet-3 wavelets gave the best classification results. The selected features were used to classify the test samples using 3-Nearest Neighbor classifiers.

The results show that the wavelet-based method out-performs two most commonly used filter based gene selection methods - the T-test method and the BSS/WSS methods in many cases tested, particularly when a higher number of features are selected to construct the classifiers.

## 1.2. DNA microarrays

DNA microarray chips allow simultaneous monitoring of the expression of thousands of genes in clinical specimens. This gives us huge sets of data points, ( The oligonucleotide microarrays give

approximately 16,000 data-points for each sample studied) which combine together to form a sort of molecular fingerprint of the state of the samples being studied.

The microarray technology is based on two basic principles: (1) Nucleotide sequences tend to hybridize to their complementary base pairs. (2) The differential levels of mRNA in the samples represent differential levels of gene expression.

The mRNA extracted from the tissue samples are hybridized onto tagged nucleotides to form tagged cDNAs. These in turn are hybridized onto an array of spotted cDNAs (cDNA microarrays which provide relative gene expression between specific cells or tissue samples) (Schena M, 1995) or oligonucleotide probes (Affymetrix Gene chips which give direct quantification of mRNA expression) (Lipshutz RJ, 1999). The excess cDNAs are washed away and the chips are scanned with laser beams which are characteristic of the fluorescent tags used.

The images obtained are processed and normalized to obtain a single number, the level of gene expression, for each gene on the array. Dudoit et al. (Yang, 2002) provides very good discussion on the normalization procedures and some of the important issues to be considered when designing normalization procedures.

Carrying out microarray assay requires special skills and it is not easy to obtain cancer tissue samples for analysis. With ever-increasing interest in the microarray dataset based studies, it becomes important that researchers, who do not have molecular biology laboratory skills or simply do not have the funds to carry out these experiments, have access to these datasets in a standard format. This has lead to the formation of microarray databases, which are maintained by major research labs like the Stanford Microarray Database and other agencies like the Gene Expression Omnibus (GEO). (Christopher J. Penkett and Jurg Bahler, 2004) A typical dataset will be grouped into training sets and testing sets having the samples listed along the columns and the genes along the rows of a 2 dimensional matrix . Each cell represents the expression of gene g listed along the row with its corresponding n th sample listed along the columns.(A screenshot of a typical dataset is shown in Chapter 7). There will also be a .cls class file carrying the correct class labels of the samples.

## 1.3. Gene expression studies in cancer

In his landmark paper on class discovery and classification of cancer (Golub T, 1999), T.R.Golub et.al. has explained the importance of discovery of subclasses in cancer. The morphology of the tumor, which is currently the most important feature for classification, has been proven to have serious limitations. Several cases were identified where the tumors, supposedly in the same class, responded differently to therapy. In that paper, he proposed for the first time, gene expression monitoring using DNA microarrays as a possible tool for class discovery and further classification in cancer. This paper created a completely new area of functional genomics, called cancer genomics, which deals with using the gene expression profile as a tool in cancer diagnosis and therapy. Several research papers followed which applied gene expression monitoring to address cancer classification, discovery of tumor specific biomarker and studying drug sensitivity. Easier availability of the gene expression datasets has allowed researchers from a wide spectrum of fields including biology, statistics, signal processing etc. to concentrate on analysis of the expression data to look for inherent information in the datasets.

## 1.4. Classification using gene expression data

The classification problem in cancer studies relates to early diagnosis of the presence or absence of the tumor (L.Dyrskjot, 2003), the correct sub-type of the tumor (Golub T, 1999) or several other criteria . A classifier should take a vector of gene expression values as input and give the class label as output. There has been a lot of literature applying different classifier algorithms to microarray gene expression data. I have referred to some of the important ones in the literature review section. One issue, which is omnipresent in all microarray studies (at least for now), is the small sample size available for study relative to the huge feature size associated with the gene expression studies.

There are three major issues identified with this mismatch between sample and feature size (E.R.Dougherty, 2001). (1) How to generalize a classifier designed using a small sample set to the general population? (2) How to estimate the error for the classifier using the small sample set? (3) How to select the most relevant features needed for classification when the sample size is small?

All these three questions are being heavily researched right now. The objective of this thesis is to address the feature selection problem, i.e. how to select the genes most relevant for classification, from the thousands of genes in the gene expression profiles available.

## 1.5. Feature selection

Feature selection, as the name suggests, refers to selecting the most relevant features from all the variables, such that it contains enough information about the samples to classify them into the defined classes.

Most of the methods used now for feature selection from microarray datasets are based on statistical methods. There are two main approaches for selecting the relevant features.

(1) Filter method – where the genes are ranked according to some general properties like correlation, discriminative power etc. that are relevant for the problem being studied. An example of the filter method is the one proposed in (Dudoit, 2000) where a ratio of between class sum of squares to within class sum of squares is used to rank the genes.

(2) Wrapper method – formed to restrain the set of factors so that a given classification method s performance is improved. The prediction accuracy of the chosen classification method for all possible combinations of the genes is considered and the best set is chosen.

## 1.6. Feature extraction

Feature extraction refers to extracting the relevant information from the dataset. In this case, the extracted features need not be in the same domain as that of the datasets and therefore need not make any intuitive sense on first look. Methods like Principal Component Analysis, Singular Value Decomposition etc have been used before for feature extraction in microarray gene expression data with considerable success.

1.7. Wavelet transforms

The wavelet transform breaks down any arbitrary function into a superposition of wavelets. The wavelets are generated from a mother wavelet by dilation and translation operations.

The 1-D Discrete Wavelet Transform (DWT) of a signal is implemented by passing an input signal simultaneously through high-pass and low-pass filters followed by down sampling. A typical decomposition is shown below and the decomposition is discussed in detail in later chapters.

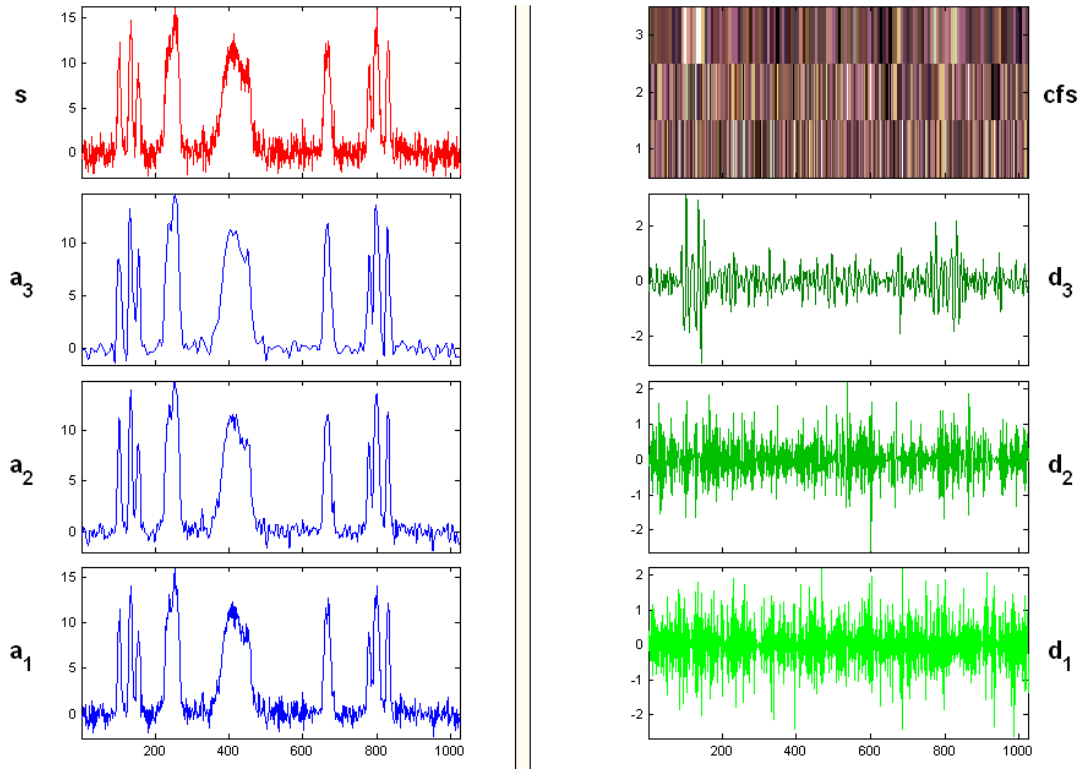


Figure 1.1. A 1,024 point signal decomposed to level 5 using the Sym 4 wavelet. The low frequency approximation signals are shown on the left pane and high frequency detail signals on the right side. Coefficients at different levels are also shown. (Implemented using wavelet toolbox of MATLAB®).

The decomposition of a signal using the wavelet transform is represented by equation (1.1). In Figure(1), decomposition of s can be represented by equation in (1.2) :

$$s = a_{n-1} + \sum_{k=1}^{n-1} d_k \tag{1.1}$$

$$s = a_5 + \sum_{k=1}^{4} d_k = a_4 + \sum_{k=1}^{4} d_k = a_3 + \sum_{k=1}^{3} d_k = a_2 + \sum_{k=1}^{2} d_k = a_1 + d_1 \tag{1.2}$$

1.8. Hypotheses

Null Hypotheses.

The probability p1 of the wavelet based method to make an accurate classification is the same as its probability p2 to make an incorrect classification. (p1=p2).

Alternate Hypotheses.

The probability p1 of the wavelet based method to make an accurate classification is greater than its probability p2 to make an incorrect classification. (p1>p2).

Apriori Significance Level $\alpha = 0.01$

# CHAPTER II

# LITERATURE REVIEW

## 2.1. Microarrays in biology

It was Schena et al (Schena M, 1995).who proposed a procedure to monitor the expression levels of multiple genes in parallel in their study of the expression profiles of 45 genes of Arabidopsis by two-colored fluorescent hybridization. They then extended the proposed method to the human genome (Mark Schena, October 1996) by creating a microarray imprinted with 1,046 human cDNAs to study the genes affected by heat shock. This was followed by the work of Lokhart et al (Lipshutz RJ, 1999) from Affymetrix where they proposed the oligonucleotide-based gene chips which were the first commercially available microarray chips. These landmark papers were followed by hundreds of publications which focused on addressing questions in oncology, cell biology, pathology, pharmacology and toxicology (J.Derisi, 1996) (D.J. Duggan, 1999) (Lettieri, 2006) (Stoughton, 2005) (Winzeler, 2000).

## 2.2. Noise removal and normalization

The fluorescent intensities showing expression levels obtained from the microarray chips at each location have to be denoised and normalized to obtain values indicating the expression level of the mRNA in that sample at the time of the assay. Bozinov and Rahnenfuhrer (Bozinov, 2002) and Smyth et al. (Smyth, 2002) have reviewed the current methods available for image processing of microarray data for noise removal. Brown et al(2001) addressed the need of a metric to evaluate the precision of each spot on a spotted microarray chip to get a measure of the reliability of the expression profile (Brown, 2001).

Different approaches are needed for cDNA arrays and Oligonucleotide arrays. When it comes to cDNA arrays, there have been several image analysis tools proposed for removing noise artifacts like circular spots and irregular shapes. ScanAlyze (Eisen, 1999), GenePix (Axon Instruments in., 1999) and Quantarray

(GSI Lumonics Inc., 1999) (for segmentation) and Spot (Buckley, 2000)(for background estimation) are some of the software. Yang et al. (Yang, 2002) has published a very interesting comparative study of the different segmentation and background estimation methods available for cDNA microarray studies. The spot software doesn't work well when artifacts fall into the background area. The Imagene (BioDiscovery Inc., 1997) has been found to be a good alternative in this case.

## 2.3. Microarray databases

Carrying out microarray assay requires special skills and it is not easy to obtain cancer tissue samples for analysis. With ever increasing interest in the microarray dataset based studies, it becomes important that researchers, who do not have molecular biology laboratory skills or simply do not have the funds to carry out these experiments, have access to these datasets in a standard format. This has lead to the formation of microarray databases which are maintained by major research labs like the Stanford Microarray Database and other agencies like the Gene Expression Omnibus (GEO). Reviews of the different microarray datasets and their formats are available.(Christopher J. Penkett and Jurg Bahler, 2004).

## 2.4. Classification

The classification problem basically requires design of a classifier which takes a vector of gene expression as input and outputs a vector indicating the class label of the input sample vector. Golub et. al. ( (T.R.Golub, 1999) were the first to address the issue of class discovery and classification when they used the expression profiles of 50 genes to classify Leukemia samples into Acute Lymphoid Leukemia and Acute Myeloid Leukemia. This was followed by numerous studies where classifiers were designed to classify between different types of cancer, different cancer stages, cancerous and noncancerous samples etc. I have listed some of the papers addressing classification problem in microarray data: (A.Ben-Dor, 2000) (M.Bittner, 2000) (G.Callagy, 2003) (L.Dyrskjot, 2003) (T.SFurey, 2000) (I.Hedenfalk, 2001).

There are three critical statistical issues related to addressing the classification problem using the expression data which are all caused by the huge dimension mismatch between the number of variables and the number of samples:

(1) How to generalize the classifier designed from a small sample set to the entire population?

(2) How to estimate the error with the limited data?

(3) How to select a subset of relevant features from the huge number of features?

There has been a lot of work in the literature addressing the importance of these issues. (E.R.Dougherty, 2001), (A.K.Jain, 1991), (D.Zongker, 1997), (J.Slansky, 2000) address the issue of feature selection and the limitations caused by the small sample sizes for designing classifiers in pattern recognition.

For the first issue of designing classifiers which will perform well on the general population, several methods of regularization have been proposed (Edward R. Dougherty, 2005). This can be achieved by either regularizing the classifier parameters or the data (Friedman, 1989). A typical case of regularizing data is 'noise injection' where a Gaussian distribution is kept at each data point and a large number of new data points are injected to the variable space, thereby shifting the balance of the variable – sample dimension mismatch (M. Skurichina,2000).

Error Estimation in the context of small sample microarray data has also been addressed in the literature (C. Sima, 2005). "Leave one out" cross validation is the most common method, but it has been found to be severely limited by high variances in its estimation (Dougherty, 2004). Several alternatives have been proposed, including bolstered error estimation where a method similar to noise injection is used to estimate the error term as a fraction (0-1) rather than just 0 or 1 (0 misclassification,1 classified correctly) (Dougherty, 2004).

## 2.5 Feature selection

The third problem is closely associated to a very common problem in machine learning and is called the 'curse of dimensionality' problem (A.K.Jain, 1991). A large number of feature selection methods are already in the machine learning literature. Many of them have been directly used in the microarray applications and many have been modified to meet the applications (Yvan Saeys, 2007).

A recent review on the different gene expression data analysis methods (Jafari, 2006) has identified t-test and ANOVA as the most common features selection tool.

The ANOVA and t-test are typical cases of univariate parametric methods, where each variable (genes) is assumed to have an underlying distribution (Churchill, 2003). The t-test and ANOVA assume the distribution to be Gaussian. Regression modeling (Thomas, 2001) and Gamma modeling (Newton, 2001) have also been used in place of Gaussian modeling in parametic testing. Several modifications of the t-test and ANOVA have been proposed to address the small sample issues (Churchill, 2003). There is ambiguity associated with the underlying assumptions about the gene expression profiles and the limited sample sizes have made it very difficult to validate the assumptions. This has lead to the use of model free non-paramtric methods for univariate gene selection (Troyanskaya, 2002). Several scoring methods like Wilcoxon rank sum method (Troyanskaya, 2002), BSS/WSS method (Dudoit, 2002) etc. have been proposed in the literature.

Univariate methods like the t-test and Wilcoxon method are simple to estimate and intuitive to comprehend, but have problems with the basic assumptions. The assumption in selecting genes using the univariate approach is that a set of the best N genes which are differentially expressed between the classes will form the best feature set of size 'N'. This approach completely ignores redundancy in the data.

There are several multivariate methods in the literature which account for gene-gene interaction by looking at the correlation between the variables. They include correlation based feature selection ranging from high dimensional (Wang, 2003), minimum redundancy – maximum relavance methods (Ding, 2003) and uncorrelated shrunken centroid approach (Yeung, 2003) to simple bivariate method where two genes are considered together at a time (Bo, 2002).

Feature selection methods which use classifiers embedded in them ( wrapper methods) have also been proposed. Here the gene space is explored for the best feature set using several methods such as the hill climbing method, forward search method, sequential floating forward search methods (Edward R.Dougherty, 2005) until the classifier performance is optimized.

Embedded methods where the feature selection methods are part of the classifier are also in use. A random forest approach where many single decision trees are combined to evaluate gene relevance (Dıaz-Uriarte, 2006) and support vector machine based approach (Guyon, 2002) are two examples.

2.6 Wavelet transform in microarray studies

There are two ways of applying the 1-dimensional wavelet transform on gene expression datasets: (1) Along the time axis, where we are looking at the expression of a particular gene at different resolutions. One very interesting case is the one by Klevec, R.R et al (Klevecz, 2000) where he discovered inherent oscillation in the expression of yeast genes. He found two periodicities, one of ~40 minutes and another ~80 minutes. Another case is the one by Rajendra Sahu et al (Subramani Prabakaran, 2006) where they used the Haar power spectrum to rank each gene. Also, (2) along the genes, one sample at a time. In a paper by Siew Hong Leong, Amit Aggarwal et al (Aggarwal Amit, 2005 ) studied the local and global effects of Aneuploidy (a genetic aberration observed in human cancers) on the genome using the continuous wavelet transform, the low scale analysis giving aberrations occurring locally in the genome and the high scale analysis revealing the aberrations occurring globally. Another very interesting case is the paper by Shutao Li et al (Li Shutao, 2006). They took the 1-D DWT of the gene expression data and applied an algorithm based on thresholding to choose the most important coefficients and used these coefficients as the feature inputs to SVM based classifiers.

# CHAPTER III

# DNA MICROARRAYS

## 3.1. The cell

The Webster dictionary defines cell as a small, usually microscopic mass of protoplasm bounded externally by a semi-permeable membrane, usually including one or more nuclei and various other organelles with their products, capable alone or interacting with other cells of performing all the fundamental functions of life, and forming the smallest structural unit of living matter capable of functioning independently. Every single event within a cell is the result of highly coordinated activity of a multitude of specific chemical transformations. These transformations involve a wide range of small organic molecules like carbohydrates and fatty acids and macromolecules like DNA and amino acids.

Most of the structural, functional and regulatory activities of life are performed and regulated by different combination of 20 unique amino-acids, called *proteins*. All the information about when, where and how to produce these proteins are carried in the genetic material called *Deoxyribonucleic acid (DNA)* contained in the nucleus of every single living cell (Lodish H, 2000). The entire genome is packaged by folding into the chromosomes in the nucleus and is replicated in a complex DNA replication process during cell division. (Figure 3.1)

## 3.2. Genes

The information in the DNA strands lies in the order in which the sequence of four nucleotides (Adenine, Guanine, Cytosine and Thymine) are arranged. Even though the DNA is millions of base pairs long (chromosome 1 of human genome has 220 million base pairs), the information bearing regions are discrete functional units ranging from a few thousand to a few tens of thousands of nucleotides called *genes*

The *coding region* of the genes encode its protein and *regulatory regions* defines when and where it gets expressed.

## 3.3. Introduction

The coded genetic information in the DNA is converted into proteins through a two step series – Transcription in the nucleus where the RNA polymerase enzyme catalyzes a copy of the information content of the DNA on to messenger RNA, and the Translation step by the ribosome in the cytoplasm, where the preprocessed mRNA is translated to amino-acid chains following a universal genetic code of translation. These series of processes constituting the transfer of information from DNA to RNA and then to proteins is called *gene expression* .

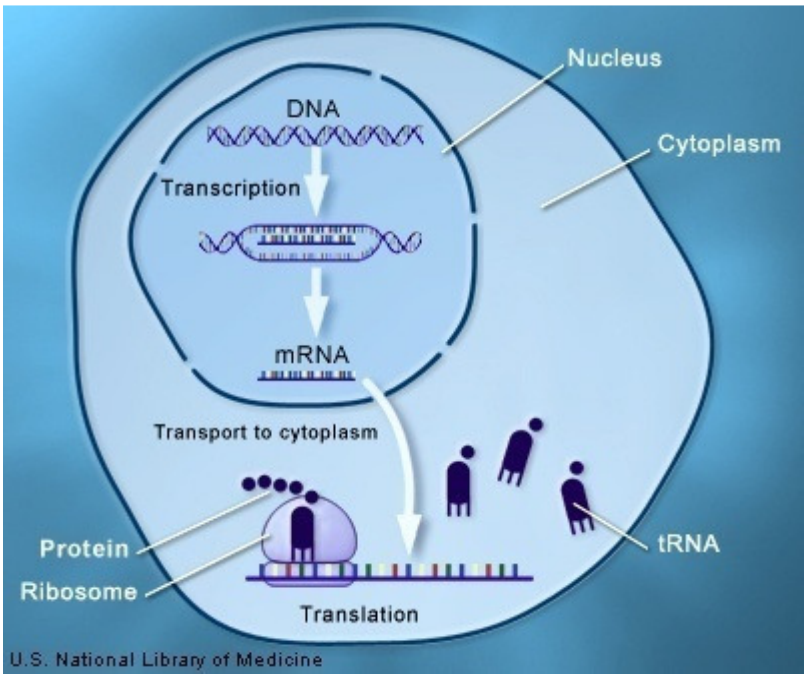


Figure 3.1. The three molecular genetics processes. Transcription, RNA preprocessing and Translation: (http://ghr.nlm.nih.gov/handbook/illustrations/proteinsyn.jpg).

## 3.4. Gene expression

The genes have coding and regulatory regions. The coding regions are divided into informative *exons* and non-informative *introns* . An enzyme called *RNA polymerase* attaches to sites upstream from

the start sites of each gene, called promoters, and catalyzes the linkage of rNTPs (ribo Nucleoside Tri Phosphate monomers) to form something called *pre-mRNA chains* . This step is *Transcription.*

In the next stage, first, a 7-methyl guanylate cap is added at the 5' ( 5 prime) end. This cap protects the mRNA from enzymatic degradation. The cap is also bound by a protein factor which is required to begin translation in the cytoplasm. Then, the poly (A) polymerase enzyme catalyzes the addition of a string of Adenylic acid residues to the 3' end. This results in a 100-250 base long poly (A) tail. This tail has importance in DNA microarray technology as it is used in the reverse transcription process of mRNA. The final product after the preprocessing of the pre-mRNA is *the messenger RNA*.

The preprocessed mRNA now leaves the nucleus and enters the cytoplasm. The mRNA attaches to the transfer RNA in the ribosomal subunits. A very complex process which involves the ribosomal RNAs, several protein complexes and transcription factors use the codons (triplets of nucleotides) on the mRNAs to assemble and connect together amino acid sequences based on a nearly universal genetic code. This process is *Translation.*

The entire process from Transcription of genes to RNAs to their subsequent translation to amino acid sequences is *gene expression*.

Table 3.1. The genetic code.

| | T | C | A | G |
|---|---|---|---|---|
| T | TTT Phe (F)<br>TTC"<br>TTA Leu(L)<br>TTG" | TCT Ser (S)<br>TCC"<br>TCA"<br>TCG" | TAT Tyr(Y)<br>TAC<br>TAA Ter<br>TAG Ter | TGT Cys(C)<br>TGC<br>TGA Ter<br>TGG Trp(W) |
| C | CTT Leu(L)<br>CTC"<br>CTA"<br>CTG" | CCT Pro (P)<br>CCC<br>CCA"<br>CCG" | CAT His (H)<br>CAC"<br>CAA Gln(Q)<br>CAG" | CGT Arg(R)<br>CGC"<br>CGA"<br>CGG' |
| A | ATT Ile(I)<br>ATC"<br>ATA"<br>ATG Met(M) | ACT Thr(T)<br>ACC"<br>ACA"<br>ACG" | AAT Asn(N)<br>AAC"<br>AAA<br>LYS(K)<br>AAG" | AGT Ser(S)<br>AGC"<br>AGA<br>ARG(R)<br>AGG" |
| G | GTT Val(V)<br>GTC"<br>GTA"<br>GTG" | GCT Ala (A)<br>GCC"<br>GCA"<br>GCG" | GATAsp (D)<br>GAC<br>GAA Glu(E)<br>GAG" | GGT Gly(G)<br>GGC"<br>GGA"<br>GGG" |

### 3.5. The need for dna microarrays

The latest estimate puts the number of protein coding genes in humans to be around 20,000 (Consortium, 2004). Some of these are expressed more or less to a constant extent as they are involved in basic reactions in the cells. They are called *housekeeping genes* . The rest of the genes are only expressed selectively! This means that the genes expressed in a cell at a given point of time depend on several factors like the stage of the organism s growth, the location of the tissue studied, the health of the tissue studied, the time of the day, etc. If we can find out simultaneously, which genes are expressed and to what extent are they expressed at any point in time, if we can somehow quantify this notion of gene expression, it can act as a very powerful tool for understanding almost all the biological processes in the cell.

### 3.6 Monitoring gene expression

As we discussed above, there are two main stages in gene expression. The important thing with cellular control is that the cellular control, and its failure, is the result of interaction between thousands of genes and their products. Even though DNA, RNA and the proteins are in different levels of the gene expression process, they show very high interaction levels. We do need to combine all three levels to get complete information about the cellular processes. But at the same time, each individual level does have considerable information available in itself, which makes study of individual levels very relevant.

As of now, we know most of the protein producing genes in us, but still have to go a long way in identifying all the proteins involved. The efforts are now focused on the mRNA expression levels because of the measurement considerations.

One issue that comes up here is to save all the genomic information is some form of library for further study. Many of the sequences in the DNA are non-informative. A better option is to save the mRNA information. It is is much easier to isolate as it is present in the cytoplasm and it is far shorter in length as all the introns and regulatory information is lost in the preprocessing stages after transcription, as was explained earlier. An enzyme called *Reverse Transcriptase* first discovered in retroviruses is used to reverse transcribe mRNA to what is called *complementary DNAs (cDNAs)* . Almost all the gene expression studies conducted today are based on the cDNAs.

Before the sequencing of the genomes of many organisms, there were several techniques available for studying one gene at a time under different conditions – the most important of this being screening using oligonucleotide probes and blotting techniques. Both these techniques are based on hybridization.

3.6.1. Hybridization

The nucleotides in DNA and RNA will preferentially associate with their corresponding complementary base-pairs. (Adenine – Thymidine/Uracil & Guanine – Cytosine). The double helix DNA strands can be denatured by heating in dilute solution at high temperatures (annealing). If the temperature is brought back and the ion concentration raised (near neutral pH, 40-65 C and .3-.6 M NaCl solution), the single strands will hybridize back to complementary strands. This re-association will take place even if other non-complementary strands are present. The hybridization property is the basis for probing and blotting techniques.

The idea of using probes to scan for the presence of specific genes in a sample, is to make a tagged, short (of the order of 20 nucleotides) oligonucleotide sequence from the cDNA being studied and hybridizing it with the sample, fragmented using restriction enzymes and separated by electrophoresis. The amount of cDNA present can be studied by autoradiography. The Southern Blotting looks for cDNA fragments while Northern Blotting looks for mRNAs.

The microarrays are basically conducting thousands of Northern Blotting experiments in parallel. Instead of distributing the probes on the samples, the microarray experiments involve attaching thousands of oligonucleotide probes onto some surface. The samples are tagged fluorescently and added on to the array surface and allowed to hybridize. The excess unhybridized samples are washed away and the hybridized material excited by laser using a laser beam scanner. We end up with the intensity of laser at each location on the chip which corresponds to the expression level of each gene (represented by the probes at each location) in the sample being studied.

This means that instead of looking at a few genes at a time, microarrays allow simultaneous monitoring of thousands of genes at the same point of time. This opens innumerable possible experimental designs for researchers. We can study change of gene expression in different stages of growth of on organism, disease condition, different stages of disease, effect of drugs, effect of a particular treatment etc.

3.6.2. Microarray technology platforms

There are three main types of DNA microarray technology platforms(Lee,2004) (Knudsen,2004): (1) spotted cDNA arrays, (2) spotted oligonucleotide arrays and (3) *in-situ* oligonucleotide arrays. They are different in the ways the arrays are produced and the types of probes used on the arrays.

Spotted cDNA microarrays involve using the full length cDNA clones or short sequences which represent the cDNA sequences (Expressed Sequence Tags) as probes. This method is of particular importance in cases where the genes are not completely known.

Spotted oligonucleotide microarrays involves referring to some genome database like GENEBANK for the gene we want to study and using synthetic oligonucleotide sequence (20 – 80 nucleotides long) which is characteristic of the gene to be studied instead of the entire cDNA to be used. This avoids a lot of the noise associated with the cloning, spotting and PCR stages of the cDNA based method. This method also gives more control over the specific part of the gene to be studied. This becomes important in cases where the transcripts are highly similar and give cross hybridization problems in cDNA microarrays.

Both the spotted methods give a tremendous amount of design flexibility for researchers, but involve considerable noise and tendency for cross hybridization.

In the case of in-situ oligonucleotide microarrays, a large number of probes (25-mer oligos to the order of 50,000 per square 1.28 square cm) are synthesized on to the surface of the chip using a combination of photolithography and oligonucleotide chemistry. The Affymetrix gene chips have much higher density compared to the spotted arrays, but there is no control for the researchers over the probes.

The test and reference samples have to be hybridized separately and the intensity values represent gene expression in the sample. The test and reference samples are mixed and hybridized together and the intensity values represent the ratio of the gene expression of one sample with respect to the other. The scanning and image analysis methods are also different.

