Reconstruction of Gene Regulatory Network from Expression Profile of Plasma RNA Data of Colorectal Cancer Patients using Soft Computing Techniques

Thesis Submitted by

VINEETHA S

In partial fulfillment of the requirements

for the award of the degree of

DOCTOR OF PHILOSOPHY

UNDER THE FACULY OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

KOCHI-682 022

INDIA

October 2012

Certificate

*This is to certify that the thesis entitled "**Reconstruction of Gene Regulatory Network from Expression Profile of Plasma RNA Data of Colorectal Cancer Patients using Soft Computing Techniques**" is a bonafide record of the research work carried out by Ms. Vineetha S in the Department of Computer Science, Cochin University of Science and Technology, Kochi-682022, under my supervision and guidance.*

Kochi-682022
October, 2012

Dr. Sumam Mary Idicula
Supervising Guide, Professor,
Department of Computer Science
Cochin University of Science and Technology,
Kochi-682022, Kerala

DECLARATION

*I hereby declare that the work presented in this thesis entitled "**Reconstruction of Gene Regulatory Network from Expression Profile of Plasma RNA Data of Colorectal Cancer Patients using Soft Computing Techniques**" is based on the original research work carried out by me in the Department of Computer Science, Cochin University of Science and Technology, Kochi-682022, under the supervision and guidance of Dr. Sumam Mary Idicula, Professor, Department of Computer Science , Cochin University of Science and Technology , Kochi-682022.The results presented in this thesis or parts of it have not been presented for the award of any other degree.*

Kochi-682022
October 2012

Vineetha S
Research Scholar


Dedicated to My Father
Shri P. R. Prabhakaran

Acknowledgements

The work on this thesis would not have been possible without the encouragement and support of many people who in one way or another have contributed in the completion of this study. It is a pleasure to convey my gratitude to all of them in my humble acknowledgment.

First and foremost, I express my utmost and profound gratitude to my supervising guide Dr. Sumam Mary Idicula, Professor, Department of Computer Science, Cochin University of Science and Technology for her valuable guidance, encouragement and support throughout the course of work. I am grateful for her constructive comments and careful evaluation of my thesis.

I express my deep gratitude and respect towards Dr. C ChandrashekaraBhat, Scientist, National Institute for Interdisciplinary Science and Technology,Thiruvananthapuram, who inspired me in doing research in the extremely dynamic field of bioinformatics. I am extremely grateful to him, for his timely guidance and unflinching support that has helped me overcome the obstacles encountered during my research work. His dedication and sincerity in reviewing the research papers and the thesis are greatly appreciated.

It is my privilege to place on record my gratitude to Dr. K Poulose Jacob, Professor and Head, Department of Computer Science, Cochin University of Science and Technology, for providing me necessary facilities in the Institute for the research work. My sincere gratitude to Dr. David Peter S and Dr. Sheena Mathew of the institute for being constant source of encouragement for me.I

also thank all the administrative and supporting staff of the Institute for their support and help.

A great deal of thanks goes to all my friends and colleagues of Rajiv Gandhi Institute of Technology, Kottayam, for all the timely help during the entire course of my work and thesis submission.

It would not have been possible to undertake the journey of my career, and to reach where I am today without the support of my family members, especially my parents. I thank them for the unconditional love and care they showered on me. Their faith on me has always been a great strength which helped me throughout the way.

I am thankful to my husband Mr. Sojan V. J, for his persistent support, affection and care throughout this journey.

Above all, I express my gratefulness to the Almighty for making me able to achieve whatever I have.

Vineetha S.

ABSTRACT

Microarray data analysis is one of data mining tool which is used to extract meaningful information hidden in biological data. One of the major focuses on microarray data analysis is the reconstruction of gene regulatory network that may be used to provide a broader understanding on the functioning of complex cellular systems. Since cancer is a genetic disease arising from the abnormal gene function, the identification of cancerous genes and the regulatory pathways they control will provide a better platform for understanding the tumor formation and development. The major focus of this thesis is to understand the regulation of genes responsible for the development of cancer, particularly colorectal cancer by analyzing the microarray expression data.

In this thesis, four computational algorithms namely fuzzy logic algorithm, modified genetic algorithm, dynamic neural fuzzy network and Takagi Sugeno Kang-type recurrent neural fuzzy network are used to extract cancer specific gene regulatory network from plasma RNA dataset of colorectal cancer patients. Plasma RNA is highly attractive for cancer analysis since it requires a collection of small amount of blood and it can be obtained at any time in repetitive fashion allowing the analysis of disease progression and treatment response.

The approaches proposed in this study extend the previous state of the art by incorporating clustering and some statistical techniques to reduce the computational complexity and processing time. The unpaired t-test has been employed to identify the genes that are differentially expressed in cancerous tissues and normal ones. The fuzzy logic algorithm models gene

regulatory network by extracting regulatory information in the form of fuzzy rules. The second approach, modified genetic algorithm, applies genetic algorithm to model regulatory network by searching for an optimum weight matrix. The reverse engineering algorithms based on neural fuzzy network exploits the advantage of neural networks, in terms of low-level learning and computational power, and those of fuzzy system, in terms of the high–level human like reasoning and results interpretability. Unlike other neural fuzzy architectures, both dynamic neural fuzzy network and TSK-type recurrent neural fuzzy network has no predefined rules. The algorithms automatically produce an adaptive number of fuzzy rules that describe the relationship between regulating genes and regulated genes. The feedback structure of TSK-type recurrent neural fuzzy network stores the prior system states that increases the learning ability of the algorithm.

The algorithms captured regulatory relationship among 27 differentially expressed genes from plasma RNA of colorectal cancer patients. Detailed pathway analysis shows that most of these genes are actively involved in the cancer related canonical pathways. The work in this thesis resulted in two interesting findings. First, upregulated genes are regulated by more genes than downregulated genes. Second, tumor activators suppress tumor suppressors strongly in the disease environment. The high degree of centrality of upregulated genes in the regulatory network indicates their key roles in cancer specific gene regulatory network.

The results of the above machine learning algorithms using the microarray dataset are compared. The regulatory relations extracted by these computational approaches are validated by comparing it with the regulatory relations extracted from the microarray dataset of colon tumor samples. It was found that TSK-type recurrent neural fuzzy network identified more gene interactions and gave better recall than other approaches. The computational efficiency of these approaches is tested using the benchmark dataset of Saccharomyces Cervisiae. The results demonstrate the effectiveness of these algorithms in retrieving biologically valid regulatory relations. It was found that 87.8% of the total interactions extracted by TSK-type recurrent neural fuzzy network are correct in accordance with the biologically proven regulatory interactions outperforming other computational approaches.

Finally, TSK-type recurrent neural fuzzy network is applied to extract microRNA-mRNA association network from paired microRNA, mRNA expression profiles of colorectal cancer patients. The algorithm achieved good performance in identifying experimentally known colorectal cancer related microRNAs and their target genes. Targeting such microRNAs may help in the early detection, prognosis and future therapy of colorectal cancer.

Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
ANOVA	-	Analysis of Variance
APC	-	Adenomatous Polyposis Coli
BINGO	-	Biological Networks Gene Ontology Tool
cDNA	-	Complementary Deoxyribo Nucleic Acid
CRC	-	Colorectal Cancer
DCC	-	Deleted in Colon Cancer
DNA	-	Deoxyribonucleic Acid
DNFN	-	Dynamic Neural Fuzzy Network
FCM	-	Fuzzy C Means
FDA	-	Food and Drug Administration
FDR	-	False Discovery Rate
FOM	-	Figure of Merit
GA	-	Genetic Algorithm
GO	-	Gene Ontology
GRN	-	Gene Regulatory Network
HAC	-	Hierarchical Agglomerative Clustering
HNPCC	-	Hereditary Non-polyposis Colorectal Cancer
miRNA	-	Micro RNA
miRO	-	Micro RNA Ontology Database
MMR	-	Mismatch-Repair Genes
mRNA	-	Messenger RNA
MSE	-	Mean Square Error

PCA	-	Principal Component Analysis
PCR	-	Polymerase Chain Reaction
RISC	-	RNA-Induced Silencing Complex
RNA	-	Ribonucleic Acid
RNN	-	Recurrent Neural Network
rRNA	-	Ribosomal RNA
SOMs	-	Self Organizing Maps
SVM	-	Support Vector Machines.
tRNA	-	Transfer RNA
TRNFN	-	Takagi Sugeno Kang-type Recurrent Neural Fuzzy Network

Chapter 1
Introduction

Contents

Microarray is a powerful technology capable of providing simultaneous measurement of expression levels of thousands of genes which can accurately represent the state of a biological cell or tissue of interest. Analysis of expression profiles from microarray experiments for understanding fundamental cellular processes represent a challenging task for bioinformatics. One of the main focuses on microarray data analysis is to unravel the interactions among various types of molecules (proteins, RNA, metabolites, etc.), by data-mining and integrating high-throughput 'omics' data.

Cancer is essentially a genetic disease, in which a group of cells display uncontrolled growth, invasion that intrudes upon and destroys adjacent tissues and spreading to other locations in the body via lymph or blood. Unlike other genetic diseases like cystic fibrosis, there is no single gene defect that directly 'causes' cancer [1]. Hereditary or acquired anomalies in several classes of genes including oncogenes, tumor

suppressor genes and stability genes can lead to the development of cancer. Thus, identifying cancerous genes and the pathways they control is a key step towards the treatment of cancer.

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in the world and contributes to over 655,000 deaths per year [2]. However, in almost all case, early diagnosis can lead to complete cure. The treatment of Cancer includes surgery, chemotherapy and radiation therapy [3]. Unfortunately, these treatments often destroy or injure normal cells and tissues by damaging their genetic material. Thus, there is a great need for identifying new biomarkers for early diagnosis and prognosis and to identify the underlying processes involved in the disease. Furthermore, such biomarkers might be useful for developing cancer therapeutics. Although many important genes responsible for the genesis of various cancers have been identified, the underlying molecular mechanism remains unclear. In this study, the efforts have been given to get a better understanding of the regulation of genes responsible for the development of cancer, particularly colorectal cancer, by analyzing some experimental data.

The cell of a living organism can be viewed as an overlay of complex system of interacting networks. These networks can be roughly divided to three types. A signal transduction network coordinates the response of a cell to the application of an external stimulus. The stimulus can either be chemical or physical, such as light, hormones, sound, smell, nutrients etc. Metabolic network includes all the metabolic and physical process that determine the physiological and biochemical properties of a cell. Regulatory networks are responsible for much of the biological functions within the

cell. Gene regulation refers to the cellular control of the amount and timing of the appearance of functional gene products like RNA, proteins etc. A Gene Regulatory Network (GRN) models the complex regulatory mechanisms that control the activity of genes in the living cells and provides the most realistic representation of gene regulation. With the advent of microarray technology, whole genome expression profiles can be used to understand the regulatory mechanisms behind cancer. The major goal of this study is to extract a cancer specific gene regulatory network by applying various soft computing algorithms on gene expression data.

On a fundamental level, reconstruction of gene regulatory network determines the influence of one gene over another. However such data mining approaches opened the path to explore many questions on the network involving the cell's genes such as:

- Which genes are expressed in a particular cell?
- What is the structure of regulatory networks?
- What are the hidden regulators governing the regulation of a particular gene?

Several techniques have been emerged in the past few years to explore these questions. Gene expression profile from high throughput microarray experiments provides first-hand information on genome wide molecular interactions under different conditions. The analysis of such data provides insight in to the regulatory relations among genes without any prior knowledge. The next session considers issues in analyzing gene expression dataset and subsequent sessions discuss how previous approaches handle

these issues and the motivation behind the selection of the topic for the study.

1.1 Issues in handling Gene Expression Dataset

A principal goal of cancer research is to identify potential biomarkers that specifically characterize a given malignancy. Microarray technology has enabled marker discovery by allowing qualitative analysis of steady-state expression levels of thousands of genes from human cells. However many challenging issues regarding the acquisition and analysis of microarray data have to be taken into account. The first among these is the high variability of data resulting from experimental process. The sources of variability include technical and biological [4]. The biological variability found in gene expression was influenced by various factors including age, sex, time of day of sampling and constituent cell subsets [5]. Technical variability could result from any of the multiple steps involved in the detection of gene expression changes using microarrays including amplification of RNA and hybridization [6]. Either biological variability or technical variability may constitute impediments difficult to surpass by current analysis techniques. Second, the complex experimental procedures in microarray data analysis can contribute to a high noise level and errors. Third issue is related to the large number of different databases with different formats. The lack of standards for presenting and exchanging data is an extremely important problem especially in the case of microarray expression data. Much of the publicly available databases may be incorrect or incomplete due to the volume of the submitted information and the nature of research (e.g., researchers move on to other projects, mistakes in the

original data go unnoticed, etc.). There are also issues of duplication with minor variations and redundancy. Because of this, a global standardization effort, at the initiative of the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI), developed the MIAME (Minimum Information about a Microarray Experiment) project, which is meant to bring uniformity to the disparate storage formats of various databases [7]. Another important issue is the severely underdetermined nature of microarray data, in which the number of variables (genes) greatly exceeds the number of samples (biological specimens), creates a significant risk of overfitting a predictor function. This situation is either due to the high cost of the technology involved in the measurements or to the scarcity of biological samples.

The computational intelligent methods used or developed for problem solving in bioinformatics must be therefore customized in such a way as to efficiently surpass the above-presented issues, as well as several others described in the following sections of this dissertation.

1.2 Extending the State of the Art

Reconstructing and modelling gene regulatory networks form the basis of dynamic analysis of gene interactions and remains one of the most challenging problems of functional genomics. For the better understanding of complex biological phenomena and disease mechanism, the interaction structure of molecular components involved in the cellular process need to be unraveled.Gene networks attempt to describe how genes or group of genes interact with each other and identify the complex regulatory

mechanisms that control the activity of genes in living cells. The inference of GRN from gene expression data is often called 'reverse engineering'. There are two classes of reverse engineering algorithms: One identifying true physical interactions among regulatory proteins and their promoters, and the other identifying regulatory influences among RNA transcripts [8]. In this thesis, the second class: gene-gene interaction network is discussed. The regulation between two genes in a GRN implies direct physical interactions as well as indirect regulations via proteins, metabolites and ncRNA that have not been measured directly [9].

Various reverse engineering methods have been developed to describe models of gene networks.GRN modelling using Kouffman Boolean network presented by Akutsu*et. al*. [10] is simple but assumes that a gene is either on or off (no intermediate expression levels allowed). Models using Bayesian [11] and regression networks [12] deals with the stochastic aspects of gene networks, they fail to consider temporal dynamic aspects that are an important part of regulatory networks modelling. The dynamic Bayesian network can deal with the temporal aspects of regulatory networks but their benefits are hindered by the high computational cost.Woolf and Wang [13] have applied fuzzy rules to every possible combination of genes to find the activator/repressor relationship among genes in a dataset. Although their results are consistent with the literature, the approach is slow and computationally complex.Researches by Shin Ando and Hitoshi Iba [14, 15] have confirmed that the Genetic Algorithm (GA) can infer network structure with significant accuracy. Since GA is a probabilistic search, several

generations and greater computation power were required to model smaller network with good sensitivity and precision.

The approaches proposed in this study extend the previous state of the art by incorporating clustering and some statistical techniques to reduce the computational complexity and processing time. The clustering algorithm used in this thesis can effectively handle noise and outliers. The inference algorithms bring the low-level learning and computational power of neural networks into fuzzy systems and provide the high–level human like reasoning of fuzzy system into neural network. The self-organized nature of the neural fuzzy algorithms produce an adaptive number of fuzzy rules that describe the relationships between the input (regulating) genes and the output (regulated) gene. Related to that, another advantage of the final algorithm is that there is no need of prior data discretization, a characteristic of many inference approaches which often leads to information loss.

1.3 Dissertation Motivation

As stated before, colorectal cancer ranks among the third most common cancers in terms of both cancer incidence and cancer-related deaths worldwide. The blood of CRC patients is known to contain increased levels of DNA fragments released from tumor cells [16]. Thus, the analysis of circulating plasma RNA expression data could be useful for the diagnosis of early stage cancer. The analysis might also constitute a tool to study the development of tumor and therapy responsiveness. A major goal of this study was to identify potential tumor markers in blood and to uncover genetic pathways.

Most of the previous efforts towards the reconstruction of cancer-specific gene networks utilized the expression profile of all genes to identify regulatory relations among genes, some of which actually had nothing to do with the observed cancer phenotype. As a result, it is difficult to detect gene interactions essentially responsible for oncogenesis. To reveal authentic patterns of gene interactions relevant to colorectal cancer, the cancer specific gene regulatory network is reconstructed by focusing on a small set of relevant genes, each of which shows good performance in distinguishing cancerous tissues from normal ones. This network will serve as a blueprint for biologist to understand cancer progression and develop cancer therapeutics.

Recently, it has been reported that changes in expression profiles of short noncoding RNAs such as miRNA play an important role in the development of many cancers, including CRC [17]. Therefore, identification of cancer related miRNAs and their target genes is a key step towards the diagnosis and treatment of cancer. Thus, the five basic motivating questions in this study are

- What are the potential tumor markers in the blood of a colorectal cancer patient?
- What are the likely regulatory relationships among these discriminative genes?
- Which microRNAs (miRNAs) or group of miRNAs regulates a specific gene?

- What are the computational methodologies that can be used to infer a regulatory network?
- How the efficiency of the computational methods can be evaluated using the biological annotations already available?

Reconstruction of gene regulatory network and miRNA-mRNA interaction network are the two computational strategies that can be used to understand the regulatory relations. GRN reconstruction is to detect the components and topology of an unknown pathway, while miRNA-mRNA association network to infer the association between known miRNAs and a gene. This thesis focuses on the reconstruction of colorectal cancer specific gene regulatory network and miRNA-mRNA interaction network from high throughput microarray data.

1.4 Goals and Objectives

Colorectal cancer is both curable and preventable if it is diagnosed early [18]. Unfortunately, many cases of colorectal cancer are not diagnosed until advanced stages because most patients do not develop noticeable symptoms. The overall goal of this study is to identify the tumor markers in blood of colorectal patients and to generate preliminary gene regulatory network for the diagnosis of early stage and /or recurrent colorectal cancer. This will be achieved by developing soft computing algorithms for extracting regulatory relations among the tumor markers by analyzing unique and comprehensive set of expression data. Our aims include extending reverse engineering techniques to generate microRNA-mRNA

interaction network to assist in improved colon cancer therapy design. The objectives of this study are:

1. Identify relevant genes, which show good performance in distinguishing cancerous tissues from normal one.
2. Observe the roles played by high class-discrimination genes in the context of cancer-specific gene regulatory networks.
3. Identify the set of microRNAs involved in the regulation of the above cancerous genes.
4. Develop effective computational methods to reconstruct Gene Regulatory network from experimental data.

1.5 Contributions

The main aim of this study is to develop methods to reconstruct multiscale gene regulatory network that reveal global patterns of gene interactions in cancer. The following are the set of algorithms proposed in this thesis for the analysis of microarray data.

1. Hybrid clustering algorithm - a framework that integrates Hierarchical Agglomerative Clustering and K-means algorithm with the specific goal of eliminating outliers from gene expression data in the process of state space partitioning in to clusters.
2. Fuzzy GRN Algorithm - a model designed to find triplets of activators, repressors and target among the set of selected genes.

3. Modified Genetic Algorithm - a method for optimizing weight matrix for gene regulatory network
4. Dynamic Neural Fuzzy Network - a framework for reverse engineering gene regulatory networks based on the discovery of interactions between genes using a Mamdani-type feed forward neural-fuzzy inference network
5. TSK-type Recurrent Neural Fuzzy Network- an approach that extends dynamic neural fuzzy network by the inclusion of feedback connections to store prior system states.

Based on these algorithms the regulatory relationships between 27 differentially expressed genes in the plasma RNA of Colorectal Cancer patients were modelled. The set of microRNAs involved in the regulation of the above differentially expressed genes were identified. These findings may provide new insights into cancer diagnosis, prognosis and therapy.

1.6 Dissertation Outline

This work focuses on the reconstruction of gene-gene interactions network and microRNA-mRNA interaction network from high throughput gene expression data. The remainder of this dissertation is organized as follows:

- Chapter 2 presents back ground information about the relevant biological concepts. The chapter provides basic information needed to understand the biological process underlying cancer, an essential prerequisite for

computer scientists developing computational tools meant to improve the capability to diagnose and treat patients

- Chapter 3 describes the basic principles and working of high throughput microarray technology used to study the mechanism and structure of gene regulatory network.
- Chapter 4 discusses some of the previous computational intelligence methods used for analyzing gene expression data. It serves as a brief literature review on three relevant research topics such as identification informative genes, clustering of gene expression data and reconstruction of gene regulatory network.
- Chapter 5 provides a brief description of data pre-processing and clustering methods employed in this study to make the data amenable to subsequent analysis.
- Chapter 6 presents the approaches used in this work for inferring the complex interactions among genes from microarray data.
- Chapter 7 compares the performance of algorithms. The results are validated using two more datasets, dataset from colon tumor samples and Yeast data set.
- Chapter 8 describes the additional work for identifying microRNAs involved in the regulation of cancerous genes.
- Chapter 9 summarizes the key contributions of this dissertation, proposes future work in this area and draws some concluding remarks.

Chapter 2

Cancer Biology and Gene Regulation

Contents

Cancer is a complex disease in which a group of cells display uncontrolled growth, invasion that intrudes upon and destroys adjacent tissues and spreads to other locations in the body via lymph or blood. It is a gene disorder that occurs in somatic tissues. Hereditary or acquired anomalies in the regulation of the genes responsible for controlling cell reproduction can lead to cancer. Multiple genes in diverse pathways are involved in its initiation, progression, invasion and metastasis. The first section of this chapter provides a general overview of the biology behind cancer, particularly colorectal cancer which is a commonly diagnosed cancer in western countries. The second section presents a brief description of genetic regulation and the last section describes the role microRNAs in gene regulation and their influence in cancer formation.

2.1 Colorectal Cancer

Colorectal cancer, commonly known as bowel cancer, originates from the inner lining of the colon or the rectum called the mucosa. In most cases, colorectal cancer progresses slowly over a period of 10 to 15 years. It may be present without symptoms for several years. The tumor typically begins as a noncancerous polyp on the inner lining of the colon or rectum (see Figure 2.1). This tumor can be benign or malignant. Benign polyps are

not cancer and are not life threatening. Malignant tumors are cancer. It invades nearby tissues and spreads to other part of the body. The polyps are an early warning sign that colorectal cancer may develop. A polyp may or may not change into cancer. The chance of the polyp turning cancer depends upon the kind of polyp. For example, a type of polyp known as an adenoma can become cancer. If polyps are not removed surgically, they can become malignant over time. Thus screening and removing polyps from large intestine reduces the risk of developing of colorectal cancer.

Once cancer forms in the large intestine, it can grow through the lining and into the wall of the colon or rectum [2].The cancer cells may invade and destroy adjacent tissues and may break away from the tumor and spread via blood or lymph vessels to form new tumor in different locations of the body. The process through which cancer cells break away from primary (original) tumor and travel to distant parts of the body through blood or lymph is called metastasis.

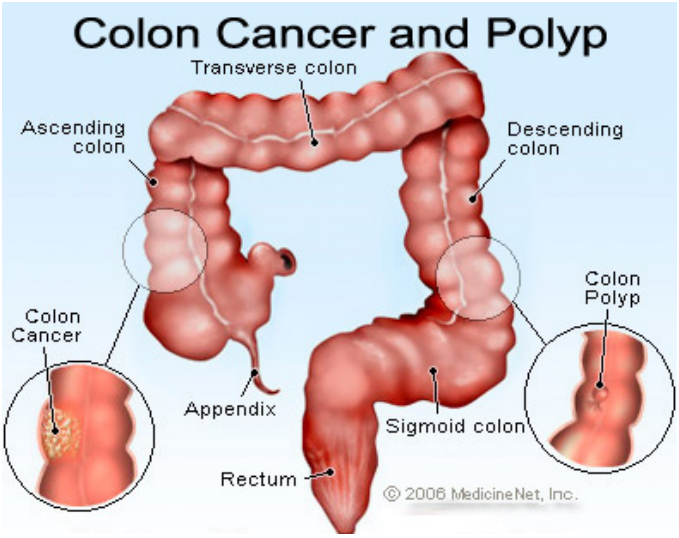


Figure 2.1 Picture of colon cancer. The source of this image is MedicineNet, Inc. http://www.medicinenet.com/colon_cancer

2.1.1 Stages of Colorectal Cancer

The extent to which a colorectal cancer has spread in the body is described as its stage. Staging is one of the most important factors in determining the choice of treatment and in assessing prognosis. It is also useful in predicting the probability of the cancer recurring after surgical removal. Colorectal cancer develops through five definable stages (0-4):

- Stage 0 (*in situ*) – Abnormal cells are found in the innermost lining of the colon and hasn't moved from where it started.
- Stage I (*Local*)– In this stage, cancer has extended beyond the innermost layer of the colon into the middle layers of the colon
- Stage II (*Local*) – Cancer has grown beyond the middle layer of the colon, but has not extended through the wall to invade nearby tissue.
- Stage III (*Regional*)– Cancer has spread through outer most layer of the colon wall and has invaded nearby tissue, or has spread to nearby lymph nodes
- Stage IV (*Distant*)- Cancers has spread through blood or lymph nodes to distant organs, such as the liver or lung

Staging helps in determining whether the treatment may be helpful in preventing or decreasing the likelihood of a cancer recurrence. Stage I colon cancers have survival ratesranging from 80 to 95 percent. Stage II cancers have a survival rate of 55 - 80 percent. A stage III tumor has about a 40 percent chance of cure and a patient with a stage IV colon cancer has only a 10 percent chance of a cure.

2.1.2 Risk factors for Colorectal cancer

The exact cause of colorectal cancer is still unknown. However, researchers have found that certain risk factors that may increase a person's chance of developing colorectal cancer. The factors that increase the risk factor of colorectal cancer includes

- Age- The risk of developing colorectal cancer increases with age. Although this disease can affect people of all ages, most people who develop colorectal cancer are over age 50
- Personal History- A person who has treated for colorectal cancer has an increased risk for developing colorectal cancer in future.
- Family History- A person, whose one or more close relatives (parents, siblings, or children) has had colorectal cancer, is at a risk for developing colorectal cancer.
- Diet- A diet that is high in red and processed meat and low in fresh vegetables and fruits increases the risk of colorectal cancer.
- Physical Inactivity- The people who follows a sedentary lifestyle may have an increased risk of developing colorectal cancer. Regular exercise will reduce the risk of developing colorectal cancer.
- Smoking- Tobacco smoking, particularly long-term smoking increases the risk of colorectal cancer.
- Alcohol- Alcoholic drink, especially drinking heavily, may be a risk factor.

2.1.3 Cancer Genes

Genetic instability is a hallmark of almost all cancers. It refers to a set of events capable of unscheduled alterations, either in temporary or permanent nature, with in the genome. Alterations in three types of genes such as oncogenes, tumor suppressor genes and DNA mismatch-repair genes are responsible for the development of cancer [18].

2.1.3.1 Oncogenes

Oncogenes functions as a positive growth regulators and has the potential to cause cancer. Oncogenes are altered forms of normal cellular genes called proto-oncogenes which produce proteins that regulate cell growth and division. When mutated, oncogenes typically produce more proteins, results in the alteration of the pathway of cell growth and proliferation. This may lead to abnormal growth of cell. For example, K-ras gene is an oncogene that is mutated in colon cancer cells.

2.1.3.2 Tumor Suppressor Genes

Tumor suppressor genes or anti-oncogenes function as a negative growth regulator and suppress tumor formation. They regulate cell growth, differentiation and promote cell suicide (apoptosis).When mutated; tumor suppressor genes produce less of their protein. Thus, apoptosis does not occur and abnormal cell growth results. Tumor suppressor genes such as DCC(Deleted in Colon cancer) and p53 are mutated in colorectal cancer.

2.1.3.3 DNA Mismatch-repair Genes

Mismatch-repair genes (MMR) play a central role in maintaining genomic stability by repairing damaged DNA. When these genes are

mutated, repair does not occur and the cell is more prone to become cancerous. Germline mutations in DNA mismatch-repair genes are associated with the inherited cancer syndrome, hereditary non-polyposis colorectal cancer (HNPCC)

2.1.4 Colon Carcinogenesis

Carcinogenesis, also called tumorigenesis, is the molecular process by which cancer develops. There are four distinct sequential mutations described in the development of colon cancer. This includes mutations of the APC (adenomatous polyposis coli), K-ras, DCC (deleted in colon cancer), and p53 genes. Each mutation causes progressive changes in the colonic epithelium. During initiation phase mutation of APC typically occurs and is sometimes inherited. Mutations in APC lead to benign polyp formation. These polyps can remain inactive for several years. When one cell in this polyp develops a second mutation, in the K-ras oncogene, it grows at a faster rate resulting in a larger tumor or intermediate adenoma. Mutation of tumor suppressor gene DCC represents the third step in genetic pathway. Loss of DCC plays a role in tumor progression, invasion and metastasis. Mutations of p53 lead to late adenoma and finally carcinoma.

2.1.5 Treatment

Treatment options of colorectal cancer depend on the stage of the tumor as well as the general state of the patient like age, medical history, overall health etc. In general, treatments include:

1. Surgery –The tumor and the nearby tissues in the diseased area are removed. In addition to removal of the primary tumor, surgery is

often necessary for estimating the penetration of disease and whether it has metastasized

2. Chemotherapy - Chemotherapy is the treatment of cancer with drugs that can kill cancer cells and thus decrease the chance of the tumor reoccurring elsewhere in the body. It targets all rapidly dividing cells and is not specific to cancer cells. Therefore chemotherapy may harm healthy tissues, especially those that have a high replacement rate.

3. Radiation therapy - High-energy radiation is used to destroy cancer tissue. Radiation destroys any remaining cancer cells after surgery and reduces the chance of cancer spread or recurrence. Although radiation is occasionally used for the treatment for colorectal cancer, in some cases radiation is used in conjunction with chemotherapy treatments to gain better results.

Cancer treatment aims at the complete removal of the cancer without damage to the rest of the body. To some extent, this can be accomplished by surgery, but invasion and spread of disease to distant locations of the body limits its effectiveness. Since chemotherapy is not specific to cancer cells, it is sometimes toxic to healthy tissues. Radiation also damage normal cells and tissues. Therefore, development of novel target specific therapeutics must be necessary for the effective treatment of cancer.

2.2 Biological Aspects of Gene Regulation

Life sciences began with Robert Hooke; who in 1665 discovered cells which are the basic unit of life for all living organisms. There are

different types of cells in our body like brain cells, liver cells, skin cells etc. All these cells have unique characteristics and functions. The nucleus of the cell stores the hereditary material, the genes, in the form of long and thin DNA (deoxyribonucleic acid) strands. The genome of an organism contains necessary information to control of all cellular process like replication of DNA, protein synthesis etc. According to central dogma of molecular biology [19], DNA, RNA and proteins are the three macromolecules essential for all known forms of life. DNA is the carrier of genetic information used in the development and functioning of all organisms. This genetic information is used to encode protein molecules. Three different processes are responsible for the inheritance of genetic information and its conversion from one form to another (see Figure 2.2):

1. Replication: Before a cell divides, its DNA is replicated to give identical copies. It is the basis for biological inheritance.DNA replication is said to be semi conservative since one strand serves as a template for the second strand.

2. Transcription: The process of making single stranded ribonucleic acid (RNA) from DNA template is called transcription. During transcription, a DNA sequence is read by an enzyme called RNA polymerase (RNA pol), which produces a complementary, antiparallel RNA strand. Several types of RNAs are synthesized in the nucleus during transcription. Of particular interest are

    - messengerRNA(mRNA)- later used for protein synthesis

- ribosomalRNA(rRNA)-major component of building ribosome, the protein making machinery
- transferRNA(tRNA)-The molecules that carry aminoacids to the growing peptide chain
- microRNA(miRNA)-tiny RNA molecules that regulate the expression of mRNA

3. Translation: Translation is a process where ribosomessynthesize proteins from the information contained the mRNA. During translation, the ribosome reads a string of three bases on the RNA (codon) and translates them into one amino acid.

Proteins are further processed in sub cellular compartments and transported in-and-out of the cell to carry out different metabolic functions. These highly coordinated activities empower cells to respond to the varying environment with both speed and precision.

Gene expression is a process by which information encoded in a gene is used for the production of gene products such as RNA or proteins. It covers the entire process from transcription through protein synthesis. If the protein is synthesized, a gene is said to be “expressed” and the expression level of gene depends on the amount of mRNA it produced. Different cell types in an organism carry out a range of specialized function depends upon the genes that are expressed only in that cell type. The factors that affect gene expression are the type of tissue, the age of the person, the presence of specific chemical signals etc.

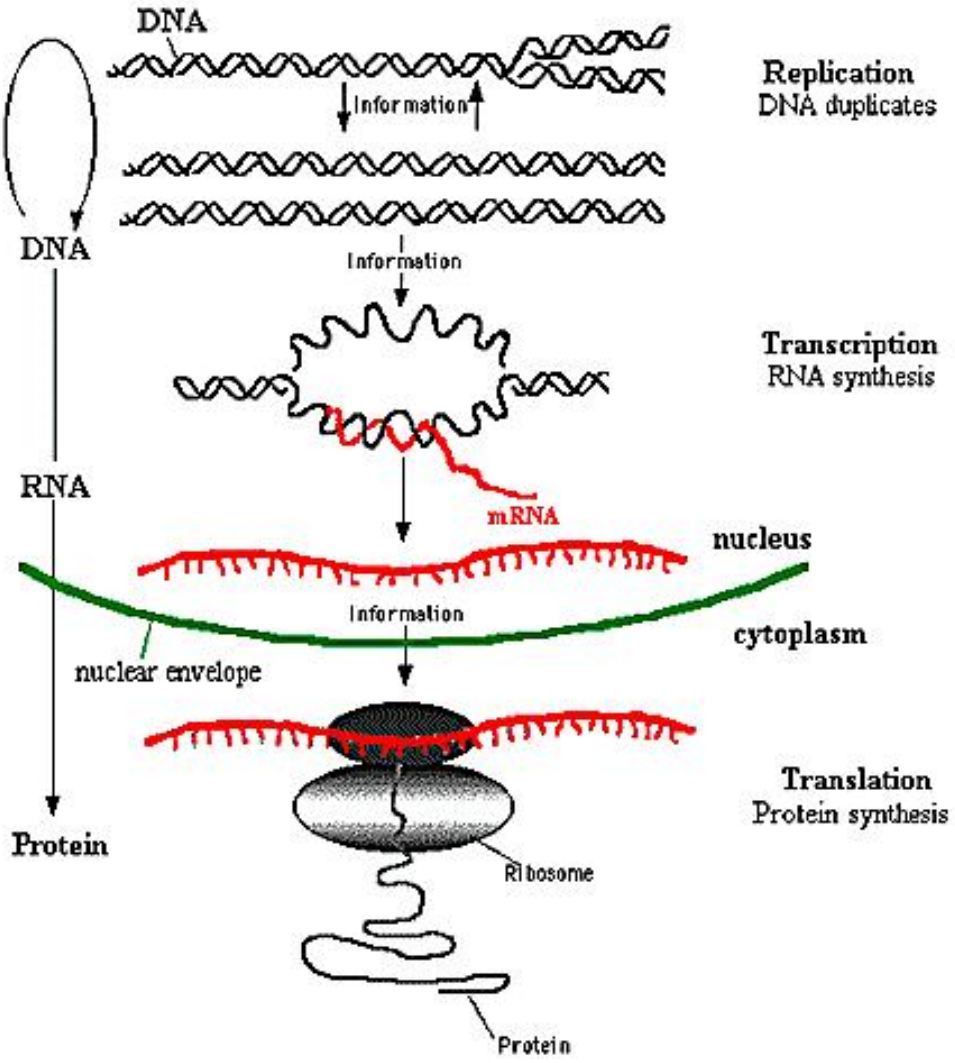


Figure 2.2 Central Dogma of Molecular biology. Genes transcribed in the nucleus are translated in to proteins in the cytoplasm. The figure is taken from http://www.accessexcellence.org/

2.2.1 Gene Regulation

Gene regulation refers to the collection of process that controls the amount and timing of appearance of the functional gene product. It is the

basis for diverse biological process including cell growth and development as well as cellular differentiation, versatility and adaptability of any organism. Gene expression is controlled at three possible levels in the production of an active gene product. First and most important mode for regulating eukaryotic gene expression is the transcriptional regulation. Regulation of transcription controls when the gene is transcribed and how much it is transcribed. Different factors that influence transcription regulation are the strength of promoter elements within the DNA sequences of a given gene, the presence or absence of enhancer sequences (which enhance the activity of RNA polymerase and increase transcription), and the interaction among multiple activator proteins and inhibitor proteins. Second is the translational regulation, controls the amount of proteins synthesized from mRNA. Third, post-transcriptional or post-translational regulation mechanisms control the level of active gene products. Active mRNA level can be controlled by addition of poly (A) tail, splicing, silencing by noncoding RNAs (miRNA, siRNA) etc. Some proteins may also undergo modifications such as folding, enzymatic cleavage, bond formation etc. These modifications can play a vital role in the regulation and control of gene expression.

2.2.2 Gene Regulatory Network

The interactions among genes, proteins and other cellular components form complex circuits that control all biological functions in a living organism. One type of such circuit is gene regulatory network which

represents the interaction structure of genes. A Gene regulatory network (GRN) models the complex regulatory mechanisms that control the activity of genes in living things and provides the most realistic representation of gene regulation. GRN models can be categorized into two classes, detailed and abstract model, according to the level of complexity in the model. In the detailed GRN model, the true physical interactions between regulatory proteins and their promoters are represented. In such models, regulator nodes are either transcriptional regulator proteins or genes and the target nodes are the mRNA levels for the target genes. The figure 2.3 shows the schematic illustration of a detailed GRN model. For instance, gene1 inhibits gene2 and activates gene3, implies that mRNA1 transcribed from gene 1 is translated to protein1 which in turn inhibits gene 2 and activates gene 3. In abstract GRN model such detailed functional descriptions are not represented explicitly. The abstract GRN model is depicted in figure 2.4. In abstract model; genes are represented as nodes and the regulatory relationships as directed edges. The regulatory relationship can be either an activation (increasing the transcription of other genes) or a repression (inhibiting the transcription level).The absence of link between two nodes implies that there is no relationship between two nodes. The regulation between two genes in a GRN implies direct physical interactions as well as indirect regulations via proteins, metabolites and noncoding RNAs that have not been measured directly [8]. This work focuses on inferring abstract GRN models from high throughput microarray data.

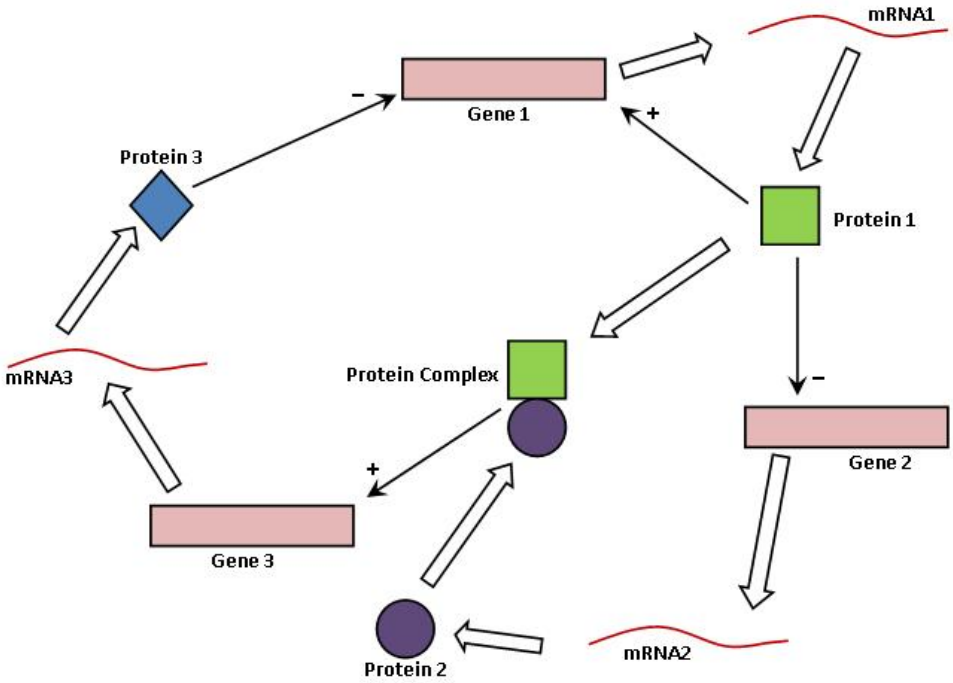


Figure 2.3 An example of a detailed GRN model. Genes can either activate or inhibit themselves or other genes (gene1 inhibit gene2 and activates itself). Often proteins form complex and regulates other genes.

