APPLICATION OF COMMITTEE NEURAL NETWORKS FOR GENE EXPRESSION BASED LEUKEMIA CLASSIFICATION

A Thesis

Presented to

The Graduate Faculty of the University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Mihir Sewak

May, 2008

APPLICATION OF COMMITTEE NEURAL NETWORKS FOR GENE EXPRESSION BASED LEUKEMIA CLASSIFICATION

Mihir Sewak

Thesis

Approved:

Accepted:

Advisor Dr. Narender P. Reddy	Department Chair Dr. Daniel B. Sheffer
Co-Advisor Dr. Zhong-Hui Duan	Dean of the College Dr. George K. Haritos
Committee Member Dr. Dale H. Mugler	Dean of the Graduate School Dr. George R. Newkome
	Date

ABSTRACT

The present study was an effort to design a Committee Neural Networks-based classification system to subcategorize leukemia cancer data. The need for automated classification arose from the limitations of the traditional techniques which are tedious, time consuming and expensive. In this study, two intelligent systems were designed that classified Leukemia cancer data into its subclasses. The first was a binary classification system that differentiated Acute Lymphoblastic Leukemia from Acute Myeloid Leukemia. The second was a ternary classification system which further considered the subclasses of Acute Lymphoblastic Leukemia. Gene expression profiles of leukemia patients were first subjected to a sequence of preprocessing steps. This resulted in filtering out approximately 95 percent of the genes. The remaining 5 percent of the informative genes were used to train a series of artificial neural networks. These networks were trained using different subsets of the preprocessed data. The networks that produced the best results were further recruited into decision making committees. The training and recruitment procedure enlists the best networks with different background information acting in parallel in a decision making process. The committee neural network systems were later evaluated using data not used in training. The systems correctly predicted the subclasses of Leukemia in 100 percent of the cases for the binary classification system and in more than 97 percent of the cases for the ternary classification system.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Dr. Reddy. He has been a great source of knowledge and inspiration throughout the course of the thesis. His timely suggestions and untiring advices helped to give a direction to my efforts and eventually take them to a successful completion. My experience working with him has been truly memorable; something that will help me for years to come.

Dr. Duan gave me the opportunity to work in the field of Bioinformatics. Her introductory course in Bioinformatics was the turning point behind me deciding to work in this field. I am really grateful to her for devoting so much time and imparting invaluable advices and suggestions.

Dr. Mugler was very supportive and helpful as a committee member. I am very thankful to him for his suggestions during the various 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 Dr. Sheffer and the rest of the faculty of the Biomedical Engineering Department for their support and guidance.

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

TABLE OF CONTENTS

PAGE

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 BIOMEDICAL INFORMATICS

Biomedical Informatics has recently been in the forefront of research and development due to its potential in disease diagnosis, discovery and classification. The change of focus from macro level (organs and tissues) to molecular level has made way for better understanding of gene function and regulation.

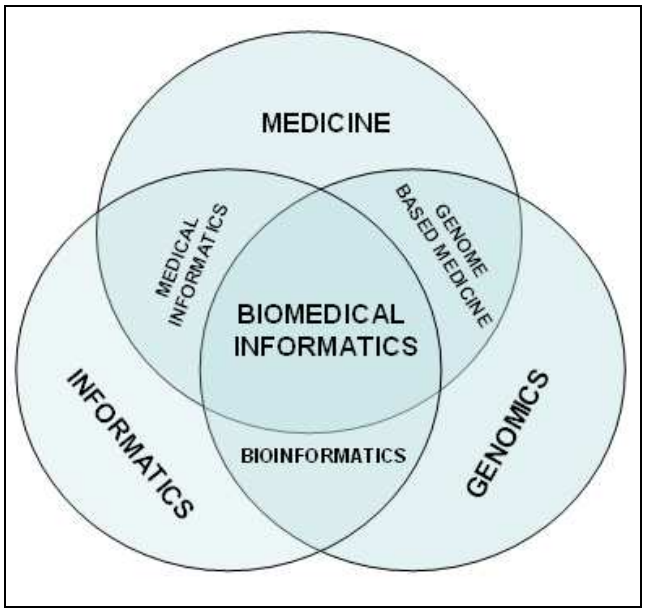

Figure 1.1: Interdisciplinary nature of Biomedical Informatics [1]

The main task of this field has been to integrate different research disciplines such as genomics, proteomics, and clinical research. This integration has helped to develop and share biomedical knowledge related to healthcare [2]. Research on computationally efficient algorithms for data mining, simulation and visualization has also helped in disease discovery and classification efforts. The study of genomics has answered some of the many questions about abnormalities and diseases that have haunted scientists for the past few decades. The basis of various studies has been gene sequencing of different organisms. It has been estimated that the 23 pairs of human chromosomes accommodate around 3 billion base pairs of DNA. These base pairs make up around 40,000 protein coding genes [3]. Answers to most of life's difficult questions like origin of diseases, color of skin, body structure and tendencies etc. are known to be hidden in underlying genetic codes. With the advancement of technology, it has been possible to store this huge amount of data into computer systems. Faster and efficient data access, storage and retrieval have also helped harness the large potential that these information warehouses have had to offer. Data visualization, mining and simulation methods are continuously improving the understanding of biological processes thus accelerating the drug development and design processes [2].

1.2 RECENT DEVELOPMENTS

Public genomic databases have been ever growing with the frequent addition of newer genome sequences. Comparing genomes of different species has helped in identifying control structures and regulatory elements in the gene sequences. A large

number of proteins are being discovered from the various genome projects. Comparative studies of gene expression profiles have aided in identification, classification and monitoring of abnormalities in different organisms. Molecular imaging techniques like Positron Emission Technology and Magnetic Resonance Imaging have enabled the imaging of cell metabolic states which has aided the study of proteins interaction. The understanding of the relation between genotypes, gene regulatory networks and biochemical pathways has vastly helped in combating diseases like obesity, hypertension etc. [4].

The present study involves application of machine learning techniques for classification of gene expression data. A brief theory about genes and genetic processes is an important prerequisite to understand the purpose of the current study.

1.3 UNDERSTANDING GENE EXPRESSIONS

Most of the cells in the body comprise of the same genetic material. However, not all this material has a function in every cell. The function of genes can be defined in terms of its level of “expression” in the corresponding cell. Some genes are turned on or “expressed” in certain cells while the rest of them remain off. Some genes remain dormant for most of the time and are triggered during exceptional circumstances. Research in the field of bioinformatics has shown that the roots to several causes of diseases and abnormalities have been traced to the deviation of gene expressions from the normal levels. Questions such as why a particular organ functions in the way it does, what makes people of certain race susceptible to certain diseases, why exposure to certain

chemicals and substance causes abnormalities in tissues and organs etc. can all be answered to a large extent by analyzing the gene expressions.

1.4 ANALYZING GENE EXPRESSIONS

Microarray technology has made it possible to study thousands of genes at once. Substantial efforts have been made to compare the expression levels of genes in normal and abnormal cells. Deviation of gene expression levels in the abnormal cells from those in normal cells can suggest that the concerned genes are either involved or affected due to that abnormality. Berkum et al. [5] have summarized various studies that have incorporated the DNA Microarrays technology. Parallel monitoring of gene expressions have given a thorough insight into the cells phenotype and internal conditions [5**]**. Quantitative information of gene expression profiles may have a very high potential in diagnosis of diseases, development of drugs and generation of extensive data repositories with information about the processes that govern normal functioning of living cells [6]. These informative elements have already been utilized in developing classifier systems that distinguish normal cells from abnormal ones, classify the abnormal cells into their sub categories or discover sub-classifications of diseases.

1.5 NEED FOR AUTOMATED ANALYSIS

Microarrays have made it possible to monitor expression levels of thousands of genes at the same time. However, studies involving microarray data analysis require only a subset of the available genetic data. Subsets are usually decided on the basis of certain trends and patterns that the expression levels follow. Manually obtaining the optimally

informative subsets from the huge collection of genes is very difficult. From the nature of the data, it is evident that manual techniques are not the solution to explore this data for information. On the contrary, automated applications which carry out these analyses have not yet realized their complete potential. The main reason has been the limitation of computational resources to manage such high amounts of data. Hence, there has been a significant need to automate and optimize gene expression analysis.

In the present study, an effort has been made to classify gene expression data of leukemia patients into its sub-categories. Leukemia is a cancer of the white blood cells. It can be classified into 2 major categories: a) Acute Lymphoblastic Leukemia and b) Acute Myeloid Leukemia. The two are further subcategorized, details of which can be found in the next chapter. Sub-classification of leukemia presents a major problem because the cells of the subclasses show similar morphological appearance but may show contrasting response to drugs and therapy. Traditional methods for sub-classification are tedious, time consuming, and expensive. A need was hence felt to automate the classification of leukemia considering other parameters.

1.6 PREVIOUS WORK

Many automated techniques have been developed for analysis and classification of gene expression data. Golub et al. [7] built a binary classification system in order to automate classification of leukemia into its two sub-classes. They used a class discovery method based on a technique called self-organizing maps to identify a group of 1100 genes that revealed sub-class information for Acute Lymphoblastic Leukemia and Acute

Myeloid Leukemia which occurred above chance levels. Mallick et al. [8] used several Bayesian classification techniques for classification purposes. Their study showed that support vector machine models with multiple shrinkage parameters produced very less misclassification errors. Peng et al. [9] used genetic algorithms and support vector machines in a leave-one-out cross validation test environment. They worked on a two class cancer classification problem and correctly classified all the test cases. Antonov et al. [10] used a maximal margin linear programming method using 132 to 149 features in their approach. Khan et al. [11] used a combination of 3-fold cross validation and Artificial Neural Networks to classify microarray data of small round blue cell tumors into its sub categories. Other authors have also used different approaches to design multi-classifiers for microarray data [12-14]. Alizadeh et al. [15] used the hierarchical clustering approach to identify two molecular distinct forms of diffuse large B-cell lymphoma (DLBCL) which had gene expression patterns at distinct stages of B-Cell maturation. Their study had revealed that microarray data could be used to discover new classes of previously undetected subtypes of cancer.

In all the aforementioned classification studies, the authors have attempted to solve the problem using a specific algorithm. However, machine learning techniques are heuristic in nature. They are aimed at getting an optimal solution as against to the best solution to a problem. It is thus evident that they may not give the best solution to a problem. Confidence and reliability of the final solution can be increased by using an ensemble of techniques [16]. Reddy et al. [17, 20], Palreddy et al. [18], and Das et al. [19] have developed and evaluated the technique of committee neural networks. They

found significant improvement in the prediction performance with committee networks when compared to individual networks. Currently, in the field of machine learning, there has been a growing trend to employ a committee of classifiers for classification purposes. This technique yields a confirmed classification or misclassification with more reliability as against using specific algorithms.

The question remains whether a system of committee neural networks can be designed to improve the sub-classification of gene expression data from leukemia patients. The purpose of this study is to address this question.

1.7 TEST OF HYPOTHESIS

Null Hypothesis:

- Committee Neural Networks cannot correctly classify leukemia using microarray based gene expression parameters.
- The probability p1 of the committee making a correct classification is the same as the probability p2 of the committee making and incorrect classification ($p1 = p2$)

Alternate Hypothesis:

- Committee Neural Networks can correctly classify leukemia using microarray based gene expression parameters.
- The probability p1 of the committee making a correct classification is greater than the probability p2 of the committee making an incorrect classification ($p1 > p2$)

Apriori Significance Level $\alpha = 0.01$

1.8 OBJECTIVES OF THE STUDY

The specific objectives of the study were to:

1. Identify the most informative genes from a collection of gene expression profiles of leukemia patients.
2. Use the identified informative genes to build a series of multi-classifier Neural Network systems that would identify the sub-class of the corresponding sample
3. Recruit the top performing networks into a committee and decide the class based on majority voting.
4. Evaluate the committee of neural networks using a fresh set of data.
5. Perform statistical analysis to test the hypotheses

CHAPTER II

LITERATURE REVIEW

2.1 OVERVIEW OF CANCER

Cancer is known to be a cause of death in about 13 percent of the people every year [21]. There are different factors attributed to cancer, some of them being exposure to radiation and carcinogenic substances like tobacco smoke, etc. The most important factors however, have been the abnormalities in the gene pool of the affected. Most of the cancer research is carried out based on the underlying assumption that mutations in the cancer genes and the tumor suppressing genes are responsible for the aggressive replication of cells. In their research paper, Hanahan et al [22] summarized the six different kinds of alterations that if taken place within the cell physiology, lead to malignant growth.

2.1.1. Elimination of Dependence of Growth Signals on Other Factors

Growth signals are required to revive dormant cells and get them into a multiplicative state. These signals originate in the transmembrane organelles namely:

i. Diffusible growth factors

ii. Extracellular matrix components

iii. Cell to cell adhesion/interaction molecules

These signals are received by transmembrane receptors which when stimulated, bring about cell proliferation. Tumor cells however, learn to mimic these growth signals. This reduces their dependence on external growth signals and hence replicate uncontrollably.

2.1.2. Developed Resistance / Immunity towards Growth-Inhibitory Signals

These signals are emitted by soluble growth inhibitors and immobilized inhibitors. They are present in the extracellular matrices and on the surfaces of the nearby cells. These signals counteract to the Growth Signals and inhibit cell proliferation. The inhibition takes place either by moving cells from the active proliferation state to a dormant state, or permanently crippling the proliferation mechanism of the cell. Tumor cells learn to evade these extraneous signals thus aiding their replication.

2.1.3. Evolution of Properties that Disrupt the Mechanism of Apoptosis

Apoptosis is a mechanism by which cells kill themselves programmatically. This functionality is present in latent form in all cells and can be triggered by different physiological signals. The mechanism works in a sequence of steps:

i. Cell membranes are disrupted

ii. Cytoplasmic and nuclear frameworks are disrupted

iii. Cytosol is extruded

iv. Chromosomes are degraded

v. Nucleus is fragmented

The whole process takes about 30 to 120 minutes. Later, the remains of the dead cells are consumed by nearby cells. Cancer cells influence this mechanism in two ways. Certain over expressed oncogenes learn to trigger apoptosis in normal cells thereby reducing their number. These oncogenes can disrupt the defense mechanisms of the body if they are over expressed in related cells. The other way is that the cancer cells themselves learn to evade apoptosis by suppressing expression of certain proapoptotic regulators. This premature activation of apoptosis in normal cells in combination with suppression of the mechanism in tumor cells, leads to limitless growth of tumor cells.

2.1.4. Aggressive Cell Replication Potential

Along with the aforementioned capabilities, cells carry built-in autonomous programs that limit multiplication. This program operates independent of the discussed signaling pathways. The limit to replication can be defined in terms of the 50-100 base pair loss of telomeric DNA's (ends of chromosomes which comprise of several thousand repeats of 6 base pair elements) at the ends of each replication cycle. The gradual loss of telomeric DNA's after each cell cycle finally leads to death of cell thus imposing a hard limit to cell replication. Some malignant cells can counteract this mechanism by over expressing the telomerase enzyme and maintaining the telomeres. Other malignant cells maintain the telomeres through recombination based inter-chromosomal exchanges of sequences. The described two mechanisms lead to maintaining the telomeres after cell cycles. This leads to limitless cell replication in cancer cells.

2.1.5. Development of Angiogenic Capabilities by Neoplasm

The process of the growth of blood vessels is termed as angiogenesis. Cells derive their nutrition from the tissue vasculature. Cells in tissues reside within 100 µm of capillary blood vessels. Angiogenesis is a transitory and controlled process. This capability is not well defined for newly born cells. However as the tissues expand, the ability is developed by neoplasias. The Growth and Inhibitory signals help control angiogenesis. Tumor cells develop the ability to induce and sustain angiogenesis by creating an imbalance between the Angiogenesis inducers and inhibitors. This is made possible via over expression and under expression of genes responsible for Angiogenesis control. In this manner tumor cells can block nutrition sources for normal cells and induce development of capillaries and blood vessels for their own nutrition.

2.1.6. Invasion of Nearby Tissues and Distant Settlement of Tumor Cells

During the development of cancer, the primary tumors give rise to cells that migrate to nearby tissues. They compete with the normal tissue cells for nutrition and proliferate by forming new colonies. At the same time they bring about reduction in the normal cell count. This process of cells moving out into adjacent tissues is termed as Metastasis and is the cause of 90 percent of human cancer deaths. Successful tissue invasion and metastasis depends on all the aforementioned capabilities. Progressive genetic changes empower cells with a series of growth advantages and capabilities. The cells also learn to manipulate the cell death mechanism (apoptosis). This results in transforming mortal human cells into cancer cells. Figure 2.1 presents the six alterations that cause cancer.

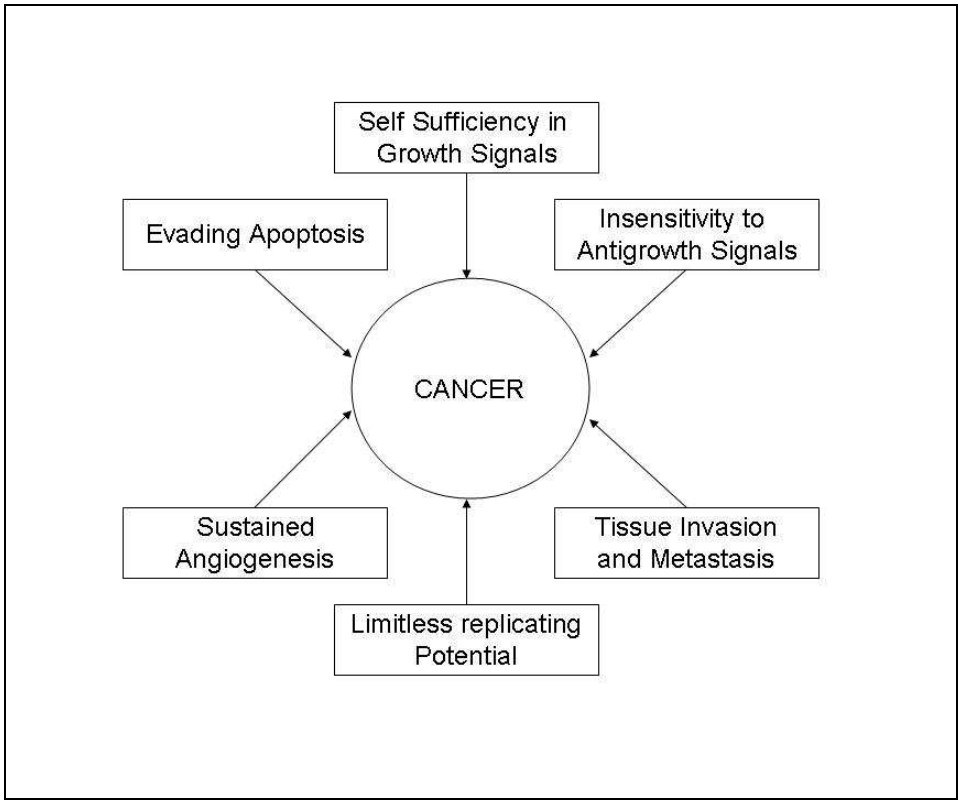

Figure 2.1: Alterations in the cells that are responsible for cancer (Source [22])

2.2 LEUKEMIA

Cancer has been known to affect many body organs and tissues, some of them being bones, stomach, lungs and blood. The cancer of the blood is called Leukemia. It is a subtype of a broad array of diseases commonly referenced as Hematological Malignancies. According to statistics provided by the Leukemia and Lymphoma Society, this disease is expected to be diagnosed in more than 44000 people in the United States this year, and it has one of the top mortality rates among different types of cancer. [23]

"Leukemia is the general name for the four different types of blood cancers." - [24]

The four different types of blood cancers that are covered under this general term:

1. Acute Myelogenous Leukemia (AML)
2. Acute Lymphocytic Leukemia (ALL)
3. Chronic Myelogenous Leukemia (CML)
4. Chronic Lymphocytic Leukemia (CLL)

If the disease occurs in the Lymphocyte forming marrow cells then it is called Lymphocytic or Lymphoblastic. If the disease occurs in the bone marrow cells that form the Red Blood Cells, White Blood Cells or platelets then the term Myelogenous or Myeloid is employed. Acute Leukemia is differentiated from Chronic Leukemia such that the former is composed of an immature blood forming cells called Lymphoblasts or Myeloblasts whereas the latter shows scarce to no presence of these blast cells. Acute Leukemia has an aggressive growth rate if not treated at the appropriate time whereas Chronic Leukemia has a slow growth rate, in comparison.