3.6.3. Selecting and synthesizing the probes

cDNAs allow great flexibility for the researchers for selecting the probes they want to explore. The cDNA can be created from the cDNA libraries discussed earlier. cDNAs obtained from the libraries are amplified using an assay for amplification of the number of nucleotide chains called Polymerase Chain Reaction (Lodish H, 2000). The strategy used for selecting templates for the studies depend on the organisms (Lee, 2004).

In the case of oligonucleotide probes, the public sequence databases like GeneBank, dbEST etc. can be used to obtain the exact sequence of the genes to be studied. But there are several questions to be answered about the sequence of the oligonucleotide made.

Firstly, how long should the oligonucleotide sequence be? Obviously, the longer the sequence, the better it is. But, it is going to be more expensive to synthesize long sequences. So how small can it be? 20 nucleotides can represent $4^{20}$genes! Therefore, it is reasonable to use 20 to 70 nucleotides.

Secondly, how similar are two oligos in the same array? If they are too similar, the cross-hybridization can take place and might cause loss of information. Therefore, care should be taken to keep all oligos as different as possible.

Thirdly, do all probes have similar hybridization properties? For example, for all the oligos to perform similarly, their melting temperatures should be similar.

Several other issues like the location of the probe in the gene and self-hybridization properties should be considered.

In the in-situ oligonucleotide arrays, around 20 nucleotide long probes are synthesized on to the surface directly. They are organized as a pair of Perfect-Match and Mismatch probes. The Perfect-match forms the complementary to the target sequence and the Mismatch probe has all but one sequence forming the complementary to the target sequence. The central nucleotide in the Mismatch probe is deliberately made a mismatch to the target sequence. The MM probe act as control for cross hybridization. A probe set of 16-20 probe pairs are used for each transcript.

3.6.4. Manufacturing the arrays

The in-situ oligonucleotide arrays use photolithography for manufacturing the arrays. The support is covered with a protective chemical which can be deactivated by light. A mask with holes is used to select the locations onto which a particular nucleotide is needed. The light through the hole will expose the locations on the surface through the holes in the mask. The process is repeated for different nucleotides by varying the location of the holes and thereby selecting the location on the surface.

In the case of spotted cDNA microarrays, cDNAs are prepared away from the surfaces. The probes are deposited and then attached to the surface robotically. Different materials ranging from plastic polymers to glass are used for making the surface materials. The surface of the materials is covered by background reducing materials like poly-lysine coatings or complex 3D molecular matrix layers.

The cDNAs or oligos are spotted using two main methods: Using capillary action for picking up the cDNAs from their solution and depositing onto the support surface. This method is called Contact spotting. The other method doesn t involve contact, as the oligos are sprayed on to the surface with high precision using a technology similar to that used in ink-jet printing.

Once the oligos or cDNAs are spotted onto the surface of the support, they have to be attached covalently to the surface. The cDNAs are attached using Ultra Violet cross-linking where attachments are formed randomly across the probe molecules. The oligonucleotide is more often linked by using linker molecules. The oligonucleotide is attached with a linker at its end, which chemically binds to the surface which has appropriate reactive groups.

3.6.5. Labeling the targets

There are several labeling techniques available. Three very common approaches used for labeling the targets are summarized below (Lee M.-L. T., 2004).

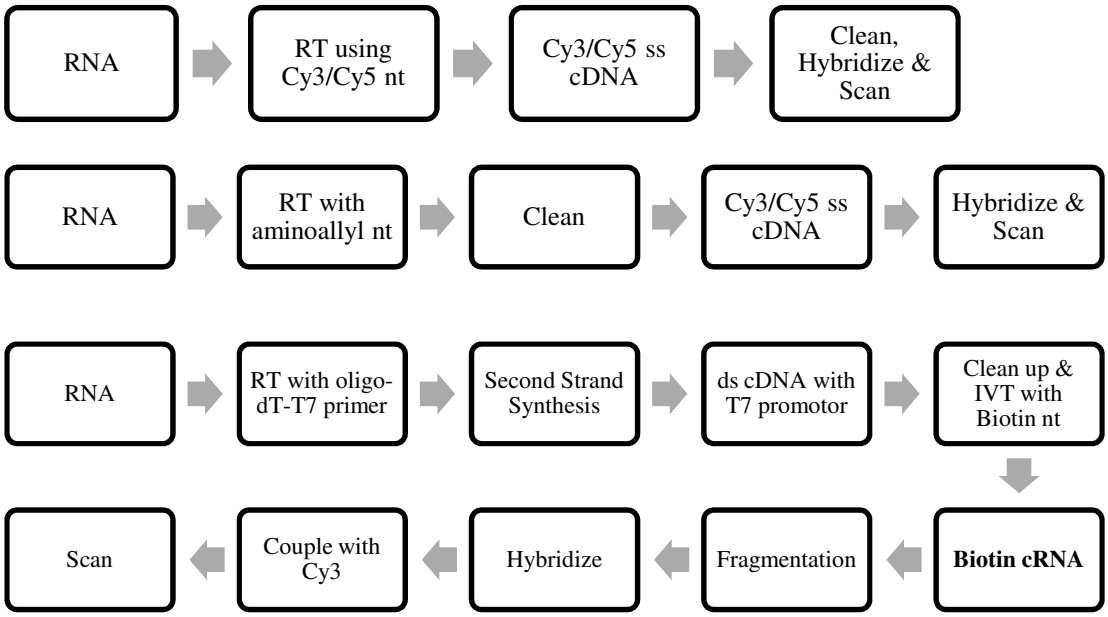


Figure 3.2. Three labeling methods used for target labeling- Direct, Using Aminoallyl nucleotides and in vitro transcription based method. RT – Reverse Transcription; IVT – In vitro transcription.

The first method basically involves reverse transcription of the mRNA to cDNA strands where the nucleotides included are modified by attaching a dye to it. For the spotted microarrays ( both cDNA and Oligonucleotide), two test and reference samples, are tagged separately, but mixed together before hybridizing onto the microarray slides. In the case of the direct reverse transcription method, the dye molecules can be incorporated to the reference and the test sample with different efficiency.

This problem is addressed in the second method by using Aminoallyl group for reverse transcription as it is less sensitive to structural bias of the dye to the cDNA molecules.

The third method, the in vitro transcription based methods has the advantage of using very small RNA quantities (of the order of 10-20µg). This involves amplification processes like PCR which comes with a disadvantage that the different mRNAs can be amplified disproportionately.

This is overcome in vitro transcription by using a oligo-dT primer with a promoter sequence extension. The cDNA generated can be used to generate any number of cRNA needed by transcription from the promoter sequence.

IVT is used for in situ oliogonucleotide arrays. Biotin-modified nucleotides are used for Affymetrix chips as biotin modifications are used for attaching the dyes and fragmenting them gives uniform hybridization properties.

### 3.6.6. Scanning and analyzing the images obtained

Once hybridization is completed, the microarrays are scanned for the fluorescence of the tags using confocal microscopes to measure light at wavelengths characteristic of the dyes. The high resolution monochrome images are captured using scanners and a high resolution image is obtained. The spots are identified using Gridding where a grid is placed on the image to obtain the approximate locations of the spots. The Segmentations process separates the background noise from the foreground signals.

This is followed by intensity extraction step which extracts intensity values for each spot which represents the expression level at each spots.

### 3.6.7. Microarray data

The final microarray data obtained after all the processing is a two-dimensional matrix with the expression levels of each gene along the rows and the samples along the columns. Either the actual numerical value can represent expression levels of each probe in the case of in situ oligonucleotide arrays and ratios or log transformed log ratios of differential expression between the test and the reference sample.

# CHAPTER IV

## FEATURE SELECTION IN MICROARRAY BASED CLASSIFICATION

In an ideal setting, i.e. in a case where we know the underlying distribution of the class labels and the samples, the expected error of the classifier will decrease monotonously with increasing number of variables in the feature set. But, in most real situations, as in the case with the microarray data, we do not know and cannot estimate the underlying distribution as we are severely limited by the available sample sets. All the gene expression values obtained from microarrays are potential features. But using the entire variable (gene) set as the feature set will result in over fitting the data (Over fitting results in the classifier performing well in the given sample set, but not on new samples). This problem is inherent in classification problems and is called curse of dimensionality. Here, the classifier error decreases with increasing size of the feature set until it reaches an optimal number of variables in the feature set after which it starts increasing. This is called peaking phenomenon (A.K.Jain, 1991).

In order to optimize the performance of the classifiers, it becomes imperative to choose a small subset of the variables which can give optimum classification results for the chosen classifier. A very important question to be addressed here is to find out the optimal size of the feature set. Empirical studies have indicated that the optimal number of variables in a feature set depends on the number of samples, the classifier being used, redundancy in the variables selected and even on the error estimation method used for evaluating the classifier (J. Hua, 2005) (Chao Sima, 2005).

A straight forward approach to select the best feature set will be to try every possible combination of variables (genes) on the chosen classifier and find out which one gives the minimum error. Considering the fact that every single gene is a potential feature, testing every combination of thousands of genes is computationally impractical.

The need for Feature Selection algorithms which can give suboptimal performance on the classifier can thus be summarized by three main reasons: (a) To overcome the curse of dimensionality problem, (b) To give cost effective, computationally efficient methods which select features sets which gives close to optimal classification performance and (c) To provide small feature sets which can act as potential diagnostic tools for a particular disease condition.

The Feature Selection techniques currently used are grouped into three categories depending on how they are linked to the classifiers. Some of the most commonly used methods are discussed below with their advantages and disadvantages.

### 4.1. Filter methods

The approach in filter methods is to calculate a score of relevance for all the variables, taken one at a time. The assumption is that most relevant features taken together will give a good feature set and thus very good classification results. They are completely independent of the classification algorithms and the computations are simple and fast.

The simplest and a naïve method for identifying differential expression will be to look for fold changes between the different conditions. This basically comes down to finding the ratio of average expression levels of the gene being considered between the different conditions. A threshold ratio is chosen and genes which cross this threshold are selected as relevant genes (MARK SCHENA, 1996). The main problem with this approach is the lack of a score which indicates the relevance of a gene.

Several univariate filter based approaches have been proposed after that. There are two subtypes of filters based on the assumption about the underlying distribution of the variables - Parametric methods where the variables are assumed to have an underlying Gaussian distribution and Non-parametric methods where there are no such assumptions.

#### 4.1.1. Parametric methods

The two sample sets might come from patients at different stages of a disease. The problem definitions have to be different for cases where more than one samples is obtained from the same patient.

The null hypotheses will be that there is no difference in the expression levels between the conditions. The interpretation of the hypotheses depends on the experimental design. In the case of cDNA microarrays where the two samples are hybridized together, the ratio of the expression between the two samples should be one. In the case of cDNA arrays where the two samples are hybridized seperately against a common reference, there should be no difference between there ratios. If log transformed expression ratios are used, then the difference should be zero. The same hypotheses will apply for the oligonucleotide data also. It is worth mentioning here that the definition of hypotheses assumes the samples to come from independent biological sources (Churchill, 2003).

#### 4.1.1.1. T-test

The 2 sample t-test is one of the most common tools used for finding differentially expressed genes in two condition cases (Jafari, 2006) (Callow MJ, 2000). The t-test does a statistical assessment to see if two groups of samples are statistically different from each other (Table 4.1). But the power of the t-test is severely limited by the small number of samples available for each condtions. If, merely by chance, the variance of a samples from a condition is very small, then the t score end up showing a false high value even when the difference in mean is small.

#### 4.1.1.2. Modified t-test

One possible solution to solving the effect of small samples has been to use a variance obtained from pooling the samples in the class across all genes. This ends up giving the same effect as a fold-change test as the variance for each genes across the class types are not accounted for (Table 4.1). In "Significance analysis of microarrays" method, a constant is to added to the denominator term which will bring the score $t_g$ down when the difference of mean is small, but the variance term in one of the classes is also small (Churchill, 2003)(Virginia Goss Tusher, 2001).

In regularized t-test, a weighted average of both local and global variances are accounted for to avoid local variances from giving high score even when the difference in mean is very small (Baldi P, 2001).

B-test statistic applies a bayesian approach to the entire data-set to estimate prior probabilites for the genes to be differentially expressed and for it to not to be differentially expressed between the classes.The ratio of these probabilites is used as a test statistic (Lonnstedt I, 2002).

Table 4.1. The summary of t-test and its modifications.

| | |
|---|---|
| t-test for gene g, classes i and j<br>$t_g = \frac{\overline{X_{ig}} - \overline{X_{jg}}}{\sqrt{\frac{s_{ig}^2}{n_i} + \frac{s_{jg}^2}{n_j}}}$ (4.1) | $\overline{X_{ig}}, \overline{X_{jg}}$ - mean $s_{ig}, s_{jg}$ -standard deviation<br>$n_i, n_j$ - Number of samples $\frac{s_{jg}^2}{n_j}$ and $\frac{s_{ig}^2}{n_i}$ are standard error terms $SE_i$ and $SE_j$. |
| Significance analysis of microarrays<br>$t_g = \frac{\overline{X_{ig}} - \overline{X_{jg}}}{s_c + \sqrt{\frac{s_{ig}^2}{n_i} + \frac{s_{jg}^2}{n_j}}}$ (4.2) | A constant term $s_c$ is added to prevent a small $SE_i$ or $SE_j$ from affecting the t-static measure. (Virginia Goss Tusher, 2001) |
| Regularized t-test<br>$t_g = \frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1) SE_g^2}{v_0 + n - 2}}}$ (4.3) | A parameter $v_0$ which determines the relative contributions of gene specefic variance to the global variance(Churchill, 2003). |

4.1.1.3. Analysis of variance

Analysis of Variance(ANOVA) is a generalization of the t-test to multiple class cases. The null hypotheses is that expression of genes in all the conditions studied have the same mean and the alternate hypotheses is that the gene is differentially expressed.

A typical gene selection method using ANOVA is described below (Churchill, 2003):

The model is for cDNA microarray data and we deal with logarithm of intensity and not the ratio here. There are two stages in the design of the model.

Stage 1: The normalization model which accounts for the effects of the arrays(i), the dyes(j) used and the measurement(r). The log transformed intensity value y is modelled as:

$$y_{ijgr} = \mu + A_i + D_j + AD_{ij} + r_{ijgr} \tag{4.4}$$

Stage 2: Gene Specefic model where the residual term $r_{ijgr}$ from the normalized model is used to model one gene at a time. G is the mean across gene $g$. VG accounts for variation caused by samples, DG by dye and AG by arrays, across the gene $g$.

$$r_{ijgr} = G + VG_{ij} + DG_j + AG_i + \varepsilon_{ijr} \tag{4.5}$$

The null hypotheses will be that the effect of interaction between the Variation across samples term and Gene term is zero and the alternate hypotheses is that there is at least one pair of samples for which the interactions are not equal. Both Fixed Effect and Mixed Effect design are possible based on the assumptions about the terms (Dragichi, 2003). All the variations of the t-test can be extended to the ANOVA also.

Regression modeling (Thomas, 2001) and Gamma modeling (Newton, 2001) have also been used in place of Gaussian modeling in the parametic testing.

4.1.2. Non-parametric methods

The uncertainity associated with the assumptions about the underlying distributions and the problems associated with the small sample settings has lead to proposal of several non-parametric methods for gene selection.

4.1.2.1. Non-parametric t-test

Non parametric t-test does not assume normal distribution in the data. Instead the probability distribution of the p-value is estimated by permutation of the dataset several thousand times and counting how many times the t statistic for each gene j exceeds the true t-statistic (Troyanskaya, 2002). The p value is then corrected for multiple testing using Bonfferoni correction(Dragichi, 2003).

$$p_j = \frac{count(t_{jperm} > t_{jobserv})}{count(permutations)} \tag{4.6}$$

$$p_{jBonferroni} = \min(m \times p_j, 1)$$

4.1.2.2. Wilcoxon rank sum test

It is a non-parametric alternative to t-test which works with the rank of the data. The genes are ranked according to their expression levels across the samples. The null hypothesis is that the difference between the mean of the ranks is zero and the alternative hypothesis is that it is not zero. The p-values are assumed to be normally distributed when sample sizes of both groups are over 8. The estimations involved are shown below (Troyanskaya, 2002).

Group 1 has n1 samples and Group 2 has n2 samples.

$$w1 = \sum ranks_{sample1} \tag{4.7}$$

$$u1 = w1 - n1 \times (n1 + 1) / 2 \tag{4.8}$$

$$mean_{u1} = n1 \times n2 / 2$$

(4.9)

$$\mathrm{var}_{u1} = n1 \times n2 \times (n1 + n2 + 1) / 12$$

(4.10)

$$z = (u1 - mean_{u1}) / \sqrt{\mathrm{var}_{u1}}$$

(4.11)

z Є N(1,0) if n1,n2 > 8

4.1.2.3. Sum of Square method

This is a very intuitive and one of the most commonly used non-parametric statistics for finding differentially expressed genes (Dudoit, 2000). The score will be high if spread of class average with respect to joint average is high and the spread within each class is small. score for $j^{th}$ gene when I is the indicator taking 1 if $i^{th}$ sample belong to $k^{th}$ group and 0 otherwise.

$$Score(j) = \frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\overline{x_{kj}} - \overline{x_{.j}})^2}{\sum_i \sum_k I(y_i = k)(\overline{x_{ij}} - \overline{x_{kj}})^2}$$

(4.12)

4.2. Multivariate methods for gene selection

There is a fundamental problem with the underlying hypotheses used in univariate approach for gene selection for classification. The assumption is that a set of genes which are most differentially expressed in the different classes will give the best feature set for classification. This is not true in practice! The univariate approach completely ignores the interactions between the selected variables and thereby results in redundant feature set. The problem with this approach is two prone. For example, if the set feature set size is 10, the univariate methods will choose 10 most relevant genes which discriminates one condition from another. If 4 of these selected genes are highly correlated with the rest of them, we are not

only wasting the 4 genes but also limiting the space of the dataset represented by the selected features by a dimension of 4. The Multivariate methods address this problem and looks to minimize correlation between the selected features to make them less redundant.

Some of the multivariate methods proposed in the literature are discussed below.

4.2.1.Correlation based feature selection

Here (Hall, 1999.) a subset of features (genes) is considered together instead of ranking one feature at a time.

$$CFS_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{4.13}$$

$CFS_s$ is score of feature subset having k features, $\overline{r_{cf}}$ is the average feature to class correlation and $\overline{r_{ff}}$ is average feature to feature correlation.

4.2.2. Minimum redundancy maximum relevance method

This method(Ding, 2003) selects feature sets which maximizes some Relevance criteria and minimizes some Redundancy criteria. Both F – statistic is used as a measure of relevance and both the absolute value of correlation criterion and Euclidean distance between the selected features are used as measure of Redundancy.

A typical optimization criterion used is:

$$MID = \max_{i \in \Omega_S} [F(i,h) - \frac{1}{abs(S)} \sum_{j \in S} abs(c(i,j))] \tag{4.14}$$

where F(i,h) is the F-statistic of $i^{th}$ feature and $h^{th}$ class and c(i,j) is the correlation between $i^{th}$ and $j^{th}$ features.

4.2.3. Uncorrelated shrunken centroid based method

Overall Centroid (Yeung, 2003) is the average expression level of a gene across all samples, class Centroid is the average expression level of a gene within each class, Δ is the shrinkage threshold and ρ is the correlation threshold. The estimations are made for a set level of Δ and ρ. The algorithm starts by selecting a number of genes for which the difference between the class Centroid and the overall Centroid is the higher than the shrinkage threshold Δ and the correlation coefficient between the genes is less than correlation threshold ρ. The features are tested for classification accuracy and the values of Δ and ρ which optimize the classification accuracy are used for the selecting the best feature set.

4.3. Wrapper methods

The wrapper method(Edward R.Dougherty, 2005) uses a multivariate approach but includes the classifier bias into the search procedure thereby optimizing the performance for a given classifier. The classifier block acts as a black box in the wrapper based feature selection methods, so it can be applied directly to a wide range of classification algorithms.

In a typical wrapper based method, the entire dataset is divided into small blocks. All except one block is given to a feature selection and parameter estimation stage where the best feature set which minimizes optimization criteria is estimated. A typical optimization criterion the sequential search algorithm looks for is given below (Edward R.Dougherty, 2005).

$$J(\vartheta,\gamma)=\frac{\sum_{t=1}^{n}(Y(t)-g(X_{i1}(t),\ldots X_{ik}(t);\vartheta))^2}{n} \tag{4.15}$$

{i1..ik} is the feature set, θ is the classifier parameters, $g(X_{i1}(t),\ldots X_{ik}(t);\vartheta)$ is the classifier error for the selected feature set and n is the number of samples.

The left out block is now used to estimate the classification error of the selected feature set. The method is repeated for all the blocks and average of the classification error is estimated.

There are several approaches for implementing the search for optimal feature in the feature selection and optimization stage. A naïve one will be the Hill Climbing approach where the gene selection starts with one gene. The entire variable space is searched and the gene which minimizes the classifier error is added to the feature set. The process is continued until none of the candidates perform better than the current feature set.

One disadvantage of this method is that once a variable (gene) is included in the feature set, it cannot be excluded later (nesting property). The Sequential Floating Forward Search is an approach where after every single feature is added a back tracking process looks for better classification results, if one of the selected genes are removed from the feature set. The forward search process resumes after this. This method has been reported to give better results but is computationally more demanding.

# CHAPTER V

## WAVELET TRANSFORM

The wavelet transform breaks a signal into a sum of scaled and translated versions of localized waveforms called wavelets. A waveform has to fulfill the following mathematical criteria for it to be classified as a wavelet.

(1) It should have finite energy. i.e. for a wavelet ,

$$E = \int_{-\infty}^{\infty} \left|\psi(t)\right|^2 dt < \infty \tag{5.1}$$

(2) Admissibility condition:

$$C_g = \int_0^{\infty} \frac{\left|\bar{\psi}(f)\right|^2}{f} df < \infty \tag{5.2}$$

where

$$\bar{\psi}(f) = \int_{-\infty}^{\infty} \psi(t) e^{-i(2\pi f)t} dt \tag{5.3}$$

(3) The Fourier transform of the wavelet should be real and it should vanish for negative frequencies.

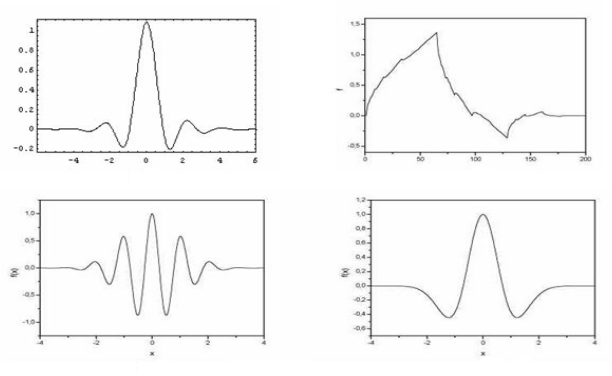

Figure 5.1. Some commonly used wavelets. Coiflet, Daubechies, Morlet and Mexican hat (Clockwise).

The wavelet transform is extremely useful in analyzing highly noisy, aperiodic and broken signals. It has found applications in a wide variety of applications ranging from climate analysis, seismology, financial measures, biological applications like ECG, EMG and recently even in studying genomic data(Addison 2002).

There are a large number of wavelets available for analyzing signals. The best wavelet for a particular application depends on the nature of the data and what the user is looking for in the data The localized small waveforms, typical cases of which are shown in the figure are called mother wavelets . Each wavelet is characterized by its pass-band center frequency.

The pass-band central frequency is given by:

$$f_c = \sqrt{\frac{\int_0^\infty f^2 \left|\hat{\psi}(f)\right|^2 df}{\int_0^\infty \left|\hat{\psi}(f)\right|^2 df}} \tag{5.4}$$

where $\hat{\psi}(f)$ gives the Fourier spectrum of the mother wavelet.

The characteristic frequency of a scaled wavelet is given by $f_c/a$ where a is the scale.

5.1. The continuous wavelet transform (cwt)

The CWT of a signal can be thought of its cross- correlation with the dilated and translated versions of a mother wavelet. The CWT of a signal x(t) is expressed as:

$$T(a,b) = \int_{-\infty}^{\infty} x(t)\psi^*_{a,b}(t)dt \tag{5.5}$$

where

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{5.6}$$

The b term accounts for the translation and a accounts for dilation of the mother wavelet. An illustration of analysis of a sine wave with varying frequency using a Mexican hat wavelet is shown below.

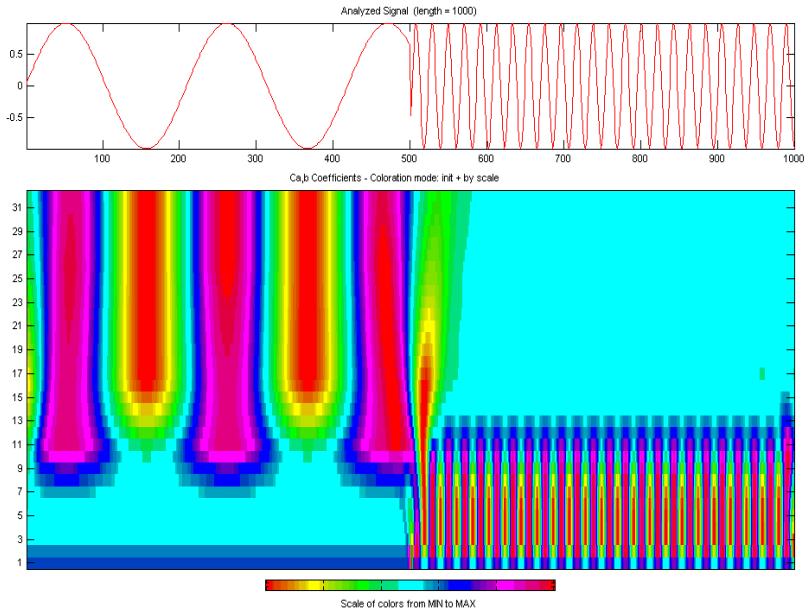


Figure 5.2. Continuous wavelet transform of the sine wave with different frequencies using a Mexican hat wavelet. The sampling frequency of the signal is 1000 Hz and the mother wavelet is scaled from 1 to 64 in step sizes of 1.The heat map scale is shown below the plot. (Implemented using Wavelet Toolbox of MATLAB®).

The mother wavelet is a Mexican hat wavelet shown in the figure 5.1. The transform operation is carried out in the computer by discretizing integral of the wavelet transform equation. Each point on the

transform plot is obtained by multiplying the signal with the mother wavelet with the scale, a ranging from 1 to 64 and the location parameter b ranging from 1 to 1000. It is clear from the figure that the lower scales emphasize the higher frequencies while the higher scales emphasize lower frequency signals.

The original signal can be recovered from the transform by:

$$x(t) = \frac{1}{C_g} \int_{-\infty}^{\infty} \int_{0}^{\infty} T(a,b) \psi_{a,b}(t) \frac{dadb}{a^2} \tag{5.7}$$

where T(a,b) is the wavelet transform of the signal, $\psi_{a,b}(t)$ is the wavelet at scale a and location b and x(t) is the reconstructed signal. The inner integral over all the values of b for each scale a, gives the reconstruction of the signal at each scale. Integrating all the reconstructed levels gives the signal.

5.2. The discrete wavelet transform (dwt)

The DWT involves using discrete values of dilation and translation parameters instead of the continuous integrals used in the CWT. A typical approach is to take logarithmic steps in the scaling parameter a and defining the steps in the location parameter b in terms of a .

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_0^{\,m}}} \psi\left(\frac{t - n b_0 a_0^{\,m}}{a_0^{\,m}}\right) \tag{5.8}$$

where $a_0$ and $b_0$ are constants.

The scaling term is represented as a power of $a_0$ and the translation term is a factor of $a_0^{\,m}$.The most common choice for the parameters $a_0$ and $b_0$ are 2 and 0 (dyadic grid scaling).

The dyadic grid wavelet is expressed as:

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}} \psi\left(\frac{t - n2^m}{2^m}\right) = 2^{-m/2} \psi(2^m t - n) \tag{5.9}$$

The dyadic grid wavelets are designed to be orthonormal. This means that they are both orthogonal and they have unit energy. These conditions can be expressed as:

$$\int_{-\infty}^{\infty} \psi_{m,n}(t)\psi_{m',n'}(t)dt = \{1 for : m = m' and n = n' / 0 : otherwise$$

(5.10)

So, the DWT for a continuous signal x(t) using a dyadic wavelet and its reconstruction from the DWT back to the signal x(t) can be expressed as:

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt$$

(5.11)

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t)$$

(5.12)

It is very clear from the equations that the transform values are taken at discrete points of m and n of the wavelet and the signal can be reconstructed from the wavelet coefficients over all values of m and n.

5.2.1. Multi resolution analysis

The wavelet functions are associated with the details in the signal. Another set of functions called scaling functions are associated with dyadic wavelets. They basically have the same form as the wavelet functions.

$$\phi_{m,n}(t) = 2^{-m/2}\phi(2^m t - n)$$

(5.13)

The scaling functions have a smoothing effect on the signal. It give a approximation of the signal in terms of the wavelet used. The $\int_{-\infty}^{\infty} \phi_{0,0}(t) = 1$ and the scaling function is orthogonal only to its translations and not its dilations.