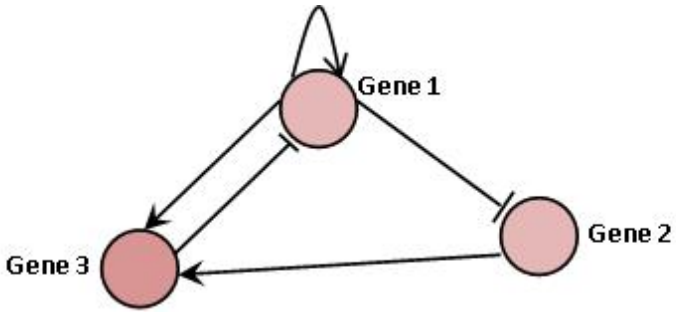


Figure 2.4 Abstract model of the GRN depicted in figure2.3 is shown. An edge ⟶ indicates activation of transcription, where as an edge ⊣ indicates repression of transcription.

2.3 Role of MicroRNAs in Gene Regulation

MicroRNAs are a class of non-coding RNAs that hybridize to mRNAs and regulate their activities at post transcriptional as well as

translational level [20]. There are at least 800 miRNAs within the human genome, which may target about 60% of mammalian genes [21, 22]. MicroRNAs bind to partially complementary sites in the messenger RNA of other genes and inhibit the translation of these genes. They have been found to regulate a wide range of biological process such as cell differentiation, proliferation, growth, mobility and apoptosis in diverse cancer-related biological processes [23, 24].

MicroRNAs were discovered in 1993 by Rosalind Lee, Rhonda Feinbaumand Victor Ambrosduring a study of the gene lin-14 in C.elegans development [25]. Since then, over 4000 miRNAs have been identified in almost all metazeon genomes including mammals, flies, worms and plants. In the human genome as many as 700 miRNAs have been identified yet and over 800 more are predicted to exist. The impact of microRNA on the proteome suggests that the microRNA acts as a rheostat, making fine-scale adjustments to protein synthesis from thousands of genes [26, 27].

2.3.1 Biogenesis of miRNAs

MicroRNA biogenesis is a stepwise process that starts in the nucleus and ends in the cytoplasm (see Figure 2.5). Most miRNAs are located in the introns of protein and non-protein coding genes or even in exons of long non-protein coding transcripts [28]. MiRNA genes are usually transcribed by RNA polymerase II (Pol II) in the nucleus [29]. The miRNA sequence and its reverse-complement base pair to form a double stranded RNA hairpin loop called pri-miRNA (primary miRNA structure). The nuclear enzyme Drosha and its cofactor DGCR8/Pasha cleave the base of the hairpin to form pre-miRNA of about 70 nucleotides in length. The pre-miRNA hairpins are transported from the nucleus into the cytoplasm by Exportin 5, a carrier protein.

In cytoplasm, RNase III enzyme Dicer cuts 20-25 nucleotides from the base of the hairpin yielding an imperfect miRNA:miRNA* duplex [30]. The functional strand of the microRNA duplex is then loaded into Argonaute protein within RNA-induced silencing complex (RISC) and becomes mature miRNA, whereas the other strand, miRNA*, is degraded [31, 32]. Finally, the mature miRNA load in RISC is potent for regulating protein production, either by translational repression or mRNA cleavage.

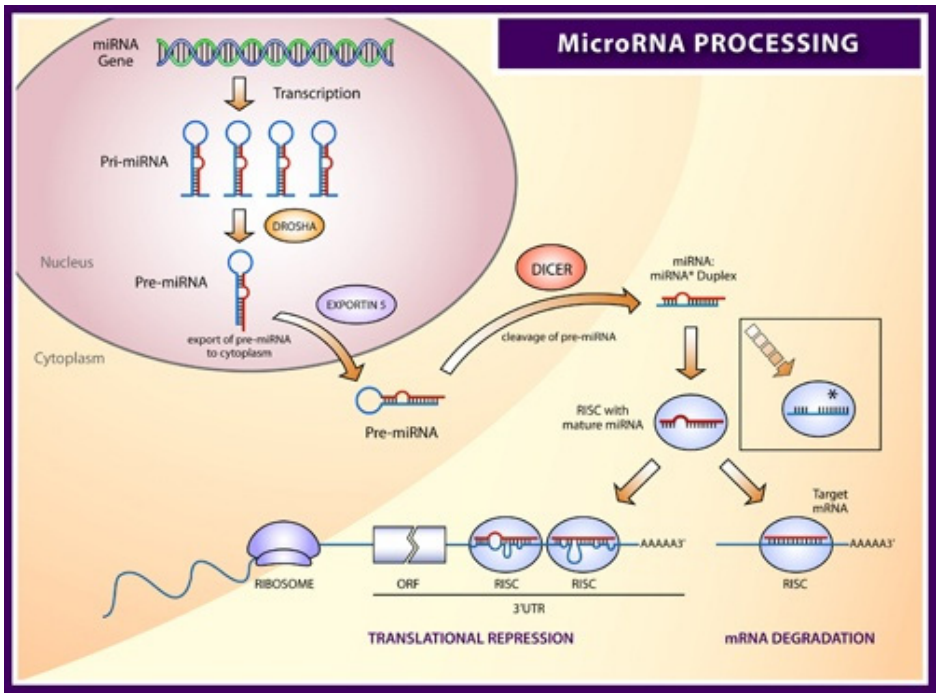


Figure 2.5Pathway from microRNA biogenesis to mRNA regulation.The microRNA gene is transcribed by RNA polymerase II into a double stranded RNA hairpin loop called the primary transcript or pri-microRNA. The nuclear enzyme Drosha cleaves the flanking sequences, resulting in the ~70 nucleotide long pre-microRNA. After the relocation of pre-miRNA into the cytoplasm by exportin-5, Dicer, RNase III enzyme, performs the second cleaving step called 'dicing' to produce the microRNA:microRNA* duplex. Subsequently the duplex is separated and one strand gets incorporated into the RISC, while the other strand is degraded. Finally the microRNA loaded RISC is potent for regulating protein production, either by translational repression or mRNA cleavage. The image is taken fromhttp://dna-rna.net

2.3.2 MicroRNA and Cancer

MicroRNAs have diverse expression pattern and might play a key role in various developmental and physiological processes like cell development, proliferation, mobility, differentiation and apoptosis [23, 24]. Accordingly, altered miRNA expression is likely to contribute to a wide range of human diseases, including cancer.The findings that miRNAs have a role in cancer are supported by the fact that many miRNA genes are located at fragile sites in the genome or regions that are commonly amplified or deleted in human cancer [33].Also, malignant tumors and tumor cell lines contain widespread deregulated miRNA expression compared to the corresponding normal tissues [34,35].

First evidence of involvement of miRNAs in cancer was reported in 1999 [36]. Calin*et. al*. identified that two miRNAs, mir-15 and mir-16,were involved in the pathogenesis of chronic Lymphocytic Leukemia. Later, in 2005, He *et. al.* [37] demonstrated that miRNAs from mir-17-19 cluster were over expressed in lymphoma cell lines. In the same year, Johnson *et. al.* [38] experimentally confirmed that loss or reduction of let-7 in lung cancer lead to the over expression of RAS oncogene which in turn results in the increased cell growth and tumerogenesis. The authors suggested let-7 act as tumor suppressor. Recent experiments also show that miRNAs upregulate genes in one condition, but act as a negative regulator in another condition. For example, let7 and the synthetic microRNA miRcxcr4-likewise upreguate target mRNAs upon cell-cycle arrest; yet, they inhibit translation in proliferating cells [39].

In general, changes in the expression pattern of miRNAs can influence carcinogenesis if their mRNA targets are encoded by oncogenes or tumor suppressor genes [17]. Recent functional studies suggest that miRNAs regulate many known oncogenic and tumor suppressor pathways involved in the pathogenesis of Colorectal Cancer [40, 41]. MiRNAs regulate many proteins involved in key signaling pathways of CRC, such as members of the Wnt/β-catenin pathway, EGFR signaling (KRAS and phosphatidylinositol-3-kinase (PI-3-K) pathways) and p53 pathways [17].Thus the analysis of such miRNAs is useful for cancer diagnosis, prognosis, treatment and drug target discovery.

Chapter 3

Microarray Technology

Contents

Microarray based gene expression analysis provides an adjuvant tool to understand the cancer-causing processes at the molecular level. This chapter aims to provide an overview of the principles of microarray technology. It has been divided in to four sections. The first section provides basic concepts on the working of microarray and the basic principles behind the microarray experiment. The second section deals with the practical concerns of the analytical processing of the gene expression data obtained. The third section focuses on the microarray applications in distinct areas of basic and clinical science. Finally, the last section provides the challenges and future prospects in the development and clinical use of microarray-based tests.

3.1 Introduction

Understanding of biological organization in the system level is a key objective in post genome era. Measuring the expression level of genes across different tissues or cells under different environmental conditions is very important and useful for understanding and interpreting the biological process. With the emergence of high throughput technologies such as

microarray, it is possible to measure simultaneously the combinatorial changes in thousands of individual genes, proteins and metabolites in cells. Microarray technology is considered to be one of the most important and powerful tools used to extract and interpret genome wide molecular interactions at specific conditions. The analysis of microarray data will provide new insights into the targets for the treatment of disease which is aiding drug development, immunotherapeutics and gene therapy.

The term microarray is synonymously used with DNA microarray, is a collection of microscopic DNA spots attached on a solid surface, usually glass. Each DNA spots contain many copies of the same single stranded DNA sequence (called probes) that uniquely represents a gene from an organism. Since the microarray contains thousands of such spots, it can accomplish many genetic tests in parallel. Besides DNA microarrays, there are different types of microarrays depending on the biological material embedded on the spot. These include protein microarray, microRNA microarray (MMchips), tissue microarray, antibody microarray, cellular microarray etc. [42]. Since all types of arrays are based on the same conceptual foundations, DNA microarray has been discussed in the rest of this chapter.

3.2 Microarray Experiment

The core principle behind the microarray technology is hybridization, the complementary nucleotidessequence stick to or “hybridise” to, one another. For example, a DNA molecule with the sequence -A-T-G-A-C- will hybridize to another with the sequence -T-A-C-T-G- to form double-stranded DNA. There are two formats of arrays prevail today: the spotted cDNA microarray and the Affymetrix oligonucleotide array.

The cDNA array (or two- color or two-channel microarrays) has been widely used and made popular through the work of Patrick Brown and his colleagues at Stanford University [43]. The complementary DNA (cDNA) is synthesized from messenger RNA (mRNA) template and copied rapidly using polymerase chain reaction (PCR).The length of cDNA sequence vary from a few hundred bases to a thousand or so. Thousands of cDNAs are spotted onto an individual array to serve as microarray probes. These probes of known identity, is used to determine complementary binding of the unknown sequence in a sample. Each spot representsa specific gene, but some genes may be represented by multiple spots. The cDNA array uses two colours, red/green, labeling cDNA from one sample with red dye and cDNA from another with green dye. Both labeled samples can be mixed and hybridized to one single array and then scanned to determine the relative binding for each probe.

The oligonucleotide microarray, also called one-color or single channel microarrays, is developed by Affymetrix Inc. under GeneChip® trademark. The oligoarray contains gene specific oligonucleotide probes of 25 nucleotides in length (25-mers), which are synthesized on the chip by a patented technology called photolithography. Many companies are manufacturing oligonucleotide based chips using alternative technologies. In oligonucleotide microarray, each gene is normally represented by more than one probe. The collection of probes designed to interrogate a given sequence of gene is usually called probe set. A probe set composed of 16-20 separate probe pairs representing the mRNA sequence of interest. Each of the probe pair consists of a perfect match sequence (PM) and a corresponding mismatch sequence (MM). The prefect match sequence is complementary to a reference sequence of interest.The mismatch sequences are same as PM except for

homomeric base change (A-T or G-C) at the 13th position.The scanned result for a particular gene is the average signal difference between PM and MM across a probe-set. The oligonucleotide microarrays are used to measure absolute value of gene expression and therefore the comparison of two conditions require two separate microarrays. The basic steps in the microarray experiment include chip fabrication, target preparation and hybridization and scanning (see figure 3.1)

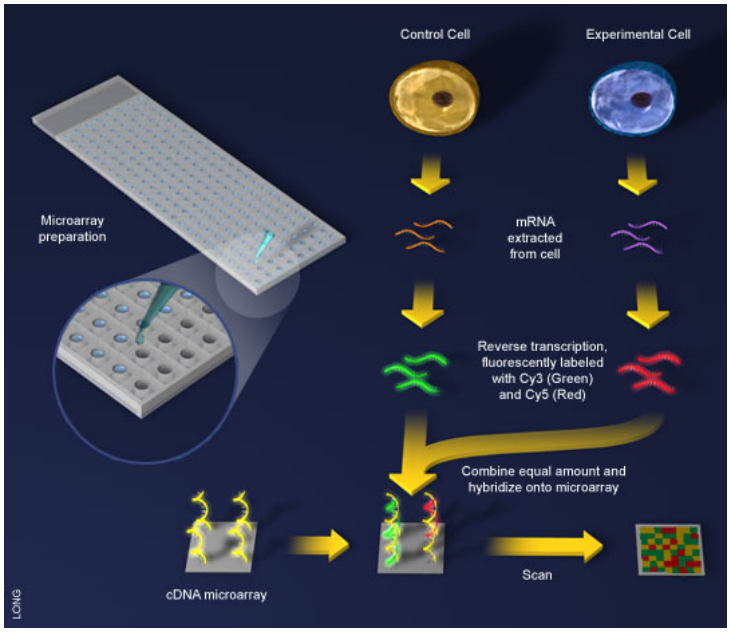


Figure 3.1 The schematic diagram of microarray experiment. Microarray technology allows the simultaneous monitoring of expression levels of thousands of genes. The mRNAs extracted from the control sample is labeled with cy3 (green) dye and from the experimental sample is labeled with cy5 (red) dye. The two labeled samples are mixed in equal proportion and hybridize onto the microarray slide. The microarray slide is scanned using a laser and the image obtained is stored for further analysis. The image is taken from http://www.scq.ubc.ca

3.2.1 Chip Fabrication

Spotted arrays are fabricated using an arrayer or a spotter, a high precision dispensing device mounted on a robotic arm. The dispensing

device can be either a pin (contact printing) or an inkjet needle (non-contact printing) [44]. The probes are synthesized off-chip and the spotter will deposit each probe at designated locations on the array surface. The precision and speed of the non-contact printing is far greater than the contact printing [45]. The performance and quality of the microarray depends on several parameters such as array geometry, spot density, morphology of the spot, probe and hybridized density, specificity etc. The factors that affect these parameters are shown in figure 3.2.

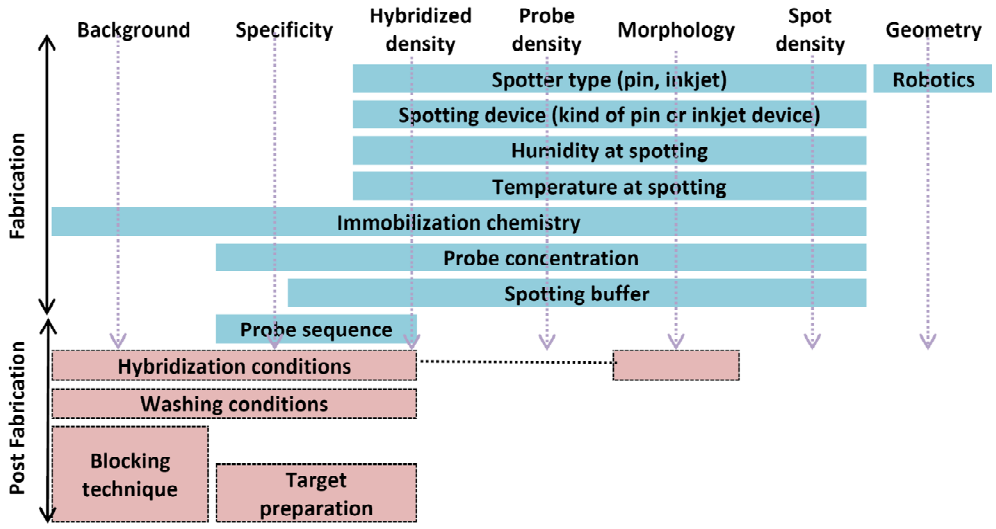


Figure 3.2 Parameters and factors that determine the performance of DNA microarrays. Array geometry is the spatial localization of spots in the microarray and spot density defined as the number of (different) spots that can be fabricated in a given area. Spot performance is affected by three parameters; morphology, probe density and hybridized density. Morphology concerns the shape and homogeneity of the spots. Probe density is a measurement of how many probe molecules that are immobilized in a given area and hybridized density is defined as the number of target molecules that can hybridize to a given area. Specificity is defined as the number of target molecules cross hybridizes with imperfectly matched probes while background is a measurement of the noise coming from the slide. The performance parameters are influenced by fabrication specific factors (marked in blue) and post fabrication factors (marked with a dotted square).

Oligonucleotide arrays are manufactured either by spotting prefabricated oligos on a substrate or by synthesizing the sequence directly (situ oligo synthesis) onto the array surface instead of depositing intact sequences. Probes can be 25-60 mer long depending on the desired purpose. The photolithographic technique using photolabile protecting groups utilized by Affymetrix is an example of situ oligo synthesis. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface to "build" a sequence one nucleotide at a time in the illuminated region. Maskless manufacturing systems are also possible, in which a digital macromirror array is used instead of masks [46].

Bead based microarray is an alternative technology emerged recently. This technology uses silica beads that self-assemble in microwells on either of two substrates: fiber optic bundles or planar silica slides. The beads are three-microns in diameter and are randomly assembled on the arrays about 5.7 microns apart. Each bead is printed with about a million copies of a specific oligonucleotide that captures specific sequences. Bead Array technology comes in 2 formats: the Sentrix Array Matrix and the Bead Chip. The table 3.1 lists the popular microarray fabrication techniques available today.

Table 3.1 Comparison of popular microarray fabrication Techniques

Class	Technology	Principle of Fabrication	Spot size (µm)	Spot Density (spots/cm²)	Max probe length	In house
Spotted arrays	Contact printing	Deposit biomaterial by robotic controlled pins	100-200	2000-4000	>100	Yes
	Non-contact printing	Deposit biomaterial by robotic controlled inkjet needles	100	4000-5000	>100	Yes
	Contact printing	Robot with Cantilevers	0.1	100 million	>100	Yes
Optical arrays	Bar coded beads (limina, Luminex, Qdots)	Coating beads with biomolecules	N/A	N/A	>100	Yes
In situ synthesis	Mask dependent fabrication	Light guided synthesis defined by masks (Affymetrix)	5	4 million	25	No
	Maskless fabrication	Spotting phosphoramidites using inkjet technology (Agilent)	100	5000	60	No
		Spotting phosphoramidites using inkjet technology (Open source)	100	9800	?	Yes
		Light guided synthesis defined by mirrors (Nimblegene)	16	97500	132	No
		Light guided synthesis defined by mirrors (FEBIT)	N/A	48000/chip (-10,000/cm²)	30	Yes

3.2.2 Target Preparation and Hybridization

To obtain the gene expression profile at genome level, the collection of probes in the microarray is hybridized with the labeled targets. The target is the mRNA population isolated from the tissue being studied. High quality RNA is important for the success of microarray experiments. For cDNA microarray, mRNA is extracted from two cells, one from the control sample and the other from the experimental sample. The isolated RNAs are reverse transcribed into complementary DNA (cDNA) using an enzyme reverse transcriptase and are labeled with either radioactive or fluorescent markers. The commonly used fluorescent markers include Alexa Fluor, Phycoerythrin and cyanine dyes.Cyanine dyes, cy3 and cy5, have enhanced photostability and their excitation and emission spectra are spatially well separated. These fluorophores emit light of different wavelengths that are usually visualized as red (cy5) or green (cy3) after excitation with appropriate lasers.

The Affymetrix technique employs a single fluorescent label. First, cDNA is synthesized from the RNA sample using reverse transcriptase and an oligo-dT primer.Next, in the presence of biotin-labeled rNTPs, RNA polymerase uses cDNA as a template to transcribe multiple copies of biotinylated antisense mRNA. This biotinylated RNA is referred to as aRNA or cRNA. The resulting cRNA is fragmented in the presence of heat and mg+2 before hybridization. Proper fragmentation facilitates efficient and reproducible hybridization.

After preparation, the labeled targets are purified and denatured with water, and are hybridized to the microarrays. Each cDNA sequence in the sample will hybridize to specific spots on the glass slide containing its complementary sequence. After hybridization, the arrays are washed under stringent conditions to remove unbounded or loosely bounded targets and are air-dried.

3.2.3 Scanning

The amount of labeled sample bound to each spot can be measured with a confocal laser scanning microscope or a charge-coupled device (CCD) camera. The laser beam is used to excite the fluorescent probes and the amount of fluorescence emitted corresponds to the amount of bound nucleic acid at a specific spot. The intensity of fluorescent emission from each spot is quantified by either a CCD camera or a photomultiplier tube (PMT), which converts light energy into analog electrical signal. A high-resolution image is obtained by use of confocal microscopy.

The areas on the array with hybridized probes will be visible on the scanned image as red or green spots. The red spots represent genes expressed in the experimental sample while green spots represent genes expressed in the control sample. Yellow spots represent genes whose expression does not vary substantially between two samples. If the gene was not expressed in both conditions, the spot would be black. The standard image format for saving microarray images is a 16-bit tagged image file format (TIFF). Figure 3.3 illustrates a cDNA microarray image.

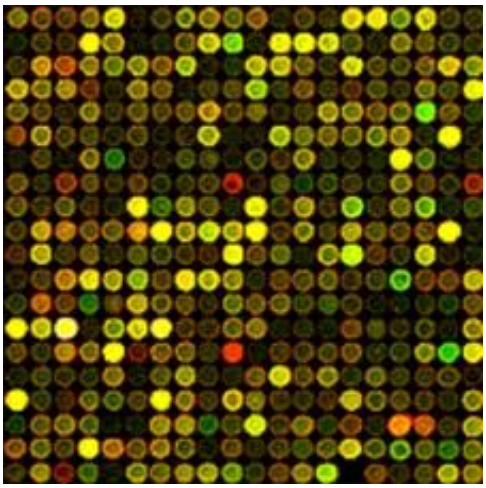


Figure 3.3 cDNA Microarray Image. The source of this image is http://www.nchu.edu.tw. The red and green spots indicate the relative abundance of the corresponding mRNA in the two cell populations. The yellow spot indicates that the specific gene is equally expressed in both samples. The dark spot represents genes that are not expressed in both samples. The images are saved in standard 16-bit TIFF format.

3.3 Microarray Analysis

Microarray experiments produce a massive amount of genome-wide data from large scale measurement of gene expression. The objective of microarray analysis is to draw biologically meaningful conclusions from this enormous data. Although microarray data analysis is a complex subject, the basic steps of analysis are image analysis, transformation, normalization and analysis of gene expression data.

3.3.1 Image Analysis

The goal of microarray image analysis is to provide the intensity levels of each mRNA detected spots which correspond to a single gene expression and input features for further analysis.

The steps include

i. The Gridding or assigning coordinates for each spot in the image.

Gridding is the primary task of microarray image processing, which consists of determining the spot locations in a microarray image. A grid covering each block is constructed so that it isolates each spot into a distinct cell, enabling the localization of each spot. The grid file contains information about the identity of each spot.

ii. Foreground separation, Segmentation of spots.

Segmentation is the process of separating the foreground pixel from the background in a microarray spot. The accuracy of segmentation affects the quality of expression data. Foreground pixel intensities will be used in calculating the signal and the background pixels are considered as noise.

iii. Intensity extraction of each spot.

This step includes extraction of foreground and background intensities for the red and green channels for each spot on the microarray. Also, the mean, median and the total values for the intensity considering all the pixels in the defined area are reported for both the spot and background. This is done for all spots in the array.

After gridding, the microarray spot is marked as a circle. The spot median is the median value of all pixels inside the circle. A square is drawn around the circle. The median value of pixels inside the square box and outside the circle is the background median value. The spot intensity is the difference between the spot median and the background median value.

3.3.2 Transformations of Expression Ratio

The relative expression level of a gene in both samples can be estimated as the amount of red or green light emitted after excitation. The most common metric used to relate this information is called expression ratio. The expression ratio $T_i$can be defined as

$$T_i = \frac{R_i}{G_i}$$

where R_i represents the spot intensity metric of gene i for experimental sample and G_i represents the spot intensity metric for control sample.

The expression ratio is a relevant way of measuring expression changes in an intuitive way. For example, the genes that do not change their expression value across the samples will have an expression ratio 1. However, the representation is troublesome as they treat up and downregulated genes differently. Genesupregulated by a factor of 2 have an expression ratio of $2\left(\frac{R}{G} = \frac{2G}{G} = 2\right)$, while those downregulated by the same factor have an expression ratio of $0.5\left(\frac{R}{G} = \frac{R}{2R} = \frac{1}{2}\right)$. As a result, the upregulated genes are expanded and between 1 and positive infinity while downregulated genes are compressed and mapped between 1 and 0.

To rectify this inconsistency in mapping, alternative transformations of the expression ratio such as inverse transformation and logarithmic transformation are used.

1. Inverse or Reciprocal Transformation

It converts the expression ratio into a fold change. For the genes with expression ratio < 1, fold change is the reciprocal of expression ratio multiplied by -1. If the expression ratio is ≥ 1 then the fold change is equal to the expression ratio. Thus, for the i^{th} gene on the array, the fold change can be defined as

$$FoldChange = \begin{cases} T_i & T_i \geq 1 \\ -\dfrac{1}{T_i} & T_i < 1 \end{cases}$$

The advantage of this approach is that upregulation and down-regulation can be represented with a uniform mapping interval. The disadvantage; however is that the mapping space is discontinuous between –1 and +1 and hence not suitable for most mathematical analysis.

2. Logarithmic transformation

Logarithmic transformation treats upregulated and downregulated genes uniformly and produces a continuous spectrum of values for differentially expressed genes. For instance, a gene upregulated by a factor of 4 has a $\log_2(\text{ratio})$of +2, a gene downregulated by a factor of 4 has a $\log_2$ (ratio) of −2, and a gene expressed at a constant level (with a ratio of 1) has a $\log_2$ (ratio) equal to zero. Thus the mapping space is continuous and up- and down regulation are comparable.

It should be noted that there are some disadvantages in using expression ratio for data analysis. Although expression ratios can reveal

some patterns in the data, they remove all information about absolute gene expression levels.

3.3.3 Normalization

In the case of microarray experiments, there may be variations in the hybridization intensities that affect the measurements of gene expression levels. The sources of variations include unequal quantities of starting mRNA, array surface chemistry, sequences used for the probes, differences in the labeling and detection efficiency of fluorescent dyes used etc. Normalization attempts to adjust the individual hybridization intensities so that meaningful biological comparisons can be made [47]. It can be thought of as the first level of filtering applied to the data. Normalization factor (scale factor) is calculated for each gene and used to adjust the data to compensate the observed technical variations. Although a number of normalization techniques can be applied to microarray data, the most commonly used are total intensity normalization, ratio statistics and regression normalization [48, 49].

3.3.4 Analysis of Gene Expression Data

Current methodologies to analyze gene expression dataset can be divided into two categories-supervised approaches and unsupervised approaches. Supervised analysis identifies genes that fit to a known class whereas unsupervised analysis characterizes genes or samples in a dataset without making use of prior knowledge.

Supervised methods are generally used for two purposes: identifying genes that are differentially expressed in different samples and finding genes

that accurately predict the sample characteristic. Only few genes in the dataset will be useful to differentiate samples among different classes, while many other genes are not relevant to this task. Such irrelevant genes will increase the dimensionality of the problem and thus results in unnecessary computational difficulties and additional noise [50, 51, 52]. Feature selection can be done in different ways, including parametric [53] and nonparametric tests [54, 55], analysis of variance (ANOVA) [56] and many others. There are several supervised methods that find genes that accurately predict sample characteristics, such as distinguishing different subclasses of given class of tumor, or a metastatic tumor from a non-metastatic one. The popular methods include decision trees [57], neural networks [58] and support vector machines [59, 60].

Unsupervised methods are usually used to identify the relationship among the components of the dataset. Within unsupervised learning, there are three classes of technique: dimensionality reduction, or projecting row or column vectors of a gene expression data matrix on to a low-dimensional space, such as principal-components analysis[61, 62]; cluster determination, or determining groups of genes or samples with similar expression pattern, such as self-organizing maps[63, 64], k-means clustering[65], hierarchical clustering[66] etc.; and network determination, or determining graphs representing gene–gene or gene–phenotype interactions using Boolean networks [10, 67], Bayesian networks [11, 12], artificial neural networks [68, 69, 70] etc.

3.4 Microarray Applications

Microarrays are a powerful genomics tool, designed for the complete analysis of genetic material by monitoring of expression changes occurring in a biological sample under various conditions. Microarrays is being widely used in various research areas such as sequencing, single nucleotide polymorphism (SNP) detection, characterization of protein-DNA interactions, miRNA profiling, and many more. The current successful application of microarrays in biosciences includes clinical diagnosis, pharmacogenomics, drug discovery, toxicogenomics, pathogen detection, and genotyping. All of these potential uses increase the demand for microarrays in both academic and industrial settings

3.4.1 Medical use of Microarrays

Global gene expression profiling of tumor cells may help early diagnosis and prognosis of various types of cancer. It may also provide valuable clues in exploring potential drug targets that would have been overlooked otherwise. Particularly exciting is the potential of DNA array technology to provide individualized therapies or in general 'personalized medicine' for a wide variety of clinical conditions based upon the molecular signature or fingerprint for each of these conditions [71]. Moreover, the combination of microarray expression profiling and tissue arrays provides tissue mapping for the disease-specific candidate genes [72]. The application of microarrays in blood testing provides an integrated platform for comprehensive donor and donation testing, replacing multiple individual

assays [73]. Microarray procedures can also be used to identify the changes in gene expression that are associated with drug abuse and alcoholism [74, 75].

3.4.2 Microarrays in Drug discovery and Development

Microarray expression data can be used for generating clues to gene function that can help to identify appropriate targets for therapeutic intervention [76]. They can also be used to monitor gene expressionchanges in response to drug treatments. The role of microarrays in drug discovery includes

1. Target identification

Drug developers can generate hypotheses for complex disease mechanisms, such as those associated with cancer, with genome-wide expression profiling using DNA microarrays, allowing them to identify targets for drugs. The microarray data obtained by profiling a diseased cell is analyzed to discover pathways. Using the novel pathways drug developers can identify new drug targets for various diseases.

2. Toxicogenomics

Toxicogenomics is the early prediction of the toxic side effect of a particular drug. By comparing the gene expression profile of a cell on a particular drug to a known database of toxic cell expression profiles, researchers can identify whether the drug will cause dangerous side effects. This would mean that drug companies could test a drug in development for its possible toxic side effects without having to test it on animals or humans [77].

3. Pharmacogenomics

A person's response to a drug is based on the differential genetics and expression characters. Pharmacogenomics involves personalizing treatments to a particular patient based on their genotype or gene expression profiles. DNA microarrays are able to produce patient's genetic information needed for pharmacogenomics. For instance, Amplichip CYP450, a microarray manufactured by Roshe molecular system and Affymetrix aids doctors in individualizing treatment options for patients. This system uses DNA extracted from a patient's blood to detect certain common genetic mutations that alter the body's ability to break down (metabolize) specific types of drugs [78].

3.4.3 Microarray based Oncology

Cancer is a genetic disease, arising from the progressive accumulation of genetic alterations in somatic cells. Since hundreds of genes are simultaneously involved in the cancer formation, microarray based tumor profiling will provide a better understanding of tumor formation and development. In addition, the identification of differences in expression profile of tumor cells in comparison with their normal counterpart will provide information that is expected to improve our understanding of the complex molecular interaction networks within the cell [79]. Such characterization of the molecular structure involved in pathology of cancerenables the identification of new drug targets and the development of new therapeutics. Moreover, substantial molecular heterogeneity among certain types of cancers like breast cancer and prostate cancer makes

prognosis and response to current treatments highly variable and difficult to predict. Expression profile will help to classify controversial tumors and provide new prognostic tools and potential therapeutic targets. They will also give new insights to the molecular events responsible for the development and progression of cancer [80].Some other applications of microarray in cancer research includes identification of single nucleotide polymorphism(SNPs) and mutations, classification of tumors, identification of target genes of tumor suppressors and identification of gene associated with chemo resistance and drug discovery.

3.5 Challenges and Future Prospects

Although microarrays have been extensively used for numerous research discoveries which have laid the ground work for evaluating disease susceptibility, diagnoses, and prognoses, only few have been translated into clinical practice. This inaction is attributable to technical, clinical and marketing challenges. One of major challenge facing current microarray technology is the interlaboratory variability. Variation in biosamples, RNA quality and target labeling, have been identified as prime sources of analytical variability. Careful experimental design and initial calibration experiments can minimize this challenge. Another challenge is for companies to overcome the high costs associated with development and production of DNA chips, thus making them affordable for the end-users.

The biggest challenge in microarray, however, is the challenge of data handling and informatics. This is due to the quality, volume, dimensionality of data generated in each experiment. Because of such large amount of data,

false positive results are likely to be obtained. Novel algorithms and software must be evolved for the analysis of these large and diverse dataset. In addition, microarray experiments can generate data set with multiple missing expression values. Accurate methods for imputing data are needed to minimize the effect of incomplete data sets on analysis [81].

One remaining hurdle when adopting microarray based test is to generate public interest or awareness. Most people are unfamiliar with such tests and few physicians are familiar in the intricacy of complex genetic test interpretation and how to advise patients based on the results. An urgent task ahead is to develop supporting information systems which offer educational and consultation programs for physicians and other health care providers as well as the general population to facilitate the adoption and use. Ultimately, these educational programs will display the impact of microarray-based tests in therapeutic decision making.

Several microarray-based tests have now come to fruition and translated in to clinical practice. The AmpliChip CYP450 from Roche [78] and MammaPrint from Agendia [82] are the first FDA approved microarray-based tests for diagnostic applications. Additional cancer related test includes Pathwork® Tissue of Origin Test [83] and AmpliChip p53 [84]. The Tissue of Origin Test has cleared by the FDA in 2010.This gene expression–based test uses a tumor's own genomic information to aid in identifying challenging tumors, including metastatic, poorly differentiated and undifferentiated tumors.The AmpliChip p53 test, currently under development at Roche, is designed to detect damage to p53 DNA in tumor cells to identify which cells might have dysfunctional p53 and thus resist

treatment. All diagnostic microarray testing is ordered by physicians and tested by a Clinical Laboratories Improvement Amendment-certified (CLIA) reference laboratory. Several companies are now offering direct-to-consumer services, whereby customers will supply a DNA sample and the company will provide them with information on a very board range of disease risk. Two US companies Navigenics and 23andMe represent pioneering companies in this field.

Many microarray-based tests are under active development. BioMerieux is applying AffymetrixGeneChip Technology towards the development of HIV genotyping and microbial contamination testing; Skyline Diagnostics is developing ALL profiler for testing acute lymphatic leukemia and Brain profiler for the classification of brain tumors including OGD subtype. Guided by these early clinical practices and driven by the explosion of new discoveries and growing marketing demands, the next wave of microarray-based tests will soon be upon us. However, there is still considerable work ahead before these tests are incorporated into routine clinical practice

Chapter 4

Microarray Data Analysis

Contents

The recent advances of microarray technologies have made it possible to monitor simultaneously the expression pattern of thousands of genes in genomes. The challenge is to analyze effectively and interpret this large volume of data. The intrinsic nature of microarray data such as high dimensionality and small sample size calls for effective computational methods. This chapter describes several computational intelligence techniques used for analyzing gene expression data. These are either well known tools successfully used for different data analysis tasks and useful also in the area under discussion or customized computational or statistical tools developed to handle the peculiarities of gene expression data. The related methods fall in three research topics

Identification of non-redundant information from gene expression data

Clustering of gene expression data

Inference of gene regulatory network

Section 4.1 discusses statistical analysis methods used for identifying discriminative genes from microarray expression data. Section 4.2 refers to

analysis. Section 4.3 ends the chapter by presenting computational methods used for inferring gene regulatory networks from microarray data.

4.1 Statistical Methods for Identifying Differentially Expressed Genes

Among a large number of genes encoded in the microarray gene expression data, only a very small fraction of them are relevant for a certain task. A very challenging problem arises as a result – how to select informative features (genes) for performing data analysis such as diagnosis, prognosis, subtype classification of a heterogeneous disease and understanding a gene regulatory network. This selection procedure is important and sometimes necessary because of three main reasons. First, it is impossible for biologist or physicians to examine the whole feature space in the laboratory experiments at a stretch. Second, irrelevant features result in unnecessary computational difficulties. Third, especially in the case of cancer (or generally disease) classification, since the activity of only few genes are responsible for the development of the disease, it is obvious that the rest of the genes measurements are irrelevant to the task of class distinction. Such irrelevant genes act as 'noise' since they confuse classifiers and thus obstruct biological information within data sets.

Identifying genes or subsets of genes that are differentially expressed in disease and normal tissues using computational algorithms will obviously increase the diagnosis accuracy. Besides this, post classification analysis of the respective discriminative (or predictive) genes may reveal important

information in what concerns the dynamics of the disease. This would be highly beneficial for drug discovery and early disease prediction.

Currently, several statistical and machine learning methods have been developed for gene selection. Among them feature ranking approaches are particularly attractive because of their simplicity, stability and empirical success. In these approaches, features are scored by a certain ranking criterion and the rank of features is used as the base of selection mechanism. For example t-statistics [85], regression model [86], χ^2 -statistics [87] and mixture model [88] can be used for feature ranking. Some other approaches utilize machine learning approaches such as support vector machines (SVM) [89, 90, 91], decision trees [57] and genetic algorithms [92] for feature ranking and selection. Based on the rank of features, subset of significant features can be selected for further analysis.

The t-statistic, also known as the student t-test, is a well-known statistical approach frequently applied in microarray data analysis. There are several versions of two sample t-test, depending on whether the sample size is large and whether it is reasonable to assume that gene expression levels have an equal variance under the two conditions [85]. Because usually sample sizes are small and there is evidence to support unequal variance [86], microarray analysis uses t-test with two independent normal samples with unequal variance.

4.2 Cluster Analysis of Gene Expression Profiles

Genes that are involved in correlated functions tend to yield similar expression patterns in microarray experiments. Analyzing these data and

learning their expression pattern can therefore reveal functional association of genes. Clustering techniques are typically used to group genes with similar expression patterns based on the organization of expression data. It generally aims to identify clusters in the data based on the similarity between genes. When there is no or little priori knowledge about the data, clustering is an appropriate tool to analyze data. Clustering can be performed on both dimensions of the expression data matrix, cluster genes or samples. From the machine learning perspective cluster analysis is unsupervised since there is no desired outcome for any particular gene or experiment. From the data mining perspective, the technique is an exploratory data analysis method.

The successful clustering results may provide researchers crucial information regarding the biological role of unknown genes. This is based on the fact that genes which show similar expression patterns (co-expressed genes) are often functionally related and they share the same regulatory mechanism at the sequence level. If a novel gene of unknown function falls into cluster containing genes with known (or partially known) functionality, it is likely that this gene serves the same functions as the other members of the cluster [93]. Since co-expressed genes have a high probability to participate in the same pathway, clustering will provide crucial information for the inference of regulatory information.

The widely adopted clustering techniques for gene expression data include hierarchical clustering [66], self-organizing maps (SOMs) [94, 63], k-means clustering [95, 96], Fuzzy C-means (FCM) [97] etc.

4.2.1 Hierarchical Agglomerative Clustering Algorithm (HAC)

The HAC algorithm is one of the earliest methods used to cluster gene expression pattern [66]. HAC successively merges of clusters until all elements belong to the same cluster from an initial partition. It connects objects to form clusters based on their distance. The clustering result of HAC can be represented by a tree structure called dendrogram, where the leaves represent individual gene patterns and the internal nodes represent clusters of similar patterns. The distance at which the two clusters merge (a measure of dissimilarity between clusters) is called the threshold distance, which is measured by the height of the node from the leaf.

Based on the calculation of similarity among the non-singleton clusters, a variety of hierarchical agglomerative techniques have been proposed. Single linkage, complete linkage, and group average linkage clustering are commonly used.

- **Single Linkage**

In single linkage (also known as nearest neighbor), the distance between two clusters is defined as the distance between the closest pair of objects that are in different clusters. In other words, the distance between two clusters is given by the length of the shortest path between the clusters.

- ***Complete Linkage***

In the complete linkage method (maximum or furthest-neighbor method), the distance between two clusters is given by the value of the longest link between the clusters.

- ***Average Linkage***

Here, the distance between two clusters is computed as the average of distances between all pairs of objects, one in each cluster.

- ***Wards Method***

In this method, cluster membership is assessed by calculating the total sum of squared deviations from the mean of the cluster. The clusters arejoined in such a manner that it produces the smallest possible increase in the sum of squared errors.

