

**A STUDY OF DYNAMIC MODEL OF GENE
REGULATORY NETWORK**

*Dissertation submitted to the Jawaharlal Nehru University
In partial fulfillment of the requirements
For the award of Degree of*

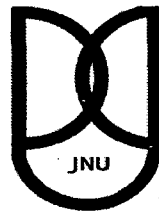
MASTER OF TECHNOLOGY

In

Computer Science and Technology

by

Piyush Kumar Srivastava



SCHOOL OF COMPUTER AND SYSTEMS SCIENCES

JAWAHARLAL NEHRU UNIVERSITY

NEW DELHI – 110067

July 2007

DECLARATION

I hereby declare that this thesis entitled “**A STUDY OF DYNAMIC MODEL OF GENE REGULATORY NETWORK**”, being submitted by me to School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi in partial fulfillment of the requirement for the award of the degree of **Master of Technology** in Computer Science & Technology, is a record of original work done by me under the supervision of **Associate Prof. R.K.Agrawal**.


The results submitted in this thesis have not been submitted in part or full at any other University or Institution for the award of any degree etc.

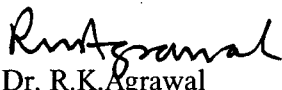

Piyush Kumar Srivastava

CERTIFICATE

This is to certify that thesis entitled “A STUDY OF DYNAMIC MODEL OF GENE REGULATORY NETWORK”, being submitted by Piyush Kumar Srivastava to School of Computer & System Sciences, Jawaharlal Nehru University in partial fulfillment of the requirement for the award of the degree of **Master of Technology** in Computer Science & Technology, is a record of original work done by him under the supervision of **Associate Prof. R.K.Agarwal**.

The results submitted in this thesis have not been submitted in part or full at any other University or Institution for the award of any degree etc.


Prof Parimala N:
Dean
School of Computer & Systems Sciences
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI-110067
Professor and Dean,
SC&SS, JNU,
New Delhi-110067


Dr. R.K. Agrawal
Associate Professor
SC&SS, JNU,
New Delhi-110067

To my Parents.....

Acknowledgements

This thesis would not have been possible without the whole hearted support of a lot of individuals. This is an attempt to acknowledge their help support and guidance, any omissions are involuntary.

This thesis work has been done under the supervision of Associate Prof. R.K.Agarwal. I am immensely grateful to him for his valuable suggestions and continuous guidance throughout the course of this study.

I would also like to thank to entire faculty and staff of SC&SS for their co-operation during the course of study.

All this work would have been impossible had it not been for the love, encouragement and constant support that I got from my family.

I also extend my thanks to my friends and classmates for their care and moral support. I can not forget the moments shared with them during my stay in Jawaharlal Nehru University.

To put into a nutshell.....**A BIG THANKS TO ALL.**

Piyush Kumar Srivastava

Abstract

In this thesis, we study different dynamic model of gene regulatory networks. Modeling, reverse engineering and analysis of macromolecular networks have spurred increasing interest in the computational biology and system biology communities. Biologists need rigorous and flexible tools to describe, infer and study biological networks. There exists a wide variety of dynamic model for gene regulatory networks. A basic review of the biology behind gene regulation is introduced along with the formalisms used for networks of such regulatory interactions. Topological measures of large-scale complex networks are discussed and then applied two different modeling paradigms to simulate gene expression data for a given microarray data.

Contents

Declaration	i
Certificate	ii
Acknowledgement	iv
Abstract	v
1 Introduction	1
2 Concepts of Genetic System	4
2.1 Bioinformatics and Computational Biology.....	4
2.2 System Biology.....	5
2.2.1 Techniques associated with System Biology.....	6
2.3 Basics of Gene Expression.....	6
2.3.1 DNA Sequence.....	6
2.3.2 Gene.....	7
2.3.3 Gene Expression.....	8
2.3.4 Regulation of Gene Expression.....	9
2.3.5 Genetic Regulatory System.....	10
2.4 Measurement Technologies.....	10
2.4.1 DNA Microarray Technology.....	10
2.4.2 Serial Analysis of Gene Expression.....	12
3 Gene Regulatory Network	13
3.1.1 Definition.....	13
3.1.2 Biological Properties of Gene Regulatory Network.....	14
3.1.3 Utility.....	16
3.2 General Properties of Modeling Formalism.....	17
3.2.1 Physical Vs Combinatorial Model.....	17
3.2.2 Dynamic Vs Static Model.....	18

3.2.3 Synchronous Model.....	18
3.2.4 Deterministic Vs Stochastic Models.....	19
3.3 Bayesian Network Model.....	20
3.4 Boolean Network Model.....	20
3.5 Petri-Net Model.....	22
3.6 Stochastic Model	23
3.7 Differential Equation Model.....	24
3.7.1 Ordinary Differential Equation.....	24
3.7.2 Weight Matrices.....	25
3.7.3 Piece-wise Linear Differential Equation Model.....	26
3.7.4 S-Systems.....	27
3.7.5 State Space Model.....	28
4 Modeling gene regulatory network	30
4.1 Inferring gene regulatory network using singular value decomposition.....	31
4.1.1 Singular value Decomposition.....	31
4.1.2 To find connectivity matrix for a single micro array dataset.....	32
4.1.3 Algorithm.....	35
4.2 Linear Gaussian State Space Models.....	35
5 Experimental Work & Results	38
5.1 Experimental Objective.....	38
5.2 Datasets Used.....	38
5.2.1 System Characteristics.....	38
5.2.2 Used Tools.....	39
5.2.3 Databases.....	39
5.2.4 S.O.S DNA Repair Network.....	39
5.3 Experimental Results.....	40
6 Conclusion	46
References	48

Chapter 1

Introduction

The study of molecular networks of molecular interaction after the availability of complete genome sequence and high throughput post genomics expression data analysis is an active research area since last few years. A major challenge to biologists is to understand functional behavior of gene regulatory network (GRN) and also the complex intermolecular interaction among the genes in a cell [2].

The main goal of genomic revolution is to understand the genetic cause behind the characteristics of organisms [7]. With the rapid development of DNA microarray technology, reverse engineering the gene regulatory network from time series gene expression data has become more important to understand the complex relation between genes, proteins and other substances, also to investigate functions of genes and to reveal cellular process in the cells. It can also help to predict the future dynamical behavior of the system [31].

A genetic regulatory system is a network which consists of a group of DNA, RNA, proteins and other molecules to describe the mechanisms of gene regulation. Deoxyribonucleic acid (DNA) is a nucleic acid molecule that contains the genetic information used in the development and functioning of organisms [11]. A gene is a small region of DNA sequence. According to central dogma of molecular biology, genes are transcribed into mRNA, which are then translated into proteins. It is important to know that which genes are regulated and which are regulator. DNA microarray technology provides an efficient and effective way to measure expression levels of thousands of genes simultaneously. The Gene Regulatory networks are described by some specific properties. Some of them are topology, transcription

control, robustness and noise. Gene networks can be model using different approaches. Once the model is chosen, the parameters need to be fit in to the data. Even the simplest network models are complex systems having a lot of parameters, and fitting them is a non trivial process, known as network inference, network identification and reverse engineering [18].

Genetic network model may be of different types depending upon the amounts and type of data available [28]. They can be physical or combinatorial, static or dynamic, deterministic or stochastic types etc. Some of the popular models for GRN are Boolean network models, Bayesian network model, State space models [28].

The disadvantages of gene network construction from microarray data is that while the gene network contains a large number of genes, the information contained in gene expression data is limited by the number of microarrays, their quality, the experimental design, noise, and measurement errors. Therefore, estimated gene networks contain some incorrect gene regulations which can not be evaluated from a biology view-point. In particular the direction of gene regulation is difficult to decide using gene expression data only [34]. Hence the use of biological knowledge, including protein-protein and protein-DNA interactions, sequences of the binding sites of the genes controlled by transcription regulators, literature and so on, are considered to be a key for microarray data analysis. The use of biological knowledge has previously received considerable attention for extracting more information from microarray data [35, 38].

This thesis is organized as follows. Chapter 2 introduces the various concepts of genetic system. The chapter 2 also includes a brief introduction of microarray and their measurement technologies. Chapter 3 reviews some of the common mathematical and computational modeling techniques for gene regulatory networks. This includes a brief introduction of Bayesian model, Boolean network model, differential equation model, state-space models and many more. Chapter 4 describes two different modeling paradigms used in our experiments i.e. modeling gene

regulatory network using state space model technique and inferring gene regulatory network using singular value decomposition techniques. Experimental setup and result are discussed in chapter 5. Finally, chapter 6 includes conclusion and future works.

Chapter 2

Concepts of Genetic Systems

2.1 Bioinformatics and Computational Biology

Bioinformatics is area of active research, which involves the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, and chemistry to solve biological problems at the molecular level. Informatics has traditionally been a discipline in which mathematicians, computer scientists, statisticians, and engineers develop technologies for supporting information management in fields like healthcare. Bioinformatics is now involved in these activities by organizing biological data related to genomes with a view to applying this information in agriculture, pharmacology, and other commercial applications. In Bioinformatics and computational biology, functions are produced by a set of macromolecules that interact with each other at different levels. Genes and their products, proteins, participate to form a regulatory network that manages the response of the cell to external input signals [28].

In literature, bioinformatics and computational biology are often used interchangeably. Bioinformatics involves the creation and development of algorithms, computational and statistical methods and theory to solve formal and practical problems from the analysis of biological data at macro level. On the other hand computational biology involves the hypothesis driven analysis and investigation of a specific problem of biology using computer technology carried out with experimental data with the main objective of discovering and advancement of biological knowledge. It can be said that bioinformatics involves the analysis of information, while computational biology is concerned with the hypothesis [31].

Major research areas in bioinformatics and computational biology are the followings [31]:

- a. Sequence Analysis
- b. Genome Annotation
- c. Sequence Matching
- d. Protein Folding
- e. Comparative Genomics
- f. Protein Analysis
- g. Computational evolutionary Biology
- h. Measuring Biodiversity
- i. Analysis of gene expression
- j. Analysis of mutation in cancer
- k. Modeling biological system
- l. High throughput image analysis

2.2 System Biology

System Biology is often used very widely in the bioscience. It can be defined in different ways [39]:

- a. According to few sources in literature, it is a field of study where we study the interaction between the components of biological system and how these interactions give rise to the function and behaviors of that system e.g. Enzymes, metabolite, E.coli etc.
- b. Other sources in literature consider systems biology as a paradigm, defined in antithesis to the so-called reductionism paradigm, although fully consistent with the scientific method.
- c. And some sources in literature consider it as a socioscientific phenomenon defined by the strategy of pursuing integration of complex data about the interactions in biological systems from diverse experimental sources using interdisciplinary tools and personnel.

2.2.1 Techniques associated with system biology

According to the interpretation of system biology as the ability to obtain, integrate and analyze complex data from multiple experimental sources using interdisciplinary tools, some typical technology platforms are [39]:

- a. Gene expression measurement
- b. Protein levels through two dimensional gel electrophoresis and mass spectrometry including phosphoproteomics to detect chemically modified proteins.
- c. Metabolomics for small molecule metabolites
- d. Glycomics for sugar
- e. Interaction for interactomes
- f. Modeling Gene Regulatory Network

A major challenge in system biology is gathering different kinds of information which one can then be used for computation.

2.3 Basics of Gene Expression

2.3.1 DNA Sequence

Deoxyribonucleic acid (DNA) is a nucleic acid that is used in the development and functioning of the body of all living organisms. DNA contains information and it is often compared to a set of blueprints.

A DNA sequence is succession of four letters which represents the structure of DNA molecule, having the capacity to carry information. The letters in sequence are *A*, *C*, *G*, and *T*, representing the four nucleotide subunits of a DNA strand - adenine, cytosine, guanine, thymine bases respectively covalently linked to phosphorus-backbone. Figure 2.1 shows detail structure of a DNA. A sequence of nucleotides greater than four is called a DNA sequence.