The approximation coefficients of a signal x(t) are given by:

$$S_{m,n} = \int_{-\infty}^{\infty} x(t)\phi_{m,n}(t)dt$$

(5.14)

The inverse transform of the approximation coefficients will give us the approximation of the signal at a particular scale given by:

$$x_m(t) = \sum_{n=-\infty}^{\infty} S_{m,n}\phi_{m,n}(t)$$

(5.15)

The original signal x(t) can be expressed as a sum of approximation coefficients at $m0^{th}$ scale and the sum of all details from $-\infty$ to m0. This can be expressed as:

$$x(t) = \sum_{n=-\infty}^{\infty} S_{m0,n}\phi_{m0,n}(t) + \sum_{m=-\infty}^{m0}\sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t)$$

(5.16)

$$d_m(t) = \sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t)$$

(5.17)

$$x(t) = x_{m0}(t) + \sum_{-\infty}^{m0} d_m(t)$$

(5.18)

So x(t) at any level m can be found by adding the approximation and detail reconstruction at the level m+1 , which can be expressed as:

$$x_m(t) = x_{m+1}(t) + d_{m+1}(t)$$

(5.19)

This idea of getting the approximation signal at a lower scale (higher resolution) by simply adding the approximation and detail reconstructions of the higher scale is called Multi Resolution Analysis .

The idea of multi resolution analysis can be illustrated with an example.

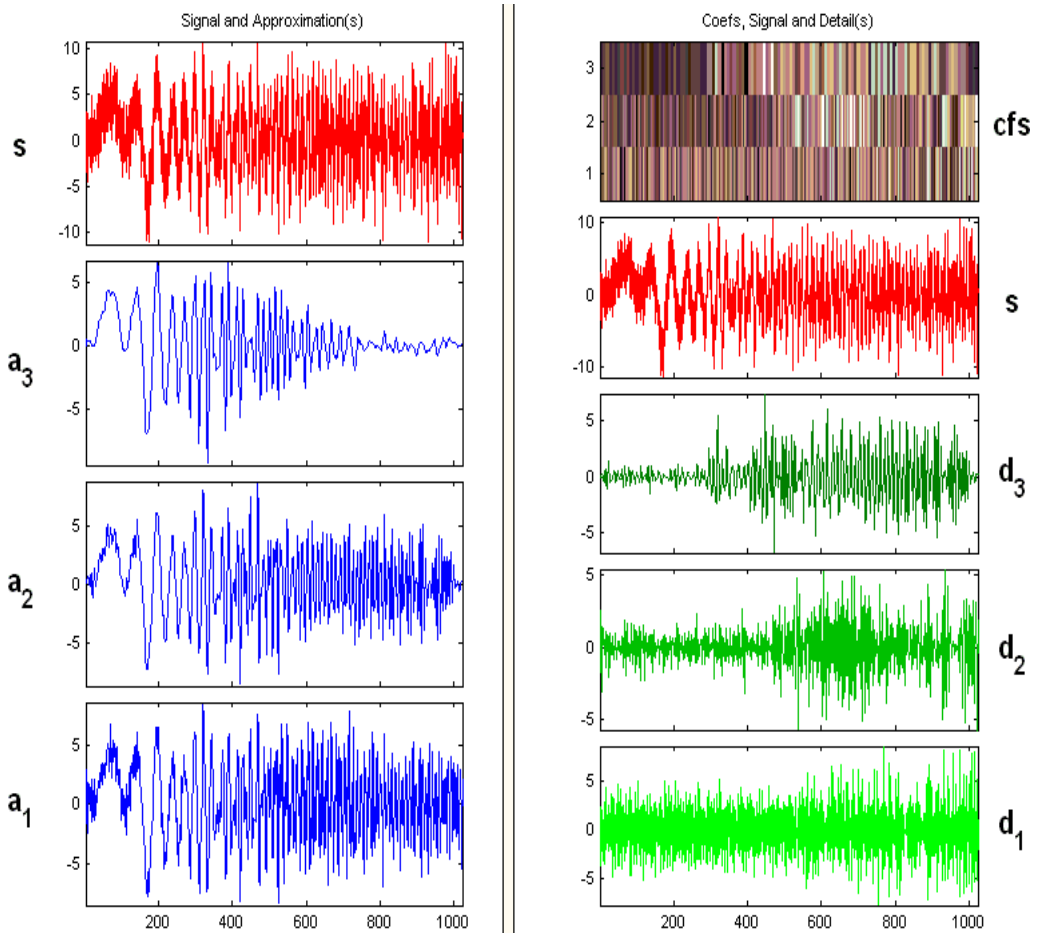


Figure 5.3. Multi Resolution Analysis of a highly noisy signal using db3 wavelet transform decomposed to the 3rd level. The original signal is shown at the upper left hand side and coefficients are shown in the upper right hand corner. The approximation signals at each level are shown on the left side and details on the right side. s = a3+d3+d1+d2. (Implemented using wavelet toolbox of MATLAB®).

5.2.2. Scaling and detail coefficients

The scaling and detail functions at a given scale can be described in terms of shifted versions of the next smaller scale, each multiplied by their respective scaling and wavelet coefficients. The wavelets are assumed to have compact support . i.e. they have a finite number of these coefficients.

$$\phi_{m+1,n}(t) = \frac{1}{\sqrt{2}} \sum_k c_k \phi_{m,2n+k}(t) \tag{5.20}$$

$$\psi_{m+1,n}(t) = \frac{1}{\sqrt{2}} \sum_k b_k \psi_{m,2n+k}(t)$$

(5.21)

The scaling functions are required to be orthogonal to each other and the ck values should add to 2.

$$\sum_k c_k c_{k+2k'} = \{2 - if : k' = 0 / 0 : otherwise$$

(5.22)

$$\sum_k c_k = 2$$

(5.23)

The scaling coefficients ck are reversed and their signs are changed before using them as wavelet coefficients. This ensures orthogonality among the scaling and its corresponding wavelet function. i.e.

$$b_k = (-1)^k c_{Nk-1-k}$$

(5.24)

This ensures that information contained in the approximation and detail coefficients are unique.

5.3. Fast wavelet transform

The use of smaller scale functions to evaluate the next higher scale function can be extended to the estimation of scaling coefficients. It can be proved(Addison 2002) that the approximation coefficients and detail coefficients at the level m+1 can be estimated from the approximation coefficients at the next lower level.

$$S_{m+1,n} = \frac{1}{\sqrt{2}} \sum_k c_k S_{m,2n+k}$$

(5.25)

$$T_{m+1,n} = \frac{1}{\sqrt{2}} \sum_k b_k S_{m,2n+k}$$

(5.26)

This allows estimation of approximation and detail coefficients of all higher levels without even knowing the signal x(t) once, we know the coefficients for a particular level. The estimation reduces to simple filtering operation, where the $(\frac{1}{\sqrt{2}})c_k$ coefficients correspond to low-pass filter and the $(\frac{1}{\sqrt{2}})$ $b_k$ terms corresponds to high pass filter. Once we know the coefficients at a particular level, the higher levels can be estimated by iteratively sending the estimated coefficients through the filters.

The schematic of signal filtering and reconstruction is illustrated below:

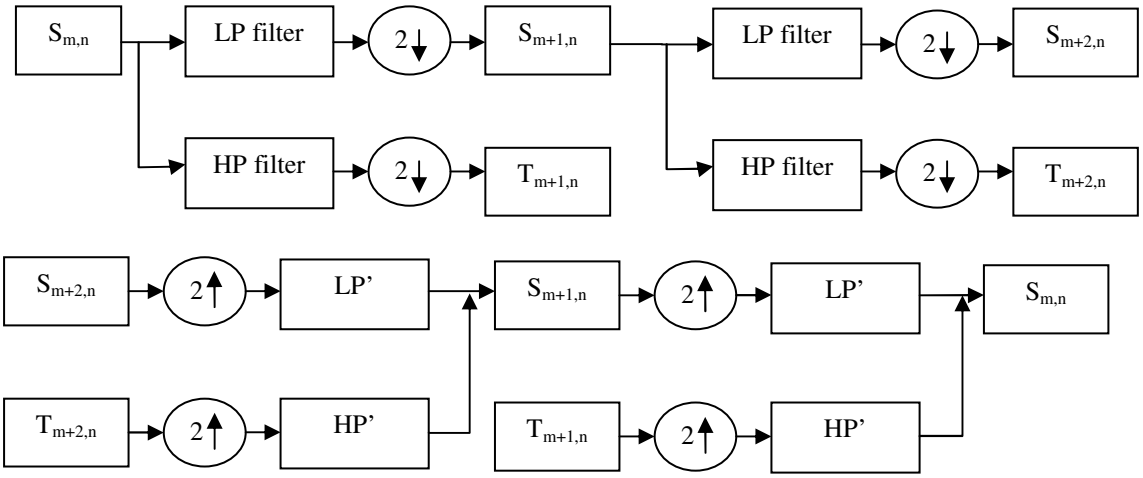


Figure 5.4. The schematic of signal filtering and reconstruction. The first block shows the approximation coefficients undergoing successive filtering and down sampling to get the next level approximation and detail coefficients and the second block shows the reconstruction of the lower level coefficients by reversing the filter coefficients and up sampling successively from the higher level coefficients (Addison 2002).

In normal cases, the input signal is discrete and of finite length. The most common approach is to extend the signal to a power of 2 and feed it to the multi resolution filters as $S_{0,n}$ approximation coefficients.

There is no loss of information during wavelet decomposition. This means that the transformation has effectively compressed the information content of the signal towards the approximation coefficients. This property of wavelets has made them a very power tool for data compression and feature extraction applications(Addison 2002). The ability to selectively reconstruct the signal using altered or chosen coefficients makes the wavelet analysis extremely useful for dealing with noise in the signals.

5.4. Smoothing using wavelets

Smoothing is illustrated in the examples shown below. All the detail coefficients which account for fluctuations in the signal are removed.

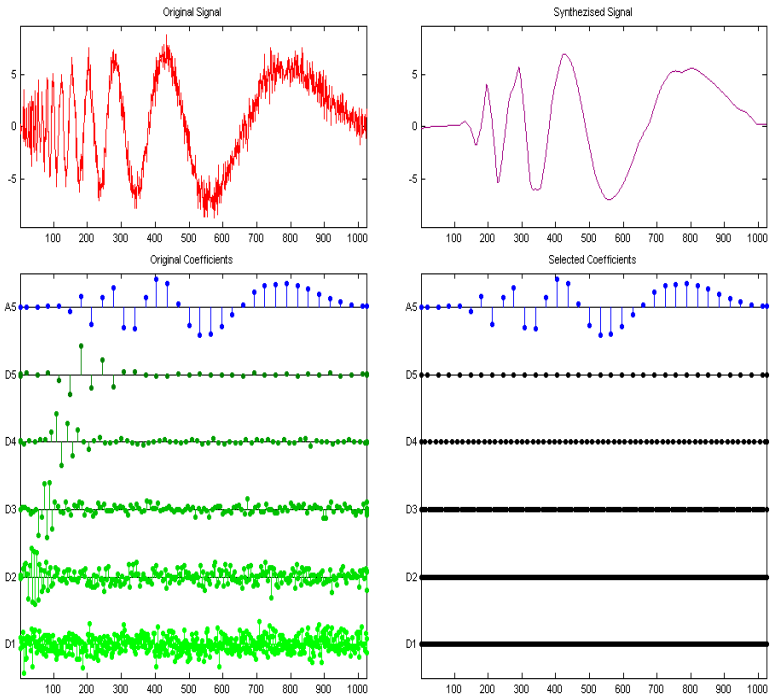


Figure 5.5. Wavelet smoothing operation. The panel on the left side shows a noisy sine wave decomposed by using sym-level 4 wavelet to the 5th level. In the right panel, only the 5th level approximation coefficients are used for reconstructing the signal. (Illustrated using wavelet toolbox of MATLAB®).

This shows how smoothing operation is carried out by choosing to reconstruct the signal using coefficients less than a threshold level.

$$T_{m,n} = \begin{cases} 0 \; if \; m \geq threshold \; 'm' \\ T_{m,n} \; otherwise \end{cases} \tag{5.27}$$

We use Daubechies wavelet – D8 and Coiflet wavelet – Coif3 for the gene selection process. The wavelets and their corresponding filters are illustrated below using MATLAB wavelet toolbox.

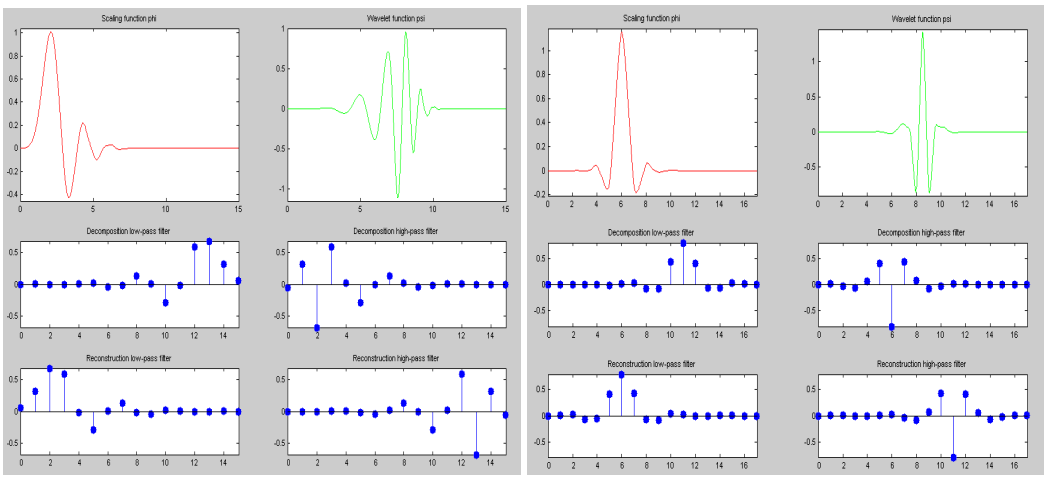


Figure 5.6. The scaling and wavelet functions of Db8 (Left) and Coif3 (Right) wavelets with their corresponding filter coefficients. (Illustrated using MATLAB® wavelet toolbox).

# CHAPTER VI

## K-NEAREST NEIGHBOR CLASSIFIER

The K-Nearest Neighbor is a very simple procedure for classifying samples based on the majority voting among their neighbors is implemented as below.

1. Each test samples are taken one at a time.
2. The Euclidean distance from the test sample to all the training samples in estimated.
3. The class labels of the K – Nearest training samples of the test labels are identified where K is always taken as an odd number.
4. The test sample is assigned to the class which has the majority number of training samples in the K- Nearest Neighbors.

The Euclidean Distance from test sample 'Test' having 'G' features to training sample 'Train' is given by:

$$Ed = \sqrt{\sum_{i=1}^{G} (Test(i) - Train(i))^2} \tag{6.1}$$

The 3-Nearest Neighbor method used in this thesis is illustrated below for a classification problem with 2 features.

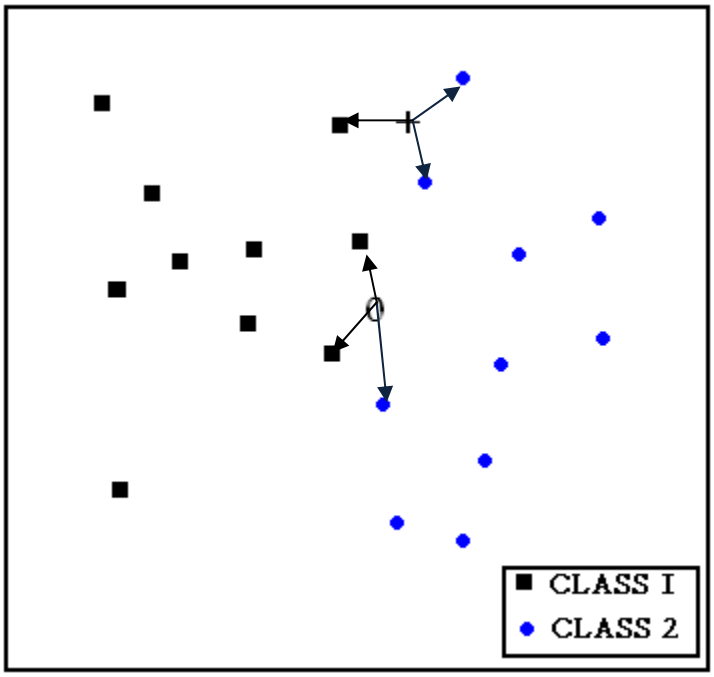


Figure 6.1. Illustration of 3-NN classifiers. The Euclidean distance from the test samples (the '+' and the '0' samples) to the entire training set (the square samples belonging to class 1 and circular samples belonging to class 2). The three samples closest to the test samples are identified and the class labels of theses samples are put to vote. If 2 from the three nearest neighbors belong to class 1, the sample is assigned to class 1, otherwise to class 2.

# CHAPTER VII

# THE METHOD

## 7.1.Preprocessing

A portion of the Leukemia dataset(Golub T, 1999), which is a typical publicly available dataset, is shown below to illustrate the arrangement of data points in the datasets.

Table 7.1. Illustration of a typical microarray dataset. The 15 genes are arranged along the rows. The 3 Acute Lymphoblastic Leukemia (ALL) samples are along the columns.

| | Description | Accession | ALL_19769_B-cell | | ALL_23953_B-cell | | ALL_28373_B-cell | |
|---|---|---|---|---|---|---|---|---|
| 2 | | CH1999021515AA | | | CH1999021511AA/scale factor=1 | | CH1999021507AA/scale factor=1 | CH1999021312A |
| 3 | 7129 | | | | | | | |
| 4 | GABAa receptor alpha-3 subunit | A28102_at | 151 | A | 484 | A | 118 | P |
| 5 | Osteomodulin | AB000114_at | 72 | A | 61 | A | 16 | A |
| 6 | mRNA | AB000115_at | 281 | A | 118 | A | 197 | M |
| 7 | Semaphorin E | AB000220_at | 36 | A | 39 | A | 39 | A |
| 8 | MNK1 | AB000409_at | -299 | A | -11 | A | 237 | P |
| 9 | VRK1 | AB000449_at | 57 | A | 274 | P | 311 | P |
| 10 | VRK2 | AB000450_at | 186 | P | 245 | P | 186 | P |
| 11 | mRNA, clone RES4-22A | AB000460_at | 1647 | P | 2128 | P | 1608 | P |
| 12 | SH3 binding protein, clone RES4-23A | AB000462_at | 137 | A | -82 | A | 204 | P |
| 13 | mRNA, clone RES4-24A, exon 1, 2, 3, 4 | AB000464_at | 803 | P | 1489 | P | 322 | P |
| 14 | mRNA, clone RES4-24C, exon 1, 2, 3 | AB000466_at | -894 | A | -969 | A | -444 | A |
| 15 | mRNA, clone RES4-25, partial cds | AB000467_at | -632 | A | -909 | A | -254 | P |

The preprocessing involves thresholding the data. This involves making all the values below a lower threshold value equal to the lower threshold value and making all values above an upper threshold value equal to the upper threshold value.

The datasets were thresholded at the levels used in the original papers. In particular, the Leukemia dataset was thresholded at 100 and 16,000. The B-cell Lymphoma dataset and the Colon dataset were each thresholded at 20 and 16,000.

The thresholded data is then normalized to the range 0-1 by the relation:

$$e'_{ij} = \frac{e_{ij} - \min_{1 \le j \le M}\{e_{ij}\}}{\max_{1 \le j \le M}\{e_{ij}\} - \min_{1 \le j \le M}\{e_{ij}\}} \tag{7.1}$$

where $e_{ij}$ is the expression value of gene $i$ in sample $j$, $e'_{ij}$ stands for the normalized expression value of gene $i$ in sample $j$ and $M$ represents the number of samples in the dataset.

7.2. The gene selection method

1. The Preprocessed and Normalized data has the gene expression values of the samples arranged along the columns with each row corresponding to the genes (An illustration is included below).

Table 7.2. A sample of the preprocessed and normalized data-points from the leukemia dataset.

| | sample 1 | sample 2 | sample 3 | sample 4 | sample 5 | sample 6 | sample 7 | sample 8 | sample 9 | sample 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| gene 1 | 0.0032 | 0.0241 | 0.0011 | 0.0225 | 0.0485 | 0.0162 | 0 | 0 | 0.0094 | 0.0004 |
| gene 2 | 0.0113 | 0.0011 | 0.0061 | 0.0042 | 0.0123 | 0.0053 | 0.0006 | 0.0046 | 0.0030 | 0.0052 |
| gene 3 | 0 | 0 | 0.0086 | 0 | 0 | 0.0018 | 0 | 0.0028 | 0 | 0.0015 |
| gene 4 | 0 | 0.0109 | 0.0132 | 0.0019 | 0 | 0.0036 | 0.0547 | 0.0250 | 0.0015 | 0.0167 |
| gene 5 | 0.0442 | 0.0873 | 0.0139 | 0.0194 | 0.0612 | 0.0332 | 0.0543 | 0.0281 | 0.0357 | 0.0646 |
| gene 6 | 0.0174 | 0.0104 | 0.0285 | 0.0133 | 0.0086 | 0 | 0.0526 | 0.0747 | 0.0161 | 0.0245 |
| gene 7 | 0.0160 | 0.0097 | 0.0294 | 0.0114 | 0.0122 | 0.0079 | 0.0418 | 0.0116 | 0.0232 | 0.0097 |
| gene 8 | 0.0659 | 0.0615 | 0.0252 | 0.0790 | 0.1043 | 0.0346 | 0.0524 | 0.0162 | 0.0457 | 0.0437 |
| gene 9 | 0.0036 | 0.0086 | 0.0050 | 0 | 0.0027 | 0 | 0.0081 | 0.0028 | 0.0022 | 0.0154 |
| gene 10 | 0.0414 | 0.0605 | 0.0454 | 0.0608 | 0.0852 | 0.0410 | 0.0394 | 0.0530 | 0.0342 | 0.1636 |
| gene 11 | 0.0283 | 0.0348 | 0.0412 | 0.0230 | 0.0574 | 0.0128 | 0.0491 | 0.0518 | 0.0238 | 0.0666 |

Table 7.2. A sample of the preprocessed and normalized data-points from the leukemia dataset. (Continued).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| gene 12 | 0.0306 | 0.0347 | 0.0565 | 0.0417 | 0.0323 | 0.0060 | 0.0337 | 0.0320 | 0.0344 | 0.0320 |
| gene 13 | 0.0018 | 0.0029 | 0.0089 | 0.0006 | 0 | 0 | 0.0232 | 0.0074 | 0.0023 | 0.0264 |
| gene 14 | 0.0646 | 0.0505 | 0.0589 | 0.0538 | 0.0418 | 0.0200 | 0.0794 | 0.0472 | 0.0242 | 0.0815 |
| gene 15 | 0.0107 | 0.0197 | 0.0186 | 0.0250 | 0.0270 | 0.0159 | 0.0123 | 0.0179 | 0.0125 | 0.0161 |

2. The Preprocessed samples are grouped such that samples belonging to each class are arranged together.

3. The expression data corresponding to each gene are decomposed using the 1-dimensional discrete wavelet transform using the selected wavelets to the third level.

4. All the detail coefficients are filtered out and the signal is reconstructed using just the approximation coefficients.

5. An absolute value of the difference between the mean of the reconstructed signal in each class is taken as the score of the gene.

$$\text{Score } (gene_i) = \left| \frac{1}{n_1} \sum_{j \in C_1} e_{ij}^{l} - \frac{1}{n_2} \sum_{j \in C_2} e_{ij}^{l} \right| \tag{7.2}$$

where $e_{ij}^{l}$ represents the expression level of gene $i$ in sample $j$ after passing the $l^{\text{th}}$ level of the wavelet filter (we note that in this study $l$ is taken to be 3); $n_1$ and $n_2$ stand for the number of samples in class 1 and 2 respectively; $C_1$ and $C_2$ represent the two sets of samples from class 1 and 2 respectively.

6. All the genes are ranked according to their corresponding scores and the required number of genes are selected from the list.

7. The samples are classified using a k-nearest neighbor based classifier.

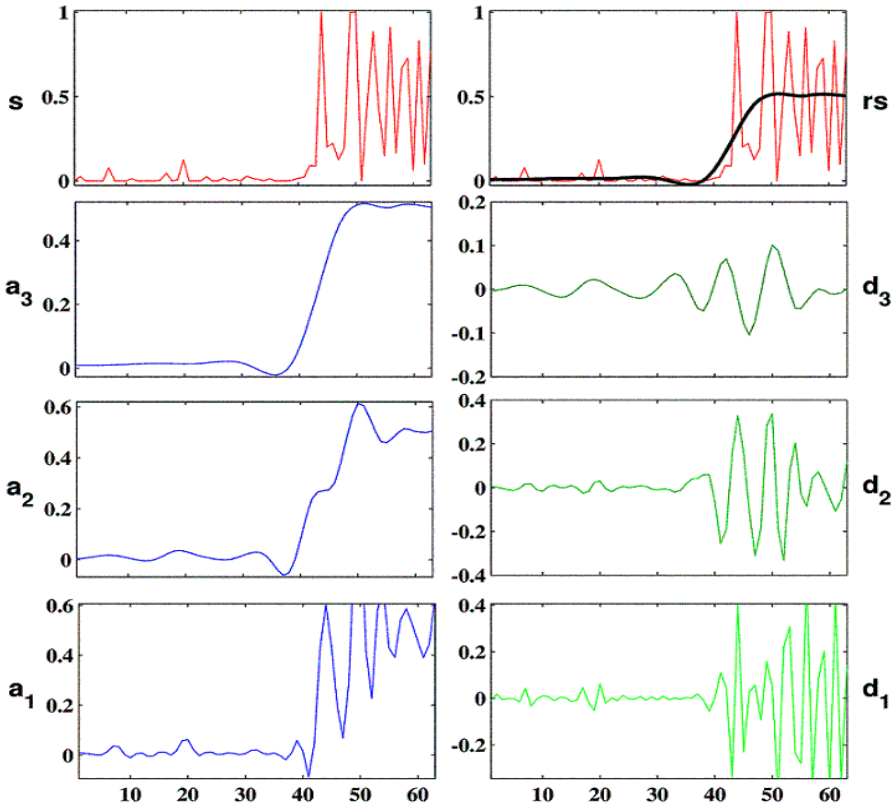


Figure 7.1. The plot shows the gene scoring process for gene CST3 (Cystatin C). The expression levels of CST3 in different samples are arranged into two groups along the *x*-axis based on their classes. The signal is decomposed using Db-8 wavelet. (S) shows the original expression levels of CST3. (a1), (a2), and (a3) represent the approximations of the signal at level 1, 2, and 3 respectively, when using Db-8 wavelets. (RS) illustrates the overlay of the original signal (S) and its level 3 approximation (a3). (d3), (d2) and (d1) represent the detail signals at their corresponding levels.

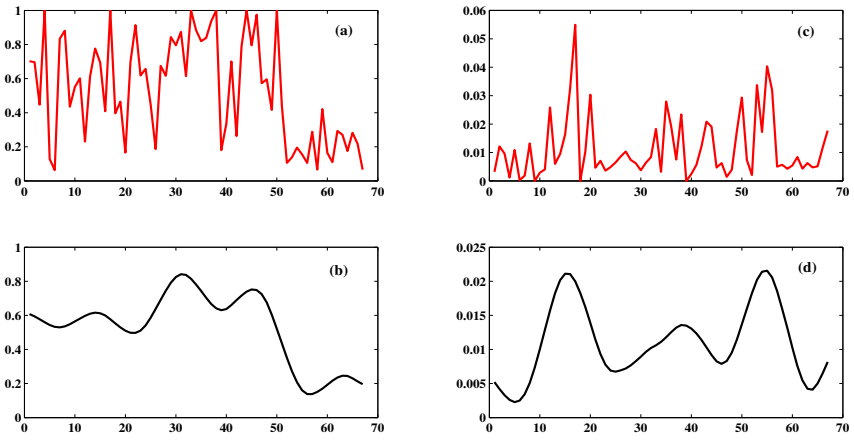


Figure 7.2. The plot illustrates the original expression signal of an informative gene LDHA (Lactate dehydrogenase A) (a) and its 3rd level approximation obtained using db-8 wavelet (b). The original expression signals of a non-informative gene SEMA3C (Semaphorin E) and its 3rd level approximation are shown in (c) and (d) respectively.

7.3. Application

The proposed method was evaluated using three publicly available datasets. The first two datasets were obtained from the Broad Institute and the third one was obtained from the Princeton University.

Table 7.3. The Datasets studied with the two classes of samples and the Number of Samples in each Class.

| **Dataset** | **Class 1** | **Number of Samples in Class 1** | **Class 2** | **Number of Samples in Class 2** |
|---|---|---|---|---|
| The Leukemia dataset (Golub T, 1999) | Acute Lymphoblastic Leukemia | 47 | Acute Myeloid Leukemia | 25 |
| The B-Cell Lymphoma dataset (Shipp, 2002). | Diffuse Large B-Cell Lymphoma | 58 | Follicular Lymphoma | 19 |
| Colon Cancer dataset (U. Alon, 1999) | Normal | 22 | Tumor | 40 |

7.4. Experiments

A series of experiments were carried out to compare the performance of the proposed gene selection method with the two standard methods – t-test and sum of square, study the genes selected and for testing the stability of the proposed method.