Eisen*et. al.* [66] used the average-link HAC algorithm to analyze the yeast cell cycle microarray expression profiles 2467 genes measured over 80 samples to discover indications of the status of cellular process. They also proposed the usage of a colored matrix to visualize the clustering result which presents anatural understanding of the clustering result of the data set. Hierarchical clustering has been used by [98, 99] for the classification of cell line, especially human cancers. They clustered the dataset on other dimension (samples) in an attempt to find new possible tumor subclasses.

Although the HAC algorithm has been widely used for clustering gene microarray expression profiles, it has several drawbacks. First, as Tamayo *et. al.* [63] have noted, HAC suffers from a lack of robustness when dealing with data containing noise, and a preprocessing data is needed to filter out noise. Second, hierarchical clustering is expensive in terms of their computational time and storage requirements when dealing with larger data. Third, since HAC is unable to reevaluate the results, some clusters of patterns are based on local decisions that will produce patterns which are

difficult to interpret when HAC is applied to a large array of data. Fourth, the number of clusters is decided by cutting the tree structure at a certain level. Biological knowledge may be needed to determine the cut.

4.2.2 K-means Clustering

The k-means algorithm is a typical partitioning based clustering technique. Given a set of N objects, a partitioning method constructs k partitions of data, where each partition represents a cluster and k ≤N. The k-means algorithm begins by k randomly selected data objects as cluster centroid. These centroids should be placed far away from oneanother as much as possible to get a better result. Next, each data object is assigned to the cluster with closest centroid. Then the centroid of each cluster is recalculated as the mean of all data objects belonging to the cluster. This process iterates until no more changes occur, or the amount of change fall below a pre-defined threshold.

The k-means method is one of the most popular methods used in DNA microarray data analysis due to its high computational performance. Tavazoie*et. al.* [100] selected the most highly expressed 3000 yeast genes from the yeast cell cycle microarray profiles and applied the k-means algorithm to cluster them into 30 clusters. They successfully found that several clusters were significantly enriched with homo functional genes.

However, k-means needs as input parameters the cluster number k and initial centroids. Its result is subject to the initialization process or in other words, different runs of k-means on the same input data might produce different solutions. To avoid the local suboptimal minimum, one should run

the k-means algorithm multiple times with different initializationof centroids, and then choose the one with smallest average dissimilarity. Furthermore, since the number of clusters is not known, to determine the proper number of clusters in the dataset, one needs to run the k-means algorithm for a range ofk value to determine the best k value.

4.2.3 Self-Organizing Maps

The self-organizing map (SOM) was developed by Kohonen [94] as a neural network based clustering method. It has been extensively used as a tool for visualizing and interpreting high dimensional data. SOM performs unsupervised learning to produce a lower-dimensional (usually 2D) representation of the input space of the training data set samples and use a neighborhood function to preserve the topological properties of the input space. As with other types of centroid based clustering, the goal of SOM is to find a set of centroids (reference vectors) and assign each object in the dataset to the centroid that provides the best approximation of that object. In neural network terminology, one neuron is associated with each centroid.

Unlike k-means, SOM imposes a topographic ordering on the centroids. During the training process, SOM uses each data point to update the closest centroid and the centroids that are nearby. In this way, SOM produces an ordered set of centroids for any given dataset. In the SOM grid, neighboring centroids will be closely related than those are farther away. This makes it easy to find cluster relationship during the visualization of the SOM. Furthermore, SOM does not keep track of the current cluster membership of an object, and unlike k-means, if an object switches clusters,

there is no explicit update of old cluster centroid. Of course, the old cluster may well be in the neighborhood of the new cluster and thus may be updated for that reason. The processing continues until some predetermined limit is reached or when the centroids are not changing considerably.

Tomayo*et. al*. [63] employed SOM to analyze yeast genes from the SaccharomycesCerevisiae dataset and different human cell culture microarray expression profiles. The SOM was able to identify the predominant gene expression pattern in these microarray expression profiles. Although, SOM proved robust performance, it does not overcome the problems of k-means such as cluster number determination and sub-optimization. Moreover, its convergence is controlled by various user supplied parameters such as learning rate, grid topology of neurons and the neighborhood function.

4.2.4 Fuzzy Clustering

In partitioned clustering algorithms such as k-means or self-organizing maps, each gene belongs to exactly one cluster. However, genes are often highly correlated with the patterns of more than one cluster. Fuzzy clustering appears to be a good candidate to reflect the genes multi-cluster participation since it can assign genes degrees of membership to a cluster. The membership value can vary between zero and one. This feature enables fuzzy clustering to provide more information about the structure of gene expression data.

The most popular fuzzy clustering approach in gene expression analysis is that of Fuzzy C-means (FCM) [97]. It allows one data object to belong to more onecluster at the same time.

It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, \qquad 1 \le m < \infty$$

where m is a real valued number which controls the fuzziness of the resulting clusters, u_{ij} is the degree of membership of gene x_i in the cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when

$$max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \epsilon$$

where$\epsilon$is a termination criterion between 0 and 1, whereas k is the number of the iterations.

Dembele and Kastner [101] applied FCM clustering approach on a variety of dataset comprising the serum data set [102] consist of 517 genes whose expressions vary in response to serum concentration in human

fibroblasts, the yeast dataset [103] and a human cancer data set represents gene expression patterns of 9703 genes in 60 human cancer cell lines [104]. Their analysis proved that the overall expression pattern for a given gene may correspond to the superimposition of distinct patterns, each corresponding to a given mode of regulation.

Besides the predetermination of the number of clusters in the data set, the FCM method requires to choose m, the fuzziness parameter. The optimal values for m vary widely from one data set to another.

Several techniques that improve the classical clustering algorithms has been proposed and presented in section 4.4 along with the contribution of this study presented in chapter 5.

4.2.5 Cluster Validation

For gene expression data, cluster analysis partitions the given dataset into groups of co-expressed genes, groups of samples with a common phenotype, or groups of genes and samples involved in specific biological processes. However, different clustering algorithms, or even a single clustering algorithm using different parameters, generally result in different sets of clusters. Therefore, it is important to compare various clustering results and select the best one that fits dataset. The process of assessing the quality and reliability of the clustering results obtained from various clustering process is called cluster validation.Validation can be either statistical or biological. Statistical validation can be done by assessing the cluster compactness and separation, by examining the predictive power of the clusters, or by testing the robustness of a cluster result against the

addition of noise. A common method to biologically validate cluster outputs is to search for enrichment of functional categories within a cluster.

4.2.5.1 Assessing Cluster Homogeneity and Separation

Since all genes within a cluster are expected to be tightly co-expressed, a clustering result can be considered reliable if the within-cluster distance is small and the cluster has an average profile well delineated from the remainder of the data set (maximal inter-cluster distance). Such criteria can be formalized in several ways, such as the silhouette coefficients [105], or Dunn's validity index [106]. Both methods can be considered as a stand-alone tool to compare cluster results. Since these statistics only require a clustered sample and a set of dissimilarities, they can be found using any clustering method and any distance metric. Dunn's validity index (D) recognizescompact and well separatedclusters. The objective is to maximize the intercluster distances and minimize the intracluster distances. Therefore, the number of cluster that maximizes*D* is taken as the optimal number of the clusters.

Whereas, the silhouette of a gene measures how well matched it is to the other objects in its own cluster versus how well matched it would be if it were moved to another cluster. It is a composite index reflecting the compactness and separation of the clusters. When all the elements are best classified, the silhouettes should all be close to 1, and the average silhouette width will be high. Therefore, the highest average silhouette width gives the strongest clustering structure, resulting in an output, which gives the optimal number of clusters.

4.2.5.2 Figure of Merit

The figure of merit (FOM) [107] compares the output of different clustering algorithms in terms of predictive power and homogeneity. This methodology is a combination of leave-one-out cross validation (LOOCV) and the Jackknife approach. The clustering algorithm is applied to all experimental conditions except for one left-out condition. If the algorithm performs well it is expected that the values of genes from a given cluster are highly coherent with those from the left out condition. Therefore, the FOM for a clustering result is computed by finding the root mean square deviation in the left-out condition of the individual gene expression levels relative to their cluster mean. The FOM estimates the within cluster similarity of the expression values of the removed experiment and therefore reflects the prediction power of the clustering. It is expected that removing one experiment from the data should not affect the cluster output if the output is robust. For cluster validation, each condition is used as a validation condition and the aggregate FOM over all conditions is used to compare cluster algorithms.

4.2.5.3 Cluster Sensitivity

Expression level of genes in the microarray dataset reflects the superposition of real biological signals and experimental errors. The confidence in cluster membership of a gene can be assessed by creating artificial datasets by adding to the original data a small amount of artificial noise (similar to the experimental noise in the data).Clustering is subsequently preformed on the artificial data. If the biological signal is stronger than the experimental noise in the measurement of a particular

gene, adding small artificial noise (in the range of experimental noise) to the expression profile of this gene will not drastically influence its overall profile and therefore will not affect the cluster membership. Thus, it is clear that the cluster membership of that particular gene is robust with respect to the sensitivity analysis and reliable confidence can be assigned to the clustering result of that gene. However, for genes with low signal-to-noise ratio, the clustering result will be more sensitive to adding artificial variants. Through some robust statistics [108], sensitivity analysis allows us to detect which clusters are robust within the range of experimental noise and therefore trustworthy for further analysis.

The key issue in the validation method is the determination of noise level for sensitivity analysis. Bittner *et. al.* [108] perturb the data by adding random Gaussian noise with zero mean and standard deviation that is estimated as the median standard deviation for the log ratios for all genes across the experiments. The bootstrap analysis methods proposed by Kerr and Churchill [109] uses the residual values of a linear analysis of variance (ANOVA) model as an estimate of the measurement error. The residuals are subsequently used to generate new replicates of the data set by bootstrapping (adding residual noise to estimated values).

4.2.5.4 Biological Significance Based on p-value

One way to biologically validate results from clustering algorithm is to compare the gene clusters with existing functional classification annotations from various ontology databases. In such databases, genes are assigned to one or more functional categories representing their biological functions, biochemical properties and so on. Finding clusters enriched

bygenes with similar function is a proof that a specific clustering technique produces biologically relevant results.

For each cluster, using hyper geometric probability distribution, a p-value can be calculated, which is the probability of observing the frequency of genes in a particular functional category in a certain cluster. The p-value of observing k genes from a functional category within a cluster of size n is

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}} = \sum_{i=k}^{\min(n,f)} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$

where f is the total number of genes within that functional category and g is the total number of genes within the genome.The p-value can be calculated for each functional category in each cluster. The lower the p-value is, higher the biological significance of a cluster.

In addition to the validity measures mentioned, combined clustering techniques are also used to validate the clustering results. This method sometimes provides clustering output which matches better with the real biological classes (prior knowledge).

4.3 Inference of Gene Regulatory Networks

The success of genome sequencing projects has enabled the biologist to identify almost all the genes responsible for the biological complexity of several organisms. The next important task is to understand the complex interactions among genes and gene products to carry out specific cell functions. Since 1960, the methods from mathematics and physics have been used to describe and simulate small gene networks more stringently.

Nowadays, biological methods and high-throughput experimental technologies make it possible to study a large number of genes and proteins in parallel enabling the inference of larger gene networks. This allows simulating regulatory networks more efficiently and has led to a new discipline called Systems Biology, which combines methods from biology with methods from mathematics, physics and engineering to describe biological systems.

A number of different approaches to gene regulatory network modeling based on large scale microarray data have been introduced, including linear models [110] Bayesian networks [11, 111], graphical model [112], neural networks [69, 113, 114], differential equations [115, 116], and models including stochastic components on the molecular level [117]. The models can be static or dynamic, continuous or discrete, linear or nonlinear, deterministic or stochastic. By analyzing the data, an appropriate learning technique has to be chosen for each model to find the best fitting network structure and parameters. Following sections present some of the popular inference models for GRN and highlight the advantages and limitations of each model.

4.3.1 Boolean Network

Boolean network (BN) model, introduced by Kauffman [118, 119, 120], is a simple computational model that may provide insight into the overall behavior of genetic networks. The main objectives of boolean network modelling is to study generic coarse- grained properties of large genetic networks and the logical interactions of genes, without knowing specific

quantitative details [121]. The biological basis for the development of Boolean networks as models of gene regulatory network lies on the observation that during the regulation of its functional states the cell often exhibits switch-like behavior. In Boolean network gene expression is quantized to only two levels: ON and OFF, which are represented as "activated" and "inhibited". The interactions between the genes are represented by Boolean functions, which determine the state of a gene on the basis of the states of some other genes. The figure 4.1 illustrates the effect of A, B, C and D on F in the form of directed graph and figure 4.2 shows the logic diagram of the activity on F as a Boolean function of 4 input variables.

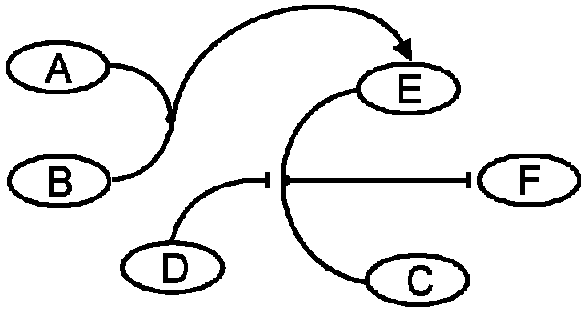


Figure 4.1A directed graph illustrating a hypothetical gene regulatory network. Arrowed lines represent activation and lines with bars in the end represent inhibition.

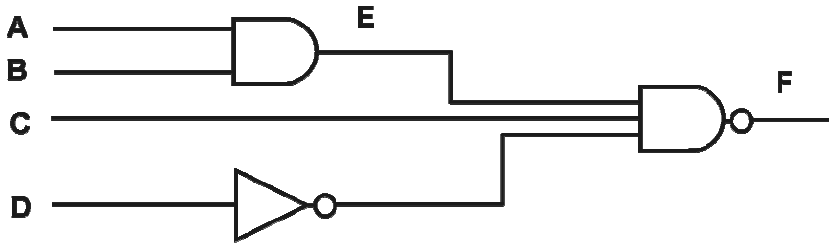


Figure 4.2 The logic diagram describing the activity of F in terms of 4 inputs A, B, C and D. The gates connecting A and B is an AND gate, the gate with input D is a NOT gate and the gate output is F is a NAND gate

A Boolean network G (V,B) is defined by a set of nodes (variables) representing genes V = $\{x_1,...,x_n\}$ and a set of Boolean functions B=$\{f_1,...,f_n\}$.Each x_i represents the state of gene i,where x_i=1 represents the fact that the gene i is expressed and x_i=0 means it is not expressed. Each Boolean function f ($x_{j1}(i)$,$x_{j2}(i)$..., $x_{jk}(i)$) with k specific input nodes is assigned to node x_i and is used to update its value. The gene state at time point t+1 is determined by the values of some other genes at previous time point t using one Boolean function f taken from a set of Boolean functions B. The values of all the nodes in V are then updated synchronously.

Boolean networks have proved successful in modelling real world regulatory networks [122, 123]. One of the appealing properties of BNs is that they are inherently simple, emphasizing generic network behavior rather than quantitative biochemical details, but are able to capture much of the complex dynamics of gene regulatory networks. However, their application in practice is hindered by a number of shortcomings. In particular, analysis can be problematic due to the exponential growth in Boolean states and the lack of tool support in this area. They are also unable to cope with the inconsistent and incomplete regulatory network data that often occurs in practice. Since Boolean network has no intermediate expression level, it cannot capture the details of the biochemical reactions involved in cellular processes. However, it is not the binary nature of the Boolean network model that is its greatest weakness, one even more important deficiency is its determinism. Deterministic models, such as the Boolean network, cannot represent the consequential perturbations due to external latent variables. In addition, the Boolean network model in its original formulation cannot be used to represent

biologically meaningful events, such as gene mutations. The stochastic extension of Boolean network - probabilistic Boolean network (PBN), accounts for those latent variables and gene perturbations while keeping the Boolean logic as the model for the gene-gene interactions [124].

4.3.2 Bayesian Network

Bayesian networks provide a language for representing joint probability distributions of many random variables. Bayesian networks have been applied extensively for modeling complex domains in different fields [125]. This success is due to both the flexibility of the models and the naturalness of incorporating expert (or prior) knowledge into the domain. Another major advantage of Bayesian model is the ability to learn from observed data. This is particularly important when the knowledge about the domain is partial; as is the case in the biological domains. Although the Bayesian network is a discrete variable model, the variables can also be continuous.

In the Bayesian Network formalism [11], the structure of GRN is modeled by a directed acyclic graph that explicitly establishes probabilistic relationships between network nodes. It is represented by a tuple $G = \langle V, E \rangle$ where V is the set of vertices corresponds to the random variables X_1, $X_2 \ldots X_n$ (X_i describes the expression level of gene i) and E describes the set of conditional distributions. For each X_i, a conditional distribution $p(X_i|$ parents$(X_i))$ is defined, where parents (X_i) denotes the variables corresponding to the direct regulators of i in G. The graph G and the

conditional distributions $p(X_i|\ parents(X_i))$ encode a unique joint probability distribution p(X), thus defining a Bayesian network

The graph encodes a set of independence statements of the form: for every gene in G, $I(X_i;\ nondescendants\ (X_i)\ |\ parents(X_i))$, that is, every node is independent of its nondescendants given its parents in graph G. By means of the Markov assumption, the joint probability distribution over X can be written as

$$p(X) = \prod_{i=1}^{n} p(X_i|parents(X_i))$$

Hence two Bayesian networks are said to be equivalent if they imply the same set of independencies among variables. Although Bayesian network is effective in dealing with noise, incompleteness and stochastic aspects of gene regulation, they fail to consider temporal dynamic aspects of gene regulation that are an important part of GRN modelling [11]. To effectively deal with the dynamic process of regulatory network, Dynamic Bayesian networks (DBN) were developed, and can yield more accurate models. However, their benefits are hindered by the high computational cost required in the cases where large numbers of genes are involved [126]. For this purpose, some supplementary methods, such as network decomposition (for dimension reduction), and Monte Carlo strategies using random sampling have been developed to enhance performance [127, 128, 129].

4.3.3 Differential Equations

Unlike the above discrete models, the models based on differential equations uses continuous variables. Several models based on differential

equations have been used to explore genetic networks, of which some uses linear ordinary differential equations (ODEs) and others nonlinear power law differential equations [130, 131, 132, 133, 134]. In general, the change in the expression level of a gene at a certain time (discrete or continuous) is characterized by rate equation that takes the regulatory influence (activation or inhibition) of other genes into account. The rate equations have the mathematical form

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \dots, x_n, p, u)$$

where x_i $(1 \leq i \leq n)$ is the expression level of gene i at time t, n is the number of genes, p is parameter set of the system and u is a set of transcriptional perturbation. The function f_i can be linear, piecewise linear, pseudo-linear, or continuously nonlinear, each describing the system dynamics with a different level of complexity.

Compared to the discrete variable models, the differential equation models can more accurately represent the system dynamics of Gene Networks by virtue of their use of continuous variables. In particular, with nonlinear ODE models (such as S-system models), a given system's steady-state evaluation, control analysis and sensitivity analysis can be established mathematically [135]. D'Haeseleer*et. al.* [136] and von Dassow*et. al.* [137] successfully applied differential equation to the modeling of regulatory networks in *Drosophila*, though the use of linear models or linearization of non-linear models does not limit the generality of methods and results.

Although the differential equations models are effective in reproducing the characteristics of the target system accurately, the benefit

comes at a significant cost: an increase in computational expense. Such expense can make certain system sizes too unwieldy to be represented by anything but the simplest of differential equation models, which may lead to lower likelihood of accurately modeling the underlying physical phenomena. Also, these models depend on numerical parameters that are often difficult to establish experimentally. It must be also noted that the modeling is very sensitive to noisy and imprecise data (as in the case of data resulting from microarray experiments) [138].

4.3.4 Neural Networks

The other commonly used type of continuous variable model to represent the dynamics of gene regulation is the neural network model. Each neuron represents genes and the connection between nodes represents the regulatory actions of one gene over other. The influence of one gene product on the expression level of other genes of the system is defined by a weight matrix. Each layer of the network represents the expression level of genes at time t. The expression level of a gene at time t+1 can be derived from the expression levels at the time t (y_j) and connection weights (w_{ij}) of all the genes connected to it. In other words, the net regulatory effect on a particular gene can be regarded as a weighted sum of expression level of all other genes capable of regulating it. Thus g_i, a regulatory effect to gene *i*, is

$$g_i \approx \sum_j w_{ij}.y_j$$

This regulatory effect is transformed by an activation function to the interval <0, 1>.For sigmoid activation function regulatory effect is defined as

$$g_i = \left\{1 + \exp\left[-(\sum_j w_{ij}.y_j + b_i)\right]\right\}^{-1}$$

where$b_i$represents an external input that can be interpreted as a reaction delay parameter. High positive or negative values of b result in a low influence of factors given in the weight matrix. In principle, all genes can regulate all others (fully connected network); in reality, only a few genes control the activity of one particular gene. The state of a gene in the network is updated in a stepwise manner; the state of each gene in the next time point$t_{i+1}$is determined by the state at timet_i.

The most successful neural network model is recurrent neural model (RNN) [139]. This model is biologically plausible and noise resistant. By adapting self-loops and feedback connections to their structure, recurrent networks can deal with temporal and spatial/temporal problems, both of which are used to memorize past information. In addition, its nonlinear characteristics provide information about the principles of control, as well as about the natural interactions of elements of the modeled system. There are several RNN architectures for GRN modelling ranging from restricted classes of feedback to full interconnection between nodes [140, 141, 142]. Greedy algorithms based on the gradient descent method, such as back-propagation through time (BPTT) [143], have been developed to efficiently update the relevant parameters of recurrent networks in discrete time steps. Recently, Xu*et. al.* [140] used an RNN combined with Particle Swarm Optimization (PSO) to capture the complex nonlinear dynamics of gene

regulatory networks. However due to its training time and computational complexity, this modelling can currently applied to only small systems.

4.3.5 Other Inference Approaches

Recent approaches tried to overcome the drawback of traditional methods in several ways [144]. Woolf and Wang [13] applied fuzzy rules to every possible combination of genes to find the activator/repressor relationship in a normalized subset of saccharomyces cervisae data. Although their method is intuitive, the results are consistent with the literature on genetic networks of Saccharomyces Cervisae. However this approach is slow and computationally complex. Keedwell and Narayanan [145] proposed a hybrid neuro-genetic algorithm to extract regulatory relationship among genes. Their method integrates genetic algorithms with a single layer artificial neural-network, where each chromosome of the GA selects a small number of regulating genesfrom the whole data set and the neural network identifies the regulatory relations among genes. However, when modelling the complex temporal dynamics of gene expression regulation, the lack of a recurrent structure and the proper training method of the ANN may pose serious problems. Also, the method is vulnerable to local minima traps.

Although RNNs are better candidates in dealing with temporal sequence production problems, it has been proved that recurrent neural fuzzy networks outperform recurrent neural networks [146, 147] in problems that involve concurrent spatial and temporal mapping like the one of regulatory networks reconstruction. Additionally, due to the high level

human like reasoning fuzzy system, fuzzy-based approaches are efficient in handling the uncertainties of modelling noisy data [148, 149]. In [148] a fuzzy data mining approach has been proposed to measure the statistically significant fuzzy dependency relationships among genes. Sokhansanj*et. al.* [149] demonstrated an approach with exhaustive search for possible rules describing gene interactions, under the framework of a linear fuzzy logic scheme that restricts the search space. However, both methods requires prior data discretization, while [149] has the additional disadvantage of not considering temporal information.

…….ଓ…….

Chapter 5

Preprocessing of Plasma RNA Dataset

Contents

Gene expression profiles obtained from microarray experiments contain information about many different aspects of gene regulation and function. As already described in the previous chapters, expression data sets are large and complex, having large number of features (genes) and unknown internal structure. In order to decipher biological information embedded in such data sets, first step is data preprocessing. A critical issue in preprocessing technique is selecting an effective and more representative feature set. The feature selection is a useful preprocessing step used to reduce dimension and improve classification accuracy. Consequently, cluster analysis can be done on the expression profile of these "representative" genes for obtaining an initial understanding of data, usually done for class discovery. This chapter presents an approach used to perform exploratory analysis of gene expression data sets.

5.1 Methods for Data Analysis

5.1.1 Dataset

The microarray dataset studied in this work is the circulating plasma RNA dataset of colorectal cancer patients, which contains 20 samples

collected from CRC patients. Among them 12 samples are from colon tumors and 8 are from normal biopsies. The dataset consists of the expression profiles of 15552 genes obtained by measuring the relative abundance of the different RNA species in plasma through cDNA microarray hybridization, by comparing RNA isolated and amplified from colorectal cancer patients and from healthy donors.Each sample was competitively hybridized against a common reference formed by a pool of blood samples from 26 healthy donors [16]. The full data set can be downloaded from the Gene Expression Omnibus Web site at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4988

5.1.2 Data Filtering

The major challenge of microarray data analysis is the comparison of large number of genes to smaller number of samples in a typical experiment. The cost and time required are linearly proportional to the sample size. For the data obtained in a specific experiment, only few genes will be useful to differentiate samples among different classes, while many other genes are irrelevant to this task. Those irrelevant genes not only increase the dimensionality, but also introduce unnecessary noise, which results computational difficulties in microarray data analysis. Elimination of those “probable noise” genes and the identification of informative genes is a feature selection problem, which is crucial in microarray data analysis. When a small number of genes are selected, their biological relations with the target gene can be more easily identified. These “marker” genes thus provide an additional understanding of the problem.

In this work a parametric method, unpaired t-test was used to select the differentially expressed genes. The t-test is used to compare the means of two samples from a normal distribution, when the standard deviations are unknown but assumed equal. The hypotheses for the comparison of two independent groups are:

H_0 : $\mu 1 = \mu 2$ (means of the two groups are equal)

H_a : $\mu 1 <> \mu 2$ (means of the two groups are not equal)

The p-value is the probability of error in rejecting the null hypothesis (H_0) of no difference between the two groups. A low p-value means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis. If the p-value is less than a critical value (less than 0.005 for example), null hypothesis will be rejected and the gene is differentially expressed.

5.1.3 Clustering of Datasets

One of the major goals of gene expression analysis is to unravel the interaction structure of genes involved in the cellular process. This task is difficult because the genes that are co-regulated or co-expressed under a subset of conditions will behave differently under other conditions. Finding genetic pathways therefore could be aided by identifying clusters of genes that are co-expressed under subsets of conditions as opposed to all conditions. A high degree of correlation among the expression levels of subsets of genes under different conditions does not of course necessarily imply causality relations. Further analysis would be required to find actual genetic pathways.

The algorithm used in this work is based on the two popular clustering algorithms- agglomerative hierarchical clustering algorithm and k-means clustering algorithm. Although these methods were mostly developed outside biological research, they still revealed biologically relevant information. Both have several properties which render them as a basis framework for building more advanced algorithms for clustering applications. They can be implemented easily, are fast, robust, and scale well to large data sets.

However, there are some characteristics that require improvement. Perhaps the strongest impact on the results of k-means is the necessity of specifying the expected number of clusters and the initial centroids. For large datasets, number of clusters is almost impossible to predict in advance [150]. In the case of hierarchical clustering major drawback is the degradation of quality of cluster as more data are joined. Moreover, both algorithms are sensitive to noise and outliers.

5.1.4 Hybrid Clustering Algorithm

In hybrid approach only the strengths of the above clustering algorithms was accepted and discarded their drawbacks. The above classic clustering algorithms were combined in order to account for the following peculiarities of gene expression data, which could not be handled by the basic algorithms:

- The number of gene clusters is unknown, although the k-means requires a predefinition of this number.

- Closely related to using a predefined number of clusters is that even genes which are not really co-expressed with other cluster members (and therefore represent noise) are forced to end up in one the clusters and by that hamper the further analysis of the clusters.
- The outcome of the clustering with the k-means is sensitive to the choice of initial partition. Under some initialization condition the solution may trapped in a local optimal.
- The number of genes in the different functional categories is unbalanced (e.g. large categories in yeast may contain more than 2200 genes and small ones as few as 4 genes). The k-means prefer clusters of approximately similar size as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters.
- It is well known that microarray data contain a large amount of noise and some outliers because of experimental error. As a small number of such data can substantially influence the mean value, both algorithms are sensitive to noise and outliers.

It is obvious that the above-mentioned problems are not exclusively k-means specific, but affect many other clustering algorithms (for example SOM). Thereby their adjustment improves not only over the basic-k-means algorithm but also over many other existing methods.

In hybrid algorithm, Pearson correlation coefficient was used to compute the distance between two data objects. It indicates both how the two objects are related and the strength of that relationship. For example, if

gene A increases and gene B decreases proportionally, their correlation value will be -1 because they are perfectly divergent. If the two sets were not perfectly divergent, but still diverged, the correlation would remain negative, but would be greater than -1. In contrast, if genes A and B increase proportionally, then their correlation will be 1. If genes A and B have absolutely no relationship to each other whatsoever, their correlation will be 0. The Pearson correlation coefficient is sensitive to not only direction of change (increasing or decreasing), but also to magnitude of change. The correlation can be any value from -1 to 1. It is invariant to scale and location of the data points, unlike Euclidean distance.

A brief description of hybrid clustering algorithm is as follows. First agglomerative hierarchical clustering algorithm was carried out and let the program stop at a terminal point. From the clusters generated, the centroid of each cluster was calculated. These centroids were used as the initial centroids for the k-means algorithm. Also, the number of clusters generated from the hierarchical clustering is the k-mean's number of clusters. The k-means algorithm was executed on the remaining objects which were not processed by hierarchical clustering. In order to deal with the outliers, a threshold distance was set for k-means. If the shortest distance from the object to the centroids exceeds threshold the object is assigned to outlier group otherwise assign it to the closest cluster.

Hybrid Algorithm

The Hybrid algorithm is summarized as follows. Inputs are data, termination percentage and outlier threshold.

Step 1: Compute the distance between all pairs of objects.

Step 2: Select two objects having closest distance among all and combine them into a new cluster

Step 3: Update the attribute of the new cluster as the average of the attribute of the two objects.

Step 4: Perform steps 2 and 3 iteratively until the termination percentage given by the user.

Step 5: Compute the centroids of the clusters generated from steps 1 to 4.

Step 6: Find the objects that are not yet the members of any cluster. For all such objects repeat the following steps

Step 7: Find the distance between the object and the centroid of existing clusters.

Step 8: If the shortest distance is greater than the threshold distance the object is assigned to the outlier group. Otherwise the object is assigned to the closest cluster, and the cluster centroid is updated accordingly.

Step 9: Repeat steps 6, 7 and 8 until no object changes belonging cluster.

5.2 Results

Each sample in the dataset is described by 15552 genes. Since there are 20 samples, each gene has 20 dimensional data. To identify the genes with differential expression under diseased and normal conditions, t-test was applied on the whole dataset. Among the 15552 genes, 100 genes that are

significant at 5% level and whose p-values less than 0.0057 were selected for clustering and analysis. Next, these 100 genes were clustered using hybrid clustering technique. The correlation distance 0.5 was set as the outlier threshold to filter out the minority group.

In order to identify the optimal termination percentage, its effect on hybrid clustering algorithm was investigated. For that, the silhouette value, which is a composite index reflecting the compactness and separation of clusters, was used to judge the quality of the clusters generated. The figure 5.1 shows the silhouette value of the clusters generated under various termination percentages. From the figure it is clear that the hybrid clustering approach can generate better result when hierarchical clustering was terminated at around 70 to 75%. The best silhouette value of 0.5753 was obtained when the termination percentage was set to 71%. With this setting hybrid clustering algorithm generated 4 clusters and identified 6 outliers from 100 genes. It was found that the quality of the clusters generated is poor if the hierarchical clustering terminates too early or too late.

In order to compare the quality of clusters obtained, the same data set was clustered using k-means and hierarchical clustering. Table 5.1 shows the clustering results at the number of clusters 4 using hybrid, k-means and hierarchical clustering algorithm. Table 5.2 shows the silhouette value of these 3 clustering methods. The values show that the hybrid method is more superior to the other two clustering methods.

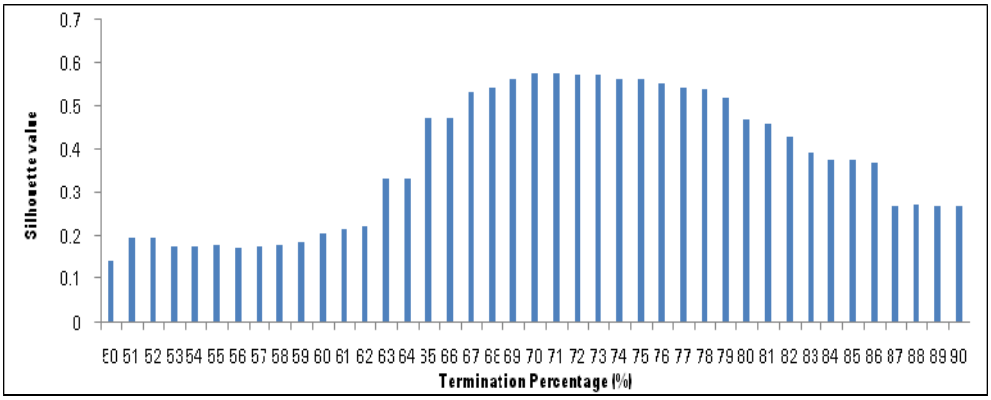


Figure 5.1 Effect of Termination Percentage of Hierarchical Clustering on the Quality of Clusters generated by Hybrid Clustering Algorithm. Outlier threshold = 0.5.

Table 5.1 Clustering Results Obtained Using Hybrid, K-means and Hierarchical Clustering Algorithms

Hybrid Clustering		K-means Clustering		Hierarchical Clustering	
Cluster	***Gene Symbol***	***Cluster***	***Gene Symbol***	***Cluster***	***Gene Symbol***
1	KIAA0801	1	KIAA0801	1	KIAA0801
	HBE1		HBE1		HBE1
	CREM		CREM		CREM
	HNRPH3		HNRPH3		HNRPH3
	PSMA3		PSMA3		PSMA3
	LOC51194		LOC51194		LOC51194
	GLRX		GLRX		GLRX
	FLJ20701		FLJ20701		FLJ20701
	LDHA		LDHA		LDHA
	OXA1L		OXA1L		OXA1L
	COX11		COX11		COX11
	ACP2		ACP2		ACP2
	CBX1		CBX1		CBX1
	ALS2CR3		ALS2CR3		ALS2CR3
	PCP4		PCP4		PCP4
	EPAS1		EPAS1		EPAS1
	DNCLI2		DNCLI2		DNCLI2
	CEP3		CEP3		CEP3
	USP9X		USP9X		USP9X
	RPL10A		RPL10A		RPL10A
	CD83		CD83		CD83
	KIAA0417		KIAA0417		KIAA0417

Hybrid Clustering		K-means Clustering		Hierarchical Clustering	
Cluster	*Gene Symbol*	*Cluster*	*Gene Symbol*	*Cluster*	*Gene Symbol*
	DGKZ		DGKZ		DGKZ
	KIAA0632		KIAA0632		KIAA0632
	LOC51023		LOC51023		LOC51023
	DKFZP434N061		DKFZP434N061		DKFZP434N061
	DKFZP564B167		DKFZP564B167		DKFZP564B167
	LOC51002		LOC51002		LOC51002
	BTF3		BTF3		BTF3
	STC1		STC1		STC1
	UBE2D3		UBE2D3		UBE2D3
	CAV2		CAV2		CAV2
	ANXA1		ANXA1		ANXA1
	ST16		ST16		ST16
	LOC51632		LOC51632		LOC51632
	PIK3C2G		PIK3C2G		PIK3C2G
	NR3C1		NR3C1		NR3C1
	SMN2		SMN2		SMN2
	SFTPA2		SFTPA2		SFTPA2
	KIAA0041		KIAA0041		KIAA0041
	DSCR5		DSCR5		DSCR5
	LOC50999		LOC50999		LOC50999
	E46L		E46L		E46L
	XRCC5		XRCC5		XRCC5
	ACP1		ACP1		ACP1
	PCBP2		PCBP2		PCBP2
	CAPZA1		CAPZA1		CAPZA1
	EIF3S6		EIF3S6		EIF3S6
	ATP5G3		ATP5G3		ATP5G3
	HLF		SMARCA5		SMARCA5
	DECR1		HLF		HLF
	HGS		DECR1		DECR1
	IL7R		HGS		HGS
	MLLT4		MAN2A1		MAN2A1
	MVP		IL7R		IL7R
	DNM2		MLLT4		MLLT4
	TIM17		MVP		MVP
	FLJ12910		DNM2		DNM2
	NOLA2		TIM17		TIM17
	PPP1CC		FLJ12910		FLJ12910
	RPC		NOLA2		NOLA2

Hybrid Clustering		K-means Clustering		Hierarchical Clustering	
Cluster	*Gene Symbol*	*Cluster*	*Gene Symbol*	*Cluster*	*Gene Symbol*
	FABP5		PPP1CC		PPP1CC
	SLBP		RPC		RPC
	LPL		FABP5		FABP5
	KIAA0758		C10RF13		C10RF13
			SLBP		SLBP
			LPL		LPL
			KIAA0758		KIAA0758
2	FLJ20037	2	FLJ10849	2	KIAA1288
	KIAA1288		KIAA1288		TRIP10
	TRIP10		TRIP10		RELN
	RELN		RELN		TNKS
	CCNT1		TNKS		RPS6KA1
	AEBP1		BDKRB1		BDKRB1
	CYP8B1		FLJ10287		FLJ10287
	TNKS		DOC1		DOC1
	TIAF1		IGLL1		IGLL1
	DOC1		COX15		DEFCAP
	HRG		DEFCAP		
	COX15				
	DEFCAP				
3	FLJ10849	3	FLJ20037	3	FLJ10849
	ERBB2		BAD		ERP28
	BDKRB1		CCNT1		FLJ20037
	LOC51242		AEBP1		BAD
	FLJ10287		C18B11		CCNT1
	IGLL1		IL2RG		AEBP1
			LOC51242		C18B11
			GCN5L1		IL2RG
			CHK		LOC51242
			HRG		GCN5L1
			SPIB		CHK
			POMT1		HRG
			RPS6KA1		SPIB
4	NMA	4	NMA	4	NMA
	ERP28		ERP28		SP38
	SP38		SP38		KIAA0524
	KIAA0524		KIAA0524		CYP8B1
	BAD		CYP8B1		ERBB2
	C18B11		ERBB2		TIAF1
	IL2RG		TIAF1		TNS
	GCN5L1		TNS		COX15
	CHK				POMT1
	TNS				

Table 5.2 Silhouette Value for 3 Clustering Algorithms

Hybrid Algorithm	K-means Clustering	Hierarchical Clustering
0.5753	0.5101	0.4814

5.3 Discussion

In this chapter explains some preprocessing steps done on the circulating plasma RNA dataset of colorectal cancer patients. By applying t-test on the whole data, 100 differentially expressed genes were identified from the whole dataset. These 100 genes were clustered and verified the advantage of using hybrid algorithm as a clustering algorithm for gene expression data. Comparison of results showed that the clustering results computed by the hybrid algorithm are better than the classic algorithms: k-means and hierarchical clustering algorithm.

Chapter 6

Computational Methods

Contents

Reconstructing and modelling gene regulatory network from experimental data is crucial in the understanding of fundamental cellular processes and disease mechanism. A GRN represents a set of regulatory interactions among genes in a cellular system. These interactions are involved directly or indirectly in controlling the production of gene products such as proteins and in mediating metabolic processes. Exploring GRNs can provide new ideas for treating complex diseases and breakthroughs for designing personalized medicine. In this chapter, four computational approaches applied to the problem of GRN reconstruction such as fuzzy logic approach, modified genetic algorithm, dynamic feed forward neural fuzzy network and TSK-type recurrent neural fuzzy network are presented.

6.1 Fuzzy Logic Approach

Fuzzy logic algorithm provides a systematic and unbiased way to transform precise numbers into qualitative descriptors in a process called fuzzification. When dealing with gene expression data the observed data is broken using fuzzy logic into discrete subsections which provides a

qualitative description of the data. Heuristic rules can be used to analyze this qualitative data, which in turn generate fuzzy solutions. Heuristic solutions can be transformed in to a precise number using defuzzification operation.

There are three main advantages of applying fuzzy logic algorithm for the analysis of gene expression data. This includes

1. Fuzzy Logic inherently accounts for the noise in the data because it extracts trends, not precise values.
2. The algorithms in fuzzy logic make use of high level human like reasoning and are cast in the same language used in day-to-day conversation. As a result, predictions made by this algorithm can be easily interpretable.
3. Fuzzy logic approaches are computationally efficient and can be scaled to include an unlimited number of components. Thus a large number of biologically relevant patterns can be recognized using this approach.

