CHAPTER 1

INTRODUCTION

Over the past decades, the availability of complete genome sequence has facilitated the development of high-throughput technologies, such as microarrays that can probe cells on a genome-wide scale. The emergence of these technologies has changed the way in which biological research is done. In the past, biologists engaged themselves in years of research and conducted different expensive experiments to uncover a few details about the cellular processes and how molecular units of living cells interact throughout this process. Today, scientists can monitor the activity levels for thousands of genes or proteins in parallel to study the effects of certain treatments, diseases and developmental stages at the molecular level. Following this abundance of data, what still need to be developed are the computational methods that improve the state of the art to better understand the biological systems which have generated these data.

In a living cell, the complex interactions among the various molecular units such as DNA, RNA, proteins and small molecules determine its structure and function. Therefore, a key goal of post-genomic research is to understand the structure and dynamics of the complex cellular interactions. There are several different kinds of interaction networks that can be distinguished at the molecular level within a cell. These are: metabolic networks, signaling networks, protein-protein interaction (PPI) networks and gene regulatory networks (GRN). Metabolic networks comprise the chemical reactions of metabolism that determines the physiological and biochemical properties of a cell. Signaling networks are the communication

channels of a cell to the outside world. PPI networks comprise both physical and genetic interactions among the proteins and play a central role in the study of the interactions among the molecules within a living cell. The GRN consists of interactions between transcription factors and target genes and control the expression of genes in a cell.

The major goal of this thesis is to develop a computational model that utilizes domain knowledge and other sources of biological data in estimating GRN from time series gene expression data. Gaining insight into such a regulatory network will assist researchers to understand a cell's function (the sequence of interactions in different developmental processes) as well as how the cell dysfunctions in the case of diseases, such as cancer. However, the most important application of such discovery is in drug design, as pathological effects of a drug can only be addressed with precise knowledge of what functions and dysfunctions really are.

The reconstruction of GRN involves identifying regulatory connections; that is, which regulators control which gene and how. Thus, computational methods can be utilized to learn both the structure and parameters of the network from experimental data. A general framework of the reconstruction process is shown in Figure 1.1. The whole process of the reconstruction of GRN can be divided into several interactive tasks. The availability of complete genome sequence has facilitated the development of high-throughput technologies which generates large quantities of genomic data. The generation and accessibility of these data have drawn attention on designing appropriate computational models to decipher the structure of GRN and understand the underlying cellular process. Though the reconstruction task appears simple and straight-forward in the figure, it faces several key issues which are elaborated in the subsequent sections. Some of these issues are addressed in this thesis and others have been left as future research directions.

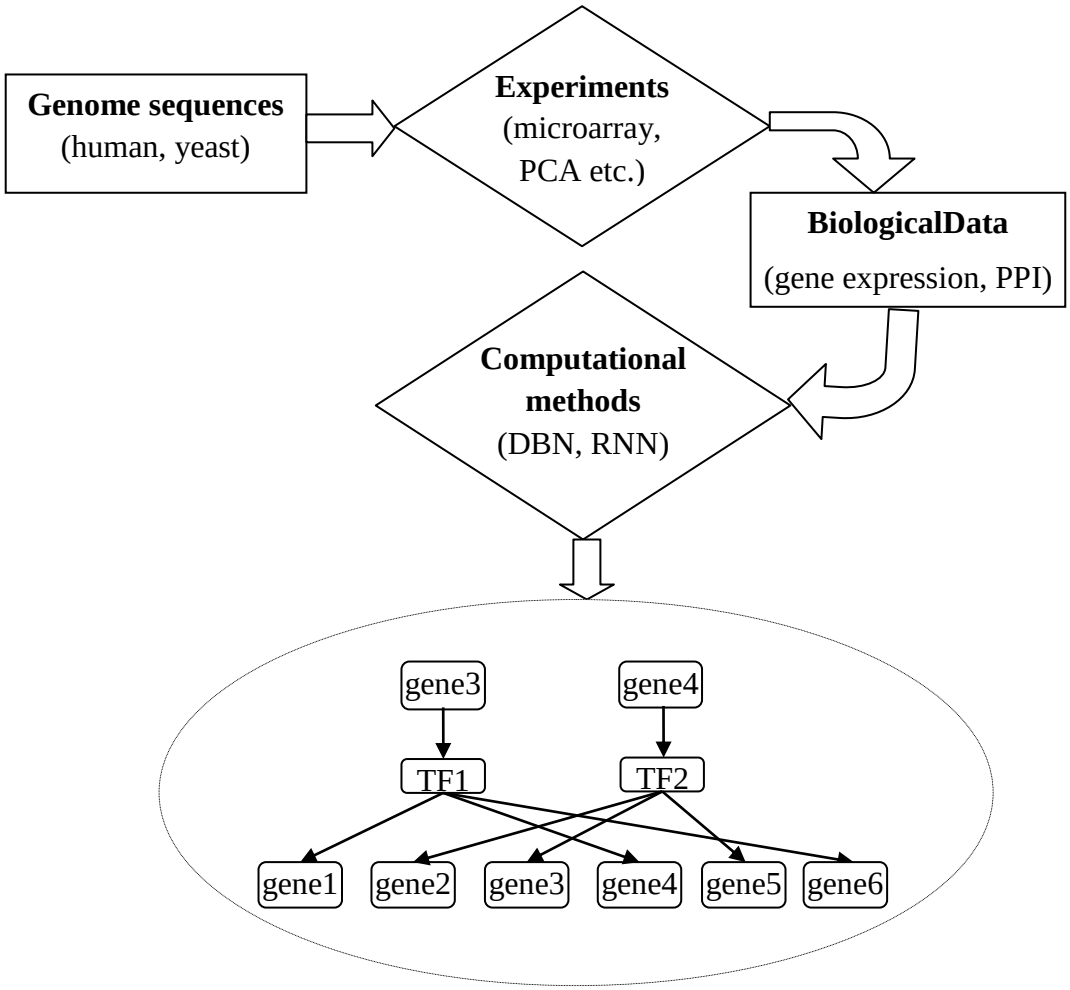


Figure 1.1: Illustration of the process of reconstructing GRN from experimental data. In this example network, each node corresponds to a gene and an edge represents the direct regulation among the genes.

1.1 Issues in Reconstructing GRN from Experimental Data

From a computational perspective, the reconstruction of a GRN from the massive high-throughput data can be considered as a large optimization problem. Such consideration derives from the fact that the computational models either optimize a scoring function over a large space or a large set of parameters. For instance, differential equations based GRN models seek to optimize a large number of parameters from the data. The other genre of models known as graphical models such as Dynamic Bayesian Network (DBN), Bayesian

Network optimize a scoring function to learn the structure of the network given the data. Therefore, the task of reconstructing GRN face a wide range of challenges, some of which originate from the perspective of the computational model being used, some from the characteristic of available data, and the others because of the lack of knowledge about the biological system are being modeled. Below we list some of the key issues that limit the effective reconstruction of GRN with the current computational methods:

1. **High dimensionality:** Dimensionality problem is one of the major challenges that researchers have come across in reconstructing GRN from experimental data. To illustrate the dimensionality problem, we consider budding yeast which is the simplest unicellular organism and has been considered as the model for studying cellular processes in eukaryotic cells. Even the yeast genome consists of more than 6400 genes, whereas the size of the human genome exceeds 40000. This numbers themselves articulate the size and intensity of the problem. Most of the computational methods learn the structure and the parameters of the network from the available data. Some models such as differential equations based models (Di Bernardo et al. 2004, Kimura et al. 2005, Kikuchi et al. 2003, Cinquemani et al. 2008) estimate the network parameters from the available data. For these models the number of parameters to be estimated grows exponentially with the number of genes in the data. Therefore, the problem of GRN construction becomes intractable with just tens of genes in the dataset whereas the models are expected to learn networks of thousands of genes. Graphical models such as Bayesian Networks (Friedman et al. 2000), DBNs (Murphy and Mian 1999, Zou and Conzen 2005) search for the best structure by optimizing a scoring function, given the data and the network parameters. An exhaustive search for the best structure considers 2^N possible set of parents for each target gene, where N is the number of genes in the data. This exponential growth of the number

of potential structures makes them intractable for analyzing networks with more than tens of genes.

2. **Low number of true positives in the reconstructed network:** The Majority of the available GRN models identify a very low number of true positives (a connection that exists in both the known and reconstructed network) in respect to the total number of known and estimated connections. This is especially true in those cases where the models have been tested on a relatively large datasets including hundreds of genes. Researchers have identified many causes that underlie such inaccurate estimation. Some of them are as follows: 1) imprecision in measuring the level of gene expression, 2) model robustness and compatibility, 3) type of data used such as discrete or continuous data, 4) inherent noise in the data and 5) lack of knowledge (very small number of regulator-gene relationships are known till date). To exemplify some of these causes, we focus on the effect of data type on the model. Some GRN models such as Boolean Networks (Akutsu et al. 1999) are discrete models; that is, they discretize the gene expression data in a preprocessing step and apply the model on the discretized data. This discretization of data causes information loss and the models are unable to identify even a reasonable number of true positive connections. Conversely, the continuous models (Recurrent Neural Networks, S-System model) are expected to perform better at the expense of high computational time. However, these models often suffer because they are not capable of distinguishing the inherent noise in the data from the actual gene expression.
3. **Data scarcity**: At present, data generated by the high-throughput experiments contain thousands of genes over only a few time points. For instance, the benchmark data of the yeast cell cycle (Spellman et al. 1998) contains the expression levels of thousands of genes over 18 time points (alpha synchronization). This insufficient data causes another

major challenge for the computational reconstruction of GRNs. Akutsu et al. (1999) discussed the amount of data required for learning the structure of a deterministic Boolean network from a computational learning theory perspective. They showed that the number of time points needed for estimating a Boolean network of N genes is lower bounded by $\Omega(2^k + klogN)$ and upper bounded by $O(2^{2k}\ 2k\ log\ N)$, where k is the fan-in of each node. Since the constant factor is exponential in k, genes with a high number of regulators would demand much more time points than what we have at present. Thus, data scarcity has a direct effect on the number of true positive connections that a model can estimate.

4. **Imprecise and incomplete data:** One pitfall of analyzing high-throughput data is that they come with an enormous amount of experimental noise (Aris et al. 2004). A range of factors can contribute to the generation of this experimental noise such as imprecision in designing the array, sample preparation, and the hybridization process. As a consequence of this inadequacy in the experimental system, the regulatory process becomes partially observable in most of the cases leaving missing values in the collected data. Therefore, the inherent noise and the incompleteness of data pose another critical challenge on the reconstruction of GRN as they provide misinformation to the computational method.
5. **Validation of the reconstructed network:** In general, the GRN models characterize quantitative knowledge concerning gene regulation which is hidden in the underlying data. However, knowledge about the underlying biological structures from which these data originate is often incomplete or unavailable. This lack of knowledge poses another critical issue which concerns the validation of the network models. Much effort has been given to address the problem of network validation. In recent years, the simulated data from DREAM (Dialogue for Reverse Engineering Assessments and Methods) project (Prill et al. 2010) are gaining interest as benchmark data for validating GRN models.

Despite the wide acceptance of simulated data, biologists are still interested in constructing GRNs from experimental data and finding a way of validating them.

We address some of the above mentioned issues in this thesis and propose ways to possibly address others in the future work.

1.2 Thesis Motivation

Modern biomedical research focuses on the understanding of the structure and functions of GRNs with the goal of identifying genetic causes of diseases and developing drugs to cure them. As discussed in the previous section, the high dimensionality and data scarcity are the two fundamental challenges associated with the computational reconstruction of GRN. The first challenge arises from the complex nature of gene regulation and contributes towards the high computation time of estimating the structure and parameters of the network. The second challenge limits the computational model from learning the network with a higher precision. One possible way of dealing with the high computational time of GRN reconstruction is to reduce the problem size by using some kind of heuristics. In case of structure learning, this reduction implies restricting the number of possible parents for a target gene. The utilization of heuristics is expected to reduce the problem dimension, however it may adversely affect the precision of the estimated network when the choice of heuristics is arbitrary and lacks biological foundation.

The negative consequence of applying such arbitrary heuristics motivates us to use domain knowledge in reducing the problem size. In this thesis, we utilize biological knowledge to restrict the number of possible regulators for each target gene. This reduction in the problem size also lessens the data requirements of the computational methods. Therefore, we put forward the theme that the exploitation of such knowledge would address both the

computational time and the data scarcity problem associated with GRN reconstruction. The objectives of the work in this thesis are:

- identify key features used by transcriptional regulators of a gene and employ that knowledge to reduce the problem size.
- improve the precision of the reconstructed network; that is, increase the number of true positives and reduce the number of false positive connections.
- apply a structure learning algorithm that has the ability to model stochasticity and provides flexibility to incorporate prior knowledge. In particular, we prefer probabilistic models that show robustness to inherent noise and can handle missing values.
- examine the effectiveness of incorporating biological domain knowledge through the analysis of both simulated and experimental data.
- investigate the scalability of the computational methods ; that is the methods maintain or even increase the performance of the estimated networks when applied on larger data.

1.3 Thesis Statement

In section 1.1, we have discussed the key challenges of reconstructing GRNs with available computational methods. In particular, our interest lies in dealing with the first two challenges which are high dimensionality and low accuracy of the reconstructed network. Therefore, this thesis aims to extend and improve the current GRN models by reducing the problem space with the incorporation of biological domain knowledge and other sources of data. This thesis also compares and quantifies to what extent the reconstruction accuracy as well as the computation time can be improved with the following hypothesis:

- exploitation of biological domain knowledge reduces the problem size and makes the reconstruction task computationally feasible.
- models that employ the key features that are used by transcriptional regulators of genes, estimate networks with a higher precision.
- integration of multiples sources of biological data to restrict the number of potential regulators for a target gene reduces both the problem dimension and the data requirements of the model; hence increasing the number of true positives in the estimated network.
- models that utilize biological domain knowledge are expected to be scalable; that is they maintain steady precision level when applied on the analysis of large datasets.

