GENE CLUSTERING AND CONSTRUCTION OF INTRA-CLUSTER GENE REGULATORY NETWORK

A thesis

submitted in partial fulfillment of the requirement for the Degree of

Master of Technology in Computer Technology

of

Jadavpur University

By

APARAJITA KHAN

Registration No.: 121256 of 2012-13

Examination Roll No.: M6TCT1518

Under the Guidance of

PROF. MITA NASIPURI

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2015

GENE CLUSTERING AND CONSTRUCTION OF INTRA-CLUSTER GENE REGULATORY NETWORK

A thesis

submitted in partial fulfillment of the requirement for the Degree of

Master of Technology in Computer Technology

of

Jadavpur University

By

APARAJITA KHAN

Registration No.: 121256 of 2012-13

Examination Roll No.: M6TCT1518

Under the Guidance of

PROF. MITA NASIPURI

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2015

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Recommendation

This is to certify that the dissertation entitled "Gene Clustering and Construction of Intra-Cluster Gene Regulatory Network" has been carried out by Aparajita Khan (University Registration No.: 121256 of 2012-13, Examination Roll No.: M6TCT1518) under my guidance and supervision and be accepted in partial fulfillment of the requirement for the degree of Master of Technology in Computer Technology. The research results presented in the thesis have not been included in any other paper submitted for the award of any degree in any other University or Institute.

……

Prof. Mita Nasipuri (Thesis Supervisor)
Department of Computer Science and Engineering
Jadavpur University, Kolkata-32

Countersigned

…….

Prof. Debesh Kumar Das
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-32.

…….

Prof. Sivaji Bandyopadhyay
Dean, Faculty of Engineering and Technology,
Jadavpur University, Kolkata-32.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval*

This is to certify that the thesis entitled "Gene Clustering and Construction of Intra-Cluster Gene Regulatory Network" is a bona-fide record of work carried out by Aparajita Khan in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Technology in the Department of Computer Science and Engineering, Jadavpur University during the period of June 2014 to May 2015. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

……..

Signature of Examiner 1

Date:

……..

Signature of Examiner 2

Date:

*Only in case the thesis is approved

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis entitled "Gene Clustering and Construction of Intra-Cluster Gene Regulatory Network" contains literature survey and original research work undertaken by me, as part of the degree of Master of Technology in Computer Technology.

All information have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Aparajita Khan
Registration No: 121256 of 2012-13
Examination Roll No.: M6TCT1518
Thesis Title: Gene Clustering and Construction of Intra-Cluster Gene Regulatory Network.

…..……..

Signature with Date

Acknowledgement

I would like to start by thanking the holy trinity for helping me deploy all the right resources and for shaping me into a better human being.

I would like to express my deepest gratitude to my advisor, **Prof. Mita Nasipuri**, Department of Computer Science and Engineering, Jadavpur University for her admirable guidance, care, patience and for providing me with an excellent atmosphere for doing research. She initiated me into the discipline of bioinformatics and taught me the important skills of how to approach a huge problem by breaking it apart into small modules to solve them more effectively and how to handle messy and huge data. She continuously kept me motivated towards striving to understand the connection between data, model and biological conclusions. Our numerous scientific discussions and her many constructive comments have greatly improved this work.

Words cannot express my indebtedness to **Dr. Ram Sarkar**, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University for his amazing guidance and supervision. I would like to especially thank him for providing key insights of taking up cluster based approach in this work. He was always there to stop me from quitting by giving me his energetic enthusiasm and leading me to believe in myself. I am deeply grateful to him for the long discussions that helped to enrich the technical content of this manuscript. He guided me throughout starting from the initial steps of understanding the basic biological and computational aspects of the work to the finalized results. Without his enthusiasm, encouragement, support and continuous optimism this thesis would hardly have been completed.

I would like to thank **Prof. Debesh Kumar Das**, Head, Department of Computer Science and Engineering, Jadavpur University, for providing me with moral support at times of need.

I would also like to extend my thanks to **Dr. Sarbani Roy**, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University for her support and guidance to this work. She was always ready to provide me with technical help required for this thesis. She provided key ideas for improving the efficiency of the algorithms and how to analyze the results step by step. I would also like to thank **Dr. Subhadip Basu**, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University for providing me with

the knowledge of several bioinformatics algorithms and access to highly efficient computational resources. In this regard I would also like to mention that I am highly grateful to Mr. Pawan Kumar Singh, for helping me out to understand the statistical aspects of the work. I would also like to extend my thanks to **Mrs. Chitrita Chaudhuri**, Associate Professor, Department of Computer Science and Engineering, Jadavpur University, and **Prof. Mahantapas Kundu**, Department of Computer Science and Engineering, Jadavpur University, for encouraging me in my endeavor to succeed.

This thesis could never have happened without Miss. Ankani Chattoraj my project partner and dearest friend. My research is the result of a close and synergistic collaboration with her, working with whom is an absolute joy and pleasure. She helped me sort out every minute mathematical details of Bayesian network theory and gave me endless support and backing during the toughest and lowest points. I feel privileged to brainstorm with her overnights handling the rigors of the work.

I am also thankful to my friends Mr. Ishan Roy Mr. Avisek Gupta, Miss Ritwika Mandal and Miss. Indrani Pramanik for always being there to support and motivate me. Without their cooperation and company this would have been a very lonely journey.

Most importantly none of this would have been possible without the love and support of my family. I extend my thanks to my parents, especially to my mother whose forbearance and whole hearted support helped this endeavor succeed. My father provided me the strength to never give up and hold on to the work until completion.

Good science is always a joint effort of many people and the research in this thesis is no exception. This thesis would not have been completed without the inspiration and support of a number of wonderful individuals — my thanks and appreciation to all of them for being part of this journey and making this thesis possible.

……..

Aparajita Khan

Registration No: 121256 of 2012-13

Examination Roll No.: M6TCT1518

Department of Computer Science & Engineering

Jadavpur University

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

DNA microarray technology is an emerging biotechnological tool that couples molecular genetics and computer science on a massive scale. This technology provides a fast way for a detailed view of the simultaneous expression of entire genome and provides new insights into gene function and disease physiology. These arrays consist of a highly ordered matrix of thousands of different DNA sequences that can be used to measure DNA and RNA expression levels. So now we are able to produce large amounts of data about many genes in a highly parallel and rapidly serialized manner. The data can then be further analyzed to identify expression patterns and variations that correlate with cellular processes. In this work with the motivation of genome wide expression data analysis, three key issued are addressed. Firstly, identification of genes exhibiting differential gene expression patterns between diseased and non-diseased population, secondly, clustering of genes with differential expression into groups of co-expressed genes and finally construction intra-cluster gene- regulatory network from expression profiles of the genes.

1.1 Scope of Present Work

Elucidating the patterns hidden in gene expression data is a very crucial step towards understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of interpreting the data resulting from microarray experiments. Therefore analysis of gene expression data is a step by step procedure, each of which involves undertaking a computational approach followed by biological interpretation of the results.

In this work a comprehensive analysis of genome wide expression profile 187 smokers with suspect of lung cancer amongst whom 90 are healthy smokers while rest 97 are smokers diagnosed with lung cancer is performed. To accomplish this, the entire work is subdivided into the following parts.

Identification of Differentially Expressed Genes among the groups of healthy and diseased population

Given the genome wide expression profile from cancerous cell or tissue of two groups of population, one group being healthy population and the other being the diseased population, the goal of this module is to identify differentially expressed genes using statistical significance tests. These are genes whose expression levels are significantly different between two groups of the experiment. These genes are relevant for discovering dis-regulated genes, potential pharmaceutical targets and diagnostic or prognostic markers. The approach used in this module includes testing each gene against the null hypothesis of showing no expression change between diseased and healthy population. This is followed by selecting genes that reject this null hypothesis with high degree of confidence. The significance level of a gene is assessed using p-values and q-values.

Cluster Analysis to identify co-expressed genes

Genes exhibiting similar expression profile are called co-expressed genes and are assigned to same cluster by clustering algorithms. Genes belonging to the same cluster are typically involved in related functions and are frequently co-regulated. Thus, grouping similar genes can provide a way to understand some of their unknown functions. After filtering out a set of differentially

expressed genes, clustering analysis is used groups the genes or samples into "clusters" based on similar expression profiles which provides clues to the functional similarity of genes via shared cluster membership. For this work hierarchical clustering algorithm is used to first cluster genes and then dissects the linkages of the resultant dendrogram to find natural non-overlapping clusters of the dataset. The clustering result differs when performed on the healthy and diseased dataset in terms of number of gene clusters present. Biological enrichment of gene clusters is addressed using Gene Ontology based terms shared between genes belonging to the same cluster.

Intra-Cluster Genetic Regulatory Network Modeling

A gene regulatory network (GRN) is a set of genes that interact with each other to control a specific cell function. In these networks the nodes typically represent genes or gene products and edges represent regulatory relationships between them. Sharing of the regulatory mechanism among genes at the sequence level, in an organism, is predominantly responsible for them being co-expressed. Genes having similar gene expression profiles are more likely to regulate one another or be regulated by some other common parent gene. Constructing GRN in healthy and diseased tissues is critical to understand cancer phenotypes, devising effective therapeutics and prioritizing drug targets. In the third module, the objective is to construct GRN between genes belonging to the same cluster since they may have regulatory relationships between them. Construction of GRN from the gene expression profile is accomplished using graph theoretic approach of Bayesian Networks. For GRN construction we use the sparse candidate algorithm for learning Bayesian network. We also propose a modification to the algorithm which results in a more extensive state space search for inferring the optimal Bayesian network structure that the dataset encodes.

1.2 Organization of Thesis Work

The thesis work is organized as follows,

Chapter 2 describes the basic biological concepts required for the study and analysis of genomic data. This includes the description of the structure of DNA, fundamental principles of molecular biology and mechanisms of gene expression and control. This chapter also provides a brief

overview on the microarray technology, which is the prime means for generating gene expression data.

Chapter 3 introduces the lung cancer gene expression dataset considered for this study. This chapter also describes the biological database consulted and tools used to select only those genes that are reported to be mutated in lung cancer.

Chapter 4 discusses the first module of this work concerning selection of significant genes from genome wide expression data of two subsets of population. It describes the problem under consideration, the related works, statistical approach undertaken to address the problem and the resultant set of significant genes filtered out.

Chapter 5 describes the second module of this work which concerns the problem of grouping genes with similar expression profile together. This chapter elaborately describes the clustering approach used for grouping the genes together and the mathematical formulations of the cluster validity criterion to determine the number of clusters in dataset. The module results in different gene clusters for healthy and diseased dataset whose functional enrichment is reported from gene ontological study.

Chapter 6 details the construction of gene regulatory network, which is the final module of this work. Here the formalisms of Bayesian network approach and the learning algorithms used to reconstruct the intra-cluster GRN are discussed. The results of this module are the network topology of interacting genes and validation of the interactions from biological literature.

Chapter 7 concludes the thesis by summarizing the entire work and the results from each module. It also sheds light on the future directions of the work.

Chapter 2

Biological Background

Genomics is the branch of science that studies the genome - the genetic material, or blue print, of a human or other species (animal, plant, and microbe) that is contained in its DNA - to better understand the workings of the organism, and what happens when certain genes interact with each other and the environment. Genomics is the study of complex sets of genes, how they are expressed in cells (what their level of activity is), and the role they play in biology. Molecular biology aims at identifying the genes and the functions of their products. While, systems biology aims at system level understanding of the biological systems. These includes understanding the components and the structure of the system, like genes and proteins and interactions among them, modeling the dynamics of such systems and finally develop methods to control and modify such systems for desired properties. Computational techniques are often applied to analyze the function, expression and interaction of genes to model the dynamics of a biological process. This chapter introduces the basic biological concepts required for the study and analysis of genomic data.

2.1 The Genetic Code

In all living cells, from unicellular bacteria to multicellular plants and animals, DNA or deoxyribonucleic acid [1] is the material in which genetic instructions are stored and transmitted from generation to generation. Proteins [2] are manufactured using the information encoded in DNA and are the molecules that direct the actual processes on which life depends. Processes essential for life, such as energy metabolism, biosynthesis, and intercellular communication, are all carried out through proteins. A gene is the information in DNA that directs the manufacture of a specific protein or RNA molecular form.

2.1.1 The Structure of DNA

DNA is a linear, un-branched polymer in which the monomeric subunits are four chemically distinct nucleotides that can be linked together in any order in chains of hundreds, thousands, or even millions of units in length. Each nucleotide in a DNA polymer is made up of three components (Figure 1.1):

- $2'$-Deoxyribose, which is a pentose, a type of sugar composed of five carbon atoms. These five carbons are numbered $1'$, $2'$,and so on. $2'$-deoxyribose, indicates that this particular sugar is a derivative of ribose, one in which the hydroxyl (–OH) group attached to the $2'$-carbon of ribose has been replaced by a hydrogen (–H) group.
- A nitrogenous base, one of 4 bases, that includes cytosine(C), thymine(T) (single-ring pyrimidines), adenine(A), or guanine(G) (double-ring purines). The base is attached to the $1'$-carbon of the sugar by a β-N-glycosidic bond attached to nitrogen number 1 of the pyrimidine or number 9 of the purine.
- A phosphate group, comprising one, two, or three linked phosphate units attached to the $5'$-carbon of the sugar. The phosphates are designated α, β and γ, with the α-phosphate being the one directly attached to the sugar.

A molecule made up of just the sugar and base is called a nucleoside; addition of the phosphates converts this to a nucleotide.

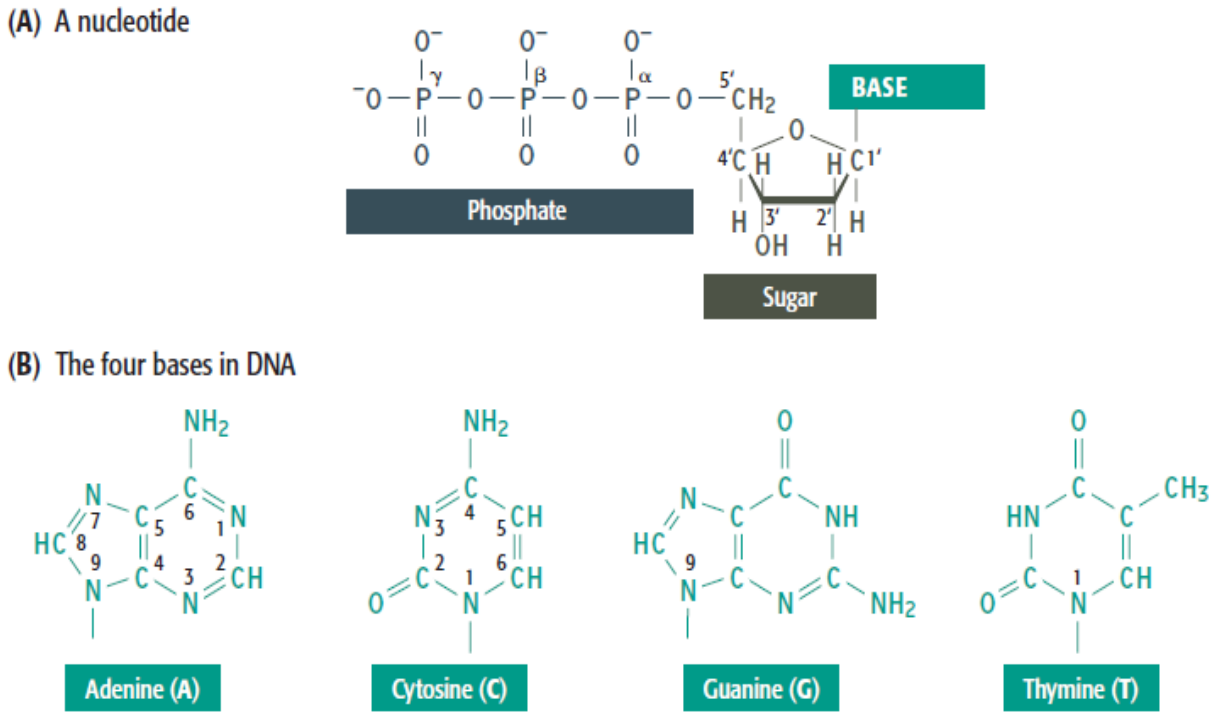

Figure 1.1 (A) General structure of a deoxyribonucleotide. (B) The four bases that occur in DNA. (Image Courtesy: *Genomes 3* [1])

The nucleotide monomers are linked together by joining the phosphate group, attached to the 5′-carbon of one nucleotide, to the 3′-carbon of the next nucleotide in the chain. Normally a polynucleotide is built up from nucleoside triphosphate subunits, so during polymerization the b and g phosphates are cleaved off. The hydroxyl group attached to the 3′-carbon of the second nucleotide is also lost. The linkage between the nucleotides in a polynucleotide is called a **phosphodiester** bond,(Figure 1.2) "phospho-" indicating the presence of a phosphorus atom and "diester" referring to the two ester (C–O–P) bonds in each linkage.

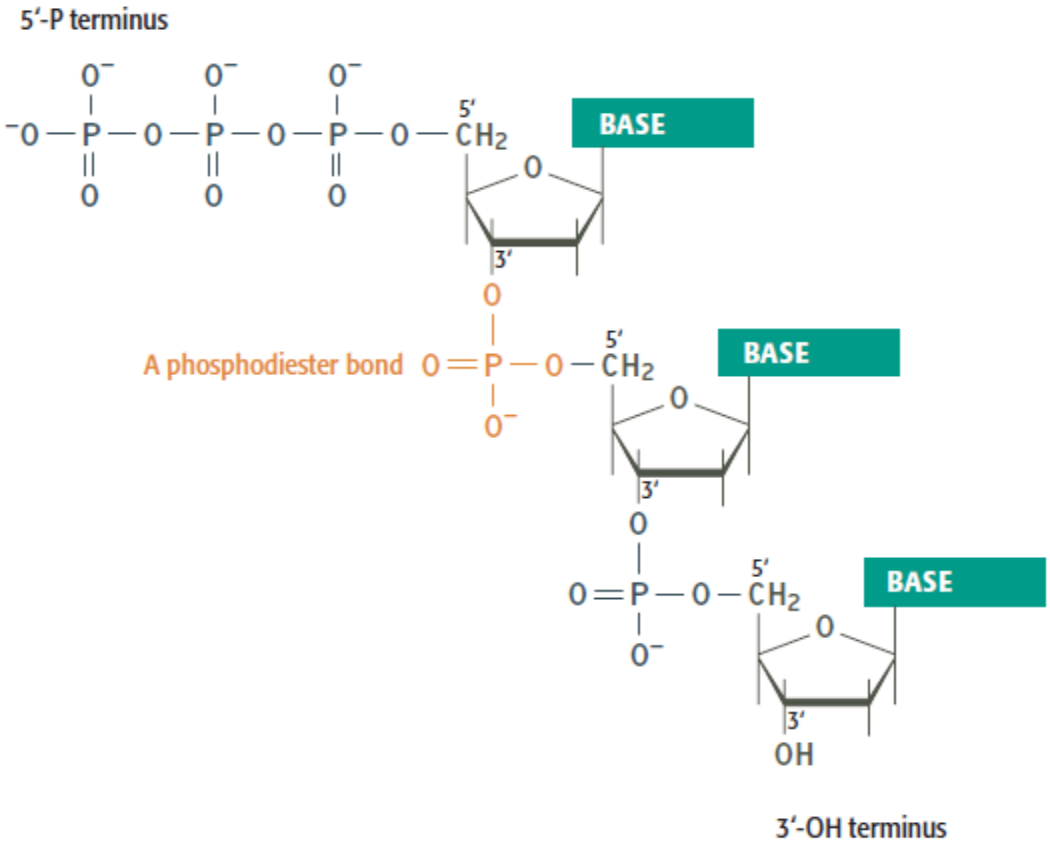

Figure 1.2 Polynucleotide structure with phosphodiester bonds. (Image Courtesy: *Genomes*[1])

The two ends of the polynucleotide chain are chemically distinct. One having an unreacted triphosphate group attached to the $5'$-carbon (the $5'$ or $5'$-P terminus), and the other having an unreacted hydroxyl attached to the $3'$-carbon (the $3'$ or $3'$-OH terminus). Thus the polynucleotide has a chemical direction, expressed as $5' \rightarrow 3'$ or $3' \rightarrow 5'$. An important consequence of the polarity of the phosphodiester bond is that the chemical reaction needed to extend a DNA polymer in the $5' \rightarrow 3'$ direction is different to that needed to make a $3' \rightarrow 5'$ extension. All natural DNA polymerase enzymes are only able to carry out $5' \rightarrow 3'$ synthesis, which adds significant complications to the process by which double-stranded DNA is replicated.

DNA is organized as a right-handed double helix structure. The two strands run in opposite directions (Figure 1.3). The helix is stabilized by two types of chemical interaction,

- **Base pairing** between the two strands involves the formation of hydrogen bonds between an adenine on one strand and a thymine on the other strand, or between a cytosine and a guanine. These are the only permissible pairs partly because of the geometries of the

nucleotide bases and the relative positions of the atoms that are able to participate in hydrogen bonds, and partly because the pair must be between a purine and a pyrimidine: a purine–purine pair would be too big to fit within the helix, and a pyrimidine–pyrimidine pair would be too small.

- **Base stacking**, also called $\pi - \pi$ interactions, which involves hydrophobic interactions between adjacent base pairs thus adding stability to the double helix structure once the strands have been brought together by base pairing. These hydrophobic interactions arise because the hydrogen-bonded structure of water forces hydrophobic groups into the internal parts of a molecule.

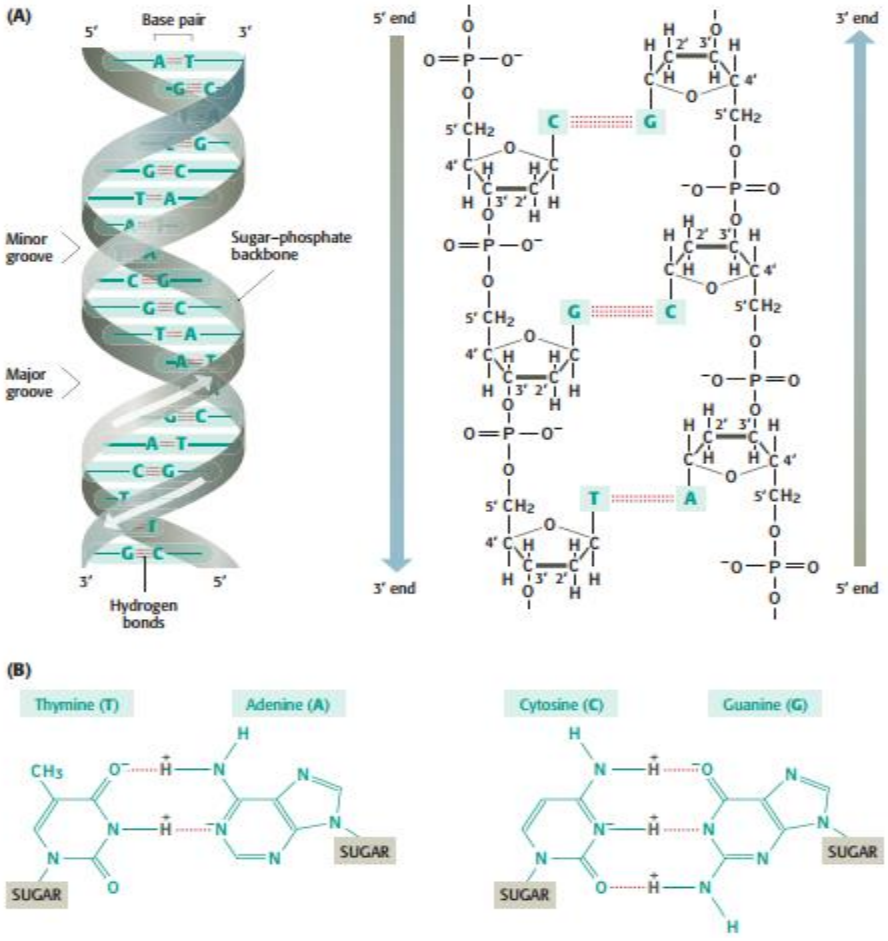

Figure 1.3 (A) The Double Helix structure of DNA. (B) A Base-Pairs with T and G Base-Pairs with C. (Image Courtesy: *Genomes*[1])

The entire information required to build and maintain a human being is contained in just 23 pairs of DNA molecules, comprising the chromosomes of the human genome. These

molecules are amongst the largest and longest known, the smallest having 47 million bases and the largest 247 million bases, with the entire human genome being composed of approximately 3 billion bases. Even bacterial genomes, which are much smaller than this, tend to have several million bases. The DNA of each chromosome encodes hundreds to thousands of proteins, depending on the chromosome, each of these protein specified by a distinct segment of DNA called **Gene**. Hence this segment being the gene for that protein. In general, a gene also includes surrounding regions of noncoding DNA that act as control regions.

2.2 The Central Dogma of Molecular Biology

The key relationship between DNA, RNA, and the synthesis of proteins, is often referred to as the central dogma of molecular biology [3] (Figure 1.4). According to Crick [6], there is essentially a single direction of flow of genetic information from the DNA, which acts as the information store, through RNA molecules from which the information is translated into proteins. This basic scheme generally holds for all known forms of life.

The sequence of bases in the DNA of a gene specifies the sequence of amino acids in a protein chain. The conversion does not occur directly, however. After a signal to switch on a gene is received, a single-stranded RNA copy of the gene is first made in a process called **transcription**. Transcription is essentially similar to the process of DNA replication, except that only one of the DNA strands acts as a template in this case, and the product is RNA not DNA. RNA synthesis is catalyzed by enzymes called RNA polymerases, which, like DNA polymerases, move along the template, matching incoming ribo-nucleotides to the bases in the template strand and joining them together to make an RNA chain. Only the relevant region of DNA is transcribed into RNA, therefore the RNA is a much smaller molecule than the DNA it comes from. So while the DNA carries information about many proteins, the RNA carries information from just one part of the DNA, usually information for a single protein. RNA transcribed from a protein-coding gene is called messenger RNA (mRNA) and it is this molecule that directs the synthesis of the protein chain, in the process called **translation**. When a gene is being transcribed into RNA, which is in turn directing protein synthesis, the gene is said to be expressed.

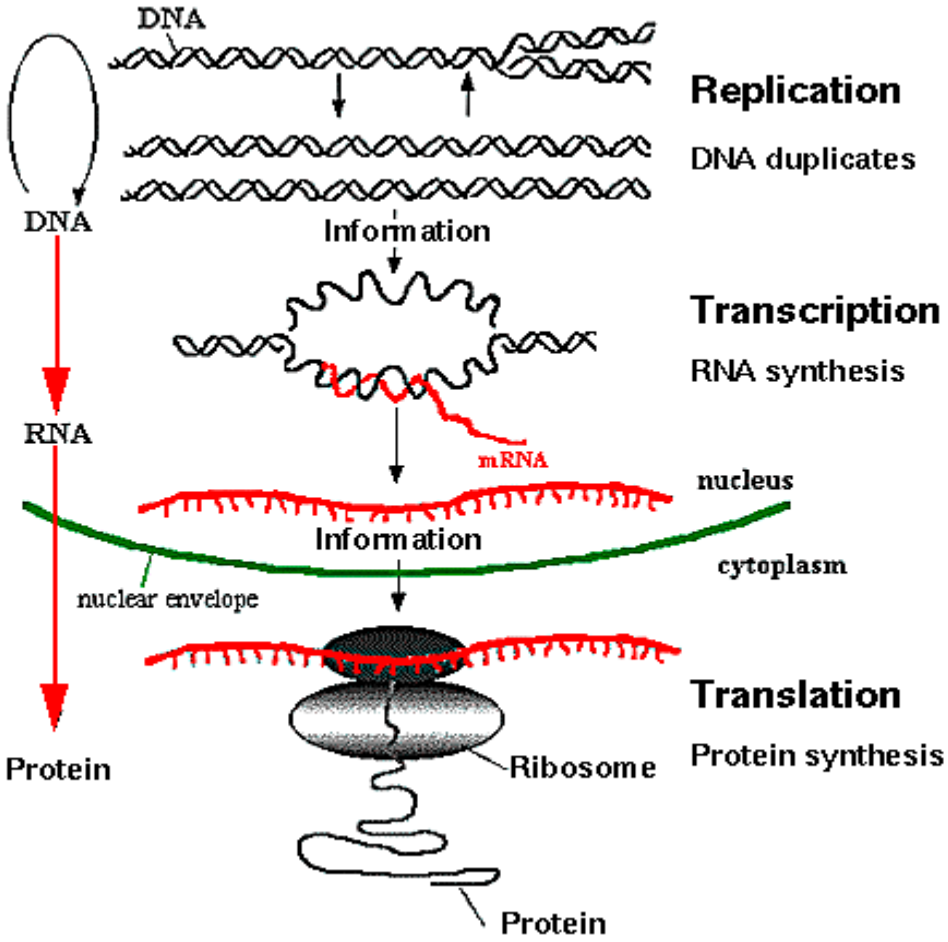

Figure 1.4 The Central Dogma of Molecular Biology. (Image Courtesy: www.bioinformatics.nl/webportal/background/translationinfo.html)

The genetic code refers to the rules governing the correspondence of the base sequence in DNA or RNA to the amino acid sequence of a protein. A code of four different bases in nucleic acids can specify proteins made up of 20 different types of amino acids. Each amino acid is encoded by a set of three consecutive bases. The three-base sets in RNA are called **codons**. The **mRNA** (messenger RNA) produced by transcription is translated into protein by ribosomes, large multimolecular complexes formed of rRNA and proteins. Amino acids do not recognize the codons in mRNA directly and their addition in the correct order to a new protein chain is mediated by the **tRNA** (transfer RNA) molecules, which transfer the amino acid to the growing protein chain when bound to the ribosome. These small tRNA molecules have a three-base anticodon at one end that recognizes a codon in mRNA, and at the other end a site to which the corresponding amino acid becomes attached by a specific enzyme. This system is the physical basis for the genetic code.