Lymphoblastic Leukemia is further subdivided into to categories:

- T-Cell Acute Lymphoblastic Leukemia

This type of leukemia is seen in 15 percent of Lymphoblastic Leukemia patients. It affects the lymphocytes that mature in the thymus and hampers their primary function, which is to mediate cellular immune responses. It is characterized by the enlargement of the thymus which causes difficulty in breathing for the patient. It is also associated with the premature diffusion of cerebrospinal fluid. [23]

- B-Cell Acute Lymphoblastic Leukemia

This is the common type of leukemia in children. This type of leukemia constitutes about 85 percent of all cases. B-Cells are responsible for manufacture of antibodies. It is characterized by the presence of immature B-Cells all through the body, lymph nodes, spleen, and liver. [23]

The origin of the four types of leukemia can be traced to the bone marrow, where the different blood cells have their origin. The bone marrow produces 3 different types of blood cells:

i. White Blood Cells (WBC): These cells fight infections. They form the defense system of the body against various diseases
ii. Red Blood Cells (RBC): These carry nutrition in the form of oxygen to tissues all through the body
iii. Platelets: These are smaller of the 3 blood cell types. They have cementing property and help in forming blood clots to control bleeding.

In leukemia, the bone marrow produces abnormal/mutant white blood cells. The leukemia cells acquire the alterations previously described. They learn to compete against the normal cells for the purpose of survival. They crowd out the normal blood cells gradually. As a result the body loses immunity due to loss of disease fighting white blood cells. Reduction in RBCs leads to scarcity of nutrition for organs and tissues through which the blood passes. Reduction of platelets leads to inability of clot formation in times of injury and hence may lead to blood loss.

2.2.1 Causes of leukemia

Several factors have been attributed to the cause of leukemia. Very extensive exposure to radiation has been studied as a factor that causes leukemia. The types of radiation involve exposure to nuclear wastes, nuclear accidents, some medical treatments that use high radiation, etc. Studies have shown that exposure to certain chemicals like benzene, formaldehyde can cause leukemia [25]. Alkylating agents that are used in chemotherapy have shown to cause leukemia in cancer patients after some years. Certain diseases both genetic and organ related have known to increase risk of leukemia in the later ages (e.g. Down syndrome, Myelodysplastic syndrome etc.). It was also believed that exposure to strong electromagnetic fields increases the risk of leukemia.

2.2.2 Diagnosis of Leukemia

Diagnosis of leukemia is carried out using a combination of different techniques. The tests involved calculating total percentages of leukocytes, erythrocytes and platelets in the total composition of the blood sample. Typically leukemia patients show abnormally high levels of leukocytes and abnormally low levels of erythrocytes, platelets and hemoglobin. It is also determining whether there is a presence of swellings of the Lymph Nodes, Spleen and Liver, Kidneys and Thymus. A series of blood tests called “Complete Blood Test (CBC)” are always carried out. These involve tests of the blood, bone marrow and additional tissues. The cell shapes are studied using morphological analysis to determine abnormalities in cell shapes and sizes. The erythrocyte abnormalities may show size differences (anisocytosis), increase of average cell size

(macrocytosis) and abnormal cell shapes. This analysis also detects giant sized platelets which lack granules. Study of the patient's bone marrow can help to explain the origin of leukemia. Bone marrow can be obtained by two ways:

1. Bone marrow Biopsy

   A combination of bone along with bone marrow are removed and studied

2. Bone Marrow Aspiration

   Only the a part of the bone marrow is extracted for study, using a fine needle

Tests are also conducted to measure the serum levels of calcium, potassium, phosphate and uric acid in the blood. Along with the CBC, there are cytogenetic techniques that have been put to use recently. These techniques involve chromosomal analysis in combination with Fluorescence in situ Hybridization and Immunophenotyping. Molecular techniques are used as supportive measures for diagnosis confirmation (e.g. Polymerase Chain Reaction).

Fluorescence in Situ Hybridization (FISH):

In this technique, the DNA sequences are labeled with fluorescent dyes. These are then detected with immunofluorescent staining. Numerous fluorescent labeled DNA probes are being manufactured, thus adding refinements to the FISH proto-cols. These refinements have increased the importance of the FISH technique in clinical applications. FISH with unique DNA sequences is one of the first molecular cytogenetic techniques. Here, the DNA segment used as a probe represents a functional gene. This technique is commonly used for detection of microdeletion syndromes and gene fusions,

rearrangements and other mutations in cancer cells. Quantification of the FISH maps is an automated process and the intensity of fluorescence can be detected using computational techniques. However the banding patterns on bent chromosomes is difficult to detect and highly experienced resources are required for this process [27].

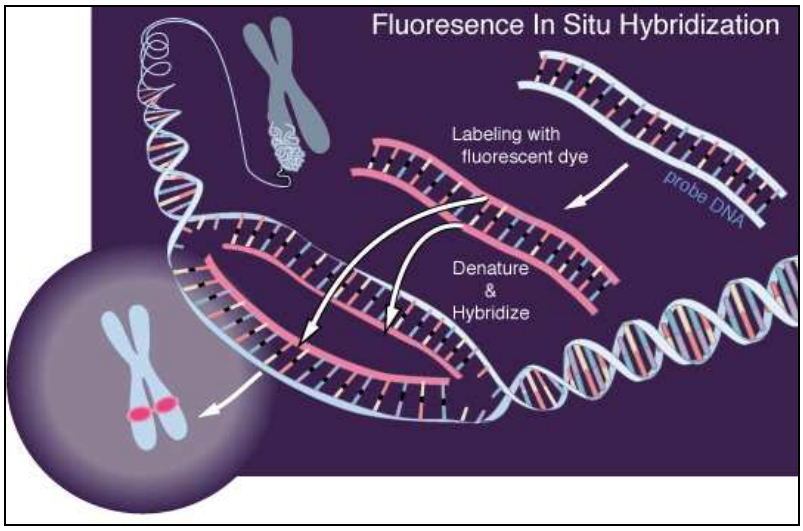

Figure 2.2: A pictorial representation of FISH (Figure reproduced from the website http://www.genome.gov)

Immunophenotyping:

In this technique, an assortment of cells, are studied in order to characterize and compare them according to the populations of interest. This technique is extensively used in the diagnosis of leukemia and Lymphoma. The samples of cells used for this study are leukocytes from peripheral blood obtained in fine needle aspirates (FNA). The leukocytes are labeled with antibodies which bind to surface proteins. Flow Cytometry is employed to detect the labeled cells [27].

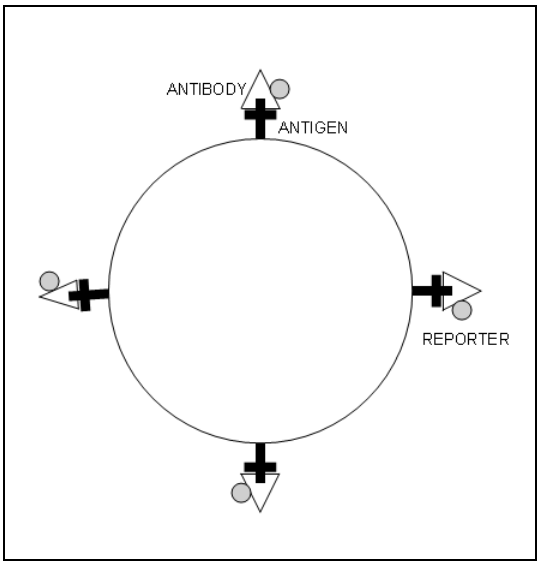

Figure 2.3: Antibodies to surface antigens

Diagnosis of leukemia is a complicated procedure. Some of the symptoms of leukemia can be confused with less trivial conditions. Early diagnosis of leukemia is important from the clinical point of view. It is important that the treatment starts immediately to prevent its spread and hence ease the curing process. As discussed there are a series of diagnostic techniques that have to be utilized before leukemia can be confirmed and sub-classified. The elaborate cytogenetic testing can be a time consuming and an expensive process [29]. It is not very easy to standardize the technique due to variation of results from lab to lab. The process takes around five days before a leukemia subgroup can be confirmed. This delay in confirmation along with the variations of data returned from individual laboratories may lead to a wrong treatment protocol in a time critical situation for the patient. Thus, there is a need to standardize the diagnosis protocol for sub-classification purposes. The growing need for standardization has led to

the emergence of gene expression profiling based sub-classification by making use of Microarray Technology.

2.3 INTRODUCTION TO MICROARRAYS

Microarrays are repositories providing extensive data about genomes, in the form of simultaneous monitoring of their expression levels. McLachlan et al. [30] have extensively studied the analysis techniques of microarray gene expression data. The following theory on microarrays has been derived using their work as reference.

2.3.1 History of Microarrays

The microarray technology was evolved out of the need to determine amounts of particular substances within a mixture of different substances. The process was first carried out using assays. Assays were used in a variety of applications like identification of blood proteins, drug screening etc. Immunoassays helped to determine the amount of antibodies bound to antigens in immunologic response processes. Fluorescent labeling and radioactive labeling were used to label either the antibodies or the antigens to which the antibodies bound. The concept of immunoassays was later extended to involve DNA analysis. The earliest microarrays involved assays in which the samples were spotted manually on test surfaces. The smallest achieved spot sizes were 300 µm. However, it was only when the spot sizes became smaller for accommodating more genes, that robotic and imaging equipments were deployed. Labeling methods involved using radioactive labels for known sequences. Another technique involved using fluorescent dyes to label biological molecules in sample solutions. The Southern Block technique,

developed later, used arrays of genetic material for the first time. The technique involved labeling of DNA and RNA strands to probe other nucleic acid molecules attached to solid surfaces. In this technique, denatured strands of DNA were transferred to nitrocellulose filters for detection by hybridization to radioactively labeled probes. Such a transfer was possible because denatured DNA strands formed covalent bonds with solid surfaces without reassociating with each other. These however easily formed bonds with complementary segments of RNA. Southern Block technique began by using porous surfaces as the solid support for DNA strands. These were later replaced by glass surfaces which sped up the chemical reactions since the substances did not diffuse into porous surfaces. In 1980 the Department of Endocrinology at the University of London used microspotting techniques to build arrays for high sensitivity immunoassay studies involving analysis of antibodies in the field of immunodiagnostics. This technique was later adopted in a wide range of applications involving biologically binding assays. The product of this technique, known as multianalyte microspot immunoassays, measured radiometric intensity by taking the ratio of the fluorescent signals to the absorbance. The technology has been evolving ever since and a lot of research has been accomplished to refine this technique. The first DNA Microarray chip was engineered at Stanford University, whereas Affymetrix Inc. was the first to create the patented DNA microarray wafer chip called the Gene Chip.

Figure 2.4: GeneChip (Source – http://www.biotaq.com/Images/genechip.jpg)

2.3.2 Types of Microarrays

There are two types of Microarrays:

- Single Channel or One color Microarrays

This technology was first introduced by Affymetrix Inc. In these microarrays, only one sample undergoes hybridization after it is labeled with a fluorescent dye. These microarrays measure the absolute intensity of expression. These are also called oligonucleotide microarrays where the probes are oligonucleotide sequences that are 15 to 70 base pairs in length. Oligonucleotides are either, synthesized separately and spotted on the chips, or they can be synthesized directly on the chip using in silico techniques. The latter technique is carried out using a process called photolithography. Experiments involving one color microarrays are characterized by simplicity and flexibility. Hybridization of a single sample per microarray not only helps to compare between microarrays but also allows analysis between groups of samples. The data used in this study was obtained using microarrays of this type.

- Dual Channel or Two color Microarrays:

These are also termed as cDNA Microarrays. In these microarrays, sample sequences and normal sequences are labeled with two different fluorescent dyes. Both these DNA sequences are hybridized together on the DNA Microarrays and a ratio of fluorescence intensities emitted by the two dyes is considered in order to evaluate differential expression level. This design of microarrays was developed to reduce variability error in microarray manufacturing. Hybridization of two samples to probes on same microarray allows for direct comparison. These microarrays are known to be highly sensitive and accurate.

2.3.3 Components of Microarray Technology

A Microarray comprises of the following components:

- The Array:

This is the solid base on which the genetic material of known sequences are arranged systematically along grids. The process of arrangement is called spotting. The array is made up of glass or nylon which bears thousands of wells to hold the different Complementary DNA (cDNA) sequences. Each spot on the microarray represents an independent experimental assay used to measure the presence and abundance of specific sequences in the sample strands of polynucleotides. The Arrays are made up of glass, nylon and sometimes coated using silicon hydride. The coating enables the microarrays to repel water and encourages hybridization of cDNA strands to the surface of the array. It also prevents the polynucleotide samples from spreading thus keeping noise in check.

- Probes:

The single stranded cDNAs that are spotted on the arrays are known as “probes”. The target polynucleotide sequences in the biological sample solutions are hybridized with the complimentary probes. Adherence of probe to the array is very crucial to maintain spot integrity and prevention of the probe from being washed away during array processing. It is also crucial as an irregularly adhered probe can cause noise to seep in thus reducing the quality of resultant image. After the probe is spotted onto the array, it is air dried and exposed to ultraviolet radiation to ensure immobility and strong adherence.

- The Spotter:

These are robotic instruments that apply the probes to the arrays using high precision. The spotters apply each spot to a grid on the array. This helps in conducting a large number of experiments in parallel. The spotting is done using either contact or non-contact techniques. Contact spotters have the spotting nozzle similar to an ink pen where applied pressure releases the probes on the arrays. Non-contact spotters use ink-jet technology or the piezoelectric capillary effects for spotting purposes. Non-contact spotters are faster than contact spotters; however contact spotters have more precision as compared to non-contacting ones.

2.3.4 Procedure of using cDNA Microarrays

Microarrays can be used to detect presence of polynucleotides of unknown sequences in the target sample. The polynucleotides of unknown sequences are first labeled with fluorescent dyes like fluorescein, rhodamine etc during reverse transcription.

Specific dye color is used for polynucleotides of control target sample and a different dye color is used for polynucleotides of unknown target samples. The different colored dyes emit varying wavelengths based on the mixture of the known and unknown samples. The scanning and imaging equipment then detect the varying intensities of fluorescence. The intensity information is later used to detect the variation of hybridization of unknown target samples from the control samples. Figure 2.5 describes the complete process.

2.3.5 Molecular Hybridization

This is defined as the association of denatured DNA strands to their complimentary strands via specific base-pair bonding. Hybridization occurs between labeled denatured DNAs of target samples and the cDNA strands of known sequences on the spots of the array. Thus, if the target sample contains a base sequence which is complimentary to a particular probe on the array, it will hybridize at that spot. Also, because the target strands are fluorescently labeled, the emitted fluorescence reveals the location of the spot on the array. If many target strands hybridize at one spot on the array, this spot will emit fluorescence of higher intensity. To summarize, the intensity of fluorescence is a measure of the amount of hybridization on the spots of the microarray.

The information can be used to detect over expressed and under expressed genes in the target sample. The fluorescence is dependent on several factors some of which are:

- Temperature
- Hybridization time
- Processing Time

- Relative Humidity
- Presence of excess sample

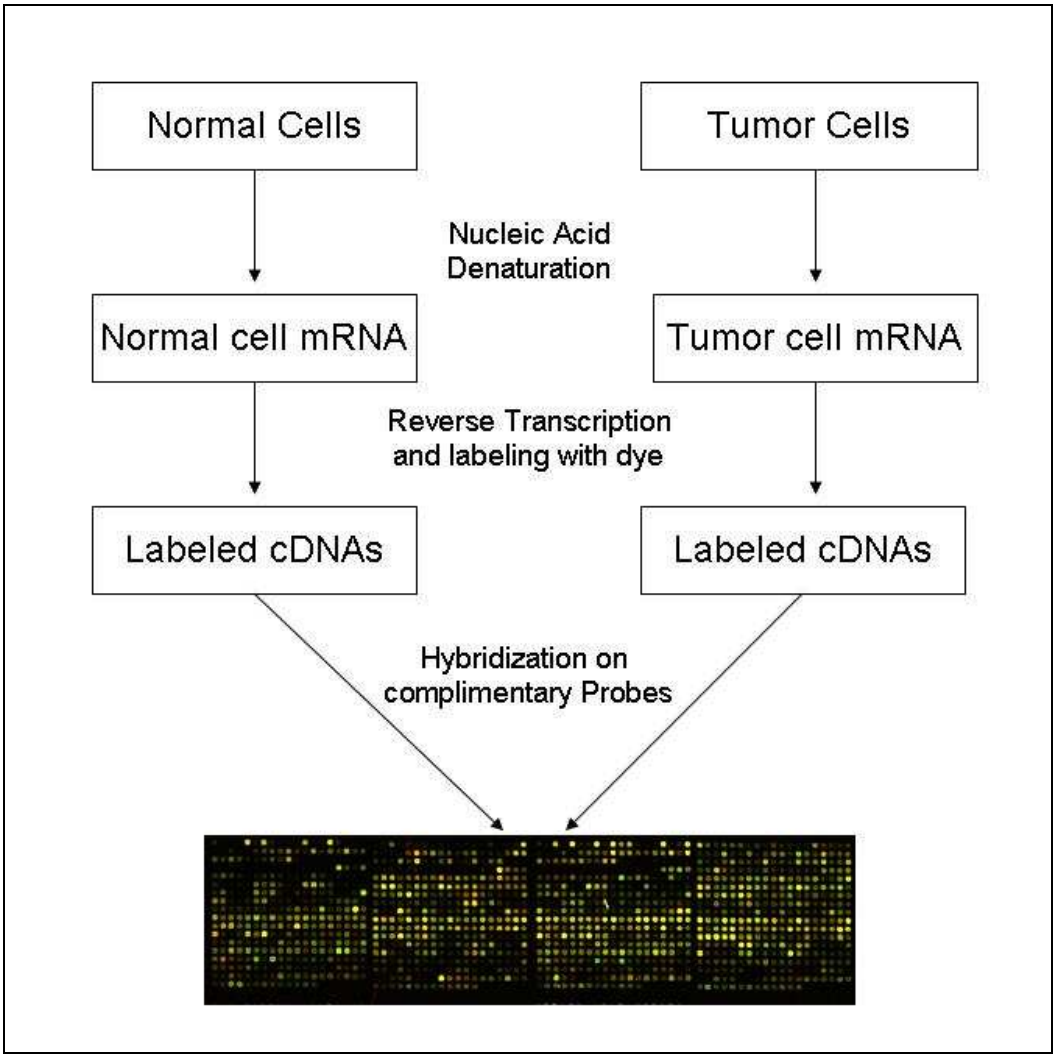

Figure 2.5: Using a cDNA Microarray

Before the chip is sent to scanning, most of these factors are neutralized. Hybridization and processing time is decided apriori. Excess sample present on the chip is washed away. Finally the humidity and moisture is eliminated by air drying the chip. The chip is then sent for scanning and imaging.