7.4.1. Classification performance of the different methods

The classification accuracy of the genes selected by the proposed method was verified using 3NN classifiers (Richard O. Duda, 2001) on the datasets. A confusion matrix was estimated for each sub-sampling step and for each increment of gene size (variable size). The confusion matrices were used to evaluate the classification performance of each gene selection methods using three parameters:

7.4.1.1. Confusion matrix and the measures of classification performances

The confusion matrix is very commonly used to study the classification performances of classifiers. The confusion matrix is illustrated below.

Table 7.4. The Confusion matrix.

| PREDICTED VALUES | ACTUAL VALUE | |
|---|---|---|
| | POSITIVE | NEGATIVE |
| POSITIVE | True Positive | False Positive |
| NEGATIVE | False Negative | True Negative |

1. The Sensitivity: It is a measure of how many of the positive samples have been identified as positive.

$$\text{Sensitivity} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \tag{7.3}$$

2. The Specificity: It is a measure of how many of the negative samples have been identified as negative.

$$\text{Specificity} = \frac{\text{True Negative}}{(\text{True Negative} + \text{False Positive})} \tag{7.4}$$

3. The Accuracy of the classification:

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{TrueNegative} + \text{False Positive} + \text{False Negative})} \tag{7.5}$$

The classification performances were compared to a standard parametric filtering method – 'T-test' and a standard non-parametric filtering method – 'Between Within method' (Jafari, 2006).

The evaluation was carried out using the MonteCarlo-Cross Validation (Sub-Sampling) strategy (A.-L.Boulesteix, 2007). 'N minus10' samples, randomly sampled from the data, are used as training data to select the relevant genes and the remaining 10 samples are used to evaluate the classification performance of the selected genes on test samples. The sub-sampling was repeated 250 times for each dataset.

### 7.4.2. Shuffling test

One of the main reasons for not using the 1-dimensional wavelet transform directly for gene selection applications is the presumed dependence of the wavelet coefficients of a signal on the order in which the samples are arranged within each class. We studied the effect of the order in which samples are arranged within groups on the proposed method by shuffling the pre-grouped training data within each class 100 times each.

### 7.4.3. Gene study

The genes selected the most number of times during the sub-sampling process were studied. The selected genes were compared with the top 100 genes selected by the standard methods and the most important genes identified in the original papers where the datasets were first discussed.

# CHAPTER VIII

# RESULTS AND DISCUSSION

## 8.1. Classification results

Sensitivity: The average sensitivity of 'Db-3', 'Db-8' and 'Coif-3' wavelets along with the standard methods for the three datasets are plotted below.

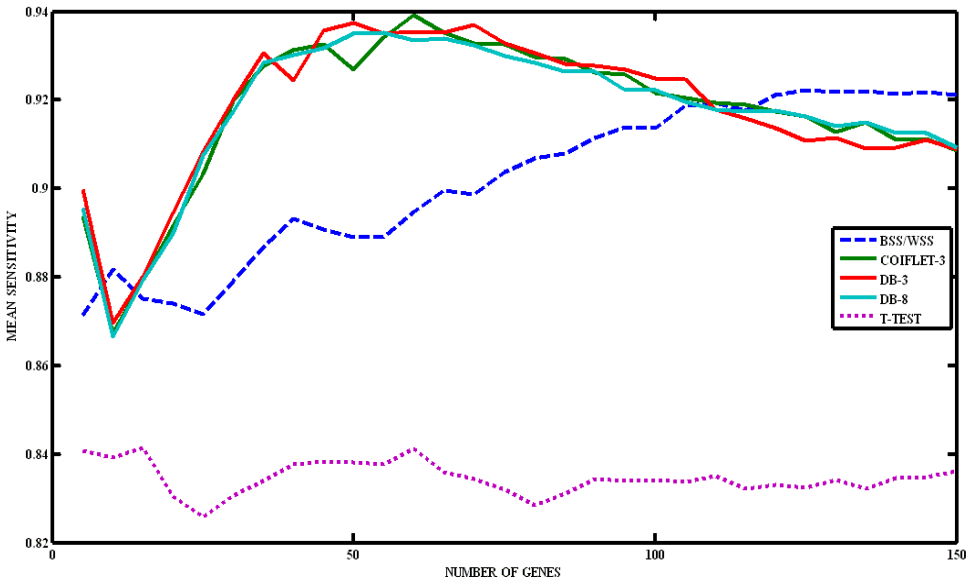


Figure 8.1. Mean Sensitivity for the different methods for the B-cell Lymphoma Dataset.

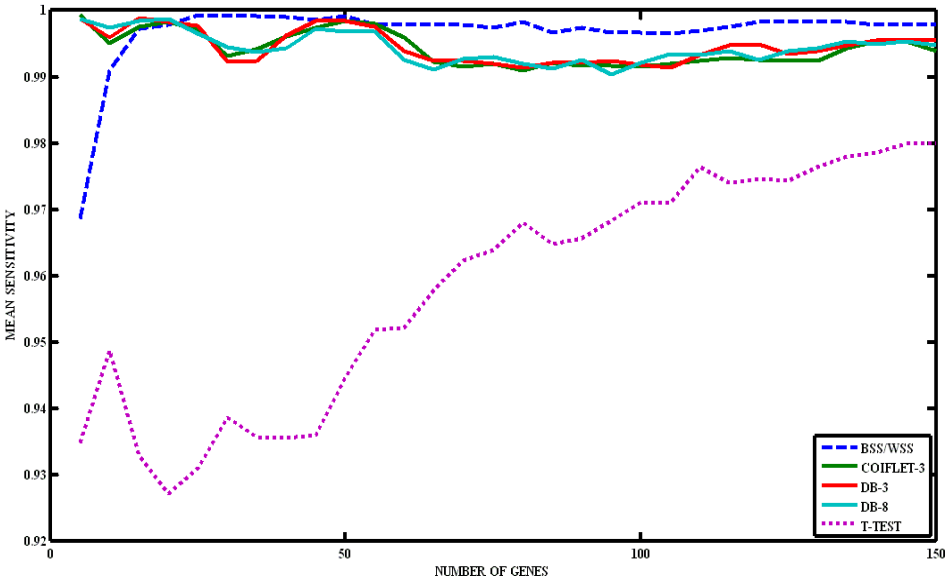


Figure 8.2. Mean Sensitivity for the different methods for the Leukemia Dataset.

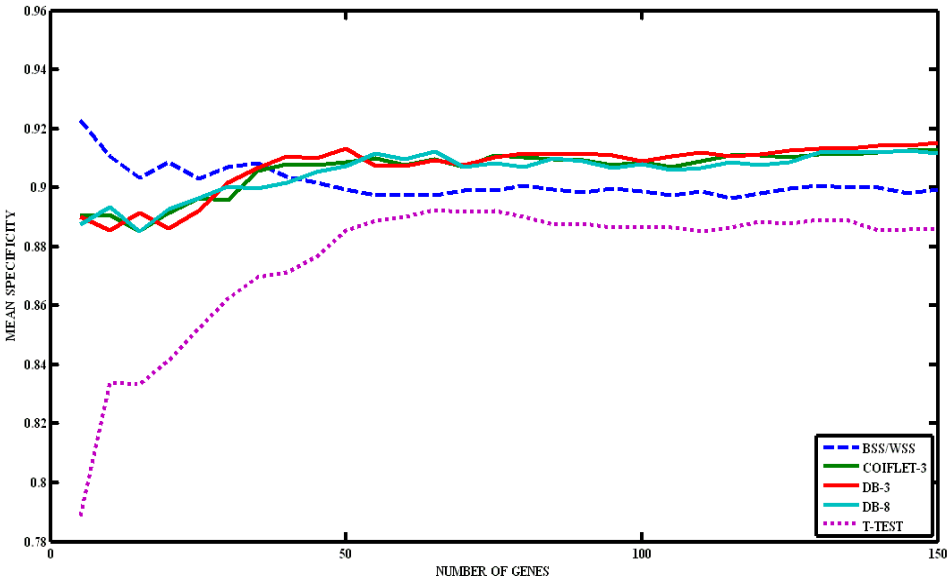


Figure 8.3. Mean Sensitivity for the different methods for the Colon Cancer Dataset.

Specificity: The average sensitivity of the best three wavelets along with the standard methods for the three datasets is plotted below.

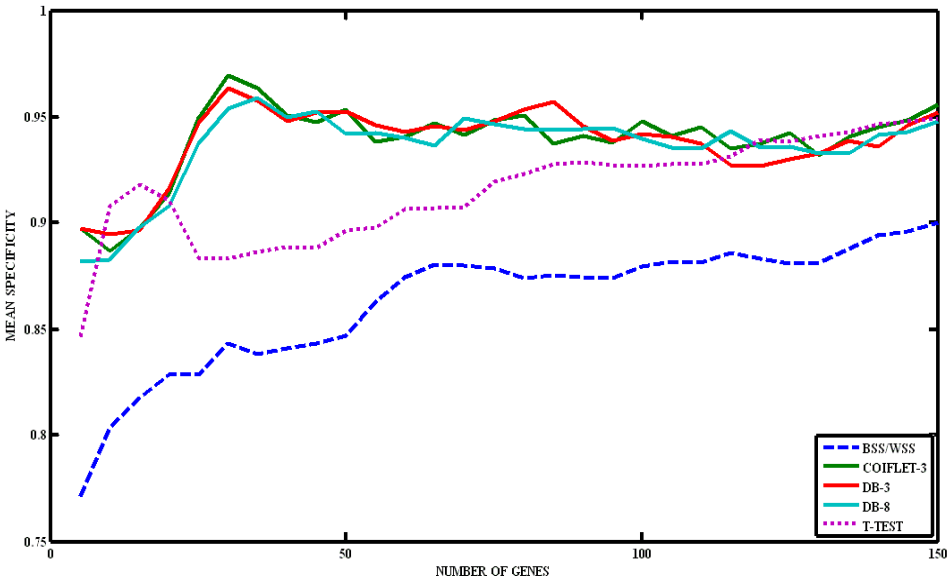


Figure 8.4. Mean Specificity of the different methods for the B-cell Lymphoma Dataset.

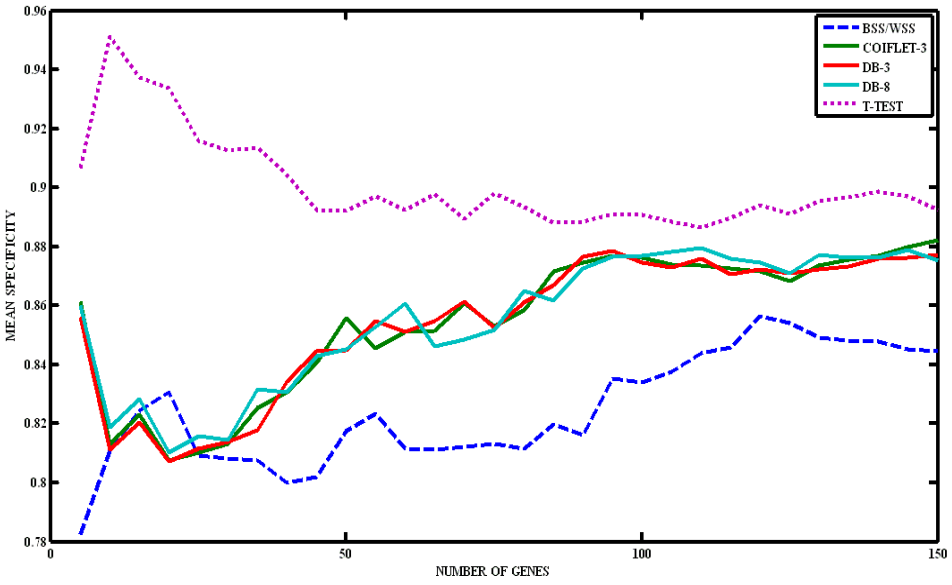


Figure 8.5. Mean Specificity of the different methods for the Leukemia Dataset.

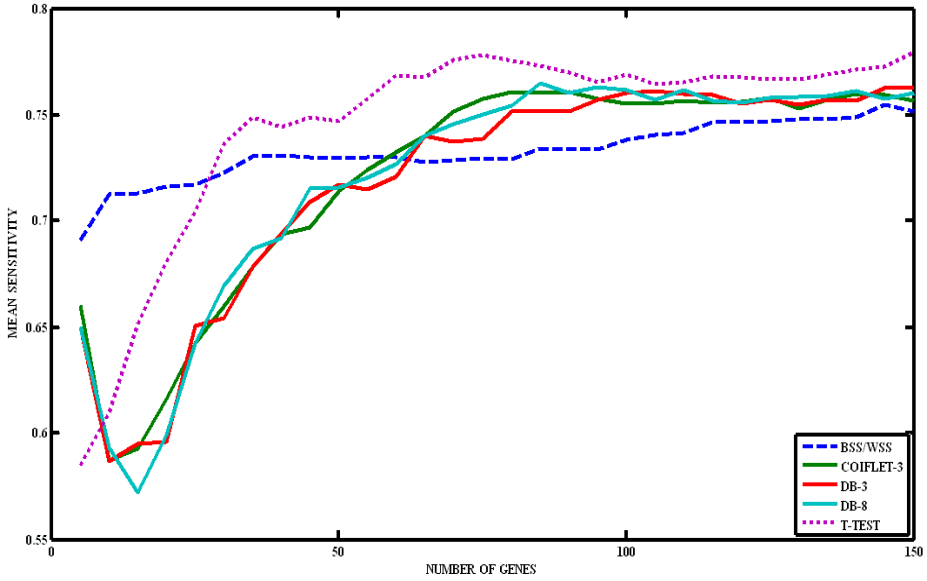


Figure 8.6. Mean Specificity of the different methods for Colon Cancer Dataset.

Mean classification accuracy: The average classification accuracy of the best three wavelets along with the standard methods for the three datasets is plotted below.

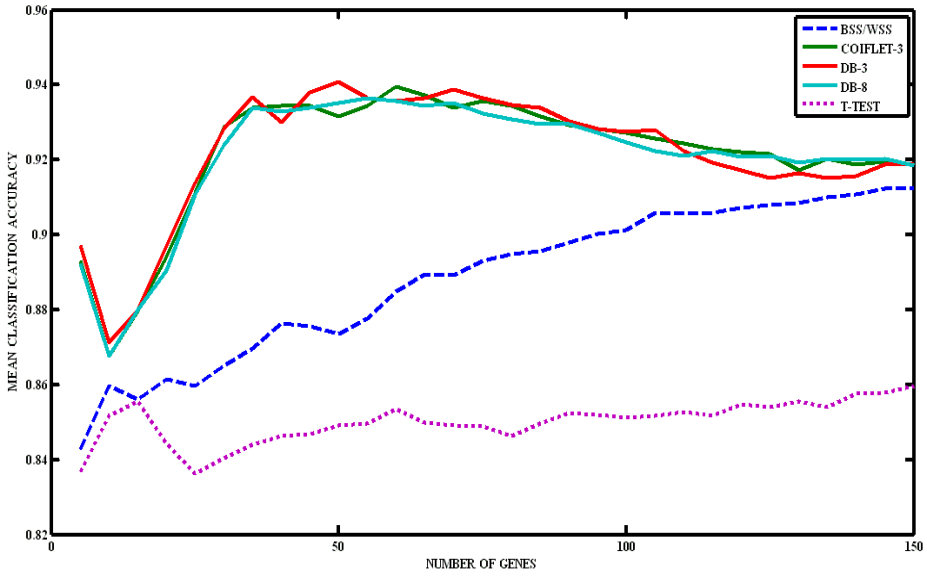


Figure 8.7. Mean classification Accuracy of the different methods for the B-cell Lymphoma Dataset.

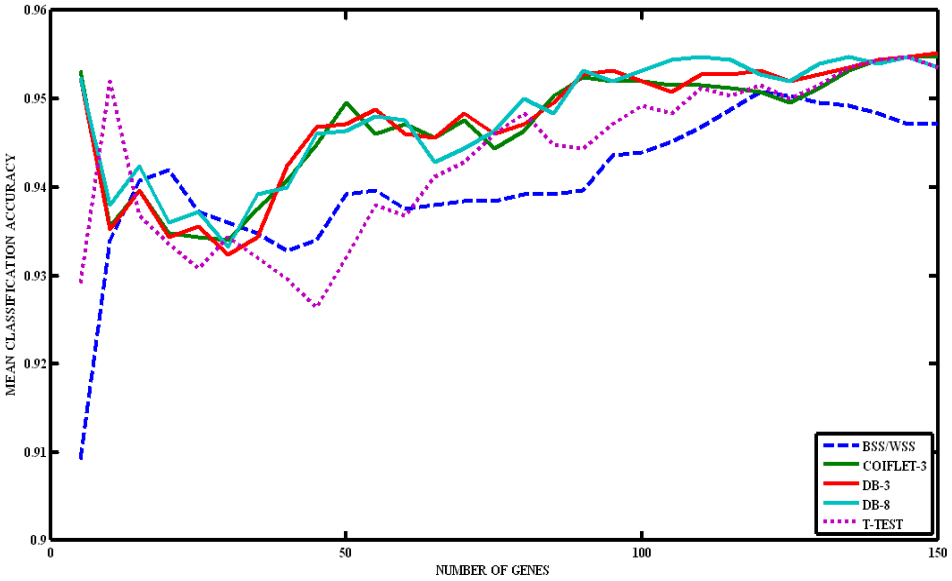


Figure 8.8. Mean classification Accuracy of the different methods for the Leukemia Dataset.

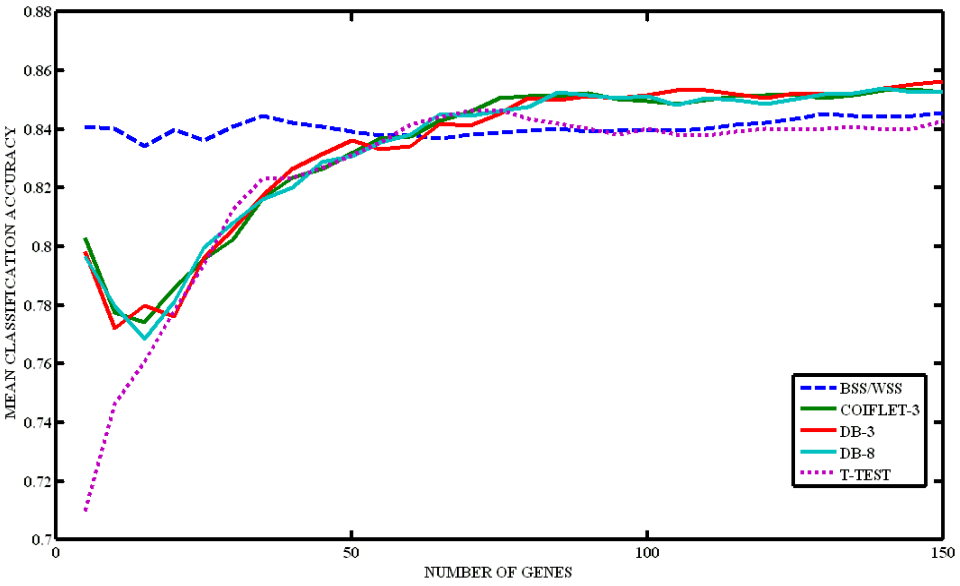


Figure 8.9. Mean classification Accuracy of the different methods for Colon Cancer Dataset.

The classification performance of the different gene selection method seems to depend on the different datasets. The sensitivity and specificity measures were used to study the performance on each class of samples, one at a time. For finding the sensitivity and specificity measures, one of the classes has

to be defined as positive and the other as negative. The table 8.1 shows the definition of the classes as positive and negative for each of the datasets.

Table 8.1. Division of Classes into Positive and Negative for estimating the Sensitivity and Specificity. The number of samples in each classes are indicated in the brackets. The ratio of true positives to total number of positives results is the sensitivity and the ratio of true negatives to the total number of negatives is specificity.

| Dataset | Positive | Negative |
|---|---|---|
| Leukemia Dataset | ALL(47) | AML(25) |
| B-Cell Lymphoma Dataset | B-Cell Lymphoma(58) | Follicular Cancer (19) |
| Colon Cancer | Tumor(40) | Normal(22) |

A two sample t-test was used to compare the classification accuracy, the sensitivity and the specificity terms between the sum of squares, the t-test and the wavelet methods in all the three datasets. The t-test was performed at 30 levels of gene size (from 5 to 150 in steps of 5). The results of the statistical analysis are provided in the Appendix B.

The sum of squares method has high sensitivity when compared to the t-test while the t-test outperforms the sum of squares method in terms of its specificity. All the positive classes have higher number of samples compared to the negative classes. The performance of the t-test and the Sum of squares method, therefore seems to depend on the number of samples in the classes being studied. This is clearly illustrated in the One-Way ANOVA and Tukey's HSD Test results (Appendix B) of B-cell Lymphoma and Leukemia datasets . The wavelet based method, on the other hand, consistently gives a more stable classification performance in most of the cases.

Sensitivity: The sensitivity term represents the performance of the three methods in classifying the positive classes (the ALL class in the Leukemia data, the B-Cell Lymphoma class in the Lymphoma dataset and the Tumor class in the Colon cancer dataset) of the datasets.

The One way ANOVA and Tukey’s HSD tests were conducted for comparing the sensitivity of the wavelet based method, the sum of squares method and the t-test. For the Lymphoma dataset, the sum of squares and the wavelets have significantly higher sensitivity when compared to the t-test at all the 30 levels of gene size (5 to 150 in steps of 5). The wavelet method has significantly higher sensitivity when

compared to the sum of squares method when the number of genes is between 25 and 75. For Leukemia data, the sum of squares method and the wavelet method again always had significantly higher sensitivity compared to the T-test. In the case of the Colon cancer data, the difference is not as clear as the other two datasets. The sum of squares and the wavelet methods has significant advantage over the t-test till around 45 genes after which there is no significant difference. There is no significant difference between the wavelet method and the sum of squares method except at 5 genes, when, the sum of squares method gave a significantly higher sensitivity compared to the wavelet method.

Specificity: The specificity term represents the performance of the three methods in classifying the negative classes (the AML class in the Leukemia data, the Follicular cancer class in the Lymphoma dataset and the Normal class in the Colon cancer dataset) of the datasets.

The One way ANOVA and Tukey's HSD tests were again used for comparing the specificity of the wavelets, the sum of squares and the t-test methods. As mentioned earlier, there was a switch in the performance of the sum of squares and the t-test methods. For the lymphoma data, the T-test significantly outperformed the sum of squares method when the number of genes ranged from 5-25and 80-150 and also at ,35,40,50. There were no significant difference for the gene sizes in between. The wavelet method always significantly outperformed the sum of squares method while its sensitivity was significantly better than t-test at some cases (number of genes from 25 to 55 and 70). There was no significant difference between the specificity of the wavelet and the t-test at the remaining gene sizes. For Leukemia data, like in the case of the Lymphoma data, the t-test showed significantly higher specificity when compared to the sum of squares method in all cases. The wavelet method outperformed the sum of squares method in five cases (for gene sizes – 5, 60, 80, 90, 100). The t-test gave significantly higher specificity compared to wavelets methods in 12 out of the 30 gene sizes (5 to 55, 65 and 75). In the remaining gene sizes, there was no significant difference in specificity between the wavelets and the t-test. There is no significant difference between the specificity of the three methods in the colon cancer data for most gene sizes. But, the pattern of switching between the sum of squares and t-test methods is clearly followed here also as can be seen on the plot. (figure 8.6). The sum of squares has significant gain over the other two methods when the gene size was less than 25.

Overall Accuracy: When it comes to overall accuracy, the wavelet method outperforms both the other methods in the case of B-cell Lymphoma dataset. But in the case of the other two datasets, there is no significant difference between the studied methods as illustrated in the statistical results (Appendix B).

In the case of B-cell Lymphoma data, the wavelet method significantly outperforms the sum of squares method when the gene size ranges from 5 to 100 (after which there is no significant difference). It outperforms the t-test in all cases except when the gene size is 10. The clear advantage in the B-cell Lymphoma is not observed in the other two datasets. Even though there is no statistically significant improvement, the average classification performs is at least as good as the other two methods in most of the cases as can been seen in the plots 8.7, 8.8 and 8.9. In the case of Leukemia, the classification accuracy is at least as good as that of the sum of squares method. The wavelet method has a significant advantage over the sum of squares method when the number of genes is 5 and 90. It has significant advantage over t-test when the number of genes is 5 and 45. In the case of colon data, the sum of squares has a significant advantage over the wavelet method till when the gene size is 35 after which there is no significant difference between the two types. The wavelet method outperforms the t-test for 5 and 10 genes after which there is no significant difference. Even though, there is no significant difference between the classification results of the different types, the wavelets gave the highest mean accuracy (86%) when compared to the sum of squares method and the t-test as illustrated in the figure 8.9. It has to be noted that the wavelet based method gave the highest average accuracy in all the three datasets (95.88% at 35 genes for B-Cell Lymphoma Data, 95.48% at 110 genes for Leukemia data and 85.4% at 140 genes for the colon cancer dataset).

## 8.2. Gene study

A record was kept of the genes selected in each of the 250 sub-samplings and a study was done of the top 100 genes which frequented the lists with the highest scores. The lists are included in the appendix. A list of the top 10 genes identified by the wavelet method from each of the three datasets are tabulated below.

Table 8.2. The top 10 genes from the three datasets identified by the Db-8 wavelet with their scores.

| Rank | Leukemia dataset | | Lymphoma dataset | | Colon cancer dataset | |
|---|---|---|---|---|---|---|
| | Gene Symbol | Average score | Gene Symbol | Average score | EST ID | Average score |
| 1 | CST3 | 0.475 | MTL5 | 0.443 | T95018 | 0.236 |
| 2 | MPO | 0.404 | LDHA | 0.428 | M63391 | 0.199 |
| 3 | AZU1 | 0.387 | ENO1 | 0.407 | T58861 | 0.170 |
| 4 | IL8 | 0.384 | CTSB | 0.370 | T61609 | 0.168 |
| 5 | FTL | 0.340 | PKLR | 0.318 | T92451 | 0.153 |
| 6 | GPX1 | 0.297 | PAPLN | 0.310 | U14971 | 0.141 |
| 7 | TCL2 | 0.294 | CLU | 0.305 | T57619 | 0.137 |
| 8 | CFD | 0.292 | MIF | 0.298 | R22197 | 0.133 |
| 9 | LYZ | 0.288 | IFI30 | 0.296 | M87789 | 0.132 |
| 10 | SDC1 | 0.285 | APOE | 0.292 | T72863 | 0.131 |

8.2.1. Leukemia data

2168 (out of 7070) genes left after preprocessing were scored by the wavelet method. Of the top ten genes CST3, IL8, AZU1, LYZ have been identified as some of the most important genes in distinguishing ALL from AML in the original study (Golub T, 1999). Myeloperoxidase and TCL2 have also been identified to be associated with AML and ALL respectively. Most of the genes reported as important in the original study were ranked very highly by the wavelet-based method and featured in the top 100 genes list in all the 100 resamplings. The list of top 100 genes with their corresponding scores are listed in the appendix.

8.2.2. Lymphoma data

2814 (out of 7070) genes left after preprocessing were scored using the wavelet method. Several of the genes found to be informative by the original study were ranked very highly by the wavelet-based method. LDHA (lactose dehydrogenase A) is a known biomarker for B-cell lymphoma as reported in the original paper (2). CTSB, CLU and ENO1 in the top 10 list had also been identified as very important. Many of the genes identified as relevant in the original study received very high scores and most of them featured in the top 100 list in all the 250 re-sampling studies. For example HMG-1 ranked $12^{th}$ and CTSD ranked $16^{th}$. The list of the top 100 genes and their corresponding scores are listed in the appendix.

8.2.3. Colon data

In the case of colon data, the original study reports the importance of soft muscle related ESTs in identifying colon cancer. They use the average intensity levels of 17 ESTs as a muscle index, with a lower index identifying tumors (3). The wavelet-based method gave high scores to most of these 17 ESTs. The list of top 100 genes and their corresponding scores are listed in their appendix.

The top 100 genes were compared with the ones identified by t-test and BSS/WSS. A matrix showing the numbers of common genes selected by different methods for the three datasets is presented in Table 6. As we can see, almost 100% of the genes found by the three wavelets are the same, indicating the consistency of the three wavelet methods. About 32% of the top 100 genes obtained by the wavelet based methods and the t-test were found to be common in the three datasets. About 38 % of the top 100 genes obtained by the wavelet based methods and BSS/WSS were common. On the other hand, the number of common genes selected by the t-test and the BSS/WSS method is much higher. The relative low percentage of common genes identified by wavelet based methods and by t-test or BSS/WSS indicates that the wavelet based methods offer a different perspective in terms of differentially expressed genes. Therefore, the proposed wavelet based gene selection method can facilitate the identification of differentially expressed genes which might be otherwise neglected. At the same time, the genes which are common in all the three methods are definitely very important for the problem being studied.