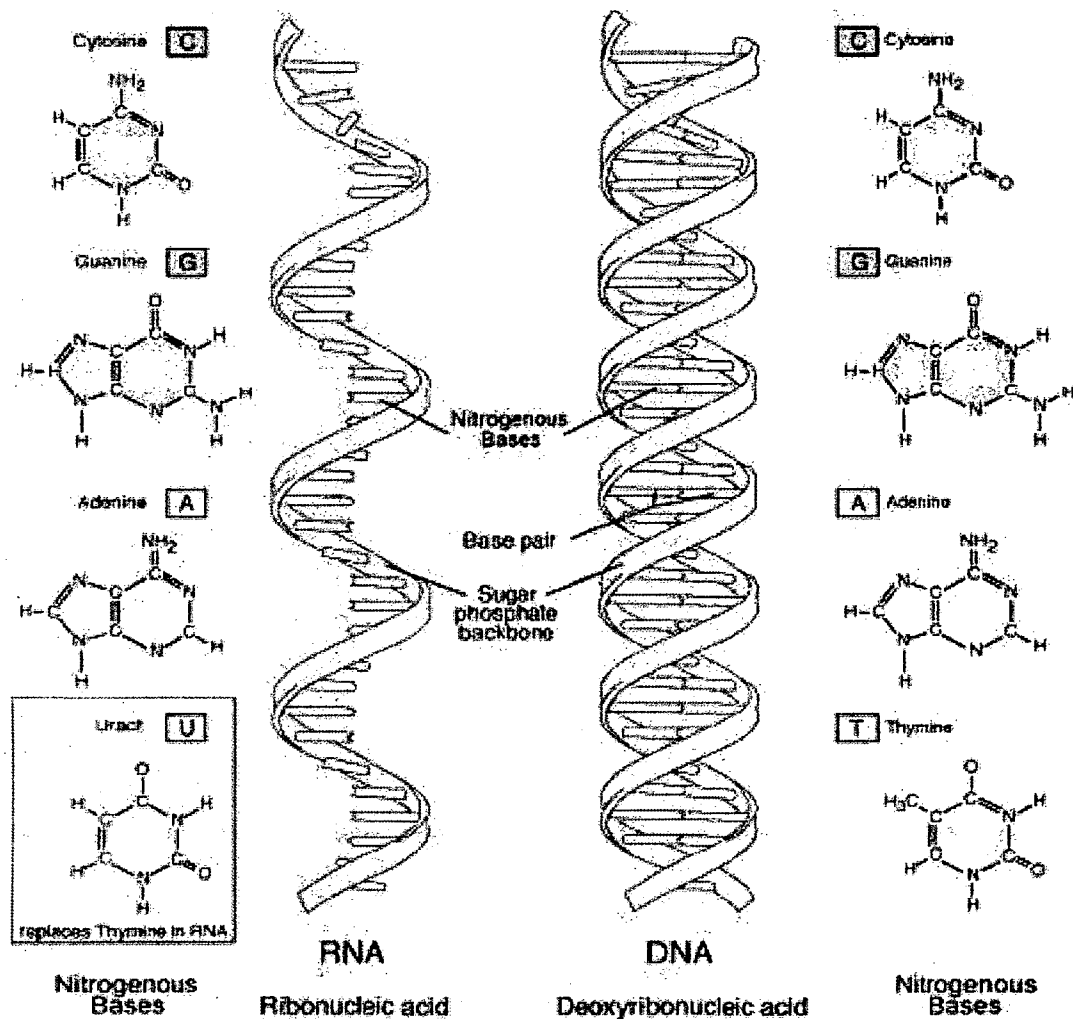


FIGURE 2.1: Structure of DNA [44]

2.3.2 Gene

A gene is a small region of DNA sequence that contains the necessary information to produce the messenger RNA (Messenger Ribonucleic acid). A gene can also be defined as a region of DNA (a sequence of nucleotides) that controls a hierarchical characteristic. A gene consists of two parts: coding DNA sequences (introns), that is transcribed into mRNA and non coding DNA sequences which can not be changed into mRNA as shown in Figure 2.2. The number of genes in a DNA sequence may vary from 1000 base pairs to several hundred thousands base pairs.

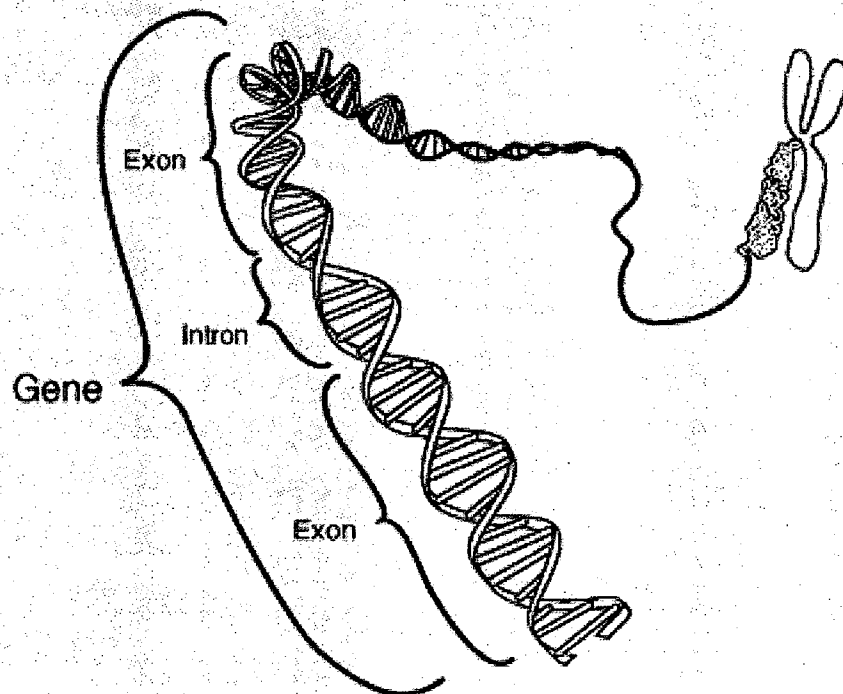


FIGURE 2.2: A gene [45]

2.3.3 Gene Expression [47]

Gene Expression is a multi step process where the information coded within a gene is changed into protein. One can describe transformation to protein from gene in terms of two processes. The process of changing of gene to messenger mRNA is known as transcription, and the process of changing from mRNA to protein is known as translation. As we know a gene consist introns and exon, while introns (coding sequences) are transcribed into mRNA. This process is followed by post transcription process and translation which translates mRNA to protein. After this folding, post translational, modification and targeting processes are performed. The different steps of gene expression process are shown in Figure 2.3.

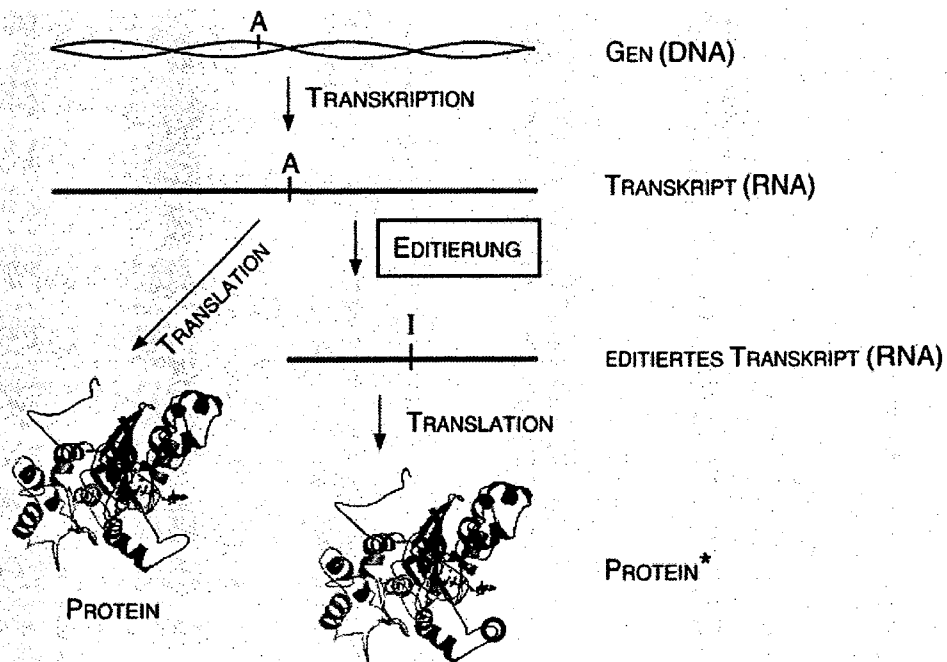


FIGURE 2.3: A Gene Regulation Process [47]

2.3.4 Regulation of Gene Expression

All genes encode protein by gene expression process and most of these proteins are enzymes. These are regulated by controlling the activity of enzymes produced. Regulation of gene expression concerned with the cellular control of the amount and timing of changes to the appearance of the functional product of a gene. Gene expression process involves the following stages [32].

- a. Chemical and structural modification of DNA sequence
- b. Transcription into mRNA
- c. Translation in Protein
- d. Post Transcriptional modification
- e. RNA transport
- f. mRNA degradation
- g. Post Translation modification

Hence a gene regulation system consists of genes, cis-element and regulators. The regulators are most often proteins called transcription factors, but small

molecules like RNA and metabolites sometimes also participate in the overall regulation. The cis-region serves to aggregate the input signal mediated by the regulators and the regulatory connection between them, together with an interpretation scheme from gene network.

2.3.5 Genetic Regulatory System

In the regulation of gene expression, the expression of a gene may be controlled during RNA processing and transport mainly in eukaryotes, RNA translation, and modification of proteins. The degradation of proteins and intermediate RNA product can also be regulated in the cell. The proteins fulfilling the above regulatory functions are produced by other genes. This give rise to genetic regulatory system structured by networks of regulatory interaction between DNA, RNA and proteins and small molecules [28].

One can interprets gene regulatory network as a dynamic network, which describes the interaction between genes and other substances of the cell. It also describes the functionality of the genes or a group of genes. In this network nodes are represented by genes and the regulation of one gene by other genes or a group of genes is represented by arcs or edge.

2.4 Measurement Technologies

High throughput technologies for measuring mRNA expression level are briefly discussed below [28]:

2.4.1 DNA-Microarray Technology

In living organism there are thousands of cell present. The entire cells are not active at all time. Only few of them are responsible for an activity of organism at one time. To study which genes are active and which genes are inactive in different cells helps to understand the cell function. With the help of DNA micro array technology one can measure the expressions of thousands of genes at same time. This is one of the most advance techniques to measure the expressions of mRNA. A DNA Microarray

known as a gene or genomes chip in which there is a group of DNA spots, where each spot represents a single gene's expression [43].

DNA microarrays work on the principle of base pairing. At first level microarrays data are measured in following steps: RNA from a cell is extracted. This RNA (targets) is then reproduced and marked with fluorescence and hybridized to existing DNA (probes) on the microarray. After hybridization, these probes that were hybridized with targets are fluorescent. The computer scanner is able to detect this fluorescence. These probes which are fluorescent correspond to the genes that were expressed in the cell [43].

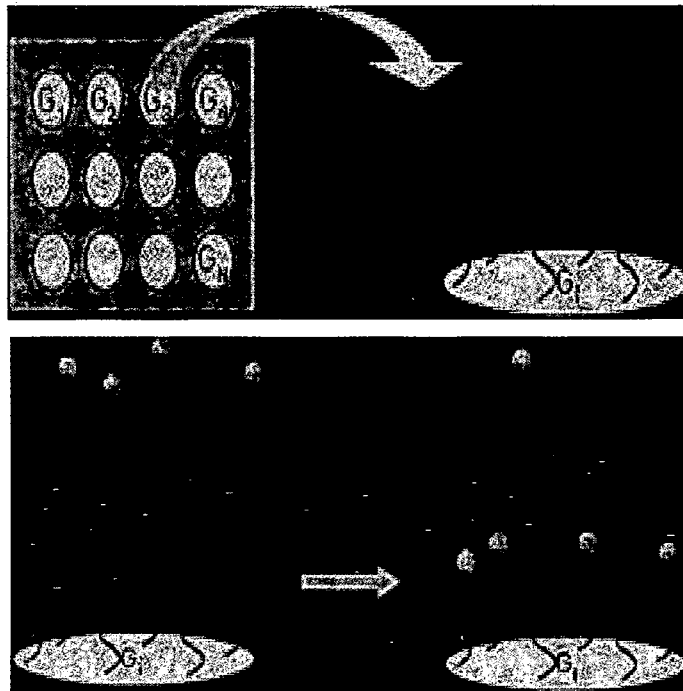


FIGURE 1.4: Illustration of Hybridization [43]

Two most common DNA microarray techniques are cDNA arrays and oligonucleotide arrays [28].

- a. cDNA Microarray Technology
- b. Oligonucleotide Arrays

2.4.2 Serial Analysis of Gene Expression

Serial analysis of gene expression (SAGE) is a powerful measurement technique with digital analysis for measuring gene expression in a given cell. The method of this technique is to first capture the mRNA molecules present in the cell. After that one determines the genes from which these mRNA are transcribed and then calculate the total number of mRNA for each gene. The profiles of mRNA expressions of different cells are very much different from those of infected cells. To analysis these gene activity, researchers can determine the gene activity related to particular disease and conditions allowing for the development of specific drug development [28].