2.3.6 Imaging and Image Processing

Scanners usually make use of lasers of specific frequency in the Ultraviolet Region. They excite the fluorescent dye attached to the samples that have hybridized on the complimentary probes. The photons emitted by the excited dyes are collected at the detectors and the intensity levels are recorded. Since the emitted intensity is the result of two different fluorescent labels (known sample and unknown target), the slide is scanned at 2 different wavelengths. The ratio of the intensity levels is the indirect measurement of the relative gene expression level. The microarray image is then scanned and digitized. The fluorescent intensities, after digitization are viewed on a computer screen in the form of translated pixel intensities. The intensities are later analyzed depending on the intensity levels. Thus, if the known sample was tagged with a red fluorescent dye, then a red spot on the image indicates that there was abundant hybridization of a known sample at the cDNA spot at that location. If the unknown target was tagged with a green fluorescent dye, then a green spot on the image indicates that there was abundant hybridization of unknown target at the cDNA spot at that location. A yellow spot would indicate that there was an equal amount of hybridization of the known sample and the unknown target at the spot of that location. Majority of the spots on the microarray would have intensity levels of different levels which correlate to the various shades of red and green. These intensities values are the informative elements in a microarray experiment.

2.3.7 Limitations of Microarray Technology

Microarray experiments depend on availability of tissue samples in sufficient

quantity. Accuracy of the resulting data depends on the quality of the RNA and cDNA samples. This quality is decided by the purity of the polynucleotides, the technology used to carry out the spotting and the experimental protocol. The efficiency of the reverse transcription of mRNA also creates a bias in the results expected out of a microarray experiment (Reverse transcription bias). The fluorescent dyes used, have varying binding affinities to the nucleotides. They lead to what is known as a sequential bias in the results. A plot of fluorescence against time suggests a non-linear relationship beyond some point. Study of gene expression levels is one way of determining activities in a cell. There are several factors which determine the expression levels. Some of these factors are faster production rate, modification due to certain chemical activities or degradation by certain proteins. Hybridization of genes is sensitive to ionic strength of the solution and temperature. Different genes have varying tolerance levels to these parameters. It is nearly impossible to have a specific experimental condition that suits all the genes. As a result some of the genes go undetected because they fail to undergo the intended hybridization. Probe attachment and cDNA hybridization are susceptible to miscellaneous errors. These errors produce variability in the results. After hybridization, the microarray is washed and dried to get rid of non-hybridized fragments. This is however not a thorough process and some non-hybridized strands persist by adhering to the glass slide. Non-hybridized target molecules are not completely washed away. The ones that persist on the microarray in a non-hybridized state also emit fluorescence. The emitted fluorescence is detected during the scanning phase and adds to the background noise. Presence of noise makes image processing more challenging. There are other

factors which also constitute to noise. Presence of foreign particles on the microarray chip such as dust, fibers etc can result in noise. Low concentration of the fluorescence labeling dye, inadequate hybridization and varying times of exposure can lead to variation in captured signal at the different probes. The mentioned factors make image processing a challenging process. Use of Microarray Data for analysis purposes can be a challenge. The data needs to undergo a sequence of transformations depending on the nature and scope of the involved study. As a result standardization of the obtained data can be a significant issue as well as a limiting factor in the involved study.

2.4 ARTIFICIAL NEURAL NETWORKS

The concept of an Artificial Neural Network (ANN) was inspired by the information processing capabilities of the biological neural networks.

2.4.1 History of Neural Networks

The concept of neural networks dates back to 1943 when a neurophysiologist Warren McMulloch and a mathematician Walter Pitts associated in a quest to determine how biological neural networks worked. They modeled the first modern neural network using simple electric circuits. In 1949, Donald Hebb reinforced their theory in his book titled “The Organization of Behavior”. He proved that neural pathways grew stronger each time that they were used. With the invention of computers in the 1950s, simple models which replicated the human thoughts and learning process began to spawn. Nathanial Rochester of IBM research labs was one of the first researchers to simulate a

neural network. During this time most of the research was concentrated in computer technology and this sidelined the efforts of neural networks research. However there were certain individuals who persistently dug deeper in the wide ocean that was machine learning. They continued making theoretical advancements and refining the definition of the artificial neural networks in the process. Marvin Minsky, in 1954, worked on his doctorate thesis, "Theory of Neural-Analog Reinforcement Systems and its Application to the Brain-Model Problem". He was also one of the premiers in discussing details of Artificial Intelligence in his publication entitled, "Steps towards Artificial Intelligence". This publication discussed in large detail, the modern neural networks architectures. Extensive research on Artificial Intelligence with an emphasis on Neural Networks was initiated at the Dartmouth Summer Research project in 1956. Bernard Wildrow and Marcian Hoff of Stanford University built one of the first models of artificial neural networks. ADALINE (Adaptive Linear Elements) and MADALINE (Multiple Adaptive Linear Elements) as they were called could adapt to real life learning problems. At around the same time, Frank Rosenblatt, a neurobiologist from Cornell University began working on the model of a perceptron. This model was simulated using electrical circuits and is still in use today. The simple model of a perceptron was used in one of the first machine learning based binary classification systems. It successfully classified a set of continuous valued inputs into two classes. In 1962, he trained the neural networks by adding weight parameters to the synapses. The weights were modified every time the network gave a misclassification. Thus the neural networks “adapted” to the input training sample so as to reduce classification error. The technology in its infancy was not

thoroughly accurate in terms of training the hidden neurons. The problem was solved by Paul Werbos in 1974, when he first introduced back-propagation networks. The algorithm that he developed allowed neurons to distribute the error compensating values to the hidden layers. Around 1982, John Hopfield of Caltech University presented a paper at the National Academy of Sciences. The paper was concerning Neural Networks application to simple useful devices. The neural network model that he came up with was asynchronous in nature that adjusted itself to find the minimum error. This highly efficient model thus used the least energy and functioned to a large extent like the human brain by using content addressing. This neural network model is very popular even today and is popularly known as the Hopfield Networks.

Neural Networks have shown promising results in a multitude of fields. As the "Data & Analysis Center for Software" puts it: *"The promise of Neural Networks seems bright as nature itself is living proof that this kind of thing works"*. Although software and simulations have yielded excellent results of Neural Network applications to different fields, the future of neural networks lies in hardware development.

2.4.2 Components of a Neural Network

A typical neural network consists of:

1. One input layer of neurons
2. One or more hidden layers of neurons
3. One output layer

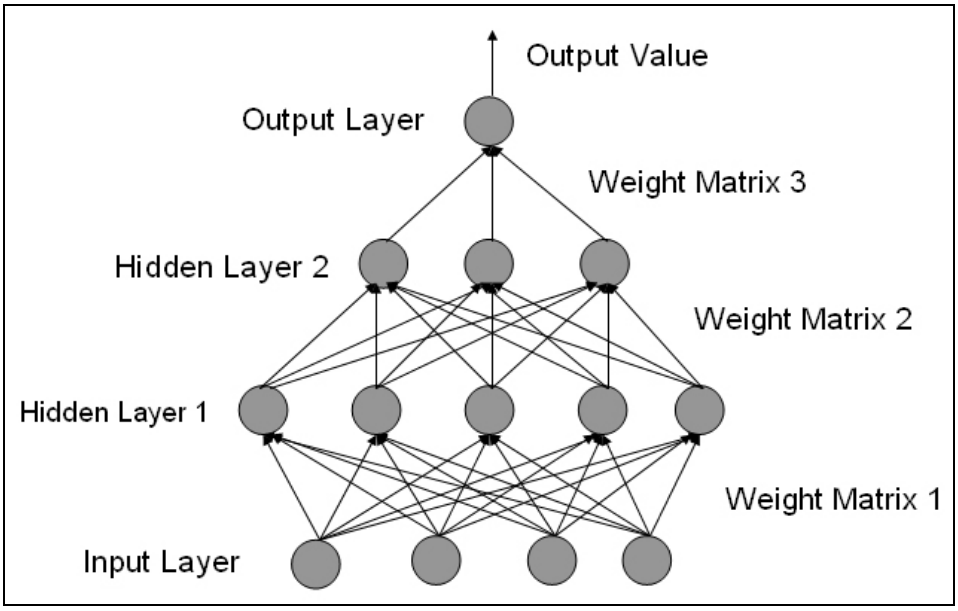

Figure 2.6: Components of a Neural Network

The neurons in each layer are connected to the neurons in the subsequent layer through synapses. In a fully connected neural network, every neuron from one layer is connected to every neuron from the subsequent layer. In a feed forward neural network the information is passed on from layer to layer until it reaches the output layer where the classification takes place. In a feedback neural network, information travels in both directions. The input parameters of a neural network may not be of equal significance from the point of view of classification. Certain parameters are more crucial in determining the output than some other parameters. To accommodate this bias, the networks are provided with a system of weights with each connection. Significant parameters are highly weighted and dominate over lesser significant parameters. The output of each neuron is a weighted sum of its inputs. This output is usually associated with a transfer function that scales it into a particular range.

2.4.3 Transfer Functions

The output of each node is usually transformed to a particular range using a transfer function. Usually the transformed output the weighted inputs at each node are added. This summation is later compared with some predefined threshold. Signal from the node is generated if and only if the weighted sum is greater than the threshold. The outputs of transfer functions can be linear or non- linear in nature. However transfer functions with linear responses are limited in application because the output is linearly proportional to the weighted inputs. Various types of transfer functions are used by researchers depending on application:

- Hard-Limit Transfer Functions:

This transfer function yields a 0 if the weighted sum is less than 0 and a 1 if the weighted sum is greater than 0

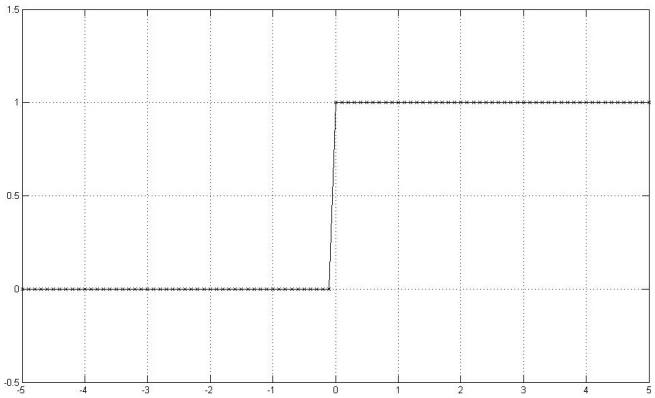

Figure 2.7: Hard Limit Transfer Function

- Hyperbolic Tangent Sigmoid transfer function:

This transfer function yields an output scaled between -1 and 1. The output is calculated according to the following equation:

$$n = \frac{2}{(1+\exp(-2*n))} - 1$$

This function when plotted for continuously varying values of n between -5 and 5 yields the following output:

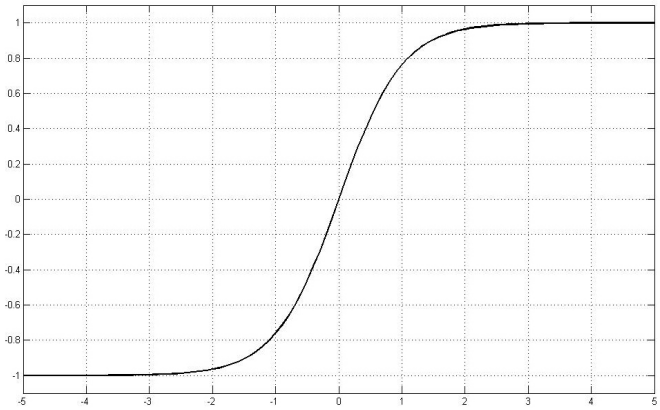

Figure 2.8: Hyperbolic Tangent Sigmoid transfer function

- Log Sigmoid transfer function:

This transfer function yields and output scaled between 0 and 1. The output is calculated according to the following equation:

$$\log sig(n) = \frac{1}{(1+\exp(-n))}$$

This function when plotted for continuously varying values of n between -5 and 5 yields the following output:

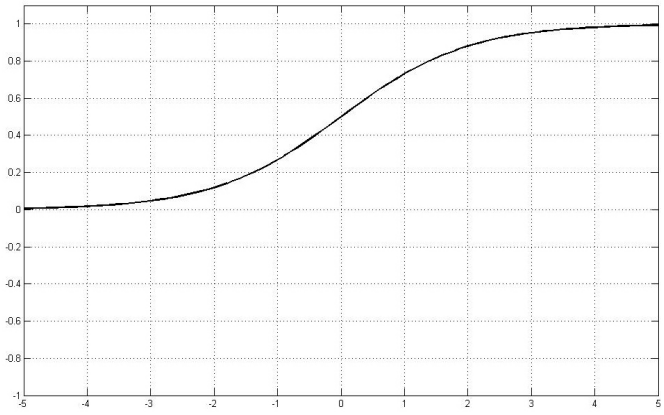

Figure 2.9: Log Sigmoid transfer function

2.4.4 Output Functions

The output of each processing element is usually decided by the transfer functions. There are times when the outputs of the processing elements are modified to consider strong or weak input parameters. This competition allows the outputs, from elements with greater strength, and masks the outputs of processing elements with less strength.

2.4.5 Error Functions

When learning is carried out, the obtained output is compared to the expected output. The difference between these two values constitutes the error. This error value is then transformed to match the network architecture using specific error functions. The

transformed error values are then back propagated to a previous layer during the weight adaptation phase of the learning process.

2.4.6 Learning Functions

These functions are triggered during the weight adaptation phase of the learning process. They modify the weights on the input side of each processing element based on certain optimization algorithms.

Some of the learning functions are:

- Learngd (Gradient descent weight and bias learning function): This function calculates the change of weight Δw for a given processing element from the input i, the error e and the learning rate lr according to the gradient descent function:

$$\Delta w = lr*gw$$

- Learngdm (Gradient descent with momentum weight and bias learning function): This function calculates the change of weight Δw for a given processing element from the input i, the error p, the weight w, learning rate lr, and the momentum constant mc. Thus:

$$\Delta w = mc* \Delta w_{prev} + (1-mc)*lr*gw$$

Δw_{prev} is the previous weight which is obtained from the previous learning state.

2.4.7 Training a Network

Training a neural network uses algorithmic methods based on optimization for network training. The most popular algorithm that has been used by researchers has been

the back-propagation algorithm. In a back-propagation algorithm, the gradient vectors of surface errors are first calculated. This vector is directed along the line of steepest descent. Traversing along this line will lead to achieving the error goal. The algorithm progresses iteratively through a sequence of steps called epochs. At every step, a new training sample is introduced and the output is obtained. The difference between the obtained and expected output is calculated. This error value and the error surface gradient are propagated backwards towards the previous layers and the weights are adjusted so as to minimize classification error. The initial network configuration is random. Training can be stopped by presetting the maximum number of epochs, presetting the target error goal to be achieved or by presetting a time limit for the training cycles. The leaning methods are broadly classified into 2 categories, supervised learning and unsupervised learning.

1. Supervised Learning

In Supervised Learning, the neural network is trained by providing both, the input vectors and the expected outputs. The obtained output is compared with the expected outputs and the error value is calculated. The objective of supervised learning is to reduce the difference between obtained output and expected output by adjusting the weights and the connections between layers in an algorithmic fashion. This type of learning is used in prediction and classification problems.

2. Unsupervised Learning

Networks that utilize unsupervised learning are also called self organizing networks. In

case of these networks the output vector is not provided. The objective of these networks is to use generalization principles for data organization. These networks do not require training and are used to identify groups or clusters from an input data set.

2.4.8 Applications of Neural Networks

Neural networks have been widely used in the following fields:

- Natural Language Processing based applications like text-to-speech conversion, voice commands feature on mobile phones and other equipment, and automatic language translation systems have used neural network based architectures
- Signal and Image processing has benefited from neural networks. Patterns in image data can be learned using unsupervised techniques and redundant and uninformative bits in a signal can be identified. The trend can then be generalized and incorporated in signal compression algorithms.
- Pattern Recognition and character recognition has been made possible using Neural Network systems. These systems can recognize handwriting; validate signatures and finger prints etc.
- Marketing strategies can be improved by training networks with current and past data of market trends. These networks can then be used for validation and estimation purposes thus helping to minimize risky market decisions.
- Defense and security departments have been using Neural Network based systems to track targets, facial recognition and signal identification
- The Gaming industry has incorporated neural networks into their software. This has

given rise to a series of intelligent games that learn from experience and adapt to situations thus delivering to the end user a whole new gaming experience.

- Robotics has benefited from neural networks especially in the fields of machine vision, trajectory control etc.

Daliakopoulos et al. [31] successfully utilized a feed forward back-propagation neural network to forecast groundwater levels. They used the Levenberg-Marquardt training function and obtained good forecast results for up to 18 months. Haeri et al. [32] designed a neural network system to model the phenomenon of pain. Their system predicted the response of the body when subjected to certain physiological excitations. Reddy et al. [33] designed a hybrid fuzzy logic committee neural network for differentiating spondyloarthropathy of the knee from rheumatoid arthritis. Pèrez-Roa et al. [34] designed a system of neural networks to model air-pollution in urban areas. The neural networks were used to predict the best eddy diffusivity functions for particular urban areas. These functions gave a good estimate of the vertical pollutant dispersion and improved the capability of the corresponding dispersion model. Vijayabaskar et al. [35] constructed an artificial neural network system to simulate the mechanical properties and volume fraction of a series of nitrile rubber compounds.