1.4 Contributions

This thesis focuses on the computational reconstruction of GRNs from microarray gene expression data by exploiting biological domain knowledge. Among other available methods, we choose DBNs as it provides a good compromise between the network relevancy and inherent noise in the data. Below, we summarize the major contributions of this thesis in addressing some of the key challenges discussed in section 1.1.

1. **Exploitation of the domain knowledge of the cellular process under study:** To address the dimensionality problem of high-throughput data, we utilize the biological features of the cellular process under study which is the yeast cell cycle in this thesis. One such feature is that, a high proportion of cell cycle regulated (CCR) genes are periodically expressed; that is genes are maximally expressed to affect and control the regulation of other genes and on completing their designated task they are repressed by some other regulator genes. Thus the whole cell cycle progresses systematically through the

successive activation and repression of CCR genes. To use this feature, we have calculated the peak time of individual genes which falls into one/more phases of the cell cycle. Therefore, genes that peak in the interval of the same phase of the cell cycle have been grouped together. In addition, each group includes known transcription factors of the previous phase with the hypothesis that they regulate some genes which facilitate the systematic initiation of the current phase. This inclusion creates clusters with overlapping genes. Finally, we employ a DBN structure learning algorithm on each individual cluster to estimate the desired network. The exploitation of the knowledge in relation to the phase-specific regulation of the cell cycle has decomposed the whole task of GRN reconstruction into several sub-problems. As a consequence of this decomposition, the computational time of the model (DBN in this thesis) has been reduced significantly. Although the model shows slight increase in the number of true positives, the precision of the estimated network is similar to that of randomly generated networks when applied on the large datasets.

2. **Application of a partitioning algorithm to find groups of co-expressed genes and their co-regulators:** In real biological networks, genes are both co-expressed and co-regulated. This implies that genes with similar temporal and spatial expression patterns are believed to be governed by a common regulatory logic. Therefore, it is equally important to find causal relationships between groups of co-expressed genes. In order to find groups of co-expressed genes, we apply a data partitioning algorithm, known as Partitioning Around the Mediods (PAM), which divides n objects (genes) into k clusters, where k is the optimal number of clusters in the dataset. The advantage of PAM is that it estimates the value of k by observing the data. Next, the algorithm computes the k representative genes, called mediods which is defined as the center point of each cluster and whose

average dissimilarity to all the objects in the cluster is minimal. To learn the causal relationships, we use a simple DBN algorithm among the mediods of the clusters and within each cluster. The partitioning of genes into k groups of co-expressed genes reduces the problem dimension by a factor of k and makes the reconstruction task computationally feasible in comparison to the existing methods. However, the precision of the estimated network solely depends on the identification of co-expressed genes.

3. **Integration of other sources of biological data to restrict the number of regulators for each target gene:** One way of improving the precision of the reconstructed network is to reduce the data requirements of the model. Since the number of time points (T) needed for estimating a network is exponential in the number of regulators (k) for each gene, we use the two sources of biological data, Protein-Protein Interaction (PPI) and transcription factors binding site (TFBS) data to restrict k. Of the two types of known interactions between proteins/genes, we assume that the genetic interactions carry some biological evidence of one interactor being regulated by the other. Then again, transcription factors which have at least one binding site in the promoter region of a target gene are considered as potential regulators along with the genetic interactors of the gene. Finally, a DBN structure learning algorithm is applied to learn the candidate regulators of a target gene from these biologically driven potential regulator sets. The removal of irrelevant genes from the potential regulator lists for each target gene not only makes the reconstruction task computationally feasible but also reduces the data requirements considerably. As a consequence, the model estimates networks with a significantly higher precision.
4. **Investigate the scalability of the DBN-based models through the analysis of networks of varying dimensions:** Scalability is one of the essential criteria for the successful reconstruction of GRNs. In the cell cycle of unicellular yeast, there are approximately 800

genes which change their expression patterns and participate in the completion of the cell division process. Therefore, to obtain a complete picture of the gene regulation program in the yeast cell cycle, the computational models need to handle all the 800 genes simultaneously. However, most of the computational methods, especially DBN-based models are only applicable to the analysis of small-scale networks because of the high computational complexity. Although some models can estimate networks that may include up to 100 genes, the precision of the models drops significantly as the network size grows. We investigate the scalability of our proposed DBN-based GRN models in comparison to some existing models. The experimental results show that incorporation of biological knowledge is a key factor in designing a scalable approach for the reconstruction of GRN.

1.5 Thesis Outline

This thesis focuses on a research problem that emerges from the multidisciplinary area of computer science, statistics and molecular biology. Therefore, we provide both the biological and computational foundation of this thesis work. The rest of this thesis is organized as follows:

Chapter 2 gives an overview of the biological background of the thesis; in particular, the core principles of transcriptional control in eukaryotic cells. It also describes the basic principles behind microarray experiments and gives a sense of what the data actually represents.

Chapter 3 introduces the computational methods that are currently being used for the reconstruction of GRN. It also discusses the key challenges that the GRN models face; some of these challenges arise from the nature of the data; others are associated with the computational models. It also briefly discusses the effectiveness of the models in addressing

such challenges.

Chapter 4 discusses the exploitation of biological domain knowledge of the yeast cell cycle in estimating GRNs. It presents a framework of the proposed GRN model based on DBN learning and discusses the detailed methodologies used. The chapter concludes with experimental results of analyzing the yeast cell cycle gene expression data with the proposed model together with two existing DBN-based models.

Chapter 5 describes the application of the Partition Around the Mediods algorithm in finding groups of co-expressed genes. It also presents a framework of the proposed GRN model based on DBN learning and discusses the detailed methodologies used for estimating co-regulators of co-expressed genes. The performance of the proposed model is compared with the model in chapter 4 in conjunction with two existing DBN-based GRN models.

Chapter 6 presents the integration of multiple sources of biological data in estimating GRN from gene expression data. It discusses the methodologies for extracting potential regulators from other sources of data and application of DBN to estimate the network through the analysis of gene expression data. The chapter also presents the experimental results of analyzing benchmark data from the yeast cell cycle with the proposed model along with the models discussed in chapters 4 and chapter 5.

Chapter 7 presents validation of the proposed models through the analysis of both simulated and experimental data. The benchmark criteria, precision and recall are used in conjunction with computation time for evaluating the performances of the models. Three additional statistical measures, namely F-measure, Negative Predictive Value and Specificity are also computed to test the performance of the models. It also compares the performance of the estimated networks with randomly generated networks to investigate the significance of the proposed models.

Chapter 8 investigates the scalability of the proposed models in comparison with two existing models. The models are applied on two different experimental datasets of the yeast cell cycle with a varying number of genes. The scalability of the models is compared in terms of computation time together with precision and recall.

Chapter 9 summarizes the key contributions of this thesis, proposes future research directions for further improvement of the GRN models, and draws some concluding remarks. We conclude this chapter with the finding that the utilization of biological domain knowledge of gene regulation is at the core of successful reconstruction of GRN.

CHAPTER 2

EUKARYOTIC GENE TRANSCRIPTION

In the recent years, significant advances have been made in the study of gene regulation, of which the control of transcription appears to be the most important component. Gene regulation drives the processes of cellular differentiation and morphogenesis, leading to the conception of different cell types in multi-cellular organisms where the different types of cells may possess different gene expression profiles though they all have the same genomic code. The control of gene regulation in the prokaryotic cells is a simple system where the expression of multiple genes is regulated by only one control point. In contrast, the eukaryotic cells have much more complex control system as every gene has usually more than one regulator and all of these regulators must be turned on for the gene to function.

With the availability of complete genome sequence, much attention has been paid in developing ways to measure the genome-wide transcriptions. One such large-scale experiment is microarray which quantifies the expression levels of thousands of genes simultaneously under a particular condition, called gene expression analysis. These experiments generate an enormous amount of data and provide a rich source for the study of gene regulation. The main purpose of this chapter is to give an overview of the biological foundation of the thesis; in particular, the core principles of transcriptional control in eukaryotic cells. We also review the basic concepts involved in a microarray experiment and what the microarray data actually represents.

2.1 Introduction

In all living organisms, biological components interact in a coordinated way to promote development and sustainability and, therefore, they play a key role in all the processes that happen in these organisms. The components of an organism range from the organelles performing processes in a cell, to cells functioning in an organ, to organs contributing in a subsystem, to subsystems performing processes in the whole system itself. The association in which these components work harmonically together is through sets of complex regulatory networks and pathways.

A transcriptional regulatory network refers to the collection of genes, their regulators and their interactions within a cell operating at the transcription level. It dynamically orchestrates the level of expression for each gene in the genome by controlling whether and how briskly that gene will be transcribed into mRNA. This mRNA carries coded information for synthesizing protein, and the functions of the protein are the determinants of the ultimate cell types. Therefore, uncovering such a regulatory network will assist researchers to understand the cell's function (the sequence of interactions in different developmental processes) as well as how a cell dysfunctions in the case of diseases, such as cancer. However, one of the most important applications of such discovery is in drug design, as the notion of accurate biological models from discovered regulatory networks or pathways can help us to predict the responses of a drug to the specific disease/infection. The derivation of such pathways would be also very beneficial for plants as it would open new options to combat plant's diseases.

Over the last two decades, molecular biology research has evolved through the development of microarray technology, such as DNA microarray, Protein microarray and Tissue microarray. A DNA microarray is a biological assay to monitor the expression levels

for thousands of genes in parallel to study the effects of certain treatments, diseases and developmental stages on gene expression within a living cell. The outcome of such experiments is massive data, which has provided a rich source allowing the scientific community to investigate and understand the fundamental aspects underlining the growth and development of life as well as exploring the genetic causes of anomalies occurring in the functions of living cells.

2.2 Biological Aspect of Transcriptional Regulation

In all living organisms, the heritable biological information, known as genetic code, is materialized within the cell nucleus as DNA (Deoxyribonucleic acid) base sequence. The information in DNA is made up of four chemical bases: adenine (A), cytosine (C), guanine (G) and thymine (T). A gene normally resides on a stretch of DNA that codes for a protein or an RNA chain that has a function in the organism. Proteins are a linear chain of monomers called amino acids bonded together by peptide bonds. There are 20 naturally occurring amino acids and the sequence of amino acids observed in a protein is defined by the sequence of nucleotides in a gene, which is encoded in the genetic code

2.2.1 Gene expression

Gene expression is the process by which biological information (genetic code) from a gene is used in the synthesis of a protein or a functional RNA. The flow of information is generally from copying of the gene into an RNA replica, known as the messenger RNA (mRNA) to the decoding of the mRNA into a polypeptide chain through two major transformation steps: transcription and translation. The biological information can be modified at each step, and its flow can be controlled within each transformation step. In

addition, there are some biological processes such as replication, RNA splicing, post-translational modification, which process the information within these transformation steps. All these steps and processes together form the core of the so-called Central dogma of molecular biology (Watson and Crick, 1958; Crick, 1970). The process of synthesizing proteins from DNA within a cell can be summarized in five major stages:

1. Replication is a process by which the nucleotides sequence in DNA is replicated in presence of many enzymes.
2. The transformation step by which nucleotides sequence information is transferred from DNA to RNA is known as transcription.
3. In eukaryotic cells, the messenger RNA (mRNA) is modified by a process known as splicing, in which introns (nonprotein-coding segments of DNA) are removed and exons (protein-coding segments of DNA) are joined.
4. mRNA carries coded information to ribosomes. On receiving this information, the ribosomes use it for protein synthesis. This transformation step is known as translation.
5. Post-translational modification is the chemical modification of a protein after translation. This involves actions such as changing the chemical nature or structure of one or more amino acids, thus allowing a range of new functions for the protein.

In summary, the biological information stored in the DNA is processed in the cell machinery and the resulting products are specific types of proteins. These proteins do not code for the production of another protein, RNA or DNA. They are the chief activators in a cell and involved in almost all biological activities, structural or enzymatic. The flow of biological information from DNA to protein through two major transformation processes (transcription, translation) is illustrated in Figure 2.1.

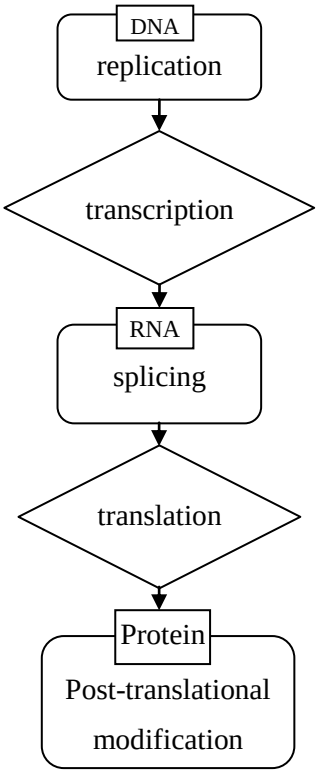


Figure 2.1: The flow of biological information from DNA to RNA to protein. Information from DNA to RNA is passed through transcription and from RNA to protein through translation. Prior to every transformation step the biological information is processed by different processes such as replication, splicing etc.

2.2.2 Genomic regulation system

All cells in an organism carry the same DNA, with the same information. The precise temporal and spatial expression of specific genes (segments of the DNA that code for proteins) governs the identity of the cells. Thus a muscle cell expresses a set of genes different (at least in part) from that expressed by a skin cell. These differences occur at the level of transcription, most commonly at the initiation of transcription. Within a cell, the transcription initiation is regulated through a gene regulation program which gives the cell the control of the timing, location, and amount of gene expression and has a profound effect on the

functions of the gene in the cell.

All living cells can be divided into two basic types: prokaryotic and eukaryotic. Animals, plants, fungi, protozoans, and algae are all eukaryotic, whereas only bacteria have prokaryotic cell types. The two cell types share some similarities and possess some differences in regard of cellular structure, for instance prokaryotic cells have no nuclei, while eukaryotic cells do have true nuclei. Likewise, many principles of transcriptional regulation in prokaryotic cells resemble that of eukaryotic cells. For instance, in both cell types the transcription is regulated by activators and repressors, which are DNA-binding proteins that help or hinder transcription initiation at specific genes in response to appropriate signals. There are however, additional regulatory machinery in eukaryotic cells and genes that complicate the action of these regulatory proteins. The most significant of these additional complexities is that there are multiple regulators and more extensive regulatory sequences (Watson et al. 2008) in eukaryotic cells. In prokaryotes, individual regulators bind short sequences to initiate transcription, but in eukaryotes, these binding sites are often more numerous and positioned further from the start site of transcription. In the following subsections, we focus on the biological aspect on the complex transcriptional regulation in eukaryotic cells.