Table 8.3. Number of genes in the list of top 100 genes, common to different methods.

| Feature Selection method | Leukemia dataset | | | | |
|---|---|---|---|---|---|
| | db-3 | db-8 | coif-3 | BSS/WSS | T-test |
| db-3 | | 99 | 100 | 42 | 31 |
| db-8 | 99 | | 99 | 42 | 31 |
| coif-3 | 100 | 99 | | 43 | 31 |
| BSS/WSS | 42 | 42 | 43 | | 53 |
| T-test | 31 | 31 | 31 | 53 | |
| Feature Selection method | Lymphoma dataset | | | | |
| | db-3 | db-8 | coif-3 | BSS/WSS | T-test |
| db-3 | | 99 | 99 | 32 | 30 |
| db-8 | 99 | | 99 | 33 | 30 |
| coif-3 | 99 | 99 | | 32 | 30 |
| BSS/WSS | 32 | 33 | 32 | | 65 |
| T-test | 30 | 30 | 30 | 65 | |

Table 8.3. Number of genes in the list of top 100 genes, common to different methods. (continued).

| Feature Selection method | Colon cancer dataset | | | | |
|---|---|---|---|---|---|
| | db-3 | db-8 | coif-3 | BSS/WSS | T-test |
| db-3 | | 99 | 99 | 34 | 33 |
| db-8 | 99 | | 100 | 33 | 33 |
| coif-3 | 99 | 100 | | 33 | 33 |
| BSS/WSS | 34 | 33 | 33 | | 82 |
| T-test | 33 | 33 | 33 | 82 | |

Table 8.4. The genes from the top 100 list common for all the three gene selection method studied.

| Dataset | Genes Common between Sum of squares, T-test and the wavelet method. |
|---|---|
| B-Cell Lymphoma Data | 25 |
| Leukemia Data | 25 |
| Colon Cancer Data | 30 |

8.4. Some observations

1.Several wavelets – Haar, Bior1.5, Bior 6.8, reverse Bior 1.5, reverse Bior 3.9 and Symlet8 were tried, but their performances were not as good as the three wavelets eventually used for the study- Db-3,Db-8 and Coiflet-3. The best wavelet and the best decomposition level might depend on the sample size of the dataset studied.

2. For the methods studied in this thesis, combining the method with multivariate methods made the results worse. Removing the most correlated genes from the list was expected to remove redundancy and give the same results at lower number of genes, but the results indicate otherwise. The method was also combined with sequential floating forward search, but the results became worse.

3. The classification accuracy was studied using 1-NN,3-NN, 5-NN and 3-NN gave the best results.

## 8.5. Limitations

1. Even though the sheer size of the microarray datasets has been growing at a very fast rate, the lack of standardization in the microarray assays and processing of the data makes the validation of a feature selection method very difficult.

2. The samples used for the analysis is obtained from patients whose disease state has already been diagnosed. There is no guarantee that the same genes might express differentially in the beginning of the disease state, when the diagnosis will be useful.

## 8.6. Significance of the Study

1. The proposed wavelet smoothing based method gives the gene expression signal a score, which quantifies the differential expression of a gene in the different cases studied. The study shows that wavelet transformation can be a very powerful tool for studying and quantifying differential expression of genes

## 8.7. Future works

1. A clustering can be done of the genes and those in the feature list which come from the same clusters can be selectively removed to see if it has an effect on the classification performances.

2. The method can be extended to multi-class problems. A study has to be conducted to find out the best wavelet and the best decomposition level for getting the best reconstructed gene expression signal.

# CHAPTER IX

# CONCLUSIONS

A 1-D discrete wavelet transform based gene selection method was proposed. It was developed based on the observation that the 3rd level 1-D wavelet approximation captures the differential microarray gene expression between sample classes. The genes that exhibit high differences between the average wavelet approximations of the expression levels are selected to form a feature set for sample classification. The experiments illustrate that:

(1) The results of the study rejected the null hypotheses and accepted the alternate hypotheses that the probability of the gene sets selected by the proposed wavelet transform based method making a correct classification is higher than it making an incorrect classification (Appendix A).

(2) A two sample t-test was used to compare the performance of the proposed method to two standard gene selection methods and the results (Appendix B) indicate that the method significantly outperforms the two standard methods for Lymphoma dataset, as good as the other methods and better than them in some cases for the Leukemia dataset and as good as the other methods in most cases for the colon cancer data (the sum of square method gives significantly better performance for gene sizes 5-30 for Colon dataset).

(3) The classification performance for both the classes were consistently high for the wavelet while the t-test seems to give high specificity and low sensitivity (higher performance in class with lesser number of samples) and the sum of square method gave high sensitivity and lower specificity (higher performance in class with higher number of samples) as illustrated by the sensitivity/specificity study in Appendix B .

(4) Shuffling the samples within the groups does not affect the accuracy of the classifier which shows that the methods does not depend on the order in which the samples are arranged.

(5) The study gave a short list of features (25 in case of B-cell Lymphoma data and Leukemia data and 30 ESTS in case of Colon cancer data) as illustrated in the gene study section (8.3).

(6) The wavelet analysis is a valuable tool for studying gene expression patterns. The wavelet based gene selection method can be used to identifying and quantifying patterns in gene expression DNA microarray data.

BIBLIOGRAPHY


1. A.Ben-Dor, L.a. (2000). Tissue classification with gene expression profiles. *7* (no. 3-4,pp.559-583).

2. A.K.Jain,S.a. (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *13* (3,pp.252-264).

3. A.-L.Boulesteix,C. &. (2007). *Evaluating microarray-based classifiers:an overview.* Department of Statistics, University of Munich.

4. A.S.Levenso,I. (2002). Molecular classification of selective oestrogen receptor modulators on the basis of gene expression profiles of breast cancer cells expressing oestrogen receptor alpha. *87* (no.4,pp.449-456).

5. A.Szabo,K.A. (2002). Variable selection and pattern recognition with gene expression data generated by the microarray technology. *176* (1, pp. 71-98).

6. Aalto,Y.E.R. (2001). Distinct gene expression profiling in chronic lymphcytic leukemia with 11q23 deletion. *15:1721-1728.*

7. Addison,P. S. (2002). *The Illustrated Wavelet Transform Handbook.* New York: Taylor & Francis.

8. Aggarwal Amit,H. L. ( 2005 ). Wavelet transformations of Tumor Expression Profiles Reveals a Pervasive Genome-Wide Imprinting of Aneuploidy on the Cancer Transcriptome. *65: (1).*

9. Alevizos,I.M. (2001). Oral cancer in vivo gene expression profiling assisted by laser capture microdissection and microarray analysis. *20* (6196-6204).

10. Alizadeh,A.E. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling . *403(6769):503-511.*

11. AlonU.,B.N. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA , 96*, 6745–6750.

12. Axon Instruments in. (1999). GenePix4000A User's Guide.

13. Baldi P, L. A. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *17:509-519.*

14. Beer,D.K. (2002). Gene expression proflies predict surival of patients with lung adenocarcinoma. *8(8):816-824*.

15. Bicciato,S.P. (2003). Pattern identification and classification in gene expression data using autoassociative neural network model. *81(5):594-606*.

16. BioDiscovery Inc. (1997). ImaGene.

17. Bittner,M.M.D. (2000). Molecular classification of cutaneous malignalt melanoma by gene expression profiling. *406:536-540*.

18. Bo,T.a. (2002). New feature subset selection procedures for classification of expression profiles. *research0017.1–research0017.11*.

19. Bozinov,D.a. (2002). Unsupervised techniques for robust target seperaton and analysis DNA microarray spots through adaptive pixel clustering. *18, 747-756*.

20. Brown,C.G. (2001). Image metrics in statistical analaysis of DNA microarray data. *98 8944-8949*.

21. Buckley, M. (2000). Spot User's Guide.

22. C. Sima, S. A.N. (2005). Impact of Error Estimation on feature selection. *38 (2005): 2472 – 2482*.

23. Callagy,G. C. (2003). Molecular classification of breast carcinomas using tissue microarrays. *12(1):27-34*.

24. Callow MJ, D. S. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. . *10:2022-2029*.

25. Celis, J. K. (2000). Gene expression profiling: Moitorin transcription and translation products using dna microarrays and proteomics. *480(1):2-16*.

26. Chao Sima, S. A.N. (2005). Impact of error estimation on feature selection. *38: 2472 - 2482*.

27. Christopher J. Penkett and Jurg Bahler. ( 2004). Navigating public microarray databases. *5: 471–479*.

28. Churchill,X. C. (2003). Statistical tests for differential expression in cDNA microarray experiments. *4:210*.

29. Consortium, I. H. (2004). Finishing the euchromatic sequence of the human genome. *431 (7011): 931–45*.

30. Culhane, A. P. (2002). Between-group analysis of microarray data. *18(12):1600-1608*.

31. D.J. Duggan, M. Y. (1999). Expression profiling using cDNA microarrays. *21* (Suppl 1,pp.10-14).

32. D.Zongker,A. a. (1997). Feature Selection:evaluation, application and small sample perfomance. *19* (2,pp.153-8).

33. Devilard, E. B. (2002). Gene expression profiling define molecular subtypes of classical hodkin's disease. *21:3095-3102*.

34. Dıaz-Uriarte, R. a. (2006). Gene selection and classification of microarray data using random forest. . *7, 3*.

35. Ding,C.a. (2003). Minimum redundancy feature selection from microarray gene expression data. *Proceedings of the IEEE Conference on Computational Systems Bioinformatics*, *pp. 523–528*.

36. Dougherty, U. B.N. (2004). Is cross-validation valid for small-sample microarray classification? *vol. 20* ( no. 3, pp. 374–380).

37. Dougherty, U. B.N. (2004). Bolstered error estimation. *vol. 37* (no. 6, pp. 1267–1281).

38. Dragichi,S. (2003). *Data Analysis Tools for DNA Microarrays.* CRC Press Company.

39. Dudoit, S. F. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *97:77-87*.

40. Dutta, E. R. (2005). Genomic Signal Processing: Diagnosis and Therapy. *January 2005*.

41. Dyrskjot, L. T.-D. (2003). Identifying distinct classes of bladder carcinoma using microarrays. . *33(1):90-96*.

42. E.R.Dougherty. (2001). Small sample issues for microarray based classification. *2* (1,pp.28-34).

43. Edward R. Dougherty, A. D. (2005). Research Issues in Genomic Signal Processing. *Nov 2005*.

44. Edward R.Dougherty, I. S. (2005). *Genomic Signal Processing and Statistics.* Hidwani Publishing Corporation.

45. Efron B, T. R. (2000). Microarrays and their use in a comparative experiment.

46. Eisen, M. (1999). ScanAlyze. http://rana.lbl.gov/EisenSoftware.htm.

47. Fathallah-Shaykh, H. R. (2002). Mathematical modelng of noise and discovery of genetic expression classes in gloimas . *21(47):7164-7174*.

48. Friedman, J. ( 1989). Regularized discriminant analysis,. *vol. 84* (no. 405, pp. 165–175).

49. Furey, T. C. (2000a). Support Vector Machine classification and validation of cancer tissue samples using microarray expression. *16(10:906-914*.

50. Furey, T. C. (2000b). Support Vector Machine classification and validation of cancer tissue samples using microarray expression data. *16(10):906-914*.

51. G.Callagy, E. a. (2003). Molecular Classification of breast carcinomas using tissue microarrays. *12* (1,pp.27-34).

52. Golub T, S. D. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,". *vol. 286* (No. 5439 pp. 531–537).

53. GSI Lumonics Inc. (1999). Quantarray.

54. Guyon, I. e. (2002). Gene selection for cancer classification using support vector machines. . *46 389–422.*

55. Hall, M. (1999.). *Correlation-based feature selection for machine learning.* Ph.D. Thesis Department of Computer Science, University ofWaikato.

56. Hedenfalk, I.R.D. ((2003). Molecular classification of familial non-brca1/brca2 breast cancer. *100(5):2532-2537*.

57. Hedvat,C.H.F. ( 2002). Application of tissue microarray technology to the study of non-hodgkin's and hodgkin's lymphoma. *33(10):968–974.*

58. Huang, S. (1999). Gene expression profiling, genetic networks, and cellular states . *77:469-480*.

59. I.Hedenfalk, D. (2001). Gene expression profile in hereditary breast cancer. *344* (8,pp.539-48).

60. J. Hua, Z.X. ( 2005). *Optimal number of features as a function of sample size for various classification rules* (Vols. vol. 21, no. 8, pp. 1509–1515). Bioinformatics.

61. J.Derisi, L.P. (1996). Use of cDNA patterns to analyze gene expresssion patterns in Human Cancers. *14* (no 4,pp.457-460).

62. J.Khan, J.(2001). Classification and diagnostic prediction of cancers using gene expression profiling and artiificial neural networks. *7* (6,pp,673-9).

63. J.Slansky,M. a. (2000). Comparison of algorithms that select features for pattern classifiers. *33* (1,pp.25-41).

64. Jafari,P.a. (2006). An assesment of recently published gene expression data analysis: reporting experimental design and statistical factors. *6,27*.

65. Jeffrey G.Thomas, J. M. ( 2001 ). An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *11: 1227-1236*.

66. Khan, J. W. (2001). Classification and diagnostic prediction of cancers using geneexpression profiling and artificial neural neral networks. *7:673-679*.

67. Kim, S. D. (2002). Identification of combination gene sets for glioma classification. *1:1229–1236*.

68. Kim, S. D. (2000). Multivariate measurement of gene expression relationships. *67(2):201–209.*

69. Kitahara, O. K. (2002). Classification of sensitivity or resistance of cervical cancers to ionizing radiation according to expression profiles of 62 genes selected by cdna microarray analysis. *4(4):295-303*.

70. Klevecz, R. (2000). Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition of expression microarray data. *1:186–19*.

71. Knudsen, S. (2004). *Guide to Analysis of DNA microarray data.* John Wiley and Sons,Inc., Publications.

72. Kohlmann, A. S. (2002). Diagnosis of leukemia using microarray technology (in german). *127(42):2216–2222.*

73. Kononen, J. B.P. (1998). Tissue microarrays for high throughput molecular profiling of tumor specimens. *4(7):844-847*.

74. Kooperberg,C. a. Improved background correction for spotted cDNA microarrays. *9,55-66*.

75. L.Dyrskjot,T. a. (2003). Identifying the disease cases of bladder carcinoma using microarrays. *33* (1,pp.90-0).

76. Lee, L. S. (2003). Gene selection: a bayesian variable selection approach. *19(1):90–97.*

77. Lee, M.L. T. (2004). *Analysis of Microarray Gene Expression Data.* Norwell, MA: Kluwer Academic Publishers.

78. Lettieri, T. (2006). Recent Applications of DNA Microarray Technology to Toxicology and Ecotoxicology. *114* (4–9).

79. Levenson, A. K. (2002). Molecular classification of selective oestrogen receptor modulators on the basis of gene expression profiles of breast cancer cells expressing oestrogen recptor alpha. *87(4):449-456*.

80. Li Shutao, L. C. (2006). Wavelet-Based Feature Extraction for Microarray Data Classification.

81. Li,W.A. (2002). Zipf's law in importance of genes for cancer classification using microarray data. *219(4):539–551*.

82. Lipshutz RJ, F. S. (1999). High density synthetic oligonucleotide arrays. *21* (20–4).

83. Liu,A.Z. (2002). Block principal component analysis with application to gene microarray data classification. *21(22):3465-3474*.

84. Lodish H, E.A. (2000). *Molecular Cell Biology* (Vol. 4th edition). New York: W.H. Freeman & Co.

85. Lonnstedt I, S. T. (2002). Replicated microarray data. *12:31*.

86. Luo, J. D. (2001). Human Prostrate cancer and benign prostatic hyperplasia molecular dissection by gene expression profiling. *61(12):4683-4688*.

87. M. Skurichina, R. D. (2000). K-nearest neighbors noise injection in multilayer perceptron training. *vol. 11* (2, pp. 504–511).

88. M.Bittner, P. a. (2000). Molecular Classification of cutaneous malignant melanoma by gene expression profiling . *406* (6795,pp.536-540).

89. Mark Schena, D. S. (October 1996). Paralell Human Genome Analysis: Microarray based expression monitoring of 1000 genes. *Vol. 93* (pp. 10614-10619).

90. Moler, E. C. (2000). Analysis of molecular profile data using regenrative and discriminative methods. *4:109–126*.

91. Nagayama, S. K. (2002). Genome-wide analysis of gene expression in synovial sarcomas using a cdna microarray. *62(20):5859-5866*.

92. Newton, M. e. (2001). On differential variability of expression ratios:improving statistical inference about gene expression changes from microarray data. . *8 : 37–52*.

93. Nguyen, D. a. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. *18(9):1216-1226*.

94. Perez-Enciso, M. a. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (plsda) approach.

95. Perou, C. S.D.L. (2000). Molecular portrailts of human breast tumors. *406:747-752*.

96. Perrou, C. J. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *96(16):9212-9217*.

97. Radmacher, M. M. (2002). A paradigm for class prediction using gene expression profiles. *9(3):505–511*.

98. Rainer Breitlinga, b. P. ( 2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *573* ( 1-3, Pages 83-92 ).

99. *References*. (n.d.). (Texas A&M University, College Station) Retrieved 10 3, 2008, from Genomic Signal Processing Lab: http://gsp.tamu.edu/web2/cv_paper/pdf/refs.pdf

100. Richard O. Duda, P. E. (2001). *Pattern Classification, Second Edition.* John Wiley & Sons, Inc.

101. S.Kim, E. Identification of combination gene sets for glioma classification. *1* (13,pp- 1229).

102. Sarkar, I. P. (2002). Characteristic attributes in cancer microarrays. . *35(2):111–122.*

103. Sasaki, H. I. (2002). Gene expression analysis of human thymoma correlates with tumor stage. *101(4):342–347.*

104. Schena M, S. D. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA microarray. *New Series, Vol. 270* (No. 5235. (Oct. 20, 1995), pp. 467-470.).

105. Schoch, C. K. (2002). Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. .

106. Shipp, M. R. (2002). Diffuse large b-cell Lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *8(1):68-74*.

107. Slonim, D. T. (2000). Class prediction and discovery using gene expression data.

108. Smyth, G. Y. (2002). Statistical issues in cDNA microarray data analysis. *111-136*.

109. Sorlie, T. P.-D. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *98(19):10869-10874*.

110. Squire, P. F. (2002). Application of Microarrays to the Analysis of Gene Expression in Cancer. *48:8 pp. 1170-1177*.

111. Stoughton, R. B. (2005). Applications of DNA microarrays in Biology. *74:53–82*.

112. Subramani Prabakaran, S. R. (2006). Feature selection using Haar wavelet power spectrum. *7:432*.

113. Sugita, M. G. (2002). Combined use of oligonucleotide and tissue microarrays identifiescancer/testis antigens as biomarkers in lung carcinoma. *62(14):3971-3979.*

114. Szabo, A. B. (2002). Variable selection and pattern recognition with gene expression data generated by the microarray technology. *176(1):71–98.*

115. T.Kobayashi, M. a. (2003). Microarray reveals differences in both tumors and vascular specific gene expression in de novo cd5+ and cd5- diffuse large b-cell lymphomas. *63* (1,pp.60-6).

116. T.R.Golub, D. (1999). Molecular Classification of cancer: class discovery and class prediction by gene expression monitoring. *286* (5439,pp.531-537).

117. T.SFurey, N. a. (2000). Support Vector Machine classification and validation of cancer tissue samples using microarray expression data. *16* (10,pp.906-14).

118. Tamayo, P. S. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *96:2907.*

119. Tan, K. Y. (2003). Evolutionary computing for knowledge discovery in medical diagnosis. *27(2):129–154*.

120. Theilhaber, J. F. Bayesian estimation of fold-changes in the analysis of gene expression:the PFOLD algorithm. *8,585-614*.

121. Thomas, J. e. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *11:1227–1236.*

122. Troyanskaya, O. e. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. . *18: 1454–1461.*

123. Tusher, V. T. (2001). . Significance analysis of microarrays applied to the ionizing radiation response. . *98(9).*

124. van de Vijver, M. H. (2002). A gene expression signature as a predictor of survival in breast cancer. *347(25):1999-2009*.

125. van't Veer, L. D. (2002). Gene expression profiling predicts clinical outcome of breast cancer . *415:530-536.*

126. Virginia Goss Tusher, R. T. ( 2001). Significance analysis of microarrays applied to the ionizing radiation response. *98:5116-5121.*

127. Wang, H. K. (2003). (2003). Analysis of gene expression profile induced by emp-1 in esophageal cancer cells using cdna microarray. *,9(3):392–398*.

128. Winzeler, D. J. (2000). Genomics, Gene Expression and DNA Arrays. *405* (15 June).

129. Yang, Y. a. (2002). Comparison of methods for image analysis on cDNA microarray data. *11, 108-136*.

130. Yeang, C. R. (2001). Molecularclassification of multiple tumor types. . *17:316–322.*

131. Yeung, K. a. (2003). Multiclass classification of microarray data with repeated measurements: application to cancer. . *4, R83.*

132. Yvan Saeys, I. I. (2007). A review of feature selection techniques in bioinformatics. *23* (19,pp:2507-2517).

133. Zhang, H. a. (2002). Tree-based analysis of microarray data for classifying breast cancer. *, 7:63–67.*

APPENDICES

APPENDIX A

STATISTICAL ANALYSIS FOR NULL HYPOTHESES

Null Hypotheses (H0):

1. The probability of the gene-sets selected by the wavelet method to classify the test sample accurately (p1) = the probability of the the gene-sets selected by the wavelet method to classify the test sample inaccurately (p2).

$$p1 = p2 = 0.5$$

Alternate Hypotheses (H1):

1. The probability of the gene-sets selected by the wavelet method to classify the test sample accurately (p1) > the probability of the the gene-sets selected by the wavelet method to classify the test sample inaccurately (p2).

$$\alpha = 0.01$$

The Hypotheses was tested using Pearson's Chi-Square testing.

The number of test samples: n = 10.

The number of Repetitions: R = 250.

The expected values are estimated for p1 = p2 = 0.5 using the equation:

$$\text{Expected Value}(r) = \frac{n!}{r! \times (n-r)!} \times p1^{r} \times p2^{n-r} \times R$$

where r – number of accurate classifications.

The $\chi^2$ term is estimated from the Expected and the Observed Values by:

$$\chi 2 = \sum_{r=0}^{n} \frac{(\text{ExpectedValue}(r) - \text{ObservedValue}(r))^2}{\text{ExpectedValue}(r)}$$

The critical value for α = 0.01 and degree of freedom '10', the critical value is 23.21.

Table A.1.The estimation for B-Cell Lymphoma Data.

| B-CELL LYMPHOMA DATA (NUMBER OF GENES = 50) | | | | |
|---|---|---|---|---|
| Accurate Classifications | Expected Value (E) | Observed Value(O) | (E - O)^2 | (E - O)^2./E |
| 10 | 0.244 | 124 | 15315.513 | 62732.340 |
| 9 | 2.441 | 95 | 8567.093 | 3509.081 |
| 8 | 10.986 | 26 | 225.410 | 20.517 |
| 7 | 29.297 | 5 | 590.338 | 20.150 |
| 6 | 51.270 | 0 | 2628.565 | 51.270 |
| 5 | 61.523 | 0 | 3785.133 | 61.523 |
| 4 | 51.270 | 0 | 2628.565 | 51.270 |
| 3 | 29.297 | 0 | 858.307 | 29.297 |
| 2 | 10.986 | 0 | 120.699 | 10.986 |
| 1 | 2.441 | 0 | 5.960 | 2.441 |
| 0 | 0.244 | 0 | 0.060 | 0.244 |
| | | Sum of Chisquare term | | 66489.120 |

Table A.2. The estimation for Leukemia Data.

| LEUKEMIA DATA (NUMBER OF GENES = 50) | | | | |
|---|---|---|---|---|
| Accurate Classifications | Expected Value (E) | Observed Value(O) | (E - O)^2 | (E - O)^2./E |
| 10 | 0.244 | 137 | 18702.165 | 76604.068 |
| 9 | 2.441 | 92 | 8020.742 | 3285.296 |
| 8 | 10.986 | 21 | 100.274 | 9.127 |
| 7 | 29.297 | 0 | 858.307 | 29.297 |
| 6 | 51.270 | 0 | 2628.565 | 51.270 |
| 5 | 61.523 | 0 | 3785.133 | 61.523 |
| 4 | 51.270 | 0 | 2628.565 | 51.270 |
| 3 | 29.297 | 0 | 858.307 | 29.297 |
| 2 | 10.986 | 0 | 120.699 | 10.986 |
| 1 | 2.441 | 0 | 5.960 | 2.441 |
| 0 | 0.244 | 0 | 0.060 | 0.244 |
| | | Sum of Chisquare term | | 80134.819 |

Table A.3. The estimation for The Colon Cancer Data.

| COLON CANCER DATA(NUMBER OF GENES = 75) | | | | |
|---|---|---|---|---|
| Accurate Classifications | Expected Value (E) | Observed Value(O) | (E - O)^2 | (E - O)^2./E |
| 10 | 0.244 | 41 | 1661.040 | 6803.620 |
| 9 | 2.441 | 89 | 7492.390 | 3068.883 |
| 8 | 10.986 | 76 | 4226.778 | 384.731 |
| 7 | 29.297 | 34 | 22.119 | 0.755 |
| 6 | 51.270 | 9 | 1786.713 | 34.849 |
| 5 | 61.523 | 1 | 3663.086 | 59.540 |
| 4 | 51.270 | 0 | 2628.565 | 51.270 |
| 3 | 29.297 | 0 | 858.307 | 29.297 |
| 2 | 10.986 | 0 | 120.699 | 10.986 |
| 1 | 2.441 | 0 | 5.960 | 2.441 |
| 0 | 0.244 | 0 | 0.060 | 0.244 |
| | | Sum of Chisquare term | | 10446.616 |

Results

1. The estimated $\chi^2$ parameter value is very high compared to the critical value **23.21** in all the three datasets studied. Therefore, the NULL HYPOTHESES IS REJECTED and THE ALTERNATE HYPOTHESES IS ACCEPTED.

2. The Probability of the gene sets selected by the wavelet based method making an accurate classification (p1) > the probability of the gene sets selected by the wavelet based method making an inaccurate classification (p1).

APPENDIX B

STATISTICAL ANALYSIS OF THE CLASSIFICATION RESULTS

The three methods studied in the thesis are compared using Analysis of Variance and Tukey's Honestly Significant Difference (HSD) post hoc test.

Here, the classification performances (Accuracy, Sensitivity and the Specificity) of each of the three methods studied, the wavelet based method (D8), the Sum of Squares method (BW) and the T-test (T) form the three samples. There are 250 values available for each gene sizes (from 5 to 150 in steps of 5). The following steps are followed for finding significant difference between there performances at different gene sizes.

(1) The ANOVA is carried out between the three samples.

(2) The critical value, Q, is obtained from the studentized range statistic table for $\alpha = 0.05$, sample size = 3 and degrees of freedom within as $\infty$.

(3) The critical difference is estimated using the equation:

$$\text{Critical Difference} = Q \times \sqrt{\frac{MS_{within}}{n}} \tag{B.1}$$

Where the Q is obtained from step 2, $MS_{within}$ is the mean square within term for the two samples considered (obtained from the ANOVA results in step 1) and n is the number of values in each sample (250 for all cases).

(4) The difference between the two mean of the two samples are estimated. If we are comparing two samples A and B, the result of the Hypotheses is obtained by:

$$Hypotheses\ Result = \begin{cases} 1 & if\ abs(M_A - M_B) > Critical\ Difference\ and M_A > M_B \\ -1 & if\ abs(M_A - M_B) > Critical\ Difference\ and M_A < M_B \\ 0 & if\ M_A - M_B < Critical\ Difference \end{cases}$$

(B.2)

The results obtained are summarized in the tables below.