CHAPTER III

METHODOLOGY

The objective of the study was to design a classifier that employed a system of artificial neural networks to categorize the types/subtypes of human cancer using gene expression data. More specifically, this study focused on the classification of subtypes of leukemia cancer. The subclasses were precisely:

1. T-Cell Acute Lymphoblastic Leukemia
2. B-Cell Acute Lymphoblastic Leukemia
3. Acute Myeloid Leukemia

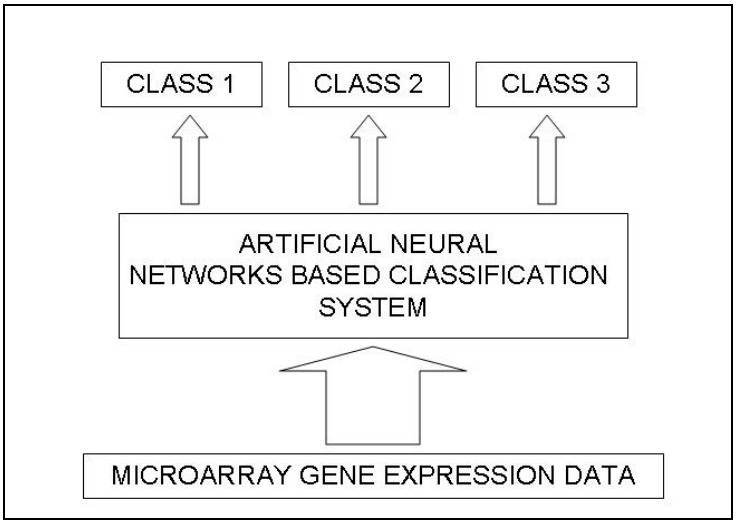

Figure 3.1: Broad outline of the ternary classification system

3.1 ABOUT THE DATASET

The data used for the present study was obtained and uploaded in the public domain by the Broad Institute of MIT and Harvard [36]. This dataset consisted of gene expression profiles of 72 patients diagnosed with leukemia cancer. Each profile consisted of quantitative expression levels for 7129 human DNA probes. The probes were spotted on high-density oligonucleotide Affymetrix Hu6800 Microarrays. In 62 patients, the genes were obtained from tissue samples collected from the bone marrow. For the remaining 10 patients, genes were obtained from the peripheral blood. In both cases the samples were collected at the time of diagnosis and before the treatment. This dataset was divided into training (38 samples) and validation (34 samples) sets and was made available on the website of the Broad Institute. Table 3.1 shows the distribution of samples as it originally appeared in the pilot study

Table 3.1: Distribution of samples as used in the original work

Class	ALL		AML					Total
Subclass	B-Cell	T-Cell	M1	M2	M4	M5	N/A	
Training Set	19	8	3	5	1	2	0	38
Validation Set	19	1	1	5	3	0	5	34
Total	38	9	4	10	4	2	5	72

3.2 FORMAT OF DATASETS

The datasets were made available in text and excel formats. Table 3.2 presents a brief snapshot of the data that was presented in the public domain.

Table 3.2: Snapshot of dataset as downloaded from the website of the Broad Institute

Gene Description	Gene Accession Number	1	call	2	call	3	call	4	call
KIAA0011 gene	D13636_at	-24	A	-134	P	-213	A	-60	A
KIAA0012 gene	D13637_at	98	A	138	A	-276	A	134	A
CCND2 Cyclin D2	D13639_at	4707	P	3367	P	-44	A	101	P
HLA-C Major histocompatibility	D13640_at	656	P	185	A	426	A	975	A
KIAA0016 gene	D13641_at	307	P	1352	P	640	P	469	P
KIAA0017 gene	D13642_at	100	A	372	A	-1	A	234	A

3.2.1 Explanation of Fields:

The Gene Description field provided information about the particular gene. In most of the cases it was the name of the gene while in some cases it provided description of the gene. The Gene Accession Number acted as an index to the gene in the public databases. Information about the gene could be searched in public databases like GenBank, located at National Center for Biotechnical Information (NCBI) at the National Library of Medicine [37]. The gene expression level fields provided intensity values of the gene expressions for each patient. Patient metadata could be obtained from the number of the columns. These fields are marked by integer numbers. Each number provides an index to the patient field in the leukemia Samples Information File located on the server [38]. The patient field described the particular patient with regards to the sub category of leukemia that the patient suffered from, the date of diagnosis, the gender of the patient and the location from where the data was obtained. The Call field was a flag field that indicated whether the particular intensity value was as a result of the actual presence of the gene at that probe or whether it was due to noise. The oligonucleotide microarrays have pairs of probe sets for every sequence. One probe set called the Perfect

Match (PM) probe set is designed to perfectly match the target transcript. The other probe set, called mismatch (MM) probe set, measures the non-specific binding signal of the partner probes. The mismatch probe set and the perfect-match probe set are identical except for one difference: the central nucleotide on the mismatch probe is replaced by its complimentary base. There are around 11 to 20 probes in the PM and MM probe sets for each gene. If the mean of the PM probe signals for one gene is significantly greater than the mean of MM probe signals for the same gene then it implies that the target transcript exists and the call is called “present”. If the mean of PM probe signals for one gene is significantly smaller than the average of MM probe signals for the same gene then it implies that the target transcript does not exist and the call is called “absent”. This could indicate that there is a high probability that the signal is affected by background noise. In the third case, if the mean of PM probe signals for one gene is neither significantly greater nor significantly smaller than the average of MM probe signals, then the call is called “marginal”, indicating that the presence or absence of the particular gene cannot be determined [39, 40].

3.3 PROCEDURE

The original study classified the data into two broad categories of leukemia (i.e. Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia) [7]. The effort of the present study was initially directed towards verification if similar or better results than the original study could be obtained using committee neural networks. Later, the technique was extended to further subcategorize acute lymphoblastic leukemia into T-Cell ALL and

B-Cell ALL. For the purpose of the current study, the data was reshuffled to accommodate the 3 classes. Sub-classification of AML was not considered because the sub-classes did not have sufficient samples. The total data of 72 patients was divided into 3 sets. Table 3.3 describes the reshuffled distribution of samples as used for the purpose of the current study.

1. Training Set: This comprised of 5 samples of T-Cell ALLs, 19 samples of B-Cell ALLs and 13 samples of AMLs to give a total of 37 samples. This dataset was used for training a series of feed forward back-propagation neural networks.
2. Initial Validation Set: This comprised of 2 samples of T-Cell ALLs, 3 samples of B-Cell ALLs and 3 samples of AMLs to account for 8 samples. This dataset was used for initial validation of the trained networks. The results were used as basis for the recruitment of the committee.
3. Final Validation Set: This comprised of 2 samples of T-Cell ALLs, 16 samples of B-Cell ALLs and 9 samples of AMLs to account for 27 samples. The recruited committee was validated using this dataset.

Table 3.3: Distribution of samples as used in the present study

Class	ALL		AML	Total
Subclass	T-Cell	B-Cell	AML	
Training	5	19	13	37
Initial Validation	2	3	3	8
Final Validation	2	16	9	27
Total	9	38	25	72

3.3.1 Data Preprocessing

The training dataset was considered for preprocessing purposes. The dataset

consisted of 7129 genes for each of 37 patients. The first step of data preprocessing was to eliminate endogenous control genes (housekeeping genes). These genes were present in all cells and were needed for day to day cellular activities. Their expression level was almost always constant across all the cells and they are used on microarrays for alignment purposes. However they exhibited very little contrast across all cells [41]. Hence they were tagged as non- informative genes and eliminated from consideration. The genes with “absent” calls across all cells were assumed to have been affected by background noise. Hence they were eliminated from consideration. Background noise or limitations of imaging equipment could also affect expression levels resulting in presence of extreme valued outliers in the dataset. Previous studies on microarray data preprocessing had shown that the imaging equipment could not measure intensity values beyond 16000 because they went into saturation. The studies had also shown that intensity values less that 20 were as a result of background noise [42, 43]. Hence the dataset was further modified by thresholding upper and lower value outliers by 16000 and 20 respectively. For microarray data to be used for classification purposes, it was necessary that the gene expressions to be classified showed differential information across the classes of interest. Previous studies had used an n-fold change technique to eliminate genes where n varied from 2 to 5. Smaller the value of n, lesser conservative was the approach and vice versa. For the purpose of this study, a moderately conservative approach was used by eliminating genes with less than 2.5 fold change. This technique for gene selection was roughly based on grounds of the process adopted by Dudoit et al. [44] in their study on methods of tumor classification.

The values for each gene in the resulting dataset were scaled between -1 and 1 using the following formula

$$normalizedVal = 2 * \frac{(currVal - smallestVal)}{(biggestVal - smallestVal)} - 1$$

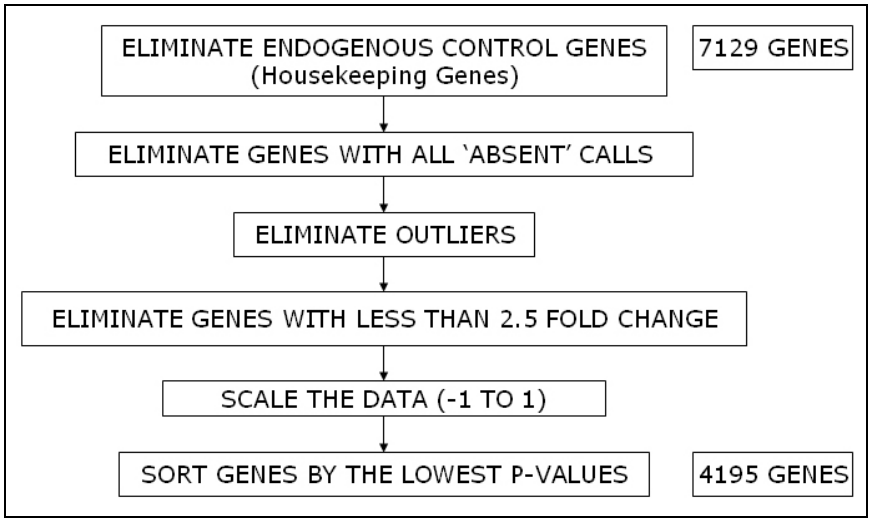

Figure 3.2: Block Diagram depicting the Data Preprocessing procedure

A statistical test was carried out between the groups for each gene expression across the samples. In the case of the binary classification problem, an independent t-test was carried out to identify significantly contrasting genes. In case of the ternary classification problem a single factor ANOVA test was deployed. The only purpose of the statistical tests was to find genes with low p-values between classes. The genes were then ordered

according to their p-values from smallest to largest. In this way the strongly differentially expressed genes were obtained in an ordered fashion. These were treated as top most informative genes and were used for further analysis. The original dataset was thus preprocessed to eliminate the obviously non- informative genes and obtain a reduced set in which genes were arranged according to decreasing levels of significance. The most informative genes lay at the top and the least informative genes were at the bottom

3.3.2 Creating Input Parameter sets

The output of the data preprocessing state was a dataset of 4795 genes arranged in decreasing order of informative potential. The order of these genes varied with the type of study (i.e. the order differed depending upon whether the system was for two classes or three classes). In the binary classification system, the top 25 genes were picked to train the classifiers. Upon obtaining satisfactory results an attempt was made to use the same approach for the ternary classification system. However, the approach did not work well and the networks misclassified consistently. The top 25 genes did not contribute adequate information for distinguishing between the three subcategories. There was a need for a different approach when more than 2 classes were to be considered. Hence, in order to accommodate differential genes that could clearly distinguish between 3 classes, the top 250 genes with a p-value cutoff of 0.001 were picked. In this way the 250 top informative genes were filtered out. These selected genes were then grouped in 10 groups of 25 genes each. The groups formed the basis of the multi-classifier building effort. Each of these gene groups were used as the input parameters for the neural network based classification

systems for initial validation and committee recruitment.

3.3.3 Initial Validation and Committee Recruitment

The preprocessed data was used as input parameters for the classification system. This system was composed of a series of artificial neural networks. The Neural Networks were implemented in MATLAB using the Neural Networks toolbox.

- The Binary Classification System

Around 36 neural networks were designed with architectural differences. These differences were in regards to the number of hidden neurons varying from 8 to 25, initial weights, hidden layers varying from 2 to 3 and different transfer functions. All the networks were fully connected and feed forward in nature. The networks took the top 25 informative gene expression data from the training dataset as an input. LOGSIG transfer functions were used for neurons in the output layer. This function scaled the output of these networks in the range of 0 to 1. Networks were trained to recognize an output of 0 as ALL and a 1 as AML. If the obtained output exceeded 0.7 then it was converted to 1 and if it was less than 0.7 then it was floored to 0. The former output indicated class-ALL and the latter indicated class-AML. The Levenberg-Marquardt training function was used for the training purposes. The training experiment was set such that training would stop if either of the following constraints were satisfied:

- The error goal of $1*10^{-17}$ was reached
- The number of epochs exceed 1000

The trained neural networks were then validated using the initial validation set consisting of 8 samples. Of the networks that were trained, the top 5 performing networks were recruited into a committee. Performance was measured on the basis of number of correct classifications.

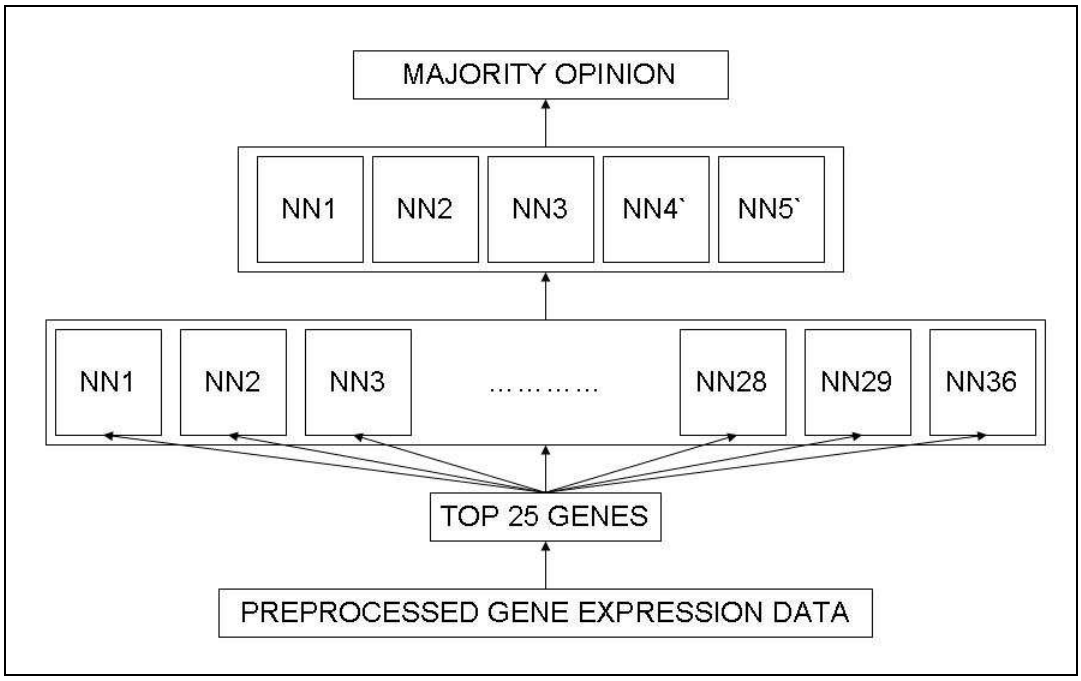

Figure 3.3: Block diagram depicting the working of the binary classification system

- The Three Class Predictor System:

A different approach was considered in the designing of the three class predictor system. The top 250 informative genes were randomly divided into 10 groups of 25 genes each. Each of the 10 groups of preprocessed data was used as input parameters. Around 10 to 12 neural networks per group were trained to give a total of more than 100 neural

networks. The architectures of the networks were different from each other in regards to the number of hidden neurons varying from 8 to 25, initial weights, hidden layers ranging from 2 to 3 and different transfer functions. All these networks were fully connected and feed forward in nature. As an input, these networks took the 25 values of each group from the training dataset that was obtained after the preprocessing effort. The networks had three nodes in the output layer. Each node represented one of the three classes of leukemia. Table 3.4 presents the output configuration for the networks in the ternary classification system.

Table 3.4: Output Configuration of the ternary classification system

Node 1 B-CELL ALL	Node 2 T-CELL ALL	Node 3 AML	O/P
0	0	0	No Classification
0	**0**	**1**	**AML**
0	**1**	**0**	**T-CELL ALL**
0	1	1	Ambiguous
1	**0**	**0**	**B-CELL**
1	0	1	Ambiguous
1	1	0	Ambiguous
1	1	1	Ambiguous

LOGSIG transfer functions were used for neurons in the output layer. This function scaled the output of these networks in the range of 0 to 1. Networks were trained to recognize output 0 as absence of a particular class and 1 as the presence of that class. If the obtained output at any of the nodes exceeded 0.7 then it was converted to 1 and if it was less than 0.7 then it was floored to 0. If for a particular case, the addition of the outputs did not equal 1 then the classification was treated as ambiguous and was not

assigned to any of the classes. When all three outputs were 0 then the classification was treated as no classification and the case was not assigned to either of the classes. The Levenberg-Marquardt training function was used for the training purposes. The training experiment was set such that training would stop if either of the following constraints were satisfied:

- The error goal of $1*10^{-17}$ was reached
- The number of epochs exceed 1000

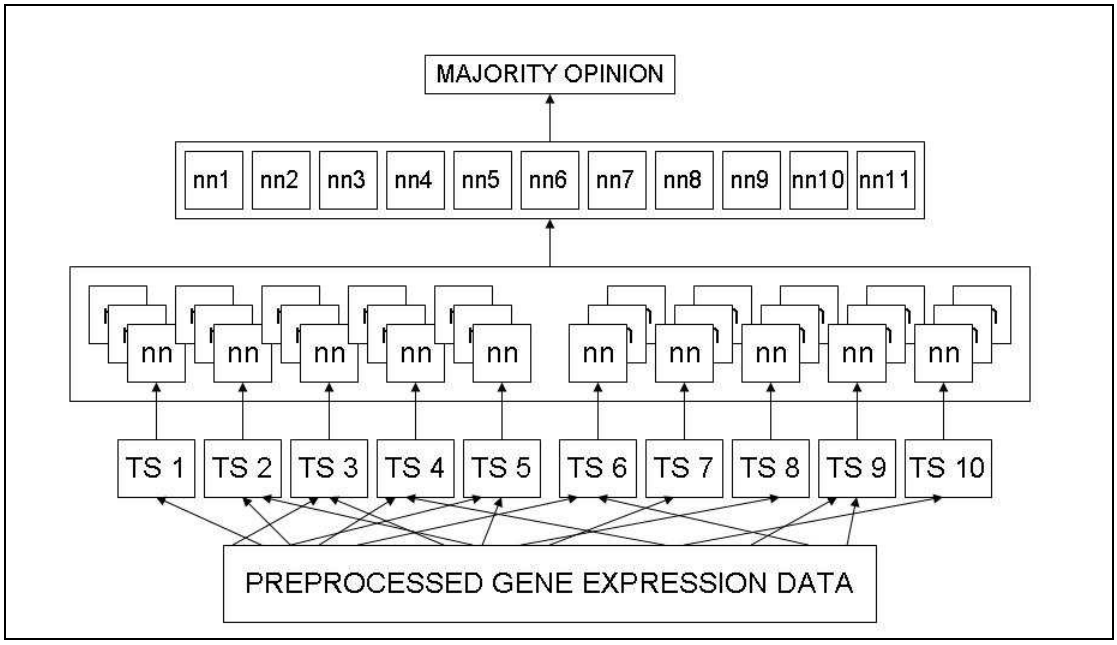

Figure 3.4: Block Diagram depicting the working of the ternary classification system

The trained neural networks were then validated using the initial validation set which comprised of 8 samples. Of the networks that were trained, committees of 5, 7, 9 and 11 members were formed. Performance was measured on the basis of number of correct classifications. The committee that validated the final validation data with the best prediction accuracy was treated as most reliable committee and results were recorded for that committee. Figure 3.4 presents a block diagram which shows the architecture of the ternary classification system.

3.3.4 Statistical Analysis

The output of the committee was compared with the expected output. Ambiguous and misclassifications were all clubbed under misclassifications category. The expected outputs were available in data files located on the website of the Broad Institute. A Binomial Test was then performed for $\alpha = 0.01$ to measure the deviation of the obtained results from those of the results expected from theory for the particular experiment.