2.3 Control of Gene Expression

An organism's DNA encodes all of the RNA and protein molecules that are needed to make its cells. Yet a complete description of the DNA sequence of an organism—be it the few million nucleotides of a bacterium or the few billion nucleotides in each human cell—does not enable us to reconstruct the organism any more than a list of all the possible codes. This is due to ***gene expression***[4].

The term *gene expression* is used to describe the transcription of genetic information contained within the Deoxyribonucleic Acid (DNA) into messenger RNA (mRNA) molecules that are later translated into proteins. Even the simplest single-celled bacterium can use its genes selectively—for example, switching genes on and off to make the enzymes needed to digest whatever food sources are available. And, in multicellular plants and animals, gene expression is under even more elaborate control. Hundreds of different cell types carry out a range of specialized functions that depend upon genes that are only switched on in that cell type: for example, the C cells of the pancreas make the protein hormone insulin, while the B cells of the pancreas make the hormone glucagon; the lymphocytes of the immune system are the only cells in the body to make antibodies, while developing red blood cells are the only cells that make the oxygen-transport protein hemoglobin. The differences between a neuron, a lymphocyte, a liver cell, and a red blood cell depend upon the precise control of gene expression.

2.3.1 Transcription Control

Promoter is a region of DNA which initiates transcription of a particular gene by attracting RNA polymerase enzyme. The promoters consists an initiation site, where transcription actually begins, and a sequence of nearly 50 nucleotides that extends upstream from the initiation site. Moreover nearly all genes have **regulatory DNA sequences** that are used to switch the gene on or off. Some regulatory DNA sequences are as short as 10 nucleotide pairs and act as simple gene switches that respond to a single signal while other regulatory DNA sequences could be very long (sometimes more than 10,000 nucleotide pairs) and act as molecular microprocessors, integrating information from a variety of signals into a command that dictates how often transcription should be initiated. Regulatory DNA sequence has to be recognized by proteins called **transcription regulators**, which bind to the DNA. It is the combination of a DNA

sequence and its associated protein molecules that acts as the switch to control transcription. The simplest bacterium codes for several hundred transcription regulators, each of which recognizes a different DNA sequence and thereby regulates a distinct set of genes. Humans make many more—several thousand—signifying the importance and complexity of this form of gene regulation in producing a complex organism.

2.3.2 Repressor and Activator Proteins

A repressor protein in its active form, switches genes off, or represses them. For example, tryptophan repressor is a transcription regulator that represses production of the tryptophan-producing enzymes. Within the promoter is a short DNA sequence (15 nucleotides in length) that is recognized by a transcription regulator. When this protein binds to this nucleotide sequence, termed the operator, it blocks access of RNA polymerase to the promoter; this prevents transcription of the operator and production of the tryptophan-producing enzymes.(Figure 1.5)

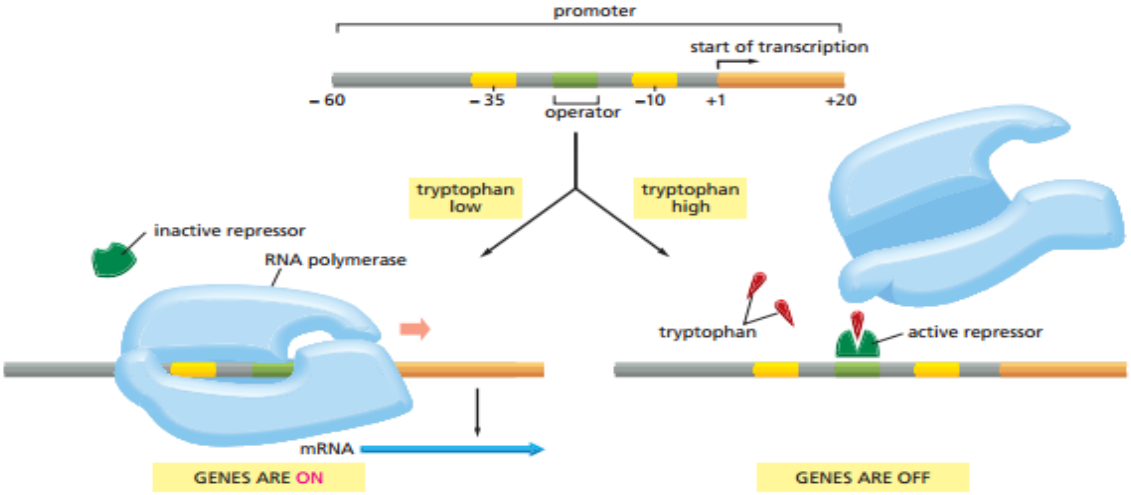

Figure 1.5 Mechanism of Transcription Control by Repressor Proteins. (Image Courtesy: *Essential cell biology* [4])

While, activator proteins often have to interact with a second molecule to be able to bind DNA. An activator protein (Figure 1.6) binds to a regulatory sequence on the DNA and then interacts with the RNA polymerase to help it initiate transcription. Without the activator, the promoter fails to initiate transcription efficiently.

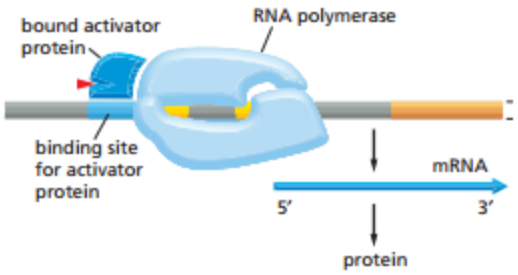

Figure 1.6 Mechanism of Transcription Initiation by Activator Proteins. (Image Courtesy: *Essential cell biology* [4])

2.4 Microarray Technology

In functional genomics large datasets of information derived from various biological experiments are analyzed. These large-scale experiments involve monitoring the expression levels of thousands of genes simultaneously under a particular condition, known as **gene expression analysis**. Microarray technology[5] has made this possible. The quantity of data generated from each such experiment is enormous, much larger the amount of data generated by genome sequencing project.

Microarray technology is a biotechnological used to monitor genome wide expression levels of genes of the organism under study. A microarray typically consists of a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called **spots** (Figure 1.6). A single microarray chip may contain several thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a particular gene. The DNA at a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to the specific gene for that spot.

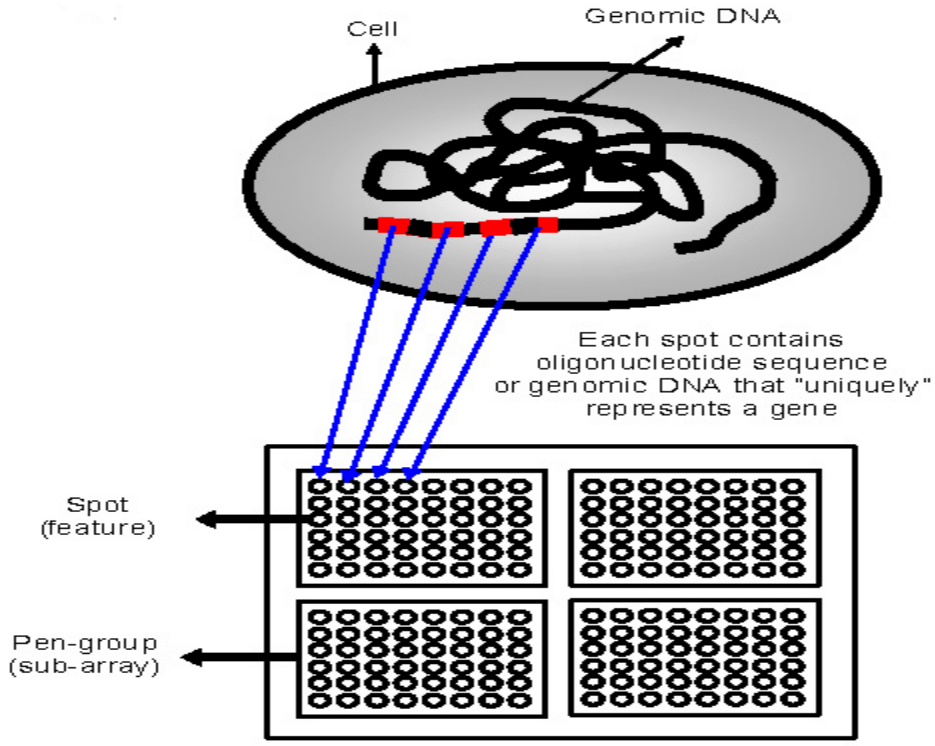

Figure 1.7 Microarray chip containing thousands of spots. (Image Courtesy: *Introduction to microarray data analysis* [5])

Microarray Technology works on the basis of **DNA hybridization**, the process by which a DNA strand binds to its unique complementary strand. One way of measuring gene expression is comparing expression of a set of genes from a cell maintained in a particular condition (say condition A) with that same set of genes from a reference cell which is maintained under normal condition (say condition B). RNA is first extracted from the cells and then the RNA molecules in the extract are reverse transcribed into cDNA using reverse transcriptase enzyme and the nucleotides are labeled with different fluorescent colored dyes. For instance, a red dye may be used to label cDNA from cells under condition A while cDNA from cells under condition B may be stained with a green dye. The samples after being differentially labeled are allowed to hybridize onto the same glass slide. Now, any cDNA sequence in the sample will hybridize to specific spots on the glass slide containing its complementary sequence and the cDNA amount bounded to a spot would be directly proportional to the initial number of RNA molecules present for that gene in both samples. After hybridization is over, the spots in the hybridized microarray are excited by a laser and scanned at suitable wavelengths to detect the red and green dyes. The

amount of fluorescence emitted upon excitation corresponds to the amount of bounded nucleic acid. For instance, if cDNA from condition A for a particular gene was in greater abundance than that from condition B, one would find the spot to be red. While for the other way, the spot would be green. If a particular gene was expressed to the same extent in both conditions, the spot would be yellow, and if the gene was not expressed in both conditions, the spot would be black. Thus the final product of the experiment is an image of the microarray, in which each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene Figure(1.8).

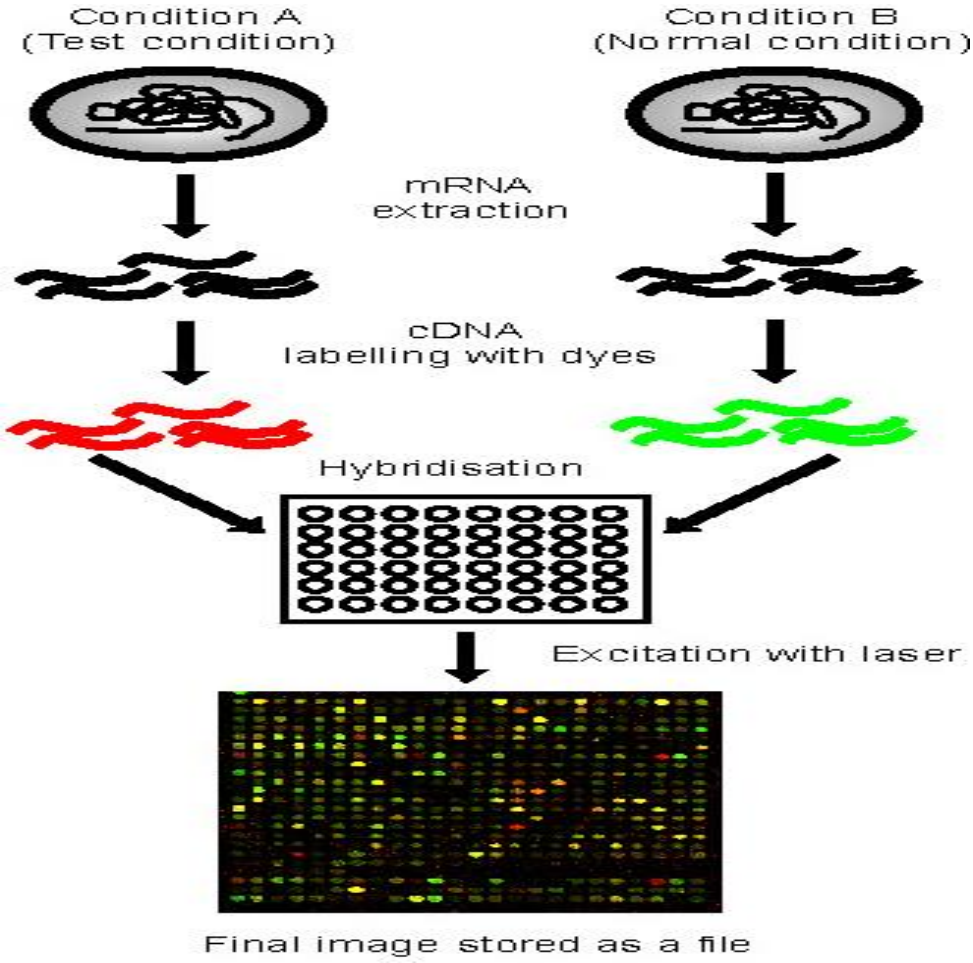

Figure 1.8 Schematic view of a Microarray experiment. (Image Courtesy: *Introduction to microarray data analysis* [5])

2.5 Gene Expression Data Representation

Gene expression data can be represented in several ways, some of them are, ***absolute measurement***, each cell in the matrix represents the expression level of the gene in abstract units. ***Relative measurement / expression ratio measurement*** where the expression level of a gene in abstract units is normalized with respect to its expression in a reference condition. ***Log ratios*** $\log_2(\text{expression ratio})$ measurement of gene expression captures up and down regulation of genes in a symmetric manner. For example, 4-fold up-regulation maps to $\log_2(4) = 2$ and a 4-fold down-regulation maps to $\log_2\left(\frac{1}{4}\right) = -2$. In discreet value representation, say for a binary expression matrix of 1 and 0 where 1 could mean that the gene is expressed above a user defined threshold, while 0 means that the gene is expressed below this threshold. In ***representation of expression profiles*** as vectors, after the individual cells in the gene expression matrix have been represented, expression profile of a gene or a sample can be thought of as a vector and can be represented in vector space. The Expression profile of a gene can be considered as a vector in n dimensional space where n corresponds to the number of conditions, and an expression profile of a sample with m genes can be considered as a vector in m dimensional space where m is the number of genes and a sample can corresponding to a particular organism. The gene expression matrix X with m genes across n conditions is considered to be an $m$xn matrix, where the expression value for gene i in condition j is given by, x_{ij}. $X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$. The expression profile of a gene i can be represented as a row vector, $G_i = (x_{i1}, x_{i2,\ \ldots}, x_{in})$ and expression profile of a sample j can be represented as a column vector, $S_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{pmatrix}$.

Chapter 3

Dataset Description and Gene Subset Selection

The entire functionality of a normal cell is achieved through a wide variety of controlled interactions involving RNA and DNA molecules with proteins. Gene expression is one of the most tightly controlled processes in the body requiring strict regulation to ensure that cells produce the correct amount of proteins when they need them. Any disruption to this regulation can lead to serious consequences, including cancer. Cancer develops due to genetic damage to DNA and other epigenetic changes. These changes affect the normal functions of the cell, including cell proliferation, programmed cell death (apoptosis) and DNA repair.

Lung cancer is the uncontrolled growth of abnormal cells in one or both of the lungs. While normal cells reproduce and develop into healthy lung tissue, these abnormal cells reproduce faster and never grow into normal lung tissue. Lumps of cancer cells (tumors) then form and grow. Other than interfering with how the lung functions, cancer cells can spread from the tumor into the bloodstream or lymphatic system where they can spread to other organs. Cigarette smoking is by far the most important cause of lung cancer, and the risk from smoking increases with the number of cigarettes smoked and the length of time spent smoking [19]. Lung cancer has been estimated to be the most common cancer in the world for a number of decades. In 2008, there were an estimated 1.61 million new cases of lung cancer worldwide, accounting

for almost 13% of the total new cancer cases, and 1.38 million deaths [15] [16]. Lung cancer has the largest proportion of cases caused by smoking: According to a recent estimate, in the UK about 85% of lung cancer cases in men are attributable to smoking (excluding environmental tobacco smoke) and about 80% of cases in women [17] [18]. From the subset of smokers who develop lung cancer, it is not easy to determine which smokers are at highest risk for cancer development. Damage caused by cigarette smoke is not limited solely to the lung but rather forms a "field of injury" throughout the entire respiratory tract [20] [21]. Tissue from this extended injured area can be used to glean clinically relevant information about smoking-induced damage and disease. Gene expression pattern variations in healthy and diseased smokers can help understand the signaling pathways that are deregulated at an early stage of lung cancer. Analysis of profile of genes that show different expression patterns in healthy smokers and lung cancer affected smokers could provide insight into smoking induced damage.

Cancer is a disease involving dis-regulation of multiple pathways governing fundamental cell processes such as death, proliferation, differentiation and migration. Thus, the activities of molecular networks that execute metabolic or cytoskeletal processes, or regulate these by signal transduction, are altered in a complex manner by diverse genetic mutations in concert with the environmental context. A major challenge therefore is how to develop actionable understanding of this multivariate dis-regulation. Traditional methods of wet lab experiments to identify dis-regulated genes and their impact and associations on other genes incurs high cost. On the other hand, availability of high throughput data from DNA microarray experiments has made it possible for monitoring expression levels of thousands of genes simultaneously. Computational methods of analysis of cancer gene expression data to identify dis-regulated genes, groups of co-expressed genes and their regulatory relationships will help infer the mechanisms of carcinogenesis from gene expression profile data and is an active area of research. In [10] authors provide an extensive review on advances in studying cancer-associated genes from a systems biology point of view.

3.1 Expression Dataset under Study

In the study [22][23] a gene expression based approach explores patterns of pathway deregulation in cytologically normal airway epithelial cells from patients with and without lung cancer. The authors published a whole genome wide expression data under Gene Expression

Omnibus GEO [24] reference series GSE4115 [25] constituting of patients undergoing flexible bronchoscopy for suspicion of lung cancer. These are the patients who are current or former smokers under suspicion of lung cancer who were undergoing diagnostic flexible bronchoscopy from four institutions [Boston University Medical Center, Boston Veterans Administration, Lahey Clinic, and St. James's Hospital (Dublin, Ireland)] and a fifth medical center (St. Elizabeth's Hospital, Boston, MA). The results of the study suggest that deregulation of the PI3K pathway in the bronchial airway epithelium of smokers is an early, measurable, and reversible event in the development of lung cancer.

The gene expression dataset related to this study bearing GEO Accession Number **GDS2771** [26] consists of genome wide expression levels of 192 smokers with suspect of lung cancer. 90 of the smokers have not been diagnosed with cancer while 97 patients are diagnosed with lung cancer and there are 5 sample patients who have been diagnosed with threat of cancer. Thus this GEO dataset GDS2771 reports genome wide expression values for 2 groups of population, one is the group of healthy smoker (90 samples) and other is the group of smokers diagnosed with lung cancer (97 samples). Genome wide expression profile of these two groups was considered for study in this thesis work.

3.2 Gene Subset Selection

The GEO gene expression dataset GDS2771 reports genome wide expression profile of 192 samples. Gene expression values of a total of 22215 genes corresponding to each samples is reported in transformed counts. To select the subset of cancer associated genes for further study, information of genes from the following database was utilized.

3.2.1 COSMIC

All cancers arise as a result of the acquisition of a series of fixed DNA sequence abnormalities, each of which ultimately confers growth advantage upon the clone of cells in which it has occurred. These abnormalities include base substitutions, deletions, amplifications and rearrangements. These genetic alteration acquired by a cell are called somatic mutations that can be passed to the progeny of the mutated cell in the course of cell division. The extent to which each of these mechanisms contributes to cancer varies markedly between different genes, and

probably also between different cancer types. Identification of the genes that are mutated in cancer is a central aim of cancer research.

COSMIC, the **Catalogue Of Somatic Mutations In Cancer** (http://cancer.sanger.ac.uk) [27] is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. The latest release of the database (v70; Aug 2014) describes 2002811 coding point mutations in over one million tumor samples and across most human genes manually curated from scientific literature. The "**Cancer Browser**" tool of COSMIC (http://cancer.sanger.ac.uk/cosmic/browse/tissue) provides helpful access to over 2500 cancer disease classifications. After selection of a disease this tool provides list of all genes reported to be mutated for that disease along with the count of the number of samples that report mutation of that gene and the total number of samples tested. From this data the entire list of genes reported to be somatically mutated for a cancer along with its corresponding mutation percentage can be obtained.

The list of gene reported to be somatically mutated in lung cancer by COSMIC is considered to be the subset of genes for rest part of the thesis. COSMIC reports a total of 24283 genes including alias to be mutated in lung cancer. Amongst this entire list gene expression profile corresponding to 11237 genes were found in the dataset GDS2771 which were considered for further study.

Chapter 4

Identification of Differentially Expressed Genes

Microarray technology has enabled the simultaneous measurement of the expression levels of genes throughout the genome. In [73], authors define differentially expressed genes as gene data determined to be statistical outliers from some standard state, and which cannot be ascribed to chance or natural variability. Use of microarray technology to identify genes, which are differentially expressed between two or more groups of patients, has many biomedical applications; including the identification of disease biomarkers that can potentially be used to understand and diagnose diseases in a better way. An inherent characteristic of microarray data is the presence of high levels of noise, high cardinality of genes, and small samples size. Therefore, selection of significant genes that truly represent the biological process or phenotype under study is an important and frequently used technique in gene expression analysis. Since the small sample count narrows down the acquirable knowledge, it reduces the probability of correct decision making. Working with a huge number of genes leads to increased processing time and memory requirements and more importantly wrong conclusions as a large subset of genes may not be representative of the microarray experiment. Therefore, a selection of the proper gene set is an important preprocessing stage for further gene expression analysis and is often directly associated with tissue category, disease state or clinical outcome.

The lung cancer dataset [26] under study consists of genome wide expression profile of two sets of population, one being healthy and the other being cancer affected. The objective of this module is to identify genes exhibiting differential expression patterns amongst the two classes of population. These genes are relevant for discovering potential pharmaceutical targets and diagnostic or prognostic markers. In this module standard statistical measure [7] of p – values corresponding to Welch's t-test is first employed for every individual gene. However with this approach the probability that a false identification (type I error) is committed increases sharply when the number of tested genes gets large. Correlation between the tests attributed to gene co-regulation and dependency in the measurement errors of the gene expression levels further complicates the problem. This problem is addressed in the current approach by adopting the false discovery rate (FDR) and q-values as measure for identifying differentially expressed genes. Both p-value and q-value gives each gene its own individual measure of significance. p-value is a measure of significance in terms of false positive rate, and q-value is a measure in terms of FDR. In the context of gene expression analysis, control of the FDR to identify differentially expressed genes means that if in reality no genes are differentially expressed and the FDR is controlled at some level of q, then the probability of erroneously detecting any differentially expressed genes is less than or equal to q. FDR already takes into account for multiple genes being tested simultaneously and is a better measure than p-value for significant gene analysis. The experimental results reported in this module support this conclusion. Finally we use a q-value cut off of 0.005 to identify a list of 168 genes differentially expressed between the two populations of healthy and diseased population.

4.1 Related Works

There exist several methods for the identification of such differentially expressed genes, and the choice of a method can profoundly affect the resultant set. One of the very earliest methods includes analysis of fold change [72, 73] which is still popular among biologists because of its computational simplicity and interpretability. But there is an inherent problem with this selection criterion, as genes of low absolute expression have a greater inherent error in their measured levels. These genes will then tend to numerically meet any given fold change cut-off even if the gene is not truly differentially expressed. The inverse also holds true, where highly expressed genes, having less error in their measured levels, may not meet an arbitrary fold-change cut-off

of 2.0 even when they are truly differentially expressed. This commonly used approach does not accommodate for background noise, variability, non-specific binding, or low copy numbers-characteristics typical of microarray data. While classic statistical approaches used for detecting differences between two groups include the parametric *t*-test and the nonparametric Wilcoxon rank sum [46]. The *t*-test has been used to compare expression profiles in microarray experiments in [47, 48]. Thomas et al. [49] proposed regression analysis and absolute value of the *Z*-score and determined 141 genes differentially expressed between AML and ALL with 1% significance at the genomic level. In [7] the authors used adjusted p-values for multiple testing to microarray data from a study of gene expression in two mouse models with very low HDL cholesterol levels. The work also suggested several data display techniques for the visual identification of genes with altered expression and of important features of these genes. Wachi et al. [11] investigated differentially expressed genes in squamous cell lung cancer which were identified by projecting microarray gene expression profiling onto a human protein interaction network.

4.2 Methodology

This module first introduces the basis idea of hypothesis testing which is used to either accept or reject the claim that randomly drawn samples come two different subsets of population having different distribution. In this study one subset of population being healthy smokers and the other being lung cancer diagnosed smokers. Samples correspond to expression profile of a gene under consideration from patients belonging to the two subsets of population. Genes supporting the hypothesis of samples being drawn from two populations with different distributions are identified as differentially expressed. Support of gene in favor or against the claimed hypothesis is quantitatively measured in terms of p-values and q-values. The following sections introduce the theory of hypothesis testing and its application to gene expression data analysis. Measures to assess the statistical significance of genes using p-values and q-values are also discussed along with algorithms used for their computation.

4.2.1 Hypothesis Testing

A statistical hypothesis is an assertion or conjecture concerning one or more populations. The truth or falsity of a statistical hypothesis is never known with absolute certainty unless we

examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the data contained in this sample to provide evidence that either supports or does not support the hypothesis. Evidence from the sample which is inconsistent with the stated hypothesis leads to a rejection of the hypothesis. However the decision procedure always involves a probability of a wrong conclusion. The rejection of a hypothesis implies that the sample evidence refutes it. That is, there is a small probability of obtaining the sample information observed when, in fact, the hypothesis is true.

Null and Alternative Hypotheses

The structure of hypothesis testing has to be formulated with the use of the term **null hypothesis**. This refers to any hypothesis we wish to test and is denoted by H_0. The rejection of H_0 leads to the acceptance of an alternative hypothesis, denoted by H_A. The alternative hypothesis H_A usually represents the theory to be tested and while the null hypothesis H_0 nullifies or opposes H_A and is often the logical complement to H_A. Through hypothesis testing, one of the two following conclusions could be reached,

Reject H_0 in favor of H_A because of sufficient evidence in the data.

Fail to reject H_0 because of insufficient evidence in the data.

The decision procedure could lead to either of two wrong conclusions. Rejection of the null hypothesis when it is true is called a **type I error**. Secondly, non-rejection of the null hypothesis when it is false is called a **type II error**.

In testing any statistical hypothesis, there are four possible situations that determine whether the taken decision is correct or in error. This can be summarized in Table 4.1

Table 4.1 Possible outcomes in Testing a Statistical Hypothesis

	H_0 is true	H_0 is false
Do not reject H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

4.2.2 Welch's t- test [28]

To identify differentially expressed genes, for each gene a null hypothesis is tested against an alternative hypothesis. A gene is declared significant if the null hypothesis is rejected in favor of the alternative hypothesis.

Let H_0 denote the null hypothesis that the expression levels in the two groups of patients namely healthy smokers and lung cancer diagnosed smokers comes from normal distributions with equal means. While the alternative hypothesis H_A is that the data comes from populations with unequal means.

Given a microarray experiment on N_A samples of Group A and N_B samples of Group B, first compute the mean $\left(\overline{X}\right)$ and variance $\left(\overline{s^2}\right)$ of both classes,

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad\qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$$

Welch's *t*-test defines the statistic *t* given by,

$$t = \frac{\overline{X_A} - \overline{X_B}}{\sqrt{\frac{{s_A}^2}{N_A} + \frac{{s_B}^2}{N_B}}}$$

where $\overline{X_A}, \overline{X_B}$ are the sample means, ${s_A}^2, {s_B}^2$ are sample variances.

