Genes Expression Analysis for TB Diagnosis

By

Sofia Aftab

2009-NUST-MS-Phd-IT-38

Supervisor

Dr. Hammad Ahmed Qureshi

NUST (SEECS)

This thesis is submitted in partial fulfillment of the requirements for the Masters degree in Information Technology (MS IT)

In

School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.
(December 2012)

APPROVAL

This is certified that all the content containing this thesis entitled **"Genes Expression Analysis for TB Diagnosis"** submitted by **Sofia Aftab** has been proved adequate and appropriate for the requirement of the Masters degree.

Advisor: Dr. Hammad Ahmed Qureshi

Signature: _____

Date: _____

Committee Member 1: Dr. Sharifullah Khan

Signature_____

Date:_____

Committee Member 2: Dr. Zahid Anwar

Signature _____

Date: _____

Committee Member 3: Dr.Kashif Rajpoot____

Signature _____

Date: _____

IN THE NAME OF ALMIGHTY ALLAH

THE MOST BENEFICENT AND THE MOST MERCIFUL

TO MY MOTHER.

CERTIFICATE OF ORIGINALITY

I hereby declare that thesis and all its content is my own work and to the best of my knowledge it has no materials previously published or written by another person, nor contain material whose substantial extent has been accepted for the award of any degree or diploma at SEECS or at any other educational institute. I acknowledge the entire person with whom I have worked in SEECS. I also declare that all the research findings is the product of my own research work.

Author Name: Sofia Aftab

Signature: _____

ACKNOWLEDGEMENTS

Firstly I am thankful to Almighty Allah for giving me courage and strength to complete this challenging task and to work with the international research community. I am also grateful to my family, who supported me throughout my research work.

I am thankful to Dr. Hammad Ahmed Qureshi for his valuable suggestions and continuous guidance throughout my research work. His foresightedness and critical analysis of things taught me a lot about conducting valuable research which will be of great help to me in my career. I am also thankful to Dr. Sharifullah Khan, Dr. Zahid Anwar and Dr.Kashif Rajpoot, for their helpful suggestions and critique during my research work. I would like to acknowledge Dr. Nguyen (at Caltech) and Dr. Reddy (at Stanford) for their kind guidance and suggestions throughout my research.

I am also very obliged towards all my teachers who have been guiding me throughout my course work and have contributed to my knowledge. Their feedback, positive criticism and guidance helped me in carrying out this research work. I would like to acknowledge www.tbdb.org for their efforts in maintaining TB databases for research purpose and Gates Foundation for their funding research in Tuberculosis Gene Expression Analysis.

Sofia Aftab

Table of Contents

List of Abbreviations

TB	Tuberculosis
MDR	Multi drug resistant
PC	Pearson correlation
FD	Fisher distance
KD	Kullback leibler distance
LTB	Latent tuberculosis
PTB	Pulmonary tuberculosis
MTB	Meningeal tuberculosis
WHO	World Health Organization
HC	Hierarchical clustering
DNA	De-oxy ribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger ribonucleic acid
FISH	Fluorescent in situ hybridization
PCR	Polymerase chain reaction
RT-PCR	Reverse transcription PCR
SAGE	Serial analysis of gene expression

LIST OF FIGURES

List of Tables

Abstract

TB, Tuberculosis is a deadly infectious disease caused by the bacteria mycobacterium tuberculosis According to an estimate, one third of the world's population is infected with TB and new infections are occurring at a rate of about one per second. Most people infected with TB never develop active disease and only a small portion of the population develops active TB in various forms. Genetic profile of a person has been known to play an important role in patient susceptibility to active TB. There are certain genes in the human body which determine body's response to TB and hence, play a critical role in different clinical manifestations of active TB in a patient. These genes control dynamic state of cells (behavior) making up the "*gene expression profile"*. This expression profile is used for the analysis of genetic data.

Previous research has led to the discovery of a gene associated with TB susceptibility namely CCl1 but there is not always a single gene involved in inducing a phenotype (disease). Genes are often co-expressed and are co-regulated producing a specific protein (condition). Therefore our study is looking to determine the genes (known and novel) which are co-expressed with and co-regulate CCl1 and become a determining factor for causing TB. We use a distance metric based clustering method to find co-regulated genes instead of merely finding differentially expressed genes as carried out in previous studies. We use three different distance metrics (Pearson correlation, Fisher discriminant and Kullback leibler distance) and use them to perform hierarchical clustering to investigate what other genes are correlated with CCL1. Our study has led to the discovery of genes (forming three clusters) which are possibly not only responsible for causing TB but can also discriminate between different clinical forms of TB.

Our results show that using the three dissimilarity metrics proposed we could reduce the data-sets by filtering based upon dissimilarity in expression with CCl1 and subsequently use various clustering techniques to cluster similar genes. Most of the 21 genes discovered in our study are found to play a role in lung functioning and development and some are also seen to be active in spread of certain tumors. Hence, some new co-regulated genes of CCl1 have been discovered which were previously unknown. The work can be extended by applying techniques such as k-means and Dbscan which could again be used employing different dissimilarity measures to discover clusters.

Chapter 1

According to an estimate one third of the world's population is believed to be infected with TB [7] [8]. Moreover, new infections occur at a rate of about one per second [5]. TB appears in different clinical forms including Latent (passive form), Pulmonary (localized to lungs) and Meningeal (affecting the nervous system) and has been declared as the global emergency of the millennium by the World Health Organization (WHO) [55]. According to a WHO estimate, by 2020 people infected with active TB will reach up to 1 billion.

In Pakistan around 60,000 people die of tuberculosis (TB) every year and it ranks sixth amongst the countries worst affected by the disease. More than one million people have TB in Pakistan. One new person is infected every two minutes and one dies every eight minutes [54]. According to a research, a person infected with active TB will infect on average between 10 to 15 people every year [6]. Each year, there are 9 million people around the world who get sick with TB and there are almost 2 million deaths due to Tuberculosis worldwide [7].

Despite the desperate need for drugs and vaccines for TB no new vaccine has been developed for the last 90 years [8]. The first anti-TB drug, streptomycin, was developed in 1944, and it has been nearly 30 years since a new TB drug has been developed [9]. The drugs being used today were developed some 40 years ago and they seem to be losing their efficacy [8]. TB requires antibiotics treatment to cure it and the treatment is six months long. However, TB patients tend to stop their medication as soon as they begin to feel well. In these cases bacteria remains in their body in the passive form and can later attack again with greater force. In developed countries people are suffering from tuberculosis because their immune systems are compromised by immunosuppressive drugs, substance abuse, AIDS etc and hence TB proves to be leading killer of people who are HIV infected [10].

In many Asian and African countries 80% of the people test positive in tuberculin tests (BCG) [8]. No vaccine is available that provides reliable protection for adults. Even BCG cannot protect against Pulmonary TB (lungs). Recent studies have also discovered that BCG is more dangerous in children born with AIDS. Referring to a study conducted by WHO, The New York Times (July 2009)

reported that BCG is a live vaccine and can cause a serious form of bacterial infection that can rage through the body of an HIV-infected infant. According to the report, infection is fatal in more than 70 percent of the cases. [8] ". In countries like South Africa, where both TB and transfer of AIDS virus from mother to child is very common, the vaccine does not protect against TB and it may kill them with BCG disease" [8]. According to WHO estimation the largest number of new TB cases in 2008 occurred in the South-East Asia Region, which makes up 35% of incident cases globally [10]. Since humans are the only host for the mycobacterium tuberculosis, eradication is considered possible. Moreover, most of the infected will not develop active TB and those who do would develop TB due to genetic susceptibility to the disease.

There are certain genes which have been found to play a significant role during TB infection. The up regulation and down regulation of these genes is responsible for production of proteins that allow or enhance the progression of TB [56]. Genes regulate and deregulate each other making up a structural relationship amongst genes referred to as co-expression or co-regulation. Co-expression usually indicates deeper similarities between the genes or their proteins [3]. For instance they might be on the same pathway, a set of proteins which interact in sequential or regular ways, in order to communicate a message or perform a function (disease) within the cell. Since pathways function as a unit, so all genes in the pathway are needed and biologists expect all the genes coding for proteins in a common pathway to be co-expressed. [3]

Genes are often co-expressed with the genes whose transcript (DNA) they regulate. For example, if Gene A is a transcription factor which activates Gene B, then when Gene A is expressed we should see a corresponding increase in the expression of Gene B. Gene up regulation and down regulation strongly influences the host genetic susceptibility to different kinds of diseases. TB is one of those diseases which is strongly influenced by the regulation of gene CCL1 [3]. Gene regulation starts with transcription (replication of DNA) and hence, transcript information is needed to understand gene regulatory networks. Measurement of genes transcription levels in organisms under various disease conditions, under different behavior, at different developmental stages is used to build up 'gene expression profiles' which characterize the dynamic functioning of each gene in the genome. Gene expression profiling indicates the quantitative and qualitative change in genes under different disease conditions. Such regulation and de-regulation provides a basis to understand the reasons behind the

underlying disease. This understanding may lead to a new way of diagnostic treatment of the disease [50].

Gene expression analysis helps in understanding gene regulation, metabolic, genetic mechanisms of disease and the response to drug treatments. For example, if over expression of certain genes is associated with a certain disease, we can explore other conditions that affect the expression of these genes and similarity of other gene expression profiles. We can also investigate which compounds (potential drugs) can lower the expression levels of such genes [35] to potentially lower the chances of the development of active disease. Identifying genetic variants that increase or decrease an individual's susceptibility to disease can potentially lead to the targeting of preventive measures at those who are at greatest risk. This may also give valuable insights into the underlying molecular processes at a cellular level that are important in disease causation, opening the way for new and novel therapies to improve the outcomes of those with disease or who are at risk of disease [63]

Majority of the previous studies pertaining TB have discovered differentially expressed genes by comparing the expression levels of genes in different samples, however one study by Nguyen et al. is unique in the sense that they have not only proved that gene CCL1 is responsible for TB susceptibility using data analysis but also proved the fact empirically. However, it is an established fact that there is not always a single gene involved in causing a disease [4] but genes usually regulate and deregulate each other (making up structural relationships) in causing a specific phenotype.

This study is different from previous studies as it does not only find the differentially expressed genes but we have filtered out the high variance genes and applied distance based clustering on them to find the most relevant set of genes. We use three different distance metrics namely Pearson Correlation (PC), Kullback leibler distance (KD) and Fisher distance (FD) to compute distance values of all genes with respect to CCL1 and filter out the highly variant genes using thresholding over the distance values. Our study has employed distance based metrics and hierarchical clustering to explore the structural relationships between genes and a list of co-regulated genes is identified which play an important role in transformation of cells from a healthy state to an unhealthy state and also discriminate between different clinical forms of TB.

CCL1 is already discovered to be associated not only with TB but is responsible for different clinical forms of TB [3] but there could be multiple genes in the common pathway to establish certain condition or disease. So we attempted to find known and novel co-expressed and co-regulated genes of CCL1 and our results show that there are 21 genes which are co-expressed with CCl1 and hence are ultimately associated with TB. For evaluation of our results we used the Pearson Correlation Coefficient to estimate the similarity in expression levels of the genes in the three clusters discovered to CCL1. In this study, we have identified three clusters of genes that are highly co-expressed with CCl1 (as indicated by the high values of Pearson Correlation Coefficient of greater than 0.9) and hence play a role in TB susceptibility. The results in detail are presented in Chapter 4.

To understand how gene regulation works and how it impacts our health, we present some discussion on the biological background of the phenomenon and how gene expression levels are measured in the next section.

1.1 Biological Background

This section deals with the biological background of genes and gene expression which is important in order to understand the context of this research.

1.1.1 Hereditary material

Nucleus of every cell contains hereditary material called the DNA and the DNA helix forms chromosomes as shown in Figure 1.1. Each chromosome contains thousands of genes **w**hich contain the genetic instructions for the generation of proteins that plays a significant role in the development and functioning of all living organisms. Transcription is the process by which DNA on which genes are located converts (transcribes) into messenger ribonucleic acid mRNA, an intermediate product of protein.

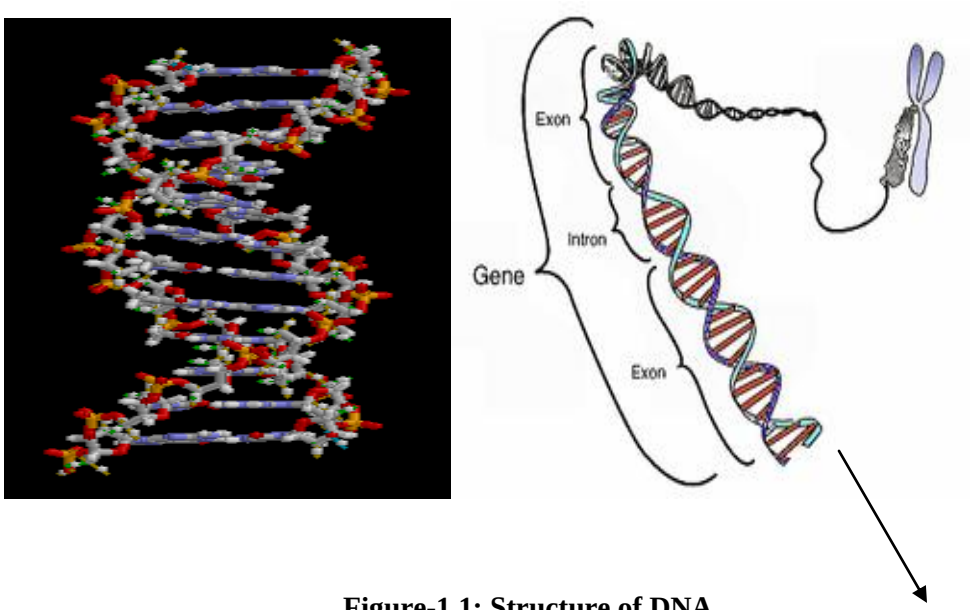


Figure-1.1: Structure of DNA

1.1.2 Gene regulation

Gene regulation starts with transcription and it provides the cell, control over all structure and function, and is the basis for all phenomenons like disease susceptibility, cellular differentiation, morphogenesis, versatility and adaptability of all organisms. There are two main phenomena's of gene regulation i.e. up-regulation and down-regulation. Up regulation of gene occurs as a result of a stimulus or signal (originating internal or external to the cell), leading to increased expression of one or more genes and as a result the proteins produced by these genes are increased. Down regulation is defined as decreased gene expression and resulting in less production of a protein.