Table B.1. Summary of results of Tukey HSD post hoc test for the classification accuracy of B-cell Lymphoma Data for α=0.05. N – Number of Genes. C – Critical Difference, MeanSS – Mean accuracy of Sum of Square method (SS). MeanD8 – Mean Accuracy of wavelet (D8). MeanT – Mean accuracy of T-test(T).

| N | B-Cell Lymphoma Data - Accuracy (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0215 | 0.8420 | 0.8920 | 0.8360 | -0.0490 | -1 | 0.0060 | 0 | 0.0550 | 1 |
| 10 | 0.0211 | 0.8590 | 0.8676 | 0.8516 | -0.0080 | 0 | 0.0080 | 0 | 0.0160 | 0 |
| 15 | 0.0208 | 0.8500 | 0.8800 | 0.8556 | -0.0240 | -1 | 0.0004 | 0 | 0.0244 | 1 |
| 20 | 0.0214 | 0.8610 | 0.8904 | 0.8444 | -0.0288 | -1 | 0.0172 | 0 | 0.0460 | 1 |
| 25 | 0.0212 | 0.8590 | 0.9108 | 0.8364 | -0.0512 | -1 | 0.0232 | 1 | 0.0744 | 1 |
| 30 | 0.0205 | 0.8650 | 0.9240 | 0.8404 | -0.0588 | -1 | 0.0248 | 1 | 0.0836 | 1 |
| 35 | 0.0200 | 0.8690 | 0.9340 | 0.8440 | -0.0644 | -1 | 0.0256 | 1 | 0.0900 | 1 |
| 40 | 0.0196 | 0.8760 | 0.9328 | 0.8464 | -0.0564 | -1 | 0.0300 | 1 | 0.0864 | 1 |
| 45 | 0.0197 | 0.8750 | 0.9340 | 0.8468 | -0.0584 | -1 | 0.0288 | 1 | 0.0872 | 1 |
| 50 | 0.0199 | 0.8730 | 0.9352 | 0.8492 | -0.0616 | -1 | 0.0244 | 1 | 0.0860 | 1 |
| 55 | 0.0201 | 0.8770 | 0.9364 | 0.8496 | -0.0588 | -1 | 0.0280 | 1 | 0.0868 | 1 |
| 60 | 0.0202 | 0.8840 | 0.9356 | 0.8536 | -0.0508 | -1 | 0.0312 | 1 | 0.0820 | 1 |
| 65 | 0.0203 | 0.8890 | 0.9344 | 0.8500 | -0.0448 | -1 | 0.0396 | 1 | 0.0844 | 1 |
| 70 | 0.0204 | 0.8890 | 0.9352 | 0.8492 | -0.0460 | -1 | 0.0400 | 1 | 0.0860 | 1 |
| 75 | 0.0207 | 0.8932 | 0.9324 | 0.8488 | -0.0392 | -1 | 0.0444 | 1 | 0.0836 | 1 |
| 80 | 0.0210 | 0.8948 | 0.9308 | 0.8464 | -0.0360 | -1 | 0.0484 | 1 | 0.0844 | 1 |
| 85 | 0.0210 | 0.8956 | 0.9296 | 0.8496 | -0.0340 | -1 | 0.0460 | 1 | 0.0800 | 1 |
| 90 | 0.0208 | 0.8980 | 0.9296 | 0.8524 | -0.0316 | -1 | 0.0456 | 1 | 0.0772 | 1 |
| 95 | 0.0209 | 0.9004 | 0.9272 | 0.8520 | -0.0268 | -1 | 0.0484 | 1 | 0.0752 | 1 |

Table B.1. Summary of results of Tukey HSD post hoc test for the classification accuracy of B-cell Lymphoma Data for α=0.05. (Continued).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 0.0212 | 0.9012 | 0.9248 | 0.8512 | -0.0236 | -1 | 0.0500 | 1 | 0.0736 | 1 |
| 105 | 0.0206 | 0.9060 | 0.9224 | 0.8516 | -0.0164 | 0 | 0.0544 | 1 | 0.0708 | 1 |
| 110 | 0.0209 | 0.9056 | 0.9212 | 0.8528 | -0.0156 | 0 | 0.0528 | 1 | 0.0684 | 1 |
| 115 | 0.0207 | 0.9060 | 0.9224 | 0.8516 | -0.0164 | 0 | 0.0544 | 1 | 0.0708 | 1 |
| 120 | 0.0208 | 0.9072 | 0.9208 | 0.8548 | -0.0136 | 0 | 0.0524 | 1 | 0.0660 | 1 |
| 125 | 0.0211 | 0.9080 | 0.9212 | 0.8540 | -0.0132 | 0 | 0.0540 | 1 | 0.0672 | 1 |
| 130 | 0.0207 | 0.9084 | 0.9192 | 0.8556 | -0.0108 | 0 | 0.0528 | 1 | 0.0636 | 1 |
| 135 | 0.0204 | 0.9100 | 0.9204 | 0.8540 | -0.0104 | 0 | 0.0560 | 1 | 0.0664 | 1 |
| 140 | 0.0207 | 0.9108 | 0.9200 | 0.8576 | -0.0092 | 0 | 0.0532 | 1 | 0.0624 | 1 |
| 145 | 0.0201 | 0.9124 | 0.9204 | 0.8580 | -0.0080 | 0 | 0.0544 | 1 | 0.0624 | 1 |
| 150 | 0.0199 | 0.9124 | 0.9184 | 0.8596 | -0.0060 | 0 | 0.0528 | 1 | 0.0588 | 1 |

Table B.2. Summary of results of Tukey HSD post hoc test for the classification accuracy of Leukemia Data for α=0.05. N – Number of Genes. C – Critical Difference, MeanSS – Mean accuracy of Sum of Square method (SS). MeanD8 – Mean Accuracy of wavelet (D8). MeanT – Mean accuracy of T-test(T).

| N | Leukemia Data - Accuracy (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0148 | 0.9092 | 0.9524 | 0.9292 | -0.0432 | -1 | -0.0200 | -1 | 0.0232 | 1 |
| 10 | 0.0147 | 0.9340 | 0.9380 | 0.9520 | -0.0040 | 0 | -0.0180 | -1 | -0.0140 | 0 |
| 15 | 0.0151 | 0.9408 | 0.9424 | 0.9368 | -0.0016 | 0 | 0.0040 | 0 | 0.0056 | 0 |
| 20 | 0.0147 | 0.9420 | 0.9360 | 0.9336 | 0.0060 | 0 | 0.0084 | 0 | 0.0024 | 0 |
| 25 | 0.0153 | 0.9372 | 0.9372 | 0.9308 | 0.0000 | 0 | 0.0064 | 0 | 0.0064 | 0 |
| 30 | 0.0156 | 0.9360 | 0.9332 | 0.9344 | 0.0028 | 0 | 0.0016 | 0 | -0.0012 | 0 |
| 35 | 0.0152 | 0.9348 | 0.9392 | 0.9320 | -0.0044 | 0 | 0.0028 | 0 | 0.0072 | 0 |
| 40 | 0.0156 | 0.9328 | 0.9400 | 0.9296 | -0.0072 | 0 | 0.0032 | 0 | 0.0104 | 0 |
| 45 | 0.0156 | 0.9340 | 0.9460 | 0.9264 | -0.0120 | 0 | 0.0076 | 0 | 0.0196 | 1 |
| 50 | 0.0149 | 0.9392 | 0.9464 | 0.9320 | -0.0072 | 0 | 0.0072 | 0 | 0.0144 | 0 |
| 55 | 0.0147 | 0.9396 | 0.9480 | 0.9380 | -0.0084 | 0 | 0.0016 | 0 | 0.0100 | 0 |
| 60 | 0.0145 | 0.9376 | 0.9476 | 0.9368 | -0.0100 | 0 | 0.0008 | 0 | 0.0108 | 0 |
| 65 | 0.0145 | 0.9380 | 0.9428 | 0.9412 | -0.0048 | 0 | -0.0032 | 0 | 0.0016 | 0 |
| 70 | 0.0142 | 0.9384 | 0.9444 | 0.9428 | -0.0060 | 0 | -0.0044 | 0 | 0.0016 | 0 |
| 75 | 0.0140 | 0.9384 | 0.9464 | 0.9460 | -0.0080 | 0 | -0.0076 | 0 | 0.0004 | 0 |
| 80 | 0.0135 | 0.9392 | 0.9500 | 0.9484 | -0.0108 | 0 | -0.0092 | 0 | 0.0016 | 0 |

Table B.2. Summary of results of Tukey HSD post hoc test for the classification accuracy of Leukemia Data for α=0.05. (Continued).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 85 | 0.0136 | 0.9392 | 0.9484 | 0.9448 | -0.0092 | 0 | -0.0056 | 0 | 0.0036 | 0 |
| 90 | 0.0136 | 0.9396 | 0.9532 | 0.9444 | -0.0136 | -1 | -0.0048 | 0 | 0.0088 | 0 |
| 95 | 0.0135 | 0.9436 | 0.9520 | 0.9472 | -0.0084 | 0 | -0.0036 | 0 | 0.0048 | 0 |
| 100 | 0.0134 | 0.9440 | 0.9532 | 0.9492 | -0.0092 | 0 | -0.0052 | 0 | 0.0040 | 0 |
| 105 | 0.0132 | 0.9452 | 0.9544 | 0.9484 | -0.0092 | 0 | -0.0032 | 0 | 0.0060 | 0 |
| 110 | 0.0131 | 0.9468 | 0.9548 | 0.9512 | -0.0080 | 0 | -0.0044 | 0 | 0.0036 | 0 |
| 115 | 0.0129 | 0.9488 | 0.9544 | 0.9504 | -0.0056 | 0 | -0.0016 | 0 | 0.0040 | 0 |
| 120 | 0.0130 | 0.9508 | 0.9528 | 0.9516 | -0.0020 | 0 | -0.0008 | 0 | 0.0012 | 0 |
| 125 | 0.0131 | 0.9504 | 0.9520 | 0.9500 | -0.0016 | 0 | 0.0004 | 0 | 0.0020 | 0 |
| 130 | 0.0129 | 0.9496 | 0.9540 | 0.9516 | -0.0044 | 0 | -0.0020 | 0 | 0.0024 | 0 |
| 135 | 0.0128 | 0.9492 | 0.9548 | 0.9536 | -0.0056 | 0 | -0.0044 | 0 | 0.0012 | 0 |
| 140 | 0.0129 | 0.9484 | 0.9540 | 0.9544 | -0.0056 | 0 | -0.0060 | 0 | -0.0004 | 0 |
| 145 | 0.0130 | 0.9472 | 0.9548 | 0.9548 | -0.0076 | 0 | -0.0076 | 0 | 0.0000 | 0 |
| 150 | 0.0130 | 0.9472 | 0.9536 | 0.9536 | -0.0064 | 0 | -0.0064 | 0 | 0.0000 | 0 |

Table B.3. Summary of results of Tukey HSD post hoc test for the classification accuracy of Colon Cancer Data for α=0.05. N – Number of Genes. C – Critical Difference, MeanSS – Mean accuracy of Sum of Square method (SS). MeanD8 – Mean Accuracy of wavelet (D8). MeanT – Mean accuracy of T-test(T).

| | Colon Cancer Data - Accuracy (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| N | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0268 | 0.8408 | 0.7968 | 0.7100 | 0.0440 | 1 | 0.1308 | 1 | 0.0868 | 1 |
| 10 | 0.0264 | 0.8404 | 0.7800 | 0.7464 | 0.0604 | 1 | 0.0940 | 1 | 0.0336 | 1 |
| 15 | 0.0271 | 0.8344 | 0.7684 | 0.7608 | 0.0660 | 1 | 0.0736 | 1 | 0.0076 | 0 |
| 20 | 0.0262 | 0.8400 | 0.7812 | 0.7784 | 0.0588 | 1 | 0.0616 | 1 | 0.0028 | 0 |
| 25 | 0.0249 | 0.8364 | 0.7996 | 0.7940 | 0.0368 | 1 | 0.0424 | 1 | 0.0056 | 0 |
| 30 | 0.0241 | 0.8408 | 0.8080 | 0.8128 | 0.0328 | 1 | 0.0280 | 1 | -0.0048 | 0 |
| 35 | 0.0231 | 0.8448 | 0.8160 | 0.8232 | 0.0288 | 1 | 0.0216 | 0 | -0.0072 | 0 |
| 40 | 0.0232 | 0.8424 | 0.8200 | 0.8236 | 0.0224 | 0 | 0.0188 | 0 | -0.0036 | 0 |
| 45 | 0.0231 | 0.8408 | 0.8288 | 0.8268 | 0.0120 | 0 | 0.0140 | 0 | 0.0020 | 0 |
| 50 | 0.0231 | 0.8392 | 0.8308 | 0.8312 | 0.0084 | 0 | 0.0080 | 0 | -0.0004 | 0 |
| 55 | 0.0225 | 0.8380 | 0.8356 | 0.8356 | 0.0024 | 0 | 0.0024 | 0 | 0.0000 | 0 |
| 60 | 0.0219 | 0.8380 | 0.8384 | 0.8416 | -0.0004 | 0 | -0.0036 | 0 | -0.0032 | 0 |
| 65 | 0.0213 | 0.8368 | 0.8452 | 0.8444 | -0.0084 | 0 | -0.0076 | 0 | 0.0008 | 0 |

Table B.3. Summary of results of Tukey HSD post hoc test for the classification accuracy of Colon Cancer Data for α=0.05. (Continued).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 70 | 0.0215 | 0.8384 | 0.8448 | 0.8464 | -0.0064 | 0 | -0.0080 | 0 | -0.0016 | 0 |
| 75 | 0.0219 | 0.8388 | 0.8464 | 0.8468 | -0.0076 | 0 | -0.0080 | 0 | -0.0004 | 0 |
| 80 | 0.0217 | 0.8396 | 0.8476 | 0.8436 | -0.0080 | 0 | -0.0040 | 0 | 0.0040 | 0 |
| 85 | 0.0215 | 0.8404 | 0.8528 | 0.8420 | -0.0124 | 0 | -0.0016 | 0 | 0.0108 | 0 |
| 90 | 0.0213 | 0.8392 | 0.8516 | 0.8404 | -0.0124 | 0 | -0.0012 | 0 | 0.0112 | 0 |
| 95 | 0.0213 | 0.8396 | 0.8508 | 0.8384 | -0.0112 | 0 | 0.0012 | 0 | 0.0124 | 0 |
| 100 | 0.0215 | 0.8396 | 0.8512 | 0.8404 | -0.0116 | 0 | -0.0008 | 0 | 0.0108 | 0 |
| 105 | 0.0215 | 0.8396 | 0.8484 | 0.8384 | -0.0088 | 0 | 0.0012 | 0 | 0.0100 | 0 |
| 110 | 0.0213 | 0.8404 | 0.8508 | 0.8380 | -0.0104 | 0 | 0.0024 | 0 | 0.0128 | 0 |
| 115 | 0.0214 | 0.8416 | 0.8500 | 0.8392 | -0.0084 | 0 | 0.0024 | 0 | 0.0108 | 0 |
| 120 | 0.0215 | 0.8424 | 0.8488 | 0.8404 | -0.0064 | 0 | 0.0020 | 0 | 0.0084 | 0 |
| 125 | 0.0211 | 0.8436 | 0.8504 | 0.8400 | -0.0068 | 0 | 0.0036 | 0 | 0.0104 | 0 |
| 130 | 0.0212 | 0.8452 | 0.8520 | 0.8404 | -0.0068 | 0 | 0.0048 | 0 | 0.0116 | 0 |
| 135 | 0.0212 | 0.8448 | 0.8524 | 0.8408 | -0.0076 | 0 | 0.0040 | 0 | 0.0116 | 0 |
| 140 | 0.0213 | 0.8444 | 0.8540 | 0.8400 | -0.0096 | 0 | 0.0044 | 0 | 0.0140 | 0 |
| 145 | 0.0214 | 0.8448 | 0.8528 | 0.8404 | -0.0080 | 0 | 0.0044 | 0 | 0.0124 | 0 |
| 150 | 0.0212 | 0.8456 | 0.8528 | 0.8428 | -0.0072 | 0 | 0.0028 | 0 | 0.0100 | 0 |

Table B.4. Summary of results of Tukey HSD post hoc test for the classification sensitivity of B-cell Lymphoma data for α=0.05. Sensitivity is the fraction of samples identified as positive which are actually positive. For B-cell Lymphoma Samples are Positive and Follicular Cancer samples are Negative.N – Number of Genes. C – Critical Difference, MeanSS – Mean sensitivity of Sum of Square method (SS). MeanD8 – Mean Sensitivity of wavelet (D8). MeanT – Mean sensitivity of T-test(T).

| N | B-Cell Lymphoma Data – Sensitivity (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0248 | 0.8715 | 0.8957 | 0.8407 | -0.0242 | 0 | 0.0308 | 1 | 0.0550 | 1 |
| 10 | 0.0253 | 0.8817 | 0.8665 | 0.8393 | 0.0152 | 0 | 0.0425 | 1 | 0.0272 | 1 |
| 15 | 0.0249 | 0.8752 | 0.8794 | 0.8414 | -0.0042 | 0 | 0.0337 | 1 | 0.0379 | 1 |
| 20 | 0.0255 | 0.8740 | 0.8898 | 0.8306 | -0.0158 | 0 | 0.0435 | 1 | 0.0593 | 1 |
| 25 | 0.0251 | 0.8716 | 0.9075 | 0.8259 | -0.0359 | -1 | 0.0457 | 1 | 0.0816 | 1 |
| 30 | 0.0247 | 0.8792 | 0.9175 | 0.8307 | -0.0383 | -1 | 0.0485 | 1 | 0.0868 | 1 |
| 35 | 0.0241 | 0.8867 | 0.9284 | 0.8340 | -0.0417 | -1 | 0.0528 | 1 | 0.0944 | 1 |
| 40 | 0.0237 | 0.8932 | 0.9301 | 0.8377 | -0.0370 | -1 | 0.0554 | 1 | 0.0924 | 1 |

Table B.4. Summary of results of Tukey HSD post hoc test for the classification sensitivity of B-cell Lymphoma Data for α=0.05. (Continued).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 45 | 0.0239 | 0.8908 | 0.9317 | 0.8383 | -0.0409 | -1 | 0.0525 | 1 | 0.0934 | 1 |
| 50 | 0.0238 | 0.8891 | 0.9350 | 0.8381 | -0.0459 | -1 | 0.0509 | 1 | 0.0969 | 1 |
| 55 | 0.0242 | 0.8890 | 0.9352 | 0.8377 | -0.0461 | -1 | 0.0513 | 1 | 0.0974 | 1 |
| 60 | 0.0244 | 0.8948 | 0.9335 | 0.8412 | -0.0387 | -1 | 0.0536 | 1 | 0.0923 | 1 |
| 65 | 0.0244 | 0.8996 | 0.9340 | 0.8360 | -0.0344 | -1 | 0.0636 | 1 | 0.0980 | 1 |
| 70 | 0.0249 | 0.8987 | 0.9324 | 0.8345 | -0.0338 | -1 | 0.0641 | 1 | 0.0979 | 1 |
| 75 | 0.0250 | 0.9038 | 0.9300 | 0.8321 | -0.0262 | -1 | 0.0717 | 1 | 0.0979 | 1 |
| 80 | 0.0251 | 0.9069 | 0.9285 | 0.8284 | -0.0216 | 0 | 0.0785 | 1 | 0.1001 | 1 |
| 85 | 0.0252 | 0.9080 | 0.9264 | 0.8311 | -0.0184 | 0 | 0.0769 | 1 | 0.0953 | 1 |
| 90 | 0.0248 | 0.9114 | 0.9264 | 0.8344 | -0.0150 | 0 | 0.0770 | 1 | 0.0919 | 1 |
| 95 | 0.0249 | 0.9138 | 0.9223 | 0.8340 | -0.0085 | 0 | 0.0798 | 1 | 0.0884 | 1 |
| 100 | 0.0248 | 0.9137 | 0.9222 | 0.8341 | -0.0085 | 0 | 0.0796 | 1 | 0.0881 | 1 |
| 105 | 0.0242 | 0.9188 | 0.9196 | 0.8337 | -0.0008 | 0 | 0.0851 | 1 | 0.0859 | 1 |
| 110 | 0.0244 | 0.9193 | 0.9179 | 0.8351 | 0.0015 | 0 | 0.0842 | 1 | 0.0828 | 1 |
| 115 | 0.0244 | 0.9177 | 0.9174 | 0.8322 | 0.0003 | 0 | 0.0855 | 1 | 0.0852 | 1 |
| 120 | 0.0241 | 0.9211 | 0.9177 | 0.8332 | 0.0034 | 0 | 0.0879 | 1 | 0.0845 | 1 |
| 125 | 0.0245 | 0.9224 | 0.9163 | 0.8324 | 0.0060 | 0 | 0.0900 | 1 | 0.0839 | 1 |
| 130 | 0.0245 | 0.9219 | 0.9142 | 0.8342 | 0.0076 | 0 | 0.0876 | 1 | 0.0800 | 1 |
| 135 | 0.0244 | 0.9220 | 0.9151 | 0.8323 | 0.0070 | 0 | 0.0897 | 1 | 0.0827 | 1 |
| 140 | 0.0245 | 0.9215 | 0.9125 | 0.8345 | 0.0090 | 0 | 0.0869 | 1 | 0.0780 | 1 |
| 145 | 0.0241 | 0.9219 | 0.9125 | 0.8348 | 0.0094 | 0 | 0.0871 | 1 | 0.0777 | 1 |
| 150 | 0.0239 | 0.9212 | 0.9092 | 0.8361 | 0.0120 | 0 | 0.0851 | 1 | 0.0731 | 1 |

Table B.5. Summary of results of Tukey HSD post hoc test for the classification sensitivity of Leukemia data for α=0.05. Sensitivity is the fraction of samples identified as positive which are actually positive. For Acute Lymphoblastic Leukemia samples(ALL) are Positive and Acute Myeloid Leukemia(AML) samples are Negative.N – Number of Genes. C – Critical Difference, MeanSS – Mean sensitivity of Sum of Square method (SS). MeanD8 – Mean Sensitivity of wavelet (D8). MeanT – Mean sensitivity of T-test(T).

| | Leukemia Data - Sensitivity (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| N | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0154 | 0.9687 | 0.9988 | 0.9348 | -0.0300 | -1 | 0.0339 | 1 | 0.0639 | 1 |
| 10 | 0.0124 | 0.9910 | 0.9974 | 0.9489 | -0.0063 | 0 | 0.0421 | 1 | 0.0485 | 1 |
| 15 | 0.0141 | 0.9973 | 0.9984 | 0.9330 | -0.0011 | 0 | 0.0643 | 1 | 0.0654 | 1 |
| 20 | 0.0144 | 0.9979 | 0.9988 | 0.9273 | -0.0009 | 0 | 0.0706 | 1 | 0.0714 | 1 |
| 25 | 0.0138 | 0.9993 | 0.9965 | 0.9311 | 0.0028 | 0 | 0.0682 | 1 | 0.0654 | 1 |

Table B.5. Summary of results of Tukey HSD post hoc test for the classification sensitivity of Leukemia Data for α=0.05. (Continued).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 0.0140 | 0.9992 | 0.9944 | 0.9388 | 0.0048 | 0 | 0.0604 | 1 | 0.0557 | 1 |
| 35 | 0.0139 | 0.9992 | 0.9938 | 0.9357 | 0.0054 | 0 | 0.0635 | 1 | 0.0581 | 1 |
| 40 | 0.0145 | 0.9991 | 0.9943 | 0.9357 | 0.0047 | 0 | 0.0634 | 1 | 0.0587 | 1 |
| 45 | 0.0142 | 0.9987 | 0.9973 | 0.9361 | 0.0013 | 0 | 0.0625 | 1 | 0.0612 | 1 |
| 50 | 0.0136 | 0.9992 | 0.9968 | 0.9447 | 0.0024 | 0 | 0.0545 | 1 | 0.0521 | 1 |
| 55 | 0.0134 | 0.9978 | 0.9968 | 0.9520 | 0.0010 | 0 | 0.0459 | 1 | 0.0449 | 1 |
| 60 | 0.0132 | 0.9978 | 0.9927 | 0.9522 | 0.0052 | 0 | 0.0456 | 1 | 0.0404 | 1 |
| 65 | 0.0128 | 0.9978 | 0.9911 | 0.9580 | 0.0067 | 0 | 0.0398 | 1 | 0.0331 | 1 |
| 70 | 0.0122 | 0.9978 | 0.9927 | 0.9623 | 0.0051 | 0 | 0.0355 | 1 | 0.0303 | 1 |
| 75 | 0.0116 | 0.9974 | 0.9931 | 0.9639 | 0.0043 | 0 | 0.0334 | 1 | 0.0291 | 1 |
| 80 | 0.0114 | 0.9983 | 0.9921 | 0.9681 | 0.0062 | 0 | 0.0303 | 1 | 0.0240 | 1 |
| 85 | 0.0119 | 0.9967 | 0.9913 | 0.9649 | 0.0054 | 0 | 0.0318 | 1 | 0.0264 | 1 |
| 90 | 0.0109 | 0.9974 | 0.9926 | 0.9657 | 0.0047 | 0 | 0.0317 | 1 | 0.0269 | 1 |
| 95 | 0.0107 | 0.9967 | 0.9904 | 0.9684 | 0.0063 | 0 | 0.0283 | 1 | 0.0221 | 1 |
| 100 | 0.0105 | 0.9967 | 0.9922 | 0.9711 | 0.0045 | 0 | 0.0256 | 1 | 0.0211 | 1 |
| 105 | 0.0102 | 0.9966 | 0.9934 | 0.9710 | 0.0031 | 0 | 0.0255 | 1 | 0.0224 | 1 |
| 110 | 0.0098 | 0.9970 | 0.9934 | 0.9764 | 0.0037 | 0 | 0.0206 | 1 | 0.0170 | 1 |
| 115 | 0.0099 | 0.9975 | 0.9939 | 0.9740 | 0.0036 | 0 | 0.0235 | 1 | 0.0199 | 1 |
| 120 | 0.0101 | 0.9983 | 0.9926 | 0.9746 | 0.0057 | 0 | 0.0237 | 1 | 0.0180 | 1 |
| 125 | 0.0096 | 0.9983 | 0.9938 | 0.9744 | 0.0045 | 0 | 0.0240 | 1 | 0.0195 | 1 |
| 130 | 0.0093 | 0.9983 | 0.9944 | 0.9765 | 0.0039 | 0 | 0.0219 | 1 | 0.0180 | 1 |
| 135 | 0.0088 | 0.9983 | 0.9954 | 0.9781 | 0.0030 | 0 | 0.0202 | 1 | 0.0172 | 1 |
| 140 | 0.0087 | 0.9978 | 0.9950 | 0.9787 | 0.0028 | 0 | 0.0191 | 1 | 0.0163 | 1 |
| 145 | 0.0085 | 0.9978 | 0.9954 | 0.9800 | 0.0024 | 0 | 0.0178 | 1 | 0.0154 | 1 |
| 150 | 0.0087 | 0.9978 | 0.9948 | 0.9799 | 0.0030 | 0 | 0.0179 | 1 | 0.0149 | 1 |

Table B.6. Summary of results of Tukey HSD post hoc test for the classification sensitivity of Colon cancer data for α=0.05. Sensitivity is the fraction of samples identified as positive which are actually positive. For Tumor samples are Positive and Normal samples are Negative.N – Number of Genes. C – Critical Difference, MeanSS – Mean sensitivity of Sum of Square method (SS). MeanD8 – Mean Sensitivity of wavelet (D8). MeanT – Mean sensitivity of T-test(T).

| N | Colon Cancer Data – Sensitivity (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0600 | 0.6910 | 0.6501 | 0.5850 | 0.0408 | 0 | 0.1059 | 1 | 0.0651 | 1 |
| 10 | 0.0583 | 0.7125 | 0.5932 | 0.6097 | 0.1194 | 1 | 0.1028 | 1 | -0.0166 | 0 |
| 15 | 0.0576 | 0.7129 | 0.5722 | 0.6516 | 0.1407 | 1 | 0.0613 | 1 | -0.0794 | -1 |
| 20 | 0.0568 | 0.7163 | 0.5994 | 0.6811 | 0.1169 | 1 | 0.0352 | 0 | -0.0817 | -1 |
| 25 | 0.0569 | 0.7170 | 0.6423 | 0.7045 | 0.0747 | 1 | 0.0125 | 0 | -0.0622 | -1 |
| 30 | 0.0552 | 0.7225 | 0.6696 | 0.7366 | 0.0529 | 0 | -0.0141 | 0 | -0.0670 | -1 |
| 35 | 0.0538 | 0.7303 | 0.6867 | 0.7488 | 0.0436 | 0 | -0.0185 | 0 | -0.0621 | -1 |
| 40 | 0.0549 | 0.7308 | 0.6919 | 0.7442 | 0.0389 | 0 | -0.0135 | 0 | -0.0523 | 0 |
| 45 | 0.0541 | 0.7300 | 0.7153 | 0.7490 | 0.0147 | 0 | -0.0190 | 0 | -0.0337 | 0 |
| 50 | 0.0543 | 0.7296 | 0.7158 | 0.7471 | 0.0139 | 0 | -0.0175 | 0 | -0.0313 | 0 |
| 55 | 0.0535 | 0.7299 | 0.7206 | 0.7575 | 0.0092 | 0 | -0.0276 | 0 | -0.0368 | 0 |
| 60 | 0.0535 | 0.7301 | 0.7268 | 0.7688 | 0.0032 | 0 | -0.0387 | 0 | -0.0419 | 0 |
| 65 | 0.0530 | 0.7279 | 0.7401 | 0.7678 | -0.0122 | 0 | -0.0399 | 0 | -0.0277 | 0 |
| 70 | 0.0529 | 0.7287 | 0.7457 | 0.7758 | -0.0169 | 0 | -0.0471 | 0 | -0.0301 | 0 |
| 75 | 0.0530 | 0.7297 | 0.7500 | 0.7783 | -0.0203 | 0 | -0.0486 | 0 | -0.0283 | 0 |
| 80 | 0.0528 | 0.7292 | 0.7545 | 0.7753 | -0.0253 | 0 | -0.0461 | 0 | -0.0208 | 0 |
| 85 | 0.0525 | 0.7341 | 0.7648 | 0.7733 | -0.0307 | 0 | -0.0392 | 0 | -0.0086 | 0 |
| 90 | 0.0521 | 0.7336 | 0.7605 | 0.7700 | -0.0269 | 0 | -0.0364 | 0 | -0.0095 | 0 |
| 95 | 0.0521 | 0.7336 | 0.7629 | 0.7653 | -0.0292 | 0 | -0.0316 | 0 | -0.0024 | 0 |
| 100 | 0.0522 | 0.7383 | 0.7615 | 0.7689 | -0.0232 | 0 | -0.0306 | 0 | -0.0074 | 0 |
| 105 | 0.0523 | 0.7406 | 0.7573 | 0.7646 | -0.0166 | 0 | -0.0240 | 0 | -0.0074 | 0 |
| 110 | 0.0520 | 0.7413 | 0.7615 | 0.7654 | -0.0202 | 0 | -0.0241 | 0 | -0.0039 | 0 |
| 115 | 0.0518 | 0.7468 | 0.7565 | 0.7679 | -0.0098 | 0 | -0.0212 | 0 | -0.0114 | 0 |
| 120 | 0.0518 | 0.7464 | 0.7555 | 0.7676 | -0.0091 | 0 | -0.0212 | 0 | -0.0121 | 0 |
| 125 | 0.0516 | 0.7469 | 0.7579 | 0.7667 | -0.0110 | 0 | -0.0198 | 0 | -0.0089 | 0 |
| 130 | 0.0520 | 0.7478 | 0.7587 | 0.7665 | -0.0109 | 0 | -0.0188 | 0 | -0.0079 | 0 |
| 135 | 0.0519 | 0.7478 | 0.7588 | 0.7689 | -0.0110 | 0 | -0.0211 | 0 | -0.0101 | 0 |
| 140 | 0.0516 | 0.7488 | 0.7613 | 0.7715 | -0.0125 | 0 | -0.0228 | 0 | -0.0103 | 0 |
| 145 | 0.0518 | 0.7548 | 0.7575 | 0.7727 | -0.0026 | 0 | -0.0179 | 0 | -0.0153 | 0 |
| 150 | 0.0518 | 0.7514 | 0.7602 | 0.7795 | -0.0088 | 0 | -0.0280 | 0 | -0.0193 | 0 |