4.2.3 p-values

The two hypotheses null and alternative specify two statistical models for the process that produced the data. The alternative hypothesis is what is expected to be true if the null hypothesis is false. The alternative hypothesis cannot always be proven to be true but we may be able to demonstrate that the alternative is much more plausible than the null hypothesis given the data. This demonstration is usually expressed in terms of a probability (a p- value) quantifying the strength of the evidence against the null hypothesis in favor of the alternative. p-value can be viewed as simply the probability of obtaining these data given that both samples come from the same distribution. In current context, p-value is a measure of how likely it is to get this spot data if no real difference existed. Therefore, a small p-value indicates that there is a small chance of

getting this data if no real difference existed and therefore we can decide that the difference in group expression data is significant. A p-value of 0.05 or less is generally considered statistically significant.

p-value is a measure of significance in terms of false positive rate. When a feature (gene) is tested for significance given a rule, the false positive rate is the rate that features supporting null hypothesis are declared significant. So a p-value cutoff of 0.05 means that on an average 5% of the features conforming the null hypothesis will be called significant.

Permutation Tests

The permutation test is a non-parametric method for calculating p-values. Here we assume a normal expression distribution, or that the two groups have an equal variance. The sample label is randomly permutated to all the possible labels, and the statistical test (say t-test) is calculated for all possible labels. Then, the likelihood of the sample being of that category is determined based on the distributions of other sample.

4.2.4 Multiple Hypothesis Testing Problem

In a microarray experiment, we analyze thousands of hypotheses simultaneously. During multiple-hypothesis testing, there are two different kinds of errors that occur: Type I error and Type II error. Type I errors occur when the null hypothesis is rejected when it is in fact true (false positive) and Type II errors occur when the null hypothesis is not rejected when it is false (false negative). Multiple-hypothesis testing involves guarding against much more complicated errors than single-hypothesis testing. Whereas we typically control the type I error rate for a single-hypothesis test, a compound error rate is controlled for multiple-hypothesis tests. The goal is to find a balance between excluding too many promising genes, and including genes that aren't significant. Using a p-value of 0.05 as our cutoff, if 15000 genes are on the microarray, and then 750 are falsely declared positive. Lowering the p-value to 0.01, we would have about 150 false positives which is still quiet high.

The Bonferroni correction

The Bonferroni correction sets the significance cut-off at α/N where α is the p-value cutoff previously set and N is the total number of hypothesis being tested. For example, while performing 20 hypothesis tests with $\alpha = 0.05$, Bonferroni correction states that, only reject a null

hypothesis if the p-value is less than 0.0025. But depending on the correlation structure of the tests, the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives.

4.2.5 False Discovery Rate (FDR) and q-Value

For large-scale multiple testing which often shows up in genomics, FDR[29] is a sensible measure of the balance between the number of true positives and false positives. This is defined as the proportion of false positives among all significant results.

Let F be the number of false positives, T be the number of true positives, and S be the total number of features called significant. FDR is the expected value of

$$FDR = \mathrm{E}\left[\frac{F}{T+F}\right] = \mathrm{E}\left[\frac{F}{S}\right]$$

A FDR of 5% means that among all features called significant, 5% of these truly support the null hypothesis on average.

q-Value

q-value [30,31] takes into account that several features are simultaneously tested while assigning significance to each feature. The q value for a particular feature is the expected proportion of false positives incurred when calling that feature significant. Therefore, calculating the q values for each feature and thresholding them at q-value level produces a set of significant features a proportion of which is expected to be false positives.

Statistical significance involves deciding between null and alternative hypotheses. While calculating multiple measures of statistical significance, it is necessary to account for the fact that decisions are made for several thousands of features simultaneously.

Let m is the total number of features being tested with m_0 being the number of features that truly support the null hypothesis, and $m_1 = m - m_0$ being the number of features supporting the alternative hypothesis. m_0 is unknown and has to be estimated, this estimation is done by $\pi_0 = \frac{m_0}{m}$, i.e., the proportion of truly null features. The height of this flat portion of the histogram density plot gives a conservative estimate of the overall proportion of null p-values.

This is quantified by,

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1,2,...,m\}}{m(1-\lambda)}$$ where λ is the tuning parameter.

The algorithm used for estimating *q*-values from a list of *p*-values is as described in [31] as follows,

1. Let $p_{(1)} \le p_{(2)} \le ... \le p_{(m)}$ be the ordered *p*-values, which is also an ordering of the features in terms of their evidence against the null hypothesis.
2. For a range of $\lambda = 0, 0.01, 0.02, ..., 0.95$ compute, $\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)}$.
3. Let $\hat{f}$ be the natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ.
4. The estimate of π_0 is set to, $\hat{\pi}_0 = \hat{f}(1)$.
5. Compute, $\hat{q}(p_{(m)}) = \min_{t \ge p_{(m)}} \frac{\hat{\pi}_0 m.t}{\#\{p_i \le t\}} = \hat{\pi}_0 . p_{(m)}$.
6. For $i = m-1, m-2, ..., 1$, find $\hat{q}(p_{(i)}) = \min_{t \ge p_{(i)}} \frac{\hat{\pi}_0 m.t}{\#\{p_i \le t\}} = \min\left(\frac{\hat{\pi}_0 m. p_{(i)}}{i}, \hat{q}(p_{(i+1)}) \right)$.
7. The estimated *q*-value for the i^{th} most significant feature is $\hat{q}(p_{(i)})$.

So if a feature with *q*-value ≤ 0.05 is called significant, then this results in a FDR of 5% among significant features.

4.3 Results

Catalogue of Somatic Mutations in Cancer, COSMIC database reports the list of genes that are somatically mutated in lung cancer. The latest version of the database COSMIC v70 reports a list of 24283 genes including alias known to be mutated in lung cancer. We use this list of genes and search for their corresponding expression profile in the dataset [26]. Expression profiles corresponding to 11237 genes were found and this gene subset was further analyzed to find differentially expressed genes. As mentioned previously, first Welch's t-statistic and its corresponding p-value was computed for each gene having 90 samples corresponding to healthy

population and 97 genes corresponding to the cancer affected population. The following histograms depict the t-statistics and p-value distribution among the genes.

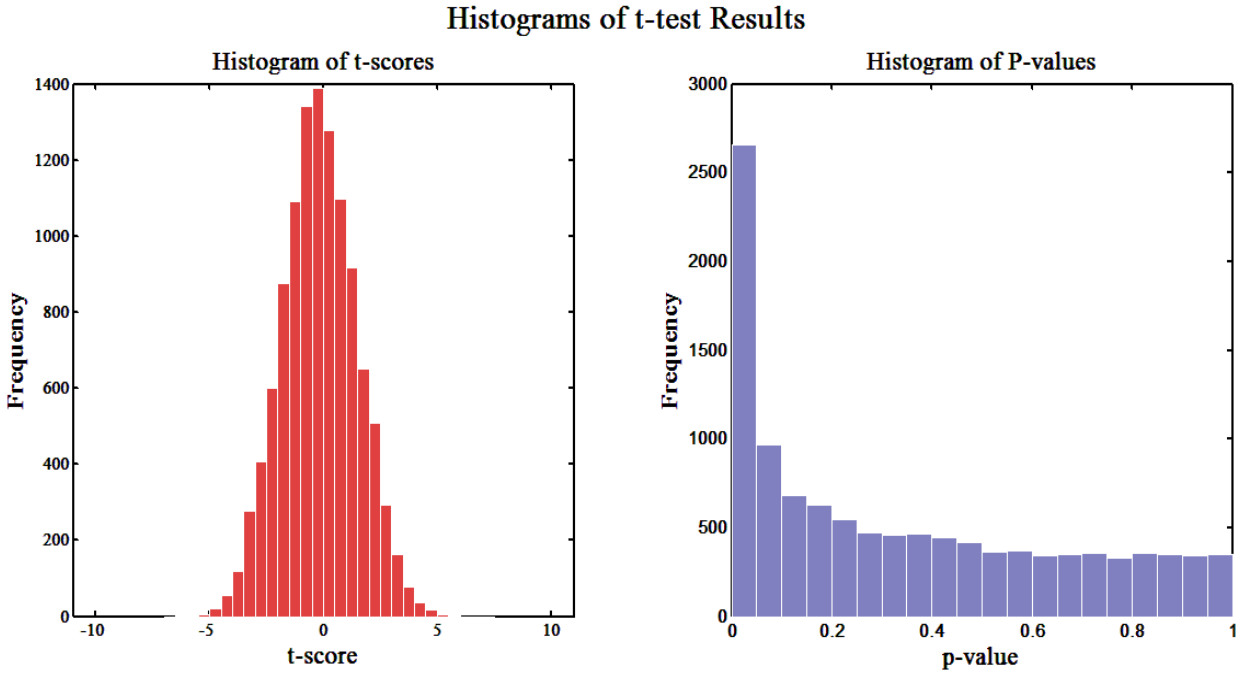

Figure 4.1: Histograms of t-test results.

From the Figure 4.1 it is evident that t-scores follow a normal distribution and genes with differential expression have absolute t-score values nearly as 4 or more.

Q-Q plots

Quantile-Quantile plots (Q-Q plots) [7] can be used for display of the test statistics for the thousands of genes under study from a microarray experiment. In a normal Q-Q plot, the quantiles of the data are plotted against the quantiles of a standard normal distribution and the plot can be used to assess whether data have a particular distribution or whether two datasets have the same distribution. Figure 4.2 gives the Q-Q plot of the t-test performed on the genes. In the t-score quantile plot, the black diagonal line represents the sample quantile being equal to the theoretical quantile. Data points of genes considered to be differentially expressed lie farther away from this line. Specifically, data points with t-scores > (1 - 1/ (2N)) or < 1/ (2N) display with red circles where N is the total number of genes.

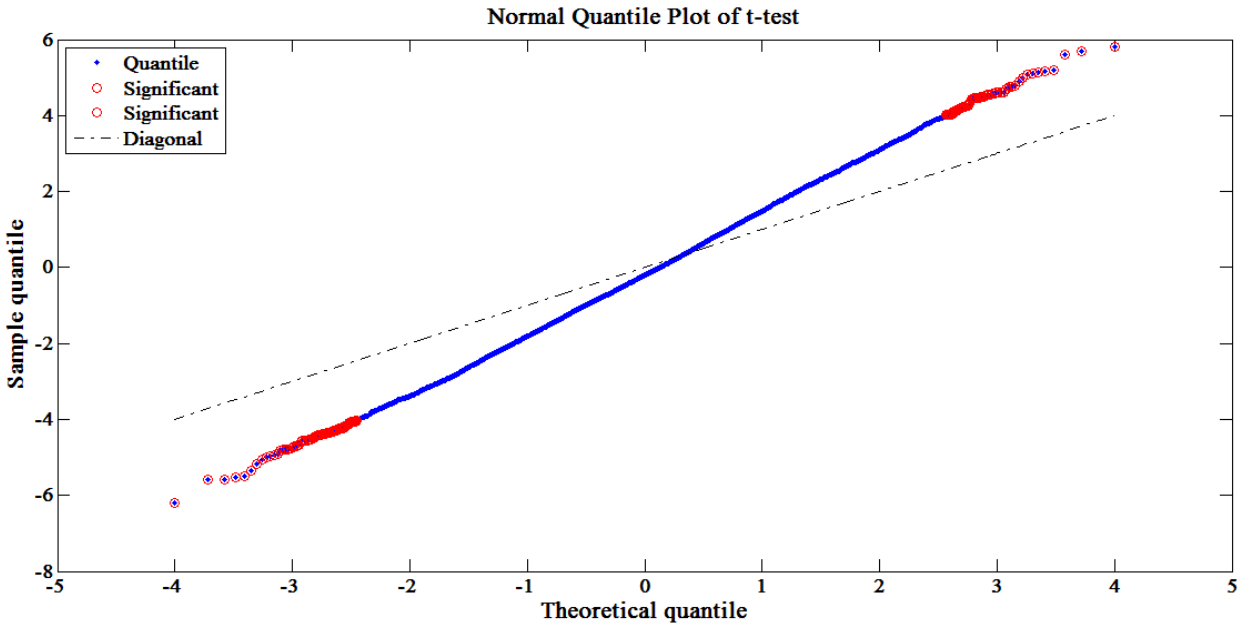

Figure 4.2 Q-Q plot of Welch's t-test statistic on gene set

Here a set of 11237 genes are simultaneously tested against the null hypothesis. So it is essential to take into consideration the errors encountered in multiple hypothesis testing. If 5% is considered to be the p-value cutoff for selecting differentially expressed genes, then taking into account the Bonferroni correction for multiple hypothesis testing, the adjusted p-value cut off is calculated to be, 0.0000044495861 which being a very stringent measure yields a list of only 30 genes. Hence the controlling the FDR is used to identify differentially expressed genes. The algorithm described in the methodology section is used to calculate q-values from the given set of p-values. The graphs in Figure 4.3 show the growth of p-value versus q-value and result of fitting the natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ, gives the resultant value of $\hat{\pi}_0(\lambda) = 0.6142$. This means that the proportion of truly null features amongst all features found significant is 61.42% which corresponds to the flat portion of the p-value histogram plot. From the graph plot of Figure 4.3 of q-values versus p-values, the proportion of false discoveries for different p-value cut offs can be estimated.

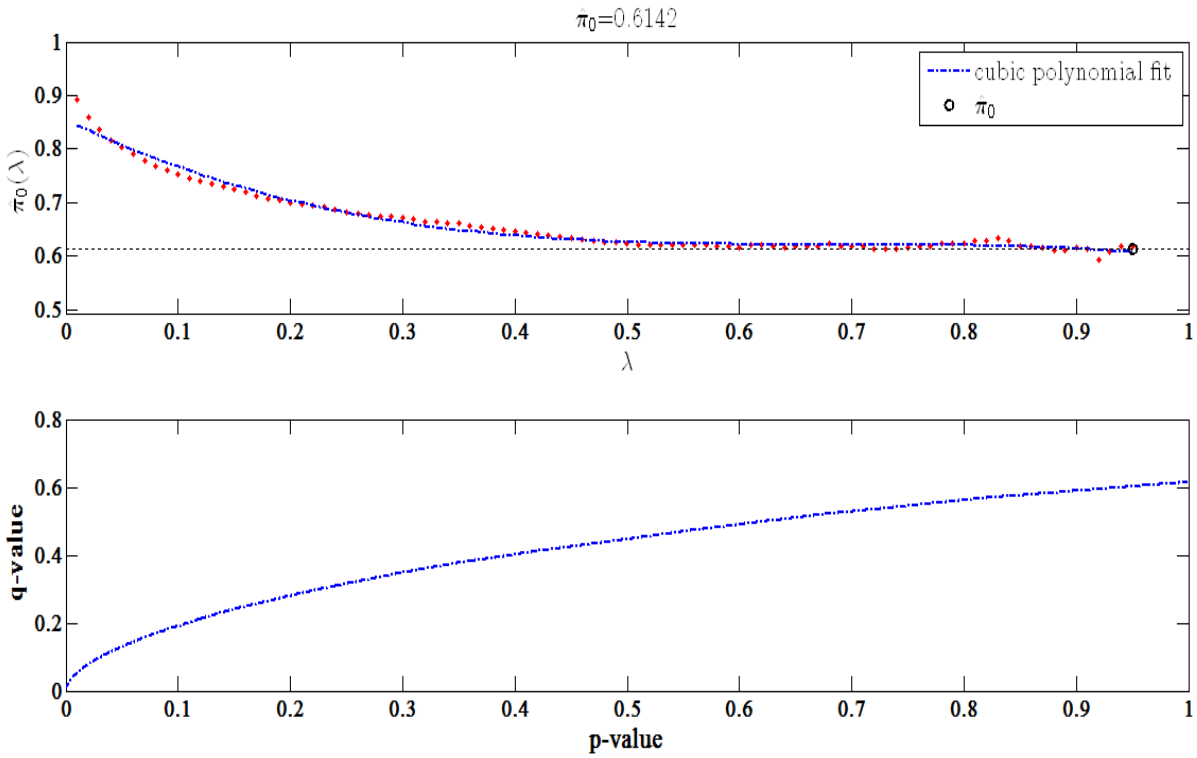

Figure 4.3 Plot of q-value versus p-value

The plot of p-values versus t-statistics and q-value versus t-statistics for the entire dataset are as follows.

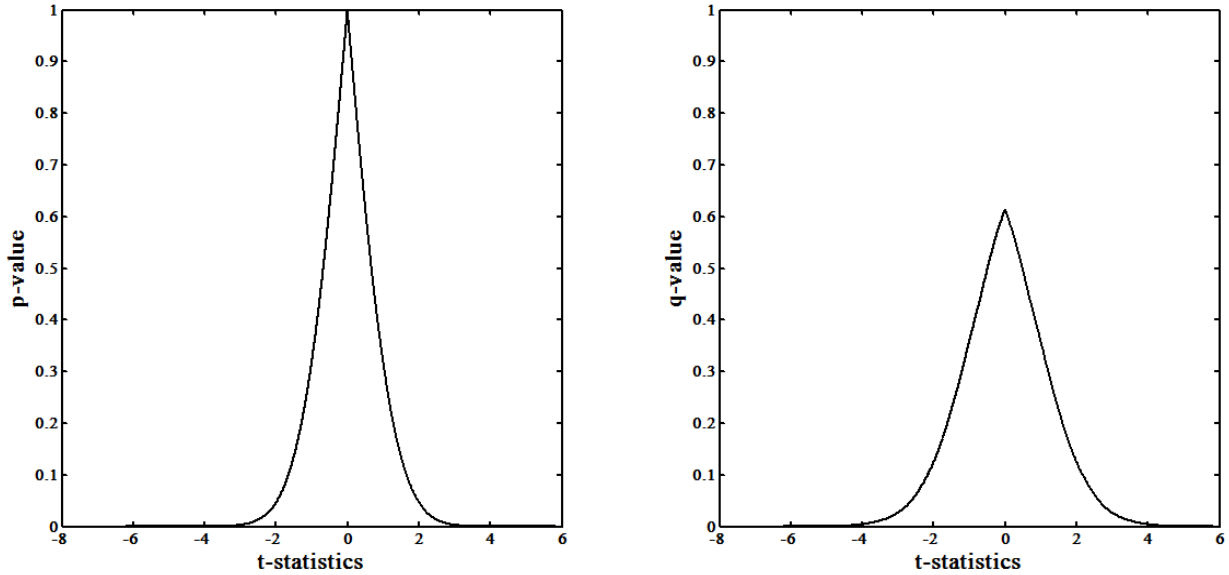

Figure 4.4: p-value versus t-statistics and q-value versus t-statistics plots.

From these plots it is evident that there is a much higher rate of false positives than false discoveries. Figure 4.4 is a plot of the variation of the number of significant genes filtered for

each p-value and q-vale cutoff versus the respective p and q value. From the plot of number of significant genes for each q value, we notice that for estimated q values slightly greater than 0.002, quiet a large increase occurs in the number of significant genes over a small increase in q value. This allows us to easily see that a slightly larger q-value cutoff results in many more significant genes. While in case of number of significant genes versus p-value cutoff, there is a uniform quadratic increase in the number of significant genes filtered.

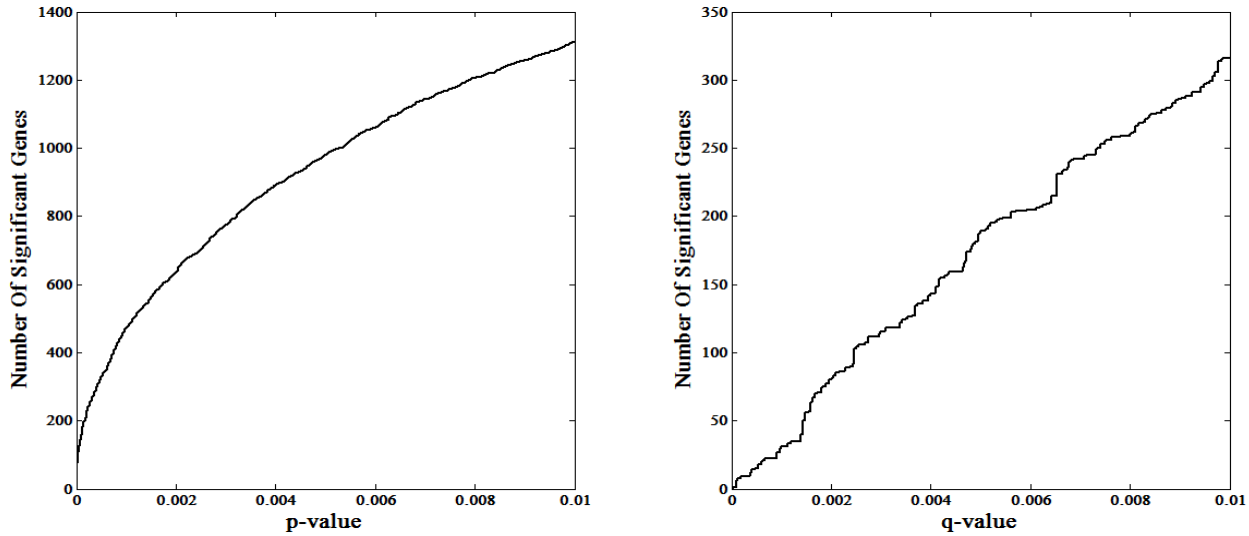

Figure 4.5: Plot of number of significant genes filtered for respective p-value and q-value cutoff.

The number of genes filtered for four statistically significant p-value and q-value cutoffs is summarized in Table 4.2 below.

Table 4.2: Number of genes filtered for 4 statistically significant p and q value cutoffs

Cut off value	No. of Significant genes based on p-value	No. of Significant genes based on q-value
0.001	374	31
0.005	968	168
0.01	1231	316
0.05	2659	1270

In this work the cutoff of q-value is set to 0.005 and the resultant list of 168 genes is identified as differentially expressed between the diseased and healthy population. Table 4.1 lists the 168 genes identified to be significant by our statistical study along with their t-statistic, p-value , q-value and mutation percentage as reported by COSMIC.

Table 4.3 List of 168 genes which are differentially expressed between healthy smokers and smokers diagnosed with lung cancer.

Serial No	Entrez Gene symbol	Entrez Gene name	t-statistic	p-value	q-value	Mutation % in Lung Cancer from COSMIC
1	HUWE1	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase	3.916639	0.000127	0.004851	4.91
2	C6	complement component 6	-5.17463	6.03E-07	0.000376	4.86
3	TRIO	trio Rho guanine nucleotide exchange factor	-4.38089	1.98E-05	0.001818	4.47
4	USP34	ubiquitin specific peptidase 34	3.996985	9.32E-05	0.004173	4.12
5	TSC2	tuberous sclerosis 2	4.473875	1.38E-05	0.001586	2.58
6	NELL2	NEL-like 2 (chicken)	-4.60188	7.76E-06	0.001397	3.03
7	BICC1	BicC family RNA binding protein 1	-3.94432	0.000115	0.004713	2.82
8	CWH43	cell wall biogenesis 43 C-terminal homolog (S. cerevisiae)	-4.07339	7.01E-05	0.003693	2.82
9	DNAJC6	DnaJ (Hsp40) homolog, subfamily C, member 6	-4.78712	3.46E-06	0.000913	2.38
10	PDZD8	PDZ domain containing 8	-4.28312	2.96E-05	0.002295	2.32
11	ALPK1	alpha-kinase 1	4.57169	8.94E-06	0.001423	2.2
12	NCOA1	nuclear receptor coactivator 1	-3.92472	0.000123	0.004801	2.24
13	ITGAL	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)	3.937816	0.000116	0.004722	1.99
14	DIP2A	DIP2 disco-interacting protein 2 homolog A (Drosophila)	4.613137	7.40E-06	0.001397	2.18
15	BCAS1	breast carcinoma amplified sequence 1	5.196561	5.38E-07	0.000371	1.89
16	RPGRIP1L	RPGRIP1-like	-4.07425	6.84E-05	0.003688	1.9
17	EVPL	envoplakin	4.263052	3.22E-05	0.002432	1.76
18	MAPK8IP3	mitogen-activated protein kinase 8 interacting protein 3	4.010311	8.82E-05	0.004113	1.6
19	PRRC2A	proline-rich coiled-coil 2A	3.992527	9.43E-05	0.004194	1.69
20	WWC3	WWC family member 3	-4.51463	1.13E-05	0.001476	1.68
21	GIGYF2	GRB10 interacting GYF protein 2	-3.92609	0.000122	0.004799	1.62
22	ATP8B1	ATPase, aminophospholipid transporter, class I, type 8B, member 1	5.678977	5.87E-08	9.02E-05	1.35
23	PRSS12	protease, serine, 12 (neurotrypsin, motopsin)	-4.24089	3.51E-05	0.002466	1.48
24	EXT2	exostosin glycosyltransferase 2	-4.09253	6.47E-05	0.003544	1.29
25	SMC6	structural maintenance of chromosomes 6	-4.00275	9.06E-05	0.004168	1.4
26	PLA1A	phospholipase A1 member A	-4.08131	6.66E-05	0.003618	1.41
27	ZNF611	zinc finger protein 611	4.985286	1.42E-06	0.000543	1.34
28	ZAP70	zeta-chain (TCR) associated protein kinase 70kDa	4.103404	6.14E-05	0.003414	1.07
29	PLEKHA5	pleckstrin homology domain containing, family A member 5	5.12791	7.80E-07	0.000412	1.33

30	ARMCX2	armadillo repeat containing, X-linked 2	-4.06017	7.31E-05	0.003727	1.26
31	HBB	hemoglobin, beta	-3.92448	0.000128	0.004889	1.27
32	DUOX1	dual oxidase 1	4.613694	7.97E-06	0.001397	1.27
33	DDX18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	-4.23184	3.64E-05	0.002466	1.26
34	SNX19	sorting nexin 19	-4.01856	8.59E-05	0.004103	1.26
35	POLR1B	polymerase (RNA) I polypeptide B, 128kDa	4.243379	3.48E-05	0.002466	1.27
36	RAB3GAP1	RAB3 GTPase activating protein subunit 1 (catalytic)	-3.94643	0.000113	0.00468	1.27
37	CRY1	cryptochrome circadian clock 1	-4.74631	4.14E-06	0.00098	1.2
38	CPE	carboxypeptidase E	-4.94078	1.77E-06	0.00061	1.2
39	CDC5L	cell division cycle 5-like	-4.34221	2.35E-05	0.001998	1.19
40	TAOK1	TAO kinase 1	3.979121	9.96E-05	0.004347	1.04
41	FUT8	fucosyltransferase 8 (alpha (1,6) fucosyltransferase)	-4.41848	1.70E-05	0.001674	1.13
42	FXR1	fragile X mental retardation, autosomal homolog 1	-4.36763	2.09E-05	0.001885	1.13
43	FGF14	fibroblast growth factor 14	-4.66194	6.00E-06	0.001182	1.03
44	SERPINI1	serpin peptidase inhibitor, clade I (neuroserpin), member 1	-4.26116	3.24E-05	0.002432	1.12
45	MPHOSPH10	M-phase phosphoprotein 10 (U3 small nucleolar ribonucleoprotein)	-4.15805	4.91E-05	0.002974	1.13
46	RIN1	Ras and Rab interactor 1	4.005729	8.98E-05	0.004157	0.92
47	TRIM36	tripartite motif containing 36	-5.52504	1.33E-07	0.000115	0.96
48	PYGB	phosphorylase, glycogen; brain	-4.42116	1.67E-05	0.001671	1.06
49	MAP2K4	mitogen-activated protein kinase kinase 4	-4.54858	1.01E-05	0.001423	0.81
50	ODF2	outer dense fiber of sperm tails 2	-4.42438	1.65E-05	0.001671	1.06
51	CLSTN3	calsyntenin 3	4.110805	5.92E-05	0.003376	1.06
52	KLHL24	kelch-like family member 24	4.455326	1.45E-05	0.001586	1.06
53	DHCR7	7-dehydrocholesterol reductase	-4.43884	1.58E-05	0.00164	1.05
54	HSPA1L	heat shock 70kDa protein 1-like	-4.1174	5.82E-05	0.003371	0.98
55	ZC3H14	zinc finger CCCH-type containing 14	-4.45005	1.48E-05	0.001599	0.99
56	ZFR	zinc finger RNA binding protein	-4.32864	2.46E-05	0.002043	0.99
57	PHACTR4	phosphatase and actin regulator 4	4.317437	2.58E-05	0.002091	0.99
58	MAP7	microtubule-associated protein 7	-4.06391	7.13E-05	0.003693	0.99
59	ZKSCAN5	zinc finger with KRAB and SCAN domains 5	-5.59651	7.80E-08	9.02E-05	0.99
60	LRRC40	leucine rich repeat containing 40	4.432574	1.59E-05	0.00164	0.99
61	NSUN3	NOP2/Sun domain family, member 3	-5.59727	7.85E-08	9.02E-05	0.92
62	ZSCAN5A	zinc finger and SCAN domain containing 5A	-4.23601	3.60E-05	0.002466	0.91
63	ZNF276	zinc finger protein 276	4.24998	3.39E-05	0.002466	0.92
64	OSBP	oxysterol binding protein	-4.21634	3.89E-05	0.002543	0.91
65	PRR11	proline rich 11	4.594356	8.02E-06	0.001397	0.84
66	MYO1C	myosin IC	5.153849	6.53E-07	0.000376	0.85