1.1.3 Gene expression profile

Gene expression profile is a dynamic process telling us what cell is actually doing now depending upon the conditions. If a gene is used to produce mRNA (protein), it is considered "on", otherwise "off".

1.1.4 Gene expression level measurement techniques

There are number of techniques that can be used to measure gene expression levels, starting with the older techniques such as northern blot and western blot etc. to more established technologies such as RT-PCR. Only one of the techniques i.e. RT-PCR is used in Pakistan. The data used in this study is also acquired using RT-PCR.

1.2 Thesis Statement

"In this thesis multiple distance metrics based hierarchical clustering is carried out to discover co-expressed and co-regulated genes responsible for TB susceptibility".

1.3 Goals and objectives

- Our study is aimed at identifying genes potentially involved in causing a specific clinical form of TB. We intend to determine the collection of genes that up regulate and\or down regulate causing the susceptibility to disease [5].
- In our study we intend to create different clusters of co-regulated genes. Each cluster representing different levels of correlation
- Given the properties of known genes we can predict the behavior of novel genes clustered with well known genes. These novel genes could also be involved in causing the phenotype (disease).

1.4 Thesis outline

Chapter 2 provides the detailed literature survey of the work in estimating expression profiles of various genes in different forms of TB and usage of data mining techniques in genes expression analysis. Chapter 3 explains the methods and techniques used in this work. Chapter 4 presents and discusses the results of our work. The last chapter concludes the thesis with future directions.

Chapter 2

Literature Review

In this chapter, we present the literature review related to our research problem. We have divided the literature review into two parts. First is the review of the various data mining techniques that have been used in analyzing genes expression data in TB infection or in similar areas. Second part is the review of the literature related to genes discovered so far known to play a role in tuberculosis infection

For carrying out analysis on gene expression data there is a need to create a database making up gene expression profiles which contributes towardse the dynamic function of each gene in the gene dataset. Different types of data mining techniques can be used for analysis of genetic data. Building a matrix of gene expression profiles can help us explore gene regulation process, genetic involvement of diseases, and response of gene expression levels to different drug compounds. We can also investigate gene regulatory networks and gene-gene interaction processes which are also the main focus of our research work.

2.1 Review of data mining techniques in genes expression analysis

- For analysis purposes, we represent the gene expression profile in a matrix called gene expression matrix where each row represents gene and column represents samples (various experimental samples often at different stages of disease development) and each cell represents expression level of a gene in a specific condition. We have also used gene expression matrix for our research where rows and columns represent genes and expression levels of a gene respectively in different forms of TB. In literature, there are two basic themes in gene expression analysis [31]: Compare the rows (genes) to find similarity between genes which ultimately leads to co-regulation of genes. It means genes are functionally related.
- Compare columns (samples) of data sets to find genes differentially expressed amongst the different samples/conditions.

For comparison we need some dissimilarity measures and to the best of our knowledge there is no one best technique that is proven to work well in all cases. We have used three different dissimilarity measures in our research to transform the data in such a way that different clusters may be constructed representing different groups of genes. By using such transformations co-regulated genes of already established genes like CCL1 can be discovered.

After computing the dissimilarity matrix using a dissimilarity measure: we perform unsupervised (clustering) analysis instead of finding merely differentially expressed genes as it has been done in majority of the previous studies. For instance, Lee, et al. [41] discovered differentially expressed genes amongst old i.e. 30 months and adult i.e. 5 months old mice. He found up regulated and down regulated genes characterizing various functions like metabolism and neuronal growth. The same process of up regulation and down regulation could also characterize different conditions and types of diseases.

One TB study [3] also discovered differentially expressed genes amongst the three different samples (Latent, Pulmonary, Meningeal) and after analyzing the up regulated and down regulated genes they found that CCL1 gene is the most differentially expressed between Latent and Pulmonary TB and hence is responsible for determining with the susceptibility to developing active TB. Our research intends to investigate gene co-regulatory networks (group of genes which affect each other) during TB infection using data mining techniques.

2.2.1 Clustering

Clustering has been used to classify genes into co-regulated groups [43-46] i.e. genes whose expression profile changes relative to each other. Clustering has been applied to group similar genes or samples to form clusters. A number of techniques and algorithms have been applied in gene expression data analysis. Hierarchical clustering [32], k-Means clustering [33, 34] and SOMs [35] are some of the techniques used for genes expression data analysis. We have employed hierarchical clustering in our analysis.

Hierarchical clustering [32] has been used for gene and sample clustering in different studies. Our study has used the technique to cluster genes as well as to cluster samples (in preprocessing); details are given in Chapter 3. Hierarchical clustering has been used by Alizadeh et al. [42] to find new

possible tumor subclasses. In this paper, diffused large B-cell lymphoma (DLBCL) was studied using 96 samples of normal and malignant lymphocytes. After application of a hierarchical clustering algorithm to the samples they showed that there is variability in genes expression amongst the tumors of DLBCL patients. They identified two groups of DLBCL patients which had different gene expression patterns at different B-cell differentiation stages. Surprisingly, these two groups have proved good co-relation with patient survival rates thus proving the validity of the clusters. This study provides a clue that a similar technique could also be used to construct different clusters of human genes which play a role during TB infection where each cluster may be responsible for different types of TB.

Clustering has also been used in identifying and classification of genes associated with Alzheimer disease (AD) [50]. In a study 67 genes were identified [50]. 17 of them were already known to be associated with AD and in this study, 20 new genes were found to be associated with AD while 30 uncharacterized sequence tags were also discovered. In this study clustering has been used to discover some novel genes by clustering them with well known genes. A similar type of analysis may lead to discovery of novel related genes clustered with genes already associated with TB (such as CCL1) as carried out in our study.

Clustering of expression profiles has been used for grouping related genes as well as for grouping samples. In the early days of gene expression analysis, DeRisi et al. [37] used clustering for grouping similar genes. Later on further studies were carried out by examining genes in different conditions such as sporulation [38], in cell cycles [39] and gene regulatory machinery [40] and various co-regulated genes were revealed. Therefore, similarly in our study we intend to determine co-regulated genes of some already established genes proven to play a role in TB susceptibility.

Tavazoie et al. [34] clustered expression profiles of 3000 most variable genes during cell cycle into 30 clusters using the K-means algorithm and discovered a strong sequence of patterns. We also tried K –means clustering in our analysis but hierarchical clustering results were better. Fuzzy K mean and principal component analysis (PCA) has been combined in hierarchical clustering for analysis of published genomic expression data which showed the relationship between response of cells and environments. Their study successfully identified clusters of those genes which are functionally related [59]. The analysis identified previously unrecognized similarities in the expression of genes and showed correlation amongst the environmental factors. The fuzzy clustering method was also

able to identify gene clusters which were not identified by other clustering technique like hierarchical or standard K-means clustering. The study identified almost 90% of the clusters in the data set but was unable to find small clusters which can be identified using hierarchical clustering and which are of great interest in genes expression analysis [58]. This is the reason that we have chosen hierarchical clustering for our research.

PCA is a technique which has been used previously in gene expression analysis. It is a dimensionality reduction method which is used for feature selection. A study by Yeung and Ruzzo shows that clustering with PCA does not necessarily improve cluster quality. PCA fails to capture all of the cluster structure. Hence, the study recommends to not to use PCA before clustering [59]. Hence, in our study we have not used PCA before clustering. We applied some variance filters for reducing our gene data set (dimensionality reduction method) before clustering our data as described in Chapter 3.

SOMs summarize the data by extracting most prominent patterns and place the similar patterns in neighborhood of each other. One study grouped up to 6000 human genes into biologically relevant clusters. Later on certain genes and pathways were highlighted which could be used in the acute promyelocytic leukemia treatment [48]. Studies such as these suggest that similar studies could lead to discovery of genes which could be used as target genes for TB treatment.

We summarize the data mining techniques and major findings in Table 2.1.

Sr. No	Studies done by	Technique	Findings	Disease/Condition
1	Alizadeh	Hierarchical clustering	Two main groups of tumor were separated out	Tumor
2	Wolkera and Smitha	Hierarchical clustering	67 genes were associated with AD	Alzheimer disease
3	P.T. Spellman	K- means clustering	Various co-regulated genes were revealed	Sporulation in cell cycles
4	Sandip and Markus	PAM	18 predictors (Genes) for AD	Alzheimer disease
	Gasch and Eisen	Fuzzy K- Mean	Clusters associated with different conditions	Yeast genetic Data
	Eisen and Spellman	SOM	genes for leukemia treatment were discovered	Leukemia

Table- 2.1 Showing summary for review of data mining techniques

Most of the studies have been conducted to find genes associated with different types of diseases like Cancer, Alzheimer and Leukemia and only two studies up till now have been specifically focused on different forms of TB which are described in the next section. Although a number of genes have been discovered that are differentially expressed in TB but only a single gene (CCL1) has been empirically proven to play a role in susceptibility to different forms of TB. As genes usually co-regulate each other hence some other genes may also be discovered to play a role in TB susceptibility especially the ones co-regulated with CCL1. Such work may lead to development of new medicines for TB. Hierarchical clustering is the most widely used technique to discover genes groups associated with different conditions [32]. Our study has also used hierarchical clustering for finding co-regulated genes for TB and we have used three different distance metrics Pearson correlation (PC), Kullback leibler distance (KD), Fisher distance (FD) for transforming gene expression data. PC measures the

importance of direct linear relationship between two variables. The value is always between [-1; 1] where 1 is strong positive relation, 0 is no relation and -1 is a strong negative correlation [62]. KD is used for measuring asymmetric distance between two variables [61] and FD is used to separate two genes by using mean and variance properties [60]. In this study, KD and FD is being used for the first time for genes expression data analysis.

We have studied different data mining techniques for gene expression analysis. But it is also very important to identify important genes involved in Tuberculosis infection. The overall literature survey of human genome associated with Tuberculosis infection is presented in the following section

2.2 Literature Review of Human Genomic Study in Different Forms of Tuberculosis

In literature, majority of the studies have engaged healthy donors, cell lines or murine cells [12, 13, and 14]. Only two previous studies have compared the gene expression profile of individuals with different clinical forms of TB. The work on which our study is based used genetic data of macrophage cells. The macrophages (defense cells) and dendrites cells (receptors for pathogens) trigger the immunity function on pathogen recognition by receptors on the surface of these cells [15]. These receptors include CD14 and Toll-like receptors (TLRS) [16, 17], nuclear factor (NF- kB) and several other macrophage cell surface molecules [18-22] that are involved in adherence and uptake of Tuberculosis [14].

In literature, one study investigated the early variation in expression of macrophage genes during and immediately after phagocytosis (entrance of pathogen) of TB [14]. This study claimed that 375 human genes are generally involved in immune regulation. Some of these genes belong to chemokine and interleukin IL-8 family of genes. After 1,6,12 hours following infection with M. tuberculosis several genes were induced significantly. This induction and variation in gene expression is responsible for causing different behaviors (diseases) which form the basis of our study as well.

Another study [12] examined 6,800 macrophage genes (defense genes) using microarrays and expression of 977 genes was considerably changed on disclosure to one or more bacteria (M. tuberculosis) and the study came up with two important proteins IL-12 and IL-15 which were found

to be significant for host defense and immunity against TB. This up regulation and down regulation of gene expressions during mycobacterial tuberculosis infection forms the main motivation of our study to find a set of genes (proteins) responsible for causing TB infection. IL-12 plays an important role in production of T helper 1 (Th1) immune responses [23] which is significant for host resistance to TB infection in mice and in humans [24-27]. Less production of IL-12 production may boost the survival of M. tuberculosis infection. This study claimed that Il-12 and Il-15 proteins could be useful in treatment for clinical TB. However, in our work we are mainly interested in genes responsible for TB susceptibility and progression. CCL1 [4] has been empirically and statistically proven to be responsible for TB susceptibility and hence we have chosen to investigate CCl1 further.

Another study [28] has shown that chemokine CCl20 is dramatically over expressed in M. tuberculosis infection and they have also confirmed that dendritic cells are a appropriate host for mycobacteria proliferation. CCL20 is a chemokine that attracts immature dendritic cells and supress the characteristic production of reactive oxygen species (ROS) induced by M. tuberculosis in monocytes which may change cell activity. This fact provides a clue that there could be some other genes (chemokine's) which are affected by CCL20 or regulated by CCL20 but our study has employed techniques to find co-regulated genes of already established genes in TB infection like CCL1 as CCl1 is empirically proven through RT-PCR but the role of CCL20 in TB progression has not been proven through such empirical studies.