Wolf and Wang [13] introduced fuzzy logic approach to search for the activator-repressor-target relationship of gene interaction from a normalized subset of Saccharomyces Cerevisiae data [103].They applied fuzzy rules to every possible activator-repressor combination of genes and the output of the model was compared to the expression levels of the remaining genes. They concluded that those combinations of genes having low error are most likely to exhibit an activator/ repressor relationship. The relations predicted by this approach agree well with the experimental results from the literature. But this algorithm suffered from high computational

time requiring 200 hours to analyze the relationship between 1898 genes [151]. In this work, fuzzy logic algorithm has been improved to infer causal relationship between genes by analyzing gene expression data. In the proposed model, clustering is used as the preprocessing step to remove the redundant computation performed by the model. Clustering will group the genes based on their similarity in the changes of expression over all samples available in the microarray data. Replacing the group with a reference expression profile, when executing fuzzy logic algorithm, will greatly reduce the computational cost.

6.1.1 The Fuzzy Logic Algorithm

The fuzzy logic algorithm [13] identifies the regulatory triplets consisting of activator, repressor and target genes by searching the microarray data set. The algorithm uses fuzzy logic to transform the gene expression levels in to qualitative descriptors such as HIGH, MEDIUM and LOW states. The Zadeh – Mamdani's fuzzy model with two inputs and one output has been adopted to realize fuzzy inference and fuzzy IF-THEN rules are used to perform a map from the input space to the output space. The inputs are the normalized expression value of genes (A and B) paired into activator and repressor. The first step is to take the inputs and determine the degree to which they belong to each of the appropriate fuzzy sets via membership function of the triangular form shown in figure 6.1. Next, these fuzzified expression values of A and B are evaluated against a set of heuristic rules summarized in the decision matrix shown in Table 6.1. The output of the fuzzy system is the fuzzified value of the predicted target C, which encompasses a range of output values, and so it must be defuzzified in order to resolve into a single output value. The predicted expression

values of C for each sample were calculated in the similar manner. The mean square error (MSE) between the entire series of predicted values and target gene's actual expression values were calculated. Those triplets that have low MSE value may exhibit the activator-repressor-target relationship.

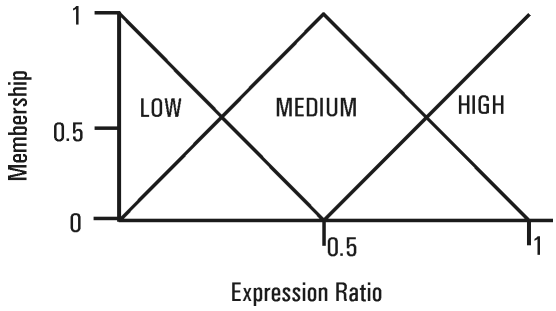


Figure 6.1 Membership Function

Table 6.1 Decision Matrix

		if repressor is		
		HIGH	**MED**	**LOW**
if activator is	LOW	Target is LOW	Target is LOW	Target is MED
	MED	Target is LOW	Target is MED	Target is HIGH
	HIGH	Target is MED	Target is HIGH	Target is HIGH

The second score is calculated based on the exploration of decision matrix. If the expression value of gene A is always high and that of gene B is always low, the dataset cannot properly explore the fuzzy rules. A hit ratio is defined as the ratio of the number of rules used to the total number of rules. A high hit ratio value implies that all of the rules are used equally in generating the output and the resulting predictions are more likely to be credible. The low value implies that all the rules have not been tested and the predictions may or may not be true. In this study only those gene pairs

with hit ratio greater than 0.8 were considered for further analysis. The algorithm implemented in Matlab requires 232secs seconds to analyze the relationship between 100 genes from Plasma RNA dataset of CRC patients.

6.1.2 Clustering to Improve Run time

Clustering is a data mining process used to reveal natural structures and identify interesting patterns in the gene expression data. Cluster analysis partitions a given set of genes in the dataset into groups so that genes having similar expression profiles form a group. In this work, hybrid clustering algorithm described in chapter 5 has been applied to cluster genes that behave similarly. The figure 6.2 illustrates the use of clustering in modelling the gene regulatory network. For example, top figure shows that an increasing activator and decreasing repressor would cause the target gene to increase quickly. These triplets make sense intuitively and should be included in the analysis. Whereas, in the bottom figure, an increasing activator and decreasing repressor would cause the target gene to decrease. These cluster triplets do not make sense intuitively and should not be included in the analysis. Using the fuzzy logic algorithm, one can easily determine whether or not large group of genes, represented by these cluster centers, are likely to fit the model. The fuzzy logic algorithm is executed using centroid of each cluster as input and the cluster triplets ranked according to how well they fit the model. Based on the knowledge of how cluster centers fit to the model, the total number of gene combination analyzed could be reduced. This will enhance the performance of the above algorithm by reducing the computation time with minimal effect on results.

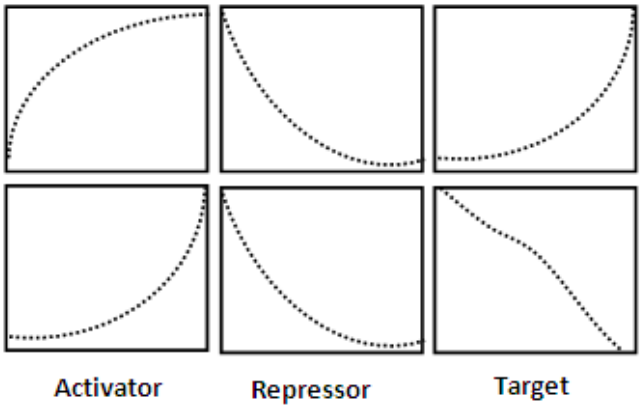
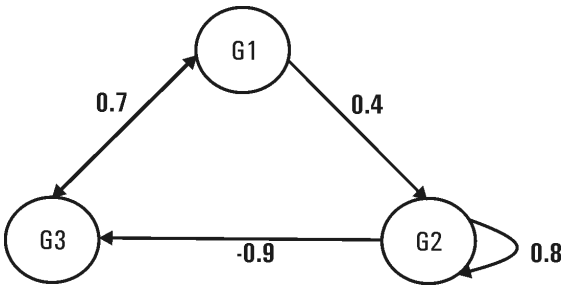


Figure 6.2 Cluster triplet that fit the model well (top) and Cluster triplet that would not fit the model (bottom)

6.2 Modified Genetic Algorithm

Regulatory relationships between genes can be represented as linear coefficients of weights, with the "net" regulation influence on a gene's expression being the mathematical summation of the independent regulatory inputs [113]. This representation is practical for analysis of microarray experiment, which provides snapshots of concentration at various samples. The regulatory networks generated from microarray data with this approach, display stable gene expression levels, consistent with known biological system [113]. In this approach GRN is represented by a weighted graph G = (V, E, W), where V is the set of nodes (genes), E is the set of edges (regulatory relationships) and W is the weight matrix. Figure 6.3 shows an example of a gene network and the corresponding weight matrix. The value of $w_{ij}$is limited to a range between -1 and 1.The positive w_{ij} means gene i activating gene j, as opposed to a negative value representing an inhibition.

Zero indicates no influence. The expression level of a transcriptional regulatory network containing N genes is represented by a vector$\vec{x}$. Each column of weight matrix W represents all regulatory inputs to a gene.



	G1	G2	G3
G1	0	0.4	0
G2	0	0.8	-0.9
G3	0.7	0	0

Figure 6.3A Sample Gene Network and the corresponding Weight Matrix.

The net regulatory input to a gene i, S_i is determined by taking the weighted sum of the expression level of other genes.

$$S_i = \sum_{j=0}^{n} w_{ij} x_j$$

The response of gene i to the regulatory input is defined using the sigmoid function as

$$x_i = \frac{m_i}{1 + e^{-s_i}}$$

where m_i is the maximal expression level.

6.2.1 Genetic Algorithm Implementation

Genetic algorithms are search procedures, based on the mechanics of natural genetics, able to provide robust search in complex problem spaces. Genetic algorithm was applied for optimizing the weight matrix for gene regulatory network. For the GA implementations, gene networks have to be coded into chromosomes. Each row of the weight matrix is aligned in an array to be the one dimensional real number array chromosome. The fitness function of the GA is defined by the Euclidean error δ between the observed expression pattern and the actual expression pattern of the target gene.

$$\delta = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Since GA is a probabilistic search, several generations and greater computation power were required to model smaller network with good sensitivity and precision. Simple regulatory networks with few numbers of genes can be inferred effectively using GA [14, 152]. However, when applied to large microarray dataset this approach is slow and computationally complex. So, in order to reduce the number of generations, some preprocessing steps were done to initialize the genotypes.

Correlation analysis can be used to capture the regulation and co-regulation among the genes and have been proven useful for identifying biologically relevant groups of genes and samples [15]. High correlation between gene A and B can be caused by (i) A regulates B or vice versa (ii) A and B are co-regulated by other genes (iii) there is no causal relationship just coincidence. Here regulations may be indirect, i.e. interactions through immediates. Pearson's correlation coefficient is widely used to measure the

expression pattern similarity between two genes. Pearson's correlation coefficient considers each gene as a random variable and calculates the similarity of expression patterns by computing the linear relationship between gene expressions.However, experimental study has shown that it is sensitive to outliers and sometimes yields a high similarity score to a pair of dissimilar patterns. If two patterns have a common peak or valley at a single feature, the correlation will be dominated by this feature, although the patterns at the remaining features may be completely dissimilar. This observation evoked an improved measure called Jackknife correlation. Given two data objects O_i,O_j,Jackknife correlation coefficient is defined as Jackknife(O_i, O_j)=min$\{\rho_{ij}(1),\rho_{ij}(2), \ldots, \rho_{ij}(p)\}$ where $\rho_{ij}(k)$ is the Pearson's correlation coefficient of data objects O_i and O_j with the k^{th}feature deleted. Use of the Jackknife correlation avoids the "dominance effect" caused by single outliers.

A Genetic Algorithm operates through a simple cycle of stages as shown in figure 6.4. Each cycle in Genetic Algorithm produces a new generation of possible solutions for a given problem. In the first phase, an initial population, set of individuals each representing a possible solution to a given problem is created to initiate the search process. For that the correlation coefficient of the target gene with every other gene is calculated. If the correlation is significant, a random number is generated to initialize the corresponding data point of each of the genotypes in the population. Based on the value of the fitness function,chromosomes are selected for a subsequent genetic manipulation process. In this work, Roulette Wheel selection algorithm was used to breed a new generation. In Roulette Wheel selection, the fitness level is used to associate a probability of selection with each individual chromosome.In order to enhance the adaptability of the GA

as well as the reverse engineering method, the genetic manipulation process consisting of two steps is carried out. In the first step, the crossover operation that interchanges the bits (genes) of two parent chromosomesat the crossover point is executed. The second step is termed mutation, where the bits at one or more randomly selected positions of the chromosomes are modified to form a new population. The mutation process avoids the algorithm to get trapped at local maxima. The offsprings produced by the genetic manipulation process form the next population to be evaluated. The process is repeated until the specified numbers of generations are completed. Finally, the chromosome with the best fitness score has the weights of the regulators of the target gene. The algorithm is repeated for each of the target gene.

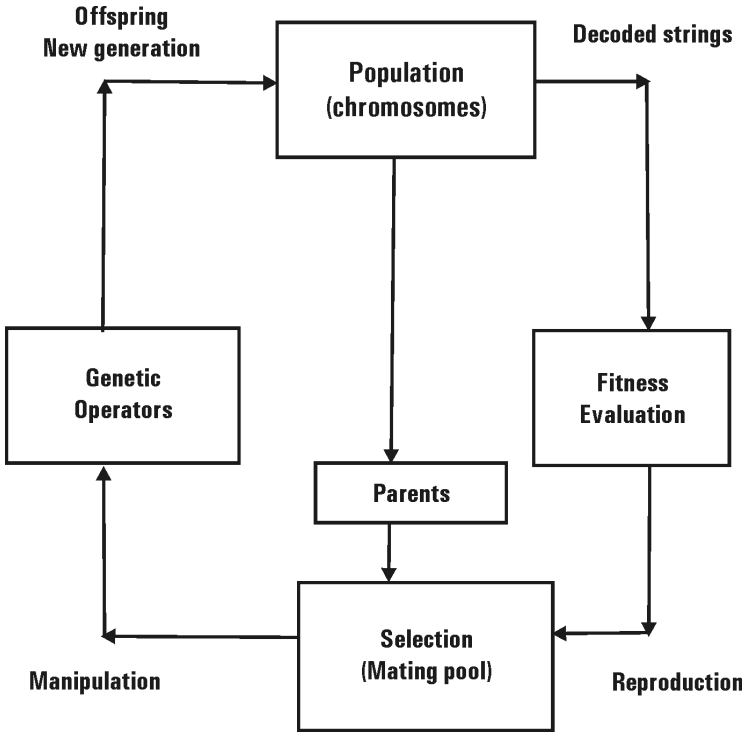


Figure 6.4 Cycle of stages in Genetic Algorithm

By incorporating statistical techniques viz. correlation analysis, while generating initial population, search space can be reduced significantly. This will improve the computational time of modified GA when compared to that of basic GA.

6.3 Dynamic Feed Forward Neural Fuzzy Network

A Neural Fuzzy Network is a learning system that finds the parameters of a fuzzy system (i.e., fuzzy sets, fuzzy rules) by exploiting approximation techniques from neural networks [153]. Both neural networks and fuzzy systems can be used for solving a problem if there does not exist any mathematical model of the given problem. They solely do have certain disadvantages which almost completely disappear by combining both concepts. For example, neural networks can only come into play if the problem is expressed by a sufficient amount of observed examples. These observations are used to train the network. On the contrary, a fuzzy system demands linguistic rules instead of learning examples as prior knowledge. If the knowledge is incomplete or wrong, then the fuzzy system must be tuned in a heuristic way. This is usually very time consuming and error-prone.

In this work, a novel dynamic neural fuzzy hybrid system has been proposed to extract information from microarray data in the form of fuzzy rules. This method combines both the advantages of neural network and fuzzy system; it brings the low-level learning and computational power of neural networks into fuzzy systems and provides the high–level human like reasoning of fuzzy system into neural network. Unlike other neural fuzzy architectures [146], where the network structure is fixed and rules should be

assigned in advance, there is no predetermination of rules; all of them are constructed during online learning.

6.3.1 Dynamic Neural Fuzzy Network Architecture

The dynamic neural fuzzy network (DNFN) adapts the ZadehMamdani's fuzzy model [154] to realize the fuzzy inference and uses fuzzy if-then rules to perform the map from input space to output space. In contrast to other neural fuzzy model, there are no fuzzy rules initially in DNFN. They are generated on-line via learning. There are two learning phase - structural and parameter learning. The structural learning phase is responsible for the construction of dynamic fuzzy if-then rules. The parameter learning phase is for tuning the free parameters such as weights of the rules, which is accomplished through the repeated training of input output patterns. Parameter learning process is done concurrently with the structural learning phase. Owing to its simplicity back propagation algorithm is adopted to adjust the weights of DNFN.

Architecture of DNFN is illustrated in figure 6.5. The network consists of 5 layers.

The detailed functions of each layer are as follows.

Layer 1(Input Layer): Dimension of this layer matches the number of input variables. No computation is of data is done in this layer. Each node transmits input values to the next layer direct.

Layer2 (Fuzzification layer): Each node in the layer corresponds to one linguistic label- low expressed, medium expressed, highly expressed (LOW, MEDIUM, HIGH) of an input variable. Each node in this layer

calculates the membership value specifying the degree to which an input value belongs to a fuzzy set. The triangular membership function shown in figure 6.1 is used in this layer.

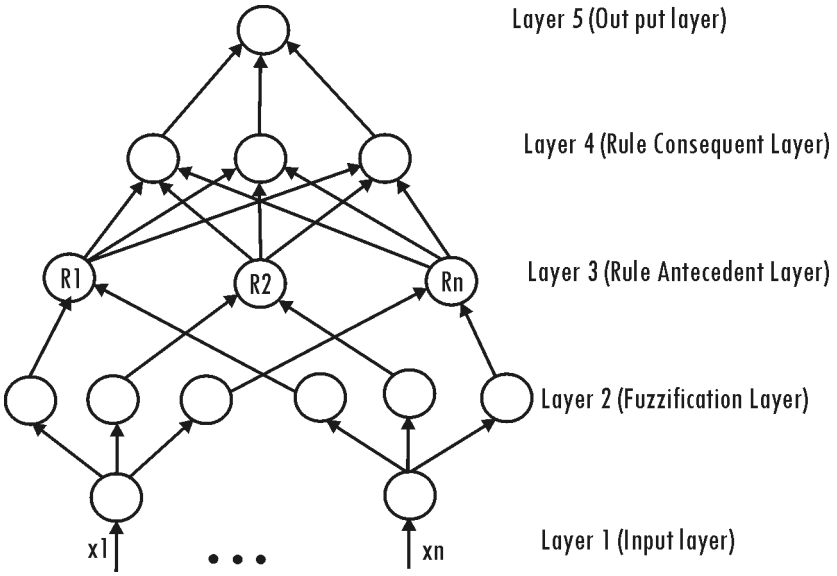


Figure 6.5 Architecture of DNFN

Layer 3 (Rule Antecedent Layer): Nodes in this layer are called rule nodes. The number of nodes in this layer will be incremented each time when a new rule is generated during the structural learning phase. Each rule node represents a fuzzy logic rule and performs precondition matching of the rule. Fuzzy AND operator max is used to evaluate the antecedent part of each rule. The output of a rule node represents the firing strength of the corresponding rule.

Layer 4 (Rule Consequent Layer): The number of nodes in this layer will be equal to the number of rule nodes. Every rule has a weight assigned to it. The weighted sum of the firing strength of each rule is calculated and it

is used to reshape the fuzzy set associated with the consequent part. The output of this layer is the reshaped fuzzy set.

Layer5 (Aggregation and Defuzzification Layer): This layer is the output layer. This node combines the fuzzy sets from the previous layer to a single fuzzy set. The fuzzy set thus obtained is then defuzzified. The centriodal defuzzification technique is used in this layer.

6.3.2 Construction of Fuzzy Rules

The way the input space is partitioned determines the number of rules. Each rule generated has the following form

R : if x_1 is A_1 and x_2 is A_2 and … and x_n is A_n then y is B

where A and B are the fuzzy sets used to perform a map from the input space vector to the output space. The algorithm for the construction of fuzzy rules is as follows.

Suppose no rule exist initially, let (x, y) be the vector representing input and output pattern respectively.

If x is the first incoming pattern

Step 1: Generate a new rule with A_i as the fuzzy set having maximum membership value for x_i and the output fuzzy set B_i is the fuzzy set having maximum membership value for y.

Else for each newly incoming pattern (x_k, y_d) do the following steps.

Step 2: Find y_k , the actual output with the existing rules.

Step3: Compute the overall error using the equation

$E_k = abs(y_d - y_k)$, where y_d is the desired output.

Step 4: If$E_k > \lambda_{error}$Create new rule, do step1.

else

Step 5: Adjust the parameters of the network using the same input pattern.

6.3.3 Deletion of Redundant Rules

In order to reduce the complexity of the model two criteria for deleting redundant rules were proposed.

1. If a fuzzy set is near zero over itsown universe of discourse for a certain number of time steps then the rule with specific fuzzy set as precondition should be removed as this fact indicates that the output of the rule is also near zero.
2. When the two fuzzy rules have same consequent and similar antecedent part, we have a case of selecting one among the rules. The solution is to select the rule whose firing strength is high for more number of samples.

The main reason for the presence of redundant rule is due to the inherent noise in the microarray input data.

The multilayered dynamic neural fuzzy network was modeled to extract regulatory relationship among genes and reconstruct gene regulatory network from microarray data.This method combines the merits of connectionist and fuzzy approaches. It encodes the knowledge learned in the form of fuzzy rules and processes data following fuzzy reasoning principles. While the dynamic aspect of gene regulation was taken into account through the on-line learning of fuzzy rules, the structural learning together with the

parameter learning form a fast learning algorithm for building a small, yet powerful, dynamic neural fuzzy network.

6.4 TSK-type Recurrent Neural Fuzzy Network

Recurrent neural networks (RNNs), in general, can deal with temporal and spatial/temporal problems by adapting feedback connections to their topologies which are used to memorize past information. Although RNNs are generally efficient for temporal sequence production problems, it has been proved that recurrent neural fuzzy networks are superior to recurrent neural networks in dealing with the problems like GRN reconstruction which involves concurrent spatial and temporal mapping [146,147]. Additionally, due to their high-level, humanlike reasoning, fuzzy-based approaches are better candidates in dealing with the uncertainties of modelling noisy data [148, 149].

In this work, a TSK type recurrent neural fuzzy network (TRNFN) has been applied to extract gene regulatory relations from microarray data. Although recurrent neural fuzzy based algorithms have been used for inferring GRN by some investigators [147], the use of TSK- type fuzzy model for the inference of gene regulatory network has not been successfully explored. The TRNFN extends DNFN by the inclusion of memory elements in the form of feedback connections. Unlike DNFN, whose output is a function of its current inputs only, TRNFN can perform dynamic mapping using their ability to store prior system states. Moreover, the rules in DNFN are of ordinary Mamdani-type fuzzy rule, whereas TRNFN uses TSK-type fuzzy rule. In [155], where TSK-type recurrent fuzzy network TRNFN is used for dynamic system control has shown that the performance and learning accuracy is superior to those of Mamdani type

fuzzy network. In TRNFN a global feedback structure is adopted where the output of all rule nodes are fed back and summed, so each rule's firing strength depends not only on its previous value but also on others. With the global feedback structure, TRNFN achieve better performance than local feedback structure where, the rule's firing strength is influenced only by its previous value. Furthermore, the inclusion of TSK- type consequence can significantly reduce the rule number [155]. TRNFN is characterized by small network size and fast learning speed.

6.4.1 Architecture

TSK-type Recurrent Neural Fuzzy Network (TRNFN) was proposed by Chia-FengJuang in 2002 [155]. TRNFN is constructed from a series of fuzzy if –then rules, with the consequence of each rule being of TSK-type fuzzy reasoning. The network precondition part includes external variables and internal variables derived from fuzzy firing strengths, and the consequence is a linear combination of them plus a constant term. Each rule i has the following form:

$$R(i) = IF x_1(t) is\ A_{i1}\ and x_2(t) is\ A_{i2}\ and \ldots x_n(t)\ is\ A_{in}\ and\ h_i(t)\ is\ G$$

$$THEN\ y(t+1) = a_{i0} + \sum_j a_{ij}.x_j(t) + a_{in+1}\,h_i(t)$$

where A and G are fuzzy sets,h is the internal variable and a is the parameter for the inference output y.

Two learning phases, structural and parameter learning are used concurrently for constructing TRNFN. The structural learning phase includes input-output space partitioning, construction of fuzzy if- then rules and the feedback structure identification. The parameter learning phase is responsible for tuning of free parameters of the network structure. TRNFN

works for supervised learning environment with available gradient information. Since the gradient information is available, the neural network learning approach is used for learning the network parameters.

Architecture of TRNFN with two external inputs and one single output is illustrated in figure 6.6. TRNFN is a six layered network, including a feedback layer that brings the temporal processing ability into a feed forward neural fuzzy network. The detailed function of each layer is described below. In the following descriptions the symbol $u_i^{(k)}$denotes the input of the i^{th} node in the k^{th} layer; correspondingly, the symbol $Y_i^{(k)}$ denotes the output of thei^{th} node of layer k.

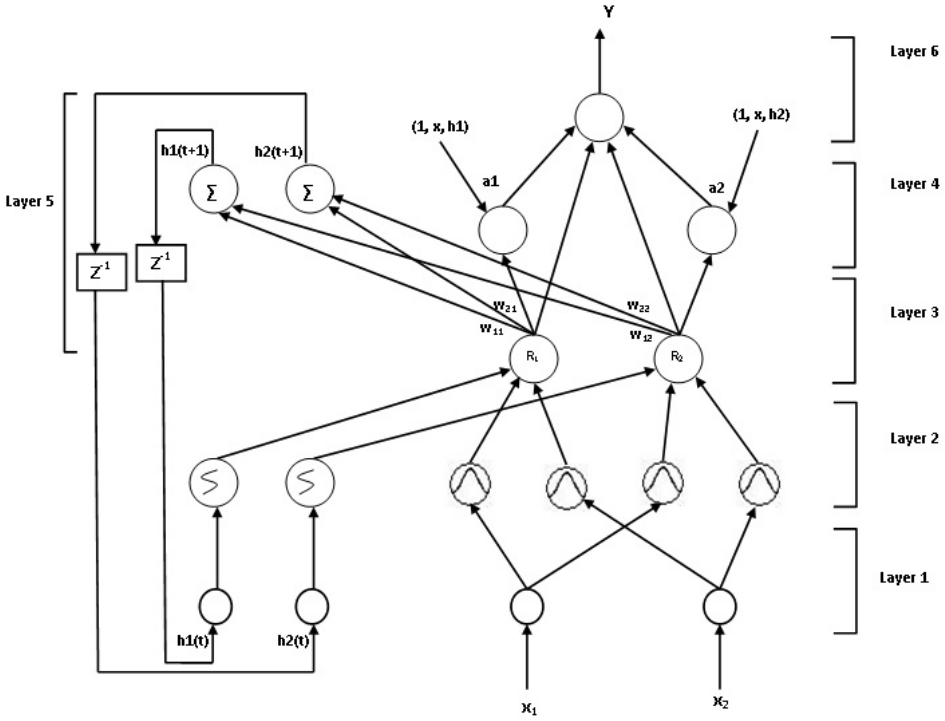


Figure 6.6 Architecture of TRNFN

Layer 1: Dimension of this layer matches the number of input variable. No computation is done in this layer. The node only transmits input values to the next layer directly. That is

$$Y_i^{(1)} = u_i^{(1)} = x_i$$

Layer 2: Node in the layer corresponds to one linguistic label- low expressed, medium expressed, highly expressed etc. of an input variable. Each node in this layer calculates the membership value specifying the degree to which an input value belongs to a fuzzy set. Gaussian membership function is employed in this layer since; it can be easily decomposed in to the product of one-dimensional membership functions. With this choice, operation performed in this layer is

$$Y_i^{(2)} = exp\left(-\frac{\left(u_j^{(2)} - m_{ij}\right)^2}{{\sigma_{ij}}^2}\right) \; and \; u_j^{(2)} = Y_j^{(1)}$$

wherem_{ij} and σ_{ij}representsthe center and width of the Gaussian membership function of the i^{th} term of the j^{th} input variable x_j.For the internal variable h_i the following sigmoid member function is used:

$$Y_i^{(2)} = \frac{1}{1 + exp\left(-u_i^{(2)}\right)} \; and \; u_i^{(2)} = Y_i^{(5)}$$

The link weights are all set to unity.

Layer 3: Nodes in this layer are called rule nodes. The number of nodes in this layer will be incremented each time when a new rule is generated during the structural learning phase. Each rule node represents a fuzzy logic rule and performs precondition matching of the rule. The input to each node

comes from two sources: one from layer 2 and the other from the feedback layer. Former, the output of the rule node represents the spatial firing strength of its corresponding rule and the latter, the output of the feedback term node, represents the rule's temporal firing degree.The output of each node in this layer is determined by fuzzy AND operation. Here, the product operation is utilized to determine the firing strength of each rule. The function of each rule is

$$Y_i^{(3)} = \prod_{j=1}^{n+1} Y_j^{(2)} = \frac{1}{1+\exp\left(-Y_i^{(5)}\right)} . exp\left\{-\left(\sum_{j=1}^{n} \frac{\left(Y_j^{(1)} - m_{ij}\right)^2}{\sigma_{ij}^2}\right)\right\}$$

where n is the number of external inputs. The link weights in this layer are set to unity.

Layer 4: Nodes in this layer perform a linear summation. The mathematical function for each node iin this layer is

$$Y_i^{(4)} = \sum_{j=0}^{n+1} a_{ij} u_j^{(4)} = a_{i0} + \sum_{j=1}^{n} a_{ij} x_j + a_{in+1} h_i$$

where n is the number of external inputs, and $a_{ij}, j = 0, \ldots n+1$ are the parameters to be tuned in the structural learning phase. All link weights from this layer to layer 6 are set to unity.

Layer 5: This layer is also called feedback layer, which calculates the value of internal variable h_i. Each rule node has a corresponding internal variable h which is used to decide the influence of the firing history to the current rule.The link weightsare assigned random values initially and are updated

during the parameter learning phase. The weighted sum of outputs from rule nodes is calculated in each node.

$$h_i = Y_i^{(5)} = \sum_{j=1}^{r} Y_j^{(3)}.w_{ij}$$

where r is the number of rules generated so far. The delayed value of $h_i$is connected to layer 2 as shown in the figure 6.6.

Layer 6: Each node in this layer is called an output linguistic node. This node performs the defuzzification operation and computes the output signal Y. The node in this layer together with the links attached to itperforms this task. The mathematical function is

$$Y = Y^{(6)} = \frac{\sum_{j=1}^{r} Y_j^{(3)}.Y_j^{(4)}}{\sum_{j=1}^{r} Y_j^{(3)}}$$

Learning Process

Upon receiving the training data TRNFN performs structural as well as parameter learning process simultaneously. In the structural learning phase TRNFN decides the number of fuzzy rules and initializes the network parameters. In the parameter learning phase it optimally adjusts the free parameters of the network structure. The way the input space is partitioned determines the number of rule. The creation of new rule corresponds to the creation of new cluster in the input space. The spatial firing strength can be regarded as the degree to which an input pattern belongs to the corresponding cluster. Based on this concept, the spatial firing strength

$$F^i(x) = \prod_{k=1}^{n} Y_k^{(2)}$$

$$= exp\left\{-\sum_{j=1}^{n} \frac{\left(x_j - m_{ij}\right)^2}{\sigma_{ij}^2}\right\} \in [0,1]$$

is used as a criterion to decide whether a fuzzy rule should be generated or not. As there was no rules initially, for the first incoming pattern x(0), a new fuzzy rule is generated with center and width of gauss membership function is assigned as

$$m_{1i} = x_i(0)\ and\ \sigma_{1i} = \sigma_{init}\ , for\ i = 1 \ldots n$$

whereσ_{init} is a constant that determines the initial width of the first cluster. For the succeeding inputsx (t), find

$$I = \arg \max_{1 \le i \le r(t)} F^i(x)$$

where r(t) is the number rules at time t. If $F_I \le F_{init}$, then new rule is generated, where$F_{init} \epsilon (0,1)$ is a pre-specified threshold that decays during learning process. The center and width of the new rule can be set according to the first-nearest-neighborheuristic as

$$m_{(r(t)+1)i} = x_i(t)\ and$$

$$\sigma_{(r(t)+1)i} = \beta \sum_{j=1}^{n} \frac{\left(x_j - m_{Ij}\right)^2}{\sigma_{Ij}^2}$$

where,i varies from 1 to n and$\beta \geq 0$ decides the overlap degree between the two clusters. Both parameters F_{init} and β decides the number of rules to be generated.

After the generation of new rule for the input data <x(t), y(t)>the consequent nodes in layer 4 and context node in layer 5 are computed. The initial value of parameter a_{i0} is set to y(t) and the other a_{ij} parameters are set to small random values in between[-0.05, 0.05]. To make initial values of h lies in the sensitive region of sigmoid function, the initial link weights are set as random values in the range [-1,1] . As a result, quick parameter learning can be reached at the beginning. The output, h, of the new context node is fed back as input in the precondition part of the newly generated rule. With this setting, each rule has its own memory elements for memorizing the temporal firing strength history. By repeating the above process for every incoming training data, a new rule is generated, one after another, and a whole TRNFN is constructed finally.

The parameter learning process is performed concurrently with the structural learning phase based on the same training pattern. The free parameters such as a, m, σ and w are tuned using the real time recurrent learning algorithm [155, 156].

The network structure of TRNFN is such that it encodes knowledge learned in the form of fuzzy if-then rules and takes into account the dynamic aspects of gene regulation through its recurrent structure. The main features of TRNFN include:

1. The learning algorithm automatically produces an adaptive number of fuzzy rules that describe the relationships between the input (regulating) genes and the output (regulated) gene.
2. It overcomes the need for prior data discretization.
3. The recurrent property , achieved by feeding the internal variables in the form of feedback connections , increases the learning ability of TRNFN

When applied to the problem of reconstruction of gene regulatory network from microarray data, TRNFN was provided good results from the accuracy and speed points of view.

……….☙❧……….

Chapter 7

Performance Evaluation

Contents

The complex molecular mechanism underlying cancer is associated with the perturbations of gene-interaction networks at some level. Therefore, identifying genes responsible for oncogenesis and the pathways they control through networks is a key step towards overcoming cancer. Inferring gene regulatory network under specific disease conditions can genuinely reflect the principal cause- effect relationship between relevant genes. The recent advances of array technologies have made it possible to find the underlying network of gene-gene interactions from gene expression dataset. Changes in expression levels of genes across different samples provide information that allows reverse engineering techniques to extract gene regulatory features like activation and inhibition, which enable to construct regulatory relations among those genes. However, these approaches suffer several difficulties including (i) dimensionality problem of microarray datasets (ii) exponential complexity of the algorithm (iii) presence of noise in expression values. A wide variety of computing techniques have been used to infer GRN from microarray data. In this work four soft computing approaches such as Fuzzy logic algorithm, Modified

genetic algorithm, Dynamic feed forward neural network and TSK type recurrent neural fuzzy network are used to extract cancer specific GRN from microarray data. In this chapter regulatory networks inferred using the above four approaches are presented. The performances of algorithms are evaluated using multiple datasets. The biological processes and cancer related pathways associated with the interested genes are also listed.

7.1 Modelling Gene Regulatory Network from Circulating Plasma RNA Dataset

This section presents the gene regulatory network extracted from circulating plasma RNA dataset of colorectal cancer patients using four computational approaches such as Fuzzy logic algorithm, Modified genetic algorithm, Dynamic feed forward neural fuzzy network and TSK type Recurrent neural fuzzy network. The main objective is to observe the roles played by some relevant genes in the context of colon cancer specific gene regulatory network.

7.1.1 Fuzzy Logic Algorithm

Gene regulatory networks are models of genes and the gene interactions. Using microarray data researchers can reverse engineer the underlying regulatory relationship between genes. The fuzzy logic algorithm has been proposed to search microarray dataset for the activator-repressor relationship among genes. To improve the algorithm in terms of reducing the computational time, clustering has been used as a preprocessing step. This will reduce the total number of gene combinations analyzed and optimizes computation time.

Among the 15552 genes in the dataset, 100 genes were selected whose p-values are found to be less than 0.005 for further analysis. By applying hybrid clustering algorithm on the expression profile of selected genes, four clusters have been obtained. The ranges of $\log_2$ (ratio) of the selected 100 genes were from -6.21 to 10.71, and they were normalized from 0.0 to 1.0 to be used as the input data for the fuzzy logic algorithm in the following analysis.

The fuzzy logic algorithm was executed with no error (MSE) or variance (hit ratio) cutoff using the cluster centroid as the input. The triplets of cluster centers are ranked according to the MSE and hit ratio. From the results obtained, six triplets having MSE value less than 0.7 and hit ratio greater than 0.5 are found to fit to the fuzzy model and are selected for further processing. Each gene in the cluster representing activator is paired with the corresponding genes in the cluster representing repressor and this gene pair determines the predicted target gene expression values based on the set of heuristic rules. The output is compared to the expression levels of the genes belonging to the target cluster. This will reduce the amount of time required and thus reduces the complexity of algorithm. Same set of gene triplets have been obtained as that obtained without using the pre-processing steps. The program implemented in matlab requires 13 seconds to cluster the dataset and 80 seconds to identify the gene triplets whereas the fuzzy logic algorithm without preprocessing step takes 232secs to identify the relationship between the same numbers of genes. The improved algorithm is four times faster than the original algorithm. Thus the use of clustering as a preprocessing step to identify genes that are likely to interact can save significant amount of time in building fuzzy logic based model. The gene triplets having MSE value less than 0.38 and a hit ratio greater than 0.7 are shown in Table 7.1.

Table 7.1 Regulatory Relations predicted by Fuzzy Logic Algorithm

Activator	Repressor	Target
LDHA	TRIP10	HBE1
UBE2D3	RELN	
KIAA0632	DOC1	DGKZ
PPP1CC	TNKS	
ACP1	TIAF1	BRP44
KIAA0758	TRIP10	PCBP2
	FLJ10849	
LDHA	FLJ10849	MVP
RPL10A		
KIAA0758		
PSMA3	SP38	PCBP2
OXA1L		
EPAS1		
ANXA1		
ACP1		
KIAA0758		
ANXA1	SP38	MVP
		IL7R
FLJ20701	C18B11	C20orf194

To evaluate the algorithm, the triplets have been examined to see whether they have made any biological sense. The expression profiles of gene EPAS1, SP38 andPCBP2 are shown in Figure 7.1. As can be seen from figure 7.1, expression profiles of EPAS1 and PCBP2 are similar. If gene A and B have similar expression profiles, there are several possible relationships: i) A and B are co regulated by other genes; ii) A regulates B or vice versa; or iii) there is no causal relationship but only coincidence. Here, regulation may be indirect, i.e., interaction through immediates. These cases cannot be differentiated solely by clustering. In this analysis, EPAS1-SP38-PCBP2 is identified as one of the best scoring triplet. From the figure 7.1 we can see that EPAS1 and PCBP2 have similar expression profiles and on increasing the expression value of EPAS1and decreasing the value of SP38 would cause the expression value of PCBP2 to increase.

The network predicted by the fuzzy logic algorithm with least error is shown in figure 7.2. Among the 27 genes, 19 genes are upregulated and 8 genes are downregulated in cancer biopsies. Red circle nodes represent upregulated genes and blue circle node denotedownregulated genes. An edge → indicates activation of transcription, whereas, an edge ⊣ indicates repression of transcription. Cytoscape software [157] has been used to draw the network.

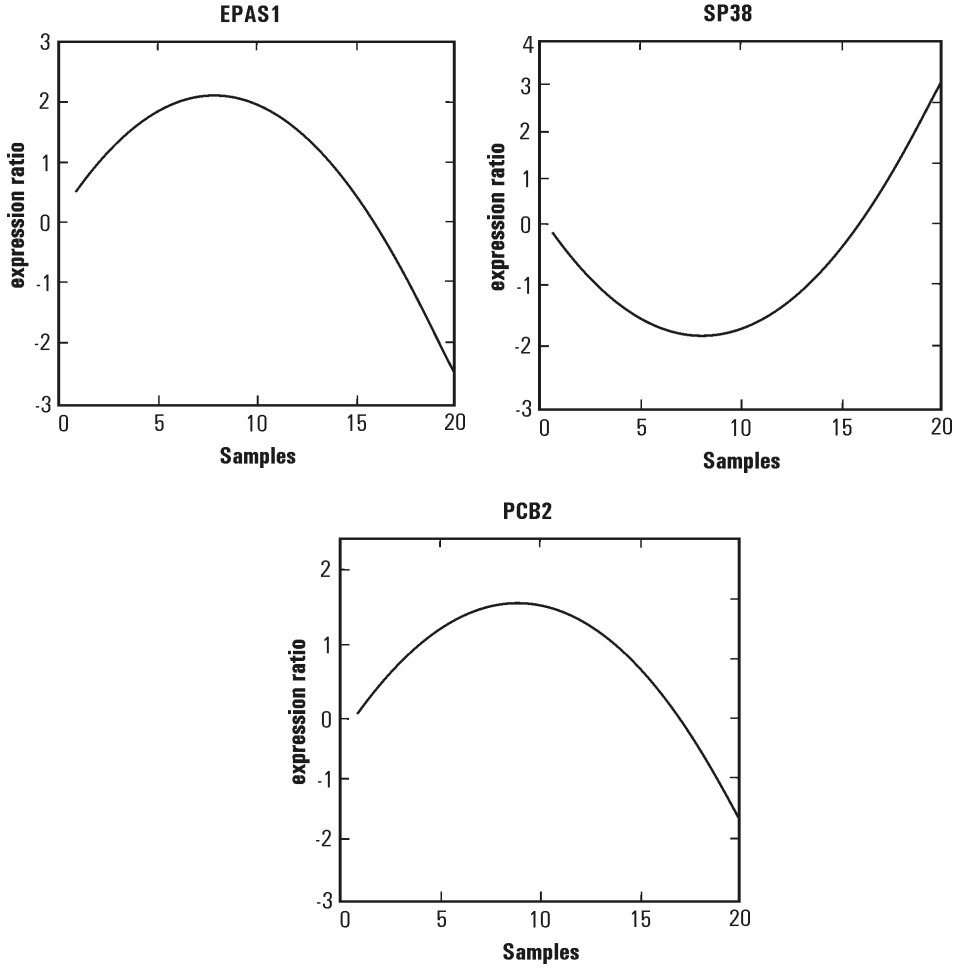


Figure 7.1 Expression profiles of EPAS1, SP38 & PCBP2

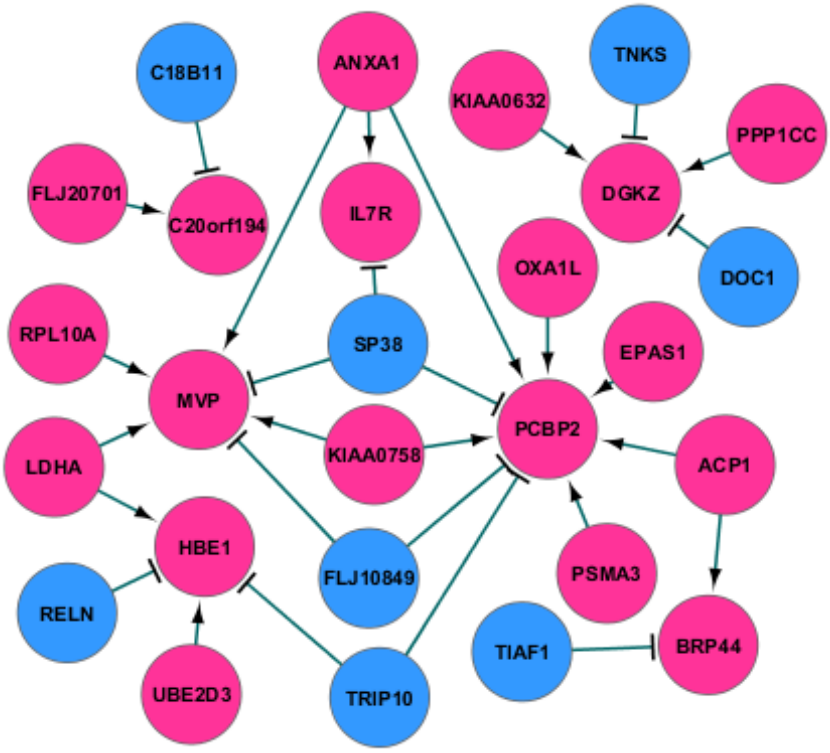


Figure 7.2 Gene Regulatory network inferred using fuzzy logic algorithm

The fuzzy logic approach was proposed for analyzing gene expression data using fuzzy logic. The algorithm has been improved in terms of computational time by using clustering as a preprocessing step. Using this method it may be possible to identify the cellular function of unknown genes by examining known genes associated with the set. This ability to extract functional information from the expression data could be particularly useful in analyzing human data, where a much larger percentage of proteins are uncharacterized.