67	LRRFIP1	leucine rich repeat (in FLII) interacting protein 1	4.148787	5.28E-05	0.003088	0.85
68	PTAFR	platelet-activating factor receptor	3.948763	0.000112	0.004674	0.85
69	ZNF160	zinc finger protein 160	5.100173	8.37E-07	0.000412	0.85
70	EIF2B3	eukaryotic translation initiation factor 2B, subunit 3 gamma, 58kDa	-4.69855	5.36E-06	0.00112	0.85
71	RRAGB	Ras-related GTP binding B	-4.38962	1.93E-05	0.001806	0.77
72	GIN1	gypsy retrotransposon integrase 1	-4.31868	2.56E-05	0.002091	0.77
73	UBQLN4	ubiquilin 4	4.46772	1.37E-05	0.001586	0.78
74	NARS2	asparaginyl-tRNA synthetase 2, mitochondrial (putative)	-4.24234	3.49E-05	0.002466	0.77
75	GSDMB	gasdermin B	4.596509	8.10E-06	0.001397	0.78
76	AGPS	alkylglycerone phosphate synthase	-5.49309	1.30E-07	0.000115	0.7
77	ACVR2A	activin A receptor, type IIA	-3.98627	9.81E-05	0.004311	0.63
78	SOX9	SRY (sex determining region Y)-box 9	-4.3548	2.27E-05	0.001961	0.7
79	B2M	beta-2-microglobulin	4.045138	7.67E-05	0.003841	0.7
80	TRDMT1	tRNA aspartic acid methyltransferase 1	-4.06607	7.17E-05	0.003693	0.7
81	CCDC81	coiled-coil domain containing 81	4.234149	3.68E-05	0.002466	0.7
82	ZNF224	zinc finger protein 224	3.934369	0.000119	0.004724	0.7
83	CARM1	coactivator-associated arginine methyltransferase 1	4.236663	3.59E-05	0.002466	0.7
84	SLC16A6	solute carrier family 16, member 6	4.15815	4.91E-05	0.002974	0.7
85	ZNF721	zinc finger protein 721	4.182838	4.44E-05	0.002737	0.7
86	HIBCH	3-hydroxyisobutyryl-CoA hydrolase	-3.95598	0.000109	0.004642	0.63
87	CLGN	calmegin	-3.93493	0.000119	0.004724	0.63
88	PIGK	phosphatidylinositol glycan anchor biosynthesis, class K	-4.22395	3.76E-05	0.002497	0.63
89	RIPK1	receptor (TNFRSF)-interacting serine-threonine kinase 1	-4.24693	3.46E-05	0.002466	0.51
90	NR2F1	nuclear receptor subfamily 2, group F, member 1	-4.29014	2.91E-05	0.002282	0.63
91	DCLRE1C	DNA cross-link repair 1C	5.816684	2.62E-08	9.02E-05	0.63
92	DLAT	dihydrolipoamide S-acetyltransferase	-4.79433	3.36E-06	0.000913	0.63
93	MFSD11	major facilitator superfamily domain containing 11	4.365797	2.10E-05	0.001885	0.63
94	EFNB2	ephrin-B2	-4.23728	3.57E-05	0.002466	0.56
95	CORO2A	coronin, actin binding protein, 2A	4.688633	5.97E-06	0.001182	0.56
96	ZNF652	zinc finger protein 652	3.97411	0.00011	0.004642	0.56
97	CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	4.031805	8.11E-05	0.00394	0.56
98	MPP5	membrane protein, palmitoylated 5 (MAGUK p55 subfamily member 5)	-3.95097	0.000111	0.004662	0.56
99	TMEM33	transmembrane protein 33	-4.06609	7.07E-05	0.003693	0.49
100	RHOT2	ras homolog family member T2	3.944158	0.000114	0.004702	0.49
101	GFPT1	glutamine--fructose-6-phosphate transaminase 1	-3.99811	9.24E-05	0.004173	0.49
102	ECI2	enoyl-CoA delta isomerase 2	-4.09408	6.34E-05	0.003496	0.49

103	PRPS1	phosphoribosyl pyrophosphate synthetase 1	-4.38722	1.94E-05	0.001806	0.49
104	TTC33	tetratricopeptide repeat domain 33	-4.40899	1.76E-05	0.001711	0.49
105	ECD	ecdysoneless homolog (Drosophila)	-4.96088	1.67E-06	0.000605	0.49
106	MTPAP	mitochondrial poly(A) polymerase	-4.39128	1.90E-05	0.001806	0.49
107	CIAO1	cytosolic iron-sulfur assembly component 1	-4.55778	9.38E-06	0.001423	0.49
108	PPM1D	protein phosphatase, Mg2+/Mn2+ dependent, 1D	-4.562	9.22E-06	0.001423	0.49
109	RAPGEFL1	Rap guanine nucleotide exchange factor (GEF)-like 1	4.033536	8.11E-05	0.00394	0.42
110	RBM4	RNA binding motif protein 4	4.112778	5.89E-05	0.003376	0.42
111	UBE2N	ubiquitin-conjugating enzyme E2N	-3.95869	0.000107	0.004629	0.38
112	ARGLU1	arginine and glutamate rich 1	4.033247	8.11E-05	0.00394	0.42
113	DCUN1D4	DCN1, defective in cullin neddylation 1, domain containing 4	-4.18481	4.44E-05	0.002737	0.42
114	P2RX4	purinergic receptor P2X, ligand-gated ion channel, 4	-4.78008	3.56E-06	0.000913	0.42
115	SLC39A14	solute carrier family 39 (zinc transporter), member 14	-4.30701	2.70E-05	0.002162	0.42
116	ARID5A	AT rich interactive domain 5A (MRF1-like)	4.487122	1.27E-05	0.001533	0.42
117	TGDS	TDP-glucose 4,6-dehydratase	-4.51507	1.17E-05	0.001476	0.42
118	XRCC4	X-ray repair complementing defective repair in Chinese hamster cells 4	-4.3352	2.39E-05	0.002009	0.42
119	SGSM2	small G protein signaling modulator 2	4.543021	1.03E-05	0.001423	0.35
120	TMEM45A	transmembrane protein 45A	-4.35485	2.21E-05	0.001957	0.35
121	PINK1	PTEN induced putative kinase 1	-4.53891	1.03E-05	0.001423	0.34
122	CRCP	CGRP receptor component	4.013196	8.76E-05	0.004111	0.35
123	DAPP1	dual adaptor of phosphotyrosine and 3-phosphoinositides	4.558413	9.52E-06	0.001423	0.35
124	HTRA1	HtrA serine peptidase 1	-5.05238	1.04E-06	0.00048	0.35
125	PTMA	prothymosin, alpha	4.50367	1.18E-05	0.001476	0.35
126	MGAT4A	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme A	-4.14047	5.26E-05	0.003088	0.35
127	NPTX1	neuronal pentraxin I	4.015251	8.63E-05	0.004103	0.35
128	UBFD1	ubiquitin family domain containing 1	-3.98555	9.68E-05	0.004281	0.35
129	MRPS15	mitochondrial ribosomal protein S15	-4.05674	7.35E-05	0.003727	0.35
130	CTSS	cathepsin S	3.917796	0.000126	0.004851	0.35
131	SLC35E1	solute carrier family 35, member E1	4.779397	3.57E-06	0.000913	0.35
132	PRKAA1	protein kinase, AMP-activated, alpha 1 catalytic subunit	-3.92165	0.000125	0.004829	0.25
133	PPP2R2D	protein phosphatase 2, regulatory subunit B, delta	-4.824	3.01E-06	0.000904	0.28
134	TSN	translin	-4.45973	1.42E-05	0.001586	0.28
135	RAB15	RAB15, member RAS oncogene family	4.460813	1.45E-05	0.001586	0.28
136	HAUS2	HAUS augmin-like complex, subunit 2	4.574222	8.75E-06	0.001423	0.28

137	ADK	adenosine kinase	-4.75622	3.96E-06	0.000976	0.28
138	CST6	cystatin E/M	-3.93171	0.000119	0.004724	0.26
139	RBCK1	RanBP-type and C3HC4-type zinc finger containing 1	4.742132	4.26E-06	0.00098	0.26
140	ICAM4	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)	4.072574	6.98E-05	0.003693	0.21
141	GGCX	gamma-glutamyl carboxylase	-4.1458	5.16E-05	0.003067	0.21
142	ZFYVE21	zinc finger, FYVE domain containing 21	-4.06278	7.16E-05	0.003693	0.21
143	GNPDA1	glucosamine-6-phosphate deaminase 1	4.461677	1.41E-05	0.001586	0.21
144	FAM193B	family with sequence similarity 193, member B	3.933077	0.000119	0.004724	0.21
145	NCR3	natural cytotoxicity triggering receptor 3	4.013645	8.68E-05	0.004103	0.21
146	RRAS2	related RAS viral (r-ras) oncogene homolog 2	-4.51599	1.15E-05	0.001476	0.11
147	LHB	luteinizing hormone beta polypeptide	4.189481	4.33E-05	0.002737	0.14
148	APOBEC3C	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C	4.033819	8.04E-05	0.00394	0.14
149	TMEM110	transmembrane protein 110	-4.92015	1.92E-06	0.000629	0.14
150	GTF2H3	general transcription factor IIH, polypeptide 3, 34kDa	4.896398	2.12E-06	0.000666	0.13
151	MAK16	MAK16 homolog (S. cerevisiae)	-3.99878	9.26E-05	0.004173	0.14
152	MED6	mediator complex subunit 6	5.066135	1.32E-06	0.000543	0.14
153	SSSCA1	Sjogren syndrome/scleroderma autoantigen 1	-3.99803	9.23E-05	0.004173	0.14
154	FSCN1	fascin actin-bundling protein 1	4.501089	1.19E-05	0.001476	0.14
155	SLC50A1	solute carrier family 50 (sugar efflux transporter), member 1	4.113336	5.98E-05	0.003384	0.14
156	APOC4	apolipoprotein C-IV	4.200397	4.17E-05	0.002688	0.14
157	SNRNP27	small nuclear ribonucleoprotein 27kDa (U4/U6.U5)	-4.69406	5.25E-06	0.00112	0.14
158	RNF146	ring finger protein 146	-3.97548	0.000101	0.004366	0.14
159	SAT1	spermidine/spermine N1-acetyltransferase 1	4.507582	1.20E-05	0.001476	0.14
160	KDELR3	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3	-4.23216	3.65E-05	0.002466	0.14
161	PBX1	pre-B-cell leukemia homeobox 1	-4.34723	2.27E-05	0.001961	0.06
162	FAM69A	family with sequence similarity 69, member A	-5.35749	2.49E-07	0.000191	0.07
163	CDK9	cyclin-dependent kinase 9	-4.18748	4.38E-05	0.002737	0.06
164	ERN2	endoplasmic reticulum to nucleus signaling 2	4.217359	3.91E-05	0.002543	0.07
165	SET	SET nuclear proto-oncogene	3.940239	0.000115	0.004713	0.06
166	NOTCH2NL	notch 2 N-terminal like	4.745427	4.49E-06	0.000998	0.06
167	NPFFR1	neuropeptide FF receptor 1	4.549352	9.77E-06	0.001423	0.07
168	IER3	immediate early response 3	-3.96441	0.000109	0.004642	0.07

4.4 Discussion

In order to analyze cancer gene expression data, the first idea is to identify gene exhibiting differential expression patterns between healthy and diseased population. This is achieved in this work by testing each gene against the null hypothesis of showing no expression change between the two population groups. Genes rejecting the null hypothesis with high degree of confidence were considered to be differentially expressed. Significance of a gene is quantified using its p-value and q-value measure. A set of 168 genes were identified by this procedure using a q-value cut off of 0.005.

Chapter 5

Clustering of Significant Genes

Cluster analysis is a statistical technique used to generate a category structure that fits a set of observations. The groups that are formed should have a high degree of association between members of the same group and that of low degrees between members of different groups. Clustering is a useful exploratory technique for gene expression data as it groups together genes with similar expression patterns and allows biologist to identify potentially meaningful relationships between genes. Sharing of the regulatory mechanism among genes at the sequence level, in an organism, is predominantly responsible for them being co-expressed. Genes having similar gene expression profiles are more likely to regulate one another or be regulated by some other common parent gene. Genes belonging to same cluster are typically involved in related functions and are frequently co-regulated. Clustering of gene expression patterns is used to identify groups of co-expressed genes [8] and generates gene interaction/gene regulatory networks. Clustering also facilitates in the functional annotation of uncharacterized genes. For instance, if an uncharacterized gene belong a cluster dominated by genes having some known similar function, the unknown gene could possibly have a similar function.

The previous chapter deals with identification of genes exhibiting differential expression profile between the two subsets of healthy and diseased population. Statistical methods of hypothesis testing using standard test statistics of Welch's t-test and its corresponding p-values

and q-values are used to assess the level of significance of a gene. A q-value cut off of 0.005 is used to identify a list of 168 differentially expressed genes between cancerous and non-cancerous population. The expression profile corresponding to these 168 genes is further analyzed in this module to identify groups of co-expressed genes using clustering techniques. The objective of this chapter is to identify gene clusters in the healthy and diseased population separately and analyze how the clustering result differs between the two cancer affected and not affected population. Such an analysis would provide insight into how the gene groupings vary between the two populations. For instance, genes belonging same cluster in healthy population but have been assigned into different groups in clustering result of diseased population may be identified as target genes which are dis-regulated thus being a cause for the disease due their altered expression amongst the two populations. The prime objective of this module is to identify gene clusters separately for the two population subsets. The approach used here to identify gene clusters is to first perform hierarchical agglomerative clustering on the gene expression data. Hierarchical clustering result is obtained for six distance measures namely Euclidean, Manhattan, Mahalanobis, Pearson correlation and Spearman rank correlation distances and using three linkage criterions namely single, average and complete linkages. All these clustering results are then evaluated using hierarchical cluster validity measure of cophenetic correlation coefficient CPCC to determine the best fitting distance measure and linkage criterion for the dataset. From the results it is found that Spearman rank correlation distance measure and average linkage criterion gives the best hierarchical clustering result for both the healthy and diseased population data sets. After finalizing the distance and linkage criterion the next step is to determine the number of natural clusters in the expression dataset corresponding to both the datasets. This issue is addressed by iteratively dissecting the linkages of the resultant dendrogram from higher to lower levels. At each depth of dissection the number of clusters returned is one more than the clustering result of previous step. At each step the clusters returned from dissection are evaluated using seven popular cluster validity indices available in literature for validating results of partitional clustering algorithms. A consensus strategy on the optimal number of clusters returned by the validity indices is employed to decide the final clustering output of the two datasets. In this work, 3 clusters are obtained corresponding to expression dataset of the healthy while on the other hand 4 clusters are obtained corresponding to the diseased population dataset. Analyzing the cluster assignment of the genes in healthy and

diseased population, we identify 12 specific genes assigned to different clusters in healthy and diseased population. A subset of 13 genes with expression profile significantly different so as to form a new cluster in diseased population is also identified. Gene ontology (GO) annotations have been used to biologically validate the clustering results and draw conclusions regarding the genes of the three major cluster of healthy population. This biological enrichment of the clusters is performed using Gene Ontology Tools namely GO Term Finder which returns the list of common GO terms shared by the queried genes. Gene belonging to each of the 3 clusters corresponding to the healthy population was analyzed using this tool, and finally the three clusters were found to be dominantly taking part in three specific groups of biological processes and molecular functions. Hence the 3 clusters were denoted as ***regulatory cluster, response and signaling cluster*** and ***cell development and maintenance cluster.***

5.1 Related Works

Cluster analysis is at present the most commonly used computational approach for analyzing microarray data. Hence it has also been extensively studied over the years using a variety of approaches. The rich literature on cluster analysis goes back to over nearly decades [32]. Various categories of clustering algorithms are possible. Hierarchical clustering approaches are tree-based approach uses distance measures between genes to group genes into a hierarchical tree. This includes works of [33, 34]. Eisen et al. [32] applied an agglomerative algorithm called UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and developed a software package called Cluster accompanying a visualization program called TreeView. The gene similarity metric used is a form of correlation coefficient. The output of the algorithm is a dendrogram and an ordered fingerprint matrix. The rows in the matrix are permuted based on the dendrogram, so that groups of genes with similar expression patterns are adjacent. While, Alon et al. [34] employed a divisive approach, called the deterministic-annealing algorithm (DAA) split the genes [35]. The other category clustering is partitional which includes K-Means and Self-Organizing Map (SOM) which clusters genes so that within-cluster variation is minimized and between-cluster variation is maximized. Tavazoie et al. [37] reported their success of identifying groups of co-regulated yeast genes with k-means algorithm. Tamayo et al. [36] used self-organizing maps (SOM) to identify clusters in the yeast cell cycle and human hematopoietic

differentiation data sets. Graph theoretic approach towards microarray data clustering maps the clustering problem into finding the minimum cut or maximal cliques in proximity graph. Sharan and Shamir [38] developed an algorithm named CLICK (CLuster Identification via Connectivity Kernels) which iteratively finds the minimum cut in the proximity graph and recursively splits the data set into a set of connected components from the minimum cut. Ben-Dor et al. [39] introduced the idea of a corrupted clique graph data model for clustering and developed a polynomial algorithm for finding true clustering with high probability. In [40] authors provide an extensive survey on cluster analysis for gene expression data.

In the work of Milligan and Cooper [54], a set of 30 cluster validity indices are compared based on the results obtained in hundreds of environments. But the work relating to cluster validation techniques for gene expression microarray data is found to have very few references in the literature. S. Datta et al. [41] consider six clustering algorithms (of various flavors) and evaluate their performances on a well-known publicly available microarray data set on sporulation of budding yeast and on two simulated data sets. J. Handl et al. [42] presented a survey of clustering validation techniques used in post-genomic data analysis. For this purpose, the different types of validation measures have been reviewed, and specific weaknesses of individual measures have been addressed. S. Datta et al. [43] proposed two performance measures for evaluating the results of a clustering algorithm in its ability to produce biologically meaningful clusters. V. Pihur et al. [44] combined the ranks of a set of clustering algorithms using a Monte Carlo cross-entropy algorithm through a weighted aggregation that optimizes a distance criterion. In one of the earlier works, A. Ghosh et al. [45] compared the performance of 19 cluster validity indices, in identifying some possible genes mediating certain cancers, based on gene expression data. While various validation measures have been proposed over the years to judge the quality of clusters produced by a given clustering algorithm including their biological relevance, unfortunately, a given clustering algorithm can perform poorly under one validation measure while outperforming many other algorithms under another validation measure.

5.2 Methodology

This section starts with the description of the distance measures employed to compare the expression profile of two genes. Five distance measures are used to compare expression profile

of genes. The next section deals with the hierarchical clustering algorithm used to group the genes into a hierarchical tree. For agglomeration of two clusters at every step of this algorithm, distance between the clusters is to be assessed. This can be done using various linkage criterions. In this work three linkage criterions are considered namely single, average and complete which is discussed next. The remaining of this section gives the mathematical formulations of the cluster validity indices used to evaluate the results of the clustering approach used in this work to find the true gene clusters in both the datasets.

5.2.1 Distance Measures

For analysis of gene expression profiles we need to quantify similarity or dissimilarity between the genes, by the variety distance measures described as follows.

Let $X = [x_1, x_2, x_3, \cdots, x_n]$ and $Y = [y_1, y_2, y_3, \cdots, y_n]$ be two n-dimensional data vectors. Here *X* and *Y* are expression profiles corresponding to two genes.

Euclidean Distance

Euclidean distance between the two data vectors *X* and *Y* is given by,

$$D_{Euc}(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan Distance

Manhattan distance between the two data vectors *X* and *Y* is given by,

$$D_{Man}(X,Y) = \sum_{i=1}^{n} | x_i - y_i |$$

Mahalanobis distance

Mahalanobis distance between *X* and *Y* is given by,

$$Dis_{Mah} = \sqrt{(X-Y)C^{-1}(X-Y)^T}$$

where *C* is the covariance matrix, whose $(i,j)^{th}$ entry is given by,

$$C_{ij} = \mathrm{cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$$

Pearson correlation coefficient

Pearson correlation coefficient (PCC) is used to measure degree of linear dependence between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1corresponds to total negative correlation. Pearson's

correlation coefficient when applied to a population is commonly represented by the Greek letter $\boldsymbol{\rho}$. It is given by,

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where, $\mathrm{cov}(X,Y)$ is the covariance between *X* and *Y*.

σ_X is the standard deviation of *X*.

The covariance between two jointly distributed real-valued random variables *X* and *Y* is defines by, $\mathrm{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$ where *E[X]* is the expected value of *X*, also known as the mean of *X*.

PCC in terms of mean and expectation thus could be written as,

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Correlation distance between *X* and *Y* is given by,

$$Dis_{Corr}(X,Y) = 1 - \rho_{X,Y}$$

In simple terms,

$$Dis_{Corr} = 1 - \frac{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)^T}{\sqrt{\left(X - \overline{X}\right)\left(X - \overline{X}\right)^T}\sqrt{\left(Y - \overline{Y}\right)\left(Y - \overline{Y}\right)^T}}$$

A PCC value of 1 essentially means that the two genes have similar expression profiles and a value of –1 means that the two genes have exactly opposite expression profiles. A value of 0 means that no relationship can be inferred between the expressions profiles of genes.

Spearman Rank correlation coefficient

Spearman Rank correlation coefficient (RCC) is a distance measure that does not take into account the actual magnitude of the expression value in each condition, but takes into account the rank of the expression value. The Spearman rank correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. An advantage of RCC is that it is not sensitive to outliers in the data.

$$Dis_{Spr} = 1 - \frac{\left(R_X - \overline{R_X}\right)\left(R_Y - \overline{R_Y}\right)^T}{\sqrt{\left(R_X - \overline{R_X}\right)\left(R_X - \overline{R_X}\right)^T}\sqrt{\left(R_Y - \overline{R_Y}\right)\left(R_Y - \overline{R_Y}\right)^T}}$$

Where, R_X is the rank of X taken over values $x_1, x_2, x_3, \cdots, x_n$ of vector X given by, $R_X = [r_{x1}, r_{x2}, r_{x3}, \cdots, r_{xn}]$. If any X values are tied, their average rank computed. Thus R_X and R_Y are co-ordinate rank vectors of X and Y respectively.

And $\overline{R_X} = \frac{1}{n}\sum_i r_{xi} = \frac{(n+1)}{2}$ and similarly, $\overline{R_Y} = \frac{1}{n}\sum_i r_{yi} = \frac{(n+1)}{2}$.

5.2.2 Hierarchical clustering

Clustering methods [50] can be hierarchical (grouping objects into clusters and specifying relationships among objects in a cluster, resembling a phylogenetic tree) or non-hierarchical (grouping into clusters without specifying relationships between objects in a cluster).

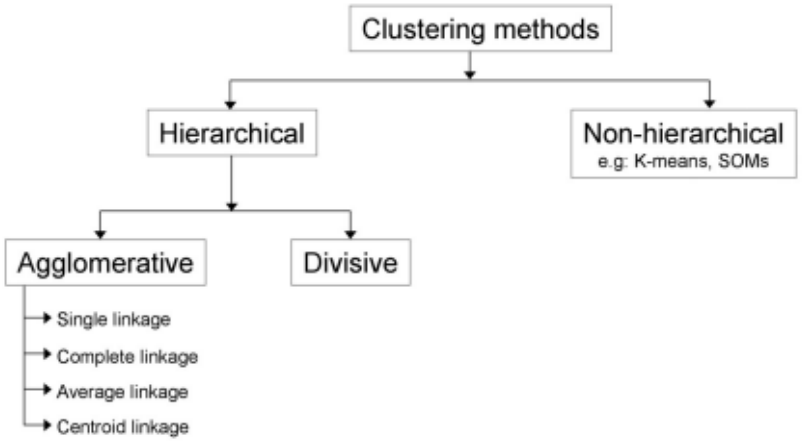

Figure 5.1 Overview of different clustering approaches. (Image Courtesy: *Introduction to microarray data analysis* [5])

Hierarchical clustering may be agglomerative (starting with the assumption of each object being a cluster in itself and grouping similar objects into bigger clusters) or divisive (starting from grouping all objects into one cluster and subsequently breaking the big cluster into smaller clusters with similar properties). Disadvantage for both agglomerative and divisive approaches is that their "greedy" nature prevents the refinement of the previous clustering. If a "bad" decision is made in the initial steps, it can never be corrected in the following steps.

Hierarchical clustering: agglomerative

In the case of a hierarchical agglomerative clustering which is a bottom-up approach, the objects are successively fused together until all the objects are included. For a hierarchical agglomerative clustering procedure, each object is considered as a cluster initially. On the first step pairwise distance measures for the objects to be clustered are calculated. Based on the pairwise distances between them, objects that are most similar to each other are grouped into clusters. After this is done, pairwise distances between the clusters are re-calculated, and clusters that are similar are grouped together in an iterative manner until all the objects are included into a single cluster. This information can be represented as a dendrogram, where the distance from the branch point indicates the distance between the two clusters or objects.

Hierarchical clustering: divisive

Hierarchical divisive clustering is the opposite of the agglomerative method and is a top-down approach, where the entire set of objects is considered as a single cluster and is broken down into two or more clusters that have similar expression profiles. After this is done, each cluster is considered separately and the divisive process is repeated iteratively until all objects have been separated into single objects. The division of objects into clusters on each iterative step may be decided upon by principal component analysis (PCA) which determines a vector that separates given objects.

5.2.3 Linkage Criterion

Comparison of clusters with another cluster or an object are carried out using three different approaches, For any two clusters C_i and C_j , and gene set given by $G = \{g_1, g_2, g_3, \cdots, g_M\}$ where each gene g_i has N samples of gene expression observation. Considering $\mathrm{dis}(g_u, g_v)$ to be the distance between two genes. The standard techniques to calculate the distance between clusters is discussed as follows.

Single linkage (Minimum distance)

In single linkage clustering, distance between two clusters is calculated as the minimum distance between all possible pairs of objects, one from each cluster. This method has an advantage that it is insensitive to outliers. This method is also known as the nearest neighbor linkage. It follows a

space-*contracting strategy*: tends to produce straggly clusters, which quickly agglomerate very dissimilar samples. Distance between two clusters C_i and C_j is given by,

$$Dis(C_i, C_j) = \min\{\text{dis}(g_{il}, g_{jm})\} \forall\, g_{il} \in C_i \text{ but } \notin C_j \text{ and } \forall g_{jm} \in C_j \text{ but } \notin C_i$$

Complete linkage (Maximum distance)

In complete linkage clustering, distance between two clusters is calculated as the maximum distance between all possible pairs of objects, one from each cluster. It follows a *space-dilating strategy*: produces clusters of very similar samples which agglomerate slowly. As clusters agglomerate, groups are moved away from each other. The disadvantage of this method is that it is sensitive to outliers. This method is also known as the farthest neighbor linkage. Distance between two clusters C_i and C_j is given by,

$$Dis(C_i, C_j) = \max\{\text{dis}(g_{il}, g_{jm})\} \forall\, g_{il} \in C_i \text{ but } \notin C_j \text{ and } \forall g_{jm} \in C_j \text{ but } \notin C_i$$

Average linkage

In average linkage clustering, distance between two clusters is calculated as the average of distances between all possible pairs of objects in the two clusters. It follows a space-conserving strategy; maximizes the cophenetic correlation, no reversals and eliminates group-size dependency. Distance between two clusters C_i and C_j is given by,

$$Dis(C_i, C_j) = \text{mean}\{\text{dis}(g_{il}, g_{jm})\} \forall\, g_{il} \in C_i \text{ but } \notin C_j \text{ and } \forall g_{jm} \in C_j \text{ but } \notin C_i$$

Figure 5.2 gives a visualization of the linkage criterion discusses above.