2.2.2.1 Transcription Regulation

Transcription is the process in which one DNA strand is copied by making a complementary RNA strand. The transcription of eukaryotic protein-coding genes is generally regulated by cis-acting elements within the regulatory regions of the DNA, and trans-acting factors that include transcription factors (TFs) and the basal transcription complex. The basal transcription complex includes one or more proteins and many of these proteins come as large

complexes. The Cis-acting elements are DNA sequences associated with the gene being regulated. These sequences influence the expression of the gene through the interaction of trans-acting factors. Two most significant cis-acting elements are: core promoter, (binding site for basal TFs and Polymerase II), and enhancer (binding site for TFs). In eukaryotic cells, one gene may have one or more enhancers and one enhancer can act on multiple genes. The transcription regulatory elements of a gene and their corresponding binding sites are illustrated in Figure 2.2.

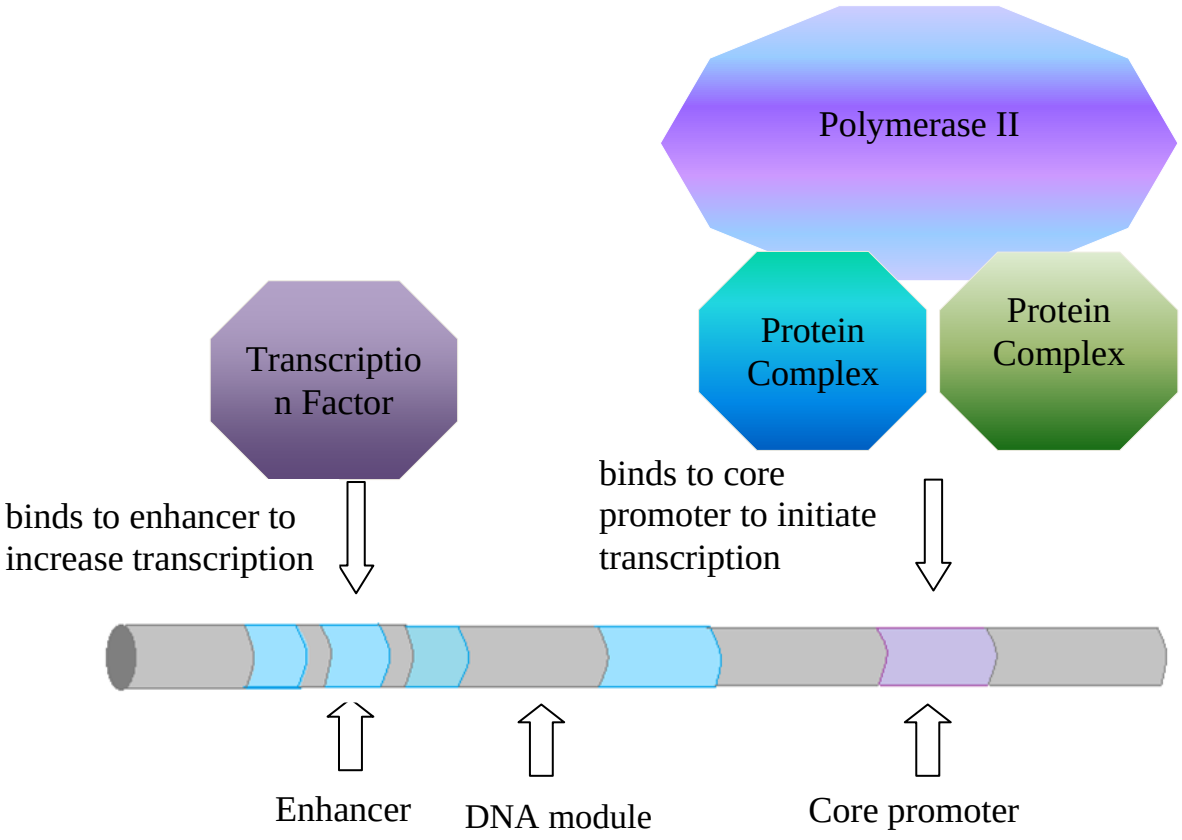


Figure 2.2: Regulatory elements on a protein-encoding DNA module. Polymerase II binds to the core promoter through the basal TF complex (one or more protein complexes) to start transcription at a very low level and determine the transcription start site. The TFs bind to the enhancers to increase expression.

In the synthesis of RNA, transcription initiation is the most pervasively regulated step

and mostly regulated on a gene-by-gene basis. In eukaryotic cells, the basal TFs (one or more protein complexes) bind to the core promoter of the gene and the enzyme complex, Polymerase II binds to the basal TFs to initiate low level of transcription and determine the transcription start site. The enhancers bind TFs (one or more protein complexes) responsible for activating the gene at a given time and place. Alternative enhancers bind different groups of regulators and control expression of the same gene at different times and places in response to different signals.

2.2.2.2 Gene Regulatory Network

A gene regulatory network (GRN) is a collection of genes in a cell which interact with each other indirectly through their expression products, generally RNAs and proteins. At the transcription level, the GRN governs the rates at which genes in the network are transcribed into mRNA. According to the central dogma of molecular biology, these mRNAs are translated into proteins. Some of these proteins are structural proteins, some are enzymes and others are regulators of other genes. The structural proteins are responsible for determining the structure of the cell; the enzymes catalyse chemical and biological reactions that occur within the cell; the regulatory proteins are called TF and play vital role in the regulatory networks. These TFs bind to the regulatory sequences of the target gene to initiate transcription and control the rate of transcription. Some of the TFs act as activators, others are inhibitors. Through the interaction to the regulatory sequences, the activators turn the gene on and the inhibitors switch it off. The regulatory interactions between genes and their gene products form a complex network, which includes both positive and negative feedback loops. A simple GRN is represented in Figure 2.3, though the control of gene expression is much more complex than it suggests. A typical GRN involves many types of proteins/protein

complexes thus allowing additional levels of control especially in multi-cellular organisms. The sample GRN in Figure 2.3 has three genes and all of them encodes for regulatory proteins (TFs). In the network, Gene1 encodes a regulatory protein which influences the regulation of Gene3. Gene3 synthesizes a regulatory protein that activates the expression of Gene2 and finally, the product of Gene2 regulates the expression of Gene1.

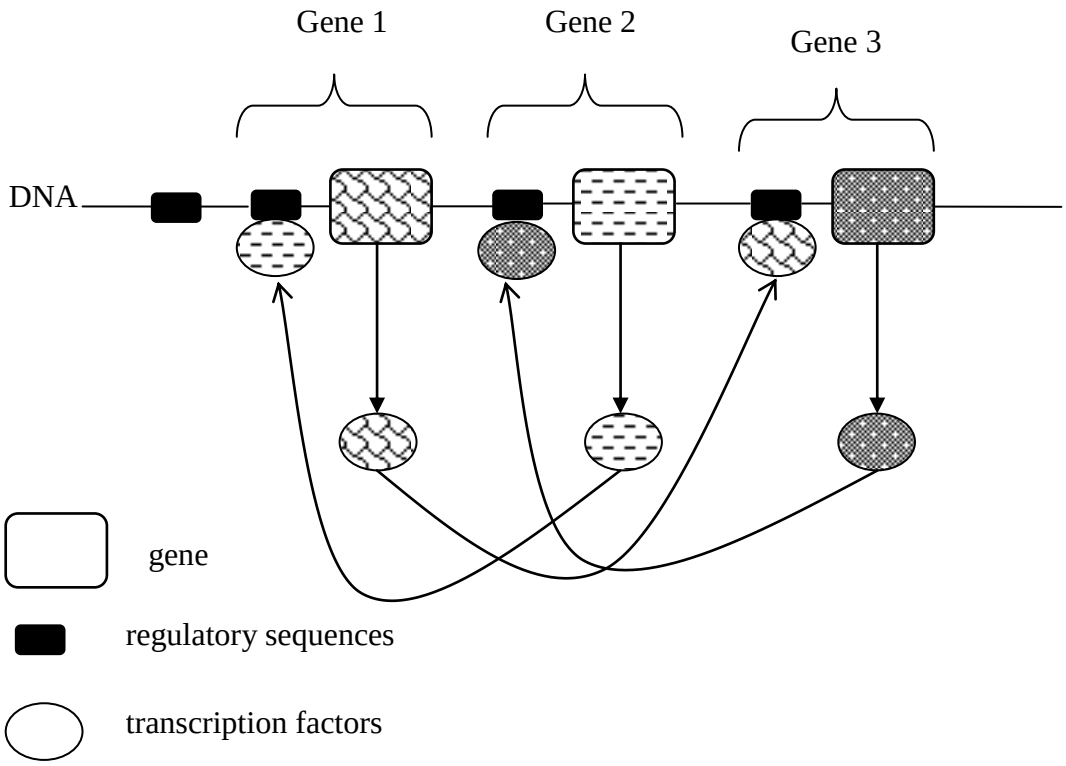


Figure 2.3: llustration of a simple GRN. The network encodes the regulatory sequence, Gene1→Gene3→Gene2→Gene1. All these interactions together with the genes and their products form GRN. Various textrues have been used to fill the shapes of the figure for distinguishing different genes and their respective products (TFs).

In this thesis, our main aim is to study gene regulation at the transcription level and we refer to the regulatory network as either "gene regulatory network", "transcriptional circuitry" or "gene regulation program".

2.3 Quantification of Gene Expression

Measuring gene expression is an important part in the life sciences. The ability to quantify the level at which a particular gene is expressed within a cell can provide an enormous amount of information, particularly in the study of GRN. There is a wide range of experimental techniques available for genome-wide mRNA quantification. These include Serial Analysis of Gene Expression (SAGE), Rapid Analysis of Gene Expression (RAGE), RT-PCR (real-time polymerase chain reaction), Northern/Southern Blotting, and Microarrays or Gene Chips. A detailed literature of some of these techniques can be found in (Roth 2002). Among these, some techniques quantify the level of expression of one gene at a time such as Northern blots. Others, such as microarrays measure the expression of many genes in a single experiment quickly and efficiently, and hence have become popular among the scientific community. A brief introduction to microarray technology is presented in the next section.

2.3.1 cDNA Microarrays

In post-genomic age, there has been a huge escalation in the availability of molecular biological data. Now-a-days, high-throughput technologies such as microarrays are capable of simultaneously monitoring the activity levels of genes/proteins within a cell in a single experiment. The two well-known microarrays are cDNA microarrays and oligonucleotide microarrays. These two technologies differ in the way in which DNA sequences are laid on the array and in the length of these sequences. In this thesis, we focus on the data generated from cDNA microarray experiments. Figure 2.4 summarizes the principles and steps that are involved in a simple microarray experiment.

A microarray is a glass microscope slide on which large numbers of DNA molecules are spotted in a grid like pattern using a robot. Each spot on the microarray contains many

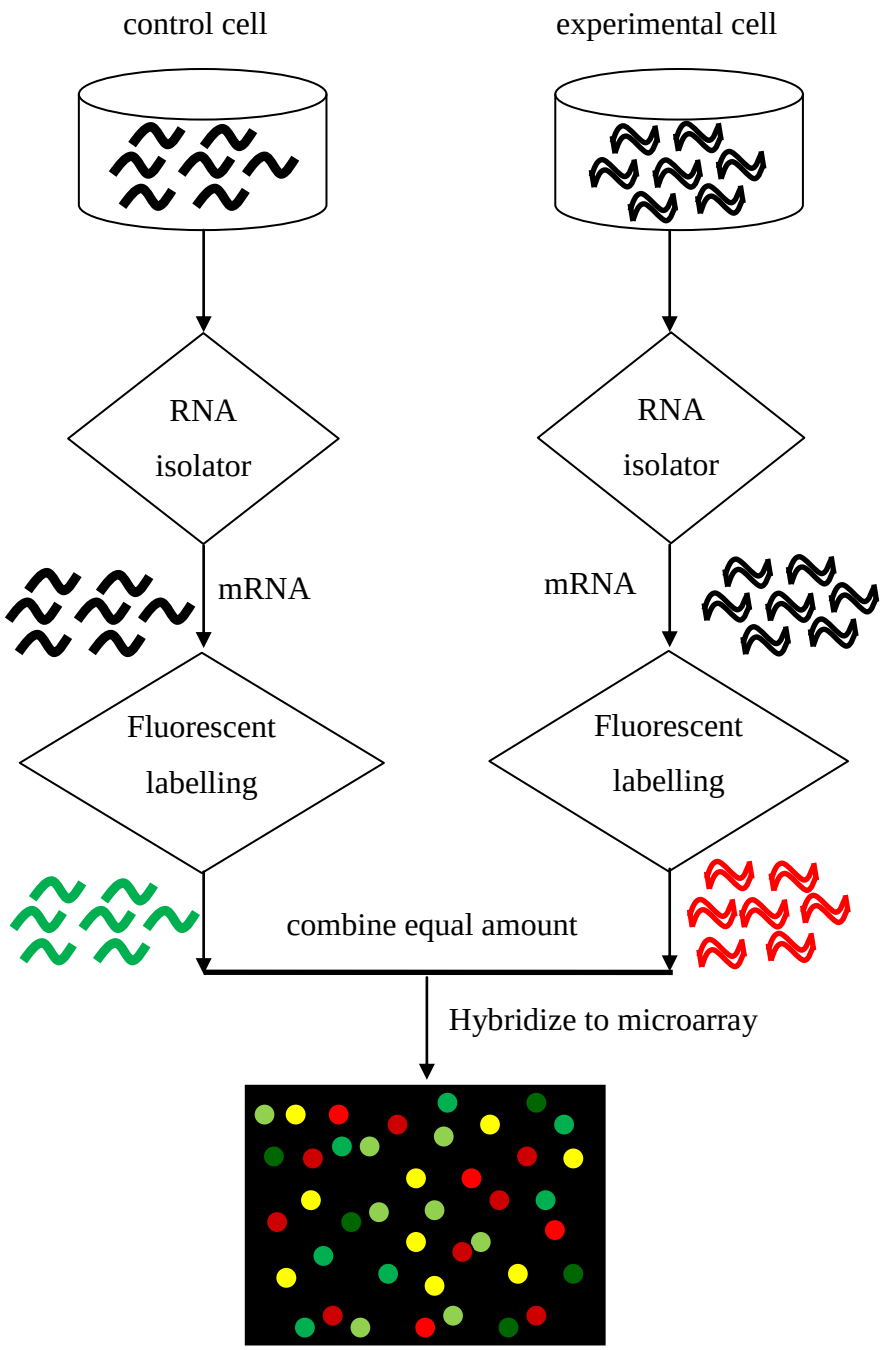


Figure 2.4: Principles and steps of a simple microarray experiment.

amplified copies of a single gene. For the experiments to be carried out, two populations of cells are grown under certain conditions or at different time points. One of these is a control cell which is used as a reference (control) for the other cell under study (experiment). Next, mRNA are harvested from each population of cells and separately converted into

complementary DNAs (cDNA). The nucleotides used to make the cDNA include either a green dye called Cy3 or a red dye known as Cy5. Red color represents experimental cDNAs and green color represents control cDNAs. Both green and red cDNAs are then mixed together and washed over the microarray. After hybridization, a laser scanner measures the intensity of the two colors at each spot on the microarray which specifies the transcription level of that gene under a particular condition.