We know that mRNA ends with a strings of As. To capture these mRNA microscopic beads are baited with strings of approximately twenty Ts. Since A and T form a strong chemical bond so the mRNAs are washed away from these beads. Then the mRNA becomes attracted to the beads. A magnet is used to extract the beads and mRNAs. These mRNA are then copied back into DNA with the use of reverse transcripts.

These DNA fragments are then quantified using genetic sequencing. In brief, we can say that SAGE works by capturing RNA molecules rewriting them as DNA [28].

Chapter 3

Gene Regulatory Network

The objective of genomic revolution is to understand the genetic effect on characteristics of living organisms. In molecular biology, functions are produced by a set of genes or their product, which interact with each other at different levels. Hence genes and their product form a network, which is known as gene regulatory network. Genes, their products (such as protein, mRNA etc.) and other substances of the cells interact with each other to control the response of the cell to external input signal [32]. It is of interest and challenge to biologists to understand the mechanism by which the regulations of genes govern and also to identify gene which plays the role of regulator and gene which plays as a regulated gene among all the genes. The development of microarray technology helped to understand these problems easily, since with the help of microarray technology expression of thousand of genes of a given organism can be measured simultaneously at the same time on the same chip. This technique helps in research of re-engineering gene regulatory network from given experimental data [11].

3.1.1 Definition

A gene regulatory network consists of genes, cis elements and regulators. Mostly the regulators are proteins also called transcription factors but some small molecules (RNA and metabolites) also participate in gene regulation process. The interaction and binding of regulators to cis regions of genes controls the gene's expression level during mRNA transcription process. So the gene, regulators (proteins) and the regulatory connections between them form a gene regulatory network. In other words gene regulatory network can be described as a set of genes, their products that interact with each other and other substances in the cell to control the rate at which genes in

the network are transcribed into mRNA. Figure 3.1 shows a hypothetical gene regulatory network with different level at which it is modeled [12].

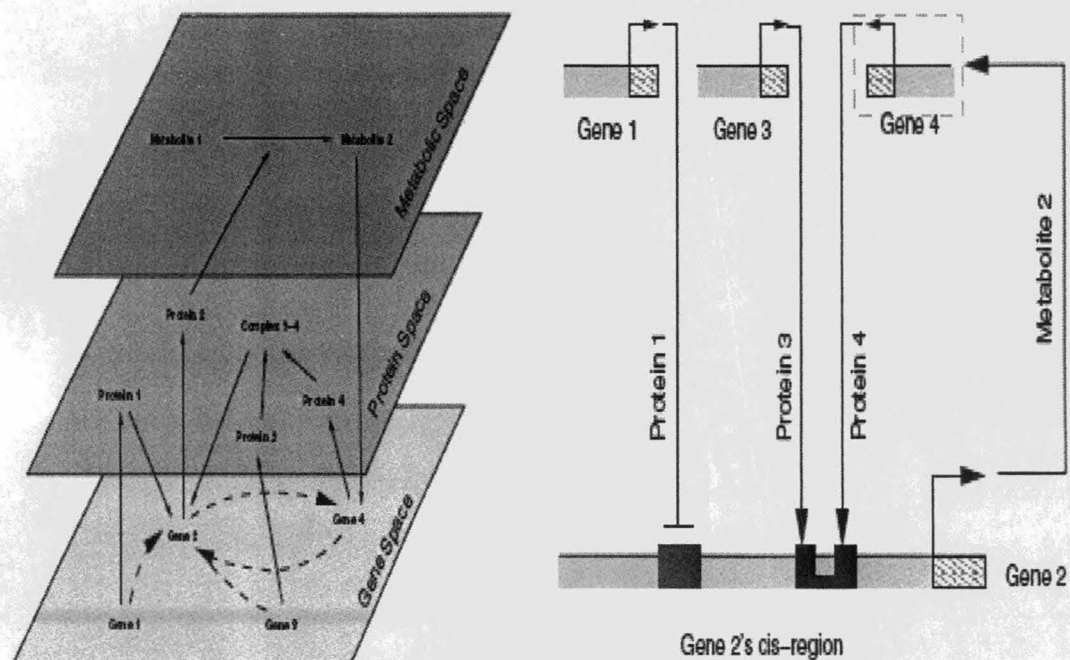


FIGURE 3.1: A hypothetical gene network [12].

Gene regulatory network can be viewed as a complex network where inputs are proteins and outputs are controlled level of gene's expressions. The node of the network can also be viewed as a function performed by a gene or a group of genes. Biologists suggest many mathematical models of gene regulatory network. Some of them are Boolean network [29], Bayesian network [20], Petri-net model [46], Graphical Gaussian state space model [8], and stochastic process models [32].

3.1.2 Biological properties of Gene Regulatory Network

The basic properties of gene regulatory network are discussed below [12]:

Topology

The topology of a network can be defined as the physical and logical arrangements of connection between the nodes of a network. It can be said as the starting point of the modeling of GRN. One of the most special feature in modeling gene regulatory network is that the topology of GRN is sparse i.e. the number of edges is very small in the network [21]. This property helps to prune the search space during network inference. In the GRN, nodes are represented by genes and the directed edges, which connect the nodes, represent effect of one gene on the other (excitation or inhibition).

There are two types of topology that appears in GRN:

1. **Scale free network topologies:** In recent work it has been seen that the frequency distribution of connectivity of nodes shows a normal distribution [28]. The appropriate distribution may belongs to a class of power law distribution shown by the given equation:

$$P(x) \approx x^{-\gamma} \quad (3.1)$$

Where $P(x)$ is the probability, x is the degree of a vertex, and γ is network specific constant.

These types of network topologies are known as scale-free network topologies

2. **Small-world network topologies:** In 2003 Watts define small-world graph topologies [28]. A graph with n -vertices and vertex degree k that exhibits

$$L \approx L_{random}(n, k) \approx \frac{\ln(n)}{\ln(k)} \text{ and } C \gg C_{random} \approx \frac{k}{n} \text{ for } n \gg k \gg \ln(n) \gg 1. C \text{ is}$$

the clustering coefficient which can be calculated by:

$$C = \frac{2}{n} \sum_{v=1}^n \left(\frac{k_v(k_v - 1)}{2} \right) \quad (3.2)$$

Where k_v the number of neighbors of vertex v . L is the average number of links connecting two nodes. L_{random} and C_{random} refer to the path length and clustering coefficient with same k and n respectively.

Transcriptional control

Transcription is the process by which DNA transcribed into mRNA. The cis-region depicts as the superposition of the effects of all transcription factors in gene regulation. The range of effects of the cis processing logic on the input transcription factor signal has been recognized for only few cases. From that one can learn that the cis function is a multi valued complex function of the input concentration even only for two inputs. However the functions become simpler and can be decomposed into linear combination of independent functional signal contributions, when the functional cis-elements are known i.e. at least when experiments are carried over the same condition and for the genes on the periphery of the network [12].

Robustness

Robustness is the quality of being capable to handle stresses, pressure or changes in the structure of the system. A system is said to be robust if it is capable to handle variations in its functional environment with minimum damage. Real gene regulatory systems are very robust in their parameter values [1]. But for only specific choice of topology guarantees such type of strong robustness [1, 5].

Noise

Noise is defined as unwanted data, which is present in a system. Noise is an integral part of a Gene Regulatory network. As we know that GRN is stochastic in nature [33], so even amount of small noise in the gene expression measurement can affect the whole network. The network controls the noise through the feedback. In some cases noise can give better result to find functional characteristic of the network.

3.1.3 Utility

A gene regulatory network can be said as a blueprint for understanding all the functional co-operativity and interaction among the genes. So GRN are representation to know the knowledge studied about the system. The gene network is used to classify that gene which act as a regulator and gene which is regulated by other genes and other influences in the cell. Using GRN the interaction among the genes can be

studied at large scale gene data. Using GRN one can predict the future behavior of the system. Using GRN one can also understand that gene which is responsible for a particular disease and it can help in identifying molecular target for specific drug or drug for specific target. Since we have GRN of different organisms so one can do comparisons among them to understand the change in GRN with respect to time (i.e. evolution of GRN) [12].

3.2 General Properties of Modeling Formalism

There exists many types of modeling formalism exists for gene regulatory network. The choice depends on the type, amount of data (gene expressions) available, prior information about the network, experimental and computational resources and other factors. Briefly, we discuss some models which are classified based on their general properties and described below [2, 12]:

3.2.1 Physical Vs Combinatorial Model

Most of the gene regulatory systems are described by differential equation, which shows the quantitative relationships between the state variable in the systems. Even through physical model are used to run the simulation to predict the future behaviors of the gene regulatory network, but using physical model it is difficult to identify even simple features (i.e. one gene effect on the other).

On the other hand, a typical combinational model can be represented by a graph. In graph, genes are represented by nodes and effect of one gene on the other is represented by an arc. These models start from higher-level features of GRN by defining features of interest like gene expression levels, the nature of relationships. Due to the higher level of modeling, the combinatorial models are most often qualitative and effective methods for their learning even for small number of observations (relative to the number of variables); hence there exists inference from a GRN.

3.3.2 Dynamic Vs Static Models

As most of the data available in modeling GRN is time series data, a dynamic gene regulatory network model can be described as the change of gene expression with respect to time, which is represented as follows:

$$\frac{dx_i(t)}{dt} = f_i(x_{i_1}(t), x_{i_2}(t), \dots) \quad (3.3)$$

Where x_i on the left is concentration of gene i at time t .

x_{i_1}, x_{i_2}, \dots are concentrations of molecules that influence x_i at time t .

$f_i(\dots)$ is the rate function, which is function of concentration of different genes.

Here, one can interpret each node in GRN as a function, which takes some inputs and gives output based on them. Dynamic model is usually more complex than static model because they characterize the exact interactions among the inputs. So it will take more amounts of input data for large number of parameters. Few examples of dynamic models are Boolean Network and linear Differential equation models.

On the other hand static models do not have time component. Static models only define the topological characteristics of the GRN. It reveals the combinatorial interactions among genes and the nature of interactions. Few examples of this modeling formalism are Graph Theoretic Model, Bayesian Network Models and Linear additive models [12].

3.3.3 Synchronous Models

In DNA-microarray technology of measuring gene expression, the observations are measured at the same time so large scale gene expression measurement drive asynchronous models for GRN. In these models time is discretized to the interval between consecutive observations. If these time interval are taken as very small the above equation (3.3) can be written as [12]:

$$\frac{x_i(t_{j+1}) - x_i(t_j)}{t_{j+1} - t_j} \approx f_i(x_{i_1}(t_j), x_{i_2}(t_j), \dots) \quad (3.4)$$

Where t_j and t_{j+1} are two consecutive observation times.

3.3.4 Deterministic Vs Stochastic Models

In deterministic models the expression states of the genes are either given by a formula or belong to a specific class. A gene's expression will remain the same when it is measured at two different times or places while keeping all other parameters the same. The accuracy of the observed expression values will depend on the experimental setup, and can be refined indefinitely with technological advances. The edges in GRN representing relationships are also deterministic, which is similar to node states.

Stochastic models, on the other hand, start from the assumption that gene expression values are described by random variables, which follow some probability distributions. The difference with the deterministic models is fundamental: randomness is modeled to be intrinsic to the observed processes, and thus all things being equal, a gene's expression on two different occasions may be different. Stochastic edges indicate probabilistic dependencies, and their absence may indicate independencies between nodes. Usually it is not easy to interpret the output of stochastic models [40].

Stochastic gene network models are generally useful for reconstructing gene networks from expression data because of the inherent noise present in them. In order to take into account of imprecision and fluctuations in the measurements, it is assumed that each observed quantity is drawn from an underlying set of values, or a statistical distribution, that the observable variable may take. Then, assessing whether a gene is differentially expressed with respect to another is transformed into well studied problems of statistical hypothesis testing.

3.3 Bayesian Network Model

A Bayesian Network Model is a class of graphical probabilistic model that represent a set of nodes and their probabilities dependencies on each other [10]. Bayesian Network relates probability with graph theory. A Bayesian Network is denoted by a directed acyclic graph $G(V, E)$ where $v_i \in V$ are random variables representing genes and the edges indicates the dependence of one node on the other. The random variables are drawn by conditional probability distribution $P(v_i/v_j)$ where v_i is dependent on v_j . With this assumption, one can find the decomposition of the joint distribution over all variables drawn to the conditional distribution of all the nodes as:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(v_i/P_a(v_i)) \quad (3.5)$$

3.4 Boolean Network Model

This model was given by Kauffman Glass and Kauffman [28]. This model is represented as a graph in which nodes are genes which states may be either 0 or 1. There is no intermediate level of transcription. In this graph each node is connected to the other shown in following figure 3.2. Incoming edges show that the binary values from other node are taken as inputs. These values are sent through a Boolean function that describes the current state of the node. So a Boolean network can be stated as a dynamic model of synchronous interaction between nodes in a network. Boolean Network is specified as NK-network, where N is the total number of nodes and K is the maximum number of incoming edges for each node. If the value of K is small then it shows that there are only few genes, which are responsible for controlling the activities of any single gene. Small K value shows that the connectivity matrix for GRN is sparse [28].