A study by Zahra et al. [29] has shown that the increased expression of gene CCL2 and TNFα (protein) in Pulmonary TB (PTB) patients may support effective leucocytes recruitment and M. tuberculosis localization and CCL2 alone is related to severity of TB. But, as explained earlier, there is not always a single gene responsible for causing TB or any other infection [4]. So there is a need to discover other genes which are co-regulated with the already established genes. As CCL1 is empirically proven through RT-PCR therefore we have decided to find co-regulated genes of CCL1 instead of CCL2 or CCL20. CC stands for cytokines or chemokines and cytokines or, chemokines are responsible regulate the migration, trafficking, homing and activation of monocytes, macrophages and other leucocytes. CCL20 is a chemokine which supresses the characteristic production of reactive oxygen species (ROS) induced by M. tuberculosis in monocytes which can affect cell activity. CCL2 plays a vital role in protection against tuberculosis in the murine model. CCL1 is responsible for differentiating different form of TB and is also associated with susceptibility of TB. *M. tuberculosis* infection of macrophages results in the generation of certain other chemokines

CCL3, CCL4 and CCL5 and these are required for decline of its growth. As these cytokines and chemokines had previously been proved to be differentially expressed according to disease location (Pulmonary and extra Pulmonary) and also disease stage [29] hence these genes may be studied separately.

To our knowledge only two previous studies have made comparative study of gene expression profiles of individuals with different clinical forms of TB [10] [32]. Mistry et al. collected whole blood samples from individuals suffering from active, Latent, cured (following 1 disease episode) and recurrent TB (following 2–3 episodes) [32]. Their discriminate analysis hypothesized that 9 genes shown in Table 2.2 could distinguish between four clinical TB groups [32] but theses 9 genes were not confirmed by a second study [10]. These genes were also not confirmed by RT-PCR. Moreover, genes discovered in this study have not been found to be relevant in any other study.

Sr. No	Differentially Expressed Genes in Human Genome During TB Infection
1	RIN3
2	LY6G6D
3	TEX264
4	C14orf2
5	SOCS3
6	KIAA2013
7	ASNA1
8	ATP5G1
9	NOLA3

Table-2.2: Differentially expressed genes in TB by Mistry et al study

Genes in table-2.2 were found to be highly expressed in human genome during TB infection by Mistry et al. This study has also suggested that these genes could distinguish the four clinical TB groups (Latent, Pulmonary, Meningeal and cured)

In another research work [3] (on which our study is also based), demonstrated that 1608 genes in three different types of diseases (Latent, Pulmonary, Meningeal) were up regulated and down regulated during mycobacterial TB infection. Some differentially expressed genes are shown in Table 2.3. The study was also confirmed empirically using RT-PCR and one gene i.e. CCL1's expression remained significantly different for the three different clinical types of disease. Later on, they also confirmed that CCL1 is the only gene involved in host susceptibility to TB [4].

Sr. No	Differentially Expressed Genes in Human Genome During TB Infection
1	CCL1
2	INHBA
3	TSLP
4	LY6K
5	IL12B
6	MMP1
7	CCL20
8	HAS1

Table-2.3: Differentially expressed genes in TB by Nguyen et al Study

Genes in table -2.3 were highly expressed in Human Genome during TB infection as found by Nguyen et al. Moreover it has been proven empirically that CCL1 is responsible for causing different forms of TB and is also associated with TB susceptibility. Table 2.4 shows a summary of the majority of the studies carried out so far for the genes associates with TB during infection. Each study is presented with its findings (genes active) during specific disease conditions and specifies which genes are involved in susceptibility to TB.

Sr. No	Studies done by	Findings	Disease	Susceptible to TB
1	Silvia Ragno et al.	375 genes involved in immuno regulation	TB	Not proven
2	Gerard	IL-12 and IL_15 protein involved in defense against TB	TB	Not proven
3	Hirsch	IFN-γ9 protein and some known and novel genes involved in immunity against TB	TB	Not proven
4	O. M. Rivero-Lezcano	CCL20 is a gene over expressed in TB patients	TB	Not proven
5	Zahra Hasan	CCL2 and TNFα involved in leukocyte recruitment and CCL2 is associated with TB	TB	Not proven
6	Mistry et al	9 genes can distinguish four TB groups	TB	Not proven
7	Nguyen	CCL1 distinguish different forms of TB and is also responsible for susceptibility to TB	TB	Proven empirically through RT-PCR

Table-2.4 Showing literature survey of human genomic data during TB infection

Genes are often co-expressed with the genes whose transcript (DNA) they regulate making up structural relationship between genes. Gene up regulation and down regulation strongly influences the host genetic susceptibility to different kinds of diseases as discussed in Chapter 1. There is only a single study which has discovered empirically as well as statistically a gene (i.e. CCL1) responsible

for TB susceptibility [4]. TB is strongly influenced by the regulation of gene CCL1 but there is not always a single gene responsible for causing any phenotype. Our study has intended to find the list of genes that get co-regulated with CCL1 (i.e. if the expression level of CCL1 increases we see corresponding increase in expression levels of other genes and vice versa).

Chapter 3

Methods and Techniques

Genes regulate and deregulate each other, forming structural relationships. Our study, explores the structural relationship of human genomic data during TB infection. Majority of the previous studies have only found differentially expressed genes during TB infection. There is only one previous study which empirically proved the significance of gene CCL1 using RT-PCR during TB infection [4] but we know that there is always a group of genes collectively responsible for causing different disease conditions. Our study intends to find genes that co-regulate with CCL1, playing a role in TB infection.

In our study, we have performed the following main steps for analysis of human genetic data during TB infection:

1. Preprocessing
2. Genes filtering
3. Hierarchical Clustering (data mining technique)

3.1 Preprocessing

Preprocessing is the first step in our analysis and starts with cleaning of data involving outlier removal and missing values imputation. For outlier detection first hierarchical clustering was applied on array samples (patients) by using Pearson correlation as a distance metric. Dissimilarity distance metrics for patients has been represented in the form of a histogram. Each bin of the histogram shows the frequency of distance values in the distance metric. Histogram of the inter array correlation (IAC) is shown in Fig. 3.1 which is right skewed having a long tail. Long tail of a histogram represents extreme distance values of a dataset. These very small distance values indicate presence of outliers. These outliers should be removed from the dataset to achieve unbiased results of analysis.

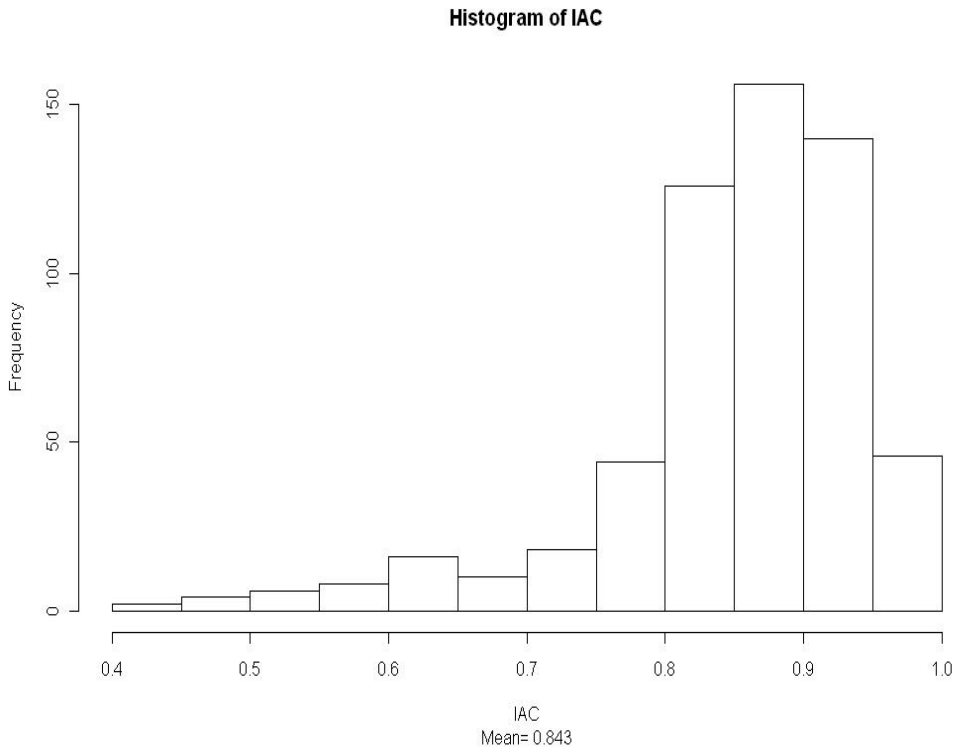


Figure-3.1: Histogram with an outlier

We took summation of the product of each correlation value with its frequency in IAC matrix (24x24) and divided it with the total number of distance values. 0.843 is the mean or average of the correlation value for all samples (24). Result of hierarchical clustering has been depicted in the form of dendrogram as shown in Fig.3.2.

Dendrogram in Figure 3.2 shows the clustering structure amongst different patients of TB. Similar samples (patients) have smaller branch height and as the similarity decreases, samples represent greater branch height. After dendrogram analysis it was found that one of the patient's samples (Latent un-stimulated) is an outlier as already confirmed by Nguyen et al. [3]. Patient 6 of Latent TB (L6) has a greatest height of branch (due to a great number of missing values) and hence it is clearly an outlier. So we removed this patient and are left with 23 patients.

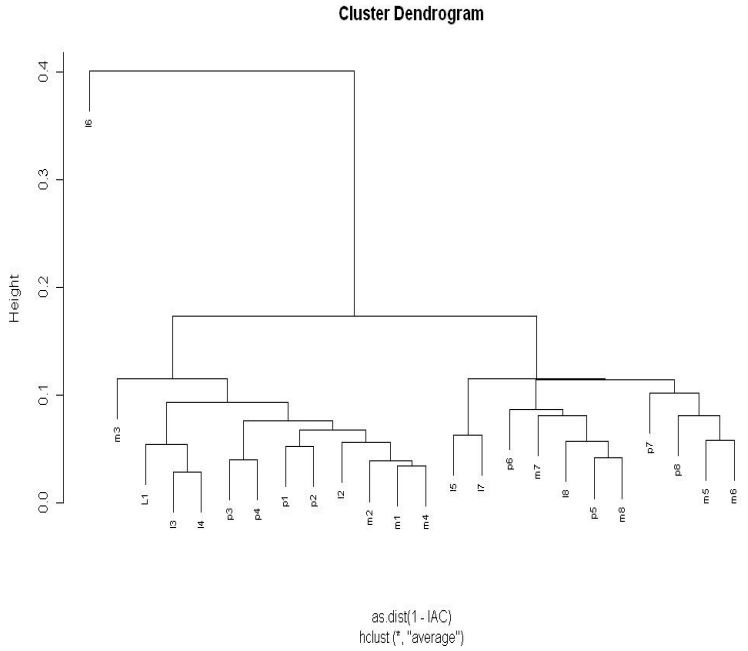


Figure-3.2: Cluster dendrogram showing an outlier

After removing this patient we again draw a histogram and dendrogram which is shown in Fig 3.3.

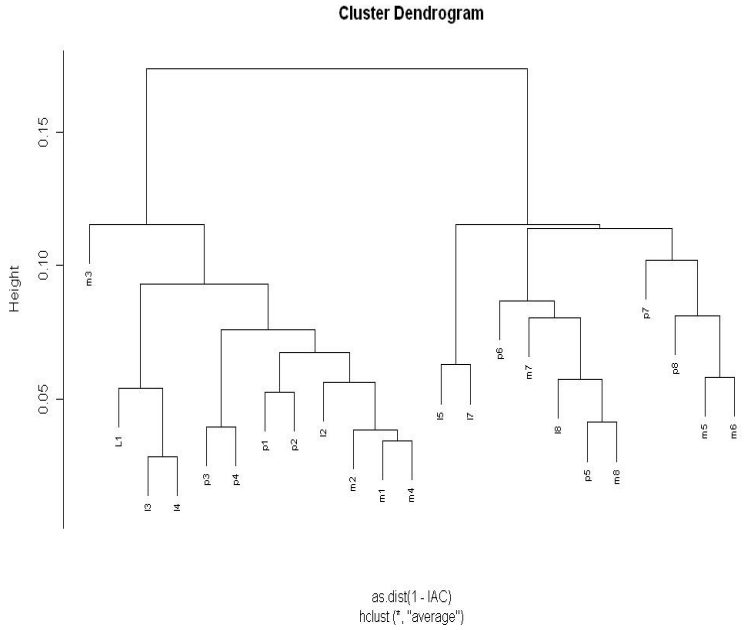


Figure-3.3: Cluster dendrogram after removal of an outlier

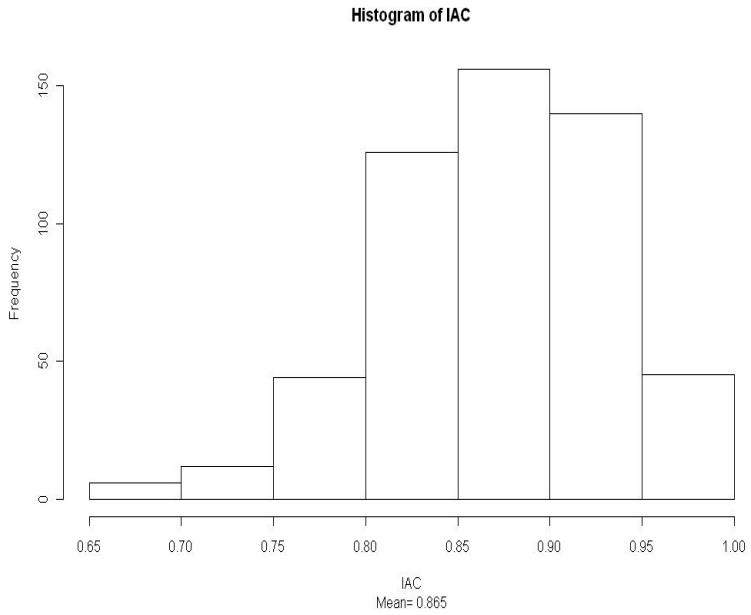


Figure-3.4: Histogram after removal of an outlier

Figure-3.4 is an improved histogram as skewness has been reduced and tail has been shortened due to removal of missing values. This short tail indicates that outlier has been removed from the dataset as we have removed patient 6 from our data.