7.1.2 Modified Genetic Algorithm

Genetic algorithm can effectively model gene interaction structure that accurately reflects the underlying biology. However this approach requires several generations and much computational power to identify the interacting genes. In this approach correlation techniques have been used to reduce the search space, thereby reducing the number of generations. The modified algorithm was applied on a set of genes whose activator/repressor relationships were identified using fuzzy GRN algorithm

The set of 27 genes from the table 7.1 are selected for modelling gene regulatory network using modified genetic algorithm. Table 7.2 shows the GA parameters used in this experiment. Due to the large search space, the basic GA requires long time to converge. By incorporating the correlation techniques, modified GA runs fast and gives good results compared to the basic GA.

Table 7.2 Algorithm Parameters

	No. of Variables	Population	Generations	Crossover	Mutation	Time (secs)
Basic GA	26	500	1500	0.8	0.2	200
Modified GA	26	100	150	0.8	0.2	50

The weight matrix representing the regulatory relationship between genes was optimized using Genetic algorithm. Instead of randomly generating initial population, Jackknife correlation coefficient has been used to create the initial population of weight vector. The genetic algorithm was executed to determine the optimal weights of the regulators of the target gene. The process is repeated for each of the target gene. The weight matrix thus generated was used to model Gene regulatory network. Each node in

the GRN represents a gene and the presence of an edge between two nodes indicates an interaction between connected genes. The edges labeled positive weight denote activation and the edges labeled negative weight denote repression.

The regulatory pathways inferred using modified genetic algorithm is shown in figure 7.3. The algorithm identified 67 relations. The relations predicted by modified GA are compared with that obtained using fuzzy GRN model on the same data set. The modified GA algorithm was successful in determining 62% of regulatory relations found by Fuzzy GRN model listed in table 7.1. The advantage of the new approach is that it can approximate the relationships among genes without heavy computation. Furthermore, the proposed algorithm is capable of predicting the nature of the relationship represented by the link through the weight. For example, the connection from EPAS1 to PCBP2 has a weight equal to +0.73, positive, so it represents activation.Another example is the connection from SP38 to PCBP2 has a weight equal to -0.87, negative, so it stands for inhibition.

Modified genetic algorithm captured more regulatory relations than that identified using fuzzy logic algorithm. This algorithm helps to integrate the biological expert knowledge with heuristic search. Further biological experiments are required to determine the validity of the genetic interactions suggested by this model.

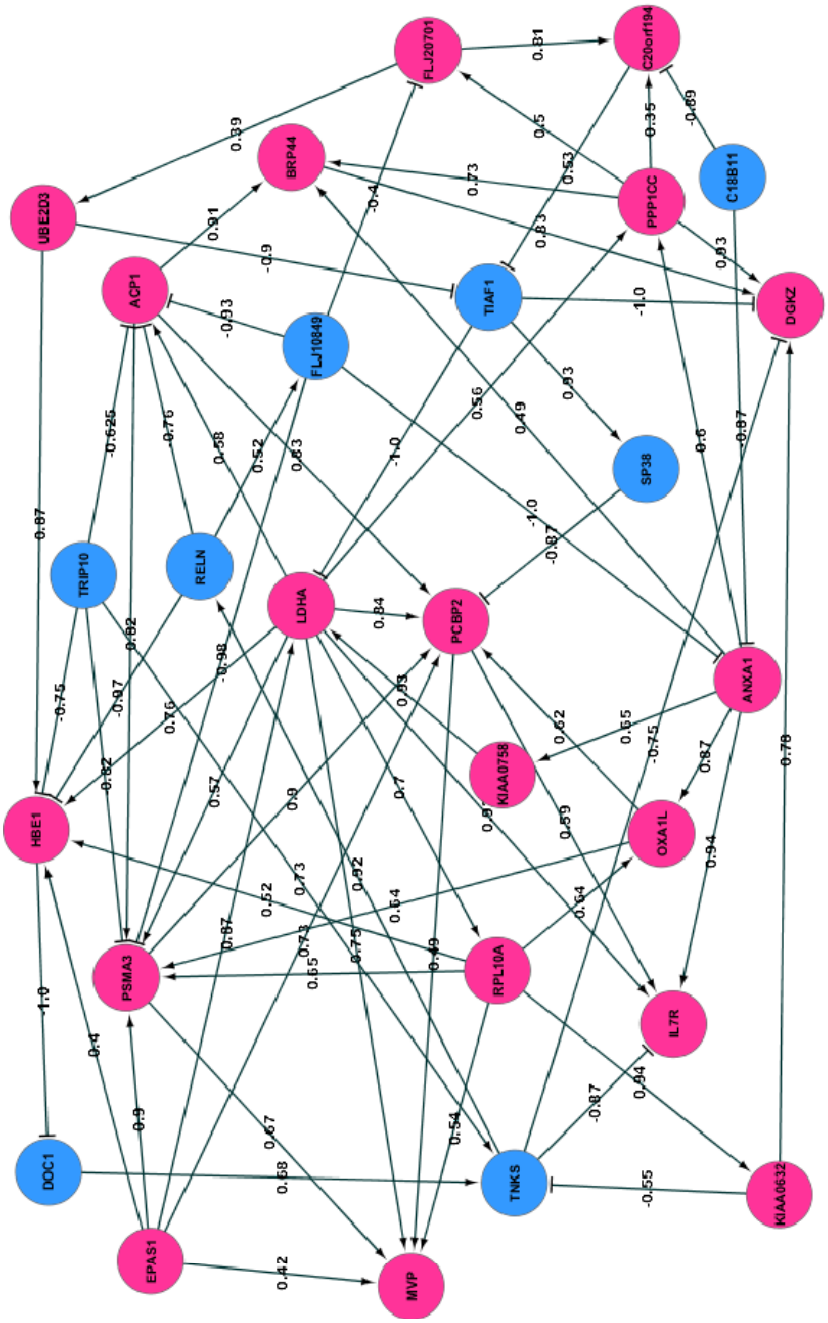


Figure 7.3 Regulatory network obtained using modified genetic algorithm.

7.1.3 Dynamic Feed Forward Neural Fuzzy Approach

Based on the neural fuzzy learning system a feed forward neural fuzzy approach (DNFN) has been proposed for inferring the complex causal relationships among genes from microarray experimental data. DNFN derives the information on the gene interactions in a highly interpretable form (fuzzy rules) and the data is processed using fuzzy reasoning principles. To determine the efficiency of the proposed approach, a set of 27 genes known to be highly regulated in plasma RNA dataset using previous approaches, were again considered. Among the 20 available samples in the dataset, 12 samples were used as training data for DNFN and the remaining 8 samples were used for testing the consistency. Twenty six multi input - single output DNFN models were implemented. Each model describes the state of output gene based on the expression value of the other 26 genes. Structure of each model is generated using the training dataset and the parameter tuning is done by the repeated learning. Table 7.3 presents the set of rules describing gene PCBP2. Each rule is read column wise. The numbers correspond to the fuzzy set to which each input gene expression level belongs, that is 3 for HIGH, 2 for MEDIUM and 1 for LOW.(e.g. the first column contains the rule 'if HBE1 is LOW and SP38 is HIGH and RPL10A is MEDIUM and…and UBE2D3 is MEDIUM then PCBP2 is LOW). Examples of rules generated by DNFN for predicting the target genes are listed in appendix 3.

Table 7.3 Rules describing the state of gene HBE1 based on the remaining 26 genes

Input Genes	Rule 1	Rule 2	Rule 3
PSMA3	1	2	3
FLJ10849	2	1	2
SP38	2	3	2
FLJ20701	1	2	3
LDHA	1	2	3
OXA1L	2	2	3
TRIP10	3	2	1
RELN	3	2	1
EPAS1	1	2	3
RPL10A	1	2	3
DGKZ	2	3	3
C20orf194	2	2	3
BRP44	2	3	3
UBE2D3	1	2	3
ANXA1	2	3	3
C18B11	2	2	2
TNKS	2	1	2
ACP1	1	2	3
TIAF1	3	2	2
PCBP2	2	2	3
IL7R	2	2	3
DOC1	3	1	2
MVP	1	2	3
PPP1CC	2	2	3
KIAA0758	2	3	3
KIAA0632	2	2	3
Prediction of HBE1	1	2	3

The fuzzy rule set derived from 27 DNFN model was used to build a gene regulatory network as shown in figure 7.4. DNFN identified 70 regulatory relations among 27 genes. The regulatory relations detected using DNFN was compared with that obtained using fuzzy GRN model and modified genetic algorithm on the same data set. DNFN predicted 72.4% relations found by Fuzzy GRN model and 64.1% relations detected by modified genetic algorithm.

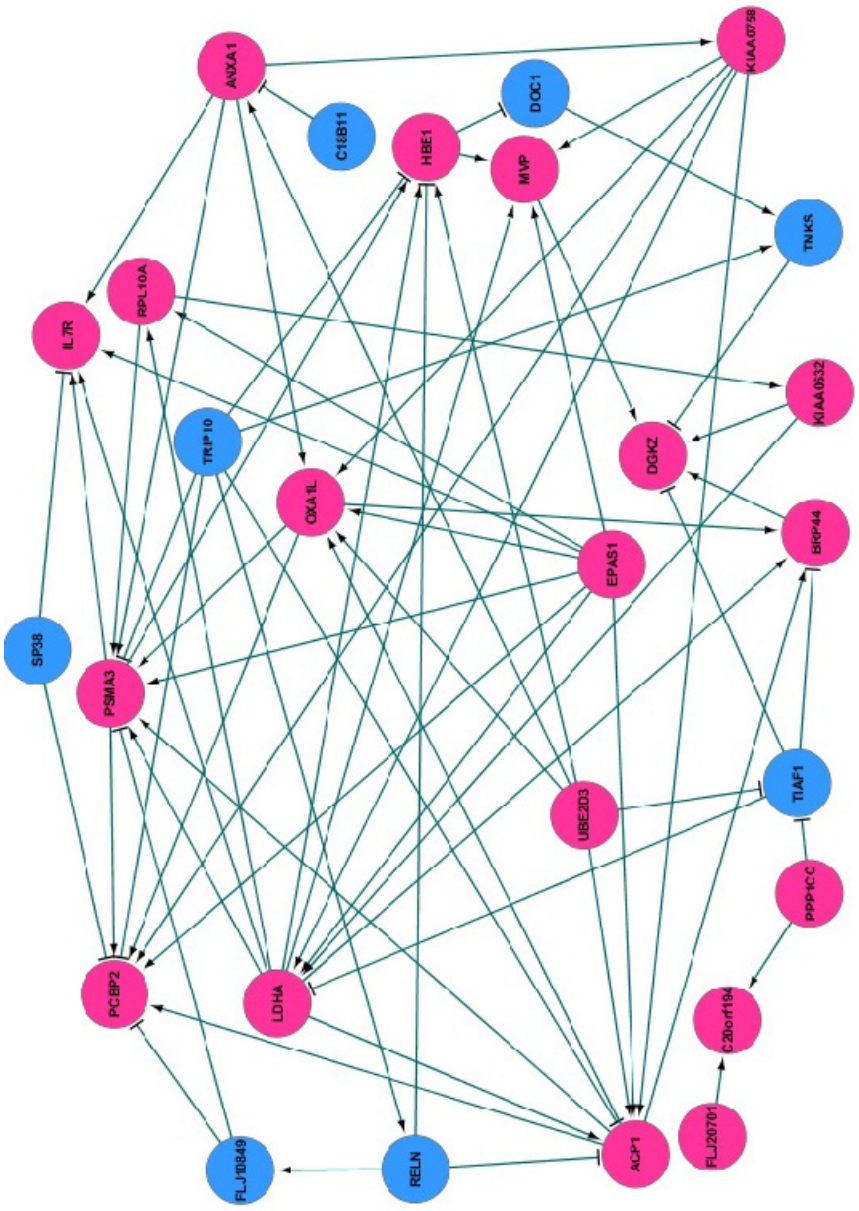


Figure 7.4 Gene Regulatory Network predicted by DNFN model.

One of the main advantages of DNFN over Fuzzy GRN model is that there is no predefined rules; since it generate fuzzy rules automatically and quickly. The online learning of fuzzy rules enables DNFN to incorporate the dynamic aspect of gene regulation while modelling regulatory network.

7.1.4 TSK type Recurrent Neural Fuzzy Network

To select potential regulators of target genes and describe their regulation type, an approach based on multilayered TSK-type recurrent neural fuzzy network (TRNFN) was proposed. The approach was tested by applying it to a set of 27 genes from table 7.1, each of which shows good performance in distinguishing cancerous samples from normal ones. TRNFN incorporates memory elements in the form of feedback connections. Unlike DNFN, whose output is a function of current input only, TRNFN can perform dynamic mapping using their ability to store prior system states. Randomly selected 12 samples from the available 20 samples where used as training data for the TRNFN and the remaining 8 samples were used for testing the consistency. Training and test dataset contains samples from both healthy and unhealthy tissues. Twenty seven multi input - single output TRNFN model was implemented. Each model describes the state of output gene based on the expression value of the other 26 genes. During the structural learning phase the structure of each model was generated from the training data set.

The parameters of the network structure were tuned by repeated learning. The fuzzy rules derived from the TRNFN model have similar structure as that of DNFN model. The fuzzy rule set derived from 27 TRNFN model was used to build gene regulatory network. Examples of rule

setsgenerated by TRNFN for describing the stateof target genes are listed in appendix 4.

The regulatory pathways inferred using TRNFN are shown in figure 7.5. TRNFN predicted 91 relations among 27 genes.

Table 7.4 presents the MSE values of 27 TRNFN, DNFN and GA models for the gene prediction. The low error values obtained in the table indicates that all the three models can perform accurate prediction of the expression values. By comparing MSE values of three models it is clear that learning accuracy of TRNFN is superior to that of DNFN and Genetic Algorithm. Table 7.5 lists the set of relations predicted by all the four approaches. When comparing the results of TRNFN to the regulatory relations predicted by modified genetic algorithm, DNFN and fuzzy logic algorithm almost all the predicted relations matched with those predicted by these three programs. Some of the predicted relations lie within the union or intersection of the other three programs. TRNFN extracted 86.2 % of the relations predicted by Fuzzy logic algorithm (when GA 62% and DNFN 72.4%), 79.1% of the relations detected by GA (when DNFN 64.1%) and 84% of the relations identified by DNFN. It seems likely that the relations that lie within intersections could be more accurate, although this claim remains to be evaluated further, experimentally.

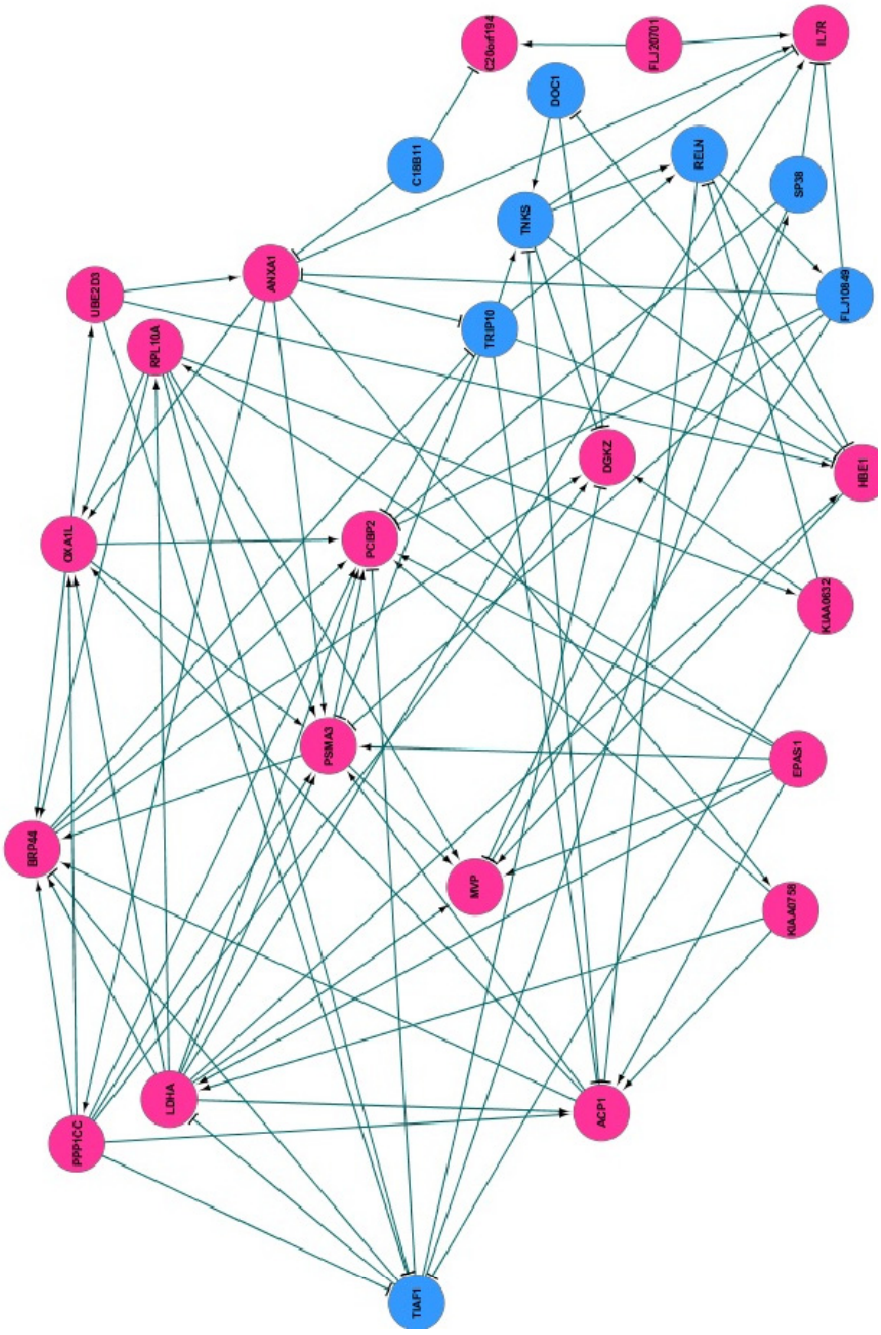


Figure 7.5 Gene Regulatory Network Predicted by TRNFN

Table 7.4 Mean Square Error (MSE) of the 27 TRNFN, DNFN and GA models for the gene prediction

Predicted Genes	Prediction Errors		
	TRNFN	DNFN	GA
HBE1	0.0053	0.0244	0.0283
PSMA3	0.0159	0.0267	0.0292
FLJ10849	0.0278	0.028	0.0294
SP38	0.014	0.0278	0.0249
FLJ20701	0.0122	0.0249	0.0293
LDHA	0.0245	0.0298	0.0327
OXA1L	0.0117	0.0305	0.0163
TRIP10	0.0103	0.0259	0.0201
RELN	0.0053	0.021	0.0254
EPAS1	0.0188	0.0247	0.0200
RPL10A	0.003	0.0339	0.0237
DGKZ	0.0123	0.0244	0.0273
C20orf194	0.0217	0.0251	0.0252
BRP44	0.0219	0.0187	0.0190
UBE2D3	0.0078	0.019	0.0159
ANXA1	0.0149	0.0218	0.0133
C18B11	0.0039	0.0242	0.0222
TNKS	0.0176	0.0288	0.0178
ACP1	0.0224	0.022	0.0131
TIAF1	0.015	0.0169	0.0232
PCBP2	0.012	0.0186	0.0239
IL7R	0.024	0.034	0.0295
DOC1	0.012	0.0209	0.0209
MVP	0.021	0.0225	0.0294
PPP1CC	0.021	0.0233	0.0262
KIAA0758	0.016	0.0179	0.0127
KIAA0632	0.017	0.0208	0.0231

Table 7.5 Set of Relations predicted by Fuzzy Logic, Genetic Algorithm, DNFN and TRNFN.

Regulator	Relation	Target	Fuzzy	GA	DNFN	TRNFN
RELN	-	HBE1	√	√	√	√
TRIP10	-	HBE1	√	√	√	√
LDHA	+	HBE1	√	√	√	√
UBE2D3	+	HBE1	√	√	√	√
TNKS	-	HBE1	-	-	-	√
EPAS1	+	HBE1	-	√	-	-
RPL10A	+	HBE1	-	√	-	-
PSMA3	+	HBE1	-	-	√	-
EPAS1	+	PSMA3	-	√	√	√
ACP1	+	PSMA3	-	√	√	√
LDHA	+	PSMA3	-	√	√	√
OXA1L	+	PSMA3	-	√	√	√
RPL10A	+	PSMA3	-	√	√	√
PPP1CC	+	PSMA3	-	-	-	√
TRIP10	-	PSMA3	-	√	√	√
ANXA1	+	PSMA3	-	-	√	√
RELN	+	FLJ10849	-	√	√	√
FLJ10849	-	PSMA3	-	√	√	√
TIAF1	+	SP38	-	√	-	√
PPP1CC	+	FLJ20701	-	√	-	-
FLJ10849	-	FLJ20701	-	√	-	-
EPAS1	+	LDHA	-	√	√	√
ACP1	+	LDHA	-	-	√	-
KIAA0632	+	LDHA	-	-	√	-
KIAA0758	+	LDHA	-	√	√	√
TIAF1	-	LDHA	-	√	√	√
RPL10A	+	OXA1L	-	√	-	√
PPP1CC	+	OXA1L	-	-	-	√
LDHA	+	OXA1L	-	-	-	√
KIAA0758	+	OXA1L	-	-	√	-
UBE2D3	+	OXA1L	-	-	√	-
EPAS1	+	OXA1L	-	-	√	-
ACP1	+	OXA1L	-	-	√	√
ANXA1	+	OXA1L	-	√	√	√

Regulator	Relation	Target	Fuzzy	GA	DNFN	TRNFN
BRP44	-	TRIP10	-	-	-	√
ANXA1	-	TRIP10	-	-	-	√
TRIP10	+	RELN	-	-	√	√
TNKS	+	RELN	-	√	-	√
KIAA0632	-	RELN	-	-	-	√
EPAS1	+	RPL10A	-	-	√	√
LDHA	+	RPL10A	-	√	√	√
PPP1CC	+	DGKZ	√	√	-	√
KIAA0632	+	DGKZ	√	√	√	√
DOC1	-	DGKZ	√	-	-	√
BRP44	+	DGKZ	-	√	√	√
MVP	+	DGKZ	-	-	√	-
TNKS	-	DGKZ	√	√	√	√
FLJ20701	+	C20orf194	√	√	√	√
C18B11	-	C20orf194	√	√		√
PPP1CC	+	C20orf194		√	√	
PPP1CC	+	BRP44	-	√	-	√
PSMA3	+	BRP44	-	-	-	√
LDHA	+	BRP44	-	-	√	√
OXA1L	+	BRP44	-	-	√	√
RPL10A	+	BRP44	-	-	-	√
ACP1	+	BRP44	√	√	√	√
TIAF1	-	BRP44	√	-	√	√
ANXA1	+	BRP44		√	-	-
FLJ20701	+	UBE2D3	-	√	-	-
OXA1L	+	UBE2D3	-	-	-	√
UBE2D3	+	ANXA1	-	-	√	√
FLJ10849	-	ANXA1	-	√	-	√
C18B11	-	ANXA1	-	√	√	√
DOC1	+	TNKS	-	√	√	√
TRIP10	+	TNKS	-	√	√	√
ACP1	-	TNKS	-	-	-	√
KIAA0632	-	TNKS	-	√	-	-
PPP1CC	+	ACP1	-	-	-	√
LDHA	+	ACP1	-	√	√	√

Regulator	Relation	Target	Fuzzy	GA	DNFN	TRNFN
TRIP10	-	ACP1	-	√	√	√
RELN	-	ACP1	-	√	√	√
EPAS1	+	ACP1	-	-	√	√
UBE2D3	+	ACP1	-	-	√	-
KIAA0758	+	ACP1	-	-	√	√
FLJ10849	-	ACP1	-	√	-	-
TIAF1	-	DGKZ	-	√	√	√
PPP1CC	-	TIAF1	-	-	√	√
KIAA0632	-	TIAF1	-	-	-	√
RPL10A	-	TIAF1	-	-	-	√
UBE2D3	-	TIAF1	-	√	√	√
C20orf194	-	TIAF1	-	√	-	-
PPP1CC	+	PCBP2	-	-	-	√
PSMA3	+	PCBP2	√	√	√	√
LDHA	+	PCBP2	-	√	-	√
OXA1L	+	PCBP2	√	√	√	√
EPAS1	+	PCBP2	√	√	√	√
KIAA0758	+	PCBP2	√	-	√	√
BRP44	+	PCBP2	-	-	-	√
ANXA1	+	PCBP2	√	-	-	-
ACP1	+	PCBP2	√	√	√	-
SP38	-	PCBP2	√	√	√	√
FLJ10849	-	PCBP2	√	-	√	√
TRIP10	-	PCBP2	√	-	√	√
TIAF1	-	PCBP2	-	-	-	√
LDHA	+	IL7R	-	√	√	√
ANXA1	+	IL7R	√	√	√	√
TNKS	-	IL7R	-	√	-	√
SP38	-	IL7R	√	-	√	√
FLJ10849	-	IL7R	-	-	-	√
EPAS1	+	IL7R	-	-	√	-
PSMA3	+	IL7R	-	-	√	-
PCBP2	+	IL7R	-	√	-	-
FLJ20701	+	IL7R	-	-	-	√
HBE1	-	DOC1	-	√	√	√

Regulator	Relation	Target	Fuzzy	GA	DNFN	TRNFN
LDHA	+	MVP	√	√	√	√
PSMA3	+	MVP	-	√	-	√
RPL10A	+	MVP	√	√	-	√
FLJ10849	-	MVP	√	-	-	√
HBE1	+	MVP	-	-	√	√
KIAA0758	+	MVP	√	-	√	-
ANXA1	+	MVP	√	-	-	-
SP38	-	MVP	√	-	-	√
EPAS1	+	MVP	-	√	√	√
PCBP2	+	MVP	-	√	-	-
ANXA1	+	PPP1CC	-	√	-	√
LDHA	+	PPP1CC	-	√	-	-
RPL10A	+	KIAA0632	-	√	√	√
ANXA1	+	KIAA0758	-	√	√	√

Upregulated genes can be referred as tumor activators and down regulated genes as tumor suppressors. From the above regulatory networks 8 regulatory patterns are observed as shown in figure 7.6. Pattern 1 represents one activator activating multiple activators. Pattern 2 represents one activator suppressing multiple suppressors. Pattern 3 represents one suppressor activating multiple suppressors. Pattern 4 represents one suppressor suppressing multiple activators. Pattern 5 represents one activator is being activated by multiple activators. Pattern 6 represents one activator is being suppressed by multiple suppressors. Pattern 7 represents one suppressor is being suppressed by multiple activators. Pattern 8 represents one suppressor is being activated by multiple suppressors. Among which the patterns 1, 5, 7 are found to be strong and the others are relatively weak.

This indicates that tumor activators activates other tumor activators and suppress tumor suppressers strongly in the disease environment. Also, the upregulated genes are regulated by more genes than the down regulated genes.The high degree of centrality of upregulated genes indicates that they play key roles in cancer specific gene regulatory network. Similar findings are made by other authors [158, 159].

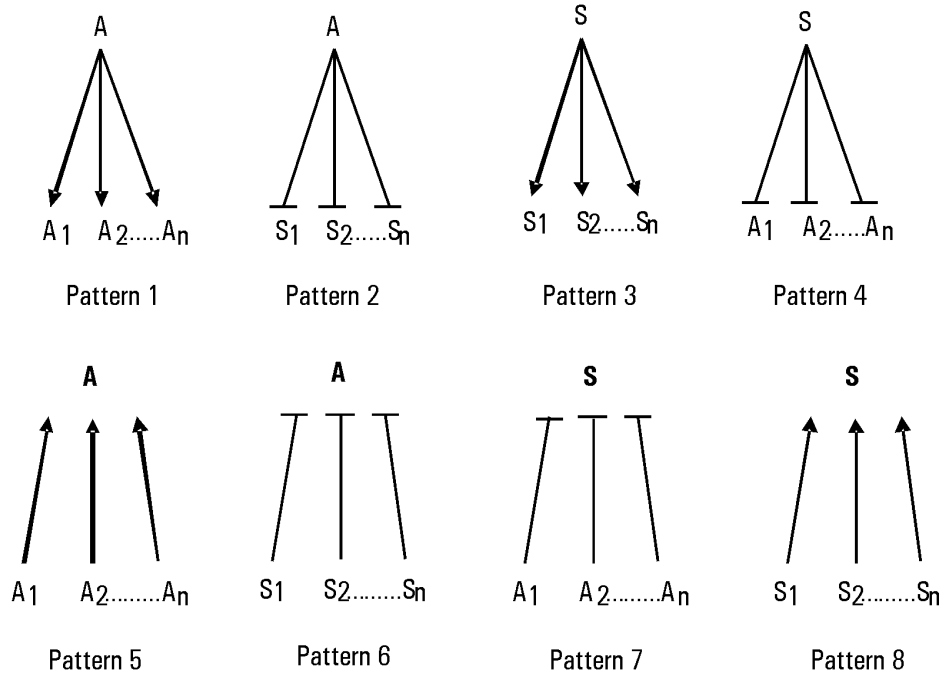


Figure 7.6 Eight Regulatory Patterns observed from the gene regulatory network. Here A denotes Activator, S denotes Suppressor, A_i is the i^{th} activator and S_i is the i^{th}suppressor.

GO-based (gene ontology) gene enrichment analysis has been done to identify the group of genes that share common biological pathways. The Cytoscape plug-in, Biological Networks Gene Ontology Tool (BINGO) [160] was used to determine the Gene Ontology terms that is significantly

over represented in these set of genes. BINGO uses the Benjamini and Hochberg correction to control the false discovery rate (FDR). The table 7.6 lists the over represented GO categories. The columns include the GO-ID term of the category and its description, the set of genes annotated to that category along with the enrichment significance (p-value).

In a biological system, genes and gene products interact with each other forming various biological pathways in order to carry out several biological processes. To better understand the cause and effect of gene expression changes in a complex biological system, a comprehensive pathway analysis is needed. So a detailed analysis was performed to identify which genes among these differentially expressed genes were involved in the canonical pathways associated with cancer. The genes like EPAS1, UBE2D3, PSMA3, LDHA etc. are actively involved in the cancer related pathways. The results are shown in table 7.7. Overall, it is clearly showed that this cancer related mRNAs may regulate many biological functions and pathways.

From the previous gene expression analysis of Colorectal Cancer, three of the selected markers – PSMA3, EPAS1 and UBE2D3 were found to be significantly overexpressed in cancer patients compared to healthy persons. The high expression of these genes in CRC has been shown to play an important role in tumor progression [16].Also these three markers showed a significant decrease after surgery, returning to the levels of healthy donors. The above models can be used to identify those genes which are affected by the up regulation of the above cancerous genes.

Table 7.6 GO terms shared by more than one gene with $p \leq 0.05$

GO –ID	Description	Genes	P-value
43161	Proteasomal Ubiquitin-dependent protein Catabolic Process	UBE2D3, PSMA3, PCBP2 ,DOC1	2.82E-05
10498	Proteasomalprotein Catabolic Process	UBE2D3, PSMA3, PCBP2, DOC1	2.82E-05
5737	Cytoplasm	OXA1L, LDHA, EPAS1, ANXA1, DOC1, PPP1CC, ACP1, FLJ20701, UBE2D3, PSMA3, PCBP2, DGKZ, RELN, BRP44, TNKS, RPL10A ,HBE1, TRIP10, MVP	1.17E-04
2698	Negative Regulation of Immune Effector Process	PCBP2, IL7R	2.70E-04
6511	Ubiquitin-dependent protein Catabolic Process	UBE2D3, PSMA3, PCBP2, DOC1	4.03E-04
43632	modification-dependent macromolecule catabolic process	UBE2D3, PSMA3, PCBP2, DOC1	4.32E-04
19941	modification-dependent protein catabolic process	UBE2D3, PSMA3, PCBP2, DOC1	4.32E-04
70979	Protein K-11 linked Ubiquitnation	UBE2D3, DOC1	4.71E-04
44260	Cellular macromolecule metabolic process	UBE2D3, EPAS1, C18B11, PSMA3, PCBP2, ANXA1, RELN ,TNKS, DOC1, RPL10A, PPP1CC, ACP1	5.16E-04
51603	Proteolysis involved in cellular protein catabolic process	UBE2D3, PSMA3, PCBP2, DOC1	6.24E-04
44257	Cellular protein catabolic process	UBE2D3, PSMA3, PCBP2, DOC1	6.47E-04

Table 7.7 Genes involved in Cancer-related Canonical Pathways

Pathway	Genes Involved
Hypoxia Signaling Pathway	EPAS1,LDHA,UBE2D3
ProteinUbiquitination Pathway	PSMA3, UBE2D3
Polyamine regulation in Colon cancer	PSMA3
Pyruvate Metabolism	LDHA
Cytokine-Cytokine Receptor Interaction	IL7R,TNF
Insulin Signaling Pathway	TRIP10,PPP1CC,EPAS1
Renal Cell Carcinoma	EPAS1
TGF beta Receptor Signaling pathway	UBE2D3
Epidermal Growth factor Receptor (EGFR1) signaling pathway	ANXA1,ACP1
TNF alpha Signaling Pathway	ACP1,UBE2D3
Wnt Signaling Pathway	ACP1
CREB signaling pathway	LDHA,PPP1CC,ANXA1
Jack-STAT signaling pathway	IL7R
Regulation of actin cytoskeleton	PPP1CC
Cycle Role of Anaphase Protein complex (APC)in in cell cycle regulation	DOC1

7.2 Analysis of Colon Tumor Sample Dataset

To determine the efficiency of the proposed approaches, microarray data from two experiments relating to colon cancer dataset and Yeast dataset have been used. The colon cancer dataset [161] contains 59 samples collected from colon cancer patients. Among them, 35 samples are from colon tissues and 24 samples are from healthy parts of the colons of the same patients. From this experimental data using the four approaches such as Fuzzy Logic, modified

Genetic Algorithm, DNFN and TRNFN, the regulatory relationship among the above set of 27 genes were inferred. The inferred GRNs using Fuzzy Logic, modified Genetic Algorithm, DNFN, and TRNFN are shown in figure 7.7, 7.8, 7.9 and 7.10. TRNFN extracted 50 interactions exist in both blood and tumor sample, whereas fuzzy logic had 15, DNFN had 42 and modified genetic algorithm had 29.

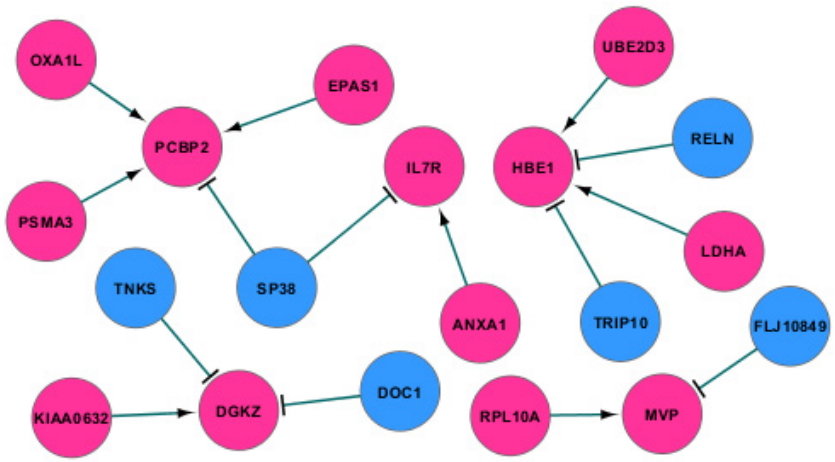


Figure 7.7 Relations common for Plasma RNA dataset and tumor sample dataset predicted by Fuzzy Logic Algorithm

Thus the proposed approaches are successful in reconstructing multiscale gene regulatory networks that reveal global patterns of gene interactions in colorectal cancer. These interaction networks will serve as a blue print for us to understand CRC progression from blood samples and to develop novel therapeutics.