The goals of reverse engineering Boolean network from gene expression data infer both the topology and the Boolean function of each node. An example of Boolean network is shown in figure 3.2. For this Boolean network $N=3$ and $K=2$.

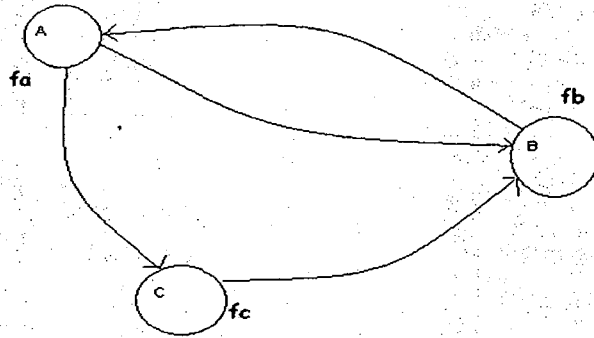


FIGURE 3.2: A simple Boolean network [12]

For this network,

$$f_A(B) = B$$

$$f_B(A, C) = A \text{ and } C$$

(3.6)

$$f_C(A) = \text{not } (A)$$

TM-14676

Wiring Diagram and Truth Table:

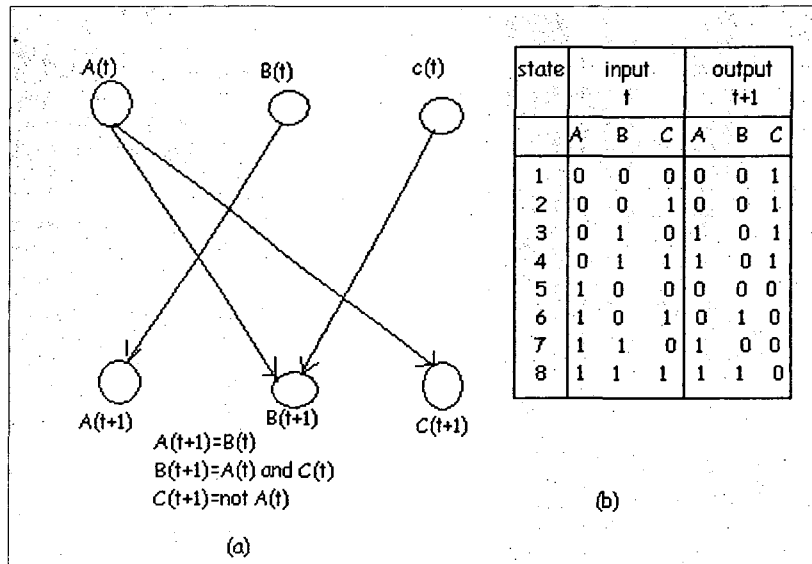


FIGURE 3.3: (a) Wiring Diagram (b) Truth Table [12]



Liang et al, Akutsu, Akutsu et al [29] and many others used Boolean network model in both forward modeling and reverse modeling GRN. They noticed that there are many genes which have different regulatory effect based on their level of expression. It is suggested that there is not a direct correspondence between the dynamic behavior of Boolean system and their continuous counterpart indicating a quantitative loss of behavior information [28].

3.5 Petri-Net Model

A Petri-Net Model [46] (also known as Place/Transition network or P/T Net) is one of the mathematical representations to describe distributed network. In modeling language one can graphically represents the structure of a complex system as a directed bipartite graph. It consists of nodes (places and transition) and arcs. Arcs are present between places and transition not present in between place to place or transition to transition. The place from which arc goes to transition is known as input place. And the place, which receives the arcs, is known as the output place of the transition. Place contains some tokens. A distribution of tokens over the places of a net is called a *marking*. Transitions performed input tokens by a process known as *firing*. A transition is *enabled* if it can fire, i.e., there are tokens in every input place. During the process of a transition fires, it consumes the tokens from its input places, performs some processing task, and places a specified number of tokens into each of its output places. The working of Petri-net network may be non-deterministic i.e. more than one transition can be enabled at the same time [46].

To represent a gene regulatory network one can represent the places by genes or proteins and transition node can be represented by the regulation or inhibition. A simple model of Petri net network is shown in following figure 3.4.

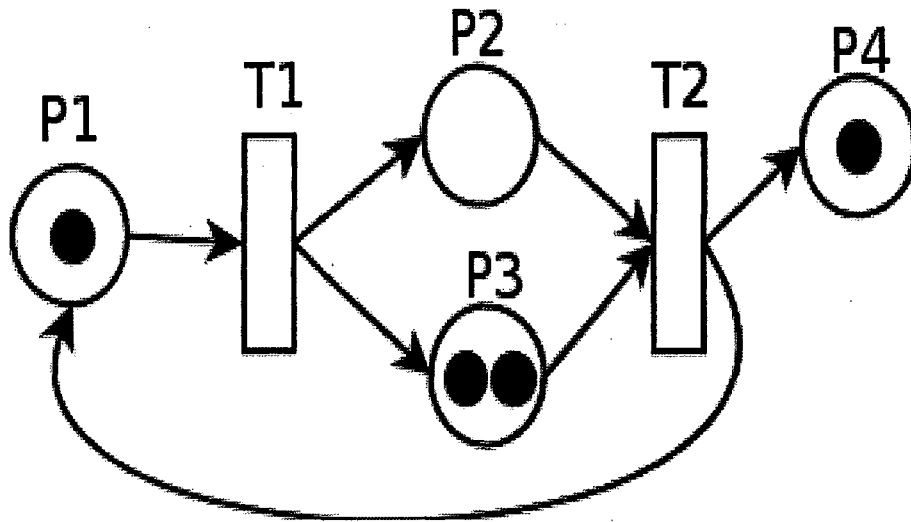


FIGURE 3.4: A simple Petri-Net Network [46].

Stochastic Models

Stochastic models remove many of the shortcomings presents in other models (differential equation models, Boolean network models). One of these drawbacks is the assumption of continuous rate of protein production. Continuous model do not represent the parameters of transcription factor in gene regulation process. In fact protein production rate are not continuous [25].

If one include the noise in the gene regulation process the regulatory path of the cells can be changed. It is possible that evolution has selected network which can produce deterministic behavior from stochastic inputs in noisy environment. So the noise can change the topology of the network [9] and noise can also act as a stabilizer for other systems [17].

There are two methods for modeling stochastic models for GRN.

1. Using Stochastic differential equation the stochastic model can be defined as [28],

$$\frac{dx_i}{dt} = f_i(x_i) + v_i(t) \quad (3.7)$$

Here $v_i(t)$ is noise present in the measurement of gene expression data.

2. Using Probability function, during each time interval there is a probability of transitioning of molecule from one state to another. So from this we can get a probability density function for the behavior of the system. This equation of probability density function is known as 'Master Equation' and can be solved by technique Gillespie algorithm [25].

3.7 Differential Equation Model

Differential Equation Model can be said as an alternative to the Boolean model described above. Differential equation may be considered as the starting point for quantitative modeling of complex gene regulatory network system. These models are continuous and deterministic modeling formalism. These models can describe non-linear and emerging phenomenon of complex dynamical systems.

Suppose that a gene regulatory network has N genes and let $C_1(t), C_2(t) \dots C_N(t)$ represents the concentrations of all N genes respectively.

Then the general form of the equation for each N gene is,

$$\frac{dC_i(t)}{dt} = f_i(C_1, C_2 \dots C_N), i = 1, 2 \dots N \quad (3.8)$$

Where the function f_i provides the aggregate effect of its argument of C_i . The arguments of f_i may be subset of all different concentrations.

However differential equation model contain many parameters, which must be obtained from observed data [28].

3.7.1 Ordinary Differential Equations

Ordinary differential equation can be derived for the chemical rate equation for GRN.

As one know that the rate law for a substrate S and a Product P is given by [14],

$$\dot{S} = -\frac{V_{\max 1} S}{K_{M1} + S} \quad (3.9)$$

$$\dot{P} = \frac{V_{\max 2} S}{K_{M2} + S} \quad (3.10)$$

The above equation is known as Michael-Menten Equation, which models the concentration of a protein and gene pair. Here V_{\max} and K_M are the parameters which control the rate of change of the substrate.

This equation can be generalized as follows,

$$\frac{dr_i}{dt} = f_i(P) \quad (3.11)$$

$$\frac{dp_i}{dt} = g_i(r) \quad (3.12)$$

Where p is a vector of protein concentrations. r is the vector of mRNA concentration. f_i and g_i are updating function and they are sigmoid in shape.

3.7.2 Weight Matrices

A weight matrix tries to model gene regulatory network using linear coefficient, which represents the relationship between genes [41]. So an individual gene's expression can be calculated by summation of all its regulatory inputs (which may be multiplied by its regulatory coefficients). This scheme can be represented as a matrix form where the entry (i, j) represents the effect of j th gene on i th gene.

So we can calculate the total regulatory input $r_i(t)$ of gene i as

$$r_i(t) = \sum_j w_{ij} u_j(t) \quad (3.13)$$

Where w_{ij} is the weight coefficient of gene i on gene j and $u_i(t)$ is the concentration of mRNA of gene i at time t .

If value of w_{ij} is positive then it means gene i activates gene j, otherwise it inhibits gene j. If the value of w_{ij} is equal to zero, then it represents that there is no effect of gene i on gene j (i.e. no interaction).

Let us take a matrix \mathbf{M} that expresses the expressions of all genes. Let \mathbf{A} represents the vector of the weight matrix row of one gene, and \mathbf{B} is the vector represented the relative expression level of the gene of interest at the given state transition. Then the system of equation $\mathbf{MA}=\mathbf{B}$ can be solved. Since there may be not sufficient data i.e. more genes than data points the equation will not provide unique solution which is not desirable [41].

3.7.3 Piece-wise Linear Differential Equation Model

One another variation of differential equation for modeling gene regulatory network is ‘Glass Network’ models with differential equation model. This model has been proposed as a simplified model of GRN [13] as well as the underlying model for the reverse engineering of gene regulatory network and to model Neural Network [13].

The dynamics of gene regulatory system can be represented by the following equation:

$$\frac{dx_i}{dt} = -\gamma_i x_i + F_i(X_{i1}(t), X_{i2}(t), \dots, X_{ik}(t)), i = 1, \dots, N \quad (3.14)$$

Where x_i is the mRNA concentration of a gene i at time t.

$X_i(t)$ - Binary variable. ($X_i = 1$ if $x_i \geq \theta_i$ and $X_i = 0$ if $x_i < \theta_i$ where θ_i is some threshold value.)

r_i - Positive decay constant.

F_i - A Boolean function, which depend on K binary input variables.

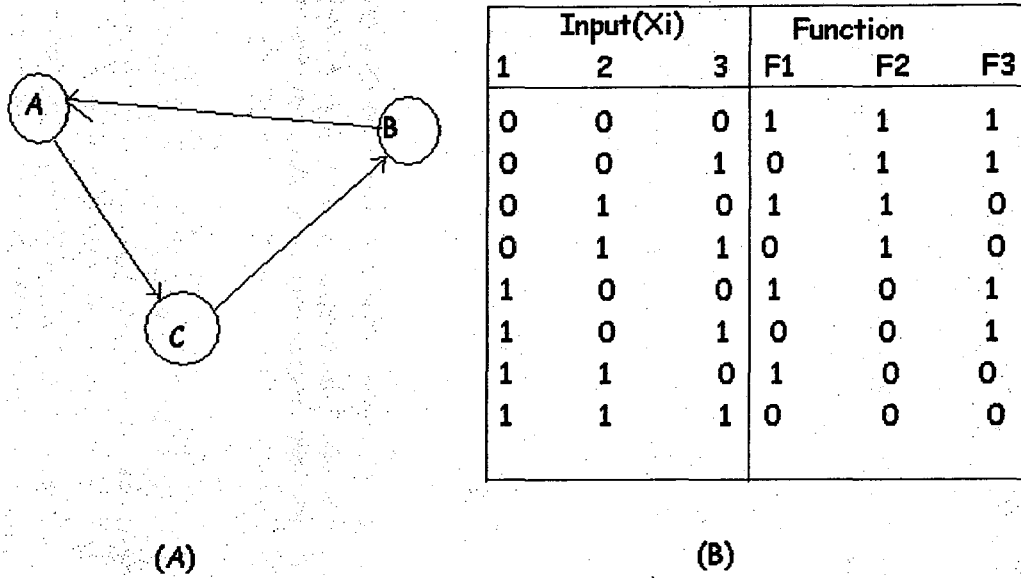


FIGURE 3.5: (A) Genetic Circuit schematic of the repressator [11].
 (B) Truth table define the function [11]

The above example is a network using a glass network, which is the repressator [42].