3.2 Genes filtering based upon dissimilarity distance

After cleaning the data we compute three different dissimilarity matrices using distance metrics Pearson correlation (PC), Fisher distance (FD), and Kullback-Leibler distance (KL). For each dissimilarity matrix we compute dissimilarity distances from CCL1 to all other genes. Pearson correlation is an established technique for finding similarity values but we have used it to find dissimilarity.

3.2.1 Pearson Correlation

Pearson correlation is used to find similarity between two variables**.** It measures the power and the direct linear relationship between two variables. The value is always between [-1; 1] where 1 is a strong positive relation, 0 is no relation and -1 is a strong negative correlation. It is the most widely used correlation coefficient [62]. The formula is as follows.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

$$PC = 1 - |r|$$

Where X and Y denotes the expression levels of genes being compared for the different samples included in the study. It is important to note here that CCl1 is usually used to measure similarity between objects but in our case we subtracted the absolute of the coefficient from 1 to compute dissimilarity.

3.2.2 Fisher Distance

It is suitable for classification as it measures the holistic distance between two classes and in our study it separates the two genes. It is an efficient criterion of divisibility between two classes and in our case two genes, which is broadly used in pattern recognition, and it computes the ratio of scatter between-class and within-class of two classes/genes. Larger ratio means larger divisibility of the two classes [60].

$$\mathbf{FD = \sum \mu_x^2 \, \mu_y^2 / \sum (\sigma_x - \sigma_y)^2}$$

Where μ_x and μ_y are means of the expression profiles of genes X and Y and σ_x and σ_y is the standard deviation of the expression profiles of genes X and Y.

3.2.3 Kullback-Leibler Distance

KL distance is a natural distance function from a "true" probability distribution, p, to a "target" probability distribution, q. It is an asymmetric distance and hence, we have calculated it in both directions, i-e between distribution 2 and 1 and 1 to 2 (gene 1 to gene 2 and gene 2 to gene 1) [61].

$$\mathbf{KL = \Sigma X_i ln X_i / Y_i + \Sigma Y_i ln Y_i / X_i}$$

It is noticed that the three distance metrics described above compare the gene expressions based upon different characteristics of the gene expression values. PC measures the difference between the expression profiles in relevance to the difference amongst the corresponding gene expression values while FD is a more holistic measure that computes the difference based upon the mean and standard deviation of the gene expression profiles while Kullback-Leibler compares the gene profiles on a per sample basis and normalizes the difference using a log of the ratio of the two values. Hence, three different measures provide insights in to how the two genes compare with respect to each other.

As CCL1 is high variant gene and we have intended to find co-regulated genes of CCL1, so we need to keep highly variant genes and discard rest of the genes to make our analysis more simple and clear. To keep highly variant genes we applied gene filtering process, which removed all non variant genes. We choose certain threshold and applied it on distance matrices, which removed all the genes having distance values above the specified threshold.

After computing dissimilarity distance of CCL1 for the whole dataset (16078 genes) using PC, KD, and FD, we use thresholding to perform filtering. This was done to find genes most similar to CCL1 using clustering. A threshold was applied on each distance metric value and the gene having a greater dissimilarity value as compared to the threshold is removed. The number of genes left for various threshold values are given in Table 3.1. As dissimilarity distance value ranges from"0" to "1", hence threshold is applied on dissimilarity distance ranging from "0.9" to "0.14". The genes were filtered gradually to get most valuable genes. Initially the threshold was large (0.9) and genes were still not distinguishable in clusters so thresholding based upon dissimilarity distance was gradually increased to extract most similar or co-regulated genes of CCl1. "0.14" was found to be a good threshold as a limited number of genes are left and the clusters obtained are of good quality.

Sr. no	Distance Dissimilarity Threshold	Genes Left
1	0.9	14872
2	0.8	12353
3	0.7	8085
4	0.6	4823
5	0.5	2831
6	0.45	1873
7	0.4	1359
8	0.35	918
9	0.3	648
10	0.25	423
11	0.2	264
12	0.18	182
13	0.15	83
13	0.14	21

Table-3.1: Relationship of dissimilarity distance with number of genes left

3.3 Hierarchical clustering

After preparing data for clustering, we perform distance based hierarchical clustering using the three dissimilarity distance values (PC, KD, FD) rather than on actual data values to obtain clusters of genes which are similar in terms of distance values to CCL1. The genes which are closer to each other i.e. have low distance, are more similar to each other and hence, are grouped in the same cluster. There are two types of algorithm used in hierarchical clustering agglomerative and divisive. We have used Agglomerative in our clustering process.

In order to decide which clusters should be merged (for agglomerative), or where a cluster should be divide (for divisive), a measure of dissimilarity is needed. Mostly hierarchical clustering uses appropriate distance metric (a measure of distance between pairs of observations), and a linkage parameter which specifies the dissimilarity of sets as a function of the pair wise distance values of observations in the data sets. We used Euclidean distance as a metric for Hierarchical clustering. Its formula is as follows.

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

It is important to note here that the Euclidean distance is measured for the three different distance values obtained earlier. There are many linkage parameters of Hierarchical clustering we have studied all separately. Linkage parameters are defined as follows.

3.3.3 Linkages

Following are the notations used by various linkage methods

- Cluster r made up of clusters p and q.
- Number of objects is represented by N_r in cluster r.
- ith object is represented by X_i in cluster r.

- *Single linkage*, also called *nearest neighbor*, takes up smallest distance between objects in the two clusters: It has got some striking theoretical properties and can be implemented relatively proficiently, so it has been widely used. However it has propensity towards forming long irregular clusters or chaining which is suitable for delineating ellipsoidal clusters but not appropriate for making spherical or weakly separated clusters.[64]

$$d(r,s) = \min(dist(x_{ri}, x_{sj})), i \in (i,\ldots,n_r), j \in (1,\ldots,n_s)$$

- *Complete linkage* also called *furthest neighbor*,considers the largest distance between objects in the two clusters: Inter cluster similarity is generated by comparing least similar pair between two clusters. It is called complete link because all data points within a cluster are grouped together with some smallest similarity. Small compactly bound clusters are uniqueness of this method.[64]

$$d(r,s) = \max(dist(x_{ri}, x_{sj})), i \in (1,\ldots,n_r), j \in (1,\ldots,n_s)$$

- *Average linkage* uses the average distance value between all data pairs in any two clusters:

All data points contributing inter cluster similarity, resulting in a structure that lies between the loosely bound single link clusters and compact bound complete link clusters. Average linkage has been considered good in evaluation studies of various clustering techniques.[64]

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj})$$

- *Centroid linkage* is characterized by using the Euclidean distance between the centroids of the two clusters:[64]

$$d(r,s) = \left\| \bar{x}_r - \bar{x}_s \right\|_2$$

where

$$\bar{x}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri}$$

- *Median linkage* takes centroids of the two clusters and use the Euclidean distance between weighted centeroids,

$$d(r,s) = \left\| \tilde{x}_r - \tilde{x}_s \right\|_2$$

For the clusters *r* and *s* $\tilde{x}_r$ and $\tilde{x}_s$ are weighted centroids. If cluster *r* was created by combining clusters *p* and *q*, $\tilde{x}_r$ is defined recursively as

$$\tilde{x}_r = \frac{1}{2}(\tilde{x}_p + \tilde{x}_q)$$

- *Ward's linkage* uses the incremental sum of squares; as a result of joining two clusters the the total within-cluster sum of squares increases. The within-cluster sum of squares is defined as the sum of the squares of the distances between the centroid of the cluster and all data points in the cluster. The sum of squares measure is the same as the following distance measure d(r,s):

$$d(r,s) = \sqrt{\frac{2n_r n_s}{(n_r + n_s)}} \left\| \bar{x}_r - \bar{x}_s \right\|_2,$$

where

- $\| \|_2$ is Euclidean distance
- $\bar{x}_r$ and $\bar{x}_s$ are the centroids of clusters *r* and *s*
- n_r and n_s are the number of elements in clusters r and s

It is also named as minimum variance method because it produces consistent clusters and symmetric hierarchy and its property of center of cluster of gravity is a useful way of representing clusters. It has been proved to be good at getting better cluster structure but it is susceptible to outliers and poor at getting elongated clusters. [64]

We also applied other clustering techniques like K- means and Density based clustering methods but we could not find prominent patterns or clusters. Although hierarchical clustering has also been used in a previous study [3] but the study employed actual data values (expression levels) for clustering to detect the patterns amongst different clinical disease phenotypes of TB. But hierarchical clustering in our study has extracted most significant co-regulated genes responsible for causing TB by using distance based clustering on metric values rather than on actual data values.

We have applied clustering with different linkage techniques (average, single, complete, median, centroid, and ward). We explored the maximum clustering limit from ‘4’ to ‘100’ based upon the number of genes in the dataset. Maximum clustering limit can be defined as the maximum number of clusters which could be attained from hierarchical clustering by cutting the tree at different appropriate levels. If we increase the maximum clustering limit, cluster size would be small and if we decrease the maximum clustering limit, cluster size would be large. We used different clustering limits in our dataset but we obtained higher quality clusters for a limit of 4. Moreover, 4 clusters were consistently found to be the best irrespective of the clustering limit, so we choose the clustering limit of 3-5 for our evaluation. By applying filtering and threshold on dissimilarity distances of genes, we got a small number of similar genes (21) to CCL1; genes which have low dissimilarity distance with respect to CCL1 are more similar to CCL1. We applied hierarchical clustering on data set of ‘21’ genes with clustering limit varying from ‘5’ to ‘3’.

Threshold is the maximum limit of dissimilarity distance of all genes with respect to CCL1. Maximum clustering limit means that specification of a maximum number of clusters which could be formed in a clustering process by cutting dendrogram at various appropriate levels. Dendrogram show how genes are grouped in different clusters and structural relationship of all the genes within each cluster.

Dissimilarity Threshold	No of genes left
0.4	1359
0.2	264
0.18	182
0.15	83
0.14	21

Table-3.2: Results of clustering with different threshold

The clustering results for dissimilarity threshold at "0.5" and "2831" genes are shown Fig 3.5. Genes are not forming any clusters so we reduced threshold to "0.4".

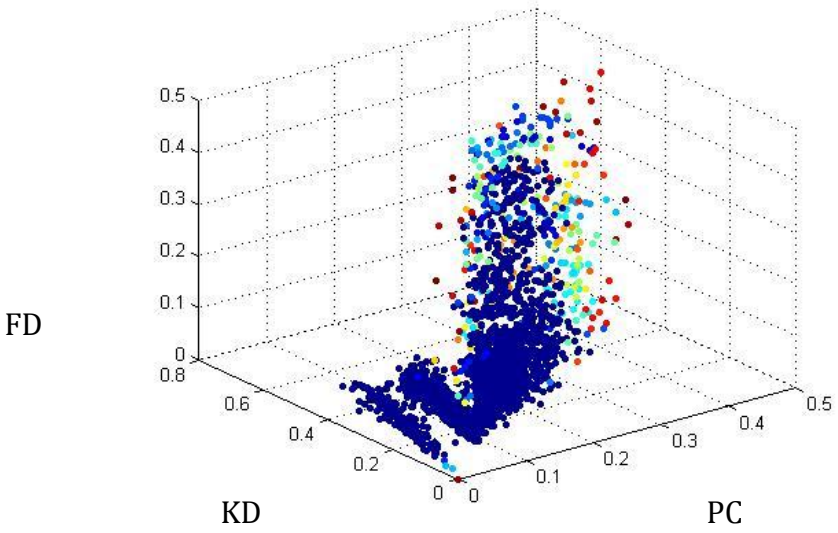


Figure-3.5: Initial results of clustering of 2831 genes

Figure-3.5 shows poor clustering when the threshold on dissimilarity distance was “0.5” and genes were “2831”. Maximum clustering limit is at “200”. There was not much improvement when the clustering limit was reduced. The clustering results for dissimilarity threshold at “0.4” and “1359” genes are shown in Fig 3.6. Still genes are not forming clear clusters so the threshold was reduced further.

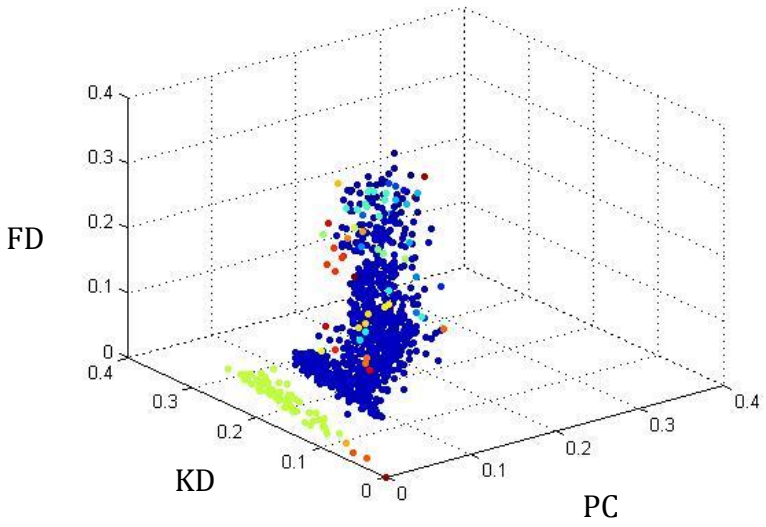


Figure-3.6 Results of clustering with 1359 genes

Figure-3.6 plot shows the clustering results when the threshold on dissimilarity distances was “0.4” and genes were “1359”. We gradually reduced the threshold value to 0.2 and got ‘264’ genes and then performed hierarchical clustering on them using different clustering limits. With a set of 264 genes, clustering was performed and results are shown in Figure-3.7 (a) and 3.7 (b). Now the genes have shown some patterns but still it needs more refinement.