CHAPTER IV

RESULTS

The first task of the project was to eliminate non- informative gene expression data. For both the binary classification and the ternary classification systems, the training set comprising of 37 patients underwent identical data preprocessing step. For the binary classification system a t-test was employed to obtain genes of differential class information, whereas for the ternary classification system, an ANOVA test was employed to obtain the differentially expressed genes. Following are the results of the data preprocessing step in both the cases.

4.1 DATA PREPROCESSING

Of the 7129 gene probes in each profile, there were 58 endogenous control genes and they were eliminated from further consideration. 2175 genes were further eliminated after genes with all ‘Absent’ calls across samples were filtered out of the list. Expression values greater than 16000 were replaced by 16000 while expression values less than 20 were replaced by 20. Genes with less than 2.5 fold changes across samples were eliminated from further consideration. In this process, 100 more genes were eliminated to yield a preprocessed dataset of 4795 genes expressions. The gene expression values were scaled to get them between -1 and 1. For the binary classification system, a t-test was

performed for each gene across the 2 classes and the p-values were recorded. The datasets then sorted according to increasing p-value. The figure 4.1 shows the values of gene expressions for the top 50 genes plotted for the 2 classes of patients. In the figure, the actual values of expressions were scaled to a full range of a hot color map and the intensity values were plotted. Distinct division between the two groups of data can be seen in the figure. In most of the cases, the intensity values of gene expressions in ALLs are higher than those in AMLs.

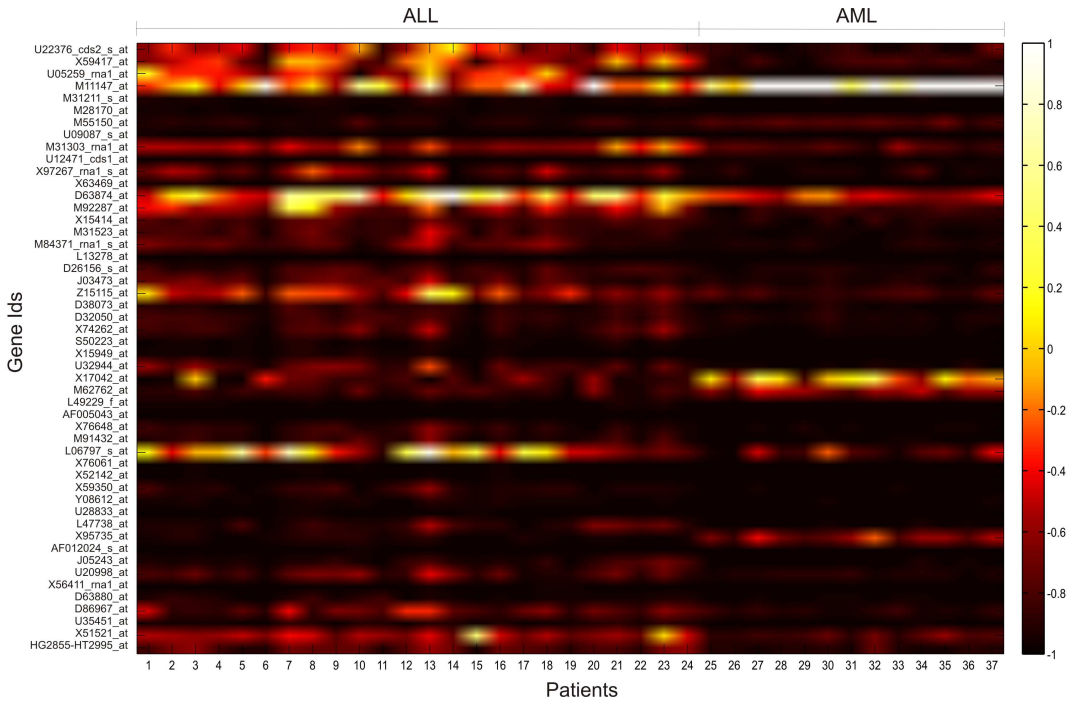

Figure 4.1: Heat map showing intensity values of top 50 gene expressions scaled to a ‘hot’ color map. The x-axis displays the patients clubbed according to disease while the y axis shows informative genes

In the case of the 3 class system, an ANOVA was performed to obtain the p-values. The sole purpose of the statistical tests was to obtain genes with differential expressions across the groups. The following set of figures (4.2, 4.3 and 4.4) gives a suggestive evidence of division of classes based on intensity values.

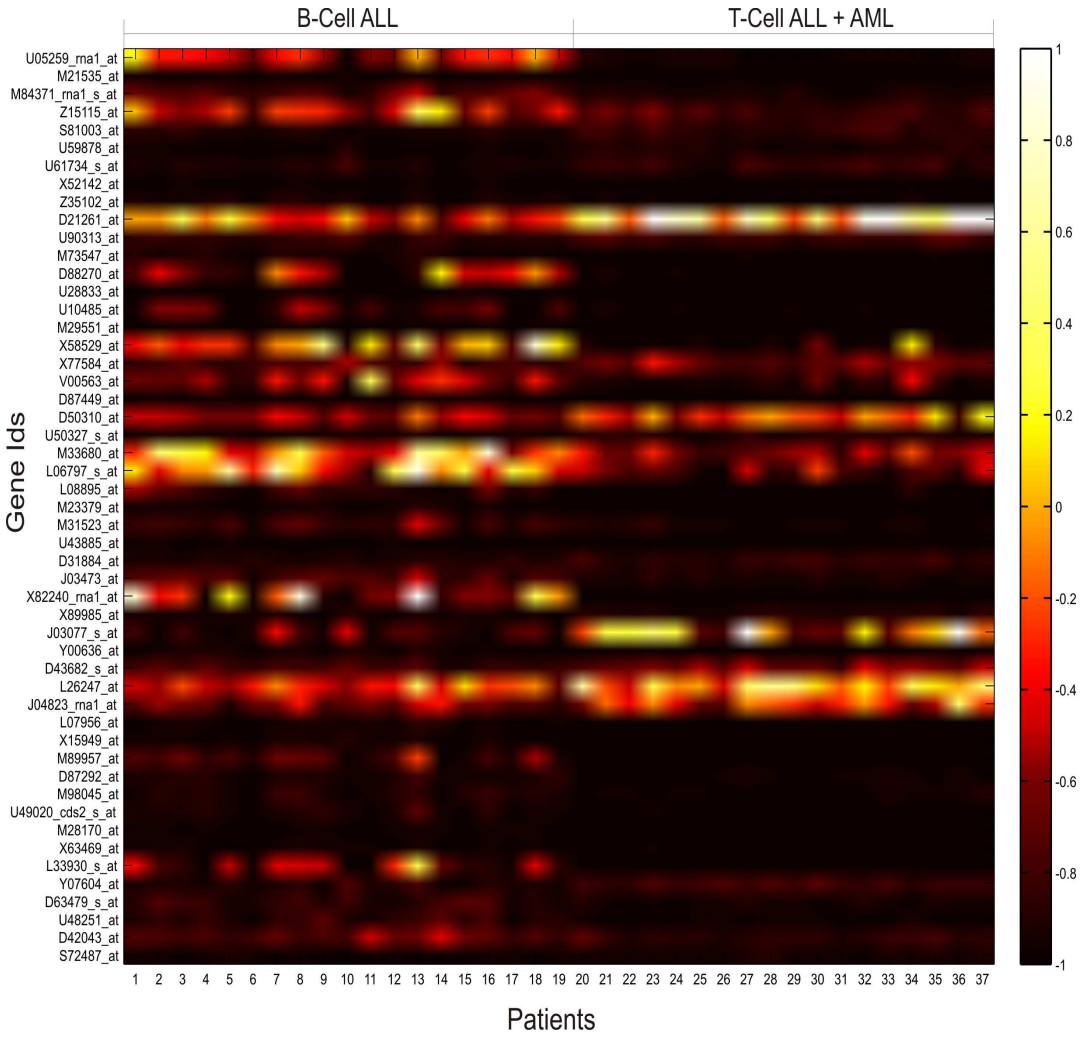

Figure 4.2: Heat map showing intensity values of gene expressions scaled to a ‘hot’ color map. There were 51 genes that showed differential expressions for B-Cell ALL and are plotted on the y-axis. The x-axis displays the patients clubbed according to disease

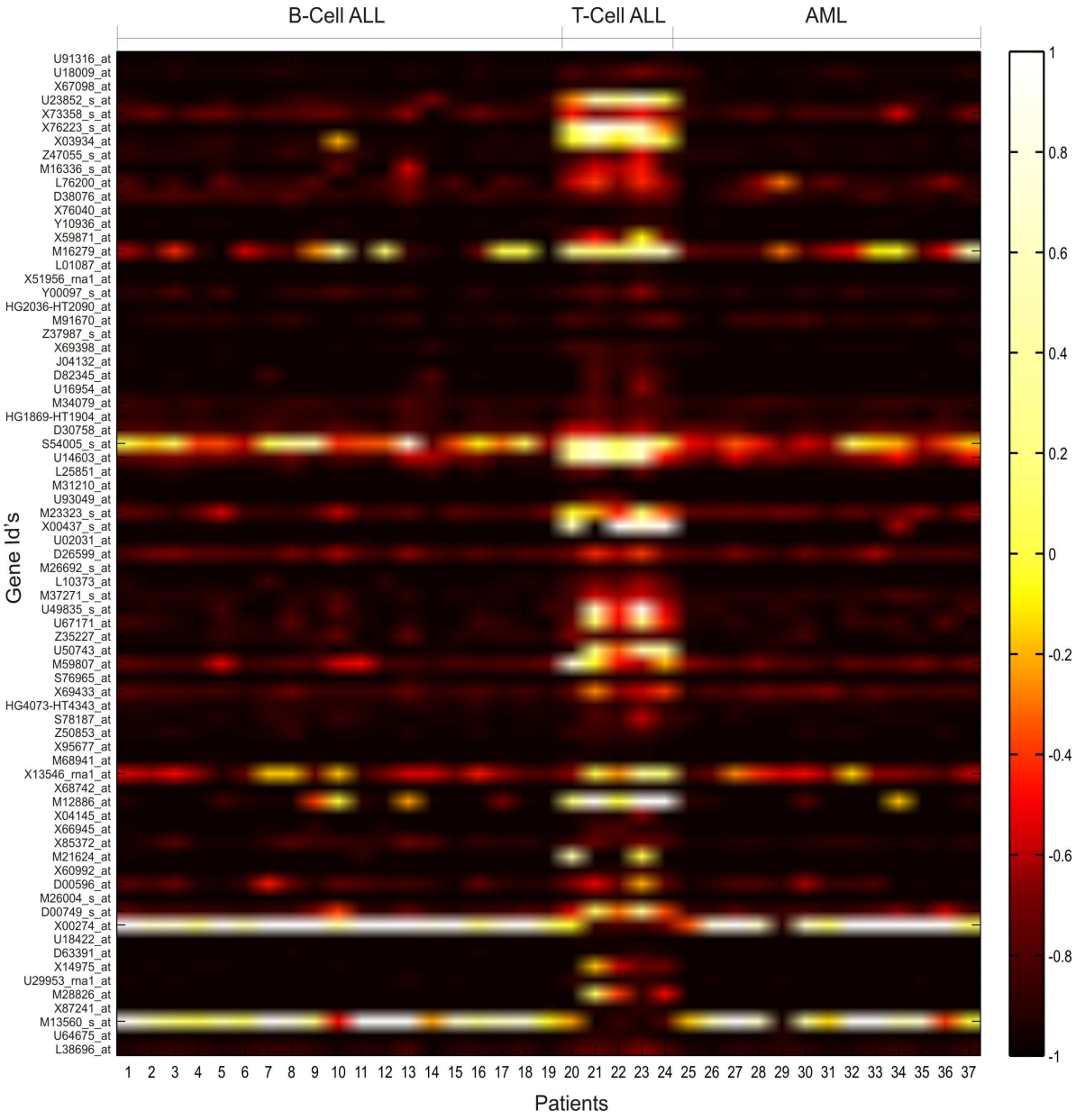

Figure 4.3: Heat map showing intensity values of gene expressions scaled to a 'hot' color map. There were 88 genes out of top 250 genes that showed differential expressions for T-Cell ALL and are plotted on the y-axis. The x-axis displays the patients clubbed according to the disease

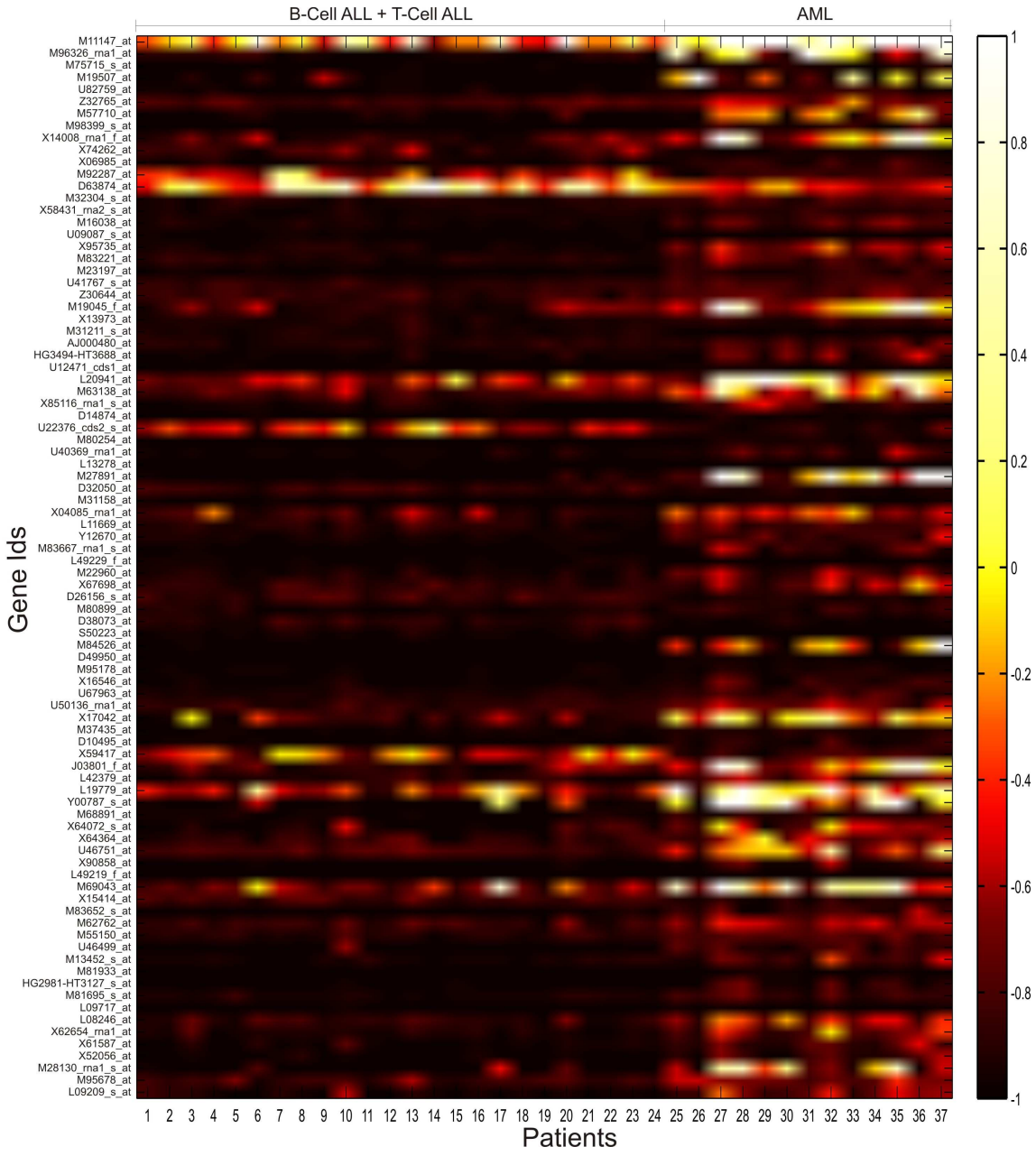

Figure 4.4: Heat map showing intensity values of gene expressions scaled to a ‘hot’ color map. There were 73 genes out of top 250 genes that showed differential expressions for AMLs and are plotted on the y-axis. The x-axis displays the patients clubbed according to the disease

4.2 RESULTS OF THE BINARY CLASSIFIER:

The preprocessed data was used to train thirty six neural networks. The trained networks were initially tested against an initial validation set comprising of eight samples. The worst case performance was from a network which correctly predicted only four of the eight samples of the initial validation set. The best case performance was achieved by one network which rightly classified all 8 samples. Five top performing networks were recruited into a committee.

The recruited committee gave 100 percent classification accuracy for the initial validation set as can be seen from Table 4.1 and Figure 4.5. This committee was then used for validating a fresh set of data comprising of 27 samples. Of the individual networks of the committee, the best case performance was obtained by three networks which rightly predicted the class in all the 27 cases. The worst case performance was obtained by one network which predicted the right class 24 out of 27 times. However, the committee decision with majority opinion correctly classified the data in 100 percent of the cases (Table 4.2 and Figure 4.6). Since neither the initial validation nor final validation datasets was used for training purposes, the combined accuracy achieved by the system for 35 samples was 100 percent (Table 4.3 and Figure 4.7).