Yellow spots on the glass slide indicate genes which are expressed in both populations. A black spot indicates no change in that gene's transcription between the control and experimental growth. Then again, the brightness of the red spot indicates how much a gene is induced in the experimental condition. Likewise, the brightness of the green spot shows the amount of repression of a gene in the experimental condition. Finally, the numerical ratios of red to green signals are calculated from the color spots on the array which represent the gene expression profile of the cell under study.

2.4 Our Model organism and the cellular process

In this thesis, we model the GRN in the yeast cell cycle, as the core machinery of the cell division process is largely conserved throughout eukaryotes, from yeast to humans. Beside this, there are other factors which have influenced our preference to yeast and the cell cycle. These are:

1. The high degree of conservation of cell cycle regulation over eukaryotics (ota et al. 2004); it is believed that a complete picture of the gene regulation in yeast cell cycle can provide useful insight in understanding the function of certain cell cycle regulated human genes.
2. Most cellular genes are constitutive genes; that is, they are transcribed at a constant level at all times. These genes are responsible for maintaining the basic function and structure

of the cell. Some genes are expressed in a periodic fashion and control the proper progression of different cellular processes such as the cell cycle. Understanding the mechanism that underlies such periodicity is important for deciphering the control of the cellular process. For cell cycle, such knowledge is even more important as it provides insight into how the control dysfunctions during the production of new tumors especially in the case of cancer.

3. Because of the expansion of regulatory sequences; that is, the increase in the number of binding sites for regulators of a gene, it requires more extensive signal integration in eukaryotes. The unicellular yeast has less extensive regulatory sequences and hence less signal integration than other multi-cellular eukaryotes.
4. The shorter cell cycle of yeast compared to other eukaryotes has made it easier to observe cell processes.

2.4.1 Cell Division in Yeast Cell Cycle

The cell division cycle is an ordered set of events whereby a cell grows and divides into two daughter cells so that each contains the information and machinery necessary to repeat the process. The whole cycle can be divided into four distinct phases as shown in Figure 2.5. These are G1 phase, S-phase, G2 phase and M-phase. The most important aspect of cell division is to replicate the genetic material (DNA sequence) accurately and then separate the two copies into two daughter cells. The process of DNA replication and Chromatid separation occurs in S–phase and M-phase respectively. In general, S and M phases are separated by two gaps, known as G1 phase and G2 phase. During phase G1, the cell ensures that everything is set for DNA replication. At G2 phase, the cell determines if the cell is prepared to proceed to chromatid separation.

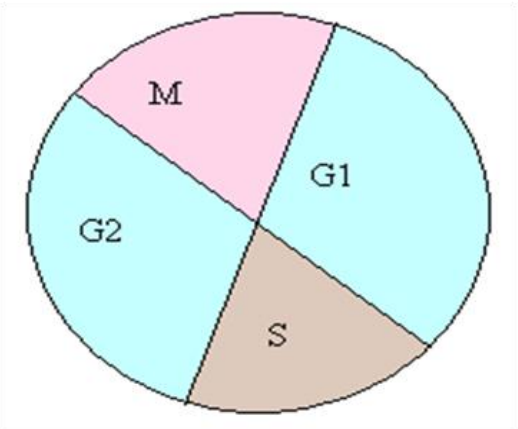


Figure 2.5: Distinct Phases of Cell Division Cycle.

2.5 Experimental Micorarray Data

In this section, we present real data from microarray experiments, in which the levels of gene expression have been measured as the yeast cell goes through the different stages of the cell cycle. At present, an enormous amount of experimental data is freely available for studying the underlying biological processes that generate these data. In this thesis, we have chosen to work with few widely accepted datasets, of which the dataset generated by Pramila et al. (2006) is complete; that is there are no missing values in the dataset. Therefore, we nominate this dataset for our primary investigation. The dataset is a time-series which represents the dynamics of gene expression over two cell cycles. The levels of gene expression have been monitored at a regular interval (5mins) resulting 22 time points in the dataset. For an ample presentation of the numbers (levels of gene expression) in microarray data, we plot them against time as in Figure 2.6. The figure shows the levels of expression of the 13 well known TFs that have been proven as cell-cycle regulated. These include G1-phase specific TFs (MBP1, SWI4, and SWI6), S-phase specific TFs (HCM1, WHI5, and YOX1) and G2/M phase TFs (FKH1, FKH2, NDD1, YHP1, MCM1, SWI5 and ACE2). Some of these TFs act as an activator (increases the level of gene expression) and others are repressors

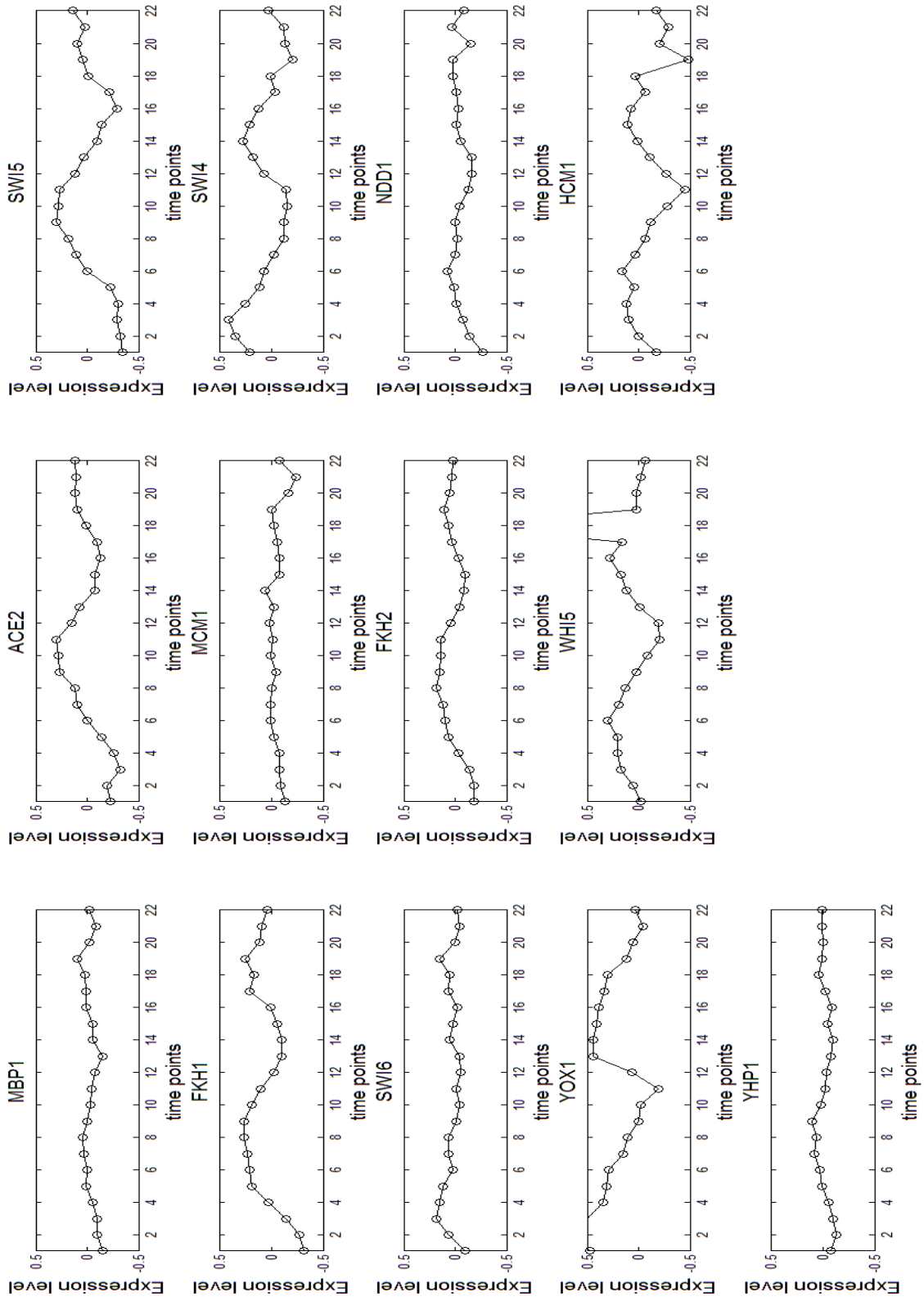


Figure 2.6: Dynamics of gene expression of the 13 known cell-cycle regulated TFs over the two cell cycles

(decrease the level of gene expression). The transcriptional regulation among these TFs has been widely studied and an incomplete framework has been published in the study of Pramila et al. (2006) as shown in Figure 2.7.

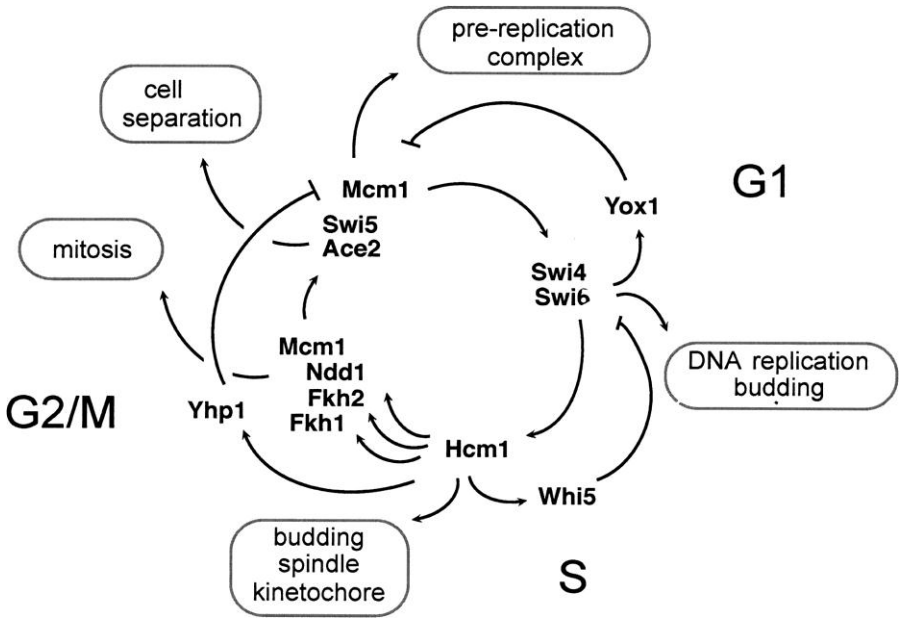


Figure 2.7: GRN among the 13 known TFs. [taken from the study of Pramila et al. (2006)].

A simple exploration of this known network accomplishes a series of regulatory actions among the TFs. For instance, at the start of a cell cycle the G1-phase specific TF, SWI4 is activated by MCM1, which is a G2/M phase specific TF of previous cell cycle. As the cell progresses through the next phase, SWI4 activates HCM1 which is responsible for regulating chromosome segregation genes and other regulators such as WHI5 during S-phase. Consequently, the S-phase TF WHI5 represses the expression of SWI4 as the cell progresses through to G2/M phase. A simple regulatory sequence among these three genes within one cell cycle has been extracted from the known network and illustrated in Figure 2.8. To explore how successfully the available experimental microarray data reflects the known regulation program, we investigate this simple network. Our experimental data contains 22 time points

over two cell cycles. However, determining the span of a single cell cycle from these time points requires detailed information about the microarray experiments that have been carried out to generate the data.

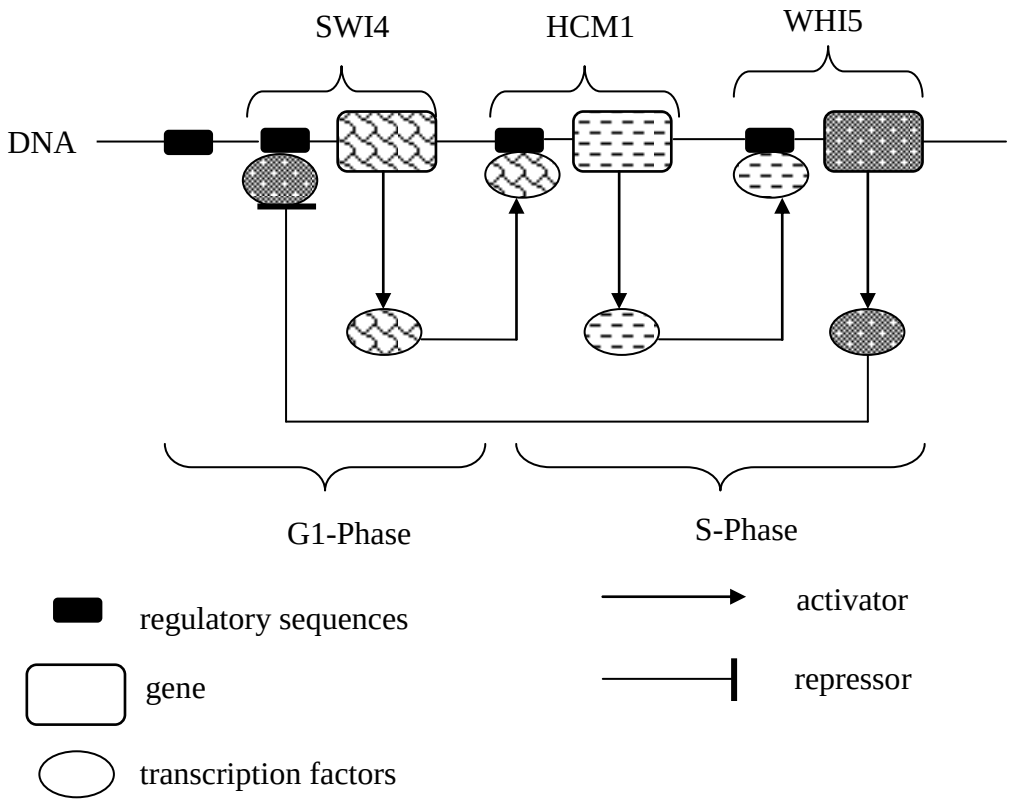


Figure 2.8: A simple transcriptional regulation network among three transcription factors during the first two phases of the cell cycle. The TFs are SWI4, HCM1, and WHI5.