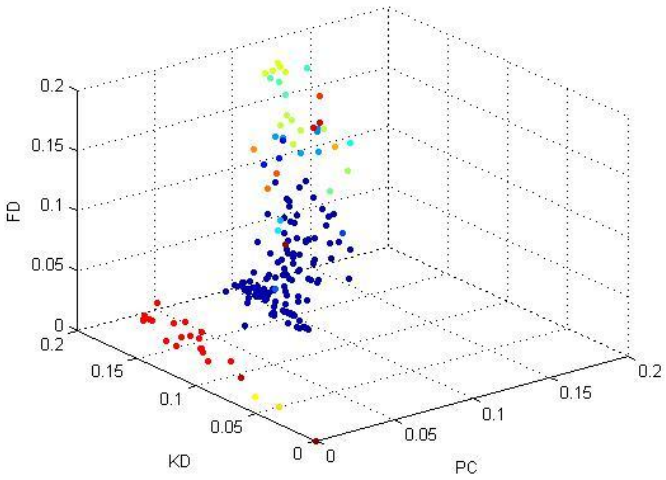
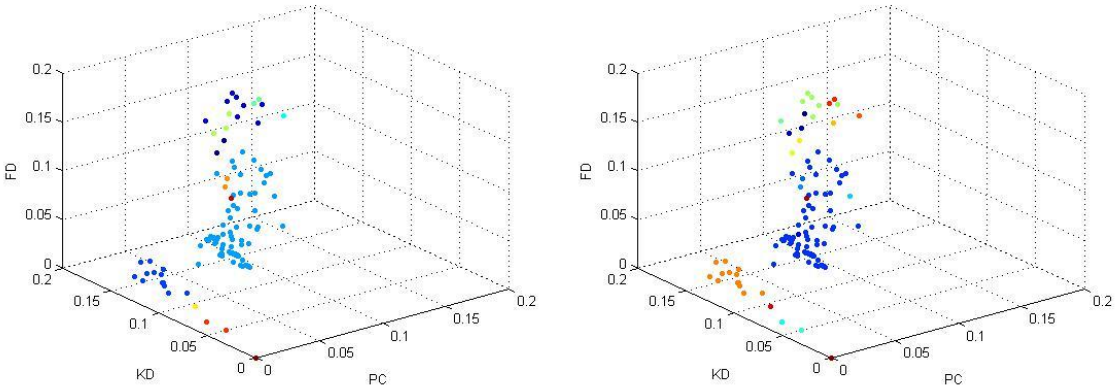


Figure-3.7: Results of clustering with 264 genes

This is better clustering with clustering limit 10 in Fig ‘a’ and 30 in Fig ‘b’ and threshold on dissimilarity distance is “0.2” and genes are “264”. We further reduced the threshold value to 0.18 and got a set of 182 genes and again applied clustering with clustering limits at (12, 20) as shown in Fig 3.8



(a) **(b)**

Figure3.8: Results of clustering with 182 genes

This is much better clustering with limit 12 in Fig ‘a’ and 20 in Fig ‘b’ and threshold is “0.18” while genes are 182. By reducing the threshold to 0.15, 83 genes are left and the following results are obtained with clustering limit set at ‘5’.

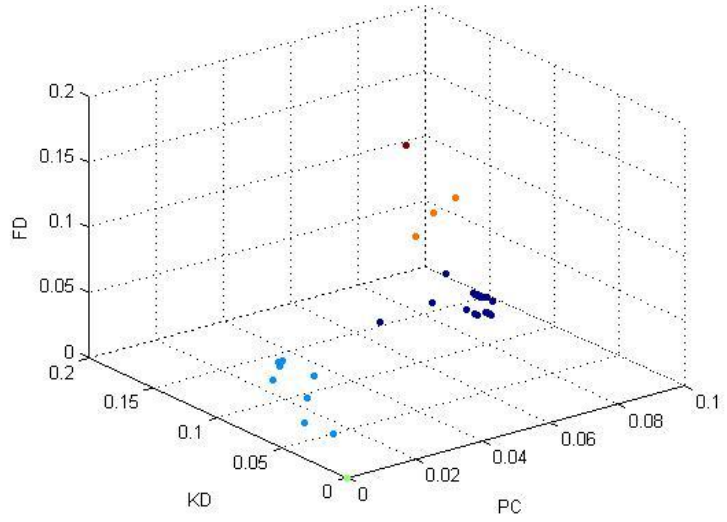


Figure-3.9: Results of clustering with 82 genes

Figure-3.9 is cluster representation with clustering limit of "5", threshold is "0.15" and genes are "82". The threshold limit is next reduced to "0.14" and 21 genes were left. We chose threshold "0.14" because decreasing the threshold further reduces the cluster quality and increasing the threshold increases the number of genes and cluster quality also decreases. Subsequently hierarchical clustering was performed with different linkages methods and clustering limits. The clustering results for average linkage are shown in Fig. 3.10 and dendrogram results are shown in Fig. 3.11. Figure-3.10 is cluster representation with clustering limit "4", threshold is "0.14" and genes are "21"

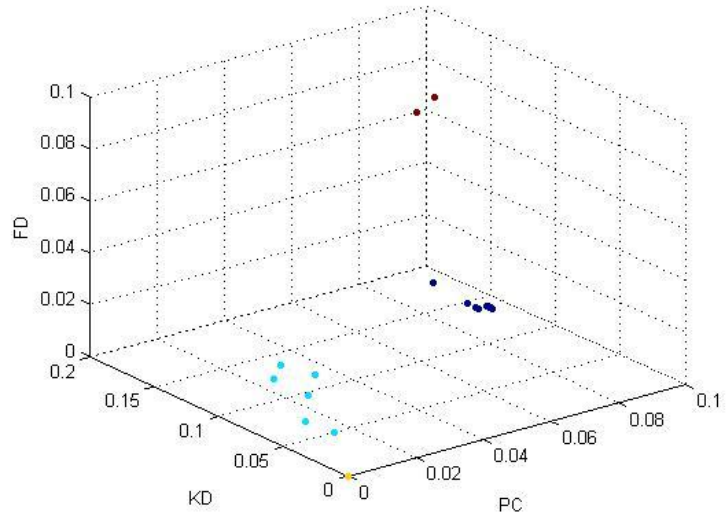


Figure-3.10: Results of clustering with 21 genes

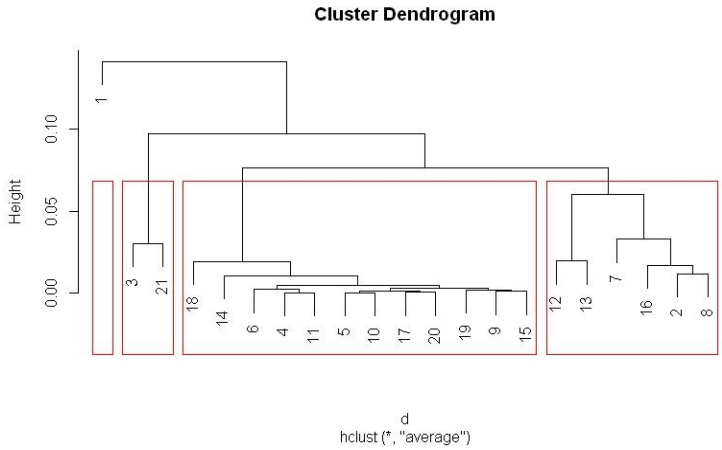


Figure-3.11: Dendrogram showing 4 clusters with CCL1 depicting as 1

Our study has explored different linkage parameters and results are summarized in Table -3.3. It is obvious from the Table-3.3 that with all linkage types, same threshold (0.14) and with clustering limit 3, 4 and 5 we obtained 4 consistent clusters. These clusters are compact containing genes that are strong correlated and co-regulate with each other.

Sr. No	Linkage Type	Dissimilarity Threshold	Clustering limit	Clusters obtained
1	Single	0.14	3,4,5	4
2	Complete	0.14	3,4,5	4
3	Ward	0.14	3,4,5	4
4	Median	0.14	3,4,5	4
5	Centroid	0.14	3,4,5	4

Table-3.3: summary of Linkage parameters

By using *single* as linkage parameter with 21 genes, 0.14' threshold and maximum clustering limit '4' obtained results and dendrogram are shown in Fig 3.12 and Fig 3.13.

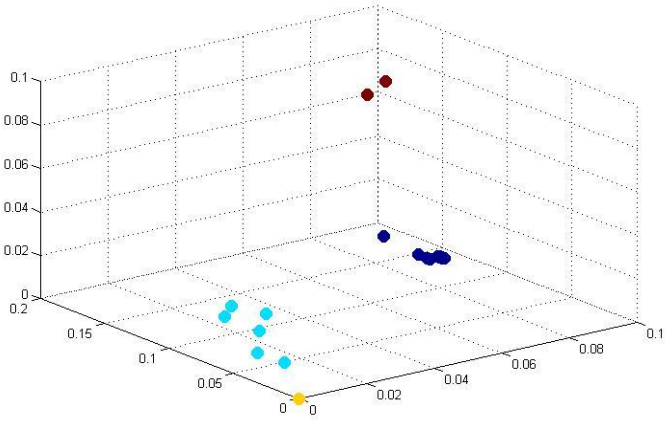


Figure-3.12: Results of clustering using single linkage

Figure-3.12 shows three main clusters other than CCL1, each cluster has different level of similarity to CCL1.

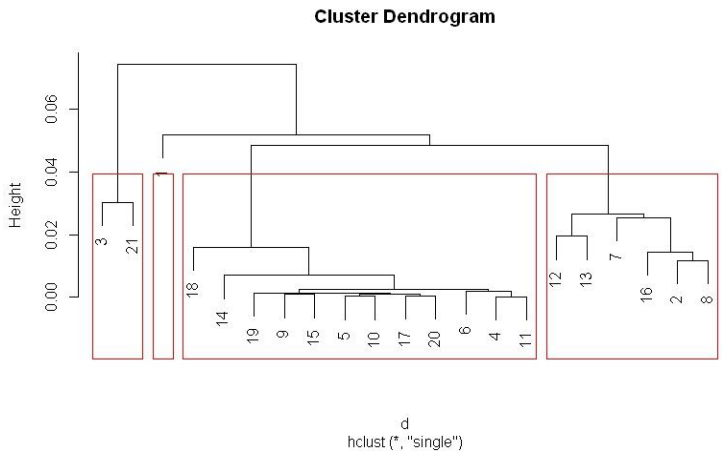


Figure-3.13: Dendrogram using single linkage with 4 clustering limit

Figure-3.13 shows CCL1 (gene 1) is clustered with other 2 clusters consisting of 12 and 6 genes respectively, but overall tree is cut down into 4 clusters. By using *complete* linkage, 21 gene dataset, 0.14 threshold and maximum clustering limit set at '4.' Results of clustering and dendrogram are shown in Fig 3.14 and Fig 3.15.

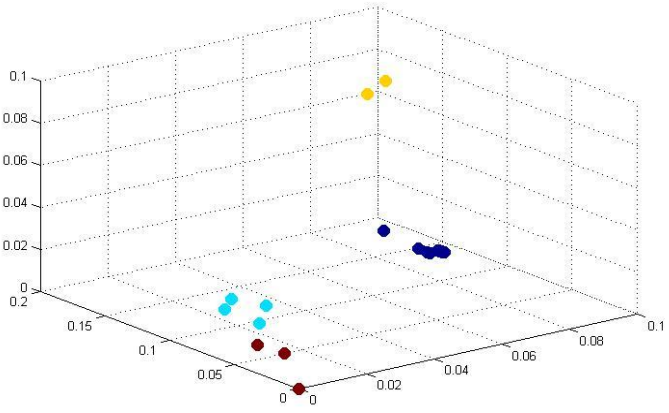


Figure-3.14: Results of clustering using complete linkage

Figure-3.15 shows CCL1 is clustered with 2 other genes to form one cluster and the remaining 18 genes are grouped in to 3 other clusters.

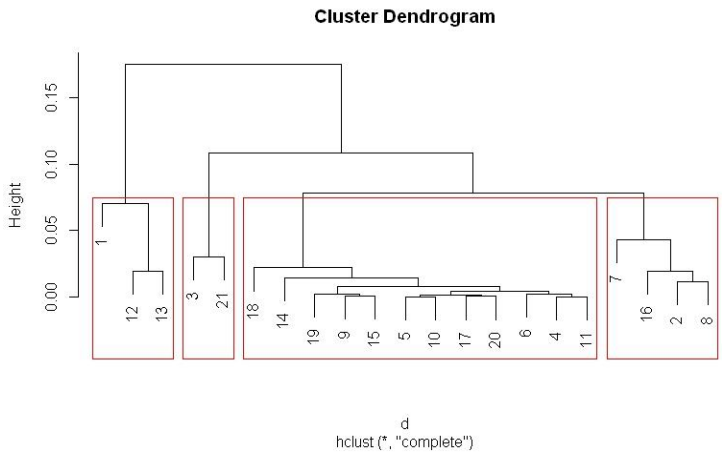


Figure-3.15: Dendrogram using complete linkage'

Figure-3.16 shows gene CCl1 is clustered with gene 12 and 13 to form one cluster and rest of the 18 genes is clustered in three different clusters. Overall tree is cut down into 4 clusters. By using *ward linkage*, 21 gene dataset , 0.14 threshold and maximum clustering limit '4'we got results as shown in Fig-3.16and Fig-3.17

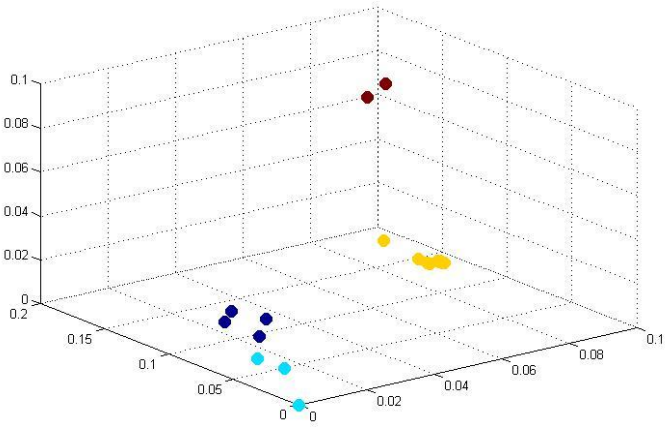


Figure-3.16: Results of clustering using linkage ward

.