3.7.4 S-Systems

S-Systems (Synergistic and saturable system) are a modeling techniques used for biochemical pathways, genetic network and immune system [29]. S-Systems can be represented by a non-linear differential equation having the form,

$$\frac{dX_i(t)}{dt} = \alpha_i \prod_{j=1}^n X_j(t)^{g_{i,j}} - \beta_i \prod_{j=1}^n X_j(t)^{h_{i,j}} \quad (3.15)$$

Where α and β are rate constants and g and h are exponential parameters (kinetic orders). $X(t)$ expresses mRNA concentration of gene i at time t . In S-systems model each dimension represents the dynamics of a single variable. The dynamics of each variable can be represented in terms of the difference of two products of power-law functions- one describing the influxes and other describing the effluxes. The

weakness of S-system is the large number of parameters required. Suppose that the system has n -dimensional vector then it requires $n*(2n+1)$ parameter to describe the dynamics of the system.

3.7.5 State Space Model

In literature, control theory or state space model for ordinary differential equation is in existence since last decade. A state space representation is a mathematical model of a physical complex system and represent in terms of as a set of inputs, outputs and states. The variables are commonly expressed by vectors and equations are written in matrix form.

Suppose that a linear system has p inputs, q outputs and n states variables. Then the general state space model can be represented by following equation,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad (3.16)$$

$$y(t) = C(t)x(t) + D(t)u(t) \quad (3.17)$$

Where $x(t) \in R^n$, $y(t) \in R^q$ and $u(t) \in R^p$.

A and C are dynamic matrices of size $n \times n$ and $q \times n$ respectively. B and D are input matrices of size $n \times p$ and $q \times p$ respectively.

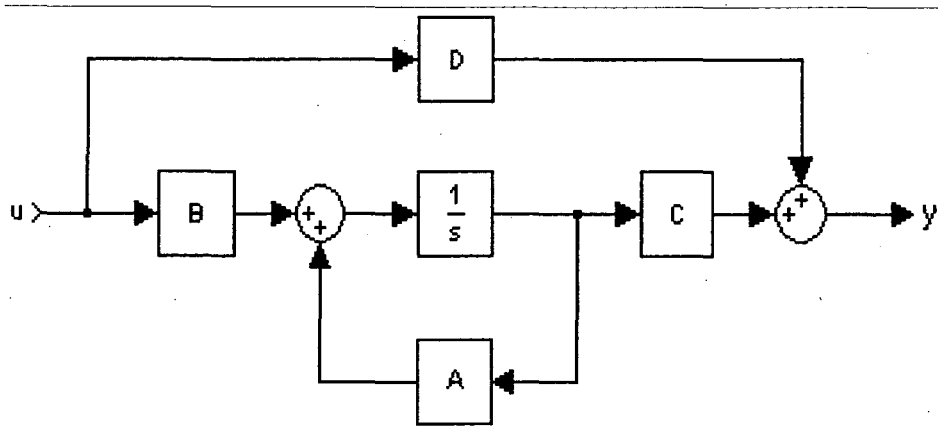


FIGURE 3.6: A typical state space model [36].

In this approach to represent GRN the state represents the concentration of mRNA of a gene. Since there are a large number of genes present in gene regulatory system so this approach take state to mean either a group of genes or certain genetic factor of the system. This is employed for reducing the computational requirement of the model. This method is used in reverse engineering of GRN by many researchers successfully [24].

Chapter 4

Modeling Gene Regulatory Network

In this we discuss two different paradigms for model gene regulatory network employed for our experiment. In these approaches, gene regulatory networks are considered as a distributed, complex and dynamic system. After that we have tried to simulate these models. To consider the GRN as a dynamic system is useful, because it yields the interaction graph between genes and the simulator of the system.

The first technique used in this chapter is to find the jacobian matrix (connectivity matrix) for a gene regulatory network using singular value decomposition methods on the given data [3].

In the second approach we have supposed that the gene regulatory network is a linear dynamic model which behaves like a complex, dynamic system, which behaves as a complex, dynamic system. In the measurement of gene's expressions levels there may be chances of noise in the measurement of gene's expressions data. So the framework of linear Gaussian state space model provides a way to take into account noise both intrinsic and extrinsic noise (noise in the observation and noise in the underlying dynamic process of gene regulation [22]). so we tried linear Gaussian state space model to simulate GRN. we will discuss these two models briefly.

4.1 Inferring Gene regulatory network using singular value decomposition [20]

The singular value decomposition (SVD) is a powerful technique in linear algebra for matrix computation and analysis [27]. This is a factorization technique for a given matrix. It is useful in a lot of applications like signal processing and statistics [27]. Using SVD of a matrix in stead of using original matrix gives more robust result. SVD also gives the geometric structure of the matrix. The spectral theorem states that a simple matrix can be factorized on the basis of eigen vectors. SVD uses eigen vectors to factorize the given matrix.

4.1.1 Definition

Suppose that X is an $m \times n$ matrix, whose all values comes from the set K . elements of K may be real numbers or complex numbers. Then the SVD decomposes this matrix in the form,

$$M = U \Sigma V^T \quad (4.1)$$

Where U is a $m \times m$ matrix over field K . U is also an orthogonal matrix known as left singular matrix. V is an $n \times n$ orthogonal matrix over K . It is known as right singular matrix. V^T is transpose of V . Σ is $m \times n$ diagonal matrix. Such that $\Sigma_{ij} = 0$ if $i \neq j$ and $\Sigma_{ii} = e_i$ if $i = j$, where $e_i \geq 0$.

There may be matrix such that $e_1 \geq e_2 \geq \dots \dots e_N \geq 0$.

These values are called singular values of X . Some times it is also possible that for $m \times n$ matrix X . one can write,

$$X = U \Sigma V^T \quad (4.2)$$

Where U is $m \times p$ matrix. Σ is $p \times p$ and V is $n \times p$ matrix. It is possible in R programming language [27].

4.1.2 Connectivity matrix for a single micro array dataset [3, 20]

This is a method of reverse engineering gene regulatory network from single micro array dataset. In this method, gene regulatory network is supposed to be typically large and sparse. It uses a singular value decomposition to find a set of solutions and after that robust regression to find out the sparsest solution for GRN. This algorithm is of $O(\log n)$ sampling complexity and $O(n^4)$ computational complexity.

Using cDNA and oligonucleotides microarray technology, it is possible to measure mRNA expression levels of thousands of genes at the same time. Since in a gene regulatory network there are thousands of genes presents, so to extract the topology of such network takes much more time. That's why it requires also a large amount of experimental data.

To overcome the problem of data deficiency many research have considered clustering method (grouping genes into hierarchal functional units). There are many attempts to model GRN such as genetic algorithm [20], neural networks [13] and Bayesian network [30]. These models requires large amount of data. To resolve this problem of less availability of data researchers have adopted linear model and used SVD techniques to reverse engineer GRN architecture.

This method involves two steps process [3].

Gene regulatory network can be represented by a system of ordinary linear differential equation [3, 20] is given by,

$$x_i(t) = -\lambda_i x_i(t) + \sum_{j=1}^N W_{ij} x_j(t) + b_i(t) + \xi_i(t), t = 1, \dots, N \quad (4.3)$$

Where x_i 's are mRNA concentration of i th gene at time t . λ_i is self degradation rate.

W_{ij} are real numbers represents the strength of gene i on gene j .

Suppose that using microarray technology we have taken m observation for n different genes, and then we will get the matrix X for concentrations of gene as,

$$X_{n \times m} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^m \\ x_2^1 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^m \end{pmatrix} \quad (4.4)$$

Subscript i indicates individual gene and superscript j indicates observation number. x_i^j is concentration of i th mRNA on j th experiment.

Suppose that there is no noise in the system and if $B = (b_1, b_2, \dots, b_n)$ is a stimulus then we can rewrite the equation (4.1) as,

$$\dot{X}_{n \times m} = A_{n \times n} X_{n \times m} + B_{n \times m} \quad (4.5)$$

Here the self degradation rates λ_i 's are absorbed into coupling constants W_{ij} .

The goal of reverse engineering GRN is to use the measured data B , X and \dot{X} to deduce A . hence the connectivity matrix W . In this context, we may take the transpose of the system and rewrite it as,

$$(X^T)_{m \times n} (A^T)_{n \times n} = [(\dot{X})_{m \times n}^T - (B^T)_{m \times n}] \quad (4.6)$$

Since $m \ll n$, because of high cost of perturbations and measurements. This is an undetermined problem to find A . So we can decompose X^T into,

$$(X^T)_{m \times n} = U_{m \times n} E_{n \times n} (V^T)_{n \times n} \quad (4.7)$$

Where U and V are orthogonal matrices and E is diagonal matrix. Where E is,

$$E = \begin{pmatrix} e_1 & 0 & \cdots & 0 \\ 0 & e_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n \end{pmatrix} \quad (4.8)$$

We may assume that all non-zero elements of E_k are listed at the end i.e., $w_1, w_2, \dots, w_L = 0$ and $w_{L+1}, w_{L+2}, \dots, w_n \neq 0$, where $L := \dim(\ker(X^T))$.

Then one particular solution may be,

$$\hat{J} = (\dot{X} - B) \cdot U \cdot E^{-1} \cdot V^T \quad (4.9)$$

Where $E^{-1} = \text{diag}(1/e_i)$ and if $e_i = 0$ then $1/e_i = 0$.

General solution may be,

$$J = (\dot{X} - B)UE^{-1}V^T + YV^T = \hat{J} + YV^T \quad (4.10)$$

Where Y is $n \times n$ matrix where $y_{ij} = 0$ if $e_j \neq 0$ and is otherwise arbitrary scalar coefficients. Solutions of (4.8) represent all of the possible networks that are consistent with the single microarray dataset.

Suppose that there are N micro array data sets then we can get N networks with N connectivity matrices as,

$$J^k = (\dot{X}_k - B)U_k E_k^{-1}V_k^T = \hat{J}^k + Y^k V_k^T \quad (4.11)$$

Where $k = 1 \dots N$ is the index of dataset k.

Since different dataset may have different qualities, so different weight coefficients will be attached with them. Weight coefficients for all datasets can be calculated as,

$$w^k = \frac{N_k}{\sum_{i=1}^N N_i} \quad (4.12)$$

Where $\sum_{i=1}^N w^k = 1$.

Now we have to find most sparse network from these networks and the algorithm [20] to determine structure of jacobian matrix is given below:

Step-0: Input the Microarray datasets $X_{n \times m}$

Step-1: Obtain J^k from equation (4.7) using SVD technique and w^k from equation (4.10). Initialize the components, $J_{ij} = 0, Y_{ij}^k = 0, J_{ij}^k = J^k, q = 0$. λ and ε are positive values.

Step-2: Fixing J_{ij} solve Y_{ij}^k by,

$$\min_{Y^k} \sum_{i=1}^n \sum_{j=1}^n |J_{ij} - J_{ij}^k| \quad (4.13)$$

Set $q=q+1$.

Step-3: Fixing J_{ij}^k solve J using,

$$\min_J \sum_{k,i,j} [w^k |J_{ij}(q) - J_{ij}^k(q)| + \lambda |J_{ij}(q)|] \quad (4.14)$$

Step-4: check for convergence if $\|J(q) - J(q-1)\| < \varepsilon$ then terminate else go to step-2.

Step-5: Output J.

4.2 Linear Gaussian State Space Models [22]

State space models are commonly used for time series analysis and longitudinal data [4]. The basic model of Gaussian state space model is shown by discrete linear equation with respect to time added with Gaussian noise. Suppose that at each time interval the system produces an output variable y_t (at time t), which is the observation vector. Suppose that the original output vector has been corrupted by some noise. [22]. so the vector x shows the property of hidden markov dynamics. The noise may be either intrinsic or extrinsic [11]. The intrinsic noise may be present in genetic expression and measurement noise can be due to acquisition techniques. So each output vector y is generated from the current state by a simple linear equation. Both the state evolution (x_t) and observation process are corrupted by hidden Gaussian noise.