For the purpose of our primary investigation, we have plotted (Figure 2.9) the expression profiles of the three TFs (SWI4, HCM1, and WHI5) over time. The plot shows that all three genes dynamically change their levels of expression as they progress through the different phases of the cycle. At the beginning of each cycle, SWI4 gets activated and gradually reaches its peak at some point of the cell cycle. Being maximally expressed, it starts to decline and get repressed at the later phases of the cell cycle. We assume that a gene is

maximally expressed when its expression level reaches its highest value. A similar pattern of expression can be observed during the second cell cycle for SWI4 as shown in Figure 2.9.

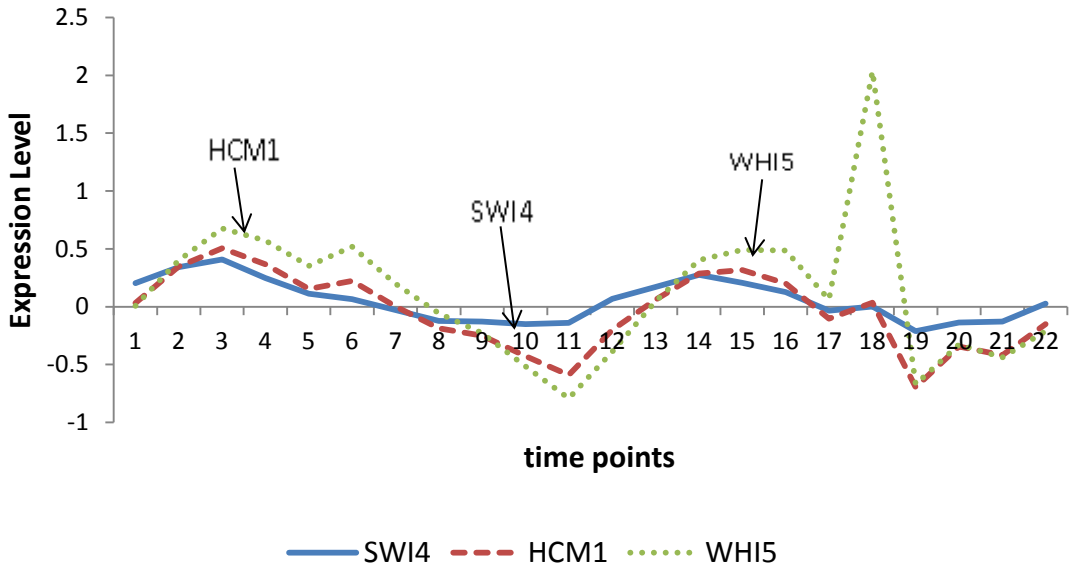


Figure 2.9: Fluctuation in the level of gene expression of the three TFs (SWI4, HCM1, and WHI5) as measured by microarray experiments. There are 22 time points in the data which covers two consecutive cell cycles.

In a typical GRN, a regulatory gene needs to be maximally expressed prior to influencing the expression of other genes. This hypothesis drives an order of expression from the known network as in Figure 2.7, which is SWI4->HCM1->WHI5. To investigate how this order of expression is revealed in the experimental data, we focus on the segment of the expression profiles when these genes are functionally active. The known network suggests that SWI4 is active during G1 phase which spans over the first 3 time points in the experimental data. The other two regulatory genes, HCM1 and WHI5 are active during the S phase which corresponds to the next 4 time points of the dataset. This mapping of time points to the phases of cell cycle has been extracted from the study of Pramila et al. (2006). Figure 2.10 magnifies the fragment of the plot (Figure 2.9) which shows the dynamics of expression

over the G1 and S phases of the first cell cycle. Roughly, this includes the first 8 time points of the dataset.

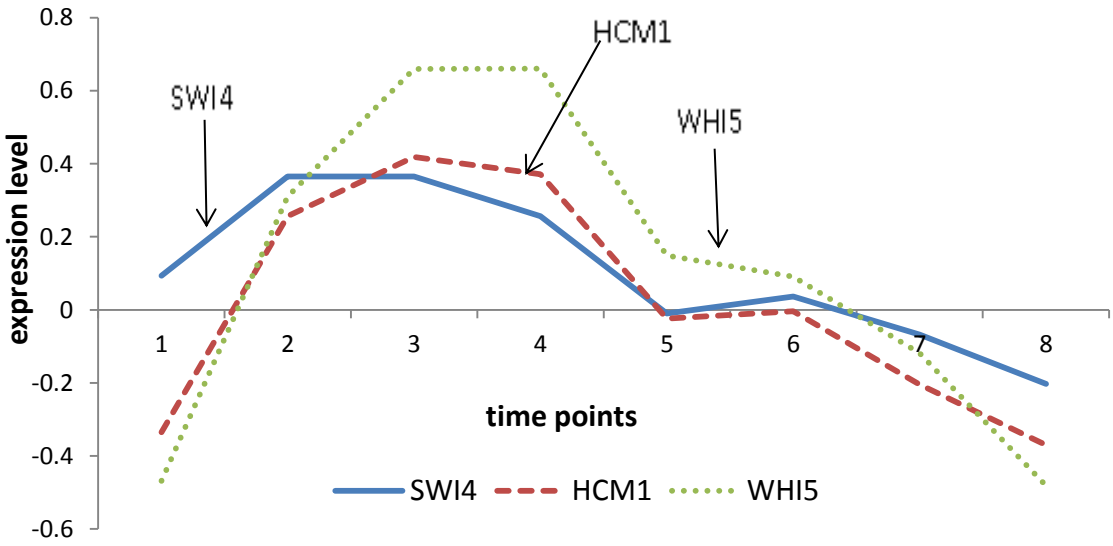


Figure 2.10: Order of expression among the three TFs (SWI4, HCM1, and WHI5) within a cell cycle. We assume, a gene is maximally expressed when the expression level reaches its highest value.

As shown in Figure 2.10, SWI4 gets activated by some regulators as the cell enters into the cell division process. It keeps activated throughout the G1 phase and starts to repress gradually as cell progresses through the S-phase. The S-phase specific TF, HCM1 starts to transcribe in the middle of G1 phase and increased its expression level during later of the phase. However, HCM1 is maximally expressed at the start of S-phase; WHI5 gets maximally activated in the middle of S-phase and declines later of the phase. The sequence and the time lag between the peaks (maximum expression) of these TFs conform to the regulatory association depicted in Figure 2.8. This conformation also encourages us to explore experimental microarray data for the study of gene regulation at the transcription level.

Despite competence, there are some pitfalls in analyzing microarray data which may not give a researcher a true picture of gene regulation. For instance, the S-phase specific

regulator WHI5 starts to transcribe at the beginning of G1 phase and becomes functionally active during S-phase, then declines expression later in the phase. Our experimental data also reveals this expression pattern over the two cell cycles as shown in Figure 2.9. However, in the second cycle, the figure shows a sharp increase in the expression level of WHI5 where it is expected to decrease gradually. This unpredicted behavior of a gene is considered as noise which is one of the major shortcomings associated with experimental microarray data. A variety of factors may contribute in making the data imperfect, such as poor sensitivity of microarrays, background noise, imprecision in designing the array etc. Nevertheless, the experimental microarray data holds great promise for understanding genes and their impact on disease, drug discovery and development.

2.6 Conclusion

Although control of gene regulation is a complex process, understanding when genes are expressed and at what levels has become an essential part of almost any biological inquiry at the cellular level. Many technologies have been in existence for measuring the levels of gene expression for decades. Microarrays have become popular in recent years because of their high-throughput quantification of gene expression. In this thesis, we analyze microarray gene expression data of the yeast cell cycle to uncover the underlying regulatory network that specifies when and how a gene is expressed during several different phases of the cell division process.

CHAPTER 3

CURRENT APPROACHES TO MODELING GENE REGULATORY NETWORKS

The recent advances in high-throughput microarray have led to an unprecedented growth in available gene expression data. This deluge of data demands for methods that can facilitate the understanding of organisms at the cellular level. In the literature, many different methods have been proposed and studied with a goal of describing the Gene Regulatory Network (GRN) in a precise and unambiguous manner. These methods include logical models such as Boolean networks, mathematical models such as differential equations, probabilistic models such as Dynamic Bayesian Network (DBN), and machine learning models such as recurrent neural networks and models that exploit the various sources of biological information in association with the gene expression data. Some of these models are applicable to only discrete data whereas others can model continuous gene expression data. However, all these models have their own advantages and limitations in describing the gene regulation program. The most common limitations are: (1) high dimensionality, (2) excessive computational complexity, (3) handling noise, and (4) small number of samples in the data set. Although DBN suffers from some of these limitations, their probabilistic nature allows them to capture the stochastic aspect of gene expression and the noisy measurements. Given microarray data are inherently noisy and have missing values in it, we choose DBN as our preferred GRN model in this thesis.

3.1 Introduction

Although many approaches have been proposed for the analysis of microarray gene expression data, the field is still evolving and there is a need for methods that can handle this data in a global fashion. The analysis of gene expression microarray data can be classified into three levels of increased complexity (Baldi et al. 2002). The levels are as follows:

1. The first level involves the analysis of a single gene. At this level, the expression profile of each gene is observed in isolation to investigate whether it behaves differently in a control versus experimental situation. The outcome of such investigation can be used to identify gene targets for drug design.
2. At the second level, the extent of complexity increases as multiple genes are included in the analysis. The most popular technique at this level is clustering. Genes are clustered in terms of common functionalities, expression patterns, interactions, co-regulations etc. to predict the behaviors of unknown genes. For instance, a cluster of co-regulated genes can provide useful insight of the regulatory mechanism of many genes for which no regulatory information is available at present.
3. The third level entails the most complex analysis of array data. It attempts to understand and reverse engineer the biological networks that are ultimately responsible for generating the dynamic patterns observed in the data.

In this thesis, we focus on the third level of data analysis that infers the interactions among the biochemical molecules in a naturally occurring large-scale biological network. Several different kinds of biological networks can be identified at the molecular level, such as gene regulatory networks (GRN), metabolic networks, signal transduction networks and protein-protein interaction (PPI) networks. Here, we focus exclusively on GRN which

regulates the amount of gene regulation at the transcription level. The transcriptional regulation network principally relies on the interactions of the transcription factors (TF) with the DNA target; nevertheless it also includes interactions among TFs and RNA polymerase complex. This implies that the transcriptional regulation itself may include PPI network in addition to the protein-DNA ones (Kepes 2007). Due to these additional complexities, it is uncertain from a biological point of view that the genome-wide mRNA concentrations, measured by the microarrays are capable of estimating the transcriptional regulation. And so, to make the data analysis task simple, we restrict our focus on estimating GRN that contains interactions between TFs and genes only.

3.2. Challenges of Analyzing Microarray Data

"The term "data deluge" is often used in association with microarray data"(Gershon 2002). A typical microarray experiment generates thousands of data items and this number increases extensively in cases of temporal experiments. Therefore, there is no doubt that analysing such massive data involves definite challenges. In this section, we highlight the major challenges associated with microarray data that hinders the application of computational models, in particular, in estimating GRN.

1. **Small sample size:** the ratio of the number of samples to the number of genes in the array data is very low. The technological and other practical limitations restrict the number of samples that can be measured. This problem leads to a severe challenge on the application of the computational methods in estimating GRN from such data. The consequences of the small ratio of genes to samples and its impact have been discussed in a number of studies (Jain et al. 1997, Dougherty 2001 and Hwang et al. 2002). In the machine learning society, this problem is mostly known as learning in almost empty spaces (Duin et al.

2000). In estimating GRN, this problem makes it quite difficult to distinguish noise from structure, unless aprior knowledge is known about the underlying interaction network generating the data.

2. **Inherent noise:** It is widely established fact that microarray data comes with enormous experimental noise (Aris et al. 2004). The changes in the measured transcript values between different experiments are caused by both biological variations such as a cell progressing through the phases of cell cycle and experimental noise. A range of factors can contribute to the generation of this experimental noise. As we briefly listed in the previous chapter, this may include imprecision in designing the array, sample preparation, the hybridization process, the normalization method, incorrect scanning of the array spots, improper filtering of data etc. In general, the noise can be injected at each single step of the whole microarray experiments. A plethora of studies (Tu et al. 2002, Aris et al. 2004, klebanov et al. 2007, Posekany et al. 2011) have been reported in the literature to analyze and decipher noise from the microarray data. This deciphering is a vital step in the GRN inference process as the presence of noise provides misinformation to the computational model.
3. **Missing Values:** The inherent noise and inadequacy in the experimental system are the two principle factors causing missing values in microarray data. A study by de Brevern et al. (2004) estimated that on average a microarray dataset has more than 5% missing values, which affects more than 60% of the genes. Most of the available computational methods are either applicable to only complete datasets or are subject to significant performance degradation when applied on the incomplete data. This supposition has led researchers to employ methods for estimating missing values which has become a crucial pre-processing step for microarray data analysis.