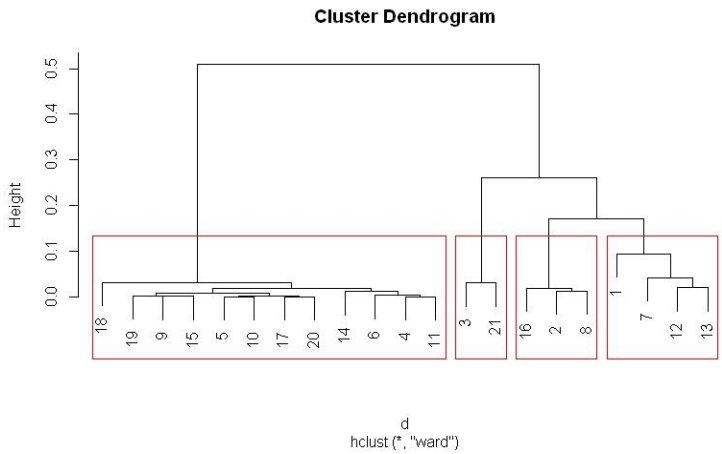


Figure-3.17: Dendrogram with clustering limit '4'

Figure-3.17 shows gene CCl1 is clustered with gene 7, 12 and 13 to form one cluster and rests of the 17 genes are clustered in three different clusters. Overall tree is cut down into 4 clusters. By using *Median linkage*, 21 gene dataset, 0.14 threshold and maximum clustering limit '4', obtained clustering results and dendrogram are shown in Figure 3.18, 3.19

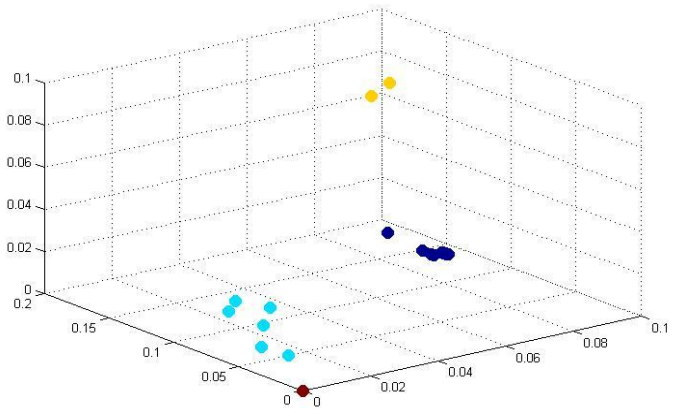


Figure-3.18: Results of clustering using Median linkage

Figure-3.18 shows three main clusters other than CCL1, each cluster has different level of similarity to CCL1.

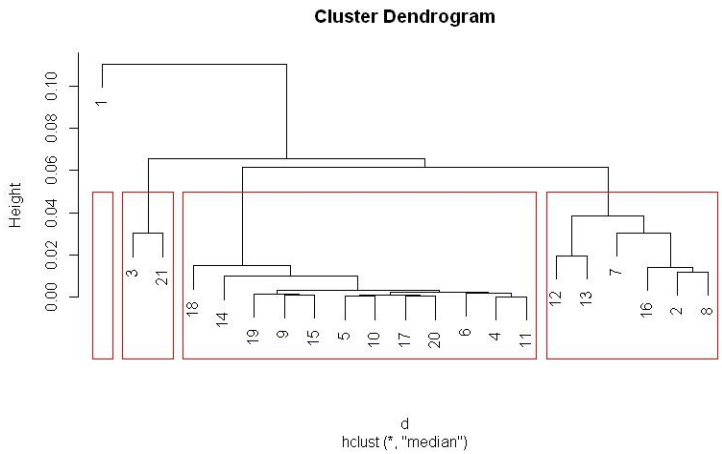


Figure-3.19: Dendrogram with clustering limit '4'

Figure-3.19 shows CCL1 (gene 1) is clustered with other 3 clusters consisting of 12 and 6 and 2 genes respectively, but overall three is cut down into 4 clusters. By using *centroid linkage*, 21 gene dataset, 0.14 thresholds and clustering limit '4' we got clustering results and dendrogram as shown in Fig 3.20, 3.21.

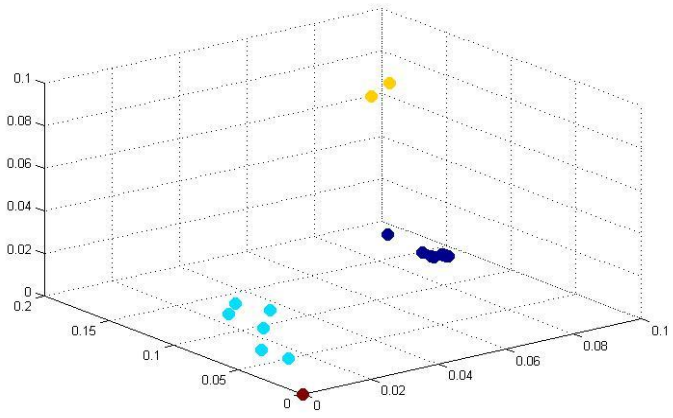


Figure-3.20: Results of clustering using linkage 'centroid'

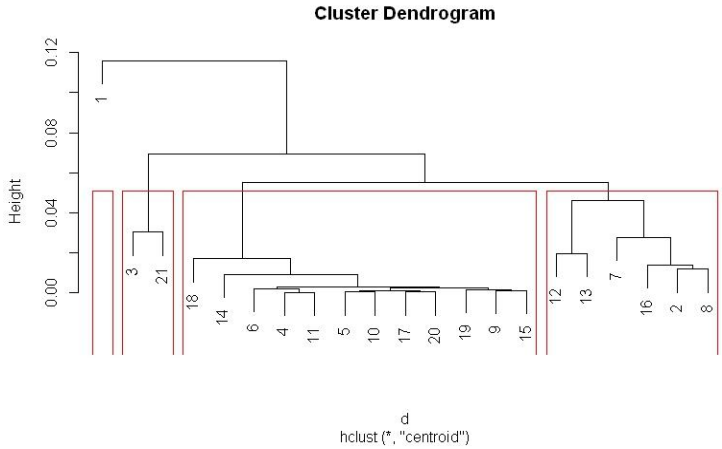


Figure3.21: Dendrogram with clustering limit '4'

Figure-3.21 shows CCL1 (gene 1) is clustered with other 2 clusters consisting of 12 and 6 genes respectively, but overall three is cut down into 4 clusters.

By comparing all the linkage results we can conclude that average, median and centroid linkage show similar results while complete, single and ward produce similar clustering results. The only difference between these two categories of linkage is that single, complete and ward linkage combine CCL1 with some other genes in the same cluster while average, centriod and median linkage place CCL1 in a separate cluster. These clusters show different level of similarity to CCL1 and would be discussed in detail in the next chapter.

Results and Discussion

Initially the data set was consisted of 16078 genes and 24 patients. After pre-processing, an outlier patient was removed. Next filtering was performed based upon dissimilarity threshold as described earlier and as shown in Table 4.1. Finally hierarchical clustering was performed with different linkage parameters such as single, complete, average, ward, median and centroid linkage. We also applied other clustering techniques like K- means and Density based clustering but hierarchical clustering provided the best results.

After getting different clusters for our dataset, we choose the results of average linkage for evaluation as it is most efficient and widely used technique as compared to other linkages [64]. We explored the gene indices in each cluster and also the actual genes. Table 4.1 represents gene indices in each cluster. There are 12 genes in the first cluster, 6 in second cluster and 2 genes in third cluster. The genes found in the three clusters are shown in Table 4.2.

Cluster1	Cluster2	Cluster3
4	2	3
5	7	21
6	8	
9	12	
10	13	
11	16	
14		
15		
17		
18		
19		
20		

Table-4.1: Gene index in three different clusters

Cluster1	Cluster2	Cluster3
NCRNA00244	C17orf72	DUSP22
CASQ1	OSGIN1	RDH14
LOC143286	MT1P2	
LOC283485	CACNA1G	
LOC257396	T || T, brachyury homolog	
ZNF155	ABCB5	
FRRS1		
ATP1B4		
C8orf42		
DYNC1LI2		
ASAP3		
LMOD3		

Table-4.2: Actual genes in three different clusters

Table-4.3 shows three clusters with similar genes clustered together. Cluster 1 contains 12 genes, cluster 2 contains 6 genes and cluster 3 contains 2 genes. These cluster show different level of similarity (co- expression and co-regulation) with CCL1. Table 4.3 is depicting actual genes which are grouped together in same cluster. 12 genes are clustered together in first cluster, 6 genes are in second cluster and 2 genes are in third cluster. We also plotted the actual expression level of genes for each cluster as shown in Fig. 4.1.

In Figure-4.1 we have plotted all patient phenotypes (L, M, P) categorized by stimulated and un-stimulated samples. On the x- axis units (1-4) represent Latent stimulated patients, (5-7) are Latent un-stimulated patients, (8-11) are Meningeal stimulated patients, (12-15) are Meningeal un-stimulated patients, (16-19) are Pulmonary stimulated patients and (20-23) are Pulmonary un-stimulated patients. Y-axis is showing the actual expression level of 21 genes as in the original data. On y-axis expression levels of genes has been plotted ranging from 0 to 2500. Expression level of 21 genes has been represented in the form of clusters. Genes belonging to each cluster are differentiated by line type and color. CCL1 is represented by blue diamond dashed line, cluster 1 (12 genes) is represented by green dashed line, cluster 2 (6 genes) is represented by red dashed line and cluster 3 (2 genes) is represented by magenta dashed line. Expression pattern of 21 genes are depicted in Figure-4.1 It is obvious from the Figure-4.1 that all the clusters are following the gene expression

pattern of CCL1. Cluster 3 (2 genes) is clearly representing expression pattern of its genes with minute variation but only few genes from cluster 1 (12 genes) and 2 (6 genes) are visible in this plot, following CCL1expression and rest of the gene expression patterns are hidden by the expression pattern of CCL1 showing no variation in expression pattern between these genes and CCL1.

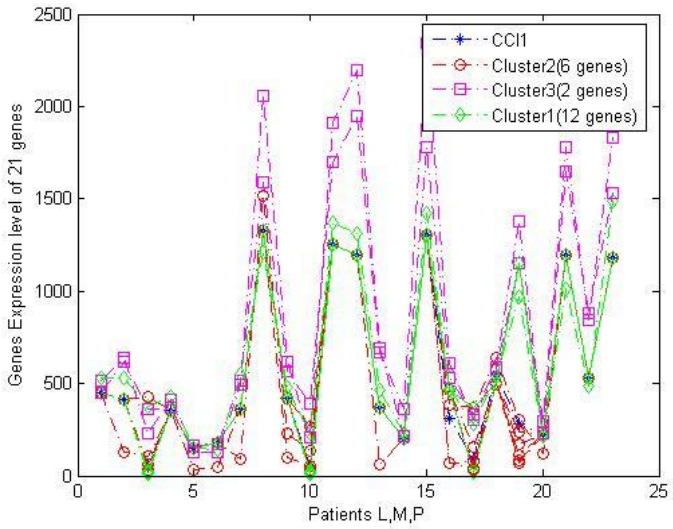
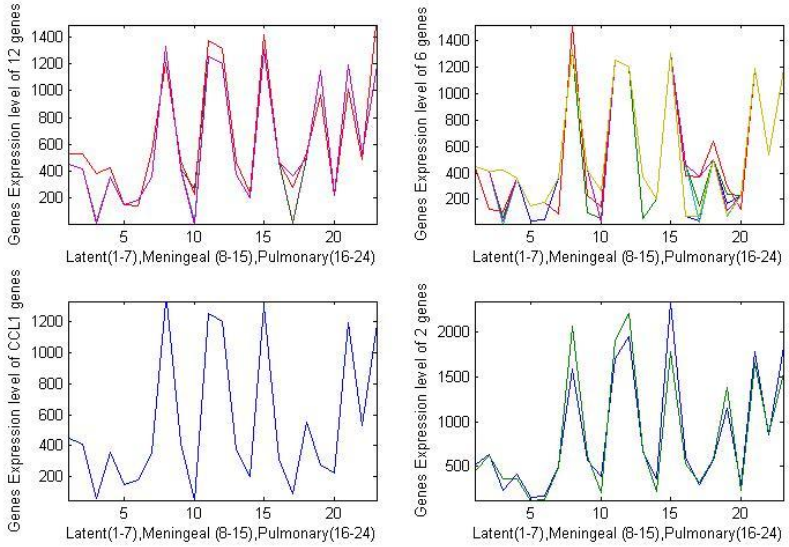


Figure-4.1: Line Plot of 3 clusters representing 21 genes in three phenotypes (L, M, P).

Expression trend of individual clusters are also represented in Figure-4.2. Figure-4.2 (c) shows expression trend of CCL1, Figure-4.2 (a) shows expression trend of genes forming cluster 1, Figure-4.2 (b) shows expression trend of genes forming cluster 2 and Figure-4.2 (d) shows expression trend of genes forming cluster 3. Each cluster genes are depicting the behavior pattern of CCL1. This study has also quantitatively proved the pattern similarity in Table 4.3.