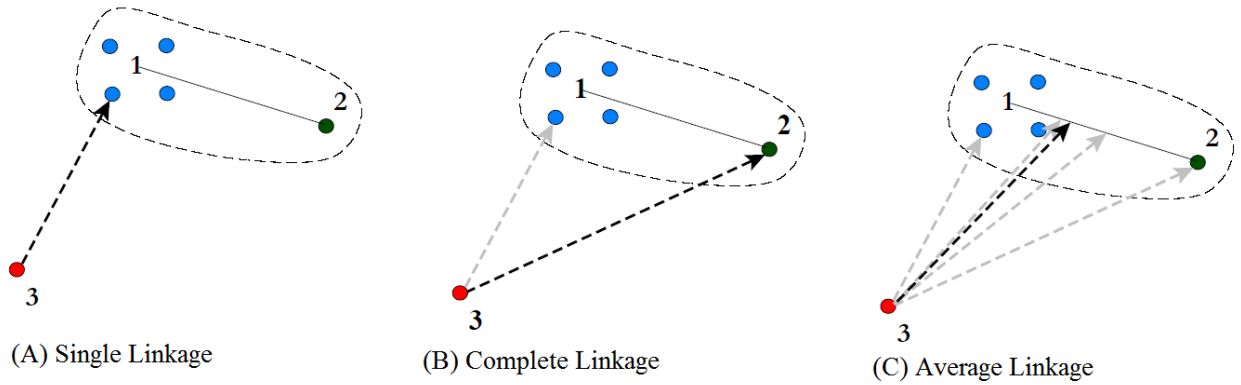

Figure 5.2 Various Linkage Criterions

5.2.4 Dendrogram

Hierarchical clustering generates a hierarchical series of nested clusters which can be graphically represented by a tree, called dendrogram. The branches of a dendrogram record the formation of the clusters as well as indicate the similarity between the clusters. Vertical lines extend up for each observation, and at various (dis)similarity values, these lines are connected to the lines from other observations with a horizontal line. The observations continue to combine until, at the top of the dendrogram, all observations are grouped together. The height of the vertical lines and the range of the (dis)similarity axis give visual clues about the strength of the clustering. Long vertical lines indicate more distinct separation between the groups. Long vertical lines at the top of the dendrogram indicate that the groups represented by those lines are well separated from one another. Shorter lines indicate groups that are not as distinct. For construction of a complete dendrogram, where leaf nodes corresponds to one data object and the root node corresponds to the whole dataset, the clustering process takes $\frac{n^2 - n}{2}$ merging or splitting steps. By cutting the dendrogram at some level, we can obtain a specified number of clusters.

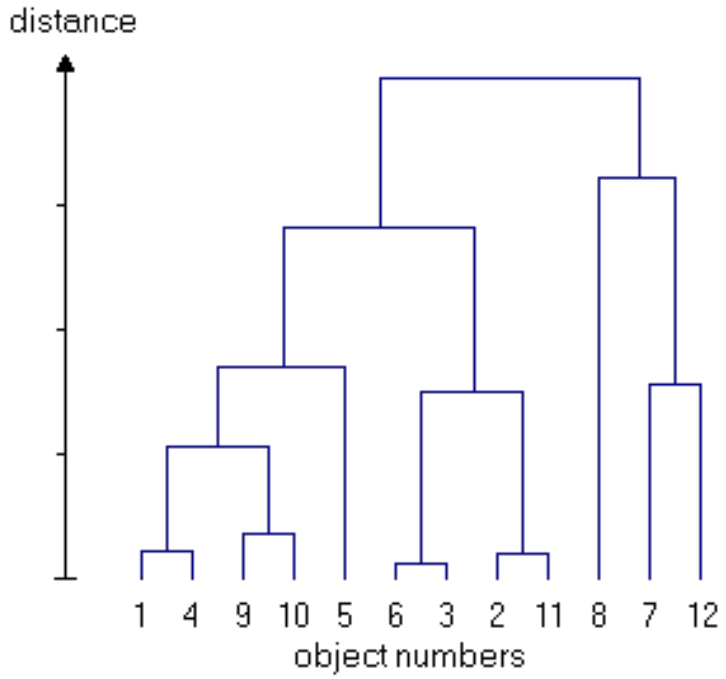

Figure 5.3 An example Dendrogram

Cutting the dendrogram at different depths produces different number of clusters as shown below.

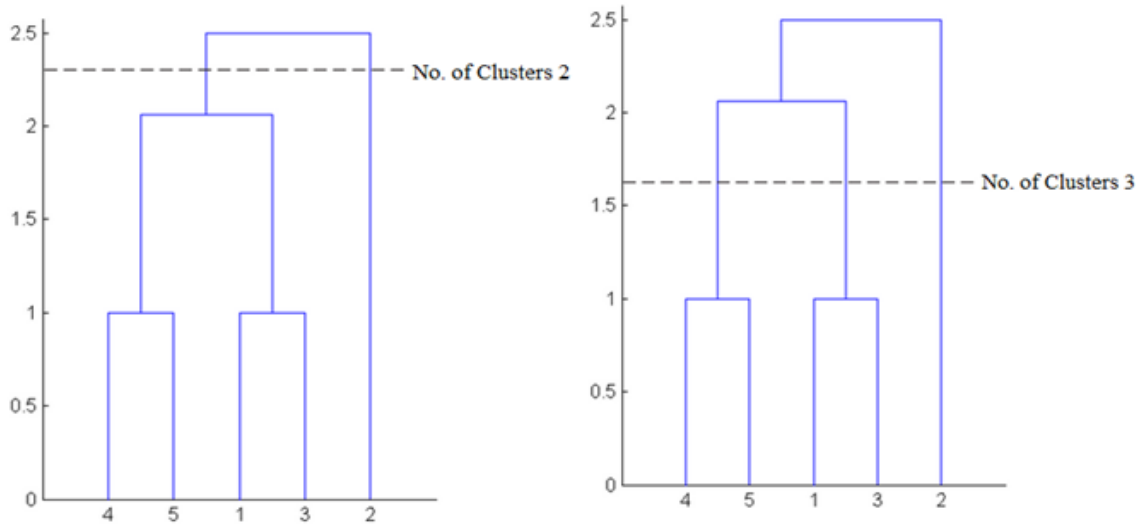

Figure 5.4 Cutting dendrogram at different depths

For example, performing agglomerative clustering on Fisher's Iris dataset [51] using centroid linkage and squared eucleidian distance measure then cutting the dendogram at second depth yeilds three distinct clusters corresponding to the three actual classes corresponding to the dataset. (Figure 5.5)

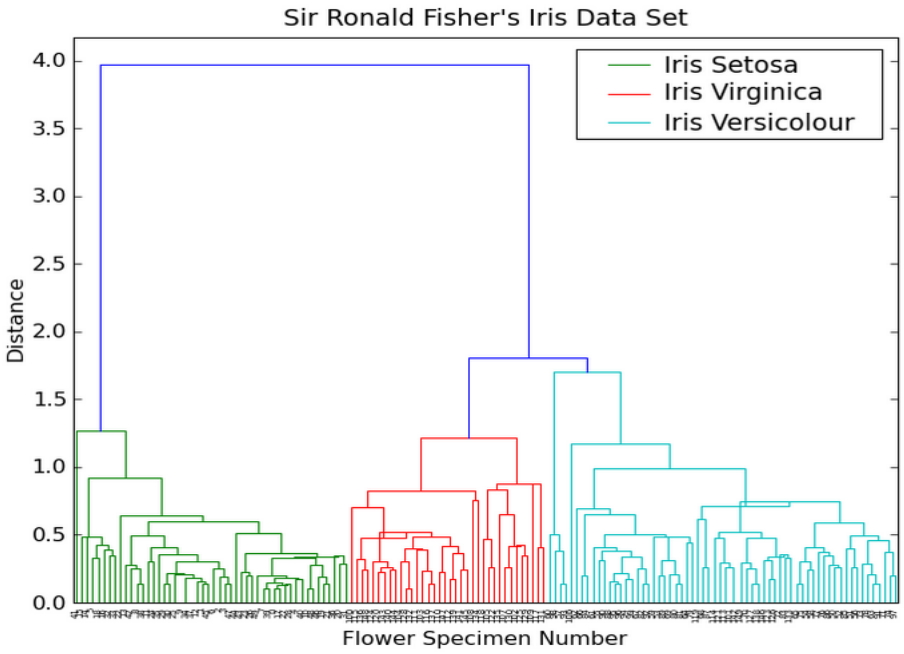

Figure 5.5 Hierarchical clustering on Iris dataset

5.2.5 Cluster Validity

The procedure of evaluating the results of a clustering algorithm is known as cluster validity. Almost every clustering algorithm depends on the characteristics of the dataset and on the input parameters. In general, clustering validity indices are usually defined in terms of compactness (i.e., measure of the closeness of the data elements of a cluster) and separability (i.e., measure of the intra-cluster distances between each of the distinct clusters). There are three approaches to study cluster validity as described below:

- ***External Criteria***

  This implies that the results of a clustering algorithm are evaluated based on a pre-specified structure imposed on a dataset, i.e. external information that is not contained in the dataset.

- ***Internal Criteria***

  This implies that the results of a clustering algorithm are evaluated using information that involves the vectors of the datasets themselves.

- ***Relative Criteria***

  This implies that the results of a clustering algorithm are evaluated by comparing them with other clustering schemes.

Cluster validity indices are used for measuring "goodness" of a clustering result comparing with other indices which are created by other clustering algorithms, or by the same algorithms but using different parameter values. A brief discussion of the validity indices is given below.

5.2.6 Cluster Validation for Hierarchical Clustering

Cophenetic Correlation Coefficient (CPCC) is a validity measure for hierarchical clustering algorithm used to represent how well the hierarchical structure from the dendrogram represents in two dimensions the multidimensional relationships within the data. CPCC is defined as the correlation between the $M = \frac{n(n-1)}{2}$ original pairwise dissimilarities between the feature vectors and the cophenetic dissimilarities from the dendrogram.

The ***cophenetic dissimilarity*** c_{ij} between the two feature vectors *i* and *j* is the inter-cluster distance at which the two feature vectors are first merged in the same cluster. CPCC is given by,

$$CPCC = \frac{\left(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^{P} c_{ij} - \mu_P \mu_c \right)}{\sqrt{\left[\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(d_{ij}^{P} \right)^2 - {\mu_P}^2 \right] \left[\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} {c_{ij}}^2 - {\mu_c}^2 \right]}}$$

Where, μ_P and μ_c are the means of the proximity and cophenetic matrices respectively. d_{ij}^{P} and c_{ij} respectively are the $(i,j)^{th}$ entry of the proximity and the cophenetic matrices. The concordance between the input data and the dendrogram is close if the value of the index is close to 1. A higher value of CPCC is regarded as a measure of successful classification. A value of 0.8 or above indicates that the dendrogram does not greatly distort the original structure in the input data.

In this work, CPCC is used to first compare clustering outputs from agglomerative hierarchical clustering of the dataset for all the aforesaid distance measures and linkage criterions. Clustering result yielding highest measure of CPCC is considered to be the optimal clustering results. And that result is further analyzed by dissection its dendrogram at different depths to determine the actual number of clusters in the dataset.

5.2.7 Cluster Validation for Partitional Clustering

For partitional clustering algorithms, most algorithms have to be initially supplied with the number of natural clusters in the dataset which is also known as the *k*-parameter. But this information is rarely known a priori, the usual approach is to run the algorithm several times with a different *k* value for each run. Then, all the partitions are evaluated and the partition that best fits the data is selected. From literature review it can be found that no single cluster validity index outperforms the rest [52, 53]. Henceforth the following set of validity indices were used in this work.

Notation

The gene expression dataset *X* of a set of N genes is represented as vectors in a d-dimensional feature space denoted as, $X = \{x_1, x_2, x_3, \cdots, x_N\} \subseteq \Re^d$. A clustering of *X* is a partition of *X* into *k*-groups given by, $C = \{c_1, c_2, c_3, ..., c_k\}$ such that, $\bigcup_{i=1}^{k} c_i = X$ and $c_i \bigcap c_j = \phi, \forall i \neq j$. Centroid of a cluster c_i, denoted by $\overline{c_i}$ is the mean vector of the data points belonging to that cluster, $\overline{c_i} = \frac{1}{|c_k|} \sum_{x_i \in c_k} x_i$. Whereas dataset centroid is given by, $\overline{X} = \frac{1}{N} \sum_{x_i \in X} x_i$. Distance between two objects x_i and x_j of the dataset is given by, $Dis(x_i, x_j)$.

5.2.7.1 Dunn's Index *[maximize]*

Dunn's Index [55] estimates the ratio of nearest neighbor to the maximum cluster diameter. For a good clustering the ratio between minimal inter-cluster distance to maximal intra-cluster distance has to be maximized.

$$Dunn's(C) = \frac{\min_{c_i \in C}\{\min_{c_j \in C \setminus c_i}\{\delta(c_i, c_j)\}\}}{\max_{c_i \in C}\{\Delta(c_i)\}}$$

where, $\delta(c_i, c_j) = \min_{x_k \in c_i}\left\{\min_{x_l \in c_j}\{Dis(x_k, x_l)\}\right\}$ and $\Delta(c_i) = \max_{x_k, x_l \in c_i}\{Dis(x_k, x_l)\}$.

5.2.7.2 Davies-Bouldin Index *[minimize]*

This index [56] aims to measure compactness by considering the distance between the points in a cluster to its centroid and separability by considering the distance between the centroids. The DB index is defined as:

$$DB(C) = \frac{1}{k} \sum_{c_i \in C} \max_{c_j \in C \setminus c_i}\left\{\frac{\sigma(c_i) + \sigma(c_j)}{Dis(\overline{c_i}, \overline{c_j})}\right\}$$

$$\text{where, } \sigma(c_i) = \frac{1}{|c_k|} \sum_{x_i \in c_k} Dis(x_i, \overline{c_i})$$

The clustering result that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best based on this criterion.

5.2.7.3 Calinski- Harabasz Index *[maximize]*

Calinski- Harabasz Index [57] measures the ration of distances from the points in a cluster to its centroid to the distance from the centroids to the global centroid. Maximizing this index produces better clustering result.

$$CH(C) = \frac{N-K}{K-1} \frac{\sum_{c_k \in C} |c_k| Dis\left(\overline{c_k}, \overline{X}\right)}{\sum_{c_k \in C} \sum_{x_i \in c_k} Dis\left(x_i, \overline{c_k}\right)}$$

5.2.7.4 Silhouette Index *[maximize]*

This index[58] is a normalized summation index which measures cohesion based on distance between all points in the same cluster and separation based on nearest neighbor distance and is defined by,

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{\mathrm{b}(x_i, c_k) - \mathrm{a}(x_i, c_k)}{\max\{\mathrm{a}(x_i, c_k), \mathrm{b}(x_i, c_k)\}}$$

$$\text{where } \mathrm{a}(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} Dis\left(x_i, x_j\right) \text{ and } \mathrm{b}(x_i, c_k) = \min_{c_l \in C \setminus c_k} \left\{ \frac{1}{|c_k|} \sum_{x_j \in c_l} Dis\left(x_i, x_j\right) \right\}$$

5.2.7.5 CS Index *[minimize]*

This ratio-type index[59] that estimates the cohesion by the cluster diameters and the separation by the nearest neighbor distance and is defined is

$$CS(C) = \frac{\sum_{c_k \in C} \left\{ \frac{1}{|c_k|} \sum_{x_i \in c_k} \max_{x_j \in c_k} \left\{ Dis\left(x_i, x_j\right) \right\} \right\}}{\sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \left\{ Dis\left(\overline{c_k}, \overline{c_l}\right) \right\}}$$

5.2.7.6 Sym-Index *[maximize]*

Sym-Index[60] is an variation of PBM index[53] based on point-symmetry distance given by,

$$SYM(\mathrm{C}) = \frac{\max_{c_k, c_l \in C} \left\{ Dis\left(\overline{c_k}, \overline{c_l}\right) \right\}}{K \sum_{c_k \in C} \sum_{x_i \in c_k} Dis_{PS}(x_i, c_k)}$$

where, $Dis_{PS}(x_i, c_k)$ is the point symmetry distance between object x_i and cluster c_k is given by, $Dis_{PS}(x_i, c_k) = \frac{1}{2}\sum \min(2)_{x_j \in c_k} \left\{ Dis\left(2\overline{c_k} - x_i, x_j\right) \right\}$ and $\sum \min(n)$ computes the sum of the n lowest values of its argument.

5.2.7.7 SV-Index *[maximize]*

SV-Index [61] estimates the separation by the nearest neighbor distance and the cohesion is based on the distance from the border points in a cluster to its centroid. It is defined as,

$$SV(C) = \frac{\sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \left\{ Dis\left(\overline{c_k}, \overline{c_l}\right) \right\}}{\sum_{c_k \in C} \frac{10}{|c_k|} \sum \max_{x_i \in c_k} \left(0.1|c_k|\right) \left\{ Dis\left(x_i, \overline{c_k}\right) \right\}}$$

5.3 Results

The previous module yields a set of 168 genes that are differentially expressed between healthy and diseased population setting the cut off q-value at 0.005. So the entire gene filtered dataset could be divided into two subsets. Dataset A corresponding to healthy smokers constitutes a set of 90 samples and Dataset B corresponding diseased smokers, constituting a set of 97 samples and both containing 168 genes. Hierarchical Agglomerative Clustering was performed on both Dataset A (healthy population dataset) and Dataset B (diseased population dataset).

Table 5.1 and Table 5.2 shows the CPCC values for hierarchical clustering performed on both Dataset A and Dataset B for various distance and linkage criterion mentioned in the previous sections.

Cluster Analysis on Healthy Population (Dataset A)

Table 5.1 Hierarchical clustering evaluation using CPCC of Dataset A

	Single Linkage	Average Linkage	Complete Linkage
Euclidean Distance	0.6515	0.8254	0.7230
Manhattan Distance	0.6669	0.8172	0.7182
Mahalanobis Distance	0.8319	0.8550	0.3216
Pearson Correlation Distance	0.8844	0.9118	0.8653
Spearman Correlation Distance	0.9046	**0.9224**	0.8843

Cluster Analysis on Diseased Population (Dataset B)

Table 5.2 Hierarchical clustering evaluation using CPCC of Dataset B

	Single Linkage	Average Linkage	Complete Linkage
Euclidean Distance	0.7024	0. 8263	0.7483
Manhattan Distance	0.7022	0.8457	0.7586
Mahalanobis Distance	0.8205	0.8437	0.3039
Pearson Correlation Distance	0.7071	0.8315	0.8128
Spearman Correlation Distance	0.6750	**0.8484**	0.7612

From Table 5.1 and 5.2 it is evident that spearman rank correlation coefficient distance with average linkage criterion yields highest CPCC value for both the datasets thus producing best clustering result among all other distances and linkage criterions considered. The dendrogram plots corresponding to the best clustering results for both datasets are as follows.

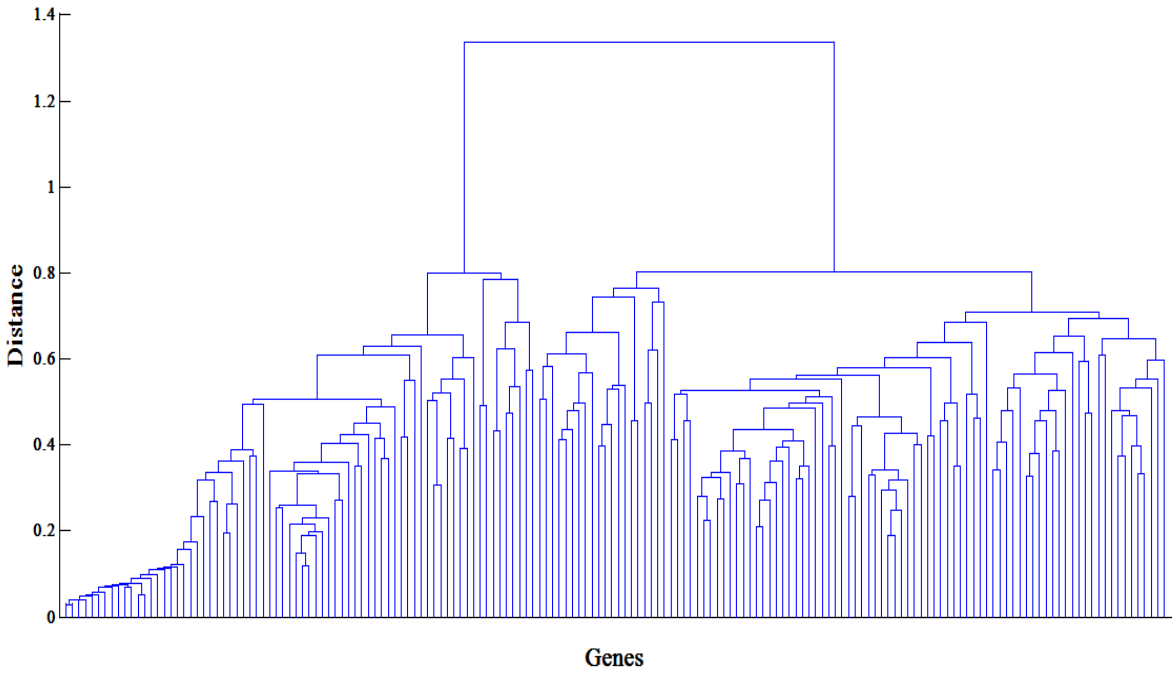

Figure 5.6 Dendrogram plot for Dataset A with Spearman Rank Correlation distance and average linkage criterion

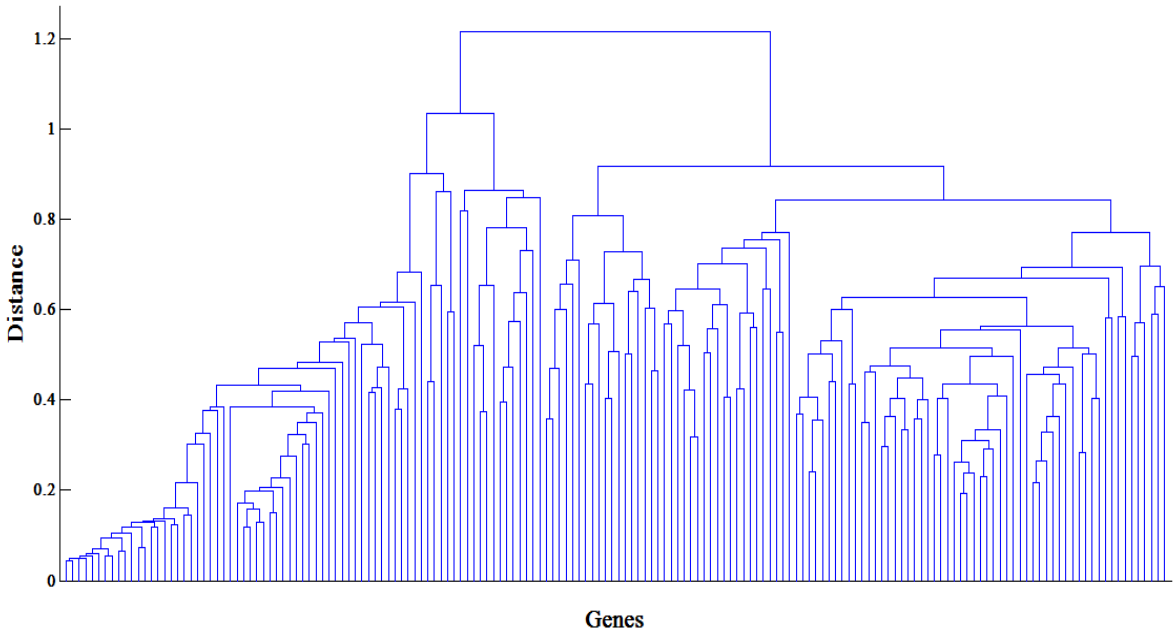

Figure 5.7 Dendrogram plot for Dataset B with Spearman Rank Correlation distance and average linkage criterion

Hierarchical clustering result of Dataset A and Dataset B using Spearman rank correlation coefficient with average linkage criterion are considered further for iteratively dissecting dendrogram linkages of the clustering results at different depths to find the natural clusters in the dataset. This procedure is done as follows; dissecting the dendrogram at depth 1 yields two clusters, while dissecting the dendrogram at depth 2 yields three clusters and so on. Thus dissecting the dendrogram linkage at depth i yields i+1 clusters in the dataset. The depth to which a dendrogram is dissected, determines the number of natural clusters within the dataset. This stopping criterion is achieved by means of the cluster validity indices for partitional clustering algorithms discussed previously. Dendrogram output of the average linkage Spearman rank correlation distance hierarchical agglomerative clustering is iteratively cut at depths to yield smaller clusters of genes. At each depth when the linkage merged at current highest distance is dissected to yield one more cluster than existing, the current clustering result is evaluated using all the validity indices. This iterative process of dissecting the linkages of the clusters is continued till a depth of 9 yielding a total of 10 clusters. Thus the clustering result was thus tested for k=2,3,4,…,10 clusters by iteratively dissecting dendrogram linkages at depths 1,2,3,…,9 respectively.

Determining the number of clusters in a dataset is intrinsically difficult because this is often a subjective process. Table 5.3 and 5.4 summarizes the results of evaluating the seven cluster validity indices considered to decide the optimal k-parameter for Dataset A and B respectively. The values in the table marked bold and shaded indicate the optimal value returned by that specific validity index over all k values considered.

Partitional clustering analysis using cluster validity indices

Table 5.3 Evaluation of cluster validity indices for different k-parameter for Dataset A

k-parameter	Dunn's Index	Davies-Bouldin Index	Calinski-Harabasz Index	Silhouette Index	CS Index	SYM Index	SV Index
2	0.519489466	**0.421794**	179.4794	0.320344	2.417697	0.123279	0.058357
3	**0.521822826**	1.285558	**240.179**	**0.341226**	3.007731	**0.202354**	**0.475018**
4	0.301566338	1.354472	127.1219	0.267722	3.691343	0.083151	0.375878
5	0.301566338	1.235196	96.22031	0.249841	2.543032	0.11748	0.397804
6	0.301566338	1.175929	77.90129	0.233563	**2.130993**	0.187262	0.417877
7	0.301566338	1.100681	65.56314	0.226116	2.929643	0.169686	0.275173
8	0.301566338	1.026924	56.36059	0.214371	4.345251	0.150266	0.187871
9	0.250779006	1.204307	51.56966	0.22683	5.736134	0.133218	0.151053
10	0.250779006	1.215522	46.92917	0.225899	6.390402	0.122311	0.132319

Table 5.4 Evaluation of cluster validity indices for different k-parameter for Dataset B

k-parameter	Dunn's Index	Davies-Bouldin Index	Calinski-Harabasz Index	Silhouette Index	CS Index	SYM Index	SV Index
2	0.323957822	**0.546018**	102.3888	0.3014	2.51132	0.0931	0.093128
3	0.339238222	0.900783	137.9075	0.3158	2.472717	0.1532	0.153163
4	**0.375366245**	1.121942	**240.9651**	**0.3796**	1.976578	**0.2221**	0.122059
5	0.339434355	1.061795	80.1508	0.2891	5.670765	0.0992	0.099245
6	0.296278403	1.031703	64.8691	0.2763	6.60191	0.1524	0.152398
7	0.296278403	0.965214	54.5249	0.2595	1.219989	0.1577	0.157732
8	0.296278403	0.883861	47.1872	0.2603	**1.067121**	0.1773	**0.177312**
9	0.352699333	0.936416	44.0637	0.2698	9.959585	0.1571	0.157077
10	0.352699333	0.850708	39.3389	0.274	8.445925	0.147	0.147001

From Table 5.3 it is found that for gene expression Dataset A corresponding to healthy population 5 cluster validity indices namely Dunn's Index, Calinski-Harabasz Index, Silhouette Index, SYM Index and SV Index have returned their corresponding optimal value corresponding to k-parameter of 3. While Davies-Bouldin Index and CS Index have returned minimum value corresponding k value of 2 and 6 respectively. Using a consensus strategy between the optimal k-parameters returned by the cluster validity indices, it is concluded that the set of 168 differentially expressed genes can be grouped into 3 gene clusters. The three clusters of this dataset consist of 20, 72 and 76 genes.

On the other hand, from Table 5.4 it can be concluded for the lung cancer affected gene expression dataset, i.e. Dataset B, after evaluating the results of all the cluster validity indices for various k-parameter that the genes should be grouped into 4 clusters as opposed to 3 clusters in case of healthy population dataset, i.e. Dataset A. This is because, 4 of 7 validity indices *viz.* Dunn's Index, Calinski- Harabasz Index, Silhouette Index and SYM Index yield an optimal k-parameter of 4 for the dataset amongst k=2 to 10. While 2 other indices namely, CS Index and SV Index decide on optimal k value of 8 and Davies-Bouldin Index decides on k-value of 2 which is also the result returned by it for healthy population Dataset A as well. Using the consensus strategy to decide the number of natural clusters in dataset, we conclude that genes corresponding to diseased dataset should be grouped into 4 clusters. The 4 clusters of this dataset consist of 13, 18, 60 and 77 genes.

Figures 5.3 and 5.4 dendrogram plots for optimal k-parameter for Dataset A and Dataset B respectively with the sub trees corresponding the 3 clusters of Dataset A and 4 clusters for Dataset B are marked using different colors. Tables 5.5 and 5.6 list the genes assigned to each of the 3 and 4 clusters of Dataset A and Dataset B respectively.