Table 4.1: Results of validation by individual networks and the integrated committee, when evaluated with an intermediate validation dataset of 8 samples

NETWORK	CORRECT CLASIFICATION OUT OF 8 EXPRESSIONS PRESENTED	ACCURACY
NN1	7	87.5
NN2	8	100
NN3	7	87.5
NN4	7	87.5
NN5	7	87.5
COMMITTEE RESULT	**8**	**100**

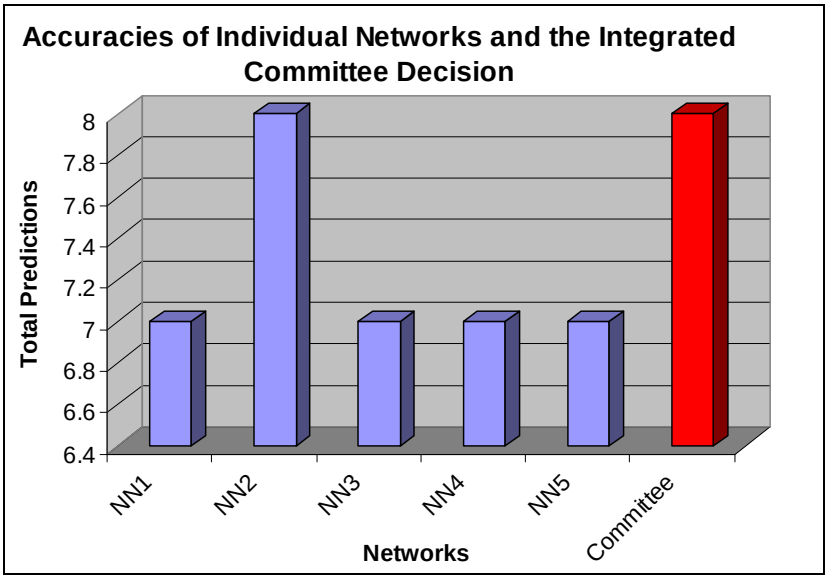

Figure 4.5: Comparative plot of prediction accuracies of individual networks versus the integrated committee decision for the initial validation dataset

Table 4.2: Results of validation by individual networks and the integrated committee, when evaluated with the final validation dataset of 27 samples

NETWORK	CORRECT CLASIFICATION OUT OF 27 EXPRESSIONS PRESENTED	ACCURACY
NN1	24	88.9
NN2	27	100
NN3	25	92.6
NN4	27	100
NN5	27	100
COMMITTEE RESULT	**27**	**100**

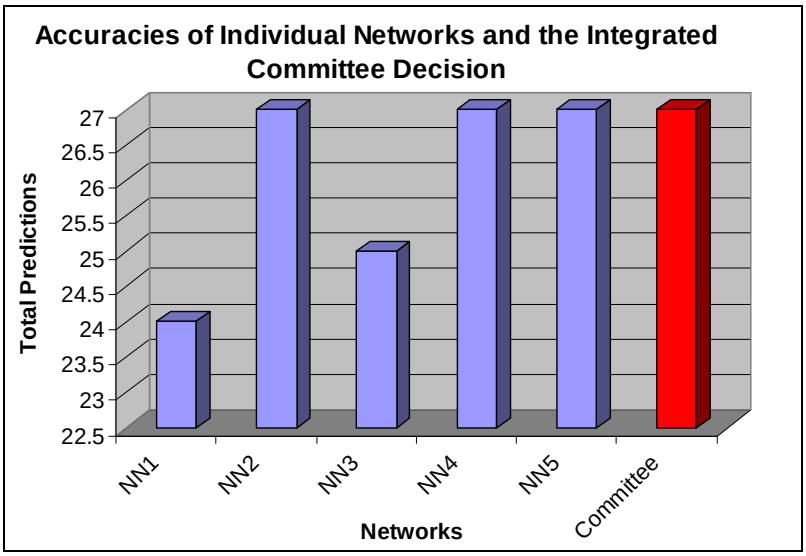

Figure 4.6: Comparative plot of prediction accuracies of individual networks versus the integrated committee decision for the final validation dataset

Table 4.3: Integrated results of intermediate and final validation by individual networks and the integrated committee, when evaluated for a combined data of 35 samples

NETWORK	CORRECT CLASIFICATION OUT OF 35 EXPRESSIONS PRESENTED	ACCURACY
NN1	31	88.6
NN2	35	100
NN3	32	91.4
NN4	34	97.1
NN5	34	97.1
COMMITTEE RESULT	35	**100**

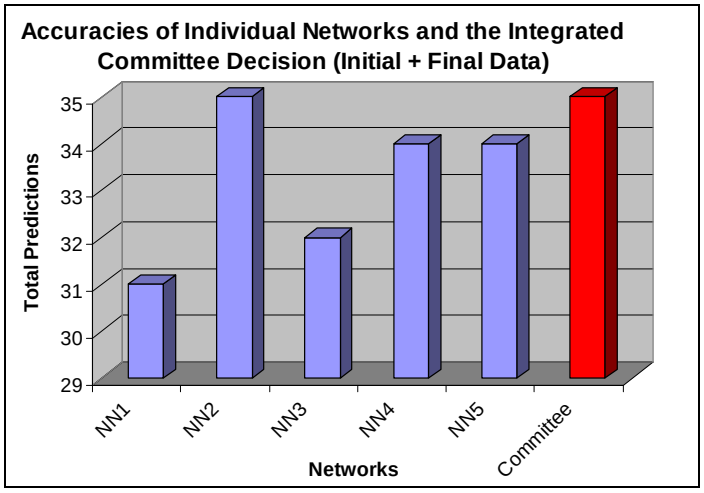

Figure 4.7: Comparative plot of prediction accuracies of individual networks versus the integrated committee decision for the combined (i.e. initial + final validation) datasets

4.3 RESULTS OF THE TERNARY CLASSIFIER:

The preprocessed data was divided into ten groups of 25 samples each. Each of these groups was used to train around 10 to 12 neural networks. A total of 115 networks were trained in this manner. Each of the trained networks was validated against the initial validation datasets. The worst case performance was from 5 networks which correctly predicted only four of the eight samples of the initial validation set. The best case performance was achieved by 20 networks which rightly classified all 8 samples. Networks which accurately predicted at least 7 out of the 8 test cases were considered. Committees of 3, 5, 7, 9 and 11 individual networks were formed and evaluated against the final testing dataset of 27 samples. The committee of 11 networks gave the best prediction accuracy. Table 4.4 presents the configuration of the individual committee members. Table 4.5 presents six classification results that explain the majority voting technique of a neural network committee. The values in bold indicate misclassifications. The actual classes were obtained from lab results and were downloaded from the website of the Broad Institute.

The recruited committees gave 100 percent classification accuracy for the initial validation set (Table 4.6 and Figure 4.8). These committees were then used for validating a fresh set of data comprising of 27 samples. The committee with eleven members yielded the best performance. It rightly predicted 26 out of the 27 test cases to yield a prediction accuracy of 96.29 percent (Table 4.7 and Figure 4.9). Of the individual networks of the committee, the best case performance was obtained by three networks which rightly predicted the class in 26 out of the 27 cases. The worst case performance

was obtained by one network which predicted the right class 20 out of 27 times. Since neither of the initial validation and final validation datasets was used for training purposes, the combined accuracy achieved by the system for 35 samples was 97.14 percent (Table 4.8 and Figure 4.10). The confusion matrix for the system can be observed in Table 4.9 and the Figure 4.11 gives a class wise distribution of accuracy of the individual networks and that of the committee.

Table 4.4: Network configuration of individual members for the Ternary Classification System

Networks	Gene Group	Hidden Layers	Layer 1		Layer 2		Output Layer	
			Neurons	Transfer Func.	Neurons	Transfer Func.	Neurons	Transfer Func.
NN1	1	1	12	TANSIG	-	-	3	LOGSIG
NN2	3	1	16	TANSIG	-	-	3	LOGSIG
NN3	3	1	20	TANSIG	-	-	3	LOGSIG
NN4	3	2	22	TANSIG	16	TANSIG	3	LOGSIG
NN5	4	1	14	TANSIG	-	-	3	LOGSIG
NN6	4	1	22	TANSIG	-	-	3	LOGSIG
NN7	4	1	26	TANSIG	-	-	3	LOGSIG
NN8	9	1	12	TANSIG	-	-	3	LOGSIG
NN9	9	1	16	TANSIG	-	-	3	LOGSIG
NN10	9	1	18	TANSIG	-	-	3	LOGSIG
NN11	9	1	20	TANSIG	-	-	3	LOGSIG

Table 4.5: Majority voting technique of a committee of neural networks

	SAMPLES					
	1	2	3	4	5	6
NN1	B-ALL	B-ALL	**B-ALL**	**NC**	AML	AML
NN2	B-ALL	**NC**	**AML**	**AML**	AML	**NC**
NN3	B-ALL	B-ALL	**AML**	**AML**	AML	AML
NN4	B-ALL	B-ALL	**AML**	**AML**	AML	**B-ALL**
NN5	B-ALL	B-ALL	**AML**	T-ALL	AML	AML
NN6	B-ALL	B-ALL	**AML**	T-ALL	AML	AML
NN7	B-ALL	B-ALL	**AML**	T-ALL	AML	AML
NN8	B-ALL	**NC**	T-ALL	T-ALL	**NC**	AML
NN9	B-ALL	B-ALL	T-ALL	T-ALL	**B-ALL**	**B-ALL**
NN10	B-ALL	B-ALL	T-ALL	T-ALL	**B-ALL**	**B-ALL**
NN11	B-ALL	B-ALL	T-ALL	T-ALL	**B-ALL**	**B-ALL**
Committee Decision	B-ALL	B-ALL	**AML**	T-ALL	AML	AML
Actual Class	B-ALL	B-ALL	T-ALL	T-ALL	AML	AML

Table 4.6: Results of validation by individual networks and the integrated committee, when evaluated with an intermediate validation dataset of 8 samples

NETWORK	NN1	NN2	NN3	NN4	NN5	NN6	NN7	NN8	NN9	NN10	NN11	COMMITTEE RESULT
CORRECT CLASSIFICATION OUT OF 8 SAMPLES PRESENTED	7	7	7	7	7	8	8	8	8	8	8	**8**
ACCURACY	87.5	87.5	87.5	87.5	87.5	100	100	100	100	100	100	**100**

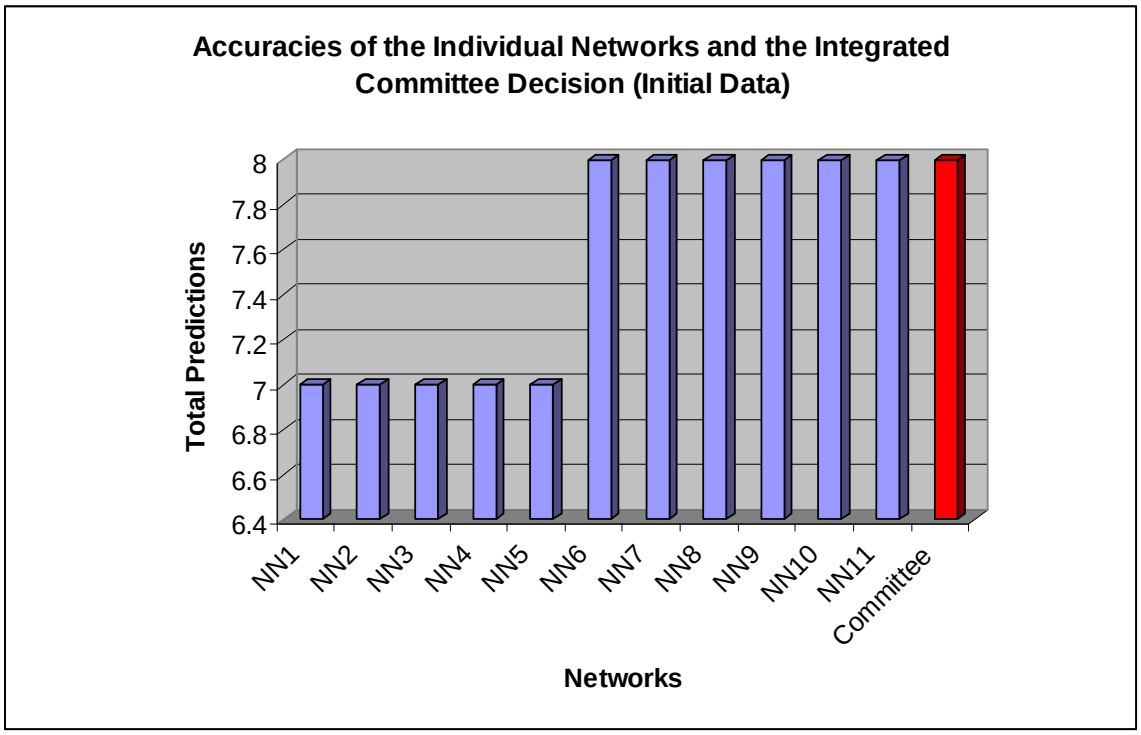

Figure 4.8: Comparative plot of prediction accuracies of individual networks versus the integrated committee decision for the initial validation dataset

Table 4.7: Results of validation by individual networks and the integrated committee, when evaluated with the final validation dataset of 27 samples

NETWORK	NN1	NN2	NN3	NN4	NN5	NN6	NN7	NN8	NN9	NN10	NN11	COMMITTEE RESULT
CORRECT CLASSIFICATION OUT OF 27 SAMPLES PRESENTED	26	20	26	23	25	25	25	26	24	23	23	**26**
ACCURACY	96.3	74.1	96.3	85.2	92.6	92.6	92.6	96.3	88.9	85.2	85.2	**96.3**

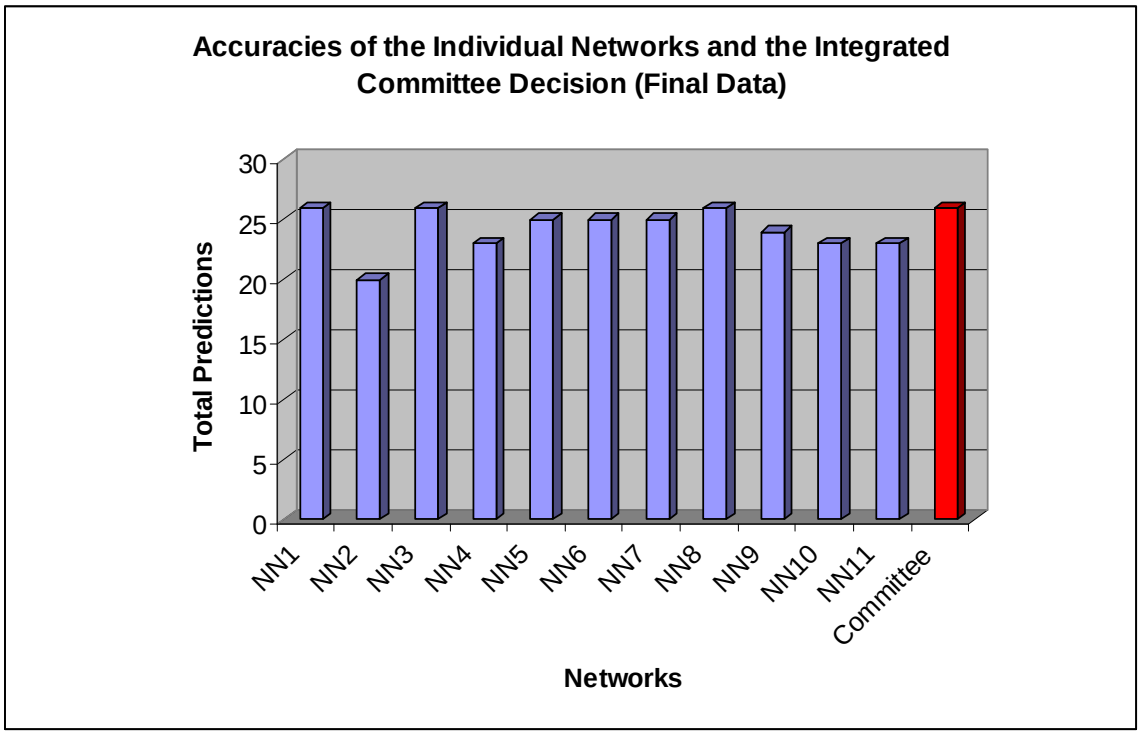

Figure 4.9: Comparative plot of prediction accuracies of individual networks versus the integrated committee decision for the final validation dataset

Table 4.8: Integrated results of intermediate and final validation by individual networks and the integrated committee, when evaluated for a combined data of 35 samples

NETWORK	NN1	NN2	NN3	NN4	NN5	NN6	NN7	NN8	NN9	NN10	NN11	COMMITTEE RESULT
CORRECT CLASSIFICATION OUT OF 35 SAMPLES PRESENTED	33	27	33	30	32	33	33	34	32	31	31	**34**
ACCURACY	94.3	77.1	94.3	85.7	91.4	94.3	94.3	97.1	91.4	88.6	88.6	**97.14**

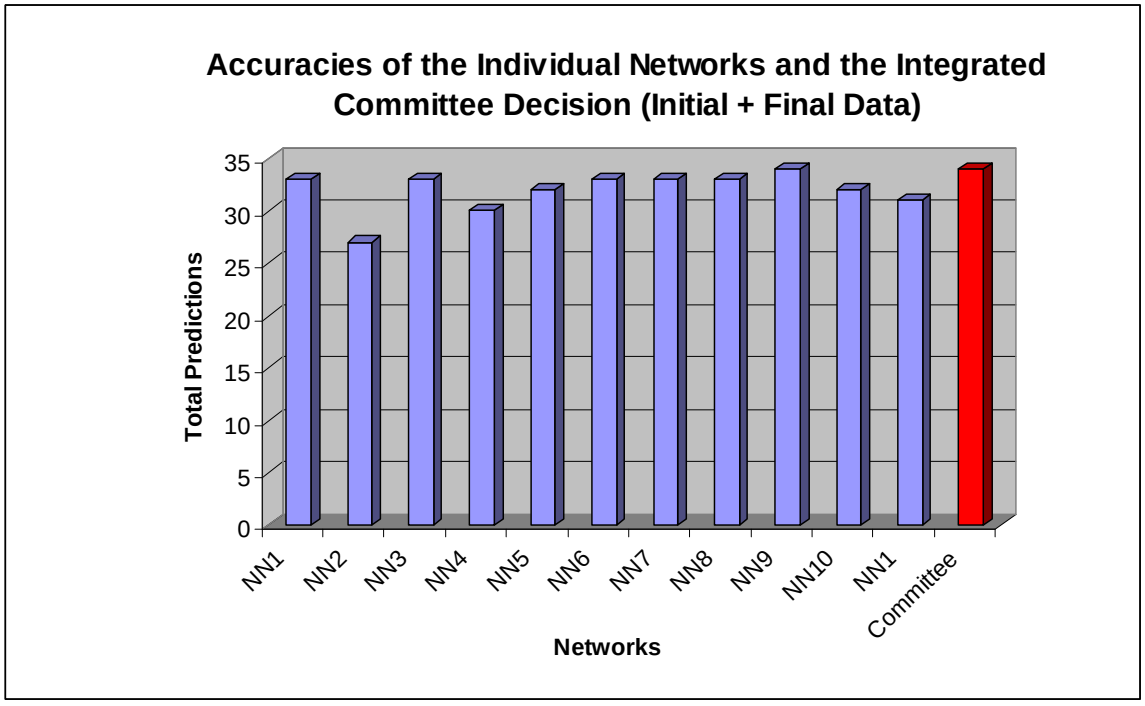

Figure 4.10: Comparative plot of prediction accuracies of individual networks versus the integrated committee decision for the combined (i.e. initial + final validation) datasets

Table 4.9: Confusion Matrix for the ternary classification system (Specificity and Sensitivity are calculated using one vs. all approach)

		Actual			
		B-ALL	T-ALL	AML	Σ
Predicted	B-ALL	19	0	0	19
	T-ALL	0	3	0	4
	AML	0	1	12	13
	Σ	19	4	12	35
Sensitivity		100	75	100	
Specificity		100	97.14	100	

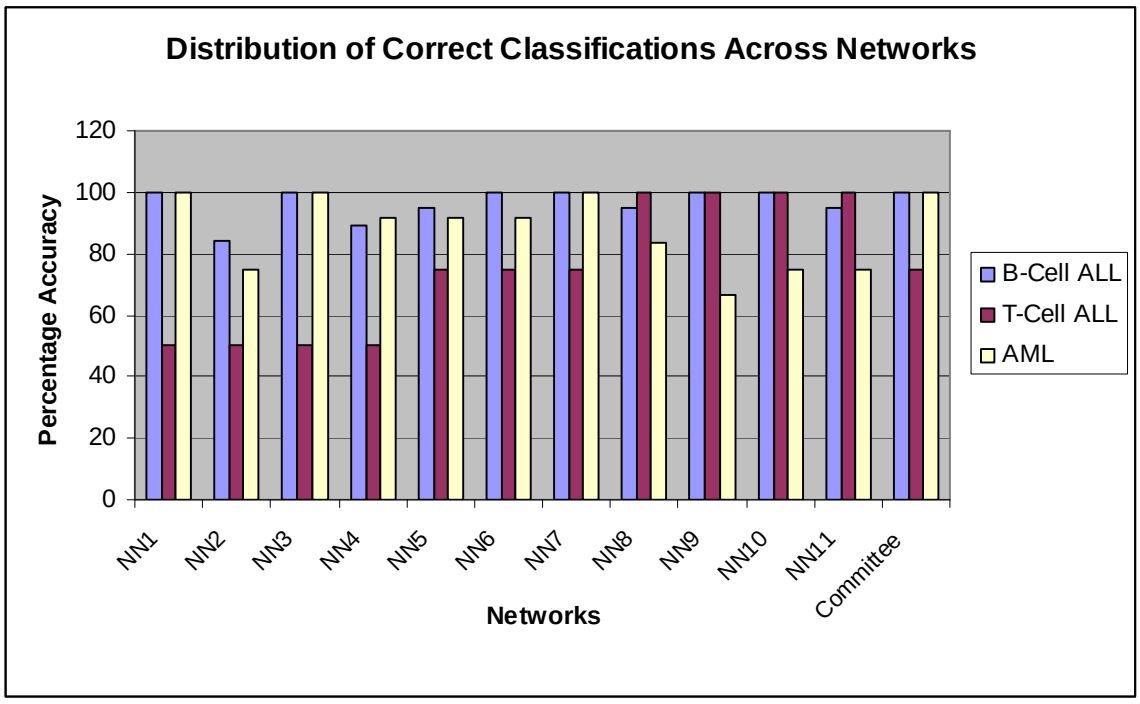

Figure 4.11: Plot of percentage accuracies of individual networks versus the integrated committee decision for the class split up

CHAPTER V

DISCUSSION

The present study provided an effective technique to classify cancer subtypes. It represents the first application of committee neural networks for cancer classification using microarray gene expression data. The binary classifier successfully classified the samples into their respective subclasses to give 100 percent prediction accuracy (Table 4.3, Figure 4.7). The ternary classifier gave an accuracy of 97.14 percent as it failed to classify 1 out of 35 validation sets (Table 4.7, Figure 4.10). Thus the results failed to support the null hypothesis of the study, and the alternate hypothesis was accepted. Only one network from the individual committees managed to achieve an equivalent accuracy as that of the committee. However, considering the heuristic nature of machine learning algorithms, the committee decision provided a highly reliable result with more confidence than the individual networks.