Suppose that x is a continuous state variable then basic linear equation model can be written as:

$$x_{t+1} = A \cdot x_t + u \quad (4.15)$$

$$y_t = Cx_t + v \quad (4.16)$$

This system is shown in following fig (4.1). Where A is transition matrix and C is projection matrix of size $n \times n$ and $m \times n$ respectively. u is the vector representing the state evolution noise and v is m vector representing observation noise. U and v are independent of each other.

$$u \sim N(0, Q) \quad (4.17)$$

$$v \sim N(0, R) \quad (4.18)$$

Both of these noise sources are temporally white (uncorrelated from time step to time step) and spatially Gaussian distributed with zero mean and covariance matrices which we have denoted by Q and R respectively. Since the state evolution noise and its dynamics are linear so x can be said as first order markov random process.

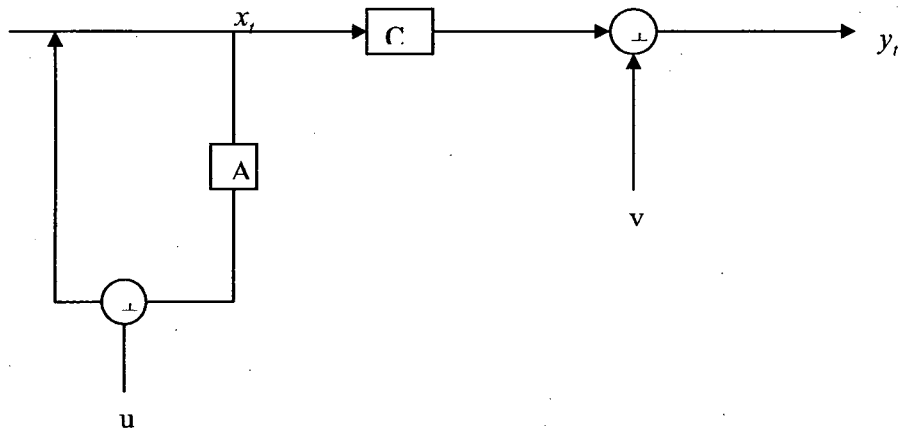


FIGURE 4.1 A simple state space Model [22]

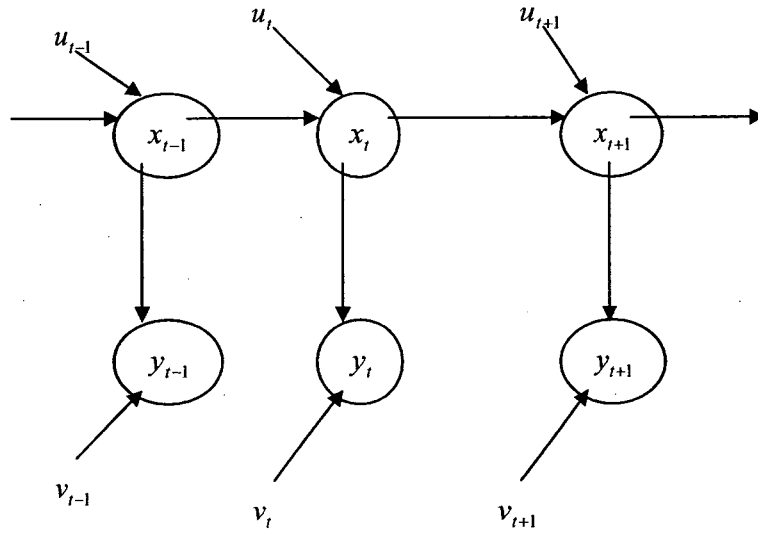


FIGURE 4.2 The network model. [22]

A dynamical Bayesian network explains the relationship of conditional dependencies on time-dependent variables. The Gaussian state space model also belongs to the family of Kalman filter models which was given by Kalman to process signal filtering and smoothing in the years of sixties.

Inference in a state space model includes computing the posterior distributions of the hidden state variables given the sequence of observations. The algorithm for computing the posterior means and covariance involves two steps: a forward pass which uses the observations from y_1 to y_t , known as the Kalman filter, and a backward pass from y_T to y_{t+1} . The combined forward and backward recursions are known as the Kalman or Rauch-Tung-Streibel (RTS) smoother Algorithm [22].

We have used toolbox available in MATLAB for our experiments. After using Kalman filter and smoother algorithms parameters can be learned using a generalized *Expectation-Maximization* (EM) algorithm [36].

Chapter 5

Experimental work and result

5.1 Experimental Objective

The Objective of our experiments is to extract structure of GRN and simulate the expression level of different genes of a given microarray datasets of organism. After that we will try to compare those models in order to determine that which the better approach to model gene regulatory network is. The objective of our experiments is to learn the parameter of the state space model in order to fit the available gene expression data.

5.2 Datasets Used

5.2.1 System Characteristics

All Experiments have performed on the system with the following configuration:

- **Processor:** Intel (R) Pentium (R) 4 CPU 2.80GHz
- **Memory:** 256 MB RAM
- **Windows Dir:** C:\WINDOWS
- **Machine name:** SCSS108_14
- **Operating System:** Microsoft Windows XP Professional Version 2002 Service Pack 2
- **Language:** English
- **System Manufacturer:** Acer Power
- **BIOS:** Default System BIOS

5.2.2 Used Tools

In this experimental work MATLAB is used as a tool. This tool is used to implement the algorithm as well as computing parts of both the experiments. MATLAB is a high performance language for technical computation. It integrates computation, visualization and programming in an easy to use environment where problems and solutions are expressed in familiar mathematical notations [23].

5.2.3 Databases

In our experiment, we have considered the database for S.O.S DNA repair network of the Escherichia coli bacterium. In second experiment from this data we have tried to simulate the model of gene regulatory network. The experimental data has been taken from the homepage of Uri Alon [26]. The experimental data are expressions levels of the main 8 genes of the S.O.S DNA repair network of E.Coli. The measurement technology used to measure the gene expression levels is property of GFP (Green Fluorescent Proteins) [11]. Measurement has done after irradiation of the cells at the initial time with UV light. Four experiments have done for various light intensities and each experiment has 50 time points spaced by 6 minutes for 8 genes. These 8 genes are *uvrD*, *lexA*, *umuD*, *recA*, *uvrA*, *uvrY*, *ruvA* and *polB* [11].

Escherichia Coli (E.Coli) is a type of bacteria, which normally lives, in the lower intestines of mammals known as gut flora. A German bacteriologist Theoder Escherich discovers it in 1885. E.Coli. is one of the most thoroughly discussed organisms. Normally, E.Coli does not cause disease although some strains frequently cause diarrhea in travelers and it may be the cause for urinary tract infections.

5.2.4 S.O.S DNA Repair Network

This genetic network consists of more than 30 genes in escherichia coli that carry out diverse functions in response to DNA damage. The S.O.S DNA repair network repairs the DNA after damage, which may be due to bacteria. Usually when no DNA damage occurs there is a master transcription factor *lexA* present in the cell, binds sites in the promoter region of these genes represses all genes of the network. Protein

recA act as a sensor of DNA damage: After binding to single strand DNA it activates and performs lexA destruction. The decrease in lexA expression level causes the de-repression (i.e. activation) of S.O.S genes. Once DNA damage is repaired the expression level of recA decreases, lexA activated and it inhibit the expression level of S.O.S genes and cells return to their initial state [11].

5.3 Experimental Results:

Before performing our experiment, we have normalized the given experimental data. Normalization is carried out using Z-score, which is given by [37],

$$x_i = \frac{x_i - \mu}{\sigma}, i = 1, 2, \dots, N \quad (5.1)$$

Where x_i is the expression level of gene at ith experiment. N is the total number of experiments. μ and σ are mean and standard deviations respectively and defined as,

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (5.2)$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=0}^N (x_i - \mu)^2} \quad (5.3)$$

We have done two experiments for 50 time points of 8 genes. The first experiment was based on Guassian state space model and second was to infer gene regulatory network using singular value decomposition technique.

Based on these two experiments we have computed the 50 expression levels of 8 genes. Then we compared the measured values with the expression levels for both the experiments.

The comparison graph between measured values and simulated values with both experiments are shown in figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8 respectively.