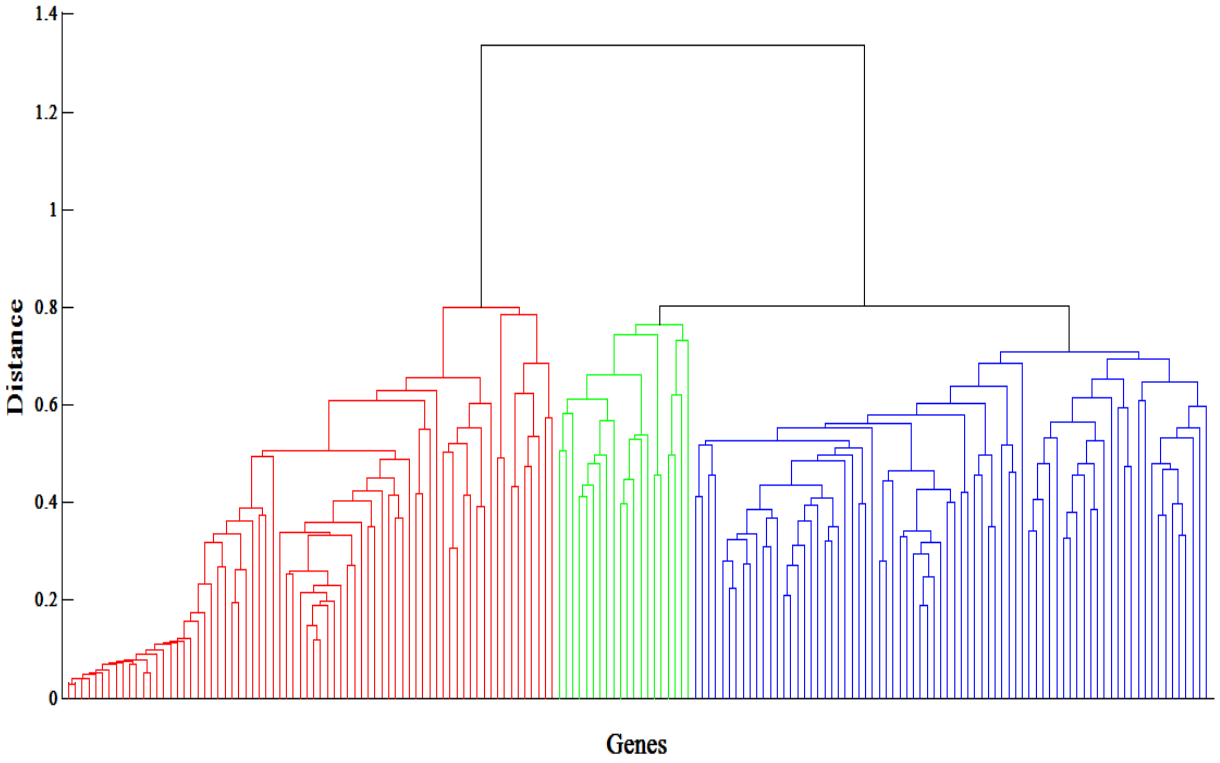

Figure 5.8 Dendrogram plot displaying the 3 clusters of Dataset A

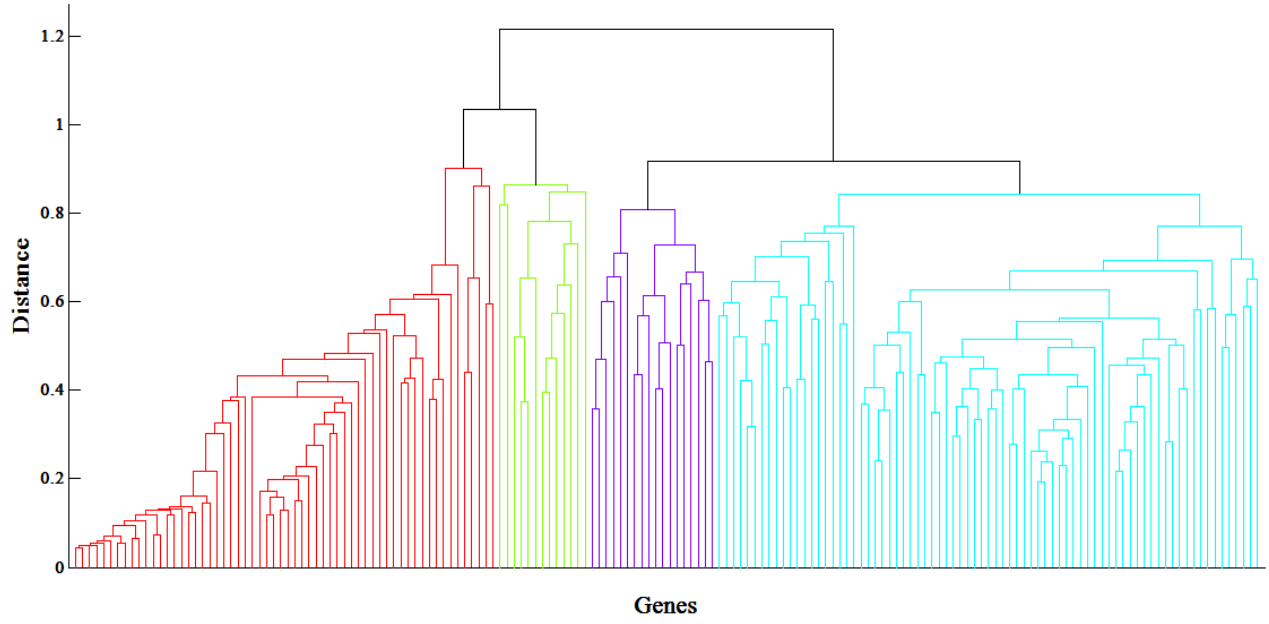

Figure 5.9 Dendrogram plot displaying the 4 clusters of Dataset B

Table 5.5 Cluster assignment table for Dataset A

Cluster #	No. of genes included	Gene belonging to the cluster					
1	20	C6	NELL2	BICC1	CWH43	DNAJC6	PLA1A
		FGF14	TRIM36	ODF2	HSPA1L	ZSCAN5A	SOX9
		TRDMT1	CLGN	NR2F1	ECI2	TMEM45A	HTRA1
		CST6	TMEM110				
2	72	HUWE1	USP34	TSC2	ALPK1	ITGAL	DIP2A
		BCAS1	EVPL	MAPK8IP3	PRRC2A	ATP8B1	ZNF611
		ZAP70	PLEKHA5	DUOX1	POLR1B	TAOK1	RIN1
		CLSTN3	KLHL24	PHACTR4	LRRC40	ZNF276	PRR11
		MYO1C	LRRFIP1	PTAFR	ZNF160	UBQLN4	GSDMB
		B2M	CCDC81	ZNF224	CARM1	SLC16A6	ZNF721
		DCLRE1C	MFSD11	CORO2A	ZNF652	CYP3A4	RHOT2
		RAPGEFL1	RBM4	ARGLU1	ARID5A	SGSM2	CRCP
		DAPP1	PTMA	NPTX1	CTSS	SLC35E1	RAB15
		HAUS2	RBCK1	ICAM4	GNPDA1	FAM193B	NCR3
		LHB	APOBEC3C	GTF2H3	MED6	FSCN1	SLC50A1
		APOC4	SAT1	ERN2	SET	NOTCH2NL	NPFFR1
3	76	TRIO	PDZD8	NCOA1	RPGRIP1L	WWC3	GIGYF2
		PRSS12	EXT2	SMC6	ARMCX2	HBB	DDX18
		SNX19	RAB3GAP1	CRY1	CPE	CDC5L	FUT8
		FXR1	SERPINI1	MPHOSPH10	PYGB	MAP2K4	DHCR7
		ZC3H14	ZFR	MAP7	ZKSCAN5	NSUN3	OSBP
		EIF2B3	RRAGB	GIN1	NARS2	AGPS	ACVR2A
		HIBCH	PIGK	RIPK1	DLAT	EFNB2	MPP5
		TMEM33	GFPT1	PRPS1	TTC33	ECD	MTPAP
		CIAO1	PPM1D	UBE2N	DCUN1D4	P2RX4	SLC39A14
		TGDS	XRCC4	PINK1	MGAT4A	UBFD1	MRPS15
		PRKAA1	PPP2R2D	TSN	ADK	GGCX	ZFYVE21
		RRAS2	MAK16	SSSCA1	SNRNP27	RNF146	KDELR3
		PBX1	FAM69A	CDK9	IER3		

Table 5.6 Cluster assignment table for Dataset B

Cluster #	No. of genes included	Gene belonging to the cluster					
1	13	HUWE1	BCAS1	HBB	LRRFIP1	SOX9	B2M
		ARGLU1	PTMA	CTSS	GNPDA1	SAT1	SET
		IER3					
2	18	C6	NELL2	CWH43	DNAJC6	PLA1A	TRIM36
		EIF2B3	CCDC81	CLGN	NR2F1	EFNB2	ECI2
		TMEM45A	HTRA1	CST6	TMEM110	SSSCA1	FAM69A
3	60	USP34	TSC2	ALPK1	ITGAL	DIP2A	EVPL
		MAPK8IP3	PRRC2A	ATP8B1	ZNF611	ZAP70	PLEKHA5
		DUOX1	POLR1B	TAOK1	RIN1	CLSTN3	KLHL24
		PHACTR4	LRRC40	ZNF276	PRR11	MYO1C	PTAFR
		ZNF160	UBQLN4	GSDMB	ZNF224	CARM1	SLC16A6
		ZNF721	DCLRE1C	MFSD11	CORO2A	CYP3A4	RHOT2
		RAPGEFL1	RBM4	ARID5A	SGSM2	CRCP	DAPP1
		NPTX1	SLC35E1	RAB15	HAUS2	RBCK1	ICAM4
		FAM193B	NCR3	LHB	APOBEC3C	GTF2H3	MED6
		FSCN1	SLC50A1	APOC4	ERN2	NOTCH2NL	NPFFR1
4	77	TRIO	BICC1	PDZD8	NCOA1	RPGRIP1L	WWC3
		GIGYF2	PRSS12	EXT2	SMC6	ARMCX2	DDX18
		SNX19	RAB3GAP1	CRY1	CPE	CDC5L	FUT8
		FXR1	FGF14	SERPINI1	MPHOSPH10	PYGB	MAP2K4
		ODF2	DHCR7	HSPA1L	ZC3H14	ZFR	MAP7
		ZKSCAN5	NSUN3	ZSCAN5A	OSBP	RRAGB	GIN1
		NARS2	AGPS	ACVR2A	TRDMT1	HIBCH	PIGK
		RIPK1	DLAT	ZNF652	MPP5	TMEM33	GFPT1
		PRPS1	TTC33	ECD	MTPAP	CIAO1	PPM1D
		UBE2N	DCUN1D4	P2RX4	SLC39A14	TGDS	XRCC4
		PINK1	MGAT4A	UBFD1	MRPS15	PRKAA1	PPP2R2D
		TSN	ADK	GGCX	ZFYVE21	RRAS2	MAK16
		SNRNP27	RNF146	KDELR3	PBX1	CDK9	

5.4 Analysis of Cluster Assignment of Genes

From the analysis of the distribution of genes to different clusters in Dataset A and Dataset B, the following conclusions could be drawn. We consider the three clusters corresponding to the healthy population Dataset A to be the true clustering result. We see that, there is an overlap of 13 genes between Cluster# 1 of Dataset A and Cluster# 2 of Dataset B which accounts for a similarity of 72.22% between these clusters with respect to Cluster# 2 of Dataset B. Similarly, overlap of all 60 genes of Cluster# 3 of Dataset B is found with Cluster# 2 of Dataset A accounting for a similarity of 100% between these clusters with respect to Cluster# 3 of Dataset B. Cluster# 3 of Dataset A and Cluster# 4 of Dataset B overlaps by a gene count of 69 attributing to similarity of 89.61% between these clusters with respect to Cluster# 4 of Dataset B. Thus there is a high one-to-one correspondence between the Cluster# 1, 2 and 3 of Dataset A and Cluster# 2, 3 and 4 of Dataset B. The extra cluster found in case of diseased population or Dataset B is majorly formed by partitioning out genes from Cluster# 2 of Dataset A. Out of 13 genes in Cluster# 1 of Dataset B eight genes belong to Cluster# 2 of Dataset A namely, HUWE1, BCAS1, LRRFIP1, B2M, ARGLU1, PTMA, CTSS, GNPDA1, SAT1 and SET. HBB and IER3 genes of Cluster# 1 of Dataset B correspond to Cluster# 3 of Dataset A and SOX9 gene comes from Cluster# 1 of Dataset A. Thus it concluded that expression profile of these genes differ significantly to form a new cluster in case of the diseased population. Six other genes namely, BICCI, FGF14, ODF2, HSPAIL, ZSCZN5A and TRDTMT1 belonging to Cluster# 1 in Dataset A has been assigned to Cluster# 4 of Dataset B thus displaying correlation of expression profile with genes with a different gene cluster in diseased dataset. Genes EIF2B3, EFNB2, FAM69A and SSSCA1 belonging to Cluster# 3 in Dataset A has correlated expression with genes of Cluster# 2 of Dataset B hence showing variation in cluster assignments. Two other genes assigned to different clusters in healthy and diseased population are CCDC81 and ZNF652. Thus we identify 12 genes showing different cluster assignments in healthy and diseased population and 13 genes showing significant change in expression profile so as branch out forming a new cluster in diseased population Dataset B.

5.5 Biological Validation of Clustering Results

In this section we attempt to make biological conclusions regarding the three major clusters of healthy population Dataset so as to make functional prediction of genes belonging to these

clusters. The incorporation of biological knowledge in gene cluster evaluation enables a better biological interpretation of clustering results. Gene ontology (GO) [9] provides a common language to describe aspects of a gene product's biology, and is represented in a taxonomic form. The use of a consistent vocabulary allows genes from different species to be compared based on their GO annotations. The gene ontology term shared by genes belonging to the same cluster is used for biological enrichment of the clustering result.

5.5.1 Gene Ontology

Gene ontology is a controlled vocabulary describing gene products and related functions. It depicts a relational structure of genes present within the gene sets according to their biological descriptions (annotations). It seems to be the most complete database which is well-curated and up-to-date. The gene ontology provides descriptions about gene products in different databases in terms of their associated biological process (BP), cellular components (CC), and molecular function (MF). It provides a grouping of genes into biologically meaningful categories, at various levels of specificity. The controlled vocabularies of terms are structured to allow annotation of gene products to GO terms at varying levels of detail and to query for gene products that are involved in similar processes, function and components. The categories are gathered in a directed acyclic graph, with the genes being annotated to GO nodes.

For biological enrichment of the clustering results, the approach is to identify the common GO terms shared between genes belonging to same cluster. We use Generic GO Term Finder tool (http://go.princeton.edu/) developed within the Bioinformatics Group at the Lewis-Sigler Institute for this purpose. This web tool finds the significant GO terms shared among a list of genes from the organism of choice, thus helping to discover what these genes may have in common. The degree of enrichment of the GO terms shared by genes of a cluster is reported by the tool in terms of p-values. A smaller p-value, close to zero, is indicative of a stronger evidence of the genes Annotated to the corresponding GO term being actually true. The list of genes belonging to a certain cluster is provided to the tool with organism of choice being Homo sapiens and the gene annotation file being ***gene_association.goa_human***.

In Table 5.7 and Table 5.8 the list of GO biological process (BP) terms and molecular function (MP) terms respectively shared between genes of Cluster# 1 of Dataset A is given. From Table 5.7 it is concluded that genes belonging to this cluster predominantly take part in

cellular metabolic process (GO: 0044237 with a corresponding p-value of 0.004534) which attributes for 75% of the genes belonging to this cluster. Other biological process involvement of genes of this cluster mostly includes regulatory mechanisms. For example, regulation of cellular component organization (GO: 0051128 with p-value 0.000635 corresponding to 30% of genes), regulation of response to stress (GO: 0080134 with p-value 0.002823 corresponding to 25% of genes), regulation of response to stimulus (GO: 0048583 with p-value 0.014725 corresponding to 30% of genes). Other regulatory mechanisms include regulation of transcription from RNA polymerase II promoter (GO: 0006357 with p-value 0.01941), regulation of immune effector process (GO: 0002697 with p-value 0.000658), positive regulation of epithelial cell proliferation (GO: 0050679 with p-value 0.001952). Hence this cluster corresponds to genes taking part in vital regulatory mechanisms. Analyzing the molecular function (MF) GO terms shared by genes of this cluster from Table 5.8 it can be concluded that the major molecular functional involvement of genes of this cluster is binding (GO:0005488 with p-value 0.047218 for 75% of genes) more specifically protein binding (GO:0005488 with p-value 0.003164 for 60% of genes). Other binding activities include sequence-specific DNA binding RNA polymerase II transcription factor activity (GO:0000981 with p-value 0.005361), enzyme binding(GO:0019899 with p-value 0.0469), RNA binding (GO:0003723 with p-value 0.041431), unfolded protein binding (GO:0051082 with p-value 0.002513), sulfur compound binding (GO:1901681 with p-value 0.006436), transcription regulatory region DNA binding(GO:0044212 with p-value 0.046727). Thus it is concluded that the genes in Cluster# 1 of healthy population Dataset A corresponds to involvement in regulatory processes with chief molecular function being binding of protein, enzyme, RNA and transcription factor. Hence we denote this to be the ***regulatory cluster***.

Table 5.9 and Table 5.10 the list of GO biological process (BP) terms and molecular function (MP) terms shared between genes of Cluster# 2 of Dataset A respectively. From Table 5.9 it can be concluded that the prime biological process involvement of genes of this cluster is response, signaling and communication mechanisms. This is evident as genes have the following shared GO biological process terms like response to stimulus (GO: 0050896 with p-value 1.42E-06 for 65.55% of genes), cell communication (GO: 0007154 with p-value 8.44E-05 for 53.05% of genes), signal transduction (GO:0007165 with p- value 0.000218 for 38.89% of genes), regulation of cell communication (GO:0010646 with p- value 0.0004 for 22.2% of genes).

Response activities of the genes corresponds to immune response (GO:0006955 with p- value 2.41E-05), defense response(GO:0006952 with p- value 0.000126) and response to stress (GO:0006950 with p- value 4.40E-05). The molecular function (MF) GO terms shared by genes of this cluster is listed in Table 5.10 and it can be concluded that the dominant function of genes of this cluster is ion binding (GO:0043167 with p-value 0.009389922 for 65.5% of genes). Heterocyclic compound binding (GO:1901363 with p-value 0.026044918 for 38.89% of genes), cation binding (GO:0043169 with p-value 0.005664816 for 29.16% of genes). Other molecular binding functions of this cluster include lipid binding (GO:0008289 with p-value 0.023521054), phospholipid binding (GO:0005543 with p-value 0.017140364), binding, bridging (GO:0060090 with p-value 0.004977341), phosphatidylinositol bisphosphate binding (GO:1902936 with p-value 0.009556571). Therefore this cluster of 72 genes is identified as the ***response and signaling cluster.***

The list of GO biological process (BP) terms and molecular function (MP) terms shared terms between genes of Cluster# 3 of Dataset A is reported in Table 5.11 and 5.12 respectively. From Table 5.11 it can be concluded that genes of Cluster# 3 primarily take part in cell development and maintenance process. This is evident from the fact that GO BP terms shared between the genes are single-organism developmental process(GO:0044767 with p-value 6.34E-06 for 35.52% of genes), anatomical structure development(GO:0048856 with p-value 6.93E-06 for 32.89% of genes), cell differentiation(GO:0030154 with p-value 0.000472 for 32.36% of genes), organ development(GO:0048513 with p-value 6.15E-05 for 21.05% of genes), cell proliferation(GO:0008283 with p-value of 0.000142), cell death(GO:0008219 with p-value 0.000648), regulation of cell death(GO:0010941 with p-value 0.00099). Hence the major biological process that genes of this cluster take part is cell development and maintenance where maintenance includes crucial functionalities like cell proliferation, cell death procedure control. The GO MP terms shared by genes of this cluster listed in Table 5.12 implies that dominant molecular functions shared by genes of this cluster include organic cyclic compound binding (GO:0097159 with p-value 1.39E-05 for 51.31% of genes), heterocyclic compound binding(GO:1901363 with p-value 3.12E-05 for 50% of genes), catalytic activity (GO:0003824 with p-value 0.002553 for 47.36% of genes) , transferase activity (GO:0016740 with p-value 3.00E-05 for 28.94% of genes), carbohydrate derivative binding (GO:0097367 with p-value 0.000288 for 25% of genes). These molecular functions of catalytic activity, carbohydrate

derivative binding are essential for supply of energy to the cell for its structural development and also for the essential process of apoptosis or cell death which when dis-regulated is the prime cause for cancer. Hence this cluster is denoted as the ***cell development and maintenance cluster.*** The following tables give detailed results of the biological process and molecular function enrichment of genes belonging to the 3 clusters of Dataset A. This functional enrichment is done by the GO Term Finder tool using the GO annotation file gene_association.goa_human.

GO terms shared by genes of Cluster 1 (20 gene cluster) of Dataset A

Table 5.7 Terms from the Process Ontology of gene_association.goa_human with p-value < 0.05 shared by genes of Cluster# 1 of Dataset A

GOID	GO TERM from biological_process Ontology	p-value	% of Genes of Cluster Annotated	Annotated Genes
GO:0044237	cellular metabolic process	0.004534	75	CST6, TRDMT1, ECI2, CLGN, SOX9, PLA1A, HTRA1, DNAJC6, CWH43, HSPA1L, ZSCAN5A, C6, FGF14, NR2F1, TRIM36
GO:0051128	regulation of cellular component organization	0.000635	30	ODF2, SOX9, HSPA1L, HTRA1, DNAJC6, NR2F1
GO:0048583	regulation of response to stimulus	0.014725	30	SOX9, HSPA1L, HTRA1, C6, FGF14, BICC1
GO:0022414	reproductive process	0.000166	25	ODF2, CLGN, SOX9, HSPA1L, TRIM36
GO:0009653	anatomical structure morphogenesis	0.008663	25	CST6, ODF2, SOX9, HTRA1, C6
GO:0080134	regulation of response to stress	0.002823	25	HSPA1L, HTRA1, C6, FGF14
GO:0009719	response to endogenous stimulus	0.005762	20	TRDMT1, SOX9, HTRA1, NR2F1
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.019413	20	SOX9, ZSCAN5A, HTRA1, NR2F1
GO:0002697	regulation of immune effector process	0.000658	15	HTRA1, C6, FGF14
GO:0050679	positive regulation of epithelial cell proliferation	0.001952	10	SOX9, HTRA1

Table 5.8 Terms from the Function Ontology of gene_association.goa_human with p-value < 0.05 shared by genes of Cluster# 1 of Dataset A

GOID	GO TERM from molecular_function Ontology	p-value	% of Genes of Cluster Annotated	Annotated Genes
GO:0005488	Binding	0.047218	75	TRDMT1, NELL2, ECI2, CLGN, SOX9, HTRA1, DNAJC6, BICC1, ODF2, HSPA1L, ZSCAN5A, C6, FGF14, NR2F1, TRIM36
GO:0005515	protein binding	0.003164	60	NELL2, ECI2, CLGN, SOX9, HTRA1, DNAJC6, ODF2, HSPA1L, C6, FGF14, NR2F1, TRIM36
GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity	0.005361	15	SOX9, ZSCAN5A, NR2F1
GO:0019899	enzyme binding	0.0469	15	ODF2, SOX9, HSPA1L
GO:0003723	RNA binding	0.041431	15	TRDMT1, SOX9, BICC1
GO:0003700	sequence-specific DNA binding transcription factor activity	0.049806	15	SOX9, ZSCAN5A, NR2F1
GO:0051082	unfolded protein binding	0.002513	10	CLGN, HSPA1L
GO:1901681	sulfur compound binding	0.006436	10	ECI2, FGF14
GO:0044212	transcription regulatory region DNA binding	0.046727	10	SOX9, NR2F1

GO terms shared by genes of Cluster 2 (72 gene cluster) of Dataset A

Table 5.9 Terms from the Process Ontology of gene_association.goa_human with p-value <0.001 shared by genes of Cluster# 2 of Dataset A

GOID	GO TERM from biological_process Ontology	p-value	% of Genes of Cluster Annotated	Some Annotated Genes
GO:0050896	response to stimulus	1.42E-06	65.55556	DAPP1, ZAP70, DCLRE1C, UBQLN4, NPTX1, LHB, PTAFR, SGSM2, DUOX1, RBM4
GO:0007154	cell communication	8.44E-05	53.05556	CRCP, DAPP1, RAPGEFL1, RIN1, RHOT2, ZAP70, NPFFR1, TAOK1, UBQLN4, KLHL24
GO:0044700	single organism signaling	0.00014	41.66667	CRCP, DAPP1, RAPGEFL1, RIN1, RHOT2, ZAP70, NPFFR1, TAOK1, CLSTN3, KLHL24
GO:0007165	signal transduction	0.000218	38.88889	CRCP, DAPP1, RAPGEFL1, RIN1, RHOT2, ZAP70, NPFFR1, TAOK1, NOTCH2NL, KLHL24
GO:0044707	single-multicellular organism process	1.64E-05	37.5	CRCP, RAPGEFL1, RIN1, ZAP70, DCLRE1C, CLSTN3, NOTCH2NL, NPTX1, LHB, RBM4
GO:0048518	positive regulation of biological process	5.19E-06	36.11111	CRCP, APOC4, RAPGEFL1, RIN1, GTF2H3, ZAP70, TAOK1, RAB15, ALPK1, RBM4
GO:0010468	regulation of gene expression	0.000363	31.94444	GTF2H3, ZNF224, ZNF652, ARID5A, ZNF611, ERN2, SET, ARGLU1, APOBEC3C, RBM4
GO:0044767	single-organism developmental process	0.00053	30.55556	RAPGEFL1, ZAP70, DCLRE1C, CLSTN3, NOTCH2NL, NPTX1, LHB, DIP2A, POLR1B, RBM4
GO:0006950	response to stress	4.40E-05	29.16667	CRCP, GTF2H3, ZAP70, TAOK1, DCLRE1C, UBQLN4, ERN2, PTAFR, APOBEC3C, RBM4
GO:0002376	immune system process	2.78E-05	22.22222	CRCP, ALPK1, TSC2, ZNF160, ZAP70, NCR3, MYO1C, RBCK1, ICAM4, APOBEC3C
GO:0010646	regulation of cell communication	0.0004	22.22222	SGSM2, RAPGEFL1, RIN1, TSC2, RHOT2, ZAP70, MAPK8IP3, MYO1C, RBCK1, KLHL24
GO:0035556	intracellular signal transduction	0.000794	22.22222	SGSM2, CORO2A, RAPGEFL1, RIN1, TSC2, RHOT2, ZAP70, MAPK8IP3, RBCK1, RBM4
GO:0006955	immune response	2.41E-05	18.05556	CRCP, TSC2, ZAP70, NCR3, MYO1C, RBCK1, ICAM4, CTSS, LRRFIP1, APOBEC3C
GO:0006952	defense response	0.000126	16.66667	CRCP, ALPK1, TSC2, ZAP70, NCR3, MYO1C, CTSS, LRRFIP1, B2M, APOBEC3C

Table 5.10 Terms from the Function Ontology of gene_association.goa_human with p-value <0.05 shared by genes of Cluster# 2 of Dataset A

GOID	GO TERM from molecular_function Ontology	p-value	% of Genes of Cluster Annotated	Annotated Genes
GO:0043167	ion binding	0.009389922	60.27778	DAPP1, GTF2H3, RHOT2, ZNF224, ZAP70, ZNF652, ZNF611, TAOK1, CLSTN3, NOTCH2NL, ERN2, NPTX1, PTAFR, APOBEC3C, RAB15, POLR1B, ALPK1, DUOX1, ZNF160, PLEKHA5, RBM4
GO:1901363	heterocyclic compound binding	0.026044918	38.88889	CRCP, GTF2H3, RHOT2, ZNF224, ZAP70, ZNF652, ARID5A, ZNF611, TAOK1, ERN2, SET, FSCN1, APOBEC3C, RAB15, PRRC2A, POLR1B, ALPK1, DUOX1, ZNF160, MYO1C, RBM4
GO:0043169	cation binding	0.005664816	29.16667	GTF2H3, RHOT2, ZNF224, ZNF652, ZNF611, CLSTN3, NOTCH2NL, ERN2, NPTX1, APOBEC3C, POLR1B, DUOX1, ZNF160, ZNF276, ZNF721, RBCK1, CYP3A4, SAT1, ATP8B1, ITGAL, RBM4
GO:0032403	protein complex binding	0.01777597	8.333333	ICAM4, CORO2A, EVPL, CTSS, FSCN1, ITGAL
GO:0008289	lipid binding	0.023521054	6.944444	DAPP1, CYP3A4, PLEKHA5, PTAFR, ATP8B1
GO:0005543	phospholipid binding	0.017140364	5.555556	DAPP1, PLEKHA5, PTAFR, ATP8B1
GO:0060090	binding, bridging	0.004977341	4.166667	EVPL, MAPK8IP3, FSCN1
GO:1902936	phosphatidylinositol bisphosphate binding	0.009556571	2.777778	DAPP1, PLEKHA5