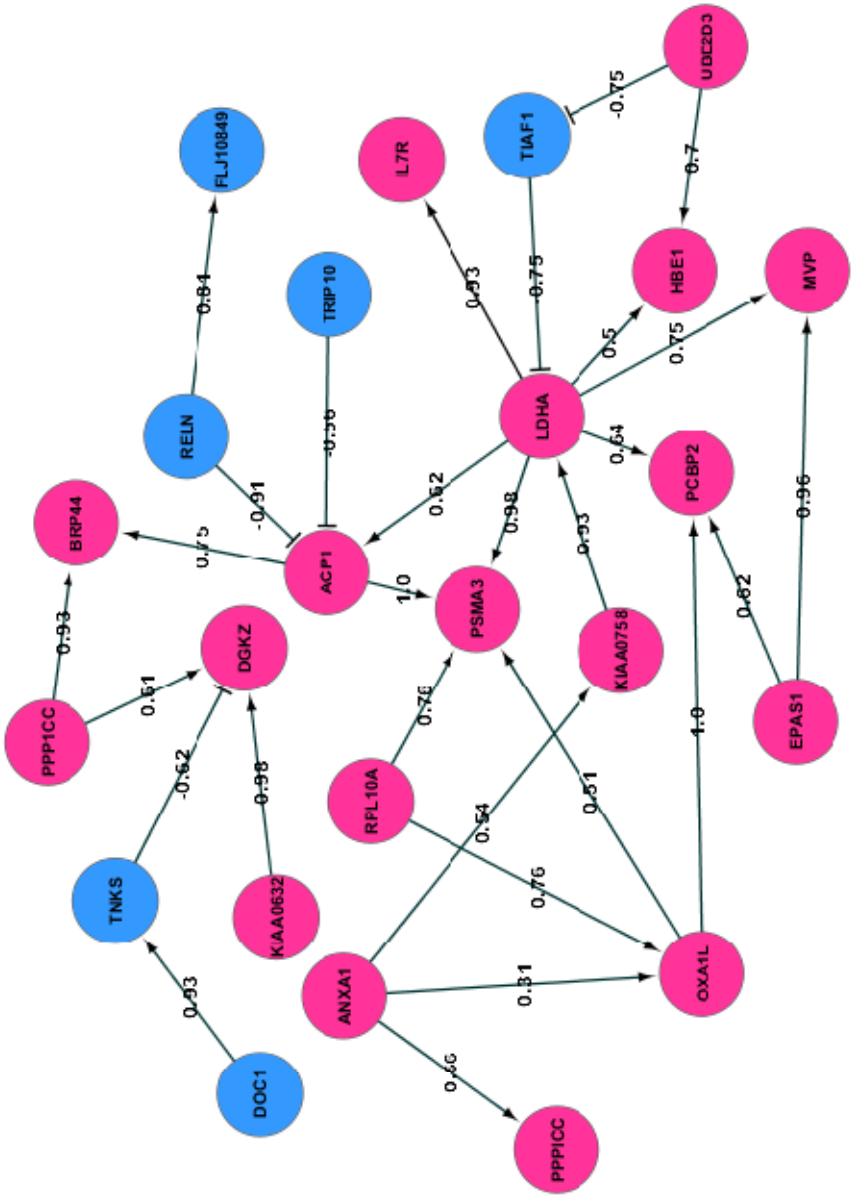


Figure 7.8 Relations common for Plasma RNA dataset and tumor sample dataset predicted by Modified Genetic algorithm

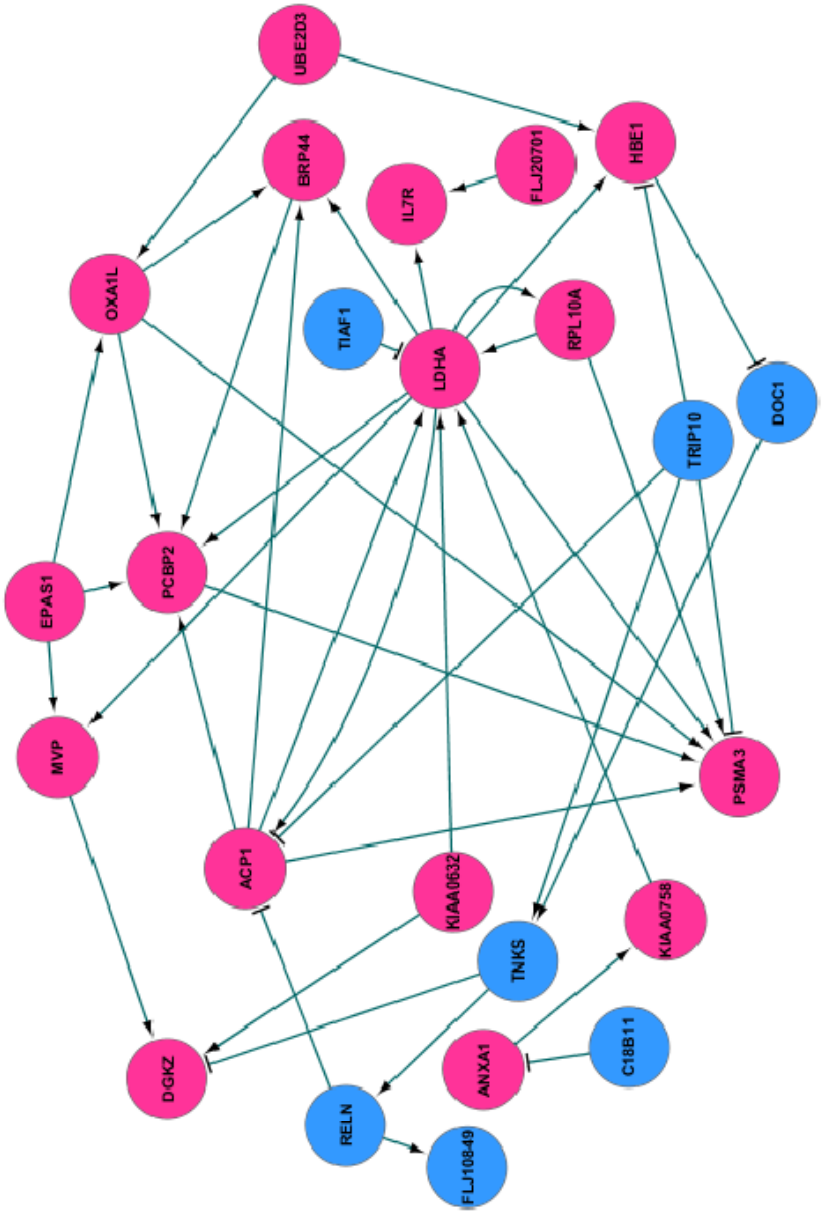


Figure 7.9 Relations common for Plasma RNA dataset and tumor sample dataset predicted by DNFN

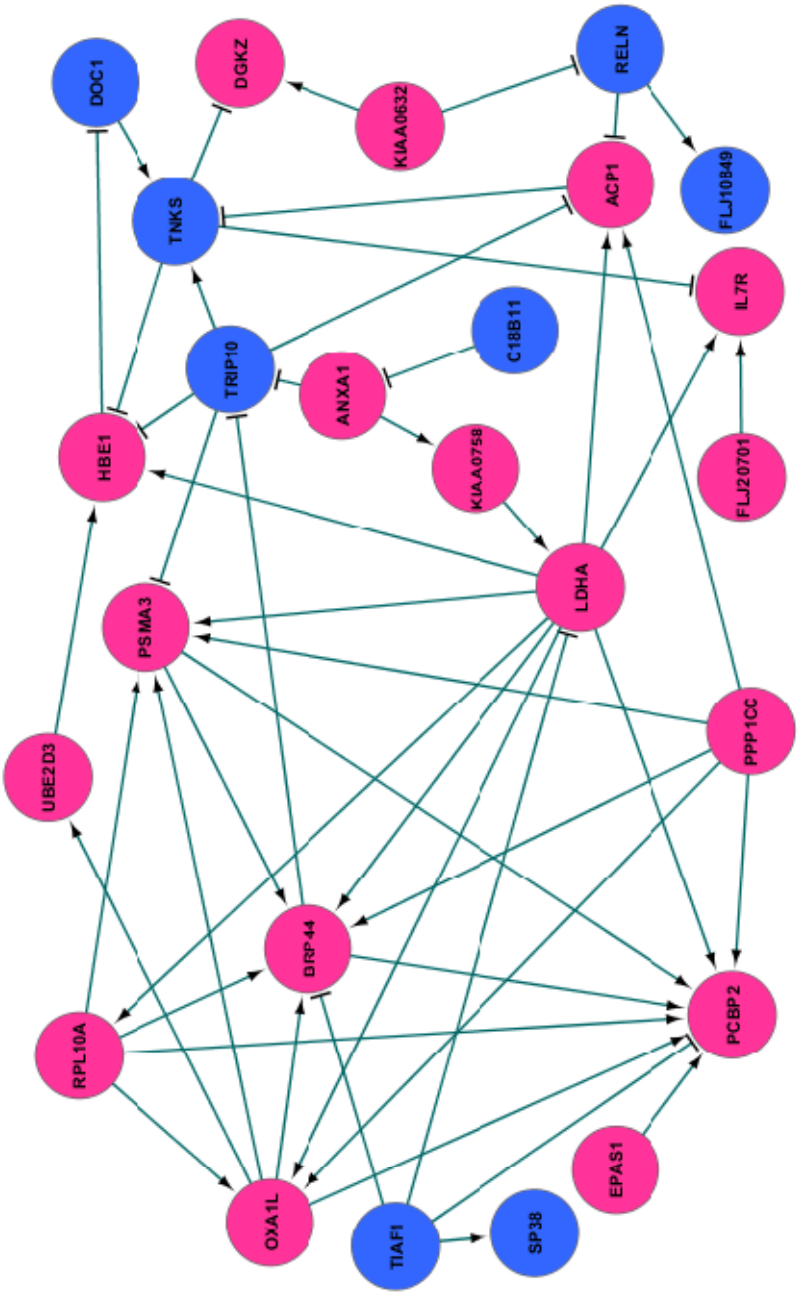


Figure 7.10 Relations common for Plasma RNA dataset and tumor sample dataset predicted by TRNFN

7.3 Analysis of Yeast Dataset

To evaluate the computational efficiency of the proposed approaches and compare the results with already published relations the dataset concerning yeast were used. For that, microarray data from Spellman *et. al.* [162] obtained for Saccharomyces Cervisiae cell was used. Saccharomyces Cervisiae cell cultures were synchronized with different methods including alpha factor arrest (two complete cycles, sampled each 7 min, totally 18 samples) temperature arrest of temperature sensitive mutants (cdc15, three complete cycles, sampled each 10 min, totally 24 samples) and temperature arrest of a cdc28 allele (17 samples taken at 10 min interval). The missing values exist in all datasets, which indicate that there was no strong enough signal in the spot. In this case, the missing data was filled using simple spline interpolation. The cdc15 experiment was chosen for the training dataset because it has the largest number of data points (sample).The remaining datasets, cdc28 and alpha-factor were used as test sets. To evaluate the performance of each approach, a well-studied pathway consist of 14 yeast genes were selected. These are the cyclin genes CLN1-3 and CLB1-6 and CDC28, MBP1, SWI4-6, CDC20, SIC1 and MCM1, which are involved in cell cycle regulation and whose interactions are well described.

To systematically present the results, the Kyoto Encyclopedia of Genes and Genomics (KEGG) map [163] was treated as the ground truth. Although there are still uncertainties, the KEGG map represents up-to-date knowledge about the dynamics of gene interaction and it should be

reasonable to serve as a benchmark of results validation. The figure 7.11 presents a part of the KEGG pathway and the regulatory network inferred using computational approaches such as TRNFN, DNFN and modified Genetic Algorithm. As it can be noticed from the figure 7.11 presenting the respective interactions, the majority of relations among genes predicted by the three methods are biologically validated interactions. This indicates the flexibility of the proposed methods over the input dataset. Indeed, it is shown that the proposed methods will adequately manage to determine the best interactions following the peculiarities of the input data. Table 7.8 presents the MSE value of fourteen TRNFN, DNFN and GA models for the gene prediction. From table 7.8 it can be observed that TRNFN manages to recover better results than that of DNFN and Genetic Algorithm.

By inspecting figure 7.11 and table 7.9 it is clear that, the TRNFN is able to predict more biologically meaningful relations than other methods. TRNFN approach extracted 33 biologically validated interactions among 14 genes whereas DNFN extracted 19 and modified GA extracted 14 valid relations. It is found that 87.8% of the total interactions extracted by TRNFN are in accordance with the biological proven regulatory interactions.

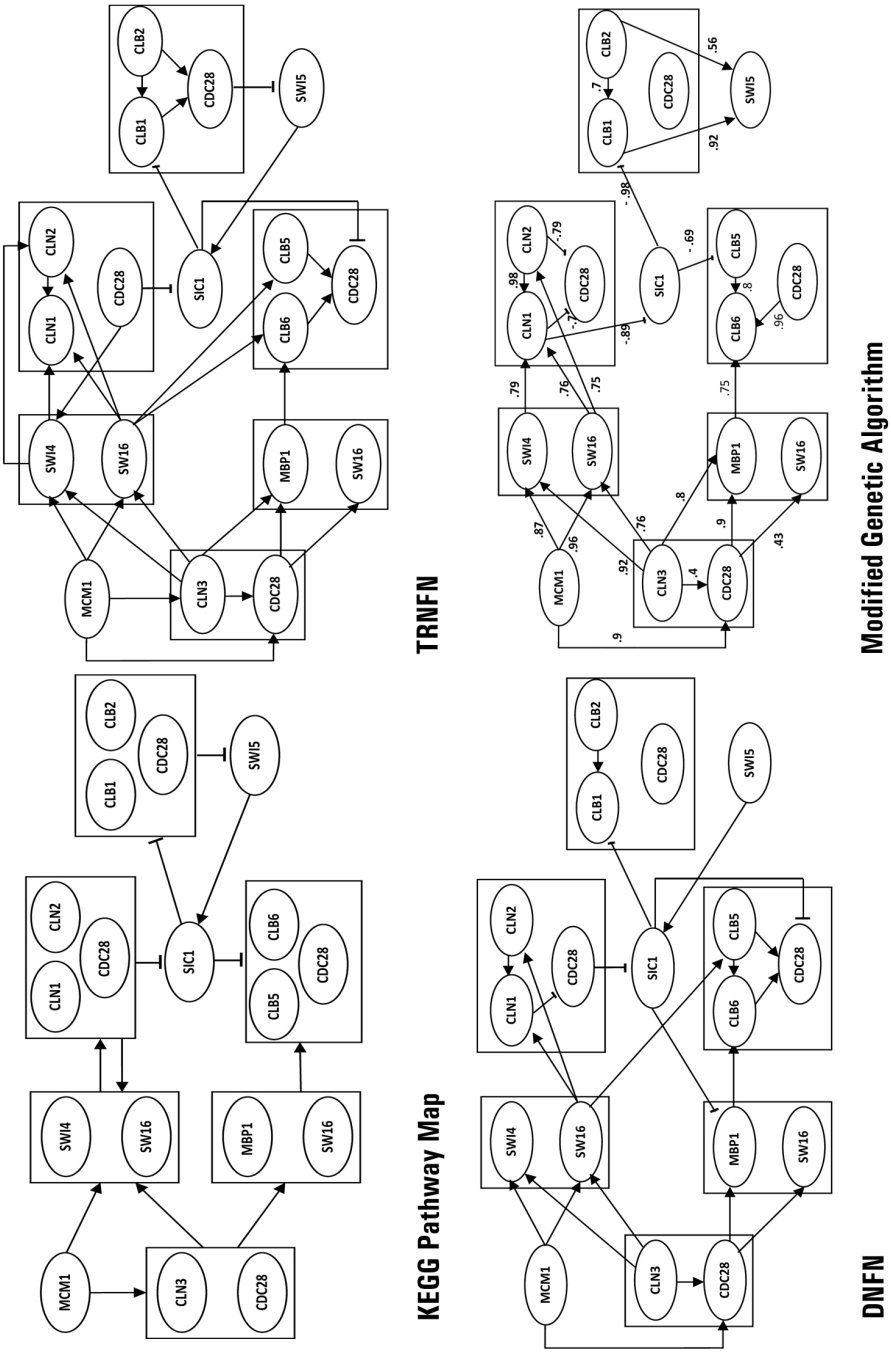


Figure 7.11 Reconstruction of KEGG pathway. The figure presents the part of the KEGG pathway and the regulatory network inferred using computational approaches such as TRNFN, DNFN and modified Genetic Algorithm. The rectangle represents genes that compose a complex.

Table 7.8 Mean Square Error obtained for predicting 14 genes using TRNFN,DNFN and modified Genetic Algorithm

Predicted gene	Prediction Errors		
	TRNFN	DNFN	GA
CLN1	0.018	0.0184	0.0213
CLN2	0.013	0.0476	0.0181
CLN3	0.015	0.0327	0.0391
CLB1	0.012	0.048	0.0216
CLB2	0.015	0.015	0.0173
CLB5	0.015	0.0266	0.0204
CLB6	0.011	0.0249	0.0239
MCM1	0.017	0.0214	0.0348
SIC1	0.011	0.0255	0.0325
SW16	0.015	0.0272	0.0348
CDC28	0.014	0.0398	0.0391
MBP1	0.013	0.0291	0.0174
SW15	0.014	0.0284	0.0186
SW14	0.012	0.0147	0.0271

Table 7.9 Biologically validated interactions detected by the three computational models TRNFN, DNFN and Modified Genetic Algorithm

a/a	Regulator	Type	Target	TRNFN	DNFN	GA
1	MCM1	+	CLN3	√	-	-
2	MCM1	+	CDC28	√	√	√
3	CLN3	+	SWI6	√	√	√
4	SIC1	-	CDC28	√	√	-
5	MCM1	+	SWI4	√	√	√
6	MBP1	+	CLB5	√	-	-
7	SWI6	+	CLB6	√	-	-
8	SWI5	+	SIC1	√	-	-
9	CLN3	+	SWI4	√	√	√
10	SWI6	+	CLB5	√	√	-
11	CDC28	+	MBP1	√	√	√
12	CDC28	-	SIC1	√	√	-
13	CDC28	+	SWI4	√	√	-
14	CDC28	-	SWI5	√	√	-
15	CDC28	+	SWI6	√	√	√
16	SIC1	-	CLB1	√	√	√
17	CLB6	+	CDC28	√	√	-
18	SWI4	+	CLN1	√	-	√
19	MCM1	+	SWI6	√	√	√
20	CLB1	+	CDC28	√	-	-
21	MBP1	+	CLB6	√	√	-
22	SWI6	+	CLN1	√	√	√
23	SWI4	+	CLN2	√	-	√
24	SWI6	+	CLN2	√	√	√
25	CLN3	+	MBP1	√	-	√
26	CLB2	+	CDC28	√	-	-
27	CLB5	+	CDC28	√	√	-
28	CLN1	-	SIC1	-	-	√
29	SICI	-	CLB5	-	-	√
30	CLB5	+	CLB6	-	√	√
31	CLN2	+	CLN1	√	√	√
32	CLN3	+	CDC28	√	√	√
33	CDC28	+	CLB6	-	-	√
34	SIC1	-	MBP1	-	√	

A highly important aspect for every method applied to the problem at hand is the one concerning the time needed for reconstructing a certain

network. The total time required for the methods to fully complete their operation and output the final GRN is recorded and listed in table 7.10. All the methods have been implemented in Matlab© and run on a 2 Core 1.83 GHz with 512MB RAM machine. Inspection of Table 7.10 proves that the TRNFN based approach is much more efficient than DNFN, modified GA and Fuzzy Logic in terms of computational time needed for the methods to fully reconstruct a regulatory network. From figure 7.12 it can deduce that although the processing time of TRNFN follows an exponential increase with the size of genes sets, still the processing time is very low (it took less than 3 minutes to reconstruct the initial structure of 200 genes).

Table 7.10 Comparison in terms of computational time of TRNFN against 2 other methods proposed (time is given in minutes)

Methods	Plasma RNA Dataset (27 genes,20 Samples)	Tumor Sample Dataset (27 genes ,59 Samples)	Yeast Dataset (14 genes,59 Samples)
TRNFN	10	38	4
DNFN	13	59	5
GA	16	67	11

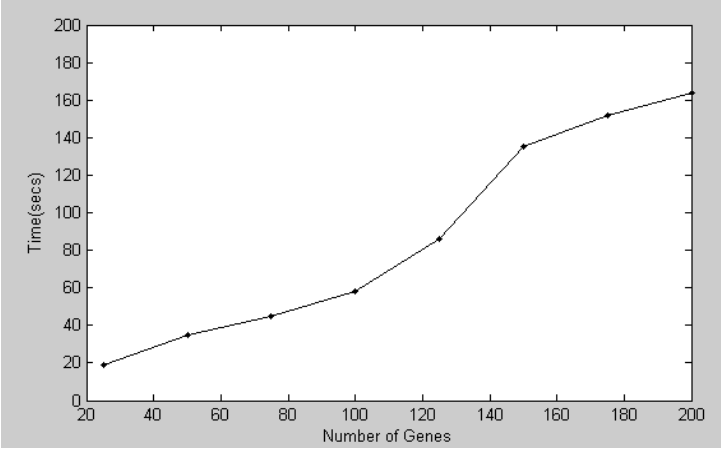


Figure 7.12 Time required for TRNFN to predict the regulators for a gene from a specific dataset. The x-axis represents the subset of genes selected ranging from 25 to 200 with a step increase of 25.

Moreover, the above algorithms followed to infer the regulators of every gene present in a dataset can be easily implemented using parallel techniques of processing. Such a framework will reduce the computational time needed for GRN reconstruction and allow the methods to be feasibly applicable to much larger datasets.

The results prove the efficiency and accuracy of TRNFN in successfully capturing the interactions among the considered genes, despite of the noise inherited from the microarray hybridization and the small amount of samples in the data. In the view of this fact, the regulatory relations suggested by TRNFN while analyzing plasma RNA data of colon cancer data set are likely to exist, but further biological experiments are required to validate this result.

Chapter 8

MicroRNA-mRNA Interaction Network

Contents

MicroRNAs are short non-coding RNAs that can regulate gene expression during various crucial cell processes such as differentiation, proliferation and apoptosis. Changes in expression profiles of miRNA play an important role in the development of many cancers, including CRC. Therefore, the identification of cancer related miRNAs and their target genes are important for cancer biology research. In this work, TSK-type recurrent neural fuzzy network (TRNFN) was applied to infer miRNA-mRNA association network from paired miRNA, mRNA expression profiles of CRC patients. It was demonstrated that TRNFN achieved good performance in recovering known experimentally verified miRNA mRNA associations.

8.1 Introduction

MicroRNAs (miRNAs) constitute a class of small non-protein-coding RNAs which regulates protein-coding genes at the post transcriptional and translational level [20]. They play important roles in the control of many biological processes, such as cell development, differentiation, proliferation and apoptosis. Accumulating evidence suggests that altered

miRNAexpression correlates with the pathogenesis of cancers. The over-expression of several miRNAs results in tumor formation; however, some miRNAs are consistently downregulated in tumors and may have tumor-suppressive effects [164, 165]. For example, microRNAs in the let-7 and miR-34 families may act as tumor suppressors by repressing certain oncogenes [38, 166] while miR-106b and miR-21 play roles in oncogenesis [167, 168]. A recent study suggested that microRNAs can identify cancer tissue origin accurately [169]. This is of great clinical importance because microRNAs may be used for tracing the tissue of origin of cancers of unknown primary origin. Thus the identification of miRNAs linked to cancer susceptibility is useful for cancer diagnosis, prognosis, treatment and drug target discovery.

Recent experiments also show that miRNAsupregulate genes in one condition, but act as a negative regulator in another condition. For example, let7 and the synthetic microRNA, miRcxcr4-likewise upregulate target mRNAs upon cell-cycle arrest; yet, they inhibit translation in proliferating cells [39]. The abundance and diversity of miRNA targets result in a large number of possible miRNA regulatory mechanisms. It is not feasible to test all the possibility through biological experiments. Therefore, the development of various computational methods to recognize crucial regulatory functions of miRNA has been widely applied to cancer research as a powerful supplement to experimental methods.

Uncontrolled growth of cells and loss of apoptosis function usually results in cancer formation. MicroRNAs have been found to regulate mechanisms such as cell growth and apoptosis [170]. Recognition of

miRNAs that are differentially expressed in tumor and normal tissues may help to identify those miRNAs that are involved in pathogenesis of human cancers. Experimental methods such as microarray profiling and qRT-PCR have been used to monitor expression levels of miRNAs in various types of cancers. Microarray profiling is a powerful technique that can be used to systematically detect the differential expression of miRNAs in cancer and normal samples. By integrating mRNA target genes, cancer genes, miRNA and mRNA expression profile information, an interaction network can be developed to link miRNAs to cancer target genes.

Several computational methods have been proposed to study miRNA regulatory mechanisms through expression data. Huang *et.al.* used Bayesian data analysis algorithm (Gene MIR++) [171] to identify miRNA targets by utilizing paired expression profiles of miRNA and mRNA. However the algorithm used in that paper is based on pairwise correlation method which may fail to undermine collinearities amongst the covariates (miRNAs).

Partial Least Square (PLS) regression approach [172] by Xiaohong*et.al.* overcomes these issues and explores likely associations between miRNA and mRNA by taking advantage of the known inverse relationship between miRNAs and target mRNAs. In this work, TSK type recurrent neural fuzzy network was used to model miRNA-mRNA interaction network from paired expression profiles of miRNA and mRNA for both colon tumors and normal tissues.The schematic diagram of whole procedure for capturing complex miRNA–mRNA association network is shown in figure 8.1.

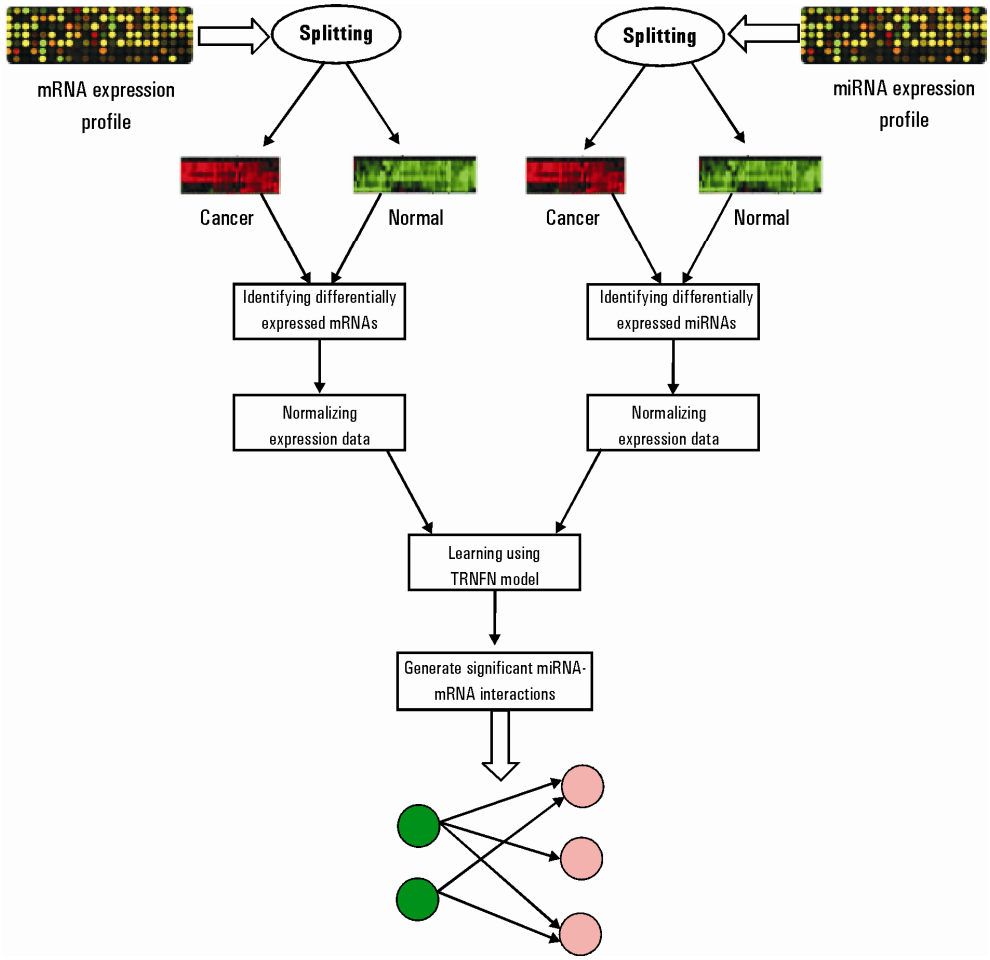


Figure 8.1 Schematic diagram of the overall procedure for generating miRNA-mRNA interaction network

8.2 miRNA-mRNA Interaction Network

TRNFN was applied to the paired miRNAand mRNA microarray data sets for both human colon tumors and normal tissues obtained from the cancer dataset provided by Broad Institute [173, 174]. The dataset originally consist of the expression profiles of 218 tumor samples representing 14

common human cancer classes out of which 10 are colon tumor tissue samples and 5 are normal colon tissue samples. Out of this, 7 colon tumor tissue samples and 4 normal colon tissue samples passed quality control criteria were taken for further analysis [175]. Unpaired t-tests were used to identify a set of miRNAs which are differentially expressed in different conditions under the investigation. We selected 56 miRNAs whose p-value less than 0.05 for further analysis.

From the previous work, 27 genes were identified, whose regulatory relations are already inferred using different approaches. In this work, the above 27 genes were analyzed to model the miRNA-mRNA association network. Twenty seven multi input - single output TRNFN models were implemented. Each model describes the state of output gene based on the expression value of the 56 miRNAs. Structure of each model is generated using the training dataset and the parameter tuning is done by the repeated learning. Training and test dataset contains samples from both healthy and unhealthy tissues. The fuzzy rule set derived from 27 TRNFN model was used to build a miRNA-mRNA interaction network. The inferred association network using TRNFN was shown in figure 8.2. There were two disjoint sets of nodes in this graph, miRNA (green circle) and mRNA genes (pink circle). A direct connection was placed from a miRNA to an mRNA indicates that the mRNA was predicted to be the target of the miRNA. An edge → indicates activation of transcription, whereas, an edge ⊣indicates repression of transcription. Cytoscape software

[157] has been used to draw the network. The resulting network had 76 nodes and 119 connections.

In order to better understand the biological processes linked to these miRNAs and their predicted target mRNAs in the context of colorectal cancer, the results were compared with the experimentally known miRNA mRNA association available in literature. MicroRNA.org [176] is a widely used web resource for the miRNA target prediction and expression profiles. In microRNA.org database, target predictions are based on a development of the miRanda algorithm [177] which incorporates current biological knowledge on target rules. The set relations predicted by TRNFN and are confirmed in mircoRNA.org database are listed in Table 8.1. For example, the down regulation of UBE2D3 by miR-107, miR-135 and miR-140 predicted correctly by TRNFN and are confirmed in the literature. Further, by searching miR2Disease: a manually curated database for microRNA deregulation in human disease [178], it was found that 17 of 56 miRNAs are known to be actively involved in the pathways associated with colorectal cancer. Table 8.2 lists each of those miRNAs and their target genes which have been previously reported to have associations with colorectal cancer.

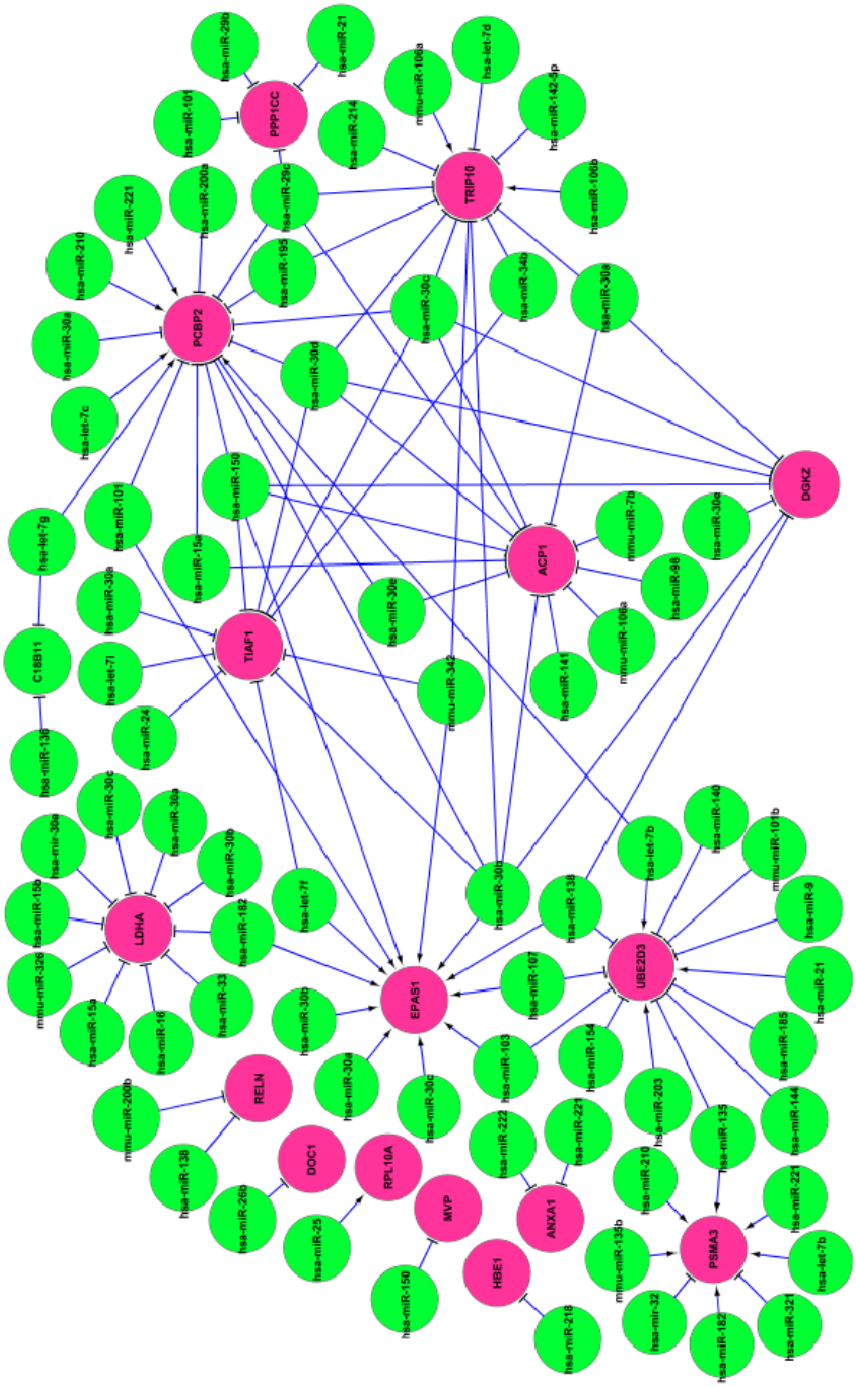


Figure 8.2miRNA -mRNA interaction network predicted by TRNFN

Table 8.1 Set of known relations predicted by TRNFN

Target Genes	miRNAs Associated
EPAS1	hsa-miR-103, hsa-miR-107, hsa-miR-138, hsa-miR-150, hsa-miR-182, hsa-miR-30b
ANXA1	hsa-miR-221,hsa-miR-222
C18B11	hsa-let-7g,hsa-miR-136
ACP1	hsa-miR-141, hsa-miR-18, hsa-miR-98, mmu-miR-106a
HBE1	hsa-miR-218
LDHA	hsa-miR-15a,hsa-miR-15b,hsa-miR-16,hsa-miR-182,hsa-miR-30a,hsa-miR-30b, hsa-miR-30c, hsa-mir-30e,hsa-miR-33
MVP	hsa-miR-150
PCBP2	hsa-let-7a,hsa-let-7b, hsa-let-7c,hsa-miR-150,hsa-miR-15a, hsa-miR-195,hsa-miR-200a
PPP1CC	hsa-miR-21
PSMA3	hsa-let-7b,hsa-miR-135,hsa-miR-182,hsa-miR-210, hsa-miR-221, hsa-mir-32,mmu-miR-135b
RELN	hsa-miR-200c, hsa-miR-138,mmu-miR-200b
TIAF1	hsa-miR-150,hsa-miR-24, hsa-miR-30a,hsa-miR-30b,hsa-miR-30c,hsa-miR-30d
TRIP10	hsa-let-7d, hsa-miR-106b, hsa-miR-142-5p, hsa-miR-214, hsa-miR-195,mmu-miR-106a
UBE2D3	hsa-let-7b, hsa-miR-103, hsa-miR-107, hsa-miR-135,hsa-miR-138, hsa-miR-140, hsa-miR-144 , hsa-miR-154,hsa-miR-185, hsa-miR-185,hsa-miR-203 , hsa-miR-21, hsa-miR-9,mmu-miR-101b

It is interesting to investigate further the biological process and cancer related canonical pathways associated with these miRNAs which are associated with colon cancer. To obtain the biological process in which the above cancer related genes are involved, the miR- Ontology database (miRo) was used [179].

The results are presented in Table 8.3. A more detailed functional analysis has been done to identify the cancer related canonical pathways in which these miRNAs and target genes are involved. The identified pathways are listed in Table 8.4.Overall, it is clear that the above CRC related miRNAs are involved in many biological process and pathways by regulating the target genes predicted by TRNFN.

Table 8.2 List of 17 miRNAs and target genes associated with colorectal cancer

CRC related miRNAs	miRNAFamiliy	Target Genes Predicted by TRNFN
hsa-let-7b	Let-7b	PCBP2,UBE2D3,PSMA3
hsa-let-7g	Let-7g	PCBP2,C18B11
hsa-miR-106b	miR-106	TRIP10
hsa-miR-107	miR-107	EPAS1,UBE2D3
hsa-miR-140	miR-140	UBE2D3
hsa-miR-141	miR-141	ACP1
hsa-miR-15b	miR-15	LDHA
hsa-miR-182	miR-182	LDHA,EPAS1,PSMA3
hsa-miR-195	miR-195	PCBP2,TRIP10
hsa-miR-203	miR-203	UBE2D3
hsa-miR-21	miR-21	PPP1CC,UBE2D3
hsa-miR-221	miR-221	ANXA1,PCBP2,PSMA3
hsa-miR-25	miR-25	RPL10A
hsa-miR-29b	miR-29	PPP1CC
hsa-miR-30c	miR-30	LDHA,PCBP2,ACP1,TIAF1,TRIP10,DGKZ,EPAS1
hsa-mir-32	mir-32	PSMA3
hsa-miR-34b	miR-34	TIAF1,TRIP10

Table 8.3 CRC related miRNAs and their associated Process

CRC related miRNA	Associated Biological Process
hsa-miR-34b	Regulation of cyclin-dependent protein kinase activity,angiogenesis,ubiquitin-dependent protein catabolic process,regulation of transcription from RNA polymerase II promoter,regulation of apoptosis
hsa-miR-32	Beta-catenin binding,regulation of apoptosis,BMP signaling pathway,regulation of cell proliferation,ubiquitin-dependent protein catabolic process
hsa-miR-21	Cell proliferation; anti-apoptosis,ubiquitin-protein ligase activity,
hsa-let-7g	Cell proliferation; anti-apoptosis,ubiquitin-protein ligase activity,Anaphase-promoting complex(APC)-dependent proteasomal ubiquitin-dependent protein catabolic process, response to oxidative stress ,regulation of transcription, DNA-dependent
hsa-miR-140	cell cycle arrest, regulation of cell proliferation
hsa-let-7b	Cell proliferation,ubiquitin-protein ligase activity,Anaphase-promoting complex(APC)-dependent proteasomal ubiquitin-dependent protein catabolic process, response to oxidative stress and regulation of apoptosis,hemopoiesis
hsa-miR-30c	Polyamine biosynthetic process, ubiquitin-dependent protein catabolic process, and regulation of apoptosis
hsa-miR-106b	Angiogenesis,regulation of cyclin-dependent protein kinase activity,regulation of cell proliferation, regulation of transcription from RNA polymerase II promoter
hsa-miR-107	Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process,anti-apoptosis,angiogenesis,BMP signaling pathway,Wnt receptor signaling pathway
hsa-miR-15b	Anti-apoptosis,regulation of cell proliferation,Anaphase-promoting complex(APC)-dependent proteasomal ubiquitin-dependent protein catabolic process
hsa-miR-182	BMP signaling pathway,apoptosis,angiogenesis,activation of MAPK activity, regulation of Wnt receptor signaling pathwaythrough beta-catenin,proteinubiquitination during ubiquitin-dependent protein catabolic process,
hsa-miR-195	Wnt receptor signaling pathway,BMP signaling pathway,anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process,regulation of apoptosis
hsa-miR-141	Angiogenesis,regulation of transcription, DNA-dependent,hemopoiesis,regulation of cell proliferation
hsa-miR-203	Regulation of cell growth,activation of MAPKK activity,anti-apoptosis,BMP signaling pathway,
hsa-miR-221	Cytokine and chemokine mediated signaling pathway,regulation of Wnt receptor signaling pathway,regulation of Wnt receptor signaling pathway
hsa-miR-25	Angiogenesis,anti-apoptosis,BMP signaling pathway,regulation of cell proliferation,regulation of Wnt receptor signaling pathway,ubiquitin-dependent protein catabolic process
has-miR-29b	Regulation of cell proliferation,angiogenesis,BMP signaling pathway,regulation of transcription from RNA polymerase II promoter,regulation of apoptosis

Table 8.4 CRC related miRNAs and target genes involved in cancer related canonical pathways

Cancer Related Canonical Pathways	Target genes in Respective Pathways	The no. of CRC Related miRNAs associated with the target Genes
Hypoxia Signaling Pathway	EPAS1,LDHA,UBE2 D3	8
ProteinUbiquitination Pathway	PSMA3, UBE2D3	8
Polyamine regulation in Colon cancer	PSMA3	4
Pyruvate Metabolism	LDHA	3
Insulin Signaling Pathway	TRIP10,PPP1CC,EP AS1	8
Renal Cell Carcinoma	EPAS1	3
Signaling by Wnt	PSMA3	4
Apoptosis	PSMA3,TIAF1	6
Cdc20:Phospho-APC/C mediated degradation of Cyclin A	PSMA3	4
Regulation of activated PAK-2p34 by proteasome mediated degradation	PSMA3	4
APC/C:Cdh1-mediated degradation of Skp2	PSMA3	4
VEGF Signaling pathway	EPAS1	3

Using TRNFN, miRNA and mRNA expression profiles derived from the same set of cancerous tissue samples was analyzed to infer the molecular mechanism for the colon cancer etiology. The above results shows that TRNFN achieved good performance in capturing known experimentally verified miRNA mRNA associations. Moreover, it was found that 17 miRNAs are directly linked with colorectal cancer. Focusing such miRNAscan benefit the development of targeted and personalized miRNA-based anti-cancer therapies.

…….ଓ୦ଓ…….

Chapter 9
Conclusion and Future Work

Cancer is a complex disease arising from the progressive accumulation of many genetic alterations. Identifying cancerous genes and the pathways they control is a key step towards the treatment of cancer. Microarray based tumor profiling provides a better platform for comparing mutant or diseased cells with normal cells and for searching differences in gene expressions that can be the potential key factors leading to diseases. The major focus on microarray analysis is the reconstruction of gene regulatory network from the measured dataset of gene expression. Modelling or understanding network of genomic regulation specific to a cellular state, such as subtype of tumor, is important for improving the understanding of the mechanisms underlying these diseases and provides potential targets for therapeutic studies. In this study, four soft computing algorithms were applied to reconstruct gene regulatory network that reveals global patterns of gene interactions in cancer, particularly colorectal cancer, which is a leading cause of cancer related death worldwide.