3.3. Current Approaches for Modeling GRN from Gene Expression Data

In the literature, hundreds of approaches have been proposed for the estimation of gene regulatory network from post-genomic data. The core principles of these approaches come from the different areas of science and engineering including mathematics, physics, biology, statistics and computing. The number of publications proposing new approaches for inferring GRN is increasing exponentially every year. Therefore, it is beyond the scope of this thesis to discuss them all. Several good reviews of current approaches for the modeling of GRN can be found in (D'haeseleer et al. 2000, de Jong 2002, Schlitt et al. 2007, Cho et al. 2007, Lee and Yang 2008, Sima et al. 2009, and Hecker et al. 2009).

D'haeseleer et al. (2000) provided an abstract of the complex gene regulation program and discussed the spectrum of models that can be used in reverse engineering such a network. The spectrum varies from discrete Boolean networks to continuous linear and non-linear networks in capturing various levels of details of the regulation program. In another review, de Jong (2002) gave mathematical formulation of gene regulation program and discussed the incorporation of this formula in a wide range of available GRN models. These include directed graphs, Bayesian networks, Boolean networks, ordinary and partial differential equations etc. Schlitt et al. (2007) classified the GRN models according to the increasing level of details they can capture. In a comprehensive survey, Cho et al. (2007) reviewed different GRN models that utilize the additional sources of biological information such as ChIP-ChIP data, protein-protein interaction data in conjunction with the microarray gene expression data. They also surveyed a range of machine learning methods (genetic algorithm, genetic programming, neural networks and fuzzy logic) in reverse engineering transcriptional regulation. Of particular interest on the time-series data, Sima et al. (2009) surveyed the GRN models that are capable of capturing a deeper insight of the regulation program than non-

temporal data do. In particular, approaches are discussed that enable the modeling of the dynamics of gene regulatory systems. In the review of Hecker et al. (2009), the authors focus on the models that incorporate other sources of biological information in reconstructing the GRN from the experimental data.

This thesis aims to study the GRN models that can infer the topology as well as the dynamics of gene regulation through the analysis of time series gene expression data. In particular, we investigate the potency of incorporating biological domain knowledge in the models to overcome the limitations associated with the data and the computational methods. Therefore, this section briefly introduces the GRN models that have been widely used in the literature on the analysis of temporal data and are capable of incorporating prior biological information. To further restrict the literature survey, we include only those models that are most relevant to our study. Then again, we categorize the models into two major types: (1) those that use discrete variables and (2) the others that use continuous variables in the inference process.

3.3.1 Models for discrete variables

The GRN models for discrete variables assume that genes only exist in a finite set of discrete states. This discretization of data makes the GRN model computationally feasible which consequently allows them to model large scale networks. Nevertheless, these models have the disadvantage that they lose information which hinders the accurate inference of the models. Therefore, discrete models are not able to capture certain system behavior that can be captured by continuous models (Hernminger et al. 2007). In the following subsections, we discuss some of the discrete models that have been widely used for GRN modeling.

3.3.1.1 Boolean Networks

Boolean network has been considered as one the pioneer model (Akutsu et al.1999, Ideker et al. 2000) in estimating GRN. This model uses the approximation that gene expression is quantized to only two states: ON and OFF. The expression state (on/off) of each gene is functionally coupled to the expression states of some other genes in the network. This implies, whether a gene will be ON or OFF, is governed by a Boolean logical function, such as AND, OR, NOR, XOR etc.

A Boolean network can be represented with a graph G (**V**, **F**), where **V** denotes a set of N nodes ($x_{1,.....}x_{N)}$ that take on binary values and F is a set of Boolean functions, **F** = (f_1,f_2,….,f_N) which describe the interaction of those nodes. Each Boolean function f_i(x_{i1},.....,x_{ik}) has k input nodes and is assigned to node x_i. Figure 3.1 shows a simple GRN and its equivalent logic circuit diagram is represented in Figure 3.2

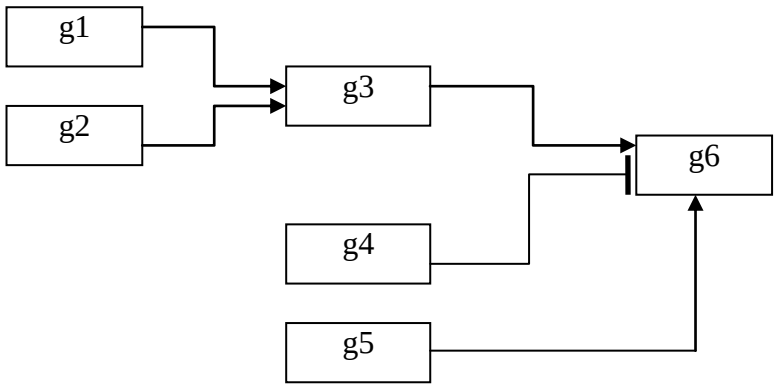


Figure 3.1: A simple GRN. The rectangles represent genes, arrowed lines represent activation and the lines end with a bar represents inhibition.

. In the regulatory network, there are 6 TFs which have either positive or negative influences on the expression of other TFs. Genes g1 and g2 work together to increase the expression of g3. Being functionally active, g3 turns on g6 in association with g5. As shown

in Figure 3.1, gene g6 is regulated by two switches: the positive switch is the g3-g5 complex which turns it ON and the negative switch (g4) inhibits it. In the circuit diagram, the regulation of gene g6 is determined by the states of g3, g4 and g5 which is represented with an AND operation. The NOT gate shows the negative influence of gene g4 on the regulation of gene g6.

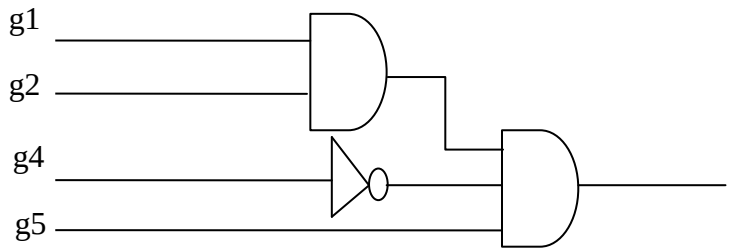


Figure 3.2: The logic circuit diagram representing the GRN. The gates having multiple gene inputs are AND gates. The NOT gate takes a single gene as input. The diagram is inspired by Shmulevich et al. (2002).

In Boolean network models, the nodes are initialized to some states through prior knowledge or estimation. Then, the nodes update their states dynamically throughout the network according to the Boolean rules assigned in the model and based on the current states of the system. This updating process continues until the system enters into a steady state or some irreducible set of states. There are two classes of approaches which are widely used in the literature to construct Boolean networks from experimental time course data. The first approach is based on correlation measurement, which estimates the relationship between genes using different methods such as mutual information. The degree of correlation is used to construct the topological connections between genes (Chen et al. 2008, Laubenbacher and Stigler 2004, Liang et al. 1998, Mehra et al. 2004). The second approach utilizes machine learning methods, of which the most novel is evolutionary modeling with Genetic Algorithm (GA) (Michalewicz 1994). Networks comprising of the Boolean functions assigned to each

node and the relationships among them are encoded in a string representation and the adaptive operations of GA are applied to generate new networks. The network that maximizes the fitness function represents the GRN estimated from the given data.

There are many advantages of modeling GRN with Boolean networks. Most importantly, they are computationally feasible and simple to expand to large scale. Another advantage is that, they are capable of modeling temporal behavior of a system through updating the states of nodes with time. However, the dynamics of Boolean networks are inherently deterministic because of the predefined static Boolean functions. The randomness of the network only depends on the initial node states. This feature restricts Boolean networks to capture some behaviors and uncertainty, which are common to gene regulation

3.3.1.2 Probabilistic Boolean Networks (PBN)

A PBN is an extension of the basic Boolean network which was proposed in (Shmulevich et al. 2002, Li et al. 2007). A new level of randomness has been introduced in this new model by providing each node with multiple logical functions, each with a predefined probability. Similar to the basic Boolean network, the initial states of the network nodes are determined randomly. Then at each time step, the function at each node is selected arbitrarily from this pool of functions according to their given probability. Thus no single predefined circuitry diagram regulates the dynamics of the network and any given set of initial nodes can result in multiple subsequent network states.

For a PBN, G (**V**, **F**), **V** is a set of n nodes ($x_{1,\ldots\ldots,}x_N$) and **F** is a vector of sets ($F_1,\ldots,F_N$), where each constituent F_i is a set of Boolean functions corresponds to the node x_i and can be written as in equation 3.1. Each member function $f_j^{(i)}$ is known as a predictor as it may predict the state of gene x_i at the next time step and l(j) is the number of predictors in F_i.

$$F_i = \left\{ f_j^{(i)} \right\}_{j=1,\dots l(i)} \tag{3.1}$$

The most important step of constructing a PBN from experimental data is to determine the set of predictors for a given gene. Shmulevich et al. (2002) suggested the employment of Coefficient of Determination (COD) for choosing the predictor sets. They considered the COD as the measurement of the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the expression levels of a target gene relative to the best possible prediction in the absence of the observed set. A detailed description of the inference algorithm can be found in Shmulevich et al. (2002). On estimating the predictor sets for each node, a simple scaling function can be employed to compute the probability of a given predictor to be chosen for a given node (Styczynski and Stephanopoulos 2005). The other parameter l(j) is chosen by the user and determines the amount of uncertainty the model can capture.

Similar to the Boolean network, PBN is a potential model for estimating GRN from experimental data as it is able to easily incorporate biological knowledge. In addition, PBN allows multiple simple predictor functions instead of one complex function which consequently facilitates the process of fitting the model to the observed data. However the main disadvantage of PBN is the increased computational complexity. The computation of predictor sets for each gene from the observed data requires excessive time which makes the model unfeasible for the inference of a large scale network.

3.3.1.3 Bayesian Networks (BN)

A Bayesian network model is a graphical representation of a joint probability distribution over a set of random variables, $\boldsymbol{X} = \{X_1, X_2, \dots\dots, X_N\}$. In the model, random variables include observed attributes, such as the expression levels of genes and hidden

variables that are estimated by the model. The graphical representation is given by a Directed Acyclic Graph consisting of nodes and directed edges. The nodes represent random variables in **X** and the directed edges represent dependencies between the variables. The joint probability distribution (θ) over **X** is represented as a product of conditional probabilities. Each variable X_i in the network is associated with a conditional probability $P(Xi|\boldsymbol{pa}(X_i))$, where $\boldsymbol{pa}(Xi)$ is the subset of **X** and is called the parent of X_i. The values of the parent set directly influence the choice of value for X_i. Under the conditional independence assumption, that is, each node X_i is independent of its non-descendants given its parents $\boldsymbol{pa}(X_i)$, the joint probability distribution of the network can be written as:

$$P(\boldsymbol{x}) = \prod_{i=1}^{N} P\left(x_i \middle| \boldsymbol{pa}(X_i)\right) \tag{3.2}$$

Figure 3.3 shows the Bayesian network representation of the GRN in Figure 3.1. In the network, the nodes correspond to the genes and the directed edges represent the direct causal relationships between the genes. For instance, in the example network, the edge (g2,g3) implies that g2 has a direct influence on the regulation of g3. The second parameter θ has been described in the model as a table which shows the probabilistic dependency as a form of conditional probability P(g3|g1,g2) associated with node g3. In the example network, we assume that the expression levels of genes are discretized to two states: 0 and 1.

The general goal of inferring a GRN from the experimental data is to learn a model that is as close to the underlying distribution as possible. This involves two major tasks: parameter estimation and model selection. The first task learns the parameters of the conditional probabilities for a given model structure and it is often considered as a maximum likelihood problem (Friedman 2004). The second task selects among different model structures to find the one that best reflects the dependencies in the data. For each possible model structure, a Bayesian scoring metric is employed which scores the model given the set

of data. Therefore, the model selection task is often referred to as an optimization problem.

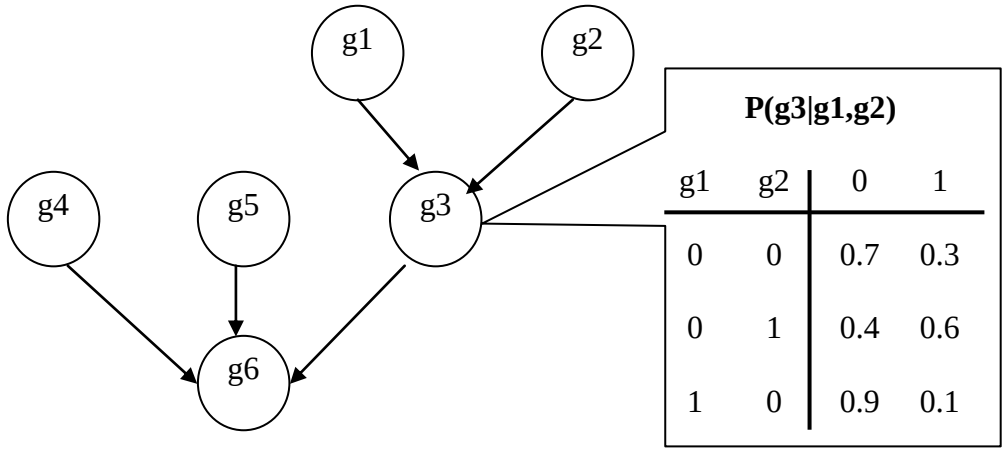


Figure 3.3: A Bayesian network representation of the GRN including 6 genes. Nodes are labeled with gene names (g1 to g6); edges correspond to direct dependencies. The table shows the conditional probability distribution that specifies P(g3|g1,g2).

also allows incorporating prior biological knowledge in the model as prior probabilities. Despite the rich statistics and probabilistic semantics in Bayesian models, they are computationally expensive and barely expandable to large scale networks. The most important disadvantage of Bayesian modeling is their acyclicity constraint; that is, no cycles are allowed in the networks. This constraint limits BN to model a GRN, because in nature there are feedback loops in biological networks. The other disadvantage is that the static nature of Bayesian models restricts it in capturing the dynamics of the biological networks as well as models the temporal data.