GO terms shared by genes of Cluster 3 (76 gene cluster) of Dataset A

Table 5.11 Terms from the Process Ontology of gene_association.goa_human with p-value <0.001 shared by genes of Cluster# 3 of Dataset A

GOID	GO TERM from biological_process Ontology	p-value	% of Genes of Cluster Annotated	Some Annotated Genes
GO:0019222	regulation of metabolic process	0.000159	43.42105	CDK9, MAK16, TSN, PRPS1, NSUN3, ACVR2A, RRAS2, MRPS15, ZFR
GO:0044767	single-organism developmental process	6.34E-06	35.52632	CDK9, MAK16, OSBP, TSN, PRPS1, NSUN3, ACVR2A, RRAS2, ZFR
GO:0048856	anatomical structure development	6.93E-06	32.89474	CDK9, MGAT4A, UBE2N, EXT2, PRKAA1, TRIO, PRPS1, NSUN3, PINK1
GO:0030154	cell differentiation	0.000472	32.36842	CDK9, RIPK1, SMC6, UBE2N, PRKAA1, TRIO, ADK, PRPS1, P2RX4
GO:0048513	organ development	6.15E-05	21.05263	CDK9, RIPK1, ACVR2A, MAP2K4, PRKAA1, PINK1, TRIO
GO:0008283	cell proliferation	0.000142	15.78947	RIPK1, ACVR2A, CDC5L, FUT8, OSBP, PINK1, MPP5
GO:0008219	cell death	0.000648	15.78947	CDK9, SMC6, UBE2N, PRKAA1, TRIO, PRPS1, ACVR2A, RRAS2, P2RX4
GO:0010941	regulation of cell death	0.00099	13.15789	CDK9, RIPK1, PRKAA1, TRIO, ADK, PRPS1, ACVR2A, MAP2K4, PINK1
GO:0043687	post-translational protein modification	2.64E-06	9.210526	CDK9, TSN, PRPS1, NSUN3, ACVR2A, RRAS2, TGDS, DDX18, MAP2K4
GO:0006281	DNA repair	0.000851	9.210526	CDK9, RIPK1, PRKAA1, TRIO, ADK, PRPS1, ACVR2A, MAP2K4, PINK1
GO:0044770	cell cycle phase transition	3.33E-05	9.210526	RIPK1, CPE, RPGRIP1L, NCOA1, EFNB2, CRY1, P2RX4
GO:0010564	regulation of cell cycle process	6.58E-05	9.210526	CDK9, PBX1, NCOA1, CRY1
GO:1901214	regulation of neuron death	0.000786	5.263158	CDK9, RIPK1, SMC6, UBE2N, PRKAA1, TRIO, ADK, PRPS1, P2RX4

Table 5.12 Terms from the Function Ontology of gene_association.goa_human with p-value <0.05 shared by genes of Cluster# 3 of Dataset A

GOID	GO TERM from molecular_function Ontology	p-value	% of Genes of Cluster Annotated	Annotated Genes
GO:0097159	organic cyclic compound binding	1.39E-05	51.31579	CDK9, MAK16, OSBP, TSN, PRPS1, NSUN3, ACVR2A, RRAS2, ZFR
GO:1901363	heterocyclic compound binding	3.12E-05	50	CDK9, MAK16, TSN, PRPS1, NSUN3, ACVR2A, RRAS2, MRPS15, ZFR
GO:0003824	catalytic activity	0.002553	47.36842	CDK9, TSN, PRPS1, NSUN3, ACVR2A, RRAS2, TGDS, DDX18, MAP2K4
GO:0016740	transferase activity	3.00E-05	28.94737	CDK9, MGAT4A, UBE2N, EXT2, PRKAA1, TRIO, PRPS1, NSUN3, PINK1
GO:0097367	carbohydrate derivative binding	0.000288	25	CDK9, SMC6, UBE2N, PRKAA1, TRIO, PRPS1, ACVR2A, RRAS2, P2RX4
GO:0005524	ATP binding	0.000405	19.73684	CDK9, RIPK1, SMC6, UBE2N, PRKAA1, TRIO, ADK, PRPS1, P2RX4
GO:0032559	adenyl ribonucleotide binding	0.000453	19.73684	CDK9, RIPK1, SMC6, UBE2N, PRKAA1, TRIO, ADK, PRPS1, P2RX4
GO:0016772	transferase activity, transferring phosphorus-containing groups	0.004367	14.47368	CDK9, RIPK1, PRKAA1, TRIO, ADK, PRPS1, ACVR2A, MAP2K4, PINK1
GO:0004674	protein serine/threonine kinase activity	0.000934	9.210526	CDK9, RIPK1, ACVR2A, MAP2K4, PRKAA1, PINK1, TRIO
GO:0044822	poly(A) RNA binding	1.15E-08	18.42105	MAK16, TSN, UBE2N, ADK, UBFD1, GIGYF2, CDC5L, MRPS15, ZC3H14
GO:0005102	receptor binding	0.037776	9.210526	RIPK1, CPE, RPGRIP1L, NCOA1, EFNB2, CRY1, P2RX4
GO:0019904	protein domain specific binding	0.000125	9.210526	RIPK1, ACVR2A, CDC5L, FUT8, OSBP, PINK1, MPP5

5.6 Discussion

In this work the objective was to employ clustering algorithm to identify groups of co-expressed genes and analyze how the gene clusters vary between the healthy and diseased population. We use hierarchical agglomerative clustering algorithm to organize the genes in a hierarchical tree. The dendrogram is then dissected iteratively at varying depths to find natural non-overlapping clusters in the dataset. We identify three clusters corresponding to the healthy gene expression data and four clusters corresponding to the gene expression data of diseased population. The healthy gene clusters were biologically enriched using gene ontology terms shared between genes of same cluster.

Chapter 6

Construction of Gene Regulatory Network

Gene Regulatory Networks (GRNs) are the most important organizational level in the cell where signals from the cell state and the outside environment are integrated in terms of activation and inhibition of genes. The genetic network or gene regulatory network (GRNs) is a kind of a biological pathway mapped to a graph of connected genes, gene products (in the form of protein complexes or protein families) or their groups, as vertices or nodes, through weighted/unweighted edges. A "connection" between two genes connotes a regulatory interaction between the genes. It represents a complex structure consisting of various gene products activating or repressing other gene products. Co-regulated genes, which encode proteins interacting among themselves and participating in common biological processes, may be grouped in the form of gene clusters. Clustering of gene expression patterns is used to identify groups of co-expressed genes. Genes having similar gene expression profiles are more likely to regulate one another or be regulated by some other common parent gene. Gene clusters can be interpreted as a network of co-regulated genes encoding interacting proteins that are involved in the same biological processes. The behavior of complex cancer cell networks cannot be deduced by intuitive approaches. Instead, it requires sophisticated and elegant network models and

computational analysis and simulation. Cancer cell network models will aid in the generation of experimentally testable hypotheses and discovery of the underlying mechanisms of tumorigenesis. Integration of microarray-generated differentially expressed cancer genes into cellular networks could help in analyzing and interpreting the biological significance of the genes in a network and their functional interrelationships. The objective is of this module is to reconstruct or reverse engineer [62, 63] the interactions between genes to elucidate the underlying intra-cluster regulatory subnetwork of the gene clusters obtained from previous module by reasoning backwards from expression level observations. Our approach is based on probabilistic generative modeling of experimental observations.

In this work we propose a modified version of the celebrated Sparse Candidate Algorithm [64] to learn a Bayesian Network from gene expression levels at a genome-wide scale of Cancer and Non - Cancer patients. Variation of the Sparse Candidate Algorithm used here is an iterative algorithm which, on every iteration removes cycles from the obtained graph to obtain Directed Acyclic Graph (DAG) in the form of Bayesian Network and also implements a scoring function via the Greedy Hill Climbing Algorithm [65] to acquire the best structure. Here, we give an alternative and efficient way of cycle removal from the cyclic graph in context to the data and also suggest a modification of the Greedy Hill Climbing Algorithm used. We prove that our modification theoretically guarantees that the output is at least as good as the original algorithm and has the same worst case complexity. Besides, evaluation of our modified algorithm on real world data shows that our algorithm can be very much useful to draw important conclusions from data. We also report some such results to support our claim.

6.1 Related Works

A number of models have been proposed for revealing the structure of the transcriptional regulation process. The simplest is the Boolean networks model [66]. In this model, each gene is modeled as a Boolean entity, which can be in one of two states: on or off. The dynamics are modeled over a discrete series of time points. The state of each gene is determined by a Boolean function of some of the other genes at the previous time step. Different algorithms have been proposed for inferring the network structure of such models from observations [67], typically by employing information-theoretic considerations. However, these ignore the effect of genes at

intermediate levels, and result in information loss during discretization. Other deterministic approaches model the expression of a gene as a linear [68] or sigmoid function of its regulators, either directly or as a solution to a set of differential equations [69]. Genetic regulatory networks modeled using ODEs relate the rate of change in gene transcript concentration (i.e., the concentrations of RNAs, proteins, and other molecules) with respect to one another as well as to external perturbation The other commonly used type of continuous variable model is the neural network based model [70] of which most successful is the recurrent neural network (RNN). This model is biologically plausible and noise-resistant and continuous in time with non-linear characteristics. Modeling GRN using a fully recurrent neural network, it assumes that each node represents a particular gene, and the wiring between the nodes defines the regulatory interactions. The level of expression of a gene at any time can be assessed by the other genes, and the output of a node at the next time step is derived from the expression levels and connection weights of all the genes connected to it. In these approaches, every gene is apriori assumed to depend on all other genes, and the connection strengths are learned through optimization, thus substituting structure learning with parameter learning. Work was also done on probabilistic models of Bayesian networks [71] to find gene regulatory networks which gave a new dimension to this field. Bayesian networks are graph models that estimate complicated multivariate joint probability distributions through local probabilities. Here, the genetic regulatory network is described as a directed acyclic graph, with vertices corresponding to genes and the edges representing the conditionally dependent interactions between genes. Basso et al. [13] developed a statistical algorithm using mutual information to model pairwise gene-gene interaction network using ARACNE algorithm on Gene expression profiles of human B-cells at different stages covering normal to cancer cells. Therefore development of network based methods for identification of diseased genes [14] remains an active area of research in systems biology.

6.2 Motivation Towards Model Selection

Bayesian networks are interpretable and flexible models for representing probabilistic relationships between multiple interacting agents. This is because firstly, at a qualitative level, the structure of a Bayesian network describes the relationships between these agents in the form of conditional independence relations. Whereas at a quantitative level, relationship between the

interacting agents are described by conditional probability distribution (CPD). Secondly, probabilistic nature of this approach is capable of handling noise inherent in both biological processes and microarray experiments. This makes Bayesian networks superior to Boolean networks which are deterministic and synchronous in nature. Thirdly, they are particularly useful for describing processes composed of locally interacting components; where the value of each component directly depends on the values of a relatively small number of components. Modeling such sparse interactions is easier compared to solving complex equations having several parameters values involved with models of continuous variables like ODEs and Neural Networks. Finally, statistical foundations for learning Bayesian networks from observations, and computational algorithms to do so are well understood and provide mechanisms of causal influence in the network.

In this work we mainly started with the very basic concepts and algorithms used for probabilistic graphical modeling of GRNs. We exploited the basic idea of widely used sparse candidate algorithm and proposed modification of steps in its ground level of execution. Our approach achieves a midway position between selecting the best scoring local optimal graph through heuristic methods and searching the global optimal graph from the whole set of exponential possible graphs and get the best scoring one.

6.3 Bayesian Networks

Bayesian networks are a language for representing joint probability distributions of many random variables. They are particularly effective in domains where the interactions between variables are fairly local: each variable directly depends on a small set of other variables. Bayesian networks have been applied extensively for modeling complex domains in different fields. A brief overview of the formalism of Bayesian networks and the algorithms for learning such models from observed data is presented as follows.

6.3.1 Model Definition

The notations used in this work are described at first. Upper case letters such as X, Y, Z represent random variables and lowercase letters such as, x, y, z represent values taken up by

random variables. Sets of random variables are denoted by, boldface capital letters, $\mathbf{X, Y, Z}$ and assignment of values to the variables in the sets is denoted using bold face lower case letters, $\mathbf{x, y, z}$. The finite set $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$ is a set of random variables where, each variable X_i can be *discrete* and can take any value, x_i from the domain $Value(X_i)$ or it can be continuous, where, it can take value from some real interval. In this work we model the Bayesian network using only discrete random variables

A Bayesian network is a structured graph representation of relationships between variables where nodes (i.e. random variables) represent objects of the problem domain and the edges often represent direct influence of one variable on another. More specifically, the graph represents conditional independencies between these variables.

Definition: $\mathbf{X}$ is ***conditionally independent*** of $\mathbf{Y}$ given $\mathbf{Z}$ if, $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})$ which is symbolically denoted by, $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$.

Definition: Let $\mathbf{G}$ be a DAG whose vertices correspond to the random variables $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$. Let $\mathbf{U}_{X_i}$ denote the parents of X_i in $\mathbf{G}$. We say that $\mathbf{G}$ encodes the local **Markov assumptions** over X : Each variable X_i is conditionally independent of its non-descendants, given its parents in $\mathbf{G}$ i.e. $\forall X_i \left(X_i \perp Non-Descendants_{X_i} \mid \mathbf{U}_{X_i}\right)$. This set of assumptions is denoted by, $Markov(\mathbf{G})$.

Definition: A ***Bayesian network*** is a representation of a joint probability distribution consisting of two components. The first component is $\mathbf{G}$, which is a DAG whose vertices correspond to the random variables $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$ and whose structure encodes the Markov assumptions $Markov(\mathbf{G})$ over $\mathbf{X}$. The second component is φ which describes a conditional probability distribution (CPD) of $P\left(X_i \mid \mathbf{U}_{X_i}\right)$ for each variable X_i in $\mathbf{X}$. Therefore Bayesian Network is given by the pair, $B = <\mathbf{G}, \varphi>$. Bayesian network components $\mathbf{G}$ and φ specifies a unique distribution of $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$.

The joint distribution that satisfies the conditional independence properties can be decomposed into the product form using the *chain rule for Bayesian networks* by,

$$P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i \mid \mathbf{U}_{X_i})$$

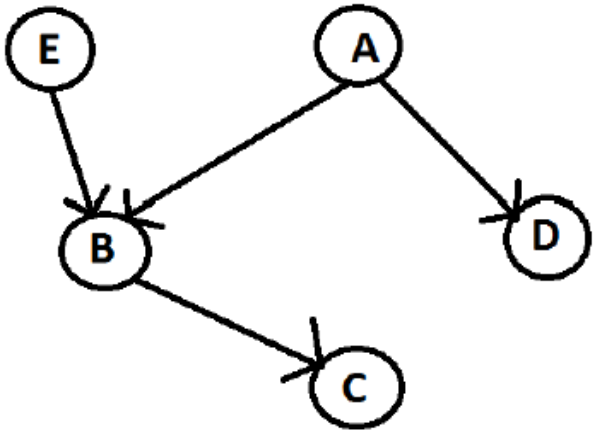

Figure 6.1: An example of a simple Bayesian network structure.

Conditional independence statements encoded are

$(A \perp E), (B \perp D \mid A, D), (C \perp A, D, E \mid B), (D \perp B, C, E \mid A)$ *and* $(E \perp A, D)$

By the chain rule of probability, joint probability distribution specified by the Bayesian network of Figure 6.1 without any independence assumptions is given by, $P(A,B,C,D,E) = P(A)P(E \mid A)P(B \mid A,E)P(C \mid A,E,B)P(D \mid A,\mathrm{E},\mathrm{B},\mathrm{C})$. While taking the conditional independencies into account joint distribution can be specified by, $P(A,B,C,D,E) = P(A)P(E)P(\mathrm{B} \mid A,E)P(C \mid B)P(D \mid A)$.

6.3.2 Parameter Representation : multinomial CPDs

The parameterization component of Bayesian networks φ defines the conditional probability distributions (CPDs) $P(X_i \mid \mathbf{U}_{X_i})$ and can be of any general form. In this work we use multinomial CPD representation as we have considered the underlying random variables to be discrete in nature. When both the variable X_i and its parents $\mathbf{U}_{X_i}$ are discrete, the most general representation for a CPD is a conditional probability table (CPT). Each row in these tables corresponds to a specific joint assignment $\mathbf{u}_{X_i}$ to $\mathbf{U}_{X_i}$ and specifies the probability distribution for X_i conditioned on $\mathbf{u}_{X_i}$. If $\mathbf{U}_{X_i}$ consists of k binary valued variables, the table will specify 2^k

distributions. For example Table 6.2 shows the CPD of random variable B for Bayesian network of Figure 6.1.,

Table 6.1 CPD of random variable B for Bayesian network of Figure 6.1

A	*E*	*P(B=0)*	*P(B=1)*
0	0	1.00	0.00
0	1	0.50	0.50
1	0	0.40	0.60
1	1	0.70	0.30

6.4 Learning Bayesian Network

The problem of learning Bayesian network can be formulated as follows, given a training set of samples, $\boldsymbol{D} = \mathrm{x}[1], \mathrm{x}[2], \cdots, \mathrm{x}[M]$ independently drawn from some unknown generating Bayesian Network $\mathbf{G}^*$ with an underlying distribution $\mathbf{P}^*$, the goal is to recover $\mathbf{G}^*$.

6.4.1 Parameter Learning

This task is stated as, assuming the correct network structure $\mathbf{G}$ is given; we need to estimate the best parameters. More formally, assuming we are given a network structure $\mathbf{G}$, and a set of data instances $\boldsymbol{D}$ for the variables represented in $\mathbf{G}$. The objective is to determine what values for the network parameters $\boldsymbol{\varphi}$ best describe the process that generated the data.

Maximum Likelihood Estimation

Without using any prior assumptions on the parameters, an intuitive and widely used measure is the probability that a model equipped with $\boldsymbol{\varphi}$ that assigns to $\boldsymbol{D}$. The likelihood function of $\boldsymbol{\varphi}$ given $\boldsymbol{D}$ which we denote here by, $L(\boldsymbol{\varphi} : \boldsymbol{D})$ is given by,

$$L(\boldsymbol{\varphi} : \boldsymbol{D}) = \prod_{m=1}^{M} P\left(\mathrm{x}[m] \mid \boldsymbol{\varphi}\right)$$

In Maximum likelihood estimation we wish to choose parameters $\bar{\boldsymbol{\varphi}}$ that maximizes the likelihood of the data.

$$\bar{\varphi} = \max_{\varphi} L(\varphi : \boldsymbol{D})$$

One of the big advantages of the Bayesian Network representation is that this likelihood decomposes into local likelihood functions. In a Bayesian Network, a variable X with its parents $\mathbf{U}$ has a parameter $\varphi_{x|\mathrm{u}}$ for each combination of $x \in Value(X)$ and $\mathrm{u} \in Value(\mathbf{U})$. The idea of decomposition is to group together all the instances in which $X = x$ and $\mathbf{U} = \mathrm{u}$ which are represented as, $\mathrm{M}[x, \mathrm{u}]$ the number of these instances, and $\mathrm{M}[\mathrm{u}] = \sum_{x \in X} \mathrm{M}[x, \mathrm{u}]$. Then,

$$\begin{aligned} L(\varphi : \boldsymbol{D}) &= \prod_{m=1}^{M} P\left(\mathrm{x}[m]\right) \\ &= \prod_{m=1}^{M} \prod_{i=1}^{N} P\left(x_i[m] \mid \mathbf{U}_{X_i}[m] : \varphi\right) \\ &= \prod_{i=1}^{N} L_i\left(\varphi_{X_i|\mathbf{U}_{X_i}} : \boldsymbol{D}\right) \end{aligned}$$

Where $L_i\left(\varphi_{X_i|\mathbf{U}_{X_i}} : \boldsymbol{D}\right) = \prod_{m=1}^{M} P\left(x_i[m] \mid \mathbf{U}_{X_i}[m] : \varphi\right)$ is the local likelihood function for X_i.

By optimizing the local likelihood functions under normalization constraints, we obtain the maximal likelihood estimators (MLE) [65] for the parameters of the multinomial table CPD by,

$$\bar{\varphi} = \frac{\mathrm{M}[x, \mathrm{u}]}{\mathrm{M}[\mathrm{u}]}$$

Where, $\mathrm{M}[x, \mathrm{u}]$ and $\mathrm{M}[\mathrm{u}]$ are called sufficient statistics as they summarize all the relevant information from the data that is needed in order to calculate the likelihood.

6.4.2 Structure Learning

The previous section deals with the method of learning the Bayesian network parameters given a known structure $\mathbf{G}$ but methods for working with real life data $\mathbf{G}$ is generally not well established. The goal of structure learning algorithms is to reconstruct from the observed data. The construction procedure takes a score-based approach. For this approach at first, a space of candidate models which we are willing to consider is defined. Secondly we define a scoring function that measures how well each model fits the observed data. Then an optimization algorithm is employed that searches for the highest scoring model.

Bayesian Score

The main principle of the Bayesian scoring approach is that whenever there is uncertainty over anything, we should place a distribution over it. Since there is uncertainty both over structure and over parameters therefore we define a structure prior $P(\mathbf{G})$ that puts a prior probability on different graph structures and a parameter prior $P(\boldsymbol{\varphi} \mid \mathbf{G})$ that puts a probability on different choice of parameters once the graph is given. By Bayes rule we have,

$$P(\mathbf{G} \mid \boldsymbol{D}) = \frac{P(\boldsymbol{D} \mid \mathbf{G})P(\mathbf{G})}{P(\boldsymbol{D})}$$

Here the denominator is simply a normalizing factor that does not help distinguish between different structures. Bayesian score is defined as,

$$score_B(\mathbf{G} : \boldsymbol{D}) = \log P(\boldsymbol{D} \mid \mathbf{G}) + \log P(\mathbf{G})$$

Where, $P(\boldsymbol{D} \mid \mathbf{G})$ takes into consideration the uncertainty over the parameters and averages the probability of the data over all possible parameter assignments to $\mathbf{G}$.

$$P(\boldsymbol{D} \mid \mathbf{G}) = \int P(\boldsymbol{D} \mid \mathbf{G}, \boldsymbol{\varphi}) P(\boldsymbol{\varphi} \mid \mathbf{G})\, d\boldsymbol{\varphi}$$

The score is ***decomposable***, and can be rewritten as the sum where the contribution of every variable X_i to the total network score depends only on its own value and the values of its parents in $\mathbf{G}$.

$$score_B(\mathbf{G} : \boldsymbol{D}) = \sum_{i=1}^{N} Score\_Contribution\left(X_i, \mathrm{U}_{X_i}^{\mathbf{G}} : \boldsymbol{D}\right)$$

The Bayesian score is well suited for situations with small sample strength. The Bayesian score is biased to more simple structures, but as it gets more data, it would support more complex structures. This bias is due to the integration over all possible parameters. Structures with many parameters are penalized, unless the probability of the true parameters is very peaked which happens when the sample size is large. Thus the Bayesian score inherently takes care of the problem of over-fitting a small sample to a complex model.

6.4.3 Search Algorithm

The Bayesian network search can be formulated as an optimization problem. The input is,

- the training set ***D***
- scoring function (including priors if needed)
- a set of *G* of possible network structures (incorporating any prior knowledge)

The desired output is a network structure (from the set of possible structures) that maximizes the score. Score - based methods view a Bayesian network as specifying a statistical model and then address learning as a model selection problem. We define a hypothesis space of potential models -the set of possible network structures we are willing to consider and a scoring function that measures how well the model fits the observed data. Our computational task is then to find the highest scoring network structure. The space of Bayesian networks is a combinatorial space, consisting of a super exponential number of structures $2^{O(n^2)}$. The search problem is NP - hard, and we resort to heuristic search techniques. Therefore, even with a scoring function, we can find the local optimum that is the local highest-scoring network.

The objective of this module is to construct intra-cluster GRN using Bayesian network approach as genes of same cluster are co-expressed and often share regulatory relationships between them. In our Bayesian network we model the nodes as discrete random variables and we modify the Sparse Candidate Algorithm to learn the best scoring network.

6.4.3.1 Initial Network Formation

In multinomial model of Bayesian network, each variable is treated as a discrete one and then learn a multinomial distribution that describes the probability of each possible state of the child variable given the state of its parents. To apply the multinomial model discretization of the gene expression values is required. Discretization is done into three categories: under-expressed(-1), normal(0), and over-expressed(1) depending on whether the expression rate is significantly lower than, similar to, or greater than control expression level, respectively. The control expression level of a gene can be either determined experimentally or it can be set as the average expression level of all the genes across the experiment. In this work we consider the discretization measure as follows, for each patient or sample, the control gene expression level is set to the mean expression level of that sample across its genome wide expression value. This can be justified by

the fact that genome wide expression profile for a patient consists of gene expression values corresponding to every gene in the genome. This includes house – keeping genes as well as crucial genes relating to cancer. Therefore mean gene expression level would generally signify the normal gene expression levels for that patient or sample. Mean and standard deviations are calculated for each sample of the dataset under consideration. Then discretization of a gene expression data is performed by computing the Z-score of that expression level with respect to the mean and standard deviation for that sample. A Z-score threshold of 0.5 is used as follows, Z-score of gene expression level greater than 0.5 is considered over-expressed (1), Z-score below -0.5 is considered under-expressed (-1) and Z-score in between is considered to be normal-expressed (0).

The idea after discretization is to use a measure of dependence, such as the mutual information, between variables to guide network construction. Here we use the idea mentioned in [64] to restrict the possible parents of each variable. Thus, instead of having *n - 1* potential parent for a variable, we only consider *k* possible parents, where $k << n$. Mutual information between each pair of nodes is used select k candidate parents, where k is any fixed natural number deciding the maximum number of parents would we prefer each node to have. For each node X_i we separately compute the mutual information value with other nodes $\mathrm{MI}(X_i, X_j)$ where $i \neq j$. This is given by,

$$\mathrm{MI}(X_i, X_j) = \sum_{x,y} \hat{P}(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$$

Where $\hat{P}$ denotes the observed frequencies in the dataset. The mutual information is always non-negative. It is equal to 0 when *X* and *Y* are independent. Higher mutual information implies stronger the dependence between *X* and *Y*.

For node X_i after computing the mutual information value with other nodes $\mathrm{MI}(X_i, X_j)$ and arrange them in descending order and select k nodes giving the highest MI values with respect to node X_i. These k nodes are the *candidate parents* for node X_i . Repeating this process for each node we generate a directed graph which is expected to have many cycles. We remove cycles from the network as our aim is to obtain a Bayesian Network and hence the graph

structure should be a DAG. On the graph obtained from mutual information the following cycle removal strategy is employed and the initial network B_0 is obtained.

Cycle Removal Strategy

The directed graph is tested weather it is acyclic or not using depth first search (DFS). For cyclic graphs, the following cycle removal algorithm used to make the graph a DAG.

Cycle_Removal(B_n)

Repeat until B_n is not a DAG

Detect cycle in graph using DFS : $C = X_{i1} \to X_{i2} \to \cdots \to X_{im} \to X_{i1}$

List nodes involved in cycle $L = \{X_{i1}, X_{i2}, \cdots, X_{im}\}$

$\forall X_{ij} \in L$ find $parent(X_{ij})$ in C

Find rank of $parent(X_{ij})$ as a candidate parent of X_{ij}

Construct $R = \{r_{i1}, r_{i2}, \cdots, r_{im} : r_{ij} = rank(parent(X_{ij}))$ in $C\}$

Find $r_{iq} = \max\{R\}$ and $X_t = parent(X_{iq})$ in C

Remove edge $X_t \to X_{iq}$ and update $B_n = B_n - edge(X_t \to X_{iq})$

Return B_n

6.4.3.2 Modified Sparse Candidate Algorithm

The main idea of Sparse Candidate algorithm is to identify a relatively small number of candidate parents for each gene based on simple local statistics. This algorithm considers only acyclic graphs as a legal solution, but we suggest to not only to consider acyclic graphs but to remove cycle using ***Cycle_Removal*** procedure for a more extensive state space search. The modified algorithm to learn Bayesian network is described in this section.