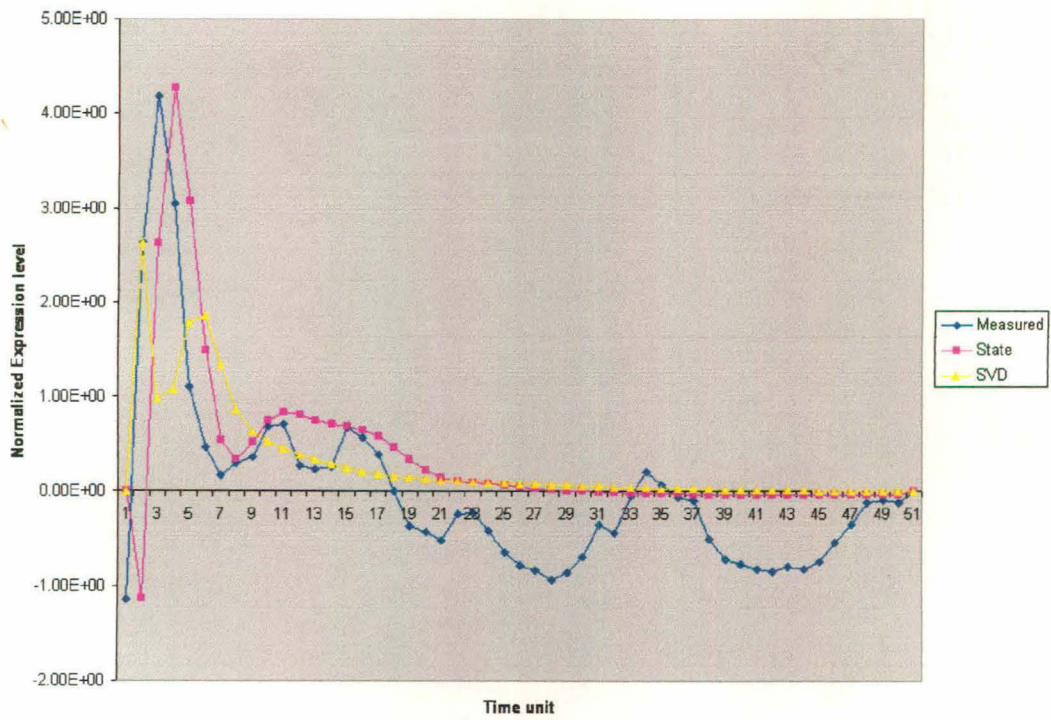


FIGURE 5.1 Variation of Expression level with time for *uvrD* gene

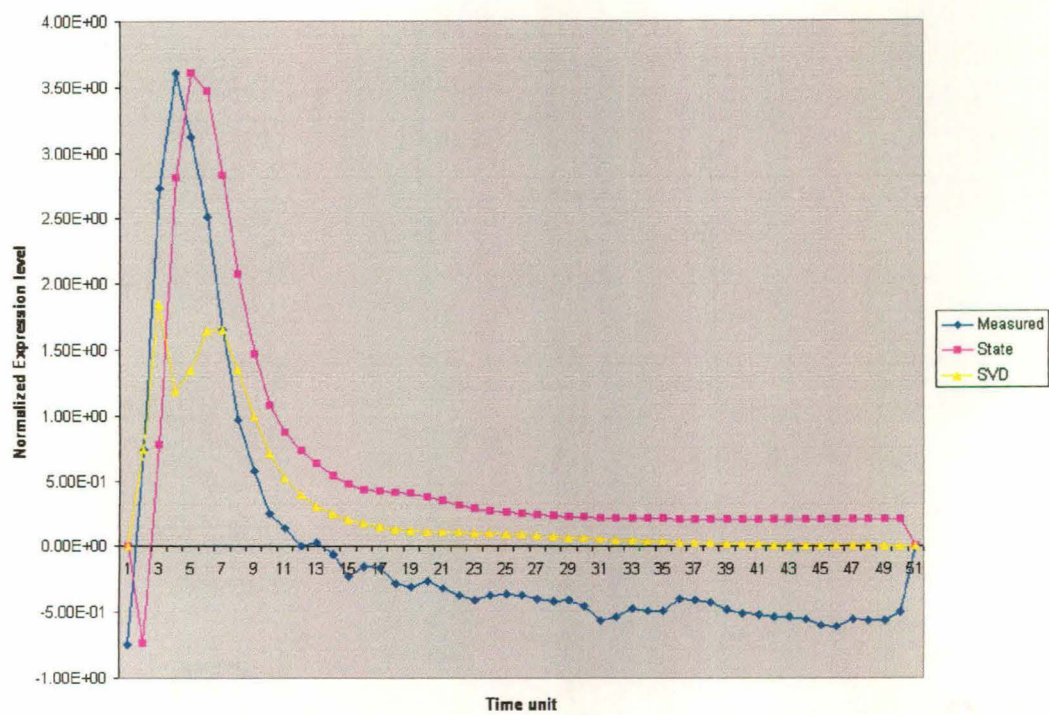


FIGURE 5.2 Variation of Expression level with time for gene lexA

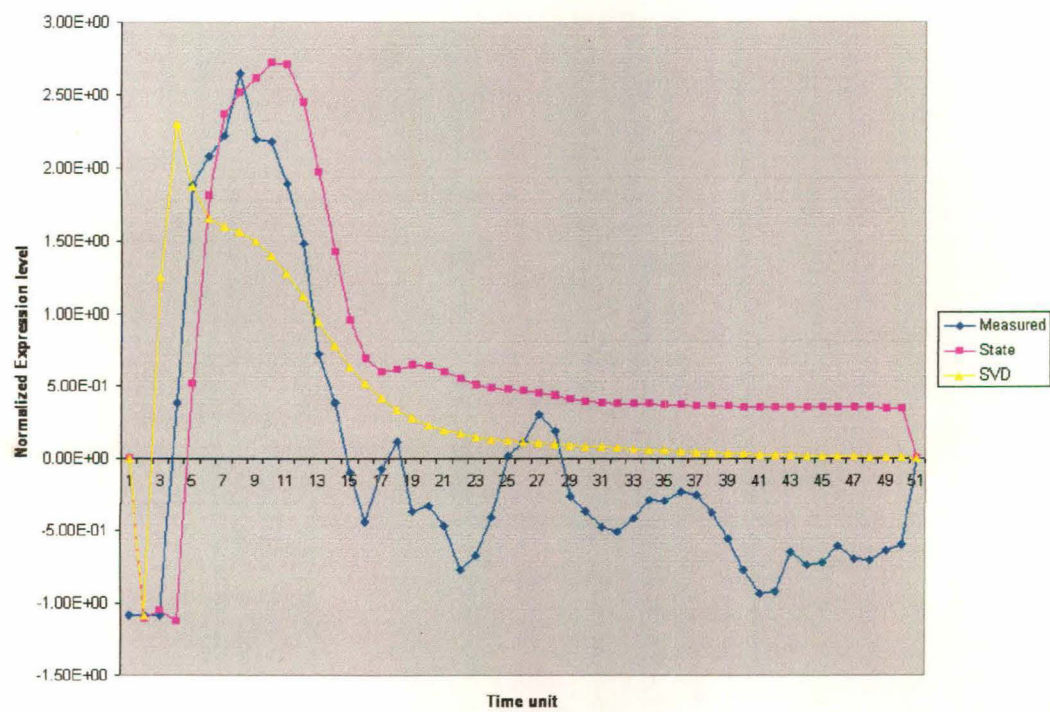


FIGURE 5.3 Variation of Expression level with time for gene umuD

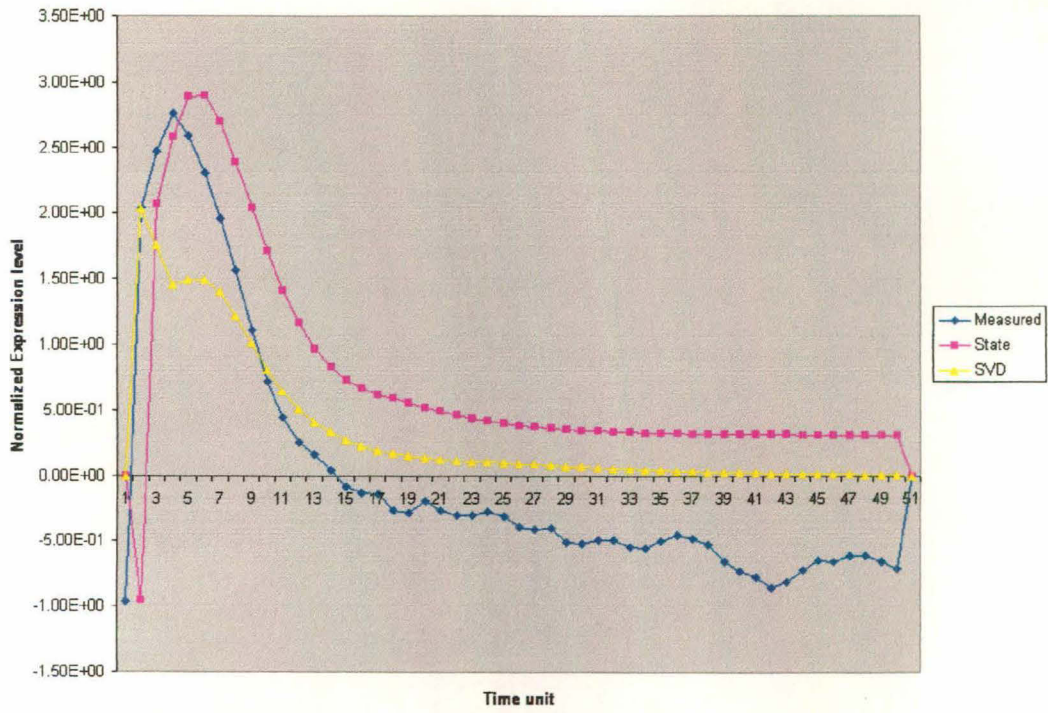


FIGURE 5.4 Variation of Expression level with time for gene recA

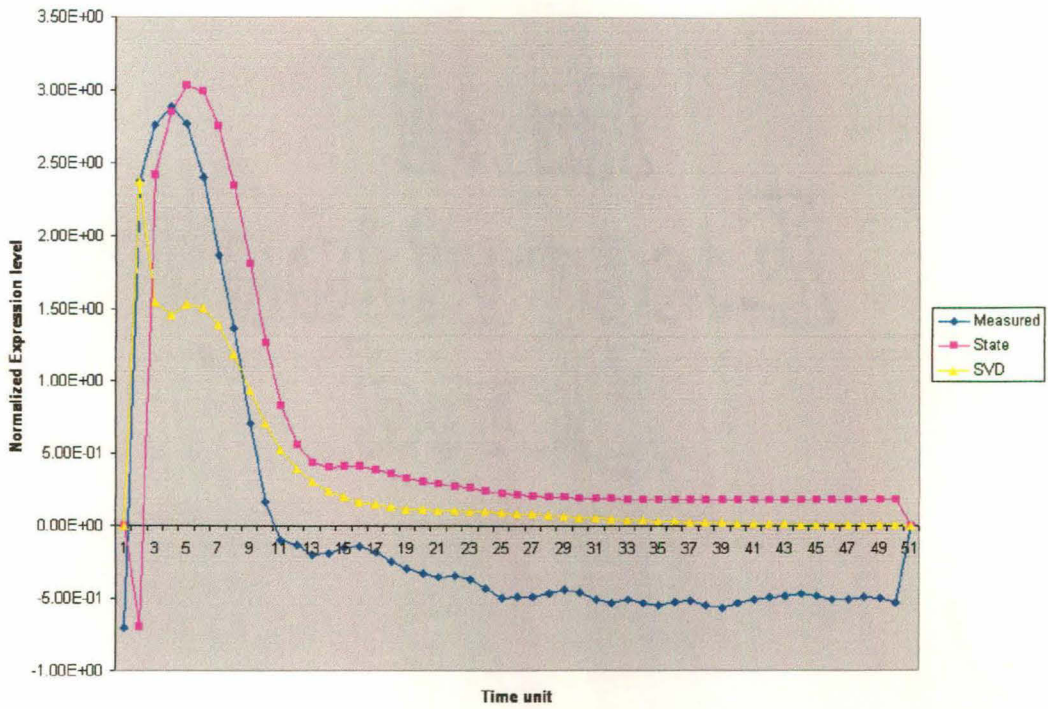


FIGURE 5.5 Variation of Expression level with time for gene uvrA

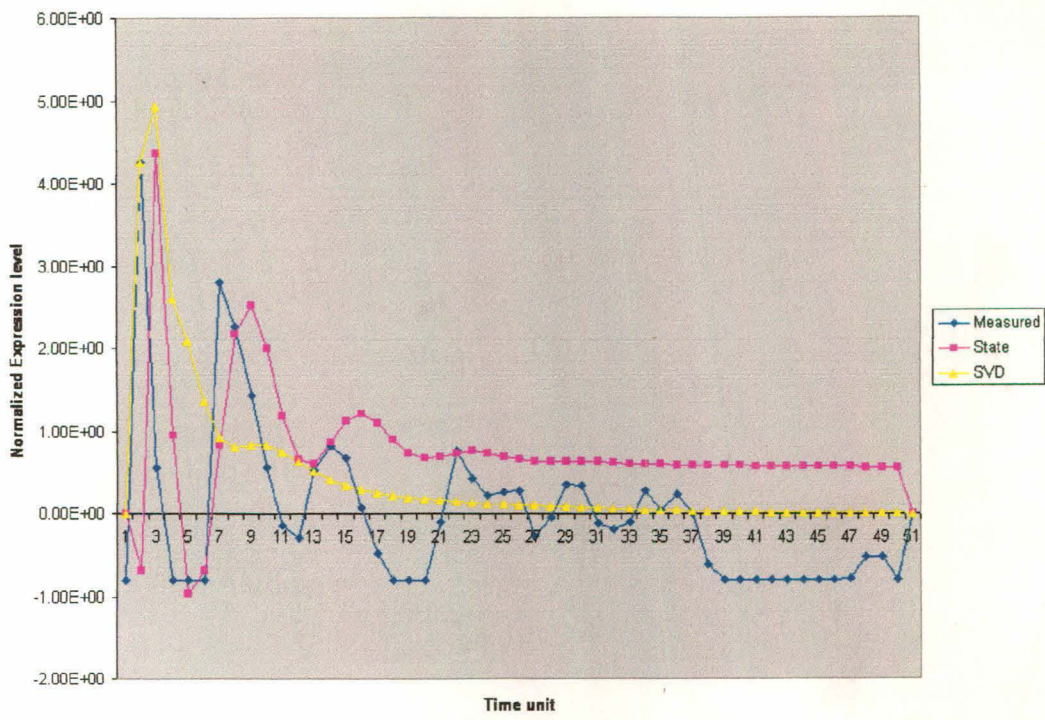


FIGURE 5.6 Variation of Expression level with time for gene *uvrY*

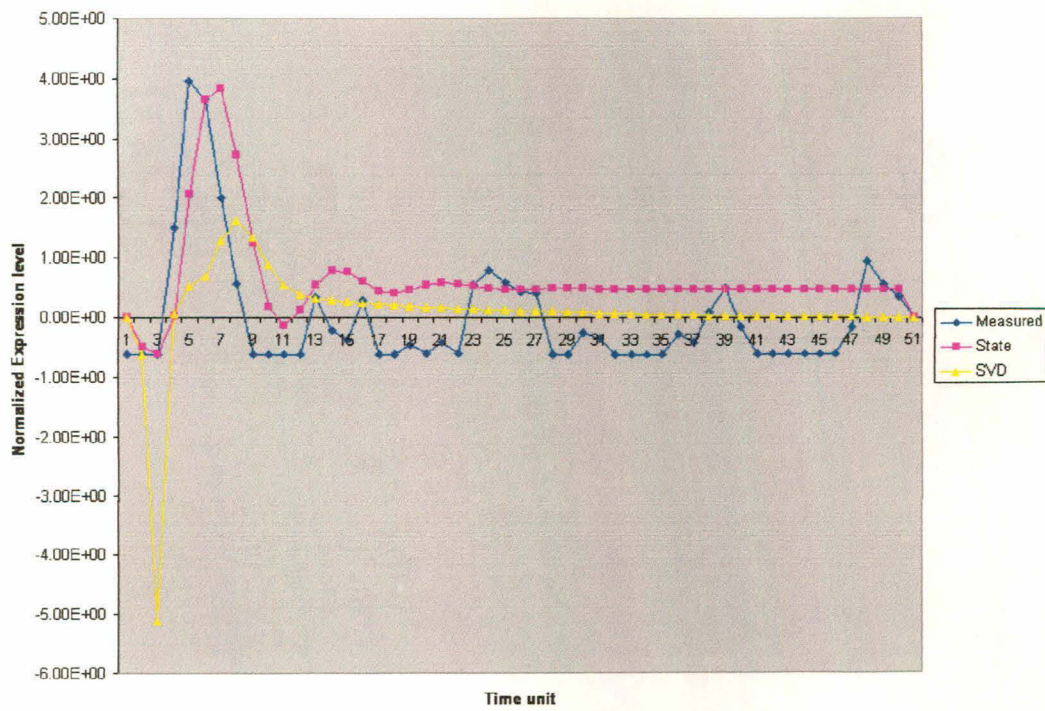


FIGURE 5.7 Variation of Expression level with time for gene *ruvA*

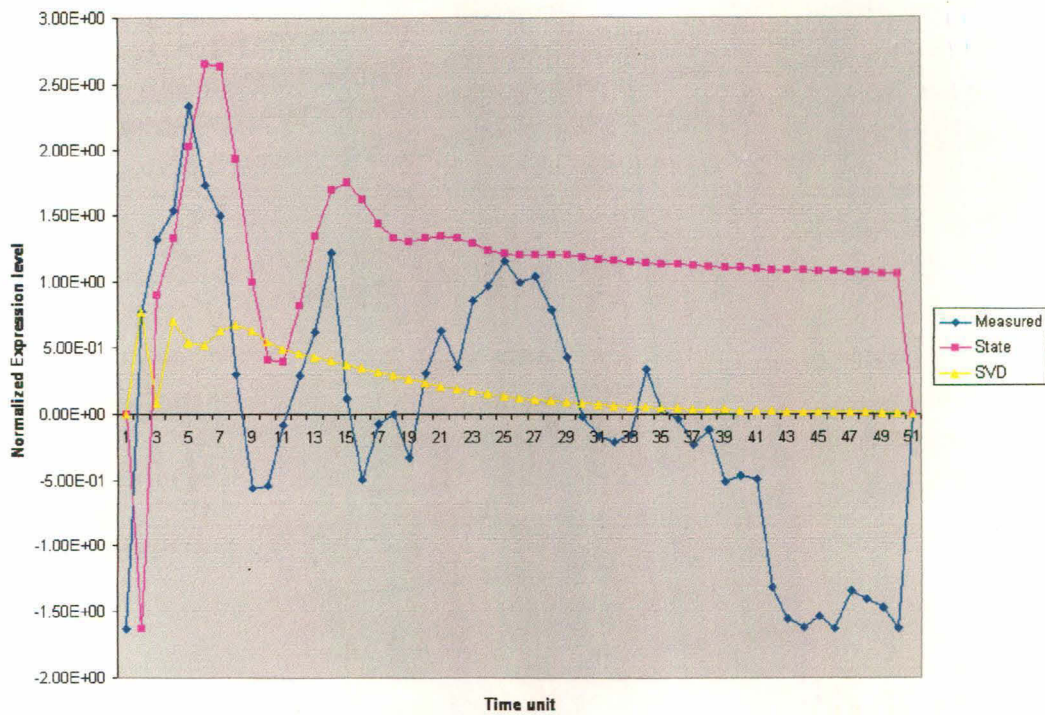


FIGURE 5.8 Variation of Expression level with time for gene polB

The following can be observed from figure 5.1 to figure 5.8 the following:

- Generally at initial stage, the difference between the measured and simulated expression levels is small while the difference between the two is prominent at later time units in case of state space model.
- Generally for all genes the difference between the measured and simulated expressions values is significant at initial time units.
- The simulated expression levels by both methods are different from the measured expression levels. However the difference between measured and simulated is small in case of SVD technique in comparison to using state space model technique.

Conclusion

In this thesis, we investigated different modeling paradigms used for biological networks and to understand their functional relations within. There are many limitations one encounters while modeling gene regulatory network. Few of these limitations are the following:

- a. The number of genes is very large compare to the number of measured time points.
- b. The data contains substantial amount of measurement noise.
- c. The goal of genetic network modeling is to extract the genetic interactions rather than to accurately predict the gene expression levels.
- d. No ground truth is known with respect to the outcome of genetic network models.
- e. No modeling techniques are correctly representing the genetic network models.

We have carried out simulation to determine the observation level of genes using two different modeling paradigms i.e. using singular value decomposition techniques and using Gaussian state space model technique. We have shown the results of two experiments. From experimental results we conclude the following:

- The technique used by state space model gives better result in comparison to SVD techniques at initial stages of time points.
- The simulated expression levels by both methods are different from the measured expression levels. However the difference between measured and simulated is small in case of SVD technique in comparison to using state space model technique.

This result can be improved by using hybrid approach, where initial structure extracted by SVD is considered state space model to obtain the structure of GRN. The

use of biological knowledge can also help to extract more information from microarray data. Hence the complexity to determine network parameter will also reduce with the prior biological knowledge.

References

- [1] G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, The segment polarity network is a robust developmental module. *Nature*, 406:188–92, 2000.
- [2] Florence d’Alch’e-Buc and Vincent Schachter, Modeling and identification of biological networks.
- [3] M. K. Stephen Yeung, Jesper Tegnér, and James J. Collins, Reverse engineering gene networks using singular value decomposition and robust regression, *PNAS*, April 30, 2002, vol. 99, no. 9, 6163–6168.
- [4] Giuseppe De Nicolao, Senior Member, IEEE, and Giancarlo Ferrari-Trecate, Member, IEEE, Regularization Networks: Fast Weight Calculation via Kalman Filtering, *IEEE Transactions on Neural Networks*, VOL. 12, NO.2, March 2001, 1045–9227.
- [5] H. Jeong, B. Tombor, R. Albert, and Z.N. Oltvai et al, The large-scale organization of metabolic networks. *Nature*, 407:651–4, 2000.
- [6] Luonan Chen, Senior Member, IEEE, and Ruiqi Wang, Designing Gene Regulatory Networks With Specified Functions, *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS*, VOL. 53, and NO.11, NOVEMBER 2006, 2444-2450.
- [7] Xutao, Deng and Hesham Ali, A Computational Approach to Reconstructing Gene Regulatory Networks, *Proceedings of the Computational Systems Bioinformatics, (CSB’03)*, 2003, IEEE.
- [8] Christoforos Anagnostopoulos, Matthew C Turnbull, A Note on Learning Linear Gaussian State-Space Models via Expectation-Maximization, March 30, 2007.
- [9] M Madan Babu1, Nicholas M Luscombe, L Aravind, Mark Gerstein and Sarah A Teichmann, Structure and evolution of transcriptional regulatory networks, *Current Opinion in Structural Biology* 2004, 14:283–291.

- [10] Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Range, and David L. Wild, A Bayesian approach to reconstructing genetic regulatory networks with hidden factors, Vol. 21, no. 3, 2005, pages 349–356 doi:10.1093/bioinformatics/bti014.
- [11] Florence d'AlcheBuc, Pierre-Jean Lahaye, Bruno-Edouard Perrin, Liva Ralaivola, Todor Vujasinovic, Aurelien Mazurie and Samuele Bottani, A dynamic model of gene regulatory networks based on inertia principle.