5.1 GENE SELECTION

Microarray chips provide a scope for parallel monitoring of gene expressions. However, not all the genes are informative from the point of view of a classification problem. The goal of gene selection was to filter out the uninformative genes and keep a

subset of genes that could be easy to handle. Of course, it was also expected that the selected genes were significant from the point of view of the disease. The list of genes can be obtained from the heat maps of figures 4.1 to 4.4. Some of the genes were investigated from the point of view of their functionality and the type of proteins they encoded. This was carried out using Entrez Gene Search Engine [45]. Table 5.1 presents the functions of some of the genes that were selected from the data preprocessing stage. Genes marked with an asterisk were found to be common with some of the other studies involving the leukemia patient data. The complete list of 250 genes can be found in the Appendix B.

However, some of the significant genes such as EVI2A, NFKB2, MLL3, etc. that have been identified as being associated with the different types of leukemia did not lie in the obtained subset. This could be attributed to the fact that these genes showed deviation from normal levels of expression. However, the present study already assumed that the genes under study were obtained from leukemia patients and the expression levels for the popular genes did not show a statistical difference between the subclasses. Genes with less than 2.5 fold change were eliminated from consideration. Past studies have carried out n fold eliminations with the values of n ranging from two to five. As the value of n increased, more genes were eliminated and vice versa. The value of n dictated the level of desired conservation of genes. The present study utilized a moderately conservative approach by eliminating all those genes whose maximum expression value was at most 2.5 times the minimum value across samples.

Table 5.1: Sample of identified genes with level of significance and Function

Gene Name	Gene Description	P-Value
FTL Ferritin, Light Polypeptide	The gene encodes an iron storage protein in prokaryotes and eukaryotes.	2.29E-05
T-Cell antigen CD7 molecule	This gene encodes a protein that is found on mature T cells. It has functions in the T-Cell / B-Cell interactions during early lymphoid development.	1.73E-11
Myeloid Cell Leukemia Sequence 1 (MCL1) *	This gene encodes a BCL-2 family protein. The protein coded by this gene functions in inhibiting apoptosis by enhancing cell survival	1.25E-05
CD79a Molecule, Immunoglobulin associated Alpha*	The CD79 is a B-Cell lymphocyte antigen receptor. This includes the antigen-specific component, surface immunoglobulin. The surface immunoglobulin plays an important role in the expression of the B-cell antigen receptor	6.5E-09
Cystatin C (Amyloid Angiopathy and Cerebral Hemorrhage)*	The proteins encoded by this gene are found in body fluids and secretions. These have many protective functions. A mutation of this gene has been associated with amyloid angiopathy	1.02E-06
CD11c, Integrin, alpha X (complement component 3 receptor 4 subunit)*	This gene encodes the integrin alpha X chain protein. The alpha X proteins are useful in differentiating lymphoid cells from myeloid cells.	4.53E-06
CD33, molecule*	The Protein encoded by this gene is useful in differentiating lymphoid cells from myeloid cells.	2.42E-07

5.2 SIZE OF INPUT VECTORS

In the case of the binary classifiers, the number of input vectors was varied from as less as 6 genes to as high as 50 genes. Intermediate validations showed that the consistent results were obtained when the size of the input vector was in the range of 18 to 30. Hence the neural networks were designed as 25 input networks.

In the case of the ternary classifiers, the approach of using top 25 genes did not show the same level of performance as that of the binary classifier. A need was hence

felt, to look beyond the top 25 genes for the classifier building effort. An important observation that affirmed the need was that there were as many as 618 genes in the preprocessed data that statistically differentiated the three classes based on a p-value cutoff of 0.01. An important criterion that decided the composition of the training set was the level of differential information that the genes imparted from the classification point of view. Hence a more stringent p-value cutoff of 0.001 was chosen for sifting the informative genes from the redundant ones. In this way the top 250 genes were picked for the classifier building purposes.

5.3 DESIGN OF NEURAL NETWORKS COMMITTEE

The preprocessed set comprised of the top 250 highly informative genes. The information content of these genes was interpreted by neural networks with varying architectures. More than 100 networks were trained using the sample sets of 25 genes each. The architectures of the individual networks varied in the number of hidden layers, the hidden layer neurons, the transfer functions and the learning functions. The training experiments were scheduled to terminate either upon reaching 1000 epochs or an error goal of $1e^{-17}$. Appropriate settings of these parameters were crucial in running a successful training session. The networks gave a consistent performance in terms of training time and accuracy for one to two hidden layers comprising of 14 to 25 hidden neurons. The performance became inconsistent when these parameters were modified. Intermediate validation was carried out on a set of eight samples. The output layer of the neural networks was transformed using a LOGSIG transfer function. This kept the output

within a range of 0 to 1. The pre-priori experiment settings were such that values of 0.6 or higher were rounded to 1 and values less than 0.7 were floored to 0.

The initial validation gave 100 percent accuracy in both the binary and ternary classification systems. However, the classifier misclassified a T-Cell ALL sample as an AML when tested against the fresh validation set. One reason for this misclassification was the limited number of samples in the training set. Although the Neural Network converged to zero, only five samples of T-Cell ALLs were present in the training set. Availability of more samples to represent T-Cell ALLs may have improved the prediction accuracy.

Although no one to date has used the committee neural networks for cancer sub classification, Reddy et al. [17, 20], Das et al. [18] and Palreddy et al. [19] have developed and used committee networks for classification of swallow acceleration signals and speech signals. However, in all theses studies, the same input vector was used to train a series of networks. The current studied was unique such that different sets of input networks were used to train different series of neural networks. This technique had its own plus points. As already mentioned, a need was felt to look beyond the top 25 informative genes. At the same time, it was not feasible both computationally and mathematically, to pass too many parameters to a neural network. Designing neural network systems with 250 input parameters would have resulted in numerical errors and instability, longer training time due to increased complexity in calculating input weights, problems related to data over fitting etc. By making use of parallel processing neural networks, all the 250 genes were incorporated in the decision making process. The

present study not only overcame the aforementioned limitations but also maintained simplicity of the architectures of the processing elements. This technique was analogous to the concept of a jury decision with members having widely differing backgrounds.

5.4 COMPARISON WITH PREVIOUS METHODS

The leukemia dataset has been utilized extensively for classifier building efforts. Most of these studies have looked at it as a two class problem.

5.4.1 Binary Classification System

In the original study, Golub et al. [7] built a binary classification system in order to automate classification of leukemia into its two sub-classes. Their classification effort yielded an accuracy of around 86 percent. Mallick et al. [8] designed a Bayesian Support Vector Machine based binary classifier to obtain an accuracy of around 97 percent. Peng et al. [9] used genetic algorithms and support vector machines in a leave-one-out cross validation test environment. They correctly classified all the test cases. Antonov et al. [10] used a maximal margin linear programming method using 132 to 149 top informative genes to obtain an accuracy of around 97 percent.

5.4.2 Ternary Classification System

Not many researchers have looked at it as a multiclass problem. Berrar et al. [12] were some of the few who worked to subcategorize the main classes in the leukemia Dataset. They designed a probabilistic neural networks based classifier to subcategorize

both ALLs and AMLs into their sub classes. However they managed a prediction accuracy of around 62 percent.

The technique utilized in the present study performed as good as or even better than some of the classification techniques used before. Also, the samples were reshuffled such that the validation sets used were larger than those used in the previous studies. Training and validation sets were completely independent in this study.

The ternary classifier performed better than the previous work on the same dataset. The classifier achieved and accuracy of as high as 97.1 percent failing to classify just one sample out of the 35 samples in the validation set.

5.5 LIMITATIONS OF THE DATA

Patients with Acute Myeloid Leukemia were not further subcategorized because training and testing sets could not be prepared from the limited samples in this category. The dataset lacked a sample set of normal patients. Presence of such a set could have helped in identifying genes that showed differential expression in leukemia cancer from normal genes. Study of these genes could have provided a knowledge element to the gene selection process. The data used in the present study was collected from patients with no information regarding the ethnic background or disease history.

5.5 SIGNIFICANCE OF THE STUDY

The present study was successful in demonstrating that a committee of neural networks could carry out a classification considering gene expressions as input

parameters. The study was the first of its kind to make use of a committee with individual networks trained with different sets of input parameters (Figure 3.4). The committee decision could be considered to be more reliable as it was obtained from members with different backgrounds participating in a majority voting scheme. Although no information was present regarding the ethnicity of the patients, the data represented a substantial size of population as it was collected from different hospitals and clinics across the United States.

CHAPTER VI

CONCLUSION

The present study successfully employed leukemia gene expression data for the classifier building effort. Based on the results following were the derived conclusions.

1. The original gene expression profiles, each more than 7000 genes, were successfully processed to identify a subset of 250 genes that could distinguish the classes from each other.
2. The identified gene sets were successfully utilized in designing a series of neural networks in an effort to design a binary and a ternary classifier for leukemia.
3. The trained networks were validated against a set of initial validation data sets to evaluate their performance. The top performing networks were then utilized to form a committee whose decision was based on majority opinion
4. The committee of neural networks was evaluated against a fresh testing data. The binary classification system yielded correct predictions in 100% of the cases while the ternary classification system correctly classified 34 of the 35 validation data sets, yielding an accuracy of 97.14 %.
5. The results of the study rejected the null hypothesis in favor of the alternative hypothesis. Committee neural networks thus give a more reliable and confident evaluation as against individual neural network systems.

REFERENCES

[1] Tsiknakis M., Kafetzopoulos D.; “ERCIM Working Group on Biomedical Informatics”; European Research Consortium for Informatics & Mathematics; 60:15-16; 2005

[2] Zobel R.; “Biomedical Informatics: The opportunity and Challenge for Multidisciplinary Research”; European Research Consortium for Informatics and Mathematics; 60:3; 2005

[3] National Human genome Research Institute, http://www.genome.gov

[4] Orphanoudakis S., Kafetzopoulos D., Tsiknakis M.; “Biomedical Informatics in Support of Individualized Medicine”; European Research Consortium for Informatics and Mathematics; 60:12-14; 2005

[5] Berkum N. L., Holstege F. C.; “DNA microarrays: raising the profile”; Current Opinion in Biotechnology; 12:48-52; 2001

[6] Young R.; “Biomedical Discovery with DNA Arrays”; Cell; 102:9-15; 2000

[7] Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., BloomÞeld C. D., Lander E. S.; “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”; Science; 286:531-537; 1999

[8] Mallick B. K., Ghosh D., Ghosh M.; “Bayesian classification of tumors by using gene expression data”; Royal Statistical Society, 67(2):219-234; 2005

[9] Peng S., Xu Q., Ling X. B., Peng X., Du W., Chen L.; “Molecular Classification of cancer types from microarray data using the combination of

genetic algorithms and support vector machines”; Federation of European Biochemical Societies Letters; 555:358-362; 2003

[10] Antonov A.V., Tetko I.V., Mader M.T., Budczies J., Mewes H.W.; "Optimization models for cancer classification: extracting gene interaction information from microarray expression data"; Bioinformatics; 20:644–652; 2004

[11] Khan J., Wei J. S., Ringnér M., Saal L. H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C. R., Peterson C., Meltzer P. S.; “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks”; Nat Med.; 7(6):673-679; 2001

[12] Berrar D.P., Downes C.S., Dubitzky W.; “Multiclass Cancer Classification using gene Expression Profiling and Probabilistic Neural networks”; Pacific Symposium on Biocomputing; 8:5-16; 2003

[13] Fu L. M., Fu-Liu C. S.; “Multi-Class cancer subtype classification based on gene expression signatures with reliability analysis”; FEBS Letters; 561:186-190; 2004

[14] O’Neill M.C., Li S.; “Neural Network Analysis of lymphoma microarray data: prognosis and diagnosis near perfect”; BMC Bioinformatics; 4:13; 2003

[15] Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J., C., Sabet H., Tran T., Yu X., Powell J., Yang L., Marti G. E., Moore T., Hudson Jr. J., Lu L., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T. C., Weisenburger D. D., Armitage J. O., Warnke R., Levy R., Wilson W., Grever M. R., Byrd J. C., Botstein D., Brown P. O., Staudt L. M.; “Distinct types of large B-cell lymphoma identified by gene expression profiling”; Nature; 403:503-511; 2000

[16] Statnikov A., Aliferis C. F., Tsamardinos I., Hardin D., Levy S.; “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis”; Bioinformatics Oxford University Press; 21:631-643; 2004

[17] Reddy N. P., Buch O.; “Committee Neural Networks for Speaker Verification”; Computer Programs and Methods in Biomedicine; 72:109-115; 2003

[18] Das A., Reddy N. P., Narayanan J.; “Hybrid Fuzzy Logic Committee Neural Networks for Recognition of Swallow Acceleration Signals”; Computer Programs and Methods in Biomedicine; 64:87-99; 2001

[19] Palreddy S., Reddy N. P., Green R., Canilang E. P.; "Neural Networks in Computer-Aided Diagnosis: Classification of Dysphagic Patients"; Proceed. of the 14th Annual International Conference of the IEEE-Engineering in Medicine and Biology Soc. Paris, France, 1517-1518; 1992

[20] Reddy N. P., Prabhu D., Palreddy S., Gupta V., Suryanarayanam S., Canilang E. P.; “Redundant Neural Networks for Medical Diagnosis: Diagnosis of Dysphagia”; Intelligent Engineering Systems through Artificial Neural Networks: Fuzzy Logic and Evolutionary Programming, 5:699-704; 1995

[21] World Health Organization, Cancer, http://www.who.int

[22] Hanahan D., Weinberg R. A.; “The Hallmarks of Cancer”, Cell, 100:57-70; 2000

[23] American Cancer Society: :Information and Resources for Cancer, http://www.cancer.org

[24] Leukemia & Lymphoma Society Website,
URL: http://www.leukeima-lyphoma.org

[25] Leukemia–Diagnosis–Oncochannel,
URL: http://www.oncologychannel.com/leukemias/diagnosis.shtml

[26] Fluorescence In Situ Hybridization, http://www.genome.gov

[27] Fan Y.; “Molecular Cytogenetics: Protocols and Applications”; Humana

Press; 204:311-312; 2003

[28] Hematopathology: Immunophenotyping Lymphomas, http://pleiad.umdnj.edu/

[29] Roche Diagnostics; “Leukemia: When the Color of Blood is fading”; Roche Diagnostics Division; 2005

[30] McLachlan G. J., Do K., Ambroise C.; “Analyzing Microarray gene expression data”; Wiley-Interscience; 2004

[31] Daliakopoulos I. N., Coulibaly P., Tsanis I. K.; “Groundwater level forecasting using artificial neural networks”; Journal of Hydrology; 309:229-240; 2005

[32] Haeri M., Asemani D., Gharibzadeh S; “Modeling of Pain Using Artificial Neural Networks”; Journal of Theoretical Biology; 220:277-284; 2002

[33] Reddy N. P., Rothchild B. M.; “Hybrid Fuzzy Logic-Committee Neural Networks for Classification in Medical Decision Support Systems”; 2nd Joint EMBS-BMES Conference; Houston, Tx; 2002

[34] Pèrez-Roa R., Castro J., Jorquera H., Pèrez-Correa J. R., Vesovic V.; “Air-pollution modeling in an urban area: Correlating turbulent diffusion coefficients by means of an artificial neural network approach”; Atmospheric Environment; 40:109-125; 2006

[35] Vijayabaskar V., Gupta R., Chakrabarti P. P., Bhowmick A. K.; ”Prediction of Properties of Rubber by using Artificial Neural Networks”; Journal of Applied Polymer Science; 100:2227-2237; 2006

[36] Broad Institute Cancer Program Publications
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

[37] Gene – Elsevier
http://www.elsevier.com/wps/find/journaldescription.cws_home/506033/authorinstructions

[38] Samples Information Files Location http://www.broad.mit.edu/mpr/publications/projects/leukemia/table_ALL_AML_samples.rtf

[39] McClintick J. N., Edenberg H. J.; “Effects of filtering by Present call on analysis of microarray experiments”; BMC Bioinformatics; 7:49; 2006

[40] Schuster E. F., Blanc E., Partridge L., Thornton J. M.; “Correcting for sequence biases in present/absent calls”; Genome Biology, 8:R125; 2007

[41] Liu D. W., Chen S. T., Liu H. P.; “Choice of endogenous control for gene expression in nonsmall cell lung cancer”; European Respiratory Journal; 26:1002-1008; 2005

[42] Yang K., Cai Z., Li J., Lin G.; “A stable gene selection in microarray data analysis”; BMC Bioinformatics; 7:228; 2006

[43] Antipova A. A., Tamayo P., Golub T.; “A strategy for Oligonucleotide microarray probe reduction”; Genome Biology; 3:12; 2002

[44] Dudoit S., Fridlyand J., Speed T.; “Comparison of discrimination methods for the classification of tumors using gene expression data”; Journal of the American Statistical Association; 97:77-87; 2002

[45] Entrez Gene, http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

APPENDICES

APPENDIX A

STATISTICAL ANALYSIS FOR NULL HYPOTHESIS

Null Hypothesis (H_0):

The probability p1 of the committee making a correct classification is the same as the probability p2 of the committee making and incorrect classification (p1 = p2)

Alternate Hypothesis (H_1):

The probability p1 of the committee making a correct classification is greater than the probability p2 of the committee making an incorrect classification (p1 > p2)

Apriori Settings: $\alpha = 0.01$

The hypothesis testing is carried out using a Binomial Test

Let n = total samples = 35 x = total misclassifications = 1

p1 = 0.5 p2 = 0.5

n*p1 = 17.5

For $\alpha = 0.01$

The critical region (i.e. region for rejection of null hypothesis) is defined by z values outside the interval -2.3263 and 2.3263

Since n * p1 > x the z value is:

$$z = \frac{((x - 0.5) - n * p1)}{\sqrt{n * p1 * p2}}$$

z = -5.74705

The z value lies in the critical region and hence I fail to accept H_0.