Input:

- The training Dataset $\boldsymbol{D} = \{\mathrm{x}[1], \mathrm{x}[2], \cdots, \mathrm{x}[M]\}$
- An initial network $\boldsymbol{B_0}$
- A decomposable score such that $score(\boldsymbol{B} \mid \boldsymbol{D}) = \sum_{i=1}^{N} Score\_Contribution\left(X_i, \mathrm{U}_{X_i}^{\boldsymbol{B}} : \boldsymbol{D}\right)$
- A parameter $\boldsymbol{k}$ = maximum parent for each node
- A parameter $\boldsymbol{l}$ = minimum number of edges in network

Output: A network **B**

Loop for $n = 1, 2, \cdots$ until convergence

Restrict:

Based on $\boldsymbol{D}$ and $\boldsymbol{B_{n-1}}$ for each variable select a set C_i^n of candidate parents with $|C_i^n| \leq \boldsymbol{k}$ by ***Candidate_Parent_Selection*** $(X_i, \boldsymbol{B_{n-1}}, \boldsymbol{D}, \boldsymbol{k})$

This defines a directed graph $\boldsymbol{H_n} = (\mathbf{X}, E)$ where $E = \{X_j \rightarrow X_i \mid \forall i, j\ X_j \in C_i^{\boldsymbol{n}}\}$

Remove cycles in $\boldsymbol{H_n}$ using ***Cycle_Removal*** $(\boldsymbol{H_n})$

Maximize:

Find network $\boldsymbol{B_n} = <\mathbf{G_n}, \boldsymbol{\varphi}>$ maximizing $score(\boldsymbol{B_n} \mid \boldsymbol{D})$ among networks satisfying $\mathbf{G_n} \subset \boldsymbol{H_n}$ (i.e. $\forall X_i,\ \mathrm{U}_{X_i}^{\mathbf{G_n}} \subseteq C_i^{\boldsymbol{n}}$) using ***Greedy_Hill_Climbing***. Restrict the minimum number of edges by $\boldsymbol{l}$ to avoid sparse graph

Return $\boldsymbol{B_n}$

The algorithm for candidate parent selection for node X_i of $\boldsymbol{B_n}$ based on dataset $\boldsymbol{D}$ is as follows.

Candidate_Parent_Selection $(X_i, \boldsymbol{B_n}, \boldsymbol{D}, \boldsymbol{k})$

Calculate $\mathrm{M}(X_i, X_j)\ \forall i \neq j$ and $X_j \notin \mathbf{U}_{X_i}$ in $\boldsymbol{B_n}$

Choose $x_1, x_2, \cdots, x_{\boldsymbol{k}-p}$ with highest ranking, where $p = |\mathbf{U}_{X_i}|$

Set $C_i = \mathbf{U}_{X_i} \bigcup x_1, x_2, \cdots, x_{\boldsymbol{k}-p}$

Return C_i

Here $M(X_i, X_j)$ is Mutual Information measure $MI(X_i, X_j)$ for $\boldsymbol{B_0}$ and discrepancy values from Kullback - Leibler Divergence [64] for $\boldsymbol{B_n}, n \neq 0$.

The greedy hill climbing algorithm to search for highest scoring Bayesian network structure is described as follows.

Input:

- Initial candidate solution S_0
- Score function *score*
- Set of search operators O

Here we use $O = \{add\_edge,\ delete\_edge,\ reverse\_edge\}$

Output: best candidate solution S_{Best}

Greedy_Hill_Climbing *(S_0, score, O)*

$S_{Best} \leftarrow S_0$
Do
  $S_{Temp} \leftarrow S_{Best}$
  $Update \leftarrow false$
  for each operator $o \in O$
    $S_0 \leftarrow o(S_{Temp})$ // Result of applying o on S_{Temp}
    if S_0 is not acyclic then
      $S_0 \leftarrow Cycle\_Removal(S_0)$
    if $score(S_0) > score(S_{Best})$ then
      $S_{Best} \leftarrow S_0$
      $Update \leftarrow true$
While *Update*
Return S_{Best}

This concludes all of the algorithms involved in learning the Bayesian network structure that best fits the data.

6.5 Results

In chapter 5, we identified 3 major clusters corresponding to the dataset of gene expression values from healthy population. Performing biological enrichment of the clusters by analysis of common GO Terms shared between genes belonging to same cluster we denoted the three clusters as *regulatory cluster*, *response and signaling cluster* and *cell development and maintenance cluster*. The objective of this module is to find intra – cluster gene regulatory network using Bayesian network approach. Co-expressed genes are often regulated by each other or are regulated by some common parent, i.e. genes belonging to same cluster are highly likely to share regulatory relationships between them. The idea here is to capture those relations using probabilistic graphical models. Cluster# 1 of Dataset A containing 20 genes namely,

C6	NELL2	BICC1	CWH43	DNAJC6	PLA1A
FGF14	TRIM36	ODF2	HSPA1L	ZSCAN5A	SOX9
TRDMT1	CLGN	NR2F1	ECI2	TMEM45A	HTRA1
CST6	TMEM110				

majorily corresponds to regulatory biological processes. This is evident from the GO biological process terms (BP) terms shared between the genes this cluster listed in Table 6.2

Table 6.2 GO terms shared from biological process ontology by genes of Cluster# 1 of Dataset A

GOID	GO TERM from biological_process Ontology	p-value	% of Genes of Cluster Annotated	Annotated Genes
GO:0050789	regulation of biological process	0.024364	65	CST6, TRDMT1, SOX9, HTRA1, DNAJC6, BICC1, ODF2, HSPA1L, TRIM36
GO:0050794	regulation of cellular process	0.04834	60	CST6, SOX9, HTRA1, DNAJC6, BICC1, ODF2, HSPA1L, ZSCAN5A, TRIM36
GO:0051128	regulation of cellular component organization	0.000635	30	ODF2, SOX9, HSPA1L, HTRA1, DNAJC6, NR2F1
GO:0048583	regulation of response to stimulus	0.014725	30	SOX9, HSPA1L, HTRA1, C6, FGF14, BICC1
GO:0080134	regulation of response to stress	0.002823	20	HSPA1L, HTRA1, C6, FGF14
GO:0002682	regulation of immune system process	0.005408	20	SOX9, HTRA1, C6, FGF14
GO:0009719	response to endogenous stimulus	0.005762	20	TRDMT1, SOX9, HTRA1, NR2F1
GO:0050793	regulation of developmental process	0.016784	20	ODF2, SOX9, C6, NR2F1

GOID	GO TERM from biological_process Ontology	p-value	% of Genes of Cluster Annotated	Annotated Genes
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.019413	20	SOX9, ZSCAN5A, HTRA1, NR2F1
GO:0032879	regulation of localization	0.02415	20	SOX9, HSPA1L, FGF14, DNAJC6
GO:0031347	regulation of defense response	0.004414	15	HTRA1, C6, FGF14
GO:0009968	negative regulation of signal transduction	0.014359	15	SOX9, HTRA1, BICC1
GO:0050679	positive regulation of epithelial cell proliferation	0.001952	10	SOX9, HTRA1
GO:0050688	regulation of defense response to virus	0.002715	10	HTRA1, FGF14
GO:0050768	negative regulation of neurogenesis	0.003733	10	SOX9, NR2F1
GO:0030178	negative regulation of Wnt signaling pathway	0.004118	10	SOX9, BICC1

Therefore we select this cluster to determine the intra-cluster gene regulatory network. After discretization, the initial network formed using Mutual Information is given in Figure 6.2. This network in consists of 20 nodes one corresponding to each gene and 60 edges. This network is not acyclic, and contains 20 cycles with minimum cycle length of 2 a maximum cycle length of 5 each of which were iteratively removed by our ***Cycle_Removal*** algorithm to yield a DAG which is $\boldsymbol{B_0}$. This $\boldsymbol{B_0}$ is provided as input to the *Modified Sparse Candidate Algorithm.*

Figure 6.2 Initial network based on mutual information between genes

After convergence of the sparse candidate algorithm, the final Bayesian network of intra-cluster GRN returned is given in Figure 6.3 with a total 39 edges.

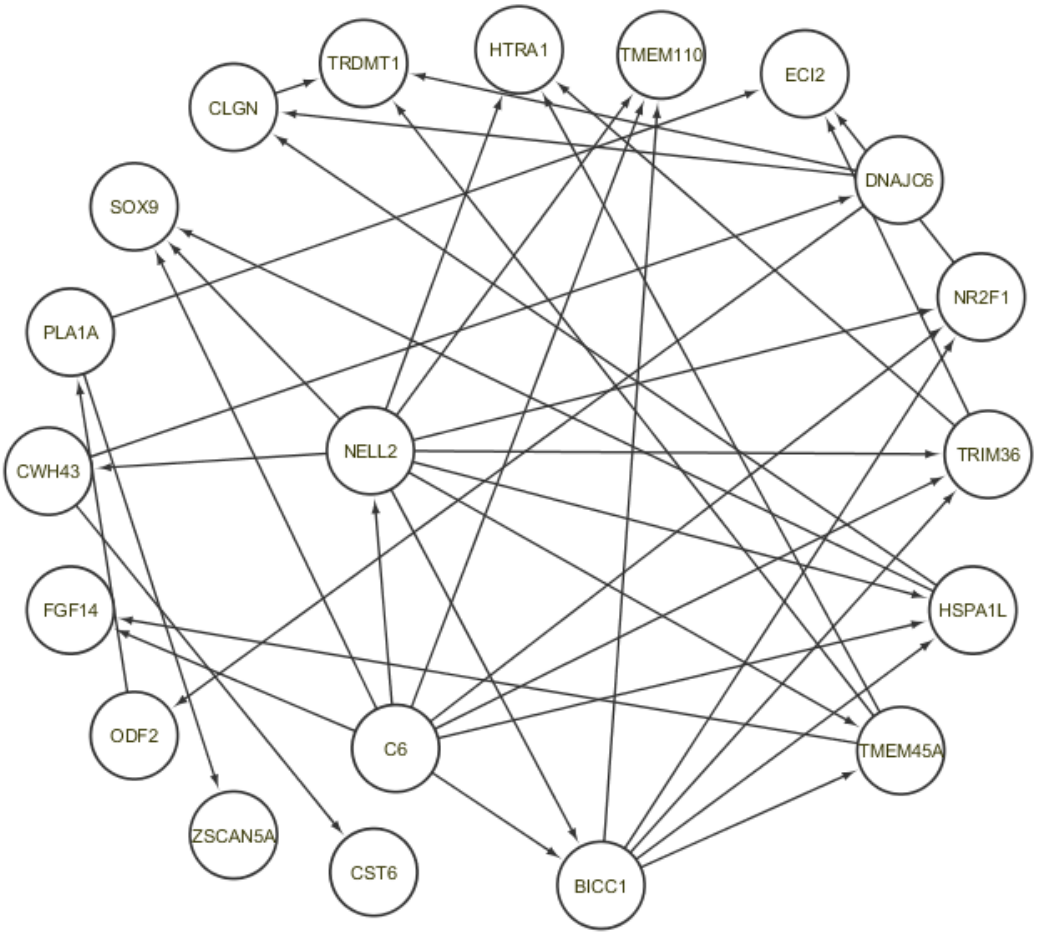

Figure 6.3 Final GRN for Cluster# 1 of Dataset A

6.6 Validation of predicted Network

Validation of the predicted network is done using GeneMANIA (http://www.genemania.org/) gene network prediction tool. GeneMANIA is a web tool that finds gene related to the set of queried genes using a very large set of functional association data. Association data include protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. The GeneMANIA web tool is provided with the queried gene list consisting of genes belonging to Cluster# 1 from Dataset A. The output is an undirected network shown in Figure 6.4. GeneMANIA return the network with many other genes relevant to the queried gene list. The network reported consists of 20 genes in addition to the queried gene list, totaling to 40 genes and 82 undirected edges with association between nodes being based on *Co-expression.*

Figure 6.4 Gene Interaction Network returned by GeneMANIA

Comparing the network returned by GeneMANIA and the network generated by our algorithm, similarities extracted are reported as follow. Direct comparison between the two networks was not possible as GeneMANIA reports an undirected network incorporating many other additional genes. While, the sparse candidate algorithm considered here returns a directed network with each edge implying regulatory relationship between parent to child node. In the network of Figure 6.3 three prime regulators are identified namely C6, NELL2 and BICC1 which also correspond to the highly connected node Figure 6.4. Corresponding to each edge in the network returned by GeneMANIA, information provided are edge weight, network group and biological literature corresponding to which edge is annotated. In the Bayesian network formed by our algorithm, the edge $C6 \rightarrow NELL2$ corresponds to the undirected edge C6 – NELL2 in GeneMANIA network and is curated in literature Ramaswamy-Golub-2001. The edge C6 –

HTRA1 in GeneMANIA network curated in Burington-Shaughnessy-2008, has correspondence with the path $C6 \rightarrow NELL2 \rightarrow HTRA1$ in the final Bayesian network. However in the initial network shown in Figure 6.2 there is a directed edge $C6 \rightarrow HTRA1$. The undirected edge BICC1 –SOX9 is reported in literature Burington-Shaughnessy-2008; Innocenti-Brown-2011, against which our algorithm detects C6 to be the common parent of SOX9 and BICC1 and in the initial network there was a directed edge $BICC1 \rightarrow SOX9$. $BICC1 \rightarrow NR2F1$ edge is returned by our algorithm which corresponds to the undirected path BICC1 –SOX9 –NR2F1 in GeneMANIA network with BICC1 –SOX9 reported in network generated by Burington-Shaughnessy-2008; Innocenti-Brown-2011 and the edge SOX9 –NR2F1 is reported in Kang-Willman-2010. Corresponding to the undirected edge CLGN –ODF2 reported in network by Mallon-McKay-2013 in our network the genes ODF2 and CLGN is regulated by common parent DNAJC6. The similarity found between the two networks is summarized in Table 6.5.

Table 6.5 Comparison of GeneMANIA interaction network and GRN formed using our approach

Serial #	**GeneMANIA Network Edge or Path**	**Network Study for the edge**	**Bayesian Network Edge Connection**	**Connectivity between nodes in Bayesian network**	**Dependence Evident From**
1	C6 – NELL2	Ramaswamy-Golub-2001	$C6 \rightarrow NELL2$	Parent-Child	Causal Reasoning
2	C6 – HTRA1	Burington-Shaughnessy-2008	$C6 \rightarrow NELL2 \rightarrow HTRA1$	Descendant	Causal flow of dependence
3	BICC1–SOX9	Burington-Shaughnessy-2008; Innocenti-Brown-2011	C6 ↙ ↘ SOX9 BICC1	Common Parent	Evidential reasoning from one child to common parent and then causal reasoning from common parent to other child
4	BICC1 –SOX9 –NR2F1	Kang-Willman-2010	$BICC1 \rightarrow NR2F1$	Parent-Child	Causal Reasoning

Serial #	GeneMANIA Network Edge or Path	Network Study for the edge	Bayesian Network Edge Connection	Connectivity between nodes in Bayesian network	Dependence Evident From
5	CLGN –ODF2	Mallon-McKay-2013	DNAJC6 ↙ ↘ ODF2 CLGN	Common Parent	Evidential reasoning followed by causal reasoning

6.7 Discussion

Genes belonging to the same cluster are highly likely to share regulatory relationships between them, in terms of one being the regulator of the other or being regulated by a common parent. With this motivation, our objective was to create GRN between genes belonging to the same cluster. We consider the regulatory cluster of the gene expression dataset corresponding to the healthy population and construct a 20 gene network directed network with edges encoding regulatory relationships. We take a Bayesian network approach to construct the GRN and propose a modification to the sparse candidate algorithm used for learning Bayesian networks. The modification lies in the development of a cycle removal strategy which removes the edge connecting weakest parent from all modes in a cycle. This cycle removal strategy enables us to evaluate the cyclic graphs generated during the greedy hill climbing based search for optimal Bayesian network structure. The resulting GRN was compared with the gene interaction network returned by the web tool GeneMANIA. Results show that several gene associations reported by GeneMANIA based on co-expression has also been encoded by GRN resulting from our algorithm.

Chapter 7

Conclusion and Future Scope

Advent of DNA microarray technology has led to complete genome expression profiling. In a cancer cell, the genome's normal functioning is distorted. Analysis of genome wide expression data can provide insights into the molecular mechanisms of carcinogenesis. In this work we target the problem of analysis of genome wide expression profile of two sets of population, one being the healthy population of smokers and the other being the group of smokers diagnosed with lung cancer. The work is divided into three main modules.

The first module aims at identification of genes exhibiting differential expression pattern between healthy and diseased population. This is achieved using statistical hypothesis testing procedure. Each gene is tested against the null hypothesis of exhibiting no differential expression change. Genes rejecting the null hypothesis with high degree of confidence which is assessed by means of q-values is identified to be significant and differentially expressed. This module concludes with a list of 168 genes showing significant expression difference between the two populations.

The second module is aimed at identification of groups of co-expressed genes between the significant genes. Hierarchical agglomerative clustering is performed based on genes expression profiles of both healthy and diseased population dataset. Dissecting the linkages of the hierarchical tree generated by the clustering algorithm, the natural clusters in the dataset is identified. This involves evaluation of the quality of clusters generated after each linkage

criterion dissection by means of internal cluster validity indices. A consensus strategy between the validity indices is followed to identify the optimal number of clusters separately for the healthy and diseased dataset. This module concludes by identification of three significant clusters corresponding to the healthy population and four clusters corresponding to the diseased population. The three major gene clusters of healthy population are biologically enriched using Gene Ontology terms shared between genes of same cluster. In this module we also identify specific genes that showed significant expression change in case of diseased population by analyzing difference in the cluster assignment of the genes between the two population dataset.

The third part of the work is concerned with the development of intra-cluster gene regulatory network as genes belonging to the same cluster are highly likely to share regulatory relationships between them or be regulated by a common parent. To infer the regulatory relationship between the genes of a cluster we use probabilistic graphical model of Bayesian network to encode the dependencies and conditional independencies between genes. In this module we propose a modification to the sparse candidate learning algorithm for Bayesian networks by adding a cycle removal strategy of deleting the weakest parent in a cycle which enables us to perform a more extensive state space search for inferring optimal Bayesian network structure that fits the data. We test this algorithm on the regulatory cluster of the healthy population dataset and show that several co-expression associations between genes reported in biological literature has been correctly encoded by Bayesian network structure reported by our algorithm.

7.1 Scope of Future Work

Each module of the work discussed in this thesis has scope of improvements and modifications described below.

In the first module of significant gene identification, the test statistic used is Welch's t-test. In future prospects of this module, several other test statistics suggested in literature of hypothesis testing could be compared and contrasted. Analysis of which genes are filtered based on different test statistic and how relevant are they with respect to showing expression change between the two populations can be done as a part of future work.

In the clustering module, the approach used in this work yields non-overlapping clusters. But genes generally tend to participate in several biological process and functions. Incorporation

of other clustering algorithms like fuzzy clustering measures can be done to find better overlapping gene groups.

A valid criticism of the work addressed in the third module of the thesis is consideration of discrete random variable based Bayesian network for GRN construction. Discretization leads a lot of information loss. Modeling regulatory network using continuous variables Bayesian network would give improvement in the learned regulatory structure and can be target as a part of future work. Advance can also be done on development of more efficient Bayesian network algorithms. Analysis of how the network connectivity changes between the healthy and diseased population clusters would lead to significant insights into identification of genes playing crucial role in carcinogenesis.

In a nutshell, in this thesis genome wide expression profile of diseased and non-diseased population is analyzed, by first identifying significant genes, then performing cluster analysis of those genes separately on diseased and non-diseased expression dataset, and finally we construct intra-cluster gene regulatory network to identify regulatory relationships between the genes the same cluster.

Bibliography

[1] T.A Brown, *Genomes*,3rd Ed, Garland Science, 2006.

[2] Mike Williamson, *How Proteins Work*, 1st Ed, Garland Science,2011.

[3] M. J. Zvelebil and J. O. Baum, *Understanding bioinformatics*, Garland Science, 2008.

[4] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology*, Garland Science, 2009.

[5]M. Madan Babu, "Introduction to microarray data analysis" in *Computational Genomics* (Ed. Richard P. Grant), Horizon Bioscience, 2004.

[6] Francis Crick, "Central Dogma of Molecular Biology", Nature vol. 227, 1970.

[7] S Dudoit, Y. H Yang, M. J. Callow, and T P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", Statistica Sinica 12(2002), pp 111-139.

[8] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. "Cluster analysis and display of genome-wide expression patterns". Proceedings of the National Academy of Science, 95, pp 14863–14868, 1998.

[9] The Gene Ontology Consortium, "Gene Ontology Consortium: going forward", *Nucl Acids Res* 43 Database issue D1049–D1056. (2015).

[10] E Wang, A Lenferink, and M O'Connor-McCourt, "Cancer systems biology: exploring cancer-associated genes on cellular networks", Cellular and Molecular Life Sciences (CMLS), 64, pp 1752-1762, 2007.

[11] Wachi, S., Yoneda, K. and Wu, R. (2005), "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues", Bioinformatics 21, pp 4205 – 4208.

[12] Franke, L., Bakel, H. v., Fokkens, L., de Jong, E. D., EgmontPetersen, M. and Wijmenga, C. "Reconstruction of a functional human gene network with an application for prioritizing positional candidate genes" Am. J. Hum. Genet. 78, pp 1011 – 1025, 2006.

[13] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., laFavera, R. and Califano, A. , "Reverse engineering of regulatory networks in human B cells". Nature Genetics 37, pp 382 – 390, 2005.

[14] Xiujuan Wang, Natali Gulbahce and Haiyuan Yu, "Network-based methods for human disease gene prediction" Briefings In Functional Genomics. Vol 10. No 5, pp 280-293,

[15] Cancer Research UK. Worldwide cancer statistics, http://www.cancerresearchuk.org/cancer-info/cancerstats/world/

[16] ASH Fact Sheet on Smoking and Cancer, http://ash.org.uk/

[17] Parkin, DM. "Tobacco-attributable cancer burden in the UK in 2010". Br J Cancer 2011; 105, pp S6-S13.

[18] Cancer Research UK. Lung cancer risk factors, http://www.cancerresearchuk.org/cancer-info/cancerstats/types/lung/riskfactors/

[19] U.S. Department of Health and Human Services. "The Health Consequences of Smoking", A Report of the U.S. Surgeon General. 2004.

[20] Wistuba II, Lam S, Behrens C, Virmani AK, Fong KM, LeRiche J, Samet JM, Srivastava S, Minna JD, Gazdar AF. "Molecular damage in the bronchial epithelium of current and former smokers" J. Natl. Cancer Inst. 1997; 89, pp 1366–1373. [PubMed: 9308707]

[21] Auerbach O, Hammond EC, Garfinkel L. "Histologic changes in the larynx in relation to smoking habits". Cancer. 1970; 25, pp 92–104. [PubMed: 5410320]

[22] Spira A, Beane JE, Shah V, Steiling K et al. "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer". Nat Med 2007 Mar;13(3). [PMID: 17334370]

[23] Gustafson AM, Soldi R, Anderlind C, Scholand MB et al. "Airway PI3K pathway activation is an early and reversible event in lung cancer development", Sci Transl Med 2010 Apr 7;2(26). [PMID: 20375364]

[24] Edgar R, Domrachev M, Lash AE. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". Nucleic Acids Res. 2002 Jan 1;30(1), pp 207-210.

[25] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4115

[26] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2771

[27] Simon A. Forbes , David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott and Peter J. Campbell, "COSMIC: exploring the world's knowledge of somatic mutations in human cancer", Nucleic Acids Research, 2015, Vol. 43, Database issue D805–D811.

[28] Welch, B. L. "The generalization of "Student's" problem when several different population variances are involved", Biometrika ,34 (1–2), pp 28–35.

[29] Benjamini, Y. & Hochberg, Y., "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", Journal of the Royal Statistical Society, Vol. 57, No. 1 (1995), pp. 289-300.

[30] Storey, J.D. (2002), "A direct approach to false discovery rates", Journal of the Royal Statistical Society 64(3), pp 479–498

[31] Storey, J.D., and Tibshirani, R.(2003), "Statistical significance for genomewide studies", PNAS 100(16), pp 9440–9445.

[32] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization." Mol Biol Cell. 1998;9, pp 3273–3297.

[33] Eisen MB, Spellman PT, Brown PO, Botstein D. "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci. 1998;95, pp 14863–14868.

[34] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array." Proc. Natl. Acad. Sci. USA, Vol. 96(12):6745–6750, June 1999.

[35] Rose, K. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems." Proc. IEEE, 96, pp 2210–2239, 1998.

[36] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." Proc Natl Acad Sci. 1999;96:2907–2912.

[37] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. "Systematic determination of genetic network architecture". Nat Genet. 1999;22, pp 281–285.

[38] Shamir R. and Sharan R. "Click: A clustering algorithm for gene expression analysis". In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00). AAAI Press., 2000.

[39] Ben-Dor A., Shamir R. and Yakhini Z. "Clustering gene expression patterns". Journal of Computational Biology, 6(3/4), pp 281–297, 1999.

[40] Daxin Jiang, Chun Tang and Aidong Zhang, "Cluster Analysis for Gene Expression Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, Volume 16 Issue 11, November 2004, pp 1370-1386.

[41] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data", In: Bioinformatics, No.4, vol.19, pp 459–466, 2003.

[42] J. Handl, J. Knowles and D. B. Kell "Computational cluster validation in post-genomic data analysis", In: Bioinformatics, No. 15, vol. 21, pp 3201–3212, 2005.

[43] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes", In: BMC Bioinformatics, 7: 397, 2006.

[44] V. Pihur, S. Datta and S. Datta, "Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach", In: Bioinformatics, No. 13, vol. 23, pp 1607-1615, 2007.

[45] A. Ghosh, B. C. Dhara and R. K. De, "Comparative Analysis of Cluster Validity Indices in Identifying Some Possible Genes Mediating Certain Cancers", In: Molecular Informatics 32, pp 347-354, 2013.

[46] Snedecor GW, Cochran WG. "Statistical methods". Ames, Iowa: The Iowa State University Press; 1980.

[47] Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, Hatfield GW., "Global gene expression profiling in *Escherichia coli* K12: The effects of integration host factor", J Biol Chem. 2000;275, pp 29672–29684.

[48] Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, et al., "Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray" ,Proc Natl Acad Sci. 2000;97, pp 9127–9132.

[49] Jeffrey G. Thomas, James M. Olson, Stephen J. Tapscott, and Lue Ping Zhao , "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles", Genome Res. 2001 Jul; 11(7), pp 1227–1236. [50] A.K. Jian and R C. Dubes , *Algorithms for clustering data*, Prentice-Hall, Inc, 1988.

[51] Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).

[52] E. Dimitriadou, S. Dolňicar, A. Weingessel, "An examination of indexes for determining the number of clusters in binary data sets" ,Psychometrika, 67 (2002), pp. 137–159.

[53] U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices",IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002), pp 1650–1654.

[54]G.W. Milligan, M.C. Cooper, "An examination of procedures for determining the number of clusters in a data set", Psychometrika, 50 (1985), pp 159–179.

[55]J.C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", Journal of Cybernetics 3 (1973)pp 32–57.

[56]D.L. Davies, D.W. Bouldin, "A clustering separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence" 1 (1979)pp 224–227.

[57]T. Calinski, J. Harabasz, "A dendrite method for cluster analysis", Communications in Statistics 3 (1974) pp 1–27.

[58] P. Rousseeuw, Silhouettes: "A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics 20 (1987), pp 53–65.

[59]C.-H. Chou, M.-C. Su, E. Lai, "A new cluster validity measure and its application to image compression" , Pattern Analysis and Applications 7 (2004), pp 205–220.

[60] S. Bandyopadhyay, S. Saha, "A point symmetry-based clustering technique for automatic evolution of clusters", IEEE Transactions on Knowledge and Data Engineering 20 (2008), pp 1441–1457.

[61] K.R. Zalik, B. Zalik, "Validity index for clusters of different sizes and densities", Pattern Recognition Letters 32 (2011), pp 221–234.

[62] Ranajit Das, Sushmita Mitra, and C. A.Murthy, "Extracting Gene-Gene Interactions through Curve Fitting", IEEE Transactions On Nanobioscience, VOL. 11, NO. 4, December 2012. pp 402-409.

[63] Alexander J Hartemink, "Reverse engineering gene regulatory networks", Nature Biotechnology Vol.23, Number. 5, pp 554 - 555 (2005).

[64] N. Friedman, I. Nachman, and D. Pe'er. "Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm". In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI), pp 196–205. 1999.

[65] Daphne Koller and Nir Friedman, *Probabilistic Graphical Models : Principles and Techniques*. MIT Press, 2009.

[66] S. A. Kauffmann, "*The Origins of Order"*. Oxford University Press, New York, 1993.

[67] S. Liang, S. Fuhrman, and R. Somogyi. "Reveal, a general reverse engineering algorithm for inference of genetic network architectures". Pac Symp Biocomput, pp 18-29, 1998.

[68] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear modeling of mRNA expression levels during CNS development and injury". Pac Symp Biocomput, pp 41-52, 1999.

[69] Kimura S, Nakayama S, Hatakeyama M, "Genetic network inference as a series of discrimination tasks", Bioinformatics 2009, 25(7), pp 918-925.

[70] Vohradsky J., "Neural network model of gene expression." FASEB J 2001; 15, pp 846–54.

[71] Nir Friedman, Michal Linial, Iftach Nachman and Dana Pe'er, "Using Bayesian Networks to Analyze Expression Data", Journal of Computational Biology. August 2000, 7(3-4), pp 601-620.

[72] R. Tibshirani, D.M. Witten, "A comparison of fold-change and the t-statistic for microarray data analysis", Analysis (2007), pp 1–17.

[73] D.M. Mutch, A. Berger, R. Mansourian, A. Rytz, M.A. Roberts, "The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data", BMC Bioinformatics 3 (2002) 17.