(a) (b)



(c) (d)

Figure-4.2 (a, b and d) show gene expression level of cluster 1, 2, 3 and CCL1

We have also made plots for different phenotypes of tuberculosis categorized by stimulated and un-stimulated samples. In Figure-4.3 and 4.4 expression levels for 21 genes including CCL1 (three clusters) are shown in Meningeal stimulated and un-stimulated samples respectively. These plots do not show much variation between stimulated and un-stimulated samples as both plots represent expression pattern ranging from 0 to 2500 which is also consistent with Nguyen's study [3].

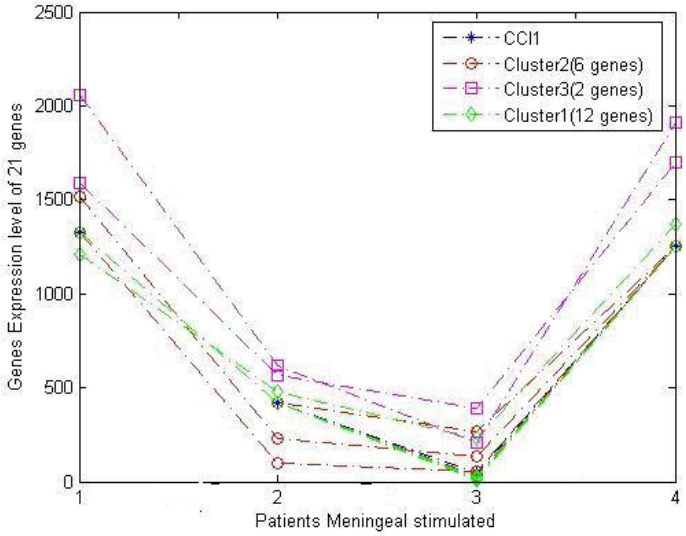


Figure-4.3: Line plot of expression levels of 21 genes in Meningeal stimulated TB

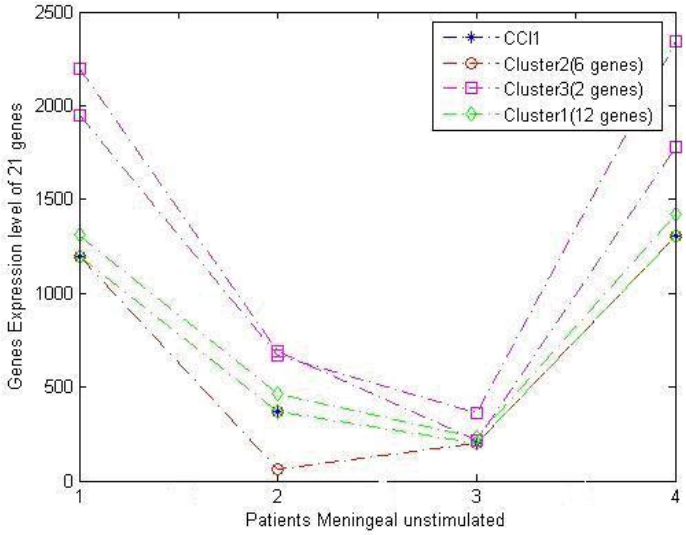


Figure-4.4: Line plot of expression level of 21 genes in Meningeal un-stimulated TB

Figure-4.5 and Figure-4.6 shows the expression levels for 21 genes in Pulmonary stimulated and un-stimulated samples respectively. These plots show great variation in the expression levels of the genes. In Figure-4.5 expression trend of Pulmonary stimulated patients ranges from 200 to 1400 and in Figure-4.6 expression trend of Pulmonary un-stimulated patient is ranging from 200 to 2000 which is also consistent with previous studies [3].

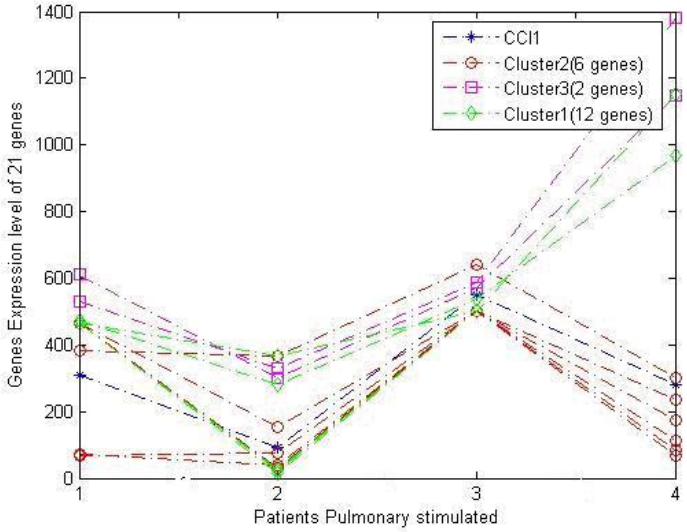


Figure-4.5: Line plot of expression level of 21 genes in Pulmonary stimulated TB

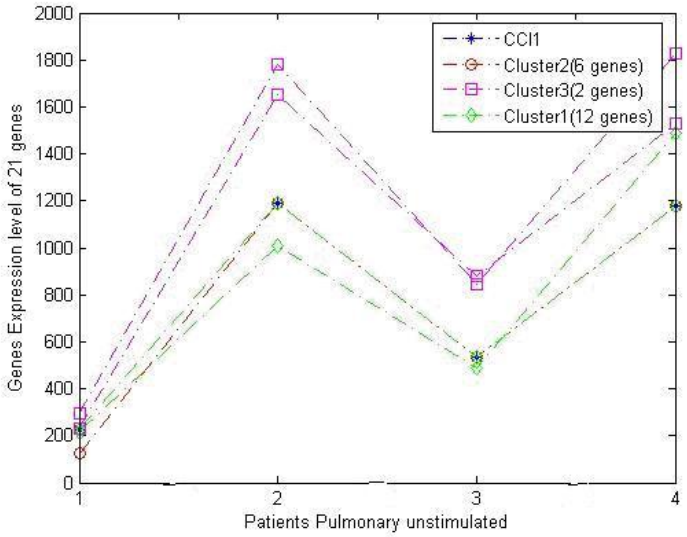
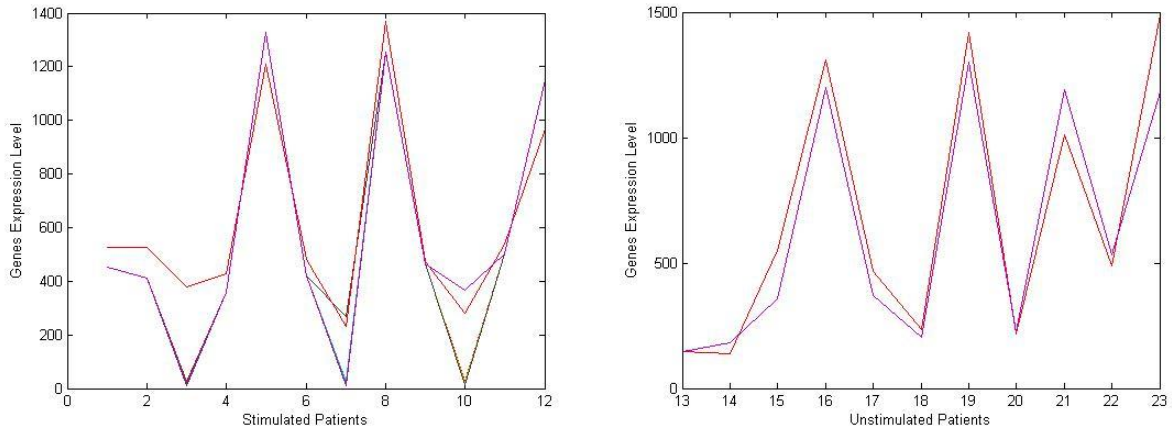


Figure-4.6: Line plot of expression level of 21 genes in Pulmonary un-stimulated TB

We have also made plots of stimulated and un-stimulated samples of 21 genes separately which are shown in Fig-4.7. These plots show that there is a great deal of variations in expression levels of these 21 genes in stimulated and un-stimulated samples. Latent stimulated samples consist of patients 1-4 in Fig-4.7 (a) and Latent un-stimulated samples consist of patients 13-15 in Fig-4.7 (b). Pulmonary stimulated and un-stimulated samples consisting of patients 9-12 in Fig-4.7 (a) and 20-23 in Fig-4.7 (b) respectively, but cannot differentiate between Meningeal stimulated and un-stimulated samples consisting of patients 5-8 and 16-19 in Fig-4.7 (a) and (b) respectively, These low and high

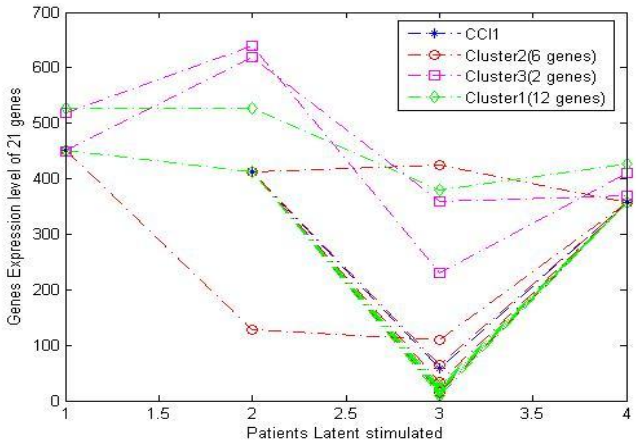
expression level of genes in stimulated and un-stimulated samples is also consistent with previous study [3].



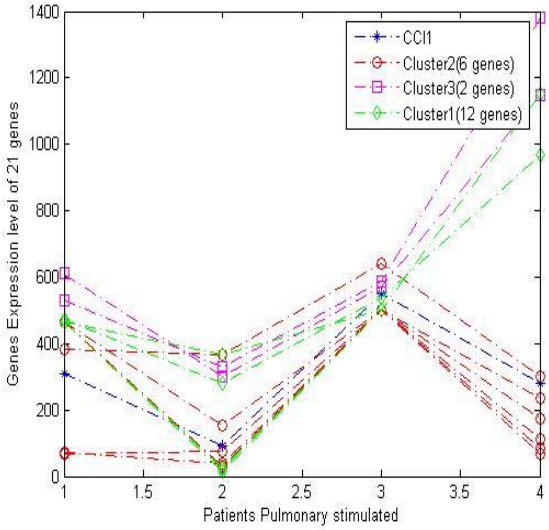
(a) **(b)**

Figure-4.7: Stimulated and un-stimulated samples plots

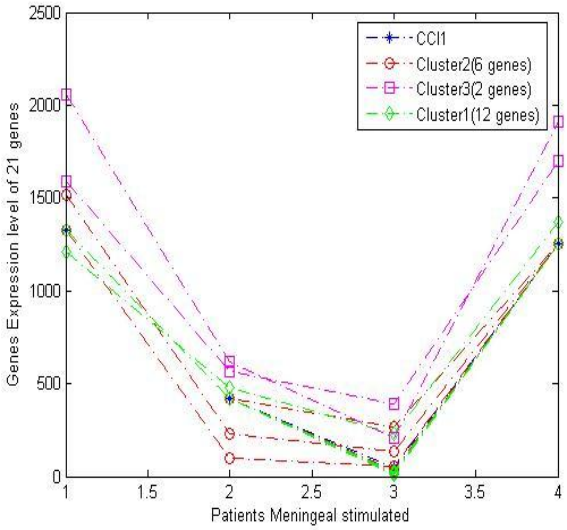
This study has compared plots of expression levels of three different types of TB i-e Latent, Pulmonary and Meningeal to discover how expression levels of the 21 genes are found to vary with respect to each other in the three different forms of TB. Figure-4.8 (a) shows that expression trend of Latent stimulated TB ranging from 0 to 700. Figure-4.8 (b) shows that expression trend of Pulmonary stimulated TB ranging from 0-1400 and Figure-4.8 (c) shows that expression pattern of Meningeal stimulated TB ranges from 0-2500. These plots also represent that these 21 genes are over expressed two times more in Pulmonary stimulated TB as compared to Latent TB and regulate three times more in Meningeal TB as compared to Latent TB and two times more as compared to Pulmonary TB. Hence, we conclude that these 21 genes fairly distinguish three different phenotypes of TB by variation in their gene regulation expression pattern (over-expression) which is consistent with the Nguyen et al. [3] study as well.



(a)



(b)



(c)

Figure-4.8: Stimulated samples of three phenotypes (M, P, L) of TB

Table 4.4 shows the average correlation of genes within each cluster and with CCL1 and proves that genes expression profiles are highly correlated with CCL1.

Sr. no	No of genes	Correlation with each other	Correlation with CCL1
All genes (without clustering)	21	0.93	0.93
Cluster1	12	0.955	0.953
Cluster2	6	0.953	0.953
Cluster3	2	0.953	0.953

Table-4.3: Result of correlation coefficient of genes

In Table-4.4, first row shows the Pearson co-relation coefficient of all genes among themselves and also with gene CCL1. Second, third and fourth row shows the Pearson co-relation coefficient between genes within the same cluster and also with gene CCL1. All Pearson co-relation coefficients prove that genes are highly co-regulated with each other and also with gene CCL1.

Our results show that using the three dissimilarity metrics proposed we could reduce the data-sets and subsequently use various clustering techniques to cluster similar genes. The distance metrics also provide a means by which filtering of the genes may be attained as shown in Table 2. We reduce the number of genes from a total of 16078 to 21 genes (found to be most similar to CCl1). The threshold is arbitrary and may be used to further reduce the number of genes. We use an optimal threshold of 0.14 as it provides a feasible number of genes to cluster and analyze. Hierarchical clustering results are depicted in Figures 3 and 4. As seen in the Table 2, three different clusters of genes most similar to CCl1 are found with each cluster having its own significance. We have explored the literature on some of the genes shown in Table 2. The literature indicates that genes namely CASQ1, ATP1B4 and ASAP3 play a role in development of soft tissue and muscle tumors amongst other things. LOC283485 and LOC257396 play a role in the development of trachea and pharynx respectively. FRRS1 is considered to be responsible for bladder carcinoma; germ cell tumour; head and neck tumour. Genes DYNC1LI2, T || T, brachyury homolog, DUSP22, RDH14 and OSGIN1 are closely associated with lung functioning and lung tumor while CACNA1G deals with lung functioning and uterine tumor. Hence, most of the genes discussed above and discovered in our study play a role in the development of the lung tissue and functioning and also impact growth of lung tumor. Finally correlation values in Table 4 prove that the genes identified in this study are highly correlated with CCl1 (a gene which is already known to play a significant role in TB infection).