3.3.1.4 Dynamic Bayesian Networks

A Dynamic Bayesian Network (DBN) extends the notion of a Bayesian Network (BN) to model a system that is dynamically changing or evolving. This model allows the user to monitor and update the system over time and further predicts the system behavior. Thus the

structure of DBN describes the qualitative nature of dependencies between the random variables over time. The inclusion of this time dimension in the classical BN enables DBN to model a system with cyclic edges. Figure 3.4 illustrates the process how of DBN can address the acyclic constraint of Bayesian networks. Consider the gene network in Figure 3.4(a) where two genes interact with each other and have feedback loops. Clearly, the network cannot be represented with a classical Bayesian network; nevertheless time-point representation of the BN model can include the feedback loops as shown in Figure 3.4(b). In this representation, there are no links between random variables within a time slice. Therefore, all interactions are among the genes at consecutive time slices and the identical network structures are duplicated over each time slice.

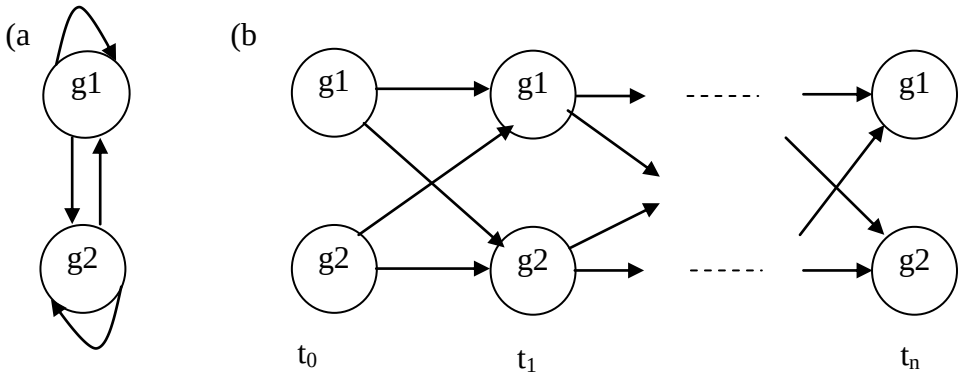


Figure 3.4: A Dynamic Bayesian Network representation of the GRN (a) A regulation network of two interacting genes, g1 and g2 having feedback loops. (b) a time-point representation of the network where the identical network structure of the genes are duplicated over each time slice (t_0, t_1, …t_n).

Assume **X**= $\{X_1,..X_N\}$ is a set of random variables that the process changes over T time points. In the DBN representation of this temporal process, $X_i[t]$ denotes the random variable X_i at time t, where $t \in \{1,\ldots,T\}$. Therefore, there are T×N interacting nodes in the model. To simplify the situation, we have assumed that the process is a first-order Markov

process. Under such an assumption the joint probability distribution in equation 3.3 can be rewritten as:

$$P(x[1], \ldots\ldots, x[T]) = \prod_{i=1}^{N} \left[P(x_i[1]) \prod_{t=2}^{T} P\left(x_i[t] \middle| \boldsymbol{pa}(X_i[t])\right) \right] \quad (3.3)$$

where the first-order Markov assumption means that the variables in the set **pa**(X_i[t]) are a subset of **X**[t-1].

The inference of DBN from experimental data can be performed using the same methods of Bayesian network learning. The additional complexity is to consider the random variables of time t-1 which makes DBN computationally infeasible. As a result, DBN models are mostly applicable to small systems compared to Bayesian network models. To ensure that the model can be inferred from data, it requires restricting the number of parameters needs to be estimated. One widely used assumption for a such restriction is to consider the dynamic process as a homogeneous Markov chain, where the transition probabilities between adjacent time slices are time-invariant and edges are not allowed within a time slice. However, in practice, especially in the case of GRN, this assumption of DBNs may not hold because some genes would interact almost instantaneously while interactions amongst some other genes could be time delayed.

3.3.2 Models for continuous variables

The Bayesian models that we discussed in section 3.3.1.3 and 3.3.1.4 can easily be generalized to the case of continuous variables by adapting a continuous distribution such as the Gaussian distribution. The only pitfall of such an adaptation is the significant increase in computational complexity. Since the structure learning of Bayesian model from discrete data

is already NP-hard (Styczynski and Stephanopoulos 2005), the additional complexity will limit the use of continuous variables in both forms of Bayesian modeling. In the following subsections, we discuss the most popular GRN models for the analysis of continuous expression data.

3.3.2.1 Ordinary differential equations (ODEs)

Ordinary Differential Equations (ODEs) is one of the most popular mathematical formalisms to model dynamical systems in science and engineering. They have been also widely studied in the reconstruction of GRN (Chen and Church 1999, Gardner et al. 2003). The ODE formalism describes the gene products, e.g., mRNAs, proteins as time-dependent variables with values contained in the set of non-negative real numbers. The functional and differential relations between the variables represent the regulatory interactions.

More specifically, the ODE model formulates the regulation of a gene as a function of other genes; that is, the change in the expression level of a gene at any time t is characterized by a function of the concentration of other genes at the same time. The formalisms have the mathematical form as in equation 3.4.

$$\frac{dx_i(t)}{dt} = F_i(x_1(t), \ldots, x_n(t)) \tag{3.4}$$

where $x_i(t)$ is the concentration of mRNA for gene *i* measured at time *t*, $dx_i(t)/d_t$ is the rate of change for the mRNA concentration of gene i, and n is the number of genes. Each function F_i represents all of the various factors that affect the expression level of a gene such as transcription rate, degradation, post-transcriptional modifications and translation rate. As a result, the model is sufficiently flexible to capture any of these detailed interactions for an accurate representation of the gene regulation. However, this flexibility introduces additional parameters to the model which are to be estimated from the small available data. Moreover,

the functional form of F_i is unknown which can be linear, piece-wise linear, pseudo linear, or continuously nonlinear (Lee and Tzou 2009). Though linear models have fewer parameters to estimate, they are unable to capture the nonlinear relationships precisely that may be present in the biological networks.

3.3.2.2 S-System model

The most popular and widely used ODE model is the S-system model that is characterized by power law functions (Savageau 1991). It has a rich structure capable of capturing the different level of dynamics present in the biological network. More specifically, the S-System model has been applied successfully in GRN modeling (Di Bernardo et al. 2004, Kimura et al. 2005, Kikuchi et al. 2003, Cinquemani et al. 2008, Savageau 1998). The model describes a set of differential equations in the following form:

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{N} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{N} X_j^{h_{ij}} \tag{3.5}$$

where X_i represents the expression level of gene i and N is the number of genes in a gene regulatory network. α_i and β_i are rate constants of activation and degradation of gene i. A positive value of g_{ij} indicates that gene *j* activates gene *i*. A negative value indicates that gene *j* inhibits the activation of gene *i*. Likewise, a positive value of h_{ij} indicates that gene *j* increases degradation of gene *i*, and a negative value represents that gene *j* inhibits the degradation of gene *i*. And the zero values of g_{ij} and h_{ij} mean gene *j* and gene i have no regulatory relations. The total number of these parameters is $2N(N + 1)$. Since the number of S-System parameters is proportional to the square of the number of genes, a large number of parameters need to be simultaneously estimated while inferring GRN from expression data. This requirement leads to the excessive computational complexities and restricts the

application of the S-System model on small scale networks only. In the literature, the parameter estimation task has been considered as a large-scale parameter optimization problem. Many methods including linear algebra methods (Hernminger et al. 2007), steady-state analysis (Curto et al. 1997), and evolutionary algorithms (Ho et al. 2007, Noman and Iba 2007) have been employed to solve the optimization problem. Of the different methods reported in the literature, evolutionary algorithms (EAs) have become popular for estimating the model parameters, as they demonstrate better performance in searching a large solution space (Baeck et al. 2000).

3.3.2.3 Neural network models

Another class of continuous variable models for GRN estimation is the neural network based models. Of these models, the recurrent neural network (RNN) has been shown to be successful in modeling GRN (Vohradsky 2001, Xu et al. 2007, Blasi et al. 2005, Lee and Yang 2008). The main strength of RNNs is that they are biologically conceivable and noise resistant. Moreover, the recurrent connections of RNN models provide the flexibility of generating oscillatory and periodic activities which allows them to represent dynamic behavior of a system over time (Lee and Yang 2008). Most importantly, the model is capable of capturing feedback loops that are naturally occurring in biological networks.

A variety of RNN architectures have been proposed in the literature ranging from restricted classes of feedback to full interconnection between nodes. Among these models, the fully recurrent neural network is the most-studied model for estimating GRN. In the fully connected recurrent network, each node has an edge to every other node, including itself. To estimate a GRN with such a neural network model, each network node represents a particular gene and the wiring between the nodes defines regulatory interactions. In the GRN, the level

of expression of genes at time t can be measured from other gene nodes and the output of a node at the next time point (t+1) can be derived from the expression levels and connection weights of all the genes connected to it at time t. This implies that in RNN modeling, the regulatory effect on a gene can be determined as a weighted sum of all the other genes which are potential regulators of this gene. To compute the expression rate of a gene from the neural network, the transformation rules in equation 3.6 and 3.7 are generally applied,

$$\frac{dx_i}{dx} = k_{1,i}G_i - k_{2,i}x_i \tag{3.6}$$

$$G_i = \left\{1 + e^{-(\sum_j w_{i,j}x_j + b_i)}\right\}^{-1} \tag{3.7}$$

where y_i is the actual concentration of the i-th gene product; $k_{1,i}$ and $k_{2,i}$ are the accumulation and degradation rate constants of gene product, respectively; G_i is the regulatory effect on each gene that is defined by a set of weights estimating the regulatory influence of gene j on gene i, and an external input b_i representing the reaction delay parameter.

The construction of the GRN with RNN involves settings of the thresholds and time constants for each neuron, and the weights of the connections between the neurons so that the network can produce the levels of gene expression as measured in the microarray experiments. This task can be considered as a parameter estimation problem that either maximizes the network performance or minimizes an equivalent error measure. For the purpose of evaluating the network performance or error estimation, a scoring function is typically introduced in the learning process.

Although RNNs are useful models for learning nonlinear relationships in time series data with complex temporal dependences, their high computational complexity limits their application on small scale network only. This is due to the number of parameters that need to be estimated from the experimental data. Moreover, the RNN model goes through hundreds of

trials with training data to stabilize the states of network nodes, which in turn makes the learning process a complex one for large scale networks. Most importantly, even though the estimated GRN generates gene expression data which are close to the experimental data, it does not guarantee the accuracy of the inferred network. This is due to the fact that there can be multiple GRN models that can generate similar expression data. This intricacy originates because the number of samples in the experimental data is insufficient for estimating the large number of parameters of the model.

3.4. Reconstruction of GRN by Incorporating Biological Information

In the previous section, we have reviewed the most widely used models for inferring gene regulation network from time-series gene expression data. All of these models suffer from several fundamental disadvantages such as high dimensionality, computational complexity, data uncertainties, small number of samples in the data set etc. Some of these difficulties can be addressed by incorporating prior knowledge in the inference process. The domain knowledge can be used to decompose the whole inference problem into smaller sub-problems and then apply the available methods to estimate the sub-networks separately. In other application of domain knowledge, a set of candidate regulators are derived from multiple sources of biological data for each gene and the resulting network represents the regulatory network. Nevertheless, the most important aspect of incorporating biological knowledge is that it reduces the data requirements of the available methods.

A variety of biological information such as TF binding DNA sequence (Tavazoie et al. 1999), gene function annotations (Friedman 2004), ChiP-chip data (Lee et al. 2002), Protein-Protein Interaction data (Chaturvedi et al. 2007) have been used in reverse engineering GRN. Tavazoie et al. (1999) reconstructed GRN from both time-series gene expression data and

sequence motif data. They first applied the K-means clustering algorithm to gene expression data which groups genes into k partitions. From the databases of MIPS (Munich Information Center for Protein Sequence), they identified functions and the sequence motif of the genes within the same cluster by using the AlignACE (Aligns Nucleic Acid Conserved Elements) (Roth et al. 1998) program. The TFs which recognize the motif sequence of the genes within the same cluster are considered as candidate regulators for the gene group. Finally, the transcriptional sub-networks are merged together to represent the ultimate GRN. However, it is difficult to identify the whole set of regulators for a cluster of genes by recognizing the sequence of motifs using only a single source of data. To overcome this problem, additional biological information such as the distance between the sequence motif and ATG (starting point of genes), the orientation of genes (Beer and Tavazoie 2004) and other sources such as genome databases, proteome databases have been incorporated (Segal et al. 2003).

The GRN reconstruction methods that use clustering algorithms to find groups of co-regulated or co-expressed genes from expression data generate many false positives. This is because of the inherent noise in the data which directly influences the clustering algorithm to find similarity in the expression pattern. This also indirectly influences the searching and identification of sequence motifs. To address this problem, a genome-wide location analysis based on ChIP-chip data (Lee et al. 2002) has been proposed to determine all the target genes that can be bound to the promoter region of a given TF (Bar-Joseph et al. 2003, Gao et al. 2004, Wang et al. 2005, Scott et al. 2005, Banerjee et al. 2003, Qian et al. 2003, Ihmels et al. 2001, Berman et al. 2002). In a promising study, Bar-Joseph et al. (2003) proposed a GRAM (Genetic Regulatory Modules) algorithm which takes protein-DNA binding data and gene expression data as sources of biological information. First, the algorithm identifies all possible subsets of regulators ($T_{1,i}$, $T2_{,i}$,…..$T_{k,i}$) for a target gene i using the protein-DNA binding data

where k is the number of possible subsets and each $T_{j,i}$ represents a subset. Next, for each $T_{j,i}$, a set of genes G_i is derived to which all the TFs in $T_{j,i}$ are commonly bound. Then, the algorithm finds a gene g_i from the set G_i that has a core expression profile and derives a collection of genes S from G_i which exhibits highly correlated expression patterns with g_i. In repeated steps, GRAM adds new genes in the set S which are similarly expressed as g_i and to which all the TFs of $T_{j,i}$ commonly bind. Following this step, the algorithm derives a module that contains similarly expressed and co-regulated genes. The above process is repeated for all possible subsets of TFs for each target gene in the dataset.