- [12] Vladimir Filkov, Identifying Gene Regulatory Networks from Gene Expression Data, 27.
- [13] Edwards, Chaos in neural and gene networks with hard switching. *Differential Equations and Dynamical Systems*, 9:187–220, 2001.
- [14] Xutao, Deng and Hesham Ali, A Computational Approach to Reconstructing Gene Regulatory Networks, *Proceedings of the Computational Systems Bioinformatics, (CSB'03)*, 2003, IEEE.
- [15] Michiel J.L. de Hoon, Sascha Ott, Seiya Imoto, Satoru Miyano, Validation of gene regulatory network models inferred from time-course gene expression data at arbitrary time intervals.
- [16] Patrik D'haeseleer, Reconstructing Gene Networks from Large Scale Gene Expression Data, December, 2000.
- [17] Hasty, J. Pradines, M. Dolnik, and J. Collins, Noise-based switches and amplifiers for gene expression. *Proceedings of the National Academy of Sciences*, 97(4):2075–2080, 2000.
- [18] HIDDE DE JONG, Modeling and Simulation of Genetic Regulatory Systems: A Literature Review, *JOURNAL OF COMPUTATIONAL BIOLOGY* Volume 9, Number 1, 2002 Mary Ann Liebert, Inc. Pp. 67–103.
- [19] Alexander J. Hartemink, David K. Giord, Tommi S. Jaakkola, Richard A. Young, Elucidating Genetic Regulatory Networks Using Graphical Models and Genomic expression Data, *IEEE Intelligent Systems*, Special Issue on Intelligent Systems in Biology.

- [20] Yong Wang, Trupti Joshi, Xiang-Sun Zhang, Dong Xu, and Luonan Chen, Inferring gene regulatory networks from multiple microarray datasets, *BIOINFORMATICS*, Vol. 22 no. 19 2006, pages 2413–2420.
- [21] Jinshan Li, Xiang-Sun Zhang, An Optimization Model for Achieving Sparsity of Gene Regulatory Networks, *The Sixth International Symposium on Operations Research and Its Applications (ISORA'06)* Xinjiang, China, August 8–12, 2006, pp. 368–379.
- [22] Sam Roweis, Zoubin Ghahramani, A Unifying Review of Linear Gaussian Models, October 1998 In Press *Neural Computation*, Vol. 11 No. 2, 1999.
- [23] Kermit Sigmon, *MATLAB Primer Third Edition*, 1989, 1992, 1993.
- [24] C. Rangel, J. Angus, Z. Ghahramani, and D. Wild, Modeling genetic regulatory networks using gene expression profiling and state space models, In D. Husmeier, S. Roberts, and R. Dybowski, editors, *Probabilistic Modeling in Bioinformatics and Medical Informatics*, pages 269–293, 2004b.
- [25] McAdams and A. Arkin. Stochastic mechanisms in gene expression, *Proceedings of the National Academy of Sciences*, 94:814–819, 1997.
- [26] <http://www.weizmann.ac.il/mcb/UriAlon/>
- [27] Sonia Leach, *Singular Value Decomposition- A Primer*, Department of Computer Science, Brown University, Providence, RI02912.
- [28] P. Dwight Kuo, *Topology and Dynamics of an Artificial Genetic Regulatory Network Model*, Department of Computer Science, Memorial University of Newfoundland, July 2005.
- [29] Akutsu, S. Miyano, and S. Kuhara, Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, pages 8–14, ACM Press, 2000b, ISBN 1-58113-186-0.
- [30] Xiaobo Zhou, Xiaodong Wang, Ranadip Pal, Ivan Ivanov, Michael Bittner and Edward R. Dougherty, A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks, *Bioinformatics*, Vol. 20 no. 17 2004, pages 2918–2927.

- [31] T K Attwood and D J Parry-Smith, Introduction to Bioinformatics, PEARSON Education, 1999.
- [32] James M. Bower and Hamid Bolouri, Computational Modeling of genetic and Biochemical Networks, Ane Books, Massachusetts Institute of Technology, 2004.
- [33] M. Thattai and A. van Oudenaarden, Intrinsic noise in gene regulatory networks, Proc. Natl. Acad. Sci. USA, 98(15):8614–8619, 2001.
- [34] Seiya imoto, Tomoyuki Higuchi, Takao Gopto, Kousuke Tashiro, Saturo Kuhara and satoru Miyano, Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks, Proceedings of the Computational System Bioinformatics, (CSB'03), IEEE, 2003.
- [35] G.D.Bader, I.Donaldson, C.Wolting, B.F.F Ouellette, T.Pawson, and C.W.V.Hogue, BIND-The biomolecular interaction network database. Nucleic acid research, 29,242-245, 2001
- [36] Sean Borman, The Expectation Maximization Algorithm: A short tutorial, July 18, 2004.
- [37] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, ELSEVIER, 2006.
- [38] T.Ideker, O.Ozier, B.Schwikowski and A.F. Seigel, Discovering regulatory and signalling circuits in molecular interaction networks, Bioinformatics, 18 (ISMB, 2002), S233-S240, 2002.
- [39] Andres Kriete and Roland Eils, Computational Systems Biology, Elsevier Academic Press, 2006.
- [40] T.Speed, editor. Statistical Analysis of Gene Expression Microarray Data, CRC Press, 2003.
- [41] Weaver, C. Workman, and G. Stormo, Modeling regulatory networks with weight matrices, In Pacific Symposium on Biocomputing, 1999.
- [42] Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators, Nature, 403(6767):335–338, 2000.
- [43]http://images.google.co.in/imgres?imgurl=http://plasticdog.cheme.columbia.edu/undergraduate_research/projects/sahil_mehta_project/images/Affymetrixarray.jpg&imgrefurl=http://plasticdog.cheme.columbia.edu/undergraduate_research/projects/sahi

l_mehta_project/work.htm&h=600&w=540&sz=61&hl=en&start=1&tbnid=5sJHfat
dqzSivM:&tbnh=135&tbnw=122&prev=/images%3Fq%3DAffymetrix%2BGeneCh
p%2BArray%2B%26gbv%3D2%26svnum%3D10%26hl%3Den%26sa%3DG

[44] <http://www.eas.slu.edu/People/DJCrossley/uniquearth/images/dna.ht1.jpg>

[45] <http://www.accessexcellence.org/RC/VL/GG/images/gene2.gif>

[46] http://en.wikipedia.org/wiki/Petri_net

[47] http://en.wikipedia.org/wiki/Gene_expression