Tumor-derived RNA was detected in the plasma or serum of patients with different types of cancer. The concentration of RNA in the blood of cancer patients is higher than that of the healthy individuals. As circulating RNAs are easily accessible in the blood and appear to be relevant surrogate materials for genetic alterations present in the primary tumor, expression

profiling of plasma RNA may have great potential especially for early detection and prognostic evaluation. By analyzing the circulating plasma RNA dataset of colorectal cancer patients, in this work, potential tumor markers in the blood and their regulatory relations were identified. The inferred relations were validated by comparing with the regulatory relations identified from the microarray data from colon tumor.

The study was set to investigate the effective methods to extract regulatory relations found in the blood of colorectal cancer patients and to devise better way to tackle the problem through computational approaches. For that in this work, the effectiveness four methods such as fuzzy logic approach, modified genetic algorithm, dynamic feed forward neural fuzzy network and TSK-type recurrent neural fuzzy network fuzzy logic algorithm were investigated. The advantages of each method were taken into consideration to make solutions. In the first application of regulatory network modelling, the fuzzy GRN model, information from microarray data was extracted in the form of fuzzy rules. Due to the high level human like reasoning, the model can deal with the uncertainties of modelling noisy data. It was shown that clustering as a pre-processing step in building a fuzzy-logic model could save a significant amount of time. The second approach, modified genetic algorithm, was based on the mechanics of natural evolution and natural genetics. Genetic algorithm was applied for optimizing the weight matrix for gene regulatory network. By incorporating statistical technique viz. correlation analysis, search space has been immensely reduced.

The other two approaches, dynamic feed forward neural fuzzy network and TSK-type recurrent neural fuzzy network, were proposed to join the soft computing approaches such as neural network and fuzzy logic to infer gene regulatory network from microarray expression data. This hybridized model combines the features of connectionist and fuzzy logic approaches and infers information on gene interactions in the form of fuzzy rules and considers the dynamic aspects of gene regulation. Unlike, fuzzy GRN model, there was no predefined rules in DNFN and TRNFN, all of them are constructed during online learning. However, DNFN has the drawback of requiring prior data discretization and has the additional disadvantage of not considering temporal information. The recurrent structure of TRNFN ensures that both spatial and temporal information are properly retained while, at the same time, the self-organizing properties of the method automatically account for the discretization process.

Each analysis method introduced in this thesis employs different modelling approaches and each captures different features. In such situations, the only way to compare the effectiveness of each algorithm is to demonstrate the methods on different benchmark datasets publically available. In all the tested cases, TRNFN identified more gene interactions and gave better recall than other computational approaches. 87.8% of the total interactions extracted by TRNFN are correct in accordance with the biological knowledge, outperforming other approaches.

The models used in this study captured the regulatory relationship among 27 differentially expressed genes from plasma RNA of CRC patient. From a detailed pathway analysis it was found that most of these genes were

actively involved in the cancer related canonical pathways. Moreover, three genes- EPAS1, PSMA3 and UBE2D3 have been already identified as commonly over expressed genes in colon cancer patients. The simulated GRNs highlight the crucial role of above genes in cancer specific regulatory network. In short, all the above models can be used to identify those genes which are affected by the up regulation of cancer markers.

In certain cases, the interactions identified from these studies could not be retrieved in the lists of experimental interactions currently existing in databases.Since the available experimental data is still far from complete, it might be possible that some of these interactions are valid but currently unknown. However, those interactions and especially the ones that lie within intersections could be more accurate should be further examined.

By analyzing the gene regulatory network extracted, two meaningful findings were obtained. First, up regulated genes are regulated by more genes than down regulated genes. Second, tumor activators activate other tumor activators and suppress tumor suppressers strongly in the disease environment. To a certain extent, these assumptions are reasonable and relevant. The utility of these approaches and the reliability of these conclusions were kept in abeyance for further experimental validation.

Recently it has been reported that the MicroRNAs play an important role in the development of many cancers, including CRC. Therefore, identifying cancer related miRNAs and their target genes is a key step towards the diagnosis and treatment of cancer. In this work, TRNFN was applied to infer miRNA-mRNA association network from microarray gene expression data of CRC patients. Using TRNFN, miRNAs which are

involved in the regulation of a set of cancerous genes were identified. It was demonstrated that the method achieved good performance in recovering known experimentally verified miRNA mRNA associations. Moreover, TRNFN was successful in identifying 17 miRNAs which are directly involved in the CRC related pathways. Targeting such miRNAs may help not only to prevent the recurrence of disease but also to control the growth of advanced metastatic tumors. These findings will help to elucidate the common molecular mechanism of colon cancer, and provide new insights into cancer diagnostics, prognostics and therapy.

Suggestions for Future Works

The computational approaches presented in this thesis hold great promise for elucidating the structure, parameters and eventually the dynamics of gene regulation networks. There are, however, limitations to these methods which should be acknowledged, some of which point at promising future research directions.

In this study, unpaired student t-test had been employed to identify genes differentially expressed under different conditions using data from microarray experiments. In microarray data analysis, t-test is often used since samples may be derived from different physical locations and may not have the same distribution. However, the power of standard t-test is limited when the number of samples is small and so some of the standard deviations will be extremely small, and therefore the test statistics will be very high, which may lead to a significant bias. In such situations, classifiers

such as support vector machine (SVM) or k-nearest neighbor (KNN) classifier are more attractive in identifying discriminative genes.

Cluster analysis of microarray data is essential for identifying biologically relevant groups of genes. The hybrid clustering algorithm used in this work shows good performance in the partitioning of genes based on their expression profiles. However, hybrid clustering algorithm performs a one-to-one mapping: one gene belongs to exactly one cluster. Genes can participate in multiple pathways and are frequently coordinated by a variety of regulatory mechanism. For the analysis of microarray data, it may expect that single genes can belong to several clusters. In this respect, fuzzy clustering approaches can be taken in consideration because of their capability to assign one gene to more than one cluster, which may allow capturing genes involved in multiple transcriptional programs and biological processes.

Like other existing methods for GRN inference, the methods used in this study also have some limitations owing to the nature of microarray gene expression data. In particular, although the inferred GRN can provide activator or inhibitory relations among genes, such a relationship does not show the exact mechanism. A predicted regulatory relationship does not always mean direct genetic regulation. Some regulation can be at the post-transcriptional or post-translational levels, which are often not reflected in mRNA expression levels detected by microarrays. Hence, the GRN models that predicted from microarray data include both direct and indirect regulations (i.e. via hidden variables). Therefore, there is a need for

integration with other information sources like proteomic data and metablomic data, to derive regulatory networks in an accurate manner.

Currently, this study is aimed to identify the potential cancer markers in the blood of colorectal cancer patients and the regulatory relationships among these tumor markers. With the similar mechanism, it can be extended to identify the biomarkers and their regulatory pathways from the datasets of different types of cancers or other diseases. By making accurate predictions on the gene network state, these models can assist in the development of early-detection tests and disease treatment.

Although in the present study, the number of genes considered has limited to be 27, the methods could be easily modified to process bigger sets.

References

[1] Bert Vogelstein and Kenneth W. Kinzler. “Cancer genes and the pathways they control.” *Nature Medicine*, vol. 10, no. 8, pp. 789-799, 2004.

[2] American Cancer Society. “Colorectal Cancer Facts & Figures 2011-2013.” *Atlanta: American Cancer Society*, 2011.

[3] A. Jemal, R. Siegel, E. Ward, T. Murray, T. Xu and M.J. Thun. “Cancer statistics.” *CA: A Cancer Journal of Clinicians*, vol. 57, no. 1, pp. 43-66, 2007.

[4] P.A Bryant, G.K. Smyth, R. Robins-Browne and N. Curtis. “Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation.” *PLoS ONE,* vol. 6, no. 5, 2011.

[5] Adeline R. Whitney, Maximilian Diehn, Stephen J. Popper, Ash A. Alizadeh, Jennifer C. Boldrick, David A. Relman and O. Patrick. “Brown Individuality and variation in gene expression patterns in human blood,” in *Proc. National Academy of Sciences of the United States of America(PNAS)*, 2003, pp. 1896-1901.

[6] Ron M. Kerkhoven, Daoud Sie, Marja Nieuwland, Mike Heimerikx, Jorma De Ronde, Wim Brugman, and Arno Velds. “The T7-primer is a source of experimental bias and introduces variability between microarray platforms” *PLoS ONE,* vol. 3, 2008.

[7] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A. Ball, Helen C. Causton, Terry Gaasterland, Patrick Glenisson, Frank C.P. Holstege, Irene F. Kim, Victor Markowitz, John C. Matese, Helen Parkinson, Alan Robinson, Ugis Sarkans, Steffen Schulze-Kremer, Jason Stewart, Ronald Taylor, Jaak Vilo1 and Martin Vingron. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data." *Nature Genetics*, vol. 29, no. 4, pp. 365-371, 2001.

[8] T. Gardner and J. Faith. "Reverse-engineering transcription control networks." *Physics of Life Reviews*, vol. 2, pp. 65-88, 2005.

[9] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato and Diego di Bernardo. "How to infer gene networks from expression profiles" *Molecular System Biology,* vol. 78, no. 3, 2007.

[10] T. Akutsu , S. Miyano and S. Kuhara. "Identification of genetic networks from a small number of gene expression patterns under the boolean network model." in *Proc. Pacific Symposium on Biocomputing*, pp. 17-28, 1999.

[11] N. Friedman, M. Linial, I. Nachman and D. Peer. "Using bayesian network to analyze expression data." *Journal for Computational Biology*, vol. 7, no. 3-4, pp. 601-620, 2000.

[12] M. Kato, T. Tsunoda and T. Takagi. "Inferring genetic networks from DNA microarray data by multiple regression analysis." *Genome Informatics*, vol. 11, pp. 118-128, 2000.

[13] P.J. Woolf and Y. Wang. “A fuzzy logic approach to analyzing gene expression data.” *Physiological Genomics*, vol. 3, no. 1, pp. 9-15, 2000.

[14] Shin Ando and Hitoshi Iba. “Inference of Gene Regulatory Model by Genetic Algorithms.” in *Proc. IEEE Congress on Evolutionary Computation*, pp. 712-719, 2001.

[15] Shin Ando and Hitoshi Iba. ”Estimation of gene Regulatory Network by Genetic Algorithm and Pairwise Correlation Analysis.” in *Proc. IEEE Congress on Evolutionary Computation*, pp. 207-214, 2003.

[16] Manuel Collado, Vanesa Garcia, Jose Miguel Garcia, Isabel Alonso, Luis Lombardia , Ramon Diaz-Uriarte, A. Luis, Angel Zaballos, Felix Bonilla and Manuel Serrano. “Genomic profiling of circulating plasma RNA for the analysis of cancer.”, *Clinical Chemistry,* vol. 53, no. 10, pp. 1860-1863, 2007

[17] Ondrej Slaby, Marek Svoboda, Jaroslav Michalek and Rostislav Vyzula. “MicroRNAs in colorectal cancer: translation of molecular biology into clinical application.” *Molecular Cancer*, vol. 8, no. 1, 2009.

[18] http://www.medicinenet.com/colon_cancer/article.htm

[19] Fransis Crick. “Central dogma of molecular biology.” *Nature*, vol. 227, no. 5258, pp. 561-563, 1970.

[20] R.W. Carthew. “Gene regulation by microRNAs.” *Current Opinion in Genetics and Development,* vol. 16, pp. 203–208, 2006.

[21] B.P. Lewi, C.B. Burge and D.P. Bartel. "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." *Cell,* vol. 120, no. 1, pp. 15–20, 2005.

[22] R.C. Friedman, K.K Farh, C.B. Burge and D.P. Bartel. "Most mammalian mRNAs are conserved targets of microRNAs." *Genome Research*, vol. 19, no. 1, pp. 92–105, 2009.

[23] N. Lynam-Lennon, S.G. Maher and J.V. Reynolds. "The roles of microRNA in cancer and apoptosis." *Biological Reviews of the Cambridge Philosophical Society*, vol. 84, no. 1, pp. 55–71, 2009.

[24] R. Schickel, B. Boyerinas, S.M. Park and M. E. Peter. "MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death." *Oncogene*, vol. 27, no. 45, pp. 5959–5974, 2008.

[25] R.C. Lee, R.L. Feinbaum and V. Ambros. "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." *Cell,* vol. 75 no. 5 pp. 843–54, 1993.

[26] D. Baek, J. Villen, C. Shin, F. D Camargo, S. P. Gygi, and D. P. Bartel. "The impact of microRNAs on protein output." *Nature*, vol. 455, pp. 64-71, 2008.

[27] M. Selbach, B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. "Widespread changes in protein synthesis induced by microRNAs." *Nature,* vol. 455 pp. 58-63, 2008.

[28] A. Rodriguez, S. Griffiths-Jones, J.L. Ashurst and A. Bradley. "Identification of mammalian microRNA host genes and

transcription units." *Genome Research*, vol. 14, no. 10A, pp. 1902–1910, 2004.

[29] Y. Lee, M. Kim, J. Han, K.H. Yeom, S. Lee, S.H. Baek and V.N. Kim. "MicroRNA genes are transcribed by RNA polymerase II." *EMBO Journal*, vol. 23 no. 20, pp. 4051–4060, 2004.

[30] E. Lund and J. Dahlberg. "Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs" in *Proc. Cold Spring Harbor symposia on quantitative biology*, no. 71, 2006 pp. 59–66.

[31] A. Khvorova, A. Reynolds and S.D. Jayasena. "Functional siRNAs and miRNAs exhibit strand bias." *Cell,* vol. 115, pp. 209-216, 2003.

[32] D. S. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin and P.D. Zamore. "Asymmetry in the assembly of the RNAi enzyme complex." *Cell*, vol. 115, pp. 199-208, 2003.

[33] George Adrian Calin, Cinzia Sevignani, Calin Dan Dumitru, Terry Hyslop, Evan Noch, Sai Yendamuri, Masayoshi Shimizu, Sashi Rattan, Florencia Bullrich, Massimo Negrini and Carlo M. Croce. "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers," in *Proc. National Academy of Sciences of the United States of America*, 2004, pp. 2999–3004.

[34] G.A. Calin and C.M. Croce. "MicroRNA signatures in human cancers." *Nature Reviews Cancer*, vol. 6, pp. 857–866, 2006.

[35] Arti Gaur, David A. Jewell, Yu Liang, Dana Ridzon, Jason H. Moore, Caifu Chen, Victor R. Ambros and Mark A. Israel. "Characterization of microRNA expression levels and their

biological correlates in human cancercell lines." *Cancer Research*, vol. 67, no. 6, pp. 2456–2468, 2007.

[36] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, Laura Rassenti, Thomas Kipps, Massimo Negrini, Florencia Bullrich and Carlo M. Croce. "Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia," in *Proc. National Academy of Sciences of the United States of America*, 2002, pp. 15524–15529.

[37] L. He, J.M. Thomson, M.T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S.W. Lowe, G.J. Hannon and S.M. Hammond. "A microRNA polycistron as a potential human oncogene." *Nature,* vol. 435, pp. 828–833, 2005.

[38] S.M Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier , K.L. Reinert, D. Brown and F.J. Slack. "RAS is regulated by the let-7 microRNA family." *Cell,* vol. 120, no. 5, pp. 635–647, 2005.

[39] S. Vasudevan, Y. Tong and J.A. Steitz. "Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation." *Science*, vol. 318, no. 5858, pp. 1931-1934, 2007.

[40] Eric R. Fearon and Bert Vogelstein. "A genetic model for colorectal tumorigenesis." *Cell*, vol. 61, no. 5, pp. 759-767, 1990.

[41] Stefano Volinia, George A. Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio,

Claudia Roldo, Manuela Ferracin, Robyn L. Prueitt, Nozumu Yanaihara, Giovanni Lanza, Aldo Scarpa, Andrea Vecchione, Massimo Negrini, Curtis C. Harris and Carlo M. Croce. "A microRNA expression signature of human solid tumors defines cancer gene targets," in *Proc. National Academy of Sciences of the United States of America*, 2006, pp. 2257–2261.

[42] Supratim Choudhuri. "Microarrays in Biology and Medicine." *Journal of Biochemical and molecular Toxicology*, vol. 18, no. 4, pp. 171-174, 2004.

[43] Mark Schena, Dari Shalon, Ronald W. Davis and Patrick O. Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science*, vol. 270 no. 5235, pp. 467-470, 1995.

[44] Martin Dufva. "Production of DNA microarrays for biomedical research." *Biotech International,* pp. 1-4, 2006.

[45] Martin Dufva. "Fabrication of high quality microarrays." *Biomolecular Engineering,* vol. 22, no. 5-6, pp. 173-184, 2005.

[46] S. Singh-Gasson, R.D. Green, Y. Yue, C. Nelson, F. Blattner, M.R. Sussman and F. Cerrina. "Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array." *Nature Biotechology,* vol. 17, no. 10, pp. 974-978, 1999.

[47] John Quackenbush. "Microarray data normalization and Transformation." "*Nature genetics supplement*, vol. 32, pp. 496-501, 2002.

[48] J. Dopazo, E. Zanders, I. Dragoni, G. Amphlett and F. Falciani. "Methods and approaches in the analysis of gene expression data." *Journal of Immunological Methods*, vol. 250, pp. 93–112, 2001.

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

[50] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer and Z. Yakhini. "Tissue Classification with Gene Expression Profiles." *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 559-583, 2000.

[51] Charles M. Perou, Stefanie S. Jeffrey, Matt van de Rijn, Christian A. Rees, Michael B. Eisen, Douglas T. Ross, Alexander Pergamenschikov, Cheryl F. Williams, Shirley X. Zhu, Jeffrey C. F. Lee, Deval Lashkari, Dari Shalon, Patrick O. Brown and David Botstein. "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," in *Proc. National Academy of Sciences of the United States of America*, 1999, pp. 9212-9217.

[52] Han-yu Chuang, Huai-kuang Tsai, Yuan-fan Tsai and Cheng-yan Kao. "Ranking Genes for Discriminability on Microarray Data." *Journal of Information Science and Engineering*, vol. 19 no. 6, pp. 953-966, 2003.

[53] V.G. Tusher, R. Tibshirani and G. Chu. "Significance analysis of microarrays applied to the ionizing radiation response," in *Proc. National Academy of Sciences of the United States of America*, 2001, pp. 5116–5121.

[54] A.J. Butte, J. Ye, G. Niederfellner, K. Rett, H.U. Haring, M.F. White and I.S. Kohane. "Determining significant fold differences in gene expression analysis," in *proc. Pacific Symposium on Biocomputing*, 2001, pp. 6–17.

[55] P.J. Park, M. Pagano and M. Bonetti. "A nonparametric scoring algorithm for identifying informative genes from microarray data," in *Proc. Pacific Symposium on Biocomputing*, 2001, pp. 52–63.

[56] Paul Pavlidis and William Stafford Noble. "Analysis of strain and regional variation in gene expression in mouse brain." *Genome Biology*, vol. 2, no. 10, pp. research0042.10–0042.15, 2001.

[57] J.R. Quinlan. "Induction of Decision Trees." *Machine Learning*, vol. 1, pp. 81-106, 1996.

[58] D.P. Berrar, C.S. Downes and W. Dubitzky. "Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks," in *Proc. Pacific Symposium on Biocomputing*, 2003, pp. 5-16.

[59] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michèl Schummer and David Haussler. "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data." *Bioinformatics*, vol. 16, no. 10, pp. 906-914, 2000.

[60] Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., and David Haussler. "Knowledge-based analysis of microarray gene expression data by using support vector machines," in *Proc.*

National Academy of Sciences of the United States of America, 2000, pp. 262-267.

[61] O. Alter, P.O. Brown and D. Botstein. "Singular value decomposition for genome-wide expression data processing and modeling," in *Proc. National Academy of Sciences of the United States of America*, 2000, pp. 10101–10106.

[62] S. Raychaudhuri, J.M. Stuart and R.B. Altman. "Principal components analysis to summarize microarray experiments :application to sporulation time series," in *Proc. Pacific Symposium on Biocomputing*, 2000, pp. 455-466.

[63] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub. "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," in *Proc. National Academy of Sciences of the United States of America*, 1999, pp. 2907–2912.

[64] P. Toronen, M. Kolehmainen, G. Wong and E. Castren. "Analysis of gene expression data using self-organizing maps." *FEBS Letters,* vol. 451, no. 2, pp. 142–146, 1999.

[65] Fang-Xiang Wu, Wen-Jun Zhang and Anotony J Kusalik. "Determination of minimum sample size in Microarray Experiments to cluster genes using K-means Clustering," in *Proc. the third IEEE Symposium on Bioinformatics and BioEngineering*, 2003, pp. 401-406.

[66] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein. "Cluster analysis and display of genome-wide expression patterns," in *Proc.*

National Academy of Sciences of the United States of America, 1998, pp. 14863–14868.

[67] S. Liang, S. Fuhrman and R. Somogyi. "Reveal, a general reverse engineering algorithm for inference of gene network architectures," in *Proc. Pacific Symposium on Biocomputing*, 1998, pp. 18-29.

[68] G. Marnellos and E.D. Mjolsness. *Modeling neural development*. MIT Press, Cambridge, 2003, pp. 27-28.

[69] J. Vohradsky. "Neural network model of gene expression." *The FASEB Journal,* vol. 15, no. 3, pp. 846-854, 2001.

[70] Zainal A. Hasibuan, Romi Fadilah Rahmat and Muhammad Fermi Pasha. "Adaptive nested neural network (ANNN) based on human gene regulatory network for gene knowledge discovery engine." *International Journal of Computer Science and Network Security*, vol. 9, no. 6, pp. 43-53, 2009.

[71] K.M. Eyster and R. Lindahl. "Molecular medicine: a primer for clinicians. Part XII: DNA microarrays and their application to clinical medicine." *South Dakota journal of medicine*, vol. 54, no. 2, pp. 57-61, 2001.

[72] Manjula Kurella, Li-Li Hsiao, Takumi Yoshida, Jeffrey D. Randall, Gary Chow, Satinder S. Sarang, Roderick V. Jensen and Steven R. Gullans. "DNA microarray analysis of complex biologic Processes." *Journal of the American Society of Nephrology*, vol. 12, no. 5, pp. 1072-1078, 2001.

[73] J Petrik. "Microarray technology: the future of blood testing?" *Vox Sang*, vol. 80, no. 1, pp. 1-11, 2001.

[74] M.J. Kuhar, A. Joyce and G. Dominguez. "Genes in drug abuse." *Drug Alcohol Depend,* vol. 62, no. 3, pp. 157-162, 2001.

[75] J.M. Lewohl, P.R. Dodd, R.D. Mayfield and R.A. Harris. "Application of DNA microarrays to study human alcoholism." *Journal of Biomedical Science,* vol. 8, no. 1, pp. 28-36, 2001.

[76] Christine Debouck and Peter N. Goodfellow. "DNA microarrays in drug discovery and development." *Nature Genetics*, vol. 21, pp. 48 – 50, 1999.

[77] R. Ulrich and S.H. Friend. "Toxicogenomics and drug discovery: will new technologies help us produce better drugs?" *Nature Reviews Drug Discovery*, vol. 1, pp. 84-88, 2002.

[78] "Roche Diagnositics Product Sheet Amplichip." Internet: http://www.roche-diagnostics.com/products_services/amplichip_cyp450.html, 2005.

[79] Kumaravel Somasundaram, Sathish Kumar Mungamuri and Narendra Wajapeyee. "DNA microarray Technology and its applications in cancer biology." *Applied Genomics and proteomics*, vol. 1, no. 4, pp. 209-218, 2002.

[80] F. Bertucci, R. Houlgatte, C. Nguyen, A. Benziane, V. Nasser, S. Granjeaud, B. Tagett, B. Loriod, A. Giaconia, J. Jacquemier, P. Viens and D. Birnbaum. "Molecular typing of breast cancer:

transcriptomics and DNA microarrays." *Bull Cancer*, vol. 88, no. 3, pp. 277-286, 2001.

[81] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman. "Missing value estimation methods for DNA microarrays." *Bioinformatics*, vol. 17, no. 6, pp. 520-525, 2001.

[82] S. Mook, L.J. Van't Veer, E.J. Rutgers, M.J. Piccart-Gebhart and F. Cardoso. "Individualization of therapy using Mammaprint: from development to the MINDACT Trial." *Cancer Genomics Proteomics*, vol. 4, no. 3, pp. 147-155, 2007.

[83] C.I. Dumur, M. Lyons-Weiler, C. Sciulli, C.T. Garrett, I. Schrijver, T.K. Holley, J. Rodriquez-Paris, J.R. Pollack, J.L. Zehnder, M. Price, J.M. Hagenkord, C.T. Rigl, L.J. Buturovic, G.G. Anderson and F.A. Monzon. "Interlaboratory performance of a microarray-based gene expression test to determine tissue of origin in poorly differentiated and undifferentiated cancers." *The Journal of Molecular Diagnostics*, vol. 10, no. 1, pp. 67-77, 2008.

[84] H.J. Lawrence, S. Truong, N. Patten, A. Nakao and L. Wu. "Detection of p53 mutations in cancer by the amplichip p53 test, a microarraybased resequencing assay in *Proc. The third international workshop on mutant p53*, 2007.

[85] J. Devore and R. Peck. "Statistics: The Exploration and Analysis of Data." 3rd edition, Duxbury press, Pacific Glove, 1997.

[86] J.G. Thomas, J.M. Olson, S.J. Tapscott and L.P. Zhao. "An efficient and robust statistical modeling approach to discover differentially

expressed genes using genomic expression profiles." *Genome Research,* vol. 11, no. 7, pp. 1227-1236, 2001.

[87] H. Liu, R. Setiono. "Chi2: Feature selection and discretization of numeric attributes," in *Proc. IEEE 7th International Conference on Tools with Artificial Intelligence*, 1995, pp.338-391.

[88] W. Pan, J. Lin and C.T. Le. "A mixture model approach to detecting differentially expressed genes with microarray data." *Functional & Integrative Genomics,* vol. 3, no. 3, pp. 117-124, 2003.

[89] Jason Weston, André Elisseeff, Bernhard Schölkopf and Pack Kaelbling. "Use of the zero-norm with linear models and kernel methods." *Journal of Machine Learning Research,* vol. 3, pp. 1439-1461, 2003.

[90] H. Frohlich, O. Chapelle and B. Schölkopf. "Feature Selection for Support Vector Machines using Genetic Algorithms," in *Proc. 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp.142-148.

[91] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. "Gene Selection for Cancer Classification using Support Vector Machines." *Machine Learning*, vol. 46, no. 1-3, pp. 389-422, 2002.

[92] J. Yang and V. Honavar. "Feature Subset Selection using a Genetic Algorithm." *IEEE Intelligent Systems*, vol. 13, no. 2, pp. 44-49, 1998.

[93] Matthias E. Futschik and Nikola K. Kasabov. "Fuzzy Clustering of Gene Expression Data," in *Proc. IEEE International Conference on Fuzzy Systems*, 2002.

[94] T. Kohonen. "The Self-Organizing Map," in *Proc. the IEEE*, 1990, pp. 1464-1480.

[95] J.B. MacQueen. "Some Methods for classification and Analysis of Multivariate Observations," in *Proc. of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press*, 1967, pp. 281-297.

[96] J.A. Hartigan and M.A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[97] J.C. Bezdek, R. Ehrlich, and W. Full. "FCM: The fuzzy c-means clustering algorithm." *Computers and Geosciences*, vol. 10, pp. 191-203, 1984.

[98] Douglas T. Ross, Uwe Scherf, Michael B. Eisen, Charles M. Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S. Jeffrey, Matt Van de Rijn, Mark Waltham, Alexander Pergamenschikov, Jeffrey C.F. Lee, Deval Lashkari, Dari Shalon, Timothy G. Myers, John N. Weinstein, David Botstein and Patrick O. Brown. "Systematic variation in gene expression patterns in human cancer cell lines." *Nature Genetics*, vol. 24, pp. 227-235, 2000.

[99] A.K. Virmani, J.A. Tsou, K.D. Siegmund, L.Y. Shen, T.I. Long, P.W. Laird, A.F. Gazdar and I.A. Laird-Offringa. "Hierarchical clustering of lung cancer cell lines using DNA methylation markers"

Cancer Epidemiol Biomarkers Prevention, vol. 11, no. 3, pp. 291-297, 2002.

[100] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church. "Systematic determination of genetic network architecture." *Nature genetics*, vol. 22, no. 3, pp. 281-285, 1999.

[101] Doulaye Dembele and Philippe Kastner, "Fuzzy C-means method for clustering microarray data." *Bioinformatics* vol. 19, no. 8, pp. 973-980, 2003.

[102] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J. Hudson Jr., M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein and P.O. Brown. "The transcriptional program in the response of human fibroblasts to serum." *Science*, vol. 283, no. 5398, pp. 83-87, 1999.

[103] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart and R. Davis. "A genome-wide transcriptional analysis of the mitotic cell cycle." *Molecular Cell,* vol. 2, pp. 65–73, 1998.

[104] "A Gene Expression Database for the Molecular Pharmacology of Cancer." Internet: http://discover.nci.nih.gov/nature2000/

[105] P.J. Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis" *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.

[106] J. Dunn. "Well separated clusters and optimal fuzzy partitions." *Journal of Cybernetics*, vol. 4, pp. 95-104, 1974.

[107] K.Y. Yeung, D.R. Haynor and W.L. Ruzzo. "Validating clustering for gene expression data." *Bioinformatics*, vol. 17, pp. 309-318, 2001.

[108] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts and V. Sondak. "Molecular classification of cutaneous malignant melanoma by gene expression profiling." *Nature,* vol. 406, no. 6795, pp. 536-540, 2000.

[109] M.K. Kerr, M. Martin and J.A. Churchill. "Analysis of variance for gene expression microarray data." *Journal of Computational Biology*, vol. 7, no. 6, pp. 819-837, 2000.

[110] P. D'Haeseleer, X. Wen, S. Fuhrman and, R. Somogyi. "Linearmodeling of mRNA expression levels during CNS development and injury," in *Proc. Pacific Symposium on Biocomputing*, 1999, pp. 41–52.

[111] K. Murphy and S. Mian. "Modelling gene expression data using dynamic Bayesian networks." Technical Report, University of California, Berkeley, 1999.

[112] A. Hartemink, D. Gifford, T. Jaakkola and R. Young. "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," in *Proc. Pacific Symposium on Biocomputing*, vol. 6, pp.422–433, 2001.

[113] D.C. Weaver, C.T. Workman and G.D. Stormo. "Modeling regulatory networks with weight matrices," in *Proc. Pacific Symposium on Biocomputing,* 1999, pp. 112–123.

[114] J. Vohradsky. "Neural model of the genetic network." *The Journal of Biological Chemistry,* vol. 276, no. 39, pp. 36168–36173, 2001.

[115] T. Chen, H.L. He and G.M. Church. "Modeling gene expression with differential equations," in *Proc. Pacific Symposium on Biocomputing*, 1999, pp. 29-40.

[116] T. Mestl, E. Plahte and S.W. Omholt. "A mathematical framework for describing and analyzing gene regulatory networks." *Journal of Theoretical Biology*, vol. 176, pp. 291-300, 1995.

[117] H.H. McAdams and A. Arkin. "It's a noisy business! Genetic regulation at the nano molarscale." *Trends in Genetics*, vol. 15, pp. 65-69, 1999.

[118] S. A. Kauffman. "Metabolic stability and epigenesist in randomly constructed genetic nets." *Journal of Theoretical Biology*, vol. 22, pp. 437-467, 1969.

[119] L. Glass and S. A. Kauffman. "The logical analysis of continuous non-linear biochemical control networks." *Journal of Theoretical Biology*, vol. 39, pp. 103-129, 1973.

[120] S. A. Kauffman. "The Origins of Order: Self-organization and Selection in Evolution," *Oxford University Press, New York*, 1993.

[121] S. Huang. "Gene expression profiling, genetic networks and cellular states: An integrating concept for tumorigenesis and drug

discovery." *Journal of Molecular Medicine,* vol. 77, pp. 469-480, 1999.

[122] Z. Szallasi and S. Liang. "Modeling the Normal and Neoplastic Cell Cycle with Realistic Boolean Genetic Networks: Their Application for Understanding Carcinogenesis and Assessing Therapeutic Strategies," in *Proc. Pacific Symposium on Biocomputing*, 1998, pp. 66–76.

[123] R. Pal, A. Datta, A.J. Fornace, M.L. Bittner and E.R. Dougherty. "Boolean relationships among genes responsive to ionizing radiation in the NCI 60 ACDS." *Bioinformatics*, vol. 21, no. 8, pp. 1542–1549, 2005.

[124] I. Shmulevich, E.R. Dougherty, S. Kim and W. Zhang. "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks." *Bioinformatics, vol. 18, no. 2, pp.* 261–274, 2002.

[125] D. Heckerman, A. Mamdani and M.P. Wellman. "Real-World Applications of Bayesian Networks." *Communications of the ACM*, vol. 38, no. 3, pp. 24-68, 1995.

[126] S.Y. Kim, S. Imoto and S. Miyano. "Inferring gene networks from time series microarray data using dynamic bayesian networks." *Briefings in Bioinformatics*, vol. 4, no. 3, pp. 228-235, 2003.

[127] M. Zou and S.D. Conzen. "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data." *Bioinformatics,* vol. 21, no. 1, pp. 71–79, 2005.

[128] A.V. Werhli and D. Husmeier. “Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge.” *Stat Appl GenetMol Biol*, vol. 6, no. 15, 2007.

[129] C.S. Kim. “Bayesian Orthogonal Least Squares (BOLS) algorithm for reverse engineering of gene regulatory networks, BMC Bioinformatics 8 (2007) 251.

[130] S. Kimura, K. Ide, A. Kashihara, M. Kano, M. Hatakeyama, R. Masui, N. Nakagawa, S. Yokoyama, S. Kuramitsu and A. Konagaya. “Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm.” *Bioinformatics*, vol.21, pp. 1154-1163, 2005.

[131] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi and M. Tomita. “Dynamic modeling of genetic networks using genetic algorithm and S-system.” *Bioinformatics*, vol. 19, pp. 643-650, 2003.

[132] Eugenio Cinquemani, Andreas Milias-Argeitis, Sean Summers and John Lygeros. “Stochastic dynamics of genetic networks: modelling and parameter identification.” *Bioinformatics*, vol. 24, no. 23, pp. 2748-2754 , 2008.

[133] M.A. Savageau. “Rules for the evolution of gene circuitry,” in *Proc. Pacific Symposium on Biocomputing*, no.3, 1998, pp. 54-65.

[134] D. Di Bernardo, T.S. Gardner and J.J. Collins, “Robust identification of large genetic networks,” in *Proc. Pacific Symposium on Biocomputing*, no.9, 2004, pp. 486-497.

[135] E.O. Voit. *Computational Analysis of Biochemical Systems A Practical Guide for Biochemists and Molecular Biologists*. Cambridge, UK, Cambridge University Press, 2000.

[136] P. D'haeseleer, S. Liang and R. Somogyi. "Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering." *Bioinformatics*, vol.16, no.8, pp. 707-726, 2000.

[137] G. Von Dassow, E. Meir, E.M. Munro and G.M. Odell. "The Segment Polarity Network is a Robust Developmental Module." *Nature,* vol. 406, pp.188–192, 2000.

[138] Guy Karlebach and Ron Shamir. "Modelling and analysis of gene regulatory networks" *Nature Reviews Molecular Cell Biology*, vol.9, pp. 770-780, 2008.

[139] Rui Xu. "Inference of genetic regulatory networks with recurrent neural network models." in *Proc. 26th Annual International Conference of the IEEE*, vol. 2, pp. 2905-2908, 2004.

[140] R. Xu, G.K. Venayagamoorthy and D.C. Wunsch. "Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization. " *Neural Networks*, vol. 20, pp. 917-927, 2007.

[141] Monica Francesca Blasia, Ida Casorellia, Alfredo Colosimob, Francesco Simone Blasic, Margherita Bignamia and Alessandro Giuliania. "A recursive network approach can identify constitutive regulatory circuits in gene expression data." *Physica A*: *Statistical Mechanics and its Applications*, vol. 348, pp. 349-370, 2005.

[142] W.P. Lee and K.C. Yang. "A clustering-based approach for inferring recurrent neural networks as gene regulatory networks." *Neurocomputing*, vol. 71, pp. 600-610, 2008.

[143] P.J. Werbos. "Backpropagation through time: what it does and how to do it," in *Proc. IEEE*, vol. 78, no. 10, pp.1550-1560, 1990.

[144] A. Bezerianos and I.A. Maraziotis. "Computational models reconstruct gene regulatory networks." *Molecular Biosystems*, vol.4 pp. 993-1000, 2008.

[145] E. Keedwell and A. Narayanan. "Discovering gene networks with a neural genetic hybrid." *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 2, no.3, pp. 231-242, 2005.

[146] C.F. Juang and C.T. Lin. "A recurrent self-organizing neural fuzzy inference network." *IEEE Transaction on Neural Networks*, vol.10, pp. 828-845 ,1999.

[147] I.A. Maraziotis, A. Dragomir and A. Bezerianos. "Gene networks reconstruction and time series prediction from microarray data using recurrent neural fuzzy networks." *IET Systems Biology*, vol. 1, pp. 41-50, 2007.

[148] P.C.H. Ma and K.C.C. Chan. "Inferring gene regulatory networks from expression data by discovering fuzzy dependency relationships." *IEEE Transactions on fuzzy Systems*, vol. 16, pp. 455-465, 2008.

[149] Bahrad A. Sokhansanj, J. Patrick Fitch, Judy N. Quong and Andrew A. Quong. "Linear fuzzy gene network models obtained from

microarray data by exhaustive search." *BMC Bioinformatics*, vol. 5, 2004.

[150] Frank De Smet, Janick Mathys, Kathleen Marchal, Gert Thijs, Bart De Moor and Yves Moreau. "Adaptive quality-based clustering of gene expression profiles." *Bioinformatics*, vol. 18, no.6, pp. 735–748, 2002.

[151] Ramesh Ram, Madhu Chetty and Trevor I. Dix. "Fuzzy Model for gene regulatory network." in *Proc. IEEE congress on evolutionary computation*, 2006, pp.1450-1455.

[152] M.E. Mamakou, G. Ch.Sirakoulis, I. Andreadis and I. Karafyllidis. "Adaptive Reverse Engineering of Gene Regulatory Networks Using Genetic Algorithms," in *Proc. EUROCON*, 2005.

[153] Rudolf Kruse. "Fuzzy Neural Network" *Scholarpedia*, vol.3, no.11, 2008.

[154] L.A. Zadeh. "Fuzzy sets." *Information control*, vol. 8, pp. 38-353, 1965.

[155] Juang C.F., A tsk-type recurrent fuzzy network for dynamic systems processing by neural network and genetic algorithms, IEEE Transactions on Fuzzy Systems 10(2) (2002) 155-170.

[156] J. Chia-Feng and L. Chin-Teng. "A recurrent self-organizing neural fuzzy inference network." *IEEE Transactions on Neural Networks*, vol. 10, no. 4, pp. 828-845, 1991.

[157] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker. " Cytoscape: a

software environment for integrated models of biomolecular interaction networks." *Genome Research*, vol. 13, no. 11, pp. 2498-2504, 2003.

[158] H. Jeong, S.P. Mason, A. Barabasi and Z.N. Oltvai. "Lethality and centrality in protein networks" *Nature,* vol. 6833, no.411, pp. 41-42, 2001.

[159] S. Wachi, K. Yoneda and R. Wu. "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues." *Bioinformatics,* vol. 21, no.23, pp. 4205-4208, 2005.

[160] S. Maere, K. Heymans and M. Kuiper. "Bingo, a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." *Bioinformatics*, vol.21, no.16, pp. 3448-3449, 2005.

[161] Shahab Uddin, Maqbool Ahmed, Azhar Hussain, Jehad Abubaker, Nasser Al-Sanea, Alaa Abdul Jabbar, Luai H. Ashari, Samar Alhomoud, Fouad Al-Dayel, Zeenath Jehan, Prashant Bavi, Abdul K. Siraj, and Khawla S. Al-Kuraya. "Genome-wide expression analysis of Middle Eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy." *The American Journal of Pathology,* vol. 178, no.2, pp. 537-547, 2011.

[162] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein and Bruce Futcher. "Comprehensive identification of cell cycle-regulated genes of the yeast sacharomyces cerevisiae by

microarray hybridization." *Molecular Biology of the Cell* vol. 12, no.9, pp. 3273-3297, 1998.

[163] "Kegg: Kyoto encyclopedia of genes and genomics." Internet : http://www.genome.jp/kegg/.

[164] B. Zhang, X. Pan, G.P. Cobb and T.A. Anderson. "MicroRNAs as oncogenes and tumor suppressors." *Developmental Biology*, vol. 302, pp. 1–12, 2007.