That is,

The probability p1 of the committee making a correct classification is greater than the probability p2 of the committee making an incorrect classification (p1 > p2)

APPENDIX B

LIST OF INFORMATIVE GENES OBTAINED FROM GENE SELECTION

GENE DESCRIPTION	ACCESSION NUMBER	RANK
FTL Ferritin, light polypeptide	M11147_at	91
HISTONE H2A.X	X14850_at	168
Acyl-CoA thioester hydrolase mRNA	U91316_at	223
Azurocidin gene	M96326_rna1_at	75
Chromosome 17q21 mRNA clone LF113	U18009_at	21
RTS beta protein	X67098_at	210
DAGK1 Diacylglycerol kinase, alpha (80kD)	X62535_at	15
MB-1 gene	U05259_rna1_at	20
EUKARYOTIC PEPTIDE CHAIN RELEASE FACTOR SUBUNIT 1	M75715_s_at	92
GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA	U23852_s_at	1
HAES-1 mRNA	X73358_s_at	213
GB DEF = MAL gene exon 4	X76223_s_at	2
GB DEF = T-cell antigen receptor gene T3-delta	X03934_at	3
GB DEF = Erg protein (ets-related gene) mRNA	M21535_at	220
MPO Myeloperoxidase	M19507_at	212
GB DEF = Homeodomain protein HoxA9 mRNA	U82759_at	97
GB DEF = CD36 gene exon 15	Z32765_at	140
P130 mRNA for 130K protein	X76061_at	234
CD19 gene	M84371_rna1_s_at	44
LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)	M57710_at	94
Snk interacting protein 2-28 mRNA	U85611_at	105
TOP2B Topoisomerase (DNA) II beta (180kD)	Z15115_at	102
GB DEF = Partial cDNA sequence, farnesyl pyrophosphate synthetase like-4	Z47055_s_at	112
L-UBC	S81003_at	218
CD36 CD36 antigen (collagen type I receptor, thrombospondin receptor)	M98399_s_at	135
Lysozyme gene (EC 3.2.1.17)	X14008_rna1_f_at	60
RETINOBLASTOMA BINDING PROTEIN P48	X74262_at	192
KIAA0030 gene, partial cds	D21063_at	77

GENE DESCRIPTION	ACCESSION NUMBER	RANK
SMT3B protein	X99585_at	197
HMOX1 Heme oxygenase (decycling) 1	X06985_at	177
Low-Mr GTP-binding protein (RAB32) mRNA, partial cds	U59878_at	122
CD5 CD5 antigen (p56-62)	X04391_at	83
CCND3 Cyclin D3	M92287_at	215
CD2 CD2 antigen (p50), sheep red blood cell receptor	M16336_s_at	33
Serine kinase SRPK2 mRNA	U88666_at	123
KIAA0128 gene, partial cds	D50918_at	70
HMG1 High-mobility group (nonhistone chromosomal) protein 1	D63874_at	202
TIMP2 Tissue inhibitor of metalloproteinase 2	M32304_s_at	80
Guanylate kinase (GUK1) mRNA	L76200_at	139
RANBP1 RAN binding protein 1	D38076_at	166
(clone S31i125) mRNA, 3' end of cds	U61734_s_at	236
HLON ATP-dependent protease mRNA, nuclear gene encoding mitochondrial protein	X76040_at	73
CTPS CTP synthetase	X52142_at	158
HOX 2.2 gene extracted from Human Hox2.2 gene for a homeobox protein	X58431_rna2_s_at	90
Protein tyrosine kinase related mRNA sequence	L05148_at	22
GB DEF = Hypothetical protein downstream of DMPK and DMAHP	Y10936_at	230
Ndr protein kinase	Z35102_at	217
SM22-ALPHA HOMOLOG	D21261_at	65
Glutathione-S-transferase homolog mRNA	U90313_at	239
TCF7 Transcription factor 7 (T-cell specific)	X59871_at	16
SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	J05243_at	23
LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	M16038_at	28
MIC2 Antigen identified by monoclonal antibodies 12E7, F21 and O13	M16279_at	153
POLYPOSIS LOCUS PROTEIN 1	M73547_at	108
Thymopoietin beta mRNA	U09087_s_at	151
Zyxin	X95735_at	19
PRKCQ Protein kinase C-theta	L01087_at	38
FLN1 Filamin 1 (actin-binding protein-280)	X53416_at	201
TRANSCRIPTION FACTOR RELB	M83221_at	127
Inducible protein mRNA	L47738_at	79
CD33 CD33 antigen (differentiation antigen)	M23197_at	36
ENO2 gene for neuron specific (gamma) enolase	X51956_rna1_at	55
Metargidin precursor mRNA	U41767_s_at	131
GB DEF = Chloride channel (putative) 2163bp	Z30644_at	150
ANX6 Annexin VI (p68)	Y00097_s_at	247
Stimulatory Gdp/Gtp Exchange Protein For C-Ki-Ras P21 And Smg P21	HG2036-HT2090_at	190

GENE DESCRIPTION	ACCESSION NUMBER	RANK
Ubiquitin carrier protein (E2-EPF) mRNA	M91670_at	176
LYZ Lysozyme	M19045_f_at	54
RNH Ribonuclease/angiogenin inhibitor	X13973_at	233
GB DEF = (lambda) DNA for immunoglobin light chain	D88270_at	120
MYL1 Myosin light chain (alkali)	M31211_s_at	144
Down syndrome critical region protein (DSCR1) mRNA	U28833_at	84
GB DEF = C8FW phosphoprotein	AJ000480_at	235
EEF1A1 Translation elongation factor 1-alpha-1	Z37987_s_at	206
Lymphoid-restricted membrane protein (Jaw1) mRNA	U10485_at	227
CD47 CD47 antigen (Rh-related antigen, integrin-associated signal transducer)	X69398_at	27
Nuclear Factor Nf-Il6	HG3494-HT3688_at	132
CD3Z CD3Z antigen, zeta polypeptide (TiT3 complex)	J04132_at	13
Thrombospondin-p50 gene extracted from Human thrombospondin-1 gene, partial cds	U12471_cds1_at	85
NB thymosin beta	D82345_at	186
FTH1 Ferritin heavy chain	L20941_at	100
(AF1q) mRNA	U16954_at	113
PROBABLE 26S PROTEASE SUBUNIT TBP-1	M34079_at	109
SERINE/THREONINE PROTEIN PHOSPHATASE 2B CATALYTIC SUBUNIT, BETA ISOFORM	M29551_at	183
IGHM Immunoglobulin mu	X58529_at	81
TXN Thioredoxin	X77584_at	231
Male Enhanced Antigen	HG1869-HT1904_at	101
CTSD Cathepsin D (lysosomal aspartyl protease)	M63138_at	62
Epb72 gene exon 1	X85116_rna1_s_at	107
KIAA0050 gene	D30758_at	48
ADM Adrenomedullin	D14874_at	162
GB DEF = Immunoglobulin mu, part of exon 8	V00563_at	241
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds	U22376_cds2_s_at	64
SELL Leukocyte adhesion protein beta subunit	M15395_at	243
THYMOSIN BETA-10	S54005_s_at	245
Protein tyrosine phosphatase PTPCAAX2 (hPTPCAAX2) mRNA	U14603_at	5
PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR	M80254_at	96
INTEGRIN ALPHA-E PRECURSOR	L25851_at	121
ECGF1 Endothelial differentiation protein (edg-1)	M31210_at	103
KIAA0260 gene, partial cds	D87449_at	219
GB DEF = SLP-76 associated protein mRNA	U93049_at	53
CD22 CD22 antigen	X59350_at	196
CANX Calnexin	D50310_at	211
Protein kinase C substrate 80K-H gene (PRKCSH)	U50327_s_at	149

GENE DESCRIPTION	ACCESSION NUMBER	RANK
Spermidine/spermine N1-acetyltransferase (SSAT) gene	U40369_rna1_at	175
CRYZ Crystallin zeta (quinone reductase)	L13278_at	195
26-kDa cell surface protein TAPA-1 mRNA	M33680_at	114
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891_at	45
T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN PRECURSOR	M23323_s_at	7
TCRB T-cell receptor, beta cluster	X00437_s_at	9
PROBABLE G PROTEIN-COUPLED RECEPTOR LCR1 HOMOLOG	L06797_s_at	63
mRNA, clone HH109 (screened by the monoclonal antibody of insulin receptor substrate-1 (IRS-1))	D23673_at	244
MEF2C MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)	L08895_at	159
RASA1 GTPase-activating protein ras p21 (RASA)	M23379_at	179
Sterol regulatory element binding protein-2 mRNA	U02031_at	184
TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	M31523_at	130
Grb2-associated binder-1 mRNA	U43885_at	154
GB DEF = Fas/Apo-1 (clone pCRTM11-Fasdelta(3,4))	X83490_s_at	232
AARS Alanyl-tRNA synthetase	D32050_at	198
Proteasome subunit HsC7-I	D26599_at	119
PRKAR2B Protein kinase, cAMP-dependent, regulatory, type II, beta	M31158_at	191
KIAA0063 gene	D31884_at	248
GB DEF = Lymphocyte-specific protein tyrosine kinase (LCK) gene, exon 1, and downstream promoter region	M26692_s_at	4
Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)	X04085_rna1_at	126
ADPRT ADP-ribosyltransferase (NAD+; poly (ADP-ribose) polymerase)	J03473_at	52
MXS1 Membrane component, X chromosome, surface marker 1	L10373_at	11
Tetracycline transporter-like protein mRNA	L11669_at	148
T-CELL ANTIGEN CD7 PRECURSOR	M37271_s_at	6
LEPR Leptin receptor	Y12670_at	141
RABAPTIN-5 protein	Y08612_at	164
CHIT1 Chitinase 1	U49835_s_at	42
GPX1 Glutathione peroxidase 1	Y00433_at	146
GB DEF = Selenoprotein W (selW) mRNA	U67171_at	17
TTF mRNA for small G protein	Z35227_at	187
NF-IL6-beta protein mRNA	M83667_rna1_s_at	136
Na,K-ATPase gamma subunit mRNA	U50743_at	14
TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA for Tcell leukemia/lymphoma 1	X82240_rna1_at	157
NATURAL KILLER CELLS PROTEIN 4 PRECURSOR	M59807_at	50

GENE DESCRIPTION	ACCESSION NUMBER	RANK
Protein kinase inhibitor [human, neuroblastoma cell line SH-SY-5Y, mRNA, 2147 nt]	S76965_at	68
BCL7 B cell lymphoma protein 7B	X89985_at	246
ACYLPHOSPHATASE, ORGAN-COMMON TYPE ISOZYME	X84194_at	214
DP2 (Humdp2) mRNA	L40386_s_at	25
GB DEF = Retinoblastoma susceptibility protein (RB1) gene, with a 3 bp deletion in exon 22 (L11910 bases 161855-162161)	L49229_f_at	204
PPGB Protective protein for beta-galactosidase (galactosialidosis)	M22960_at	88
IDH2 Isocitrate dehydrogenase 2 (NADP+), mitochondrial	X69433_at	29
PSAP Sulfated glycoprotein 1	J03077_s_at	66
Tissue specific mRNA	X67698_at	205
Transcriptional activator hSNF2b	D26156_s_at	180
AHNAK AHNAK nucleoprotein (desmoyokin)	M80899_at	222
CD58 CD58 antigen, (lymphocyte function-associated antigen 3)	Y00636_at	143
MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3	D38073_at	163
HKR-T1	S50223_at	199
Very-long-chain acyl-CoA dehydrogenase (VLCAD)	D43682_s_at	82
RPL3 Ribosomal protein L3	L26247_at	228
Cytochrome c oxidase subunit VIII (COX8) mRNA	J04823_rna1_at	216
DF D component of complement (adipsin)	M84526_at	106
Liver mRNA for interferon-gamma inducing factor(IGIF)	D49950_at	74
ALPHA-ACTININ 1, CYTOSKELETAL ISOFORM	M95178_at	249
GBE1 Glucan (1,4-alpha-), branching enzyme 1 (glycogen branching enzyme, Andersen disease, glycogen storage disease type IV)	L07956_at	169
Homologue of yeast sec7 mRNA	M85169_at	160
IRF2 Interferon regulatory factor 2	X15949_at	93
RNS2 Ribonuclease 2 (eosinophil-derived neurotoxin; EDN)	X16546_at	99
IGB Immunoglobulin-associated beta (B29)	M89957_at	125
Lysophospholipase homolog (HU-K5) mRNA	U67963_at	161
Leukotriene C4 synthase (LTC4S) gene	U50136_rna1_at	40
Cytosolic Acetoacetyl-Coenzyme A Thiolase	HG4073-HT4343_at	134
PRG1 Proteoglycan 1, secretory granule	X17042_at	39
CSF1 Colony-stimulating factor 1 (M-CSF)	M37435_at	156
M-PHASE INDUCER PHOSPHATASE 2	S78187_at	78
PRKCD Protein kinase C, delta	D10495_at	129
PROTEASOME IOTA CHAIN	X59417_at	87
Cytoplasmic dynein light chain 1 (hdlc1) mRNA	U32944_at	224
CLPP	Z50853_at	59
LYZ Lysozyme	J03801_f_at	58
GB DEF = ArgBPIB protein	X95677_at	32

GENE DESCRIPTION	ACCESSION NUMBER	RANK
Quiescin (Q6) mRNA, partial cds	L42379_at	142
Histone H2A.2 mRNA	L19779_at	182
INTERLEUKIN-8 PRECURSOR	Y00787_s_at	71
GB DEF = Protein-tyrosine phosphatase mRNA	M68941_at	203
Put. HMG-17 protein gene extracted from Human HMG-17 gene for non-histone chromosomal protein HMG-17	X13546_rna1_at	167
GATA2 GATA-binding protein 2	M68891_at	147
SELL Leukocyte adhesion protein beta subunit	X64072_s_at	111
SMT3A protein	X99584_at	37
BSG Basigin	X64364_at	208
GB DEF = Integrin, alpha subunit	X68742_at	57
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA	U46751_at	61
TCRB T-cell receptor, beta cluster	M12886_at	12
CD3G CD3G antigen, gamma polypeptide (TiT3 complex)	X04145_at	116
Uridine phosphorylase	X90858_at	238
FGFR1 Basic fibroblast growth factor (bFGF) receptor (shorter form)	X66945_at	72
GB DEF = Retinoblastoma susceptibility protein (RB1) L486W 4 bp deletion mutant (resulting in premature stop at amino acid 490) gene, exon 16 (L11910 bases 76983-77136)	L49219_f_at	237
MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3	M69043_at	145
ALDR1 Aldehyde reductase 1 (low Km aldose reductase)	X15414_at	115
PFC Properdin P factor, complement	M83652_s_at	89
Rhodanese	D87292_at	207
Folylpolyglutamate synthetase mRNA	M98045_at	240
Sm protein F	X85372_at	200
ATP6C Vacuolar H+ ATPase proton channel subunit	M62762_at	46
FAH Fumarylacetoacetate	M55150_at	30
GB DEF = Plectin	Z54367_s_at	173
TCRD T-cell receptor, delta	M21624_at	194
MEF2A gene (myocyte-specific enhancer factor 2A, C9 form) extracted from Human myocyte-specific enhancer factor 2A (MEF2A) gene, first coding	U49020_cds2_s_at	225
CD19 CD19 antigen	M28170_at	35
Max gene extracted from H.sapiens max gene	X66867_cds1_at	242
T-CELL DIFFERENTIATION ANTIGEN CD6 PRECURSOR	X60992_at	8
PPP3CA Protein phosphatase 3 (formerly 2B), catalytic subunit, alpha isoform (calcineurin A alpha){alternative products}	L14778_s_at	174
Heterochromatin protein p25 mRNA	U35451_at	98
TYMS Thymidylate synthase	D00596_at	193

GENE DESCRIPTION	ACCESSION NUMBER	RANK
GTF2E2 General transcription factor TFIIE beta subunit, 34 kD	X63469_at	110
GLUTATHIONE S-TRANSFERASE, MICROSOMAL	U46499_at	181
LMNA Lamin A	M13452_s_at	209
LPAP gene	X97267_rna1_s_at	152
Allograft inflammatory factor-1 (AIF-1) mRNA	U19713_s_at	138
CDC25A Cell division cycle 25A	M81933_at	69
MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3	X62153_s_at	189
Epican, Alt. Splice 11	HG2981-HT3127_s_at	226
SNRPN Small nuclear ribonucleoprotein polypeptide N	J04615_at	165
ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)	M81695_s_at	56
CD24 signal transducer mRNA and 3' region	L33930_s_at	229
CR2 Complement component (3d/Epstein Barr virus) receptor 2	M26004_s_at	41
PTPN7 Protein tyrosine phosphatase, non-receptor type 7	D11327_s_at	26
LAMP2 Lysosome-associated membrane protein 2 {alternative products}	L09717_at	172
INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1	L08246_at	76
T-CELL ANTIGEN CD7 PRECURSOR	D00749_s_at	18
HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DR ALPHA CHAIN PRECURSOR	X00274_at	47
ME491 gene extracted from H.sapiens gene for Me491/CD63 antigen	X62654_rna1_at	104
Putative protein kinase C inhibitor (PKCI-1) mRNA	U51004_at	185
DP2 (Humdp2) mRNA	U18422_at	10
Platelet activating factor acetylhydrolase IB gamma-subunit	D63391_at	137
Nucleoside-diphosphate kinase	Y07604_at	86
DAGK4 Diacylglycerol kinase delta	D63479_s_at	188
ARHG Ras homolog gene family, member G (rho G)	X61587_at	155
SPI1 Spleen focus forming virus (SFFV) proviral integration oncogene spi1	X52056_at	133
GB DEF = CD1 R2 gene for MHC-related antigen	X14975_at	31
ADA Adenosine deaminase	M13792_at	24
Oncoprotein 18 (Op18) gene	M31303_rna1_at	51
Pigment epithelium-derived factor gene	U29953_rna1_at	171
Protein kinase C-binding protein RACK7 mRNA, partial cds	U48251_at	118
Interleukin 8 (IL8) gene	M28130_rna1_s_at	124
PLCB2 Phospholipase C, beta 2	M95678_at	117
CD1B CD1b antigen (thymocyte antigen)	M28826_at	49
APLP2 Amyloid beta (A4) precursor-like protein 2	L09209_s_at	95

GENE DESCRIPTION	ACCESSION NUMBER	RANK
GLRX Glutaredoxin (thioltransferase)	X76648_at	178
HFat protein	X87241_at	34
KIAA0084 gene, partial cds	D42043_at	221
YMP mRNA	U52101_at	250
PROBABLE PROTEIN DISULFIDE ISOMERASE ER-60 PRECURSOR	M13560_s_at	128
Platelet-derived endothelial cell growth factor mRNA	S72487_at	170
SRI Sorcin	U64675_at	43
Autoantigen p542 mRNA, 3' end of cds	L38696_at	67