Table B.7. Summary of results of Tukey HSD post hoc test for the classification specificity of B-cell Lymphoma data for α=0.05. Specificity is the fraction of samples identified as negative which are actually negative. Follicular cancer samples are labeled as Negative and B-cell Lymphoma samples as Positive. N – Number of Genes. C – Critical Difference, MeanSS – Mean specificity of Sum of Square method (SS). MeanD8 – Mean Specificity of wavelet (D8). MeanT – Mean specificity of T-test (T).

| N | B-cell Lymphoma Data - Specificity (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0570 | 0.7715 | 0.8820 | 0.8466 | -0.1106 | -1 | -0.0752 | -1 | 0.0354 | 0 |
| 10 | 0.0514 | 0.8041 | 0.8829 | 0.9082 | -0.0788 | -1 | -0.1041 | -1 | -0.0254 | 0 |
| 15 | 0.0485 | 0.8178 | 0.8981 | 0.9182 | -0.0803 | -1 | -0.1004 | -1 | -0.0201 | 0 |
| 20 | 0.0489 | 0.8287 | 0.9079 | 0.9106 | -0.0792 | -1 | -0.0819 | -1 | -0.0027 | 0 |
| 25 | 0.0494 | 0.8286 | 0.9374 | 0.8838 | -0.1088 | -1 | -0.0552 | -1 | 0.0536 | 1 |
| 30 | 0.0462 | 0.8434 | 0.9538 | 0.8831 | -0.1105 | -1 | -0.0398 | 0 | 0.0707 | 1 |
| 35 | 0.0465 | 0.8384 | 0.9588 | 0.8866 | -0.1205 | -1 | -0.0482 | -1 | 0.0722 | 1 |
| 40 | 0.0460 | 0.8410 | 0.9497 | 0.8889 | -0.1087 | -1 | -0.0479 | -1 | 0.0608 | 1 |
| 45 | 0.0457 | 0.8432 | 0.9524 | 0.8881 | -0.1092 | -1 | -0.0450 | 0 | 0.0642 | 1 |
| 50 | 0.0455 | 0.8470 | 0.9421 | 0.8964 | -0.0951 | -1 | -0.0493 | -1 | 0.0457 | 1 |
| 55 | 0.0456 | 0.8633 | 0.9424 | 0.8977 | -0.0792 | -1 | -0.0345 | 0 | 0.0447 | 0 |
| 60 | 0.0438 | 0.8744 | 0.9402 | 0.9065 | -0.0658 | -1 | -0.0321 | 0 | 0.0337 | 0 |
| 65 | 0.0436 | 0.8807 | 0.9362 | 0.9070 | -0.0555 | -1 | -0.0263 | 0 | 0.0293 | 0 |
| 70 | 0.0419 | 0.8801 | 0.9495 | 0.9076 | -0.0694 | -1 | -0.0276 | 0 | 0.0418 | 0 |
| 75 | 0.0413 | 0.8786 | 0.9464 | 0.9196 | -0.0678 | -1 | -0.0411 | 0 | 0.0268 | 0 |
| 80 | 0.0414 | 0.8743 | 0.9442 | 0.9230 | -0.0700 | -1 | -0.0487 | -1 | 0.0212 | 0 |
| 85 | 0.0407 | 0.8753 | 0.9439 | 0.9276 | -0.0686 | -1 | -0.0524 | -1 | 0.0162 | 0 |
| 90 | 0.0406 | 0.8746 | 0.9442 | 0.9286 | -0.0696 | -1 | -0.0541 | -1 | 0.0156 | 0 |
| 95 | 0.0407 | 0.8741 | 0.9445 | 0.9273 | -0.0705 | -1 | -0.0533 | -1 | 0.0172 | 0 |
| 100 | 0.0412 | 0.8797 | 0.9395 | 0.9267 | -0.0598 | -1 | -0.0470 | -1 | 0.0128 | 0 |
| 105 | 0.0418 | 0.8821 | 0.9354 | 0.9277 | -0.0534 | -1 | -0.0457 | -1 | 0.0077 | 0 |
| 110 | 0.0420 | 0.8815 | 0.9349 | 0.9277 | -0.0534 | -1 | -0.0463 | -1 | 0.0071 | 0 |
| 115 | 0.0412 | 0.8861 | 0.9432 | 0.9313 | -0.0572 | -1 | -0.0453 | -1 | 0.0119 | 0 |
| 120 | 0.0411 | 0.8832 | 0.9357 | 0.9394 | -0.0526 | -1 | -0.0563 | -1 | -0.0037 | 0 |
| 125 | 0.0417 | 0.8808 | 0.9360 | 0.9381 | -0.0552 | -1 | -0.0573 | -1 | -0.0021 | 0 |
| 130 | 0.0420 | 0.8813 | 0.9329 | 0.9409 | -0.0516 | -1 | -0.0596 | -1 | -0.0080 | 0 |
| 135 | 0.0405 | 0.8876 | 0.9330 | 0.9429 | -0.0454 | -1 | -0.0553 | -1 | -0.0099 | 0 |
| 140 | 0.0395 | 0.8943 | 0.9417 | 0.9463 | -0.0474 | -1 | -0.0521 | -1 | -0.0047 | 0 |
| 145 | 0.0392 | 0.8961 | 0.9427 | 0.9479 | -0.0466 | -1 | -0.0518 | -1 | -0.0052 | 0 |
| 150 | 0.0378 | 0.9001 | 0.9480 | 0.9492 | -0.0479 | -1 | -0.0491 | -1 | -0.0012 | 0 |

Table B.8. Summary of results of Tukey HSD post hoc test for the classification specificity of Leukemia data for α=0.05. Specificity is the fraction of samples identified as negative which are actually negative. Acute Myeloid Leukemia(AML) samples are labeled as Negative and Acute Lymphoblastic Leukemia(ALL) samples as Positive. N – Number of Genes. C – Critical Difference, MeanSS – Mean specificity of Sum of Square method (SS). MeanD8 – Mean Specificity of wavelet (D8). MeanT – Mean specificity of T-test (T).

| N | Leukemia Data - Specificity (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0424 | 0.7826 | 0.8602 | 0.9066 | -0.0776 | -1 | -0.1240 | -1 | -0.0465 | -1 |
| 10 | 0.0424 | 0.8110 | 0.8190 | 0.9511 | -0.0079 | 0 | -0.1401 | -1 | -0.1322 | -1 |
| 15 | 0.0415 | 0.8246 | 0.8284 | 0.9375 | -0.0039 | 0 | -0.1130 | -1 | -0.1091 | -1 |
| 20 | 0.0417 | 0.8306 | 0.8101 | 0.9337 | 0.0205 | 0 | -0.1031 | -1 | -0.1237 | -1 |
| 25 | 0.0448 | 0.8092 | 0.8160 | 0.9158 | -0.0067 | 0 | -0.1066 | -1 | -0.0999 | -1 |
| 30 | 0.0440 | 0.8082 | 0.8144 | 0.9127 | -0.0061 | 0 | -0.1045 | -1 | -0.0984 | -1 |
| 35 | 0.0431 | 0.8074 | 0.8317 | 0.9137 | -0.0243 | 0 | -0.1063 | -1 | -0.0819 | -1 |
| 40 | 0.0442 | 0.8000 | 0.8308 | 0.9043 | -0.0309 | 0 | -0.1044 | -1 | -0.0735 | -1 |
| 45 | 0.0448 | 0.8021 | 0.8429 | 0.8923 | -0.0408 | 0 | -0.0903 | -1 | -0.0494 | -1 |
| 50 | 0.0438 | 0.8174 | 0.8454 | 0.8920 | -0.0280 | 0 | -0.0746 | -1 | -0.0466 | -1 |
| 55 | 0.0424 | 0.8235 | 0.8528 | 0.8969 | -0.0293 | 0 | -0.0735 | -1 | -0.0441 | -1 |
| 60 | 0.0431 | 0.8116 | 0.8606 | 0.8925 | -0.0490 | -1 | -0.0809 | -1 | -0.0319 | 0 |
| 65 | 0.0435 | 0.8112 | 0.8462 | 0.8979 | -0.0350 | 0 | -0.0867 | -1 | -0.0517 | -1 |
| 70 | 0.0439 | 0.8123 | 0.8485 | 0.8893 | -0.0362 | 0 | -0.0770 | -1 | -0.0408 | 0 |
| 75 | 0.0430 | 0.8133 | 0.8518 | 0.8981 | -0.0385 | 0 | -0.0847 | -1 | -0.0462 | -1 |
| 80 | 0.0431 | 0.8117 | 0.8650 | 0.8936 | -0.0533 | -1 | -0.0819 | -1 | -0.0286 | 0 |
| 85 | 0.0429 | 0.8197 | 0.8616 | 0.8883 | -0.0419 | 0 | -0.0687 | -1 | -0.0267 | 0 |
| 90 | 0.0432 | 0.8161 | 0.8727 | 0.8884 | -0.0566 | -1 | -0.0723 | -1 | -0.0157 | 0 |
| 95 | 0.0417 | 0.8352 | 0.8765 | 0.8911 | -0.0413 | 0 | -0.0559 | -1 | -0.0146 | 0 |
| 100 | 0.0419 | 0.8341 | 0.8768 | 0.8909 | -0.0427 | -1 | -0.0568 | -1 | -0.0141 | 0 |
| 105 | 0.0417 | 0.8376 | 0.8782 | 0.8885 | -0.0406 | 0 | -0.0510 | -1 | -0.0104 | 0 |
| 110 | 0.0411 | 0.8439 | 0.8797 | 0.8865 | -0.0358 | 0 | -0.0426 | -1 | -0.0068 | 0 |
| 115 | 0.0411 | 0.8459 | 0.8759 | 0.8899 | -0.0301 | 0 | -0.0441 | -1 | -0.0140 | 0 |
| 120 | 0.0400 | 0.8565 | 0.8747 | 0.8940 | -0.0182 | 0 | -0.0375 | 0 | -0.0193 | 0 |
| 125 | 0.0403 | 0.8542 | 0.8711 | 0.8910 | -0.0169 | 0 | -0.0368 | 0 | -0.0199 | 0 |
| 130 | 0.0400 | 0.8491 | 0.8772 | 0.8953 | -0.0281 | 0 | -0.0462 | -1 | -0.0181 | 0 |
| 135 | 0.0400 | 0.8483 | 0.8763 | 0.8969 | -0.0280 | 0 | -0.0485 | -1 | -0.0206 | 0 |
| 140 | 0.0401 | 0.8477 | 0.8762 | 0.8987 | -0.0285 | 0 | -0.0510 | -1 | -0.0225 | 0 |
| 145 | 0.0401 | 0.8451 | 0.8788 | 0.8971 | -0.0336 | 0 | -0.0519 | -1 | -0.0183 | 0 |
| 150 | 0.0406 | 0.8445 | 0.8754 | 0.8924 | -0.0310 | 0 | -0.0479 | -1 | -0.0170 | 0 |

Table B.9. Summary of results of Tukey HSD post hoc test for the classification specificity of Colon Cancer data for α=0.05. Specificity is the fraction of samples identified as negative which are actually negative. Normal samples are labeled as Negative and Tumor samples as Positive. N – Number of Genes. C – Critical Difference, MeanSS – Mean specificity of Sum of Square method (SS). MeanD8 – Mean Specificity of wavelet (D8). MeanT – Mean specificity of T-test (T).

| N | Colon cancer data - Specificity (α=0.05) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C (α= 0.05) | MeanSS | MeanD8 | MeanT | MeanSS - MeanD8 | SS v/s D8 | MSS - MT | SS v/s T | MD8-MT | D8 v/s T |
| 5 | 0.0300 | 0.9227 | 0.8876 | 0.7887 | 0.0351 | 1 | 0.1340 | 1 | 0.0989 | 1 |
| 10 | 0.0283 | 0.9107 | 0.8936 | 0.8338 | 0.0171 | 0 | 0.0768 | 1 | 0.0598 | 1 |
| 15 | 0.0290 | 0.9032 | 0.8854 | 0.8334 | 0.0178 | 0 | 0.0698 | 1 | 0.0520 | 1 |
| 20 | 0.0290 | 0.9085 | 0.8928 | 0.8415 | 0.0158 | 0 | 0.0671 | 1 | 0.0513 | 1 |
| 25 | 0.0281 | 0.9030 | 0.8966 | 0.8522 | 0.0064 | 0 | 0.0507 | 1 | 0.0444 | 1 |
| 30 | 0.0267 | 0.9068 | 0.9003 | 0.8624 | 0.0066 | 0 | 0.0444 | 1 | 0.0378 | 1 |
| 35 | 0.0253 | 0.9083 | 0.8998 | 0.8698 | 0.0085 | 0 | 0.0385 | 1 | 0.0300 | 1 |
| 40 | 0.0250 | 0.9037 | 0.9019 | 0.8714 | 0.0019 | 0 | 0.0323 | 1 | 0.0305 | 1 |
| 45 | 0.0247 | 0.9018 | 0.9053 | 0.8766 | -0.0035 | 0 | 0.0253 | 1 | 0.0288 | 1 |
| 50 | 0.0245 | 0.8995 | 0.9074 | 0.8856 | -0.0079 | 0 | 0.0139 | 0 | 0.0218 | 0 |
| 55 | 0.0239 | 0.8973 | 0.9115 | 0.8887 | -0.0143 | 0 | 0.0086 | 0 | 0.0228 | 0 |
| 60 | 0.0238 | 0.8975 | 0.9095 | 0.8901 | -0.0120 | 0 | 0.0073 | 0 | 0.0193 | 0 |
| 65 | 0.0232 | 0.8973 | 0.9121 | 0.8926 | -0.0148 | 0 | 0.0047 | 0 | 0.0195 | 0 |
| 70 | 0.0235 | 0.8991 | 0.9071 | 0.8919 | -0.0081 | 0 | 0.0071 | 0 | 0.0152 | 0 |
| 75 | 0.0238 | 0.8991 | 0.9082 | 0.8921 | -0.0092 | 0 | 0.0070 | 0 | 0.0162 | 0 |
| 80 | 0.0239 | 0.9006 | 0.9069 | 0.8901 | -0.0062 | 0 | 0.0105 | 0 | 0.0168 | 0 |
| 85 | 0.0238 | 0.8995 | 0.9101 | 0.8876 | -0.0107 | 0 | 0.0119 | 0 | 0.0225 | 0 |
| 90 | 0.0238 | 0.8985 | 0.9091 | 0.8877 | -0.0106 | 0 | 0.0108 | 0 | 0.0214 | 0 |
| 95 | 0.0239 | 0.8998 | 0.9067 | 0.8867 | -0.0070 | 0 | 0.0131 | 0 | 0.0201 | 0 |
| 100 | 0.0240 | 0.8988 | 0.9080 | 0.8865 | -0.0092 | 0 | 0.0123 | 0 | 0.0214 | 0 |
| 105 | 0.0241 | 0.8973 | 0.9061 | 0.8866 | -0.0088 | 0 | 0.0107 | 0 | 0.0195 | 0 |
| 110 | 0.0243 | 0.8986 | 0.9066 | 0.8852 | -0.0081 | 0 | 0.0133 | 0 | 0.0214 | 0 |
| 115 | 0.0244 | 0.8966 | 0.9087 | 0.8865 | -0.0122 | 0 | 0.0100 | 0 | 0.0222 | 0 |
| 120 | 0.0242 | 0.8980 | 0.9077 | 0.8884 | -0.0097 | 0 | 0.0096 | 0 | 0.0193 | 0 |
| 125 | 0.0241 | 0.8997 | 0.9088 | 0.8879 | -0.0091 | 0 | 0.0118 | 0 | 0.0209 | 0 |
| 130 | 0.0239 | 0.9007 | 0.9121 | 0.8892 | -0.0114 | 0 | 0.0114 | 0 | 0.0228 | 0 |
| 135 | 0.0240 | 0.9002 | 0.9121 | 0.8887 | -0.0119 | 0 | 0.0115 | 0 | 0.0233 | 0 |
| 140 | 0.0242 | 0.9001 | 0.9121 | 0.8855 | -0.0120 | 0 | 0.0146 | 0 | 0.0266 | 1 |
| 145 | 0.0243 | 0.8980 | 0.9126 | 0.8857 | -0.0146 | 0 | 0.0123 | 0 | 0.0269 | 1 |
| 150 | 0.0244 | 0.8993 | 0.9115 | 0.8863 | -0.0122 | 0 | 0.0130 | 0 | 0.0252 | 1 |

APPENDIX C

GENE LISTS

Table C.1. Top 100 genes selected by Db -8 wavelet from Leukemia dataset.

| Rank | Gene Name | Average Score |
|---|---|---|
| 1 | "'CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage' | 0.475 |
| 2 | "'MPO Myeloperoxidas' | 0.404 |
| 3 | "'Azurocidin gen' | 0.387 |
| 4 | "'INTERLEUKIN-8 PRECURSO' | 0.384 |
| 5 | "'FTL Ferritin, light polypeptid' | 0.340 |
| 6 | "'GPX1 Glutathione peroxidase ' | 0.297 |
| 7 | "'TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA for Tcell leukemia/lymphoma ' | 0.294 |
| 8 | "'DF D component of complement (adipsin' | 0.292 |
| 9 | "'Lysozyme gene (EC 3.2.1.17' | 0.288 |
| 10 | "'PRG1 Proteoglycan 1, secretory granul' | 0.285 |
| 11 | "'LYZ Lysozym' | 0.282 |
| 12 | "'Cystic fibrosis antigen mRN' | 0.277 |
| 13 | "'LYZ Lysozym' | 0.268 |
| 14 | "'PROBABLE G PROTEIN-COUPLED RECEPTOR LCR1 HOMOLO' | 0.258 |
| 15 | "'MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD' | 0.246 |
| 16 | "'Interleukin 8 (IL8) gen' | 0.244 |
| 17 | "'GLUL Glutamate-ammonia ligase (glutamine synthase' | 0.240 |
| 18 | "'SM22-ALPHA HOMOLO' | 0.240 |
| 19 | "'ENO1 Enolase 1, (alpha' | 0.238 |
| 20 | "'IGHM Immunoglobulin m' | 0.236 |
| 21 | "'CTSD Cathepsin D (lysosomal aspartyl protease' | 0.234 |
| 22 | "'FTH1 Ferritin heavy chai' | 0.231 |
| 23 | 'MB-1 gene' | 0.227 |
| 24 | "'Histone H2A.2 mRN' | 0.227 |

Table C.1. Top 100 genes selected by Db -8 wavelet from Leukemia dataset. (Continued).

| | | |
|---|---|---|
| 25 | "'(hybridoma H210) anti-hepatitis A IgG variable region, constant region, complementarity-determining regions mRN' | 0.226 |
| 26 | "'26-kDa cell surface protein TAPA-1 mRN' | 0.223 |
| 27 | "'Terminal transferase mRN' | 0.219 |
| 28 | "'ALDOA Aldolase ' | 0.213 |
| 29 | "'PSAP Sulfated glycoprotein ' | 0.207 |
| 30 | "'ELA2 Elastatse 2, neutrophi' | 0.205 |
| 31 | "'ADA Adenosine deaminas' | 0.203 |
| 32 | "'CCND3 Cyclin D' | 0.196 |
| 33 | "'CATHEPSIN G PRECURSO' | 0.194 |
| 34 | "'LYZ Lysozym' | 0.194 |
| 35 | "'Metallothionein isoform ' | 0.191 |
| 36 | "'Neutrophil elastase gene, exon ' | 0.188 |
| 37 | "'HMG1 High-mobility group (nonhistone chromosomal) protein ' | 0.185 |
| 38 | "'PROTEASOME IOTA CHAI' | 0.182 |
| 39 | "'Immunoglobulin lambda gene locus DNA, clone:123E' | 0.182 |
| 40 | "'Ig alpha 2=immunoglobulin A heavy chain allotype 2 {constant region, germ line} [human, peripheral blood neutrophils, Genomic, 1..." <' | 0.181 |
| 41 | "'LGALS1 Ubiquinol-cytochrome c reductase core protein I' | 0.180 |
| 42 | "'PAGA Proliferation-associated gene A (natural killer-enhancing factor A' | 0.178 |
| 43 | "'mRNA fragment encoding beta-tubulin. (from clone D-beta-1' | 0.178 |
| 44 | "'CYBA Cytochrome b-245, alpha polypeptid' | 0.177 |
| 45 | "'GRN Granuli' | 0.171 |
| 46 | "'Polyadenylate binding protein I' | 0.169 |
| 47 | 'Zyxin' | 0.168 |
| 48 | 'YMP mRNA' | 0.163 |
| 49 | "'PTMA gene extracted from Human prothymosin alpha mRN' | 0.163 |
| 50 | "'GAMMA-INTERFERON-INDUCIBLE PROTEIN IP-30 PRECURSO' | 0.160 |
| 51 | "'THYMOSIN BETA-1' | 0.159 |
| 52 | "'Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRN' | 0.158 |
| 53 | 'CYSTATIN A' | 0.155 |
| 54 | "'KIAA0085 gene, partial cd' | 0.154 |
| 55 | 'Calcyclin' | 0.153 |
| 56 | "'SELL Leukocyte adhesion protein beta subuni' | 0.152 |
| 57 | "'PLACENTAL CALCIUM-BINDING PROTEI' | 0.152 |
| 58 | "'TOP2B Topoisomerase (DNA) II beta (180kD' | 0.150 |
| 59 | "'NPM1 Nucleophosmin (nucleolar phosphoprotein B23, numatrin' | 0.148 |
| 60 | "'C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cd' | 0.148 |

Table C.1. Top 100 genes selected by Db -8 wavelet from Leukemia dataset. (Continued).

| | | |
|---|---|---|
| 61 | "'CALGRANULIN ' | 0.146 |
| 62 | "'HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, F ALPHA CHAIN PRECURSO' | 0.145 |
| 63 | 'Macmarcks' | 0.142 |
| 64 | "'LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol' | 0.141 |
| 65 | "'PERIPHERAL-TYPE BENZODIAZEPINE RECEPTO' | 0.133 |
| 66 | "'60S RIBOSOMAL PROTEIN L1' | 0.129 |
| 67 | "'TCRB T-cell receptor, beta cluste' | 0.127 |
| 68 | "'Catalase (EC 1.11.1.6) 5"flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS' | 0.126 |
| 69 | "'APLP2 Amyloid beta (A4) precursor-like protein ' | 0.124 |
| 70 | "'VIL2 Villin 2 (ezrin' | 0.116 |
| 71 | 'LPAP gene' | 0.115 |
| 72 | "'SELL Leukocyte adhesion protein beta subuni' | 0.112 |
| 73 | "'ZFP36 Zinc finger protein homologous to Zfp-36 in mous' | 0.111 |
| 74 | "'Omega light chain protein 14.1 (Ig lambda chain related) gene, exon ' | 0.109 |
| 75 | "'INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL' | 0.107 |
| 76 | "'LTB Lymphotoxin-bet' | 0.107 |
| 77 | "'CD24 signal transducer mRNA and 3" regio' | 0.100 |
| 78 | "'HSPB1 Heat shock 27kD protein ' | 0.097 |
| 79 | "'Fc-epsilon-receptor gamma-chain mRN' | 0.097 |
| 80 | "'54 kDa protein mRN' | 0.097 |
| 81 | "'Kazal-type serine proteinase (HUSI-II) gen' | 0.096 |
| 82 | "'Immunoglobulin mu, part of exon ' | 0.094 |
| 83 | "'RPS3 Ribosomal protein S' | 0.089 |
| 84 | "'PLCB2 Phospholipase C, beta ' | 0.085 |
| 85 | "'Oncoprotein 18 (Op18) gen' | 0.083 |
| 86 | "'G-gamma globin gene extracted from H.sapiens G-gamma globin and A-gamma globin genes'" | 0.075 |
| 87 | "'IL7R Interleukin 7 recepto' | 0.072 |
| 88 | "'VIM Vimenti' | 0.071 |
| 89 | "'Histone H1' | 0.068 |
| 90 | "'CALM1 Calmodulin 1 (phosphorylase kinase, delta' | 0.067 |
| 91 | "'EIF4A2 Eukaryotic translation initiation factor 4A (eIF-4A) isoform ' | 0.066 |
| 92 | "'NF-IL6-beta protein mRN' | 0.065 |
| 93 | "'SAT Spermidine/spermine N1-acetyltransferas' | 0.063 |
| 94 | "'High mobility group protein (HMG-I(Y)) gene exons 1-' | 0.060 |

Table C.1. Top 100 genes selected by Db -8 wavelet from Leukemia dataset. (Continued).

| | | |
|---|---|---|
| 95 | "'X BOX BINDING PROTEIN-' | 0.058 |
| 96 | "'PIM1 Pim-1 oncogen' | 0.057 |
| 97 | "'SOD-2 gene for manganese superoxide dismutas' | 0.057 |
| 98 | "'T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRN' | 0.054 |
| 99 | "'JUNB Jun B proto-oncogen' | 0.050 |
| 100 | "'SOX4 SRY (sex determining region Y)-box ' | 0.049 |

Table C.2. Top 100 probes selected by D-8 wavelets from Lymphoma Dataset.

| Rank | Gene Name | Average Score |
|---|---|---|
| 1 | 'Metallothionein isoform 2' | 0.443 |
| 2 | 'LDHA Lactate dehydrogenase A' | 0.428 |
| 3 | 'ENO1 Enolase 1, (alpha)' | 0.407 |
| 4 | 'Cathepsin B' | 0.370 |
| 5 | 'PKM2 Pyruvate kinase, muscle' | 0.318 |
| 6 | 'PSAP Sulfated glycoprotein 1' | 0.310 |
| 7 | 'CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)' | 0.305 |
| 8 | 'Macrophage migration inhibitory factor (MIF) gene' | 0.298 |
| 9 | 'GAMMA-INTERFERON-INDUCIBLE PROTEIN IP-30 PRECURSOR' | 0.296 |
| 10 | 'APOE Apolipoprotein E' | 0.292 |
| 11 | '60S RIBOSOMAL PROTEIN L13' | 0.288 |
| 12 | 'High mobility group protein (HMG-I(Y)) gene exons 1-8' | 0.284 |
| 13 | 'Tubulin, Beta 2' | 0.283 |
| 14 | 'mRNA fragment for elongation factor TU (N-terminus)' | 0.265 |
| 15 | 'mRNA fragment encoding beta-tubulin. (from clone D-beta-1)' | 0.264 |
| 16 | 'CTSD Cathepsin D (lysosomal aspartyl protease)' | 0.261 |
| 17 | 'Triosephosphate Isomerase' | 0.254 |
| 18 | 'Alpha-1 collagen type I gene, 3" end' | 0.252 |
| 19 | 'LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)' | 0.249 |
| 20 | 'Humig mRNA' | 0.249 |
| 21 | 'PAGA Proliferation-associated gene A (natural killer-enhancing factor A)' | 0.247 |
| 22 | 'HSPD1 Heat shock 60 kD protein 1 (chaperonin)' | 0.246 |
| 23 | 'Brain-expressed HHCPA78 homolog [human, HL-60 acute promyelocytic leukemia cells, mRNA, 2704 nt]' | 0.242 |
| 24 | 'CLTA Clathrin light chain A' | 0.238 |

Table C.2. Top 100 probes selected by D-8 wavelets from Lymphoma Dataset. (Continued).