Chapter 5

Conclusion and Future Directions

In our thesis, we have investigated gene up regulation and down regulation to discover genes responsible for progression of TB. There are certain genes in every phenotype which co-regulate each other making up structural relationship. Our study has found 21 genes which are co-regulated with CCL1 and like CCL1 also influence the host susceptibility to TB and discriminate between different clinical types of TB. Studies such as these would facilitate further studies to understand the implication of gene-gene interactions for TB development and may lead to the introduction and discovery of new drugs and vaccines for TB. In this chapter, we summarize the whole research work carried out in this thesis. A discussion on possible future directions is also presented at the end of this chapter.

5.1 Conclusion

In this thesis, we have employed hierarchical clustering to find functionally related genes and propose the use of different distance metrics to analyze and compare gene expression. We have explored two different dissimilarity measures (Fisher distance and Kullback-leibler distance) in addition with Pearson correlation coefficient and used the dissimilarity estimates for hierarchical clustering. We identify 21 new genes which are highly correlated with CCl1. Most of the 21 genes discovered are found to play a role in lung functioning and development and some are also seen to be active in spread of certain tumors. Hence, some new co-regulated genes of CCl1 have been discovered which were previously unknown. By targeting such co-regulated genes of CCl1 there will be a hope in drug discovery process. Studies such as these would facilitate the introduction of new TB drugs and vaccines.

5.2 Future Directions

In this research, we have used hierarchical clustering and different dissimilarity measures to find co-regulated genes responsible for TB susceptibility. There are some other techniques like k-means and Dbscan which could be used employing different dissimilarity measures to discover further clusters.

Moreover, other genes responsible for causing TB could also be found and can help in differentiating between different forms of TB.

REFERENCES

[1] Kumar, Vinay; Abbas, Abul K.; Fausto, Nelson; & Mitchell, Richard N. (2007). Robbins Basic Pathology (8th ed.). Saunders Elsevier. pp. 516–522. ISBN 978-1-4160-2973-1

[2] http://www.tbdb.org

[3] Nguyen Thuy Thuong Thuong, Sarah J. Dunstan,Tran Thi Hong Chau, Vesteinn Thorsson, Cameron P. Simmons, Nguyen Than Ha Quyen, Guy E. Thwaites,Nguyen Thi Ngoc Lan, Martin Hibberd, Yik Y. Teo, Mark Seielstad, Alan Aderem, Jeremy J. Farrar, Thomas R. Hawn, .(2008) Identification of Tuberculosis Susceptibility Genes with Human Macrophage Gene Expression Profiles. PLoS Pathog 4(12): e1000229. doi:10.1371/journal.ppat.1000229

[4] Take-Home Final Exam: Mining Regulatory Modules from Gene Expression Data 36-350, Data Mining; Fall 2009_ http://www.stat.cmu.edu/~cshalizi/350/exams/final/final.pdf

[5] Tiffany HL, Lautens LL, Gao JL, Pease J, Locati M, et al. (1997) Identification of CCR8:a human monocyte and thymus receptor for the CC chemokine I-309. J Exp Med 186: 165-170

[6] Almut Schulze1 & Julian Downward. Navigating gene expression using microarrays — a technology review. Nature Cell Biology 3, E190 - E195 (2001) .doi:10.1038/35087138

[7] WHO Tuberculosis Factsheet, World Health Organization, March 2010 http://www.who.int/mediacentre/factsheets/fs104/en/

[8] Jasmer RM, Nahid P, Hopewell PC (December 2002). "Clinical practice. Latent tuberculosis infection". N. Engl. J. Med. 347 (23): 1860- 6. doi:10.1056/NEJMcp021045.PMID 12466511

[9] Tuberculosis http://www.pbs.org/wgbh/rxforsurvival/series/diseases/tuberculosis.html

[10] Tuberculosis Fact sheet No104 March 2012 http://www.who.int/mediacentre/factsheets/fs104/en/index.html

[11] Bayesian Networks for Analysing Gene Expression Data Dirk Husmeier Biomathematics & Statistics Scotland SCRI, Invergowrie, Dundee, DD2 5DA United Kingdom August 2001 http://www.bioss.ac.uk/~dirk/essays/GeneExpression/bayes_net.html

[12] Gerard J. Human macrophage activation programs induced by bacterial pathogens. Whitehead Institute, 9 Cambridge Center, Cambridge, MA 02142; Infectious Disease Unit, Massachusetts General Hospital, Boston, MA 02114; December 5, 2001 (received for review October 2, 2001).

[13] Damien Chaussabel, Roshanak Tolouei Semnani, Mary Ann McDowell, David Sacks, Alan Sher, and Thomas B. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites Published online before print March 27, 2003, doi: 10.1182/blood-2002-10-3232

[14] Silvia ragno. Changes in gene expression in macrophages infected with Mycobacterium tuberculosis: a combined transcriptomic and proteomic approach. Immunology 2001 104 99±108.
[15] Staunton, J.E. et al. Chemosensitivity prediction by transcriptional profiling.proc. Natl Acad. Sci. USA 98, 10787–10792 (2001).
[16] Brightbill HD, Libraty DH, Krutzik SR et al. Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. Science 1999; 285:732±6.
[17] Means TK, Wang S, Lien E, Yoshimura A, Golenbock DT, Fenton MJ. Human toll-like receptors mediate cellular activation by Mycobacterium tuberculosis. J Immunol 1999; 163:3920±7.
[18] Hirsch CS, Ellner JJ, Russell DG, Rich EA. Complement receptor mediated uptake and tumor necrosis factor-alpha-mediated growth inhibition of Mycobacterium tuberculosis by human alveolar macrophages. J Immunol 1994; 152:743±53.
[19] Schlesinger LS. Macrophage phagocytosis of virulent but not attenuated strains of Mycobacterium tuberculosis is mediated by mannose receptors in addition to complement receptors. J Immunol 1993; 150:2920±30.
[20] Stokes RW, Haidi ID, Jeffries WA, Speert DP. Mycobacteria macrophage interactions: macrophage phenotype determines the nonopsonic binding of Mycobacterium tuberculosis to murine macrophages. J Immunol 1993; 151:7067±76.
[21] Zimmerli S, Edwards S, Ernst JD. Selective receptor blockade during phagocytosis does not alter the survival and growth of Mycobacterium tuberculosis in human macrophages. Am J Respir . Cell Mol Biol 1996; 15:760±70.
[22] Fratazzi C, Manjunath N, Arbeit RD, Carini C, Gerken TA, Ardman B, Eremold-O'Donnell E, Remold HG. A macrophage invasion mechanism for mycobacteria implicating the extracellular. domain of CD43. J Exp Med 2000; 192:183±92.
[23] Magram, J., Connaughton, S. E., Warrier, R. R., Carvajal, D. M., Wu, C. Y., Ferrante, J., Stewart, C., Sarmiento, U., Faherty, D. A. & Gately, M. K. Transcription Factors: Normal and Malignant Development of Blood Cells (1996) Immunity 4, 471–481.
[24] IL-12 increases resistance of BALB/c mice to Mycobacterium tuberculosis infection Flynn, J. L., Goldstein, M. M., Triebold, K. J., Sypek, J., Wolf, S.&Bloom, B. R. (1995) J. Immunol. 155, 2515–2524

[25] Interleukin 12 (IL-12) is crucial to the development of protective immunity in mice intravenously infected with Mycobacterium tuberculosis Cooper, A. M., Magram, J., Ferrante, J. & Orme, I. M. (1997) J. Exp. Med. 186, 39–45.

[26] Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency Altare, F., Durandy, A., Lammas, D., Emile, J. F., Lamhamedi, S., Le Deist, F., Drysdale, P., Jouanguy, E., Doffinger, R., Bernaudin, F., et al. (1998) Science 280, 1432–1435.

[27] IL-12 receptor deficiency revisited: IL-23-mediated signaling is also impaired in human genetic IL-12 receptor β1 deficiency de Jong, R., Altare, F., Haagen, I. A., Elferink, D. G., Boer, T., van Breda Vriesman, P. J., Kabel, P. J., Draaisma, J. M., van Dissel, J. T., Kroon, F. P., et al. (1998) Science 280, 1435–1438.

[28] O. M. Rivero-Lezcano. CCL20 is over expressed in Mycobacterium tuberculosis-infected monocytes and inhibits the production of reactive oxygen species (ROS) cei_4168 289..29

[29] Zahra Hasan Z, Cliff JM, Dockrell HM, Jamil B, Irfan M, et al. (2009) CCL2 Responses to Mycobacterium tuberculosis Are Associated with Disease Severity in Tuberculosis. PLoS ONE 4(12): e8459. doi:10.1371/journal.pone.0008459

[30] Michael B. Eisen. Genetics cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA Vol. 95, pp. 14863–14868, December 1998

[31] Alvis Brazma , Jaak Vilo . Functional genomics gene expression data analysis. Volume 480, Issue 1, 25 August 2000,Pages 17-24

[32] M. Eisen, P.T. Spellman, D. Botstein and P.O. Brown. Proc. Natl. Acad. Sci. Cluster analysis and display of genome-wide expression patterns. USA 95 (1998), pp. 14863–14867.

[33] John Wiley and Sons, Clustering Algorithms, New York. Hartigan, J.A. (1975)

[34] S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church. Systematic determination of genetic network architecture. Nature Genet. 22 (1999), pp. 281–285.

[35] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander and T. Golub. Using GenePattern for Gene Expression Analysis. Proc. Natl. Acad. Sci. USA 96 (1999), pp. 2907–2912.

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

[37] J.L. DeRisi, V.R. Iyer and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278 (1997), pp. 680–686.

[38].S. Chu, J.L. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown and I. Herskowitz. Using DNA microarrays to study host-microbe interactions Science 282 (1998), pp. 699–705.

[39] P.T. Spellman, G. Sherlock, M. Zhang, V.R. Iyer, K. Anders, M. Eisen, P.O. Brown, D.Botstein and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9 (1998), p. 3273.

[40] Holstege, F.C.P., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95, 717-728

[41] C. Lee, R.G. Klopp, R. Weindruch and T.A. Prolla. Gene expression profile of aging and its retardation by caloric restriction. Science 285 (1999), pp. 1390–1393.

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

[43] DeRisi, J., Iyer, V. & Brown, P. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale (1997) Science 680–686.

[44]. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. & Herskowitz, I. The transcriptional program of sporulation in budding yeast. (1998) Science 282, 699–705.

[45] Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J., Trent, J. M., Staudt, L. M., Hudson, J., Boguski, M. S. The Transcriptional Program in the Response of Human Fibroblasts to Serum (1999) Science 283, 83–87.

[46] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, And David Botstein. Cluster analysis and display of genome-wide expression patterns Proc. Natl. Acad. Sci. USA 95, 14863–14868.

[47] Wen, X., Fuhrman, S., Michaels, G., Carr, D., Smith, S., Barker, J. & Somogyi, R. Large-scale temporal gene expression mapping of central nervous system development (1998) Proc. Natl. Acad. Sci. USA 95, 334–339.

[48] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Interpreting patterns of gene expression with self-organizing maps: Proc. Natl. Acad. Sci. USA 95, 14863–14868

[49] Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA

Vol. 96, pp. 6745–6750, June 1999

[50] P. Roy Walkera, Brandon Smitha, Data mining of gene expression changes in Alzheimer brain .2004 Published by Elsevier B.V. doi:10.1016/j.artmed.2004.01.008

[51] Sandip Ray1,16, Markus Britschgi2,16, Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins published online 14 October 2007; doi:10.1038/nm1653

[52] Walter J. Lukiw. Gene Expression Profiling in Fetal, Aged, and Alzheimer Hippocampus: A Continuum of Stress-Related Signaling 0364-3190/04/0600–1287/0 © 2004 Plenum Publishing Corporation

[53] The Differential Gene Expression Pattern of Mycobacterium tuberculosis in Response to Capreomycin and PA-824 versus First-Line TB Drugs Reveals Stress- and PE/PPE-Related Drug Targets.14 nternational Journal of Microbiology Volume 2009 (2009), Article ID 879621, 9 pages doi:10.1155/2009/879621

[54] Asia News. http://www.asianews.it/news-en/more-than-60,000-tb-deaths-a-year-5737.html

[55] Tuberculosis control in Pakistan rational, objectives & study design. pub.uni-bielefeld.de/download/2305289/2305295

[56] http://www.nature.com/icb/journal/v78/n4/full/icb200042a.htm

[57] Northern Blot http://www.escience.ws/b572/L13/north.html

[58] Audrey P Gasch and Michael B Eisen. Exploring the conditional co regulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology* 2002, 3(11):research0059.1–0059.22

[59] Ka Yee Yeung , Walter L. Ruzzo Model-Based Clustering and Data Transformations for Gene Expression Data. Revised May 16, 2001. 70 vol 17 no. 9 2001

[60] Fisher Discriminant .http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_fisher_discriminant.htm

[61] Kullback Leibler http://www.csse.monash.edu.au/~lloyd/tildeMML/KL/

[62]Pearson Correlation http:// www.btluke.com/pearson.html

[63] http:// http://www.phgfoundation.org/tutorials/variantsDisease

[64] Information retrieval :data structure and algorithms by William B Frakes.