[165] P.M. Voorhoeve. "MicroRNAs: Oncogenes, tumor suppressors or master regulators of cancer heterogeneity? " Biochimica et Biophysica Acta, vol. 1805 pp. 72–86, 2010.

[166] H. Tazawa, N. Tsuchiya, M. Izumiya, and H. Nakagama. "Tumor-suppressive miR-34a induces senescence-like growth arrest through modulation of the E2F pathway in human colon cancer cells," in *Proc. The National Academy of Sciences U S A*, 2007, pp. 15472-15477.

[167] C.S. Chang, O. Elemento and S. Tavazoie. "Revealing post transcriptional regulatory elements through network-level conservation." *PLoS Computational Biology ,* vol.1, no.7, 2005.

[168] I. Ivanovska, A.S. Ball, R.L. Diaz, J.F. Magnus, M. Kibukawa, J.M. Schelter, S.V. Kobayashi, L. Lim, J. Burchard, A.L. Jackson, P.S. Linsley and M.A. Cleary. "MicroRNAs in the miR-106b family regulate p21/CDKN1A 1 and promote cell cycle progression." *Molecular and Cellular Biology* ,vol.28, no.7, pp. 2167-74 2008.

[169] N. Rosenfeld, R. Aharonov, E. Meiri, S. Rosenwald, Y. Spector, M. Zepeniuk, H. Benjamin, N. Shabes, S. Tabak, A. Levy, D. Lebanony, Y. Goren, E. Silberschein, N. Targan, A. Ben-Ari, S. Gilad, N. Sion-Vardy, A. Tobar, M. Feinmesser, O. Kharenko, O. Nativ, D. Nass, M. Perelman, A. Yosepovich, B. Shalmon, S. Polak-Charcon, E. Fridman, A. Avniel, I. Bentwich, Z. Bentwich, D. Cohen, A. Chajut and I. Barshack. "MicroRNAs accurately identify cancer tissue origin." *Nature Biotechnology* vol.26 no.4, pp. 462-469, 2008.

[170] A.M. Cheng, M.W. Byrom, J. Shelton and L.P. Ford. "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis." *Nucleic Acids Research*, vol.33, pp. 1290-1297, 2005.

[171] J.C. Huang, T. Babak, T.W. Corson, G. Chua, S. Khan, B.L. Gallie, T.R. Hughes, B.J. Blencowe, B.J. Frey and Q.D. Morris. "Using expression profiling data to identify human microRNA targets." *Nature methods*, vol.4, pp. 1045-1049, 2007.

[172] X. Li, R. Gill, N.G. Cooper, J.K. Yoo and S. Datta. "Modeling microRNA-mRNA Interactions Using PLS Regression in Human Colon Cancer." *BMC Medical Genomics*, vol. 4, no.44, 2011.

[173] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[174] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H.R. Horvitz and R.R. Golub. "MicroRNA expression profiles classify human cancers."*Nature,* vol. 435, pp. 834-838, 2005.

[175] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P. Mesirov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S. Lander and Todd R. Golub. "Multiclass cancer diagnosis using tumor gene expression signatures," in *Proc. The National Academy of Sciences of the United States of America*, 2001, pp. 15149-15154.

[176] Doron Betel, Manda Wilson, Aaron Gabow, Debora S. Marks and Chris Sander. "The microRNA.org resource: targets and expression." *Nucleic Acids Research* vol. 36, no.1, pp. D149-D153, 2008.

[177] B. John, A.J. Enright, A. Aravin, T. Tuschl, C. Sander and D.S. Marks. "Human MicroRNA targets." *PLoS Biology,* vol.2, 2004.

[178] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang and Yunlong Liu. "miR2Disease: a manually curated database for microRNA deregulation in human disease." *Nucleic Acids Research,* vol.37, no. 1, pp. D98-D104,2009.

[179] A. Lagana, S. Forte, A. Giudice, M.R. Arena, P.L Puglisi, R. Giugno, A. Pulvirenti, D. Shasha and A. Ferro. "miRo: a miRNA knowledge base." *Database: the journal of biological databases and curation*, vol. 6, 2009.

List of Publications

International Journal Papers

1. Vineetha S, C. ChandrashekaraBhatand Sumam Mary Idicula. "Gene Regulatory Network from Microarray data using Fuzzy Logic Approach."*International Journal of Recent Trends in Engineering & Technology,* vol. 4, no.1, pp.54-57,2010.
2. Vineetha S, C. ChandrashekaraBhatand Sumam Mary Idicula."Modelling Gene Regulatory Network from Microarray Data Using Modified Genetic Algorithm."*Journal Of Computational Intelligence in Bioinformatics,*vol. 4,no.2-3,pp. 221-231,2011.
3. Vineetha S, C. ChandrashekaraBhatand Sumam Mary Idicula, "Reverse Engineering of colon cancer specific Gene Regulatory Network using TSK-type Recurrent Neural Fuzzy Network" , *GENE, Elsevier,*vol. 506, no.2, pp. 408-416, 2012.
4. Vineetha S, C. ChandrashekaraBhatand Sumam Mary Idicula."MicroRNA-mRNA Interaction Network Using TSK-type Recurrent neural Fuzzy Network.",communicated to*Computational Biology and Chemistry*, Elsevier, 2012.

International Conference Papers

1. Vineetha S, C ChandrashekaraBhatand Sumam Mary Idicula."Analysis of plasma RNA data from Colon Cancer patients Using Hybrid algorithm."in *proc. International Conference on Advances in Optoelectronics, Information and Communication Technologies (ICOICT)*, Thiruvananthapuram, 2009, pp.26-27.
2. Vineetha S, C ChandrashekaraBhatand Sumam Mary Idicula,"Gene Regulatory Network For Circulating Plasma RNA Data From Colon Cancer Patients Using Dynamic Neural Fuzzy Approach." in *proc.International Symposium on Computational Biology Structural Bioinformatics and Systems Biology (BioinformaticaIndica), University of Kerala, Kariavattom, Thiruvananthapuram,*2010, pp. 11-13.
3. Vineetha S, C ChandrashekaraBhatand Sumam Mary Idicula,"Gene Regulatory Network For Circulating Plasma RNA Data From Colon Cancer Patients Using Fuzzy Logic Approach."in *proc.Eighth Asia Pacific Bioinformatics Conference (APBC), Indian Institute of Science, Bangalore*, 2010, pp. 18-21.
4. Vineetha S, CChandrashekaraBhatand Sumam Mary Idicula,"Gene Regulatory Network from Microarray Data using Dynamic Neural Fuzzy Approach."in *proc.ACM Sponsored International Symposium on Biocomputing (ISB), National Institute of Technology, Calicut,*2010, pp.15-27.

…….ଓଡ଼……..

Appendix

Appendix 1: Gene symbol and description of 100 differentially expressed genes identified using parametric t-test from plasma RNA dataset

	Gene Symbol	Description
1	ACP1	Acid phosphatase 1, soluble
2	ACP2	Acid phosphatase 2, lysosomal
3	AEBP1	AE-binding protein 1
4	ALS2CR3	Amyotrophic lateral sclerosis 2 (juvenile) chromosome region, Candidate 3
5	ANXA1	Annexin A1
6	ATP5G3	ATP synthase, H+ transporting, mitochondrial FO complex, subunit c (subunit 9) isoform 3
7	BAD	BCL2-antagonist of cell death
8	BDKRB1	Bradykinin receptor B1
9	BRP44	Brain Protein 44
10	BTF3	Basic transcription factor 3
11	C18B11	C18B11 homolog (44.9kD)
12	C1ORF13	Hypothetical protein similar to swine acylneuraminatelyase
13	C20orf194	Chromosome 20 open reading frame 194
14	CAPZA1	Capping protein (actin filament) muscle Z-line, alpha 1
15	CAV2	Caveolin 2
16	CBX1	Chromoboxhomolog 1 (Drosophila HP1 beta)
17	CCNT1	Cyclin T1
18	CD83	CD83 antigen (activated B lymphocytes, immunoglobulin superfamily)
19	CEP3	Cdc42 effector protein 3
20	CHK	Choline kinase
21	COX11	COX11 (yeast) homolog, cytochrome c oxidase assembly protein
22	COX15	COX15 (yeast) homolog, cytochrome c oxidase assembly protein

	Gene Symbol	Description
23	CREM	CAMP responsive element modulator
24	CYP8B1	Cytochrome P450, subfamily VIIIB (sterol 12-alpha-hydroxylase), polypeptide 1
25	DECR1	2,4-dienoyl CoA reductase 1, mitochondrial
26	DEFCAP	Death effector filament-forming Ced-4-like apoptosis protein
27	DGKZ	Diacylglycerol kinase, zeta (104kD)
28	DNCLI2	Dynein, cytoplasmic, light intermediate polypeptide 2
29	DNM2	Dynamin 2
30	DOC1	Downregulated in ovarian cancer 1
31	DSCR5	Down syndrome critical region gene 5
32	E46L	Like mouse brain protein E46
33	EIF3S6	Eukaryotic translation initiation factor 3, subunit 6 (48kD)
34	EPAS1	Endothelial PAS domain protein 1
35	ERBB2	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 (neuro/glioblastoma derived oncogene homolog)
36	ERP28	Endoplasmic reticulum lumenal protein
37	FABP5	Fatty acid binding protein 5 (psoriasis-associated)
38	FLJ10287	Hypothetical protein
39	FLJ10849	Hypothetical protein FLJ10849
40	FLJ12910	Hypothetical protein FLJ12910
41	FLJ20037	Hypothetical protein FLJ20037
42	FLJ20701	Hypothetical protein FLJ20701
43	GCN5L1	GCN5 (general control of amino-acid synthesis, yeast, homolog)-like 1
44	GLRX	Glutaredoxin (thioltransferase)
45	HBE1	Hemoglobin, epsilon 1
46	HGS	Hepatocyte growth factor-regulated tyrosine kinase substrate
47	HLF	Hepatic leukemia factor
48	HNRPH3	Heterogeneous nuclear ribonucleoprotein H3 (2H9)
49	HRG	Histidine-rich glycoprotein
50	IGLL1	Immunoglobulin lambda-like polypeptide 1

	Gene Symbol	Description
51	IL2RG	Interleukin 2 receptor, gamma (severe combined immunodeficiency)
52	IL7R	Interleukin 7 receptor
53	KIAA0041	Centaurin beta2
54	KIAA0417	KIAA0417 gene product
55	KIAA0524	KIAA0524 protein
56	KIAA0632	KIAA0632 protein
57	KIAA0758	KIAA0758 protein
58	KIAA0801	KIAA0801 gene product
59	KIAA1288	KIAA1288 protein
60	LDHA	Lactate dehydrogenase A
61	LOC50999	CGI-100 protein
62	LOC51002	CGI-121 protein
63	LOC51023	CGI-134 protein
64	LOC51194	Ran binding protein 11
65	LOC51242	Hypothetical protein
66	LOC51632	CGI-76 protein
67	LPL	Lipoprotein lipase
68	MAN2A1	Mannosidase, alpha, class 2A, member 1
69	MLLT4	Myeloid/lymphoid or mixed-lineage leukemia (trithorax (Drosophila) homolog); translocated to, 4
70	MVP	Major vault protein
71	NMA	Putative transmembrane protein
72	NOLA2	Nucleolar protein family A, member 2 (H/ACA small nucleolar RNPs)
73	NR3C1	Nuclear receptor subfamily 3, group C, member 1
74	OXA1L	Oxidase (cytochrome c) assembly 1-like
75	PCBP2	Poly(rC)-binding protein 2
76	PCP4	Purkinje cell protein 4
77	PIK3C2G	Phosphoinositide-3-kinase, class 2, gamma polypeptide
78	POMT1	Protein-O-mannosyltransferase 1

	Gene Symbol	Description
79	PPP1CC	Protein phosphatase 1, catalytic subunit, gamma isoform
80	PSMA3	Proteasome (prosome, macropain) subunit, alpha type, 3
81	RELN	Reelin
82	RPC	RNA 3'-terminal phosphate cyclase
83	RPL10A	Ribosomal protein L10a
84	RPS6KA1	Ribosomal protein S6 kinase, 90kD, polypeptide 1
85	SFTPA2	Surfactant, pulmonary-associated protein A2
86	SLBP	Stem-loop (histone) binding protein
87	SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5
88	SMN2	Survival of motor neuron 2, centromeric
89	SP38	Zonapellucida binding protein
90	SPIB	Spi-B transcription factor (Spi-1/PU.1 related)
91	ST16	Suppression of tumorigenicity 16 (melanoma differentiation)
92	STC1	Stanniocalcin 1
93	TIAF1	TGFB1-induced anti-apoptotic factor 1
94	TIM17	Translocase of inner mitochondrial membrane 17 (yeast) homolog A
95	TNKS	Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase
96	TNS	Tensin
97	TRIP10	Thyroid hormone receptor interactor 10
98	UBE2D3	Ubiquitin-conjugating enzyme E2D 3 (homologous to yeast UBC4/5)
99	USP9X	Ubiquitin specific protease 9, X chromosome (Drosophila fat facets related)
100	XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining; Ku autoantigen, 80kD)

Appendix 2: Expression ratio of 27 differentially expressed genes from plasma RNA dataset(D – denotes diseased sample and H – denotes healthy samples)

	Gene Symbol	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	H1	H2	H3	H4	H5	H6	H7	H8
1	ACP1	2.46	0.69	-0.73	-2.44	-0.62	-1.1	0.03	1.54	2.72	1.64	1.97	2.72	-2.75	-0.92	-1.47	-0.94	-0.82	-0.9	-0.49	-0.98
2	ANXA1	3.12	-0.96	0.6	-2.74	0.24	3.15	3.67	4.08	4.37	3.09	3.25	2.79	0.25	-1.23	-0.59	-1.56	-0.96	-0.2	-0.49	0.39
3	BRP44	1.98	0.57	-0.27	-1.23	-0.78	-0.18	0.36	1.5	3.08	2.02	2.46	2.37	-2.33	-0.78	-1.49	-4.02	-0.96	-0.93	0.35	-0.7
4	C18B11	-0.89	-1.37	-1.25	-0.23	-0.7	-3.44	-2.4	-1.98	-1.45	0.24	-0.16	0.02	-0.44	-0.16	-0.73	0.84	0.25	0.58	1.03	0.31
5	C20orf194	0.57	1.62	0.42	0.51	-0.64	3.47	0.03	1.28	2.46	2.32	1.59	1.78	0.34	-0.18	-0.31	-1.61	0.13	0.14	-0.11	-0.08
6	DGKZ	1.09	0.31	-0.85	-0.03	0.55	0.87	0.37	2.32	0.97	0.92	1.42	1.49	-0.75	-0.66	-0.44	-3.14	-0.47	-0.08	-0.3	-1.85
7	DOC1	-1.66	-2.86	-0.68	0.01	-1.29	1.3	3	0.6	-0.02	1.35	1.69	0.36	2.07	0.66	3	4.15	1.61	2.12	1.17	2.16
8	EPAS1	3.1	-1.85	0.56	-1.63	2.81	3.15	2.17	3.62	3.64	2.56	2.85	2.64	-2.16	0.47	-3.08	1.71	-1.09	-0.66	-1.34	-1.86
9	FLJ10849	-0.55	-0.85	-0.17	2.47	0.64	-0.46	0.77	0.65	0.16	-0.23	-1.11	-0.39	2.39	2.13	0.92	1.53	1.03	0.93	1.52	1.58
10	FLJ20701	1.06	2.04	0.71	0.2	-1.13	0.46	-0.96	0.85	2.27	1.96	1.08	1.8	-0.94	-0.92	-0.5	-0.77	0.08	-0.19	-0.78	-1.28
11	HBE1	1.87	-1.09	0.87	1.4	2.49	0.63	1.69	1.73	4.92	0.87	1.14	1.38	-2.97	-0.09	-4.28	-2.99	-3.63	-1.23	-0.3	-2.58
12	IL7R	1	0.33	-0.92	-0.82	0.14	1.7	0.67	2.47	2.62	1	1.24	-0.39	-2.52	-0.67	-0.81	-0.94	-0.8	-0.44	-0.27	0.57
13	KIAA0632	1.48	-1.68	-0.98	0.19	2.54	2.11	1.71	2.56	3.05	2.61	2.13	2.05	0.41	-0.87	-1.52	-3.18	-0.49	0.82	-0.69	-1.28
14	KIAA0758	1.52	-2.16	1.24	-1.47	-0.56	0.91	-0.6	1.84	2.3	1.84	1.96	1.96	-2.58	0.43	-0.96	-1.46	-0.8	-0.22	-0.54	-2.08
15	LDHA	1.18	-1.81	-0.45	0.22	-0.21	1.82	1.18	2.66	3.13	1.68	1.94	1.5	-2	-0.06	-2.35	0.11	-1.18	-0.4	-1.06	-1.31
16	MVP	1.31	0.55	-0.66	-1.95	1.4	1.21	1.58	1.91	3.89	2.21	2.19	1.75	-3.21	0.3	-1.71	-0.97	-1.11	-0.52	0.96	-0.01
17	OXA1L	0.18	-0.27	1.26	-0.8	0.65	1.04	2.2	2.16	2.43	1.97	0.98	1.09	-1.49	-0.44	-0.11	-0.99	0.3	-0.42	0.07	-0.07
18	PCBP2	0.13	0.55	-0.92	-1.65	1.45	1.38	1.96	2.02	3.85	2.77	2.05	1.78	-3.62	-0.11	-0.41	-0.11	-1.09	-0.74	0.33	-0.8
19	PPP1CC	0.52	0.7	-0.43	-0.18	-2.65	2.32	0.13	2.69	2.95	2.22	2.37	2.22	-0.61	-0.45	-1.06	-3.61	-0.78	-0.33	0.17	-3.93
20	PSMA3	0.95	-0.4	-0.09	-3.09	-0.12	1.48	1.33	1.97	2.63	1.62	1.09	1.72	-2.37	-1.13	-2.69	-1.47	-1.47	-1.14	-1.12	-0.74
21	RELN	-1.71	-0.67	-0.2	-0.97	-2.11	-0.09	0.53	-0.08	-1.45	-0.3	0.28	-0.39	1.27	0.08	0.63	1.8	0.32	0.27	0.15	0.98
22	RPL10A	1.54	-1.42	-0.12	-0.05	1.66	1.69	2	2.14	3.28	2.22	2.18	1.45	0.31	-0.7	-1.53	-0.29	-0.24	-0.32	0.32	0.14
23	SP38	3.08	-2.13	1.39	-4.35	-1.41	-2.98	-0.99	-1.15	-1.28	-2.49	0.42	-1.77	5.25	2.37	0.53	2	0.49	4.16	1.87	1.81
24	TIAF1	-1.19	-1.28	-0.13	-2.51	-1.57	-0.69	-1.91	-1.75	-1.3	-1.53	-1.71	-2.39	-1.02	-0.73	0.45	0.82	-0.97	-0.96	-0.65	-0.1
25	TNKS	-1.64	-1.63	0.31	0.11	-2.27	-0.42	0.07	-0.56	0.37	-0.9	0.59	-0.21	2.78	1.37	2.15	3.66	0.4	0.21	0.25	0.92
26	TRIP10	-1.02	-2.79	1	0.97	-1.79	1.05	-1.51	-0.17	-1.55	0.62	0.15	0.59	2.53	1.96	1.47	2.23	0.68	0.34	0.96	2.43
27	UBE2D3	-1	1.33	0.78	-0.01	-1.83	0.51	1.23	2.21	2.58	1.1	1.16	1.4	0.78	-0.91	-3.15	-0.92	-2.83	-1.11	-0.88	-1.5

Appendix 3 : Examples of set of rules generated by DNFN for predicting the state of an output gene based on remaining 26 genes.

Rules Generated for Predicting ACP1											
		Rules									
INPUTS	TIAF1	2	2	2	2	1	1	1	3	3	
	PCBP2	1	3	2	2	2	3	3	2	2	
	IL7R	1	3	2	2	2	3	2	2	2	
	DOC1	2	2	2	1	3	2	2	3	3	
	MVP	1	3	2	2	2	2	3	1	2	
	PPP1CC	2	3	2	1	2	3	3	2	1	
	KIAA0758	1	3	2	2	2	3	3	1	1	**Activator**
	KIAA0632	2	3	2	3	3	3	3	2	1	
	HBE1	1	3	2	2	2	2	2	1	1	
	PSMA3	1	3	2	2	3	3	2	1	2	
	FLJ10849	3	2	2	2	2	2	1	2	2	
	SP38	3	2	2	2	2	2	2	2	2	
	FLJ20701	1	3	2	1	1	2	2	1	1	
	LDHA	1	3	2	2	2	3	3	1	2	**Activator**
	OXA1L	1	3	2	2	3	3	2	2	1	
	TRIP10	3	1	2	2	2	1	1	3	3	**Repressor**
	RELN	3	1	2	1	2	1	1	2	2	**Repressor**
	EPAS1	1	3	2	2	3	3	3	1	2	**Activator**
	RPL10A	2	3	2	2	2	3	3	1	2	
	DGKZ	2	3	2	2	2	3	3	2	1	
	C18B11	2	3	2	1	2	2	2	2	1	
	BRP44	1	3	2	2	2	3	3	2	1	
	UBE2D3	1	3	2	2	3	3	3	1	2	**Activator**
	ANXA1	2	3	2	2	3	3	3	2	1	
	C18B11	2	2	2	2	1	2	2	2	3	
	TNKS	3	2	2	1	2	2	2	2	3	
OUTPUT	ACP1	1	3	2	2	2	3	3	1	2	

Rules Generated for Predicting HBE1											
		Rules									
INPUTS	PSMA3	1	1	3	2	2	2	2	2	2	**Activator**
	FLJ10849	2	3	2	1	1	2	2	1	2	
	SP38	2	3	2	3	1	2	2	2	3	
	FLJ20701	1	1	3	2	3	2	1	2	2	
	LDHA	1	1	3	2	1	2	2	2	2	**Activator**
	OXA1L	2	1	3	2	2	2	1	2	2	
	TRIP10	3	3	1	2	2	2	1	2	2	**Repressor**
	RELN	3	3	1	1	2	2	2	2	2	**Repressor**
	EPAS1	1	1	3	3	1	2	3	3	2	
	RPL10A	1	2	3	2	1	2	2	3	2	
	DGKZ	2	2	3	3	2	2	2	3	2	
	C18B11	2	2	3	2	2	2	1	2	2	
	BRP44	2	1	3	3	2	2	2	3	2	
	UBE2D3	1	1	3	2	2	2	2	3	2	**Activator**
	ANXA1	2	2	3	3	2	2	2	3	2	
	C18B11	2	2	2	2	2	2	2	2	3	
	TNKS	2	3	2	1	1	2	1	2	2	
	ACP1	1	1	3	3	2	2	2	3	2	
	TIAF1	3	2	2	2	2	2	2	1	2	
	PCBP2	2	1	3	2	2	2	2	3	2	
	IL7R	2	1	3	2	2	2	2	2	2	
	DOC1	3	2	2	1	1	2	1	2	2	
	MVP	1	1	3	2	2	2	1	3	2	
	PPP1CC	2	2	3	2	2	2	1	3	2	
	KIAA0758	2	1	3	3	1	3	2	3	2	
	KIAA0632	2	2	3	2	1	2	3	3	2	
OUTPUT	HBE1	1	1	3	2	2	2	2	2	2	

Rules Generated for Predicting MVP											
		Rules									
	PPP1CC	2	2	2	3	2	2	3	1	2	
	KIAA0758	1	1	1	3	2	2	3	2	2	**Activator**
	KIAA0632	2	2	2	3	1	2	3	1	2	
	HBE1	1	1	1	3	2	2	2	1	2	**Activator**
	PSMA3	1	1	1	3	2	2	2	2	2	
	FLJ10849	3	3	2	2	1	2	1	2	2	
	SP38	3	1	2	2	1	2	2	2	2	
	FLJ20701	1	2	1	3	3	2	2	1	1	
	LDHA	1	1	1	3	1	2	3	2	2	**Activator**
	OXA1L	1	1	2	3	2	2	2	1	2	
	TRIP10	3	2	3	1	1	2	2	3	2	
	RELN	3	2	2	1	2	2	2	3	2	
	EPAS1	1	1	1	3	2	2	3	2	2	**Activator**
INPUTS	RPL10A	2	2	1	3	1	2	3	2	2	
	DGKZ	2	2	2	3	2	2	3	1	2	
	C18B11	2	2	2	3	2	2	2	1	2	
	BRP44	1	2	2	3	2	2	3	1	2	
	UBE2D3	2	2	1	3	3	2	3	2	2	
	ANXA1	2	1	2	3	2	2	3	1	2	
	C18B11	2	2	2	2	2	2	2	3	3	
	TNKS	3	2	2	2	1	2	2	3	2	
	ACP1	1	1	1	3	2	2	3	1	3	
	TIAF1	2	1	3	2	2	2	1	3	2	
	PCBP2	1	2	2	3	2	2	3	2	2	
	IL7R	1	2	2	3	2	2	2	2	2	
	DOC1	2	2	3	2	1	2	2	3	2	
OUTPUT	MVP	1	1	1	3	2	2	3	2	2	

Rules Generated for Predicting PSMA3												
		Rules										
	FLJ10849	3	3	1	2	2	2	1	1	2	2	**Repressor**
	SP38	1	2	2	1	2	2	2	1	2	2	
	FLJ20701	2	1	3	3	2	1	2	3	1	1	
	LDHA	1	1	3	1	2	2	3	2	1	1	**Activator**
	OXA1L	1	1	3	2	2	2	3	3	2	2	**Activator**
	TRIP10	3	3	1	2	2	1	1	1	2	2	**Repressor**
	RELN	2	2	1	2	2	1	2	2	2	3	
	EPAS1	1	1	3	2	2	3	3	3	2	2	**Activator**
	RPL10A	1	1	3	1	2	2	3	3	2	2	**Activator**
	DGKZ	2	2	3	2	2	2	3	2	2	2	
	C18B11	2	2	3	2	2	1	2	3	2	2	
	BRP44	2	2	3	2	2	2	3	3	2	2	
	UBE2D3	2	1	3	3	2	1	3	2	2	2	
INPUTS	ANXA1	1	1	3	2	2	2	3	3	2	2	**Activator**
	C18B11	2	2	2	2	2	2	2	3	3	3	
	TNKS	2	2	2	1	2	1	2	1	2	2	
	ACP1	1	1	3	2	2	2	3	3	2	2	**Activator**
	TIAF1	1	3	2	2	2	2	1	2	2	2	
	PCBP2	2	2	3	2	2	2	3	3	2	2	
	IL7R	2	2	3	2	2	2	3	2	2	2	
	DOC1	2	3	2	1	2	1	2	2	2	2	
	MVP	1	1	3	2	2	2	2	3	1	2	
	PPP1CC	2	2	3	2	2	1	3	3	2	1	
	KIAA0758	1	2	3	1	3	2	3	3	2	1	
	KIAA0632	2	2	3	1	2	3	3	3	2	2	
	HBE1	2	1	3	2	2	2	2	2	2	1	
OUTPUT	PSMA3	1	1	3	2	2	2	3	3	2	2	

Rules Generated for Predicting DGKZ											
		Rules									
INPUTS	C20orf194	1	2	2	2	1	3	2	2	2	
	BRP44	1	3	2	2	2	3	3	2	2	Activator
	UBE2D3	2	3	2	3	1	3	3	1	2	
	ANXA1	1	3	2	2	2	2	3	2	2	
	C18B11	3	2	2	2	2	2	3	2	3	
	TNKS	3	1	2	1	2	1	1	2	2	Repressor
	ACP1	2	3	2	2	2	3	3	1	2	
	TIAF1	3	1	2	2	2	1	1	3	2	Repressor
	PCBP2	2	3	2	2	2	3	2	2	2	
	IL7R	2	3	2	2	2	3	2	2	2	
	DOC1	3	2	2	1	1	2	2	3	2	
	MVP	1	3	2	2	2	3	3	1	2	Activator
	PPP1CC	1	3	2	3	1	3	3	2	2	
	KIAA0758	1	3	3	1	2	3	3	2	2	
	KIAA0632	1	3	2	1	2	3	3	2	2	Activator
	HBE1	1	2	2	2	2	3	2	1	2	
	PSMA3	2	3	2	2	2	3	3	1	2	
	FLJ10849	2	2	2	1	2	2	1	2	2	
	SP38	2	2	2	1	2	2	2	2	2	
	FLJ20701	1	2	2	3	1	3	3	1	1	
	LDHA	2	3	2	1	2	3	2	1	1	
	OXA1L	1	3	2	2	2	3	2	1	2	
	TRIP10	3	2	2	1	1	1	2	3	2	
	RELN	3	2	2	2	1	1	2	2	2	
	EPAS1	2	3	2	1	3	3	3	1	2	
	RPL10A	2	3	2	1	2	3	2	1	2	
OUTPUT	DGKZ	1	3	2	2	2	3	3	2	2	

Appendix 4 : Examples of set of rules generated by TRNFN for predicting the state of an output gene based on remaining 26 genes.

Rules Generated for Predicting ACP1								
Rules								
INPUTS	TIAF1	2	2	2	2	3	2	
	PCBP2	3	3	1	2	2	2	
	IL7R	3	2	1	2	2	2	
	DOC1	2	2	2	2	3	2	
	MVP	3	3	1	2	2	2	
	PPP1CC	3	3	2	2	1	1	**Activator**
	KIAA0758	3	3	1	2	1	1	**Activator**
	KIAA0632	3	3	2	2	1	2	
	HBE1	3	2	1	2	1	1	
	PSMA3	3	3	1	2	2	2	
	FLJ10849	1	1	3	3	2	3	**Repressor**
	SP38	2	1	3	2	2	2	
	FLJ20701	3	3	1	1	1	2	
	LDHA	3	3	1	2	1	1	**Activator**
	OXA1L	3	3	1	2	2	2	
	TRIP10	1	2	3	2	3	3	**Repressor**
	RELN	1	1	3	2	3	3	**Repressor**
	EPAS1	3	3	1	2	1	1	**Activator**
	RPL10A	3	3	2	1	1	2	
	DGKZ	3	3	2	2	1	2	
	C18B11	3	3	2	2	1	2	
	BRP44	3	3	1	2	1	2	
	UBE2D3	3	3	2	2	1	1	**Activator**
	ANXA1	3	3	2	1	1	2	
	C18B11	2	3	2	2	3	3	
	TNKS	2	1	3	2	3	2	
OUTPUT	ACP1	3	3	1	2	1	1	

Rules Generated for Predicting HBE1									
		Rules							
	PSMA3	2	3	3	1	1	2	2	
	FLJ10849	2	2	1	3	2	2	3	
	SP38	2	2	1	3	2	2	2	
	FLJ20701	2	3	3	1	1	2	1	
	LDHA	2	3	2	1	1	2	1	**Activator**
	OXA1L	2	3	2	1	2	1	2	
	TRIP10	2	1	2	3	3	3	3	**Repressor**
	RELN	2	1	2	3	2	3	3	**Repressor**
	EPAS1	2	3	3	1	1	2	1	
	RPL10A	2	3	2	2	1	1	2	
	DGKZ	2	3	3	2	2	1	2	
	C18B11	2	3	2	2	1	1	2	
	BRP44	2	3	3	1	2	1	2	
INPUTS	UBE2D3	2	3	2	2	1	1	1	**Activator**
	ANXA1	2	3	3	2	2	1	2	
	C18B11	2	2	3	2	2	3	3	
	TNKS	2	1	2	3	3	3	2	**Repressor**
	ACP1	2	3	3	1	1	2	2	
	TIAF1	2	2	1	2	3	3	2	
	PCBP2	2	3	3	1	2	2	2	
	IL7R	2	3	2	1	2	2	2	
	DOC1	2	2	2	2	3	3	2	
	MVP	2	3	2	1	1	2	2	
	PPP1CC	2	3	3	2	2	1	1	
	KIAA0758	3	3	3	1	2	1	1	
	KIAA0632	2	3	3	2	2	1	2	
OUTPUT	HBE1	2	3	2	1	1	1	1	

		Rules Generated for Predicting MVP							
		Rules							
	PPP1CC	2	3	3	3	2	2	2	
	KIAA0758	3	3	3	3	1	2	2	
	KIAA0632	2	3	3	3	2	2	2	
	HBE1	2	2	2	2	1	1	2	**Activator**
	PSMA3	2	3	3	3	1	1	2	**Activator**
	FLJ10849	2	1	1	2	3	2	2	**Repressor**
	SP38	2	1	1	1	3	3	2	**Repressor**
	FLJ20701	2	3	2	3	1	1	2	
	LDHA	2	3	3	2	1	1	1	**Activator**
	OXA1L	2	3	2	2	1	2	2	
	TRIP10	2	2	2	2	3	3	2	
	RELN	2	2	2	2	3	2	2	
	EPAS1	2	3	3	3	1	1	2	**Activator**
INPUTS	RPL10A	2	3	3	2	2	1	2	**Activator**
	DGKZ	2	3	3	3	2	2	2	
	C18B11	2	3	2	2	2	1	2	
	BRP44	2	3	3	3	1	2	2	
	UBE2D3	2	3	3	3	2	1	1	
	ANXA1	2	3	3	3	2	2	1	
	C18B11	2	3	3	3	2	2	3	
	TNKS	2	1	2	2	3	3	2	
	ACP1	2	3	3	3	1	1	3	
	TIAF1	2	2	1	1	2	3	2	
	PCBP2	2	3	3	3	1	2	2	
	IL7R	2	2	3	2	1	2	2	
	DOC1	2	2	2	2	2	3	2	
OUTPUT	MVP	2	3	3	2	1	1	2	

		Rules Generated for Predicting PSMA3							
		Rules							
	FLJ10849	2	2	1	1	3	3	2	**Repressor**
	SP38	2	2	2	1	3	2	2	
	FLJ20701	2	2	3	3	2	1	2	
	LDHA	2	3	3	3	1	1	1	**Activator**
	OXA1L	2	3	3	3	1	2	1	**Activator**
	TRIP10	2	1	1	2	3	3	2	**Repressor**
	RELN	2	2	1	2	3	2	2	
	EPAS1	2	3	3	3	1	1	2	**Activator**
	RPL10A	2	3	3	3	2	1	2	**Activator**
	DGKZ	2	3	3	3	2	2	2	
	C18B11	2	2	3	3	2	1	2	
	BRP44	2	3	2	3	1	2	2	
	UBE2D3	2	3	3	2	2	1	1	
INPUTS	ANXA1	2	3	3	3	1	2	2	**Activator**
	C18B11	2	2	2	3	2	2	3	
	TNKS	2	2	2	1	3	3	2	
	ACP1	2	3	3	3	1	1	2	**Activator**
	TIAF1	2	1	2	2	2	3	2	
	PCBP2	2	3	2	3	1	2	2	
	IL7R	2	3	3	2	1	2	2	
	DOC1	2	2	2	2	2	3	2	
	MVP	3	2	3	3	1	1	2	
	PPP1CC	2	3	3	3	1	2	2	**Activator**
	KIAA0758	3	3	3	3	1	2	2	
	KIAA0632	2	3	3	3	2	2	2	
	HBE1	2	2	3	2	1	1	1	
OUTPUT	PSMA3	2	3	3	3	1	1	2	

		Rules Generated for Predicting DGKZ						
		Rules						
	C20orf194	2	3	3	1	2	1	
	BRP44	3	3	3	1	2	2	**Activator**
	UBE2D3	3	3	3	2	1	2	
	ANXA1	3	2	3	1	2	1	
	C18B11	2	2	3	3	3	3	
	TNKS	1	2	1	3	2	2	**Repressor**
	ACP1	3	2	3	2	2	2	
	TIAF1	1	2	1	3	2	2	**Repressor**
	PCBP2	3	2	3	2	2	2	
	IL7R	3	3	2	2	2	2	
	DOC1	1	1	2	3	2	2	**Repressor**
	MVP	2	3	3	2	2	2	
INPUTS	PPP1CC	3	3	3	1	2	2	**Activator**
	KIAA0758	3	2	3	2	2	2	
	KIAA0632	3	3	3	1	2	2	**Activator**
	HBE1	2	3	2	1	2	2	
	PSMA3	3	2	3	2	2	2	
	FLJ10849	2	2	1	2	2	3	
	SP38	2	2	1	2	3	2	
	FLJ20701	2	3	3	1	2	1	
	LDHA	3	3	3	2	2	1	
	OXA1L	3	3	3	1	2	1	
	TRIP10	2	1	2	3	2	2	
	RELN	2	1	2	3	2	2	
	EPAS1	3	3	3	2	2	1	
	RPL10A	3	3	3	2	1	2	
OUTPUT	DGKZ	3	3	3	1	2	2	

Appendix 5: Details of fourteen genes selected from yeast dataset

	ORF	Gene Name	Description
1	YMR199W	CLN1	Cycline, G1/S-Specific
2	YPL256C	CLN2	Cycline, G1/S-Specific
3	YAL040C	CLN3	Cycline, G1/S-Specific
4	YGR108W	CLB1	Cycline, G2/M-Specific
5	YPR119W	CLB2	Cycline, G2/M-Specific
6	YPR120C	CLB5	Cycline, B-type
7	YGR109C	CLB6	Cycline, B-type
8	YMR043W	MCM1	Transcription Factor of the MADS box Family
9	YLR079W	SIC1	Inhibitor of Cdc28p-Clb protein kinase complex
10	YLR182W	SW16	Transcription Factor, subunit of SBF and MBF factors
11	YBR160W	CDC28	Cyclin-dependent protein kinase
12	YDL056W	MBP1	Transcription Factor, subunit of MBF factor
13	YDR146C	SW15	Transcription Factor
14	YER111C	SW14	Transcription Factor, subunit of SBF factor

Appendix 6: Expression ratio of fourteen genes from Spellman yeast alpha factor dataset

Gene Symbol	alpha0	alpha7	alpha14	alpha21	alpha28	alpha35	alpha42	alpha49	alpha56	alpha63	alpha70	alpha77	alpha84	alpha91	alpha98	alpha105	alpha112	alpha119
CLN1	-1.56	-0.93	1.29	1.6663	0.94	0.48	0.07	-0.54	-1.11	-0.77	0.66	1.14	0.99	0.3	0.35	-0.02	-0.41	-0.88
CLN2	-1.41	-0.69	1.39	1.98	0.74	0.21	-0.36	-1.32	-1.5	-1.07	0.35	1.57	1.1	0.56	0.18	-0.32	-0.38	-1.04
CLN3	1.04	0.19	0.47	-1.03	-0.63	-0.68	0.1	-0.02	0.53	0.68	0.51	-0.2	-0.34	-0.59	-0.31	-0.25	0.11	0.44
CLB1	-1.88	-0.95	-1.22	-1.1	-0.91	-0.06	0.5	1.2	1.11	0.22	0.47	-0.02	-0.12	-0.12	0.42	0.98	0.7	0.78
CLB2	-2.36	-1.5154	-1.96	-2.29	-1.36	0.4	1.09	1.54	1.5	0.92	0.05	-0.23	-0.42	-0.29	0.12	0.73	1.35	1.2
CLB5	-0.98	-0.38	0.92	0.97	0.26	-0.09	-0.18	-0.4	-0.35	-0.33	0.7	0.61	0.31	-0.23	0.1	-0.2	-0.21	-0.5
CLB6	-1.79	0.814	2.13	1.75	0.23	0.15	-0.66	-0.77	-0.73	-0.15	1.14	1.41	0.46	-0.21	-0.77	-0.56	-1.04	-0.58
MCM1	-0.27	0.74	-0.49	-1.0886	-0.36	0.81	-0.32	-0.29	-0.24	0.16	0.76	-0.02	0.4	-0.36	-0.51	0.0285	0.4578	-0.02
SIC1	-0.66	0.1	-0.57	-0.21	-0.45	-0.42	-0.77	-0.53	-0.47	0.81	1.41	1.01	0.68	0.21	-0.04	-0.09	-0.28	0.27
SW16	-0.06	-0.18	-0.14	-0.13	0.34	0.13	0.28	-0.03	-0.23	0.1	-0.35	0.11	0.08	-0.16	0.14	0.04	0.17	-0.09
CDC28	0.12	-0.57	0.22	-0.06	0.03	-0.33	0.26	-0.15	-0.18	0.37	-0.18	0.38	0.14	-0.05	-0.03	0.28	-0.33	0.08
MBP1	-0.39	-0.21	-0.46	-0.31	-0.07	0.16	-0.29	0.01	0.22	0.43	0.56	-0.16	-0.33	0.09	0.12	0.49	0.07	0.07
SW15	-1.29	-0.7	-0.33	-0.88	-0.19	0.05	0.02	0.68	0.75	0.64	0.42	-0.07	-0.79	-0.8651	-0.19	0.73	0.64	0.51
SW14	-1.21	-0.26	1.36	1.37	0.54	0.18	-0.85	-0.82	-0.75	0.07	0.89	0.78	0.37	-0.19	-0.11	-0.54	-0.34	-0.47

INDEX