| | | |
|---|---|---|
| 25 | 'ALDOA Aldolase A' | 0.237 |
| 26 | 'PGAM1 Phosphoglycerate mutase 1 (brain)' | 0.228 |
| 27 | 'FTH1 Ferritin heavy chain' | 0.227 |
| 28 | 'PABPL1 Poly(A)-binding protein-like 1' | 0.219 |
| 29 | 'HLA-A MHC class I protein HLA-A (HLA-A28,-B40, -Cw3)' | 0.211 |
| 30 | 'Proteasome activator hPA28 subunit beta' | 0.210 |
| 31 | 'PGK1 Phosphoglycerate kinase 1' | 0.199 |
| 32 | 'SLC' | 0.198 |
| 33 | 'Metallothionein isoform 2' | 0.197 |
| 34 | 'HMG1 High-mobility group (nonhistone chromosomal) protein 1' | 0.197 |
| 35 | 'SNRPB Small nuclear ribonucleoprotein polypeptides B and B1' | 0.196 |
| 36 | 'Nucleolin gene' | 0.190 |
| 37 | 'Cytochrome c oxidase subunit VIII (COX8) mRNA' | 0.187 |
| 38 | 'MMP2 Matrix metalloproteinase 2 (gelatinase A; collagenase type IV)' | 0.182 |
| 39 | 'CTSH Cathepsin H' | 0.180 |
| 40 | '26-kDa cell surface protein TAPA-1 mRNA' | 0.178 |
| 41 | 'Cystatin B gene' | 0.177 |
| 42 | 'Omega light chain protein 14.1 (Ig lambda chain related) gene, exon 3' | 0.174 |
| 43 | 'TCRB T-cell receptor, beta cluster' | 0.172 |
| 44 | 'BRCA2 region, mRNA sequence CG037' | 0.170 |
| 45 | 'Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA' | 0.168 |
| 46 | 'CD20 RECEPTOR' | 0.166 |
| 47 | 'PTMA gene extracted from Human prothymosin alpha mRNA' | 0.166 |
| 48 | 'ANT2 Adenine nucleotide translocator 2 (fibroblast)' | 0.165 |
| 49 | 'CAPG Capping protein (actin filament), gelsolin-like' | 0.162 |
| 50 | 'C1QB Complement component 1, q subcomponent, beta polypeptide' | 0.159 |
| 51 | 'HSPB1 Heat shock 27kD protein 1' | 0.153 |
| 52 | 'ANT3 Adenine nucleotide translocator 3 (liver)' | 0.152 |
| 53 | 'PHAPI2b protein' | 0.151 |
| 54 | 'COX7C Cytochrome c oxidase VIIc subunit' | 0.149 |
| 55 | 'Elongation factor-1-beta' | 0.148 |
| 56 | 'CD37 CD37 antigen' | 0.146 |
| 57 | 'Liver mRNA fragment DNA binding protein UPI homologue (C-terminus)' | 0.143 |
| 58 | 'Nucleoside Diphosphate Kinase Nm23-H2s' | 0.141 |
| 59 | 'TXN Thioredoxin' | 0.140 |
| 60 | 'MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3' | 0.140 |
| 61 | 'DbpB-like protein mRNA' | 0.138 |

Table C.2. Top 100 probes selected by D-8 wavelets from Lymphoma Dataset. (Continued).

| | | |
|---|---|---|
| 62 | 'Stimulator of TAR RNA binding (SRB) mRNA' | 0.138 |
| 63 | '54 kDa protein mRNA' | 0.138 |
| 64 | 'FN1 Fibronectin 1' | 0.136 |
| 65 | 'Myosin, Light Chain, Alkali, Smooth Muscle (Gb:U02629), Non-Muscle, Alt. Splice 2' | 0.135 |
| 66 | 'NME1 Non-metastatic cells 1, protein (NM23A) expressed in' | 0.135 |
| 67 | 'SOD1 Superoxide dismutase 1 (Cu/Zn)' | 0.133 |
| 68 | 'MLN50 mRNA' | 0.132 |
| 69 | 'ATP5B ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide' | 0.129 |
| 70 | 'NMB Neuromedin B' | 0.127 |
| 71 | 'LTB Lymphotoxin-beta' | 0.124 |
| 72 | 'GAPD Glyceraldehyde-3-phosphate dehydrogenase' | 0.124 |
| 73 | 'ATP5A1 ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle' | 0.123 |
| 74 | 'CYTOCHROME C OXIDASE POLYPEPTIDE VIA-LIVER PRECURSOR' | 0.117 |
| 75 | 'ARH9 Aplysia ras-related homolog 9' | 0.117 |
| 76 | '"L44L gene (L44-like ribosomal protein) extracted from Human Bruton"s tyrosine kinase (BTK), alpha-D-galactosidase A (GLA), L44-l..." <Preview truncated at 128 characters>' | 0.112 |
| 77 | 'Small Nuclear Ribonucleoprotein, Polypeptide C, Alt. Splice 2' | 0.101 |
| 78 | 'Immunoglobulin mu, part of exon 8' | 0.101 |
| 79 | 'IMMUNOGLOBULIN J CHAIN' | 0.100 |
| 80 | '5-aminoimidazole-4-carboxamide-1-beta-D-ribonucleoti de transformylase/inosinicase' | 0.098 |
| 81 | 'EIF4A1 Eukaryotic translation initiation factor 4A (eIF-4A) isoform 1' | 0.097 |
| 82 | 'LGALS1 Ubiquinol-cytochrome c reductase core protein II' | 0.094 |
| 83 | 'Protein Phosphatase 1, Alpha Catalytic Subunit' | 0.093 |
| 84 | 'RPS3 Ribosomal protein S3' | 0.091 |
| 85 | 'COX4 Cytochrome c oxidase subunit IV' | 0.091 |
| 86 | 'GARS Glycyl-tRNA synthetase' | 0.089 |
| 87 | 'Arp2/3 protein complex subunit p41-Arc (ARC41) mRNA' | 0.089 |
| 88 | 'Major Histocompatibility Complex, Class I, E (Gb:M21533)' | 0.089 |
| 89 | 'Lactate dehydrogenase B gene exon 1 and 2 (EC 1.1.1.27) (and joined CDS)' | 0.087 |
| 90 | 'JunD mRNA' | 0.083 |
| 91 | 'SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 1-ALPHA/BETA' | 0.082 |
| 92 | 'FBP1 Fructose-bisphosphatase 1' | 0.082 |
| 93 | '"COX6B gene (COXG) extracted from Human DNA from overlapping chromosome 19 cosmids R31396, F25451, and R31076 containing COX6B an..." <Preview truncated' | 0.078 |
| 94 | 'Bcl-2 related (Bfl-1) mRNA' | 0.076 |

Table C.2. Top 100 probes selected by D-8 wavelets from Lymphoma Dataset. (Continued).

| | | |
|---|---|---|
| 95 | 'GRN Granulin' | 0.074 |
| 96 | 'Ribosomal Protein S12' | 0.073 |
| 97 | 'Mitochondrial ATP synthase subunit 9, P3 gene copy, mRNA, nuclear gene encoding mitochondrial protein' | 0.072 |
| 98 | 'SAT Spermidine/spermine N1-acetyltransferase' | 0.066 |
| 99 | 'THYMOSIN BETA-10' | 0.064 |
| 100 | 'TCRB T-cell receptor, beta cluster' | 0.061 |

Table C.3. Top 100 ESTs selected by DB-8 wavelet from Colon Dataset.

| Rank | Gene Name | Average Score |
|---|---|---|
| 1 | '"Hsa.4689 T95018 3" UTR 2a 120032 40S RIBOSOMAL PROTEIN S18 (Homo sapiens) | 0.135 |
| 2 | '"Hsa.8147 M63391 gene 1 "Human desmin gene, complete cds. | 0.099 |
| 3 | '"Hsa.140 M87789 gene 1 IG GAMMA-1 CHAIN C REGION (HUMAN);. | 0.094 |
| 4 | '"Hsa.5398 T58861 3" UTR 2a 77563 60S RIBOSOMAL PROTEIN L30E (Kluyveromyces lactis) | 0.085 |
| 5 | '"Hsa.878 T61609 3" UTR 1 78081 LAMININ RECEPTOR (HUMAN);. | 0.085 |
| 6 | '"Hsa.1534 J00231 gene 1 Human Ig gamma3 heavy chain disease OMM protein mRNA. | 0.081 |
| 7 | '"Hsa.1131 T92451 3" UTR 1 118219 "TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN);. | 0.077 |
| 8 | '"Hsa.3004 H55933 3" UTR 1 203417 H.sapiens mRNA for homologue to yeast ribosomal protein L41. | 0.075 |
| 9 | '"Hsa.3002 R22197 3" UTR 1 130829 60S RIBOSOMAL PROTEIN L32 (HUMAN);. | 0.072 |
| 10 | '"Hsa.3087 T65938 3" UTR 1 81639 TRANSLATIONALLY CONTROLLED TUMOR PROTEIN (HUMAN);. | 0.071 |
| 11 | '"Hsa.539 U14971 gene 1 "Human ribosomal protein S9 mRNA, complete cds. | 0.071 |
| 12 | '"Hsa.8068 T57619 3" UTR 2a 75437 40S RIBOSOMAL PROTEIN S6 (Nicotiana tabacum) | 0.068 |
| 13 | '"Hsa.750 T72863 3" UTR 1 84277 FERRITIN LIGHT CHAIN (HUMAN);. | 0.066 |
| 14 | '"Hsa.1737 T72175 3" UTR 1 85528 IG KAPPA CHAIN PRECURSOR V-III REGION (HUMAN);. | 0.063 |
| 15 | '"Hsa.43279 H64489 3" UTR 2a 238846 LEUKOCYTE ANTIGEN CD37 (Homo sapiens) | 0.062 |
| 16 | '"Hsa.10755 R78934 3" UTR 2a 146232 ENDOTHELIAL ACTIN-BINDING PROTEIN (Homo sapiens) | 0.062 |
| 17 | '"Hsa.1130 Z24727 gene 1 "H.sapiens tropomyosin isoform mRNA, complete CDS. | 0.061 |

Table C.3. Top 100 ESTs selected by DB-8 wavelet from Colon Dataset. (Continued).

| | | |
|---|---|---|
| 18 | "'Hsa.5444 T48804 3" UTR 1 70269 40S RIBOSOMAL PROTEIN S24 (HUMAN). | 0.059 |
| 19 | "'Hsa.538 T56940 3" UTR 1 68306 P24050 40S RIBOSOMAL PROTEIN. | 0.059 |
| 20 | "'Hsa.1221 T60155 3" UTR 1 81422 "ACTIN, AORTIC SMOOTH MUSCLE (HUMAN);. | 0.059 |
| 21 | "'Hsa.957 M26697 gene 1 "Human nucleolar protein (B23) mRNA, complete cds. | 0.058 |
| 22 | "'Hsa.467 H20709 3" UTR 1 173155 "MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN);. | 0.058 |
| 23 | "'Hsa.316 M94132 gene 1 Human mucin 2 (MUC2) mRNA sequence. | 0.057 |
| 24 | "'Hsa.6080 J02763 gene 1 "Human calcyclin gene, complete cds. | 0.056 |
| 25 | "'Hsa.285 T62972 3" UTR 1 80738 P02403 60S RIBOSOMAL PROTEIN ;. | 0.056 |
| 26 | "'Hsa.891 M19045 gene 1 "Human lysozyme mRNA, complete cds. | 0.055 |
| 27 | "'Hsa.13491 R39465 3" UTR 2a 23933 EUKARYOTIC INITIATION FACTOR 4A (Oryctolagus cuniculus) | 0.053 |
| 28 | "'Hsa.2597 T49423 3" UTR 1 67494 BREAST BASIC CONSERVED PROTEIN 1 (HUMAN). | 0.050 |
| 29 | "'Hsa.20836 R02593 3" UTR 2a 124094 60S ACIDIC RIBOSOMAL PROTEIN P1 (Polyorchis penicillatus) | 0.050 |
| 30 | "'Hsa.1977 T51496 3" UTR 1 71488 60S RIBOSOMAL PROTEIN L37A (HUMAN). | 0.048 |
| 31 | "'Hsa.1985 T52185 3" UTR 1 71940 P17074 40S RIBOSOMAL PROTEIN. | 0.047 |
| 32 | "'Hsa.692 M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. | 0.047 |
| 33 | "'Hsa.692 M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. | 0.047 |
| 34 | "'Hsa.692 M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. | 0.046 |
| 35 | "'Hsa.474 L28809 gene 1 "Homo sapiens dbpB-like protein mRNA, complete cds. | 0.046 |
| 36 | "'Hsa.8125 T71025 3" UTR 1 84103 Human (HUMAN);. | 0.045 |
| 37 | "'Hsa.2688 X60489 gene 1 Human mRNA for elongation factor-1-beta. | 0.045 |
| 38 | "'Hsa.954 T72938 3" UTR 1 84350 QM PROTEIN (HUMAN);. | 0.045 |
| 39 | "'Hsa.832 T51023 3" UTR 1 75127 HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN). | 0.044 |
| 40 | "'Hsa.5363 R01182 3" UTR 1 123748 60S RIBOSOMAL PROTEIN L38 (HUMAN);. | 0.044 |
| 41 | "'Hsa.3566 T57633 3" UTR 1 75467 40S RIBOSOMAL PROTEIN S8 (HUMAN). | 0.044 |

Table C.3. Top 100 ESTs selected by DB-8 wavelet from Colon Dataset. (Continued).

| | | |
|---|---|---|
| 42 | "'Hsa.7877 R86975 3" UTR 1 197282 40S RIBOSOMAL PROTEIN S28 (HUMAN);. | 0.044 |
| 43 | "'Hsa.31 T57780 3" UTR 1 80626 IG LAMBDA CHAIN C REGIONS (HUMAN). | 0.0434 |
| 44 | "'Hsa.489 T47144 3" UTR 1 74837 JN0549 RIBOSOMAL PROTEIN YL30. | 0.043 |
| 45 | "'Hsa.1119 T59954 3" UTR 1 79441 THYMOSIN BETA-4 (HUMAN);. | 0.043 |
| 46 | "'Hsa.831 M22382 gene 1 MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN);. | 0.043 |
| 47 | "'Hsa.13491 R39465 3" UTR 2a 23933 EUKARYOTIC INITIATION FACTOR 4A (Oryctolagus cuniculus) | 0.042 |
| 48 | "'UMGAP | 0.042 |
| 49 | "'UMGAP | 0.042 |
| 50 | "'UMGAP | 0.042 |
| 51 | "'UMGAP | 0.042 |
| 52 | "'Hsa.678 H55758 3" UTR 1 203413 ALPHA ENOLASE (HUMAN);. | 0.041 |
| 53 | "'Hsa.2800 X55715 gene 1 Human Hums3 mRNA for 40S ribosomal protein s3. | 0.041 |
| 54 | "'Hsa.733 M14200 gene 1 "Human diazepam binding inhibitor (DBI) mRNA, complete cds. | 0.040 |
| 55 | "'Hsa.31 T57780 3" UTR 1 80626 IG LAMBDA CHAIN C REGIONS (HUMAN). | 0.040 |
| 56 | "'Hsa.1832 J02854 gene 1 "MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element | 0.040 |
| 57 | "'Hsa.3016 T47377 3" UTR 1 71035 S-100P PROTEIN (HUMAN). | 0.039 |
| 58 | "'Hsa.1447 T55131 3" UTR 1 73931 "GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (HUMAN). | 0.0382 |
| 59 | "'Hsa.2794 T48904 3" UTR 1 70455 HEAT SHOCK 27 KD PROTEIN (HUMAN). | 0.037 |
| 60 | "'Hsa.2948 H54676 3" UTR 1 203220 60S RIBOSOMAL PROTEIN L18A (HUMAN);. | 0.036 |
| 61 | "'Hsa.1902 L05144 gene 1 "PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC (HUMAN);contains Alu repetitive el | 0.036 |
| 62 | "'Hsa.558 R34698 3" UTR 1 136738 INTERFERON-INDUCIBLE PROTEIN 9-27 (HUMAN);. | 0.035 |
| 63 | "'Hsa.2221 T52015 3" UTR 1 72642 ELONGATION FACTOR 1-GAMMA (HUMAN). | 0.034 |
| 64 | "'Hsa.5710 T63484 3" UTR 1 81437 "Human ornithine decarboxylase antizyme (Oaz) mRNA, complete cds. | 0.034 |
| 65 | "'Hsa.451 D21261 gene 1 SM22-ALPHA HOMOLOG (HUMAN);. | 0.033 |
| 66 | "'Hsa.24464 H09263 3" UTR 2a 46514 ELONGATION FACTOR 1-ALPHA 1 (Homo sapiens) | 0.033 |

Table C.3. Top 100 ESTs selected by DB-8 wavelet from Colon Dataset. (Continued).

| | | |
|---|---|---|
| 67 | "'Hsa.41315 U37012 gene 1 "Human cleavage and polyadenylation specificity factor mRNA, complete cds. | 0.033 |
| 68 | "'Hsa.3006 T61602 3" UTR 1 78084 40S RIBOSOMAL PROTEIN | 0.032 |
| 69 | "'Hsa.3835 H79852 3" UTR 2a 239944 60S ACIDIC RIBOSOMAL PROTEIN P2 (Babesia bovis) | 0.032 |
| 70 | "'Hsa.541 U14973 gene 1 "Human ribosomal protein S29 mRNA, complete cds. | 0.031 |
| 71 | "'Hsa.951 M36981 gene 1 "Human putative NDP kinase (nm23-H2S) mRNA, complete cds. | 0.030 |
| 72 | "'Hsa.2555 X63432 gene 1 H.sapiens ACTB mRNA for mutant beta-actin (beta"-actin). | 0.029 |
| 73 | "'HSAC07 | 0.029 |
| 74 | "'HSAC07 | 0.029 |
| 75 | "'Hsa.2588 H40560 3" UTR 1 175410 THIOREDOXIN (HUMAN);. | 0.029 |
| 76 | "'HSAC07 | 0.029 |
| 77 | "'HSAC07 | 0.029 |
| 78 | "'Hsa.3306 X12671 gene 1 Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1. | 0.028 |
| 79 | "'Hsa.37937 R87126 3" UTR 2a 197371 "MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus) | 0.027 |
| 80 | "'Hsa.2361 T51534 3" UTR 1 72396 CYSTATIN C PRECURSOR (HUMAN). ..." Preview truncated at 128 characters>' | 0.027 |
| 81 | "Hsa.5821 X57351 gene 1 INTERFERON-INDUCIBLE PROTEIN 1-8D (HUMAN);contains MSR1 repetitive element ;. | 0.027 |
| 82 | "'Hsa.36952 H43887 3" UTR 2a 183264 COMPLEMENT FACTOR D PRECURSOR (Homo sapiens) | 0.026 |
| 83 | "'Hsa.27685 R50158 3" UTR 2a 153229 MITOCHONDRIAL LON PROTEASE HOMOLOG PRECURSOR (Homo sapiens) | 0.025 |
| 84 | "'Hsa.1836 T51574 3" UTR 1 72258 40S RIBOSOMAL PROTEIN S24 (HUMAN). | 0.025 |
| 85 | "'Hsa.1254 M18216 gene 1 "Human nonspecific crossreacting antigen mRNA, complete cds. | 0.024 |
| 86 | "'Hsa.20836 R02593 3" UTR 2a 124094 60S ACIDIC RIBOSOMAL PROTEIN P1 (Polyorchis penicillatus) | 0.024 |
| 87 | "'Hsa.2665 T68848 3" UTR 1 82178 PEPTIDYL-PROLYL CIS-TRANS ISOMERASE A (HUMAN);. | 0.023 |
| 88 | "'Hsa.773 H40095 3" UTR 1 175181 MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN);. | 0.022 |
| 89 | "'Hsa.37254 R85482 3" UTR 2a 180093 SERUM RESPONSE FACTOR (Homo sapiens) | 0.022 |
| 90 | "'Hsa.1205 R08183 3" UTR 1 127228 "Q04984 10 KD HEAT SHOCK PROTEIN, MITOCHONDRIAL ;. | 0.021 |
| 91 | "'Hsa.3348 X15880 gene 1 Human mRNA for collagen VI alpha-1 C-terminal globular domain. | 0.021 |

Table C.3. Top 100 ESTs selected by DB-8 wavelet from Colon Dataset. (Continued).

| | | |
|---|---|---|
| 92 | "'Hsa.2357 T52342 3" UTR 1 72028 Human tra1 mRNA for human homologue of murine tumor rejection antigen gp96. | 0.019 |
| 93 | "'Hsa.8831 T49941 3" UTR 1 69828 PUTATIVE INSULIN-LIKE GROWTH FACTOR II ASSOCIATED (HUMAN). | 0.019 |
| 94 | "'Hsa.98 T93094 3" UTR 1 118704 ANNEXIN II (HUMAN);. | 0.019 |
| 95 | "'Hsa.5346 T63370 3" UTR 2a 81523 GUANINE NUCLEOTIDE-BINDING PROTEIN BETA SUBUNIT-LIKE PROTEIN 12.3 (Homo sapiens) | 0.019 |
| 96 | "'Hsa.1098 M33680 gene 1 "Human 26-kDa cell surface protein TAPA-1 mRNA, complete cds. | 0.018 |
| 97 | "'Hsa.45604 H88360 3" UTR 2a 252849 "GUANINE NUCLEOTIDE-BINDING PROTEIN G(OLF), ALPHA SUBUNIT (Rattus norvegicus) | 0.018 |
| 98 | "'Hsa.1978 T72879 3" UTR 1 84299 60S RIBOSOMAL PROTEIN L7A (HUMAN);. | 0.016 |
| 99 | "'Hsa.2753 X68314 gene 1 H.sapiens mRNA for glutathione peroxidase-GI. | 0.016 |
| 100 | "'Hsa.1732 U12255 gene 1 "Human IgG Fc receptor hFcRn mRNA, complete cds. | 0.014 |

Table C.4. Genes Common between all the three methods for B-Cell Lymphoma Data.

| Gene Rank | Gene ID | Score |
|---|---|---|
| 1 | 'ENO1 Enolase 1, (alpha)' | 0.408 |
| 2 | 'LDHA Lactate dehydrogenase A' | 0.429 |
| 3 | 'High mobility group protein (HMG-I(Y)) gene exons 1-8' | 0.284 |
| 4 | 'PKM2 Pyruvate kinase, muscle' | 0.318 |
| 5 | 'Metallothionein isoform 2' | 0.442 |
| 6 | 'Macrophage migration inhibitory factor (MIF) gene' | 0.299 |
| 7 | 'Triosephosphate Isomerase' | 0.254 |
| 8 | 'Cathepsin B' | 0.371 |
| 9 | 'ALDOA Aldolase A' | 0.236 |
| 10 | '60S RIBOSOMAL PROTEIN L13' | 0.288 |
| 11 | 'GAMMA-INTERFERON-INDUCIBLE PROTEIN IP-30 PRECURSOR' | 0.295 |
| 12 | 'HSPD1 Heat shock 60 kD protein 1 (chaperonin)' | 0.248 |
| 13 | 'PGAM1 Phosphoglycerate mutase 1 (brain)' | 0.229 |
| 14 | '5-aminoimidazole-4-carboxamide-1-beta-D-ribonucleotide transformylase/inosinicase' | 0.124 |
| 15 | 'CTSD Cathepsin D (lysosomal aspartyl protease)' | 0.259 |
| 16 | 'PAGA Proliferation-associated gene A (natural killer-enhancing factor A)' | 0.247 |

Table C.4. Genes Common between all the three methods for B-Cell Lymphoma Data. (Continued).

| | | |
|---|---|---|
| 17 | 'Tubulin, Beta 2' | 0.282 |
| 18 | 'PSAP Sulfated glycoprotein 1' | 0.309 |
| 19 | 'CLTA Clathrin light chain A' | 0.237 |
| 20 | 'LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)' | 0.248 |
| 21 | 'Proteasome activator hPA28 subunit beta' | 0.211 |
| 22 | 'NME1 Non-metastatic cells 1, protein (NM23A) expressed in' | 0.137 |
| 23 | 'SNRPB Small nuclear ribonucleoprotein polypeptides B and B1' | 0.196 |
| 24 | 'APOE Apolipoprotein E' | 0.291 |
| 25 | 'Bcl-2 related (Bfl-1) mRNA' | 0.121 |

Table C.5. Genes Common between all the three methods for Leukemia Data.

| Gene Rank | Gene ID | Score |
|---|---|---|
| 1 | 'CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)' | 0.476 |
| 2 | 'MPO Myeloperoxidase' | 0.404 |
| 3 | 'Azurocidin gene' | 0.388 |
| 4 | 'FTL Ferritin, light polypeptide' | 0.341 |
| 5 | 'GPX1 Glutathione peroxidase 1' | 0.296 |
| 6 | 'DF D component of complement (adipsin)' | 0.293 |
| 7 | 'PRG1 Proteoglycan 1, secretory granule' | 0.285 |
| 8 | 'PROBABLE G PROTEIN-COUPLED RECEPTOR LCR1 HOMOLOG' | 0.260 |
| 9 | 'CTSD Cathepsin D (lysosomal aspartyl protease)' | 0.234 |
| 10 | 'MB-1 gene' | 0.228 |
| 11 | '26-kDa cell surface protein TAPA-1 mRNA' | 0.2224 |
| 12 | 'Terminal transferase mRNA' | 0.221 |
| 13 | 'ALDOA Aldolase A' | 0.214 |
| 14 | 'CCND3 Cyclin D3' | 0.196 |
| 15 | 'PROTEASOME IOTA CHAIN' | 0.184 |
| 16 | 'Immunoglobulin lambda gene locus DNA, clone:123E1' | 0.183 |
| 17 | 'Zyxin' | 0.169 |
| 18 | 'TOP2B Topoisomerase (DNA) II beta (180kD)' | 0.151 |
| 19 | 'C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds' | 0.149 |
| 20 | 'Macmarcks' | 0.143 |
| 21 | 'APLP2 Amyloid beta (A4) precursor-like protein 2' | 0.125 |
| 22 | 'VIL2 Villin 2 (ezrin)' | 0.119 |

Table C.5. Genes Common between all the three methods for Leukemia Data. (Continued).

| | | |
|---|---|---|
| 23 | 'LPAP gene' | 0.116 |
| 24 | 'Oncoprotein 18 (Op18) gene' | 0.103 |
| 25 | 'ATP6C Vacuolar H+ ATPase proton channel subunit' | 0.092 |

Table C.6. ESTs Common between all the three methods for Colon Cancer Data.

| Gene Rank | EST ID | Score |
|---|---|---|
| 1 | "'Hsa.4689 T95018 3" UTR 2a 120032 40S RIBOSOMAL PROTEIN S18 (Homo sapiens) ..." | 0.236 |
| 2 | "'Hsa.8147 M63391 gene 1 "Human desmin gene, complete cds. ..." | 0.199 |
| 3 | "'Hsa.5398 T58861 3" UTR 2a 77563 60S RIBOSOMAL PROTEIN L30E (Kluyveromyces lactis) ..." | 0.168 |
| 4 | "'Hsa.878 T61609 3" UTR 1 78081 LAMININ RECEPTOR (HUMAN);. ..." | 0.168 |
| 5 | "'Hsa.1131 T92451 3" UTR 1 118219 "TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN);. ..." | 0.154 |
| 6 | "'Hsa.539 U14971 gene 1 "Human ribosomal protein S9 mRNA, complete cds. ..." | 0.139 |
| 7 | "'Hsa.8068 T57619 3" UTR 2a 75437 40S RIBOSOMAL PROTEIN S6 (Nicotiana tabacum) ..." | 0.136 |
| 8 | "'Hsa.5444 T48804 3" UTR 1 70269 40S RIBOSOMAL PROTEIN S24 (HUMAN). ..." | 0.119 |
| 9 | "'Hsa.957 M26697 gene 1 "Human nucleolar protein (B23) mRNA, complete cds. ..." | 0.117 |
| 10 | "'Hsa.692 M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. ..." | 0.095 |
| 11 | "'Hsa.692 M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. ..." | 0.095 |
| 12 | "'Hsa.1985 T52185 3" UTR 1 71940 P17074 40S RIBOSOMAL PROTEIN. ..." | 0.095 |
| 13 | "'Hsa.692 M76378 gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6. ..." | 0.092 |
| 14 | "'Hsa.8125 T71025 3" UTR 1 84103 Human (HUMAN);. ..." | 0.091 |
| 15 | "'Hsa.832 T51023 3" UTR 1 75127 HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN). ..." | 0.089 |
| 16 | "'Hsa.831 M22382 gene 1 MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN);. ..." | 0.087 |
| 17 | "'Hsa.678 H55758 3" UTR 1 203413 ALPHA ENOLASE (HUMAN);. ..." | 0.083 |

Table C.6. ESTs Common between all the three methods for Colon Cancer Data. (Continued).

| | | |
|---|---|---|
| 18 | "'Hsa.2800 X55715 gene 1 Human Hums3 mRNA for 40S ribosomal protein s3. ..." | 0.082 |
| 19 | "'Hsa.1832 J02854 gene 1 "MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element..." | 0.080 |
| 20 | "'Hsa.3016 T47377 3" UTR 1 71035 S-100P PROTEIN (HUMAN). ..." | 0.078 |
| 21 | "'Hsa.37937 R87126 3" UTR 2a 197371 "MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus) ..." | 0.062 |
| 22 | "'Hsa.951 M36981 gene 1 "Human putative NDP kinase (nm23-H2S) mRNA, complete cds. ..." | 0.062 |
| 23 | "'Hsa.2588 H40560 3" UTR 1 175410 THIOREDOXIN (HUMAN);. ..." | 0.061 |
| 24 | "'Hsa.36952 H43887 3" UTR 2a 183264 COMPLEMENT FACTOR D PRECURSOR (Homo sapiens) ..." | 0.061 |
| 25 | "'Hsa.3306 X12671 gene 1 Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1. ..." | 0.060 |
| 26 | "'Hsa.1205 R08183 3" UTR 1 127228 "Q04984 10 KD HEAT SHOCK PROTEIN, MITOCHONDRIAL ;. ..." | 0.057 |
| 27 | "'Hsa.773 H40095 3" UTR 1 175181 MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN);. ..." | 0.0566 |
| 28 | "'Hsa.5346 T63370 3" UTR 2a 81523 GUANINE NUCLEOTIDE-BINDING PROTEIN BETA SUBUNIT-LIKE PROTEIN 12.3 (Homo sapiens)..." | 0.054 |
| 29 | "'Hsa.4252 T51529 3" UTR 2a 72384 ELONGATION FACTOR 1-DELTA (Artemia salina) ..." | 0.053 |
| 30 | "'Hsa.33965 H05803 3" UTR 2a 44039 "DIHYDROPRYRIDINE-SENSITIVE L-TYPE, SKELETAL MUSCLE CALCIUM CHANNEL | 0.050 |