The various sources of biological data have been employed in different levels in the process of reconstructing GRN. Mobini et al. (2009) showed the importance of employing information extracted from various sources of biological data to narrow down the search space after a GRN is constructed from the gene expression data. They suggested that the various databases including pathway databases such as KEGG, PATHWAY studio, DAVID etc. (Nikitin et al. 2009, Dennis et al. 2003, Okuda et al. 2008, Yeung et al. 2008, Bindea et al. 2009), TF binding site databases such as Genomatix, oPOSSUM, PAP, Amadue, GeneXPress (Sui et al. 2005, Chang et al. 2006, Linhart et al. 2008, Zheng et al. 2003) and PPI database such as GRID, PIPS, DIP (Loots et al. 2006, Breitkreutz and Stark 2003, Xenarios and Eisenberg 2001) can be used in different orders and levels of the inference process to answer a ranges to biological inquiry at the cellular level.

The transcriptional regulation is a complicated process. Though, the employment of diverse biological information is a promising way to address the current challenges associated the reconstruction of GRN, the available experimental datasets have intrinsic errors. This limitation raises a question on the precision of the biological knowledge that is generally derived from the noisy experimental data.

3.5 Discussion

To reconstruct a network structure with current computational approaches, experimental iterations and prior knowledge are required until sufficient data is available. For the reconstruction of GRN, the experiment involves altering the network structure in some way such as adding or removing a regulatory relationship, observing its effect, and using mathematics and logic to infer the underlying principles of the network. The prior knowledge includes any known biological phenomenon that is frequently observed in the GRN. For instance, there is a high degree of homology between yeast and the human genome; a known regulatory relationship of among two yeast genes can be used as a prior knowledge in the estimation of GRN for a human. Therefore, we can simplify the process of inferring GRN as a sequence of two major iterative steps. These are: (1) select a network model and (2) fit network parameters and structures into the available data.

As discussed in the previous section, all the current approaches for GRN modeling have advantages and certain limitations. Selecting an ideal model for GRN is a difficult task as it has to take into account the current challenges associated with the experimental data as well as the nature of gene regulation program. Below we list some of the key issues that govern the effective application of the GRN models on the microarray gene expression data.

1. **Ability to model both discrete and continuous Expression data:** The first issue in modeling GRN is to decide whether discrete or continuous expression data is used. In general, discretization of data causes information loss and the use of continuous data makes the model computationally expensive. Though Ott et al. (2004) argued that the form of data either discrete or continuous has no effect on the results obtained; they have not shown any side-by-side comparison to establish their argument. Among the aforementioned GRN models, some are strictly applicable to discrete data such as Boolean

networks, PBNs; others can model only continuous data such as RNN and S-System models. However, all variations of Bayesian models such as BN, DBN can be readily applicable on both data types with increased level of computational complexity.

2. **Prior knowledge:** A number of studies (Shatkay et al. 2000, Spieth et al. 2005) have recommended accumulating as much biological knowledge as possible and incorporating them as prior knowledge in the modeling process for the successful reconstruction of GRN. This pre-existing knowledge of the network structure reduces the number of possible alternatives and hence narrows down the number of free parameters that the model needs to estimate. This reduction consequently reduces the data requirements of the model. All the GRN models that have been presented in section 3.3 are able to incorporate prior knowledge. In general, the prior knowledge of the system is included as the initial state of the network structure.
3. **Performance:** Performance is a key issue in GRN modeling as it indicates how good a model is. Two major criteria of evaluating model performance are: (1) computation time and (2) accuracy. Although computational complexity is a theoretical criterion for evaluating model performance, the researchers have mainly confined their interest on computation time. This is because the computational complexity remains exponential with the size of the network despite the diversity in the modeling techniques. In general, discrete models are computationally feasible but less accurate, whereas continuous models are computationally expensive and can estimate GRN more accurately. All the continuous models discussed in section 3.3 require estimating a large number of parameters from small time series data which imposes a severe bottleneck on the performance of these models. Amongst the discrete models, Boolean networks are the most computationally feasible approach, whereas PBN models require estimating a huge number of parameters

which makes the model computationally expensive. As stated in section 3.3, the structure learning of Bayesian models is NP-hard. Therefore, computational complexity is a key challenge for all types of GRN models. Studies (Segal et al. 2003, Ott et al. 2004) have shown that the use of other sources of biological databases contributes significantly in improving the performance of GRN model. The biological data sources can be used to restrict the search space which subsequently reduces the computational time of the learning process. It is expected that the inclusion of biological knowledge also improves the accuracy of the estimated GRN. This issue needs to be further investigated through experiments.

4. **Robustness:** The robustness of a GRN model, as we define it, is its ability to demonstrate stable performance against external disturbances such as inherent noise and missing values in the data. It is expected that the GRN models show a certain level of robustness to noise and incomplete data. Amongst the aforementioned models, RNNs are ideally suited to handle random noise. To investigate the influence of noise on the inference of network structures from short time-series data, Streib et al. (2005) studied three discrete models, namely Boolean Network, PBN and DBN. The authors concluded that an increasing amount of external noise reduces significantly the overall performance of all three models; however, DBN performs better than the logical models as the noise level increases. Among these models, Boolean networks show least robustness to noise because of their deterministic nature. In another study, Noman and Iba (2007) examined the performance of the S-System model with different levels of injected noise and found that model performance declines significantly with the increased amount of noise.

   In case of incomplete data, the GRN models opt to handle missing values either in a pre-processing step in the inference process or within the inference algorithm. In the

former case, the missing values are estimated using available imputation methods such as Nearest Neighbor Averaging imputation (Hastie et al. 1999). This pre-processing step transforms incomplete data into a complete one and can be used with any of the aforementioned models. However, some of the GRN models such as BN and DBN can directly handle missing values while learning the structure and parameters from incomplete data.

5. **Scalability:** In general, scalability is defined as a characteristic of a model that maintains its level of performance or efficiency when tested by larger operational demands. In case of GRN reconstruction, a model is designated as scalable if it can infer networks of different sizes (from tens to hundreds genes) by preserving a steady accuracy level. Due to the high computation time, none of the abovementioned models is truly scalable. Most of them are able to model small scale networks including tens of genes only.
6. **Stability:** Stability is the sensitivity of the model to the variations of the gene expression data used for reconstructing the network. Despite the application of a wide range of techniques in modeling GRN, the stability analysis of the reconstructed network has not received much attention. In a recent study, Jagath and Piyushkumar (2011) investigated the effect of the number of time points on the stability of the reconstructed network. Through analysis, the authors showed that the ratio of the number of time points to the size of the network has a significant impact on the stability of the model.

3.6 Conclusion

In the post-genomic era, the inference of GRN with computational methods relies on the development of high-throughput technologies as well as techniques of information science, engineering and biology. Many computational models have been studied in the

literature ranging from very abstract level to concrete ones. In general, Abstract models capture the qualitative dynamics of system behavior and include less biological details. They have the advantage of being computationally feasible and can model large-scale networks including hundreds of genes. On the other hand, concrete models describe biological facts in detail and capture network dynamics as close to reality as possible. However, these latter models are computationally expensive and are able to infer small-scale networks including tens of genes only.

In this thesis, one of our key goals is to propose and study GRN models that are scalable. Essentially, a scalable GRN model required the incorporation of biological prior knowledge in the inference process. Here, we choose DBN as our preferred GRN model with the consideration that they are succinctly capable of capturing the stochastic process of gene expression and noisy measurements because of their probabilistic nature. Below we list some of the other incentives that have been taken into account for such a choice:

1. DBNs are able to model temporal behavior of a system such as the periodic activities of cell cycle. In a recent study, Li et al. (2007) compared two probabilistic GRN models, PBN and DBN. They applied the models on the same experimental time-series datasets and concluded that DBN identified more gene interactions and gave better performance compared to PBN.
2. Friedman (1998) proposed Structural EM algorithm for learning Bayesian models from incomplete data. Given the data, the algorithm searches for best the model in a joint space (structure × parameter). In each step, the algorithm either finds a better parameter for the current structure or finds a new structure that maximizes the expected score instead of the actual score. The successful application of this algorithm (Tofigh et al. 2011, Friedman et

al. 2002) has established Bayesian models as a competent approach for analyzing incomplete data.

3. Bayesian models are readily available to model both discrete and continuous gene expression data. Friedman et al. (2000) claimed that the use of continuous data leads to a network including the inherent noise in the data. To avoid this situation, we choose to work on discrete data in this thesis.

4. Scalability and performance are two key issues for GRN modeling. As discussed in the earlier sections, all models including DBN suffer from high computational complexity. This is due to the fact that the search space grows exponentially with the size of the dataset. This limitation restricts them from modeling large scale networks. However, biological networks are naturally scale-free (Han 2008); that is they have a few highly connected nodes in the network. In the context of GRNs, those nodes represent the TFs and regulate the expression of the majority of the genes. This fact suggests that not all the genes in the dataset are TFs; in other words, they have no regulatory affect on other genes. We can utilize this feature in the DBN learning process to restrict the number of possible structures in the search space. It is assumed that such a restriction would improve the performance of the estimated GRN and makes the model computationally feasible. In the following chapters, we investigate this assumption through a series of models and experiments.

CHAPTER 4

PHASE-SPECIFIC REGULATION IN THE YEAST CELL CYCLE[1]

This chapter discusses the employment of the biological features of the cellular process under study in reconstructing GRNs from microarray data. In this thesis, our model of the cellular process is that of the yeast cell cycle. One important feature of cell cycle regulation is that, a high proportion of cell cycle regulated genes are periodically expressed; that is, genes are maximally expressed to affect and control the regulation of other genes and on completing certain tasks; they are repressed by some other regulator genes. Thus the whole cell cycle progresses systematically through the successive activation and inactivation of cell cycle regulated (CCR) genes. We exploit this feature to decompose the entire problem of estimating gene regulation into several smaller sub-problems. As a consequence of the biological knowledge driven decomposition, both the accuracy and the computational time of our proposed model have been improved in comparison with two existing models. Most importantly, we analyze two real experimental datasets of the yeast cell cycle, containing gene expression profiles of 150 genes to study the performance of the proposed model.

---

[1]This chapter presents the results of a conference paper (Shermin and Orgun, 2009a) published in the Proc. of the 24th Annual ACM Symposium on Applied Computing.

4.1 Introduction

Cell division is the process in reproduction and growth by which a parent cell is divided into two or more daughter cells. The rate of cell growth and division is controlled by a cell cycle regulation program. Disruption to this regulation program can lead to medical problems, such as cancer where the cells start to divide uncontrollably. Therefore, cell division and its regulation are identified as one of the most fundamental activities in living organisms. As the control system for the timing and coordination of cell cycle events in eukaryotic cells is undoubtedly complex, one of the most important organisms for the study of the basic cell cycle control mechanism is the unicellular yeast. Among the two species of yeast, namely budding yeast, *Saccharomyces Cerevisiae* and fission yeast, *Schizosaccharomyces Pombe*, the former has been widely used by most laboratories. The studies on this organism are therefore supported by a stronger foundation of biological and methodological knowledge. Although yeast is unicellular, its shorter cell cycle has established it as a valuable model organism for the study of cell cycle control mechanism.

The genome wide microarray analysis of gene expression during the cell division cycle has led to the finding that about 15% of budding yeast genes are cell cycle regulated (Spellman et al. 1998). That is, these genes are subject to transcriptional regulation during the cell cycle. The expression patterns of these CCR genes shift dramatically as cells transit from one phase to another. This phase-specific gene expression articulates that a significant amount of CCR genes are periodically expressed to control the regulation of other genes and perform the phase-specific task. On completion, they are repressed by other regulator genes of either the same phase or the subsequent phases. Therefore the whole cell cycle progresses systematically through the successive activation and inactivation of the CCR genes.

In this chapter, we study the role of the phase-specific regulation of gene expression to

better understand the transcriptional circuitry in the yeast cell cycle. Through experiments on real gene expression data, we demonstrate the effectiveness of such a study in dealing with the key issues that the work in this thesis aims to address. However, all the experiments are conducted on complete data; that is there are no missing values in it.

4.2 Related Work

Gene expression profiles generated by the high-throughput DNA microarray technology represent the dynamic behavior of genes in the transcriptional circuitry. An enormous effort has already been given into designing appropriate Dynamic Bayesian Network (DBN) based models for estimating regulation in the budding yeast cell cycle.

To the best of our knowledge, Friedman et al. (1998) and Murphy and Mian (1999) are to be credited with studying the applicability of DBN first in learning casual orderings in biological processes. Murphy and Mian (1999) discussed the advantages of employing DBN in estimating GRNs from time series data, which include the ability to model stochastic processes, incorporation of prior knowledge in the model, and the ability to learn from incomplete data with hidden variables. Ong et al. (2002) described a DBN-based approach that combines prior biological knowledge with gene expression data to model interactions between sets of genes. They analyzed the time series gene expression data measured in response to physiological changes that affect tryptophan metabolism in E. coli. An initial DBN structure of gene regulation was built from the operon map which shows the operon and their associated genes and a final structure is learnt from the observation data. Kim et al. (2002) studied a statistical approach based on DBN and non-parametric regression model to estimate GRNs. They developed a non-parametric regression model with Gaussian noise to estimate a density function which allowed them to analyze continuous gene expression data

and capture nonlinear interactions among genes. The study showed that the statistical model is effective in reducing the number of false positive in estimated network and can identify a few correct regulatory relationships. Perrin et al. (2003) proposed a DBN-based model which is capable of handling the biological and measurement noise inherent in the microarray data. They analyzed E. coli gene expression data with hidden variables and used a penalized maximum likelihood measure to estimate the parameters of the learning algorithm. Husmeirer (2003) investigated the accuracy of employing DBN on gene expression data through a simulation based analysis. In the study of Yu et al. (2004), the authors examined a range of scoring metrics and search heuristics to find an effective DBN algorithm for reconstructing GRNs. Dojer et al. (2006) showed that the incorporation of perturbation experiment data in Bayesian learning improves the quality of the estimated network. In a more recent study, Nguyen et al. (2012) introduced a deterministic global optimization approach for reconstructing GRN from time course gene expression data. For DBN models that consist only of inter time slice arcs, the authors proposed a polynomial time algorithm that employes the information theoretic scoring metric namely mutual information test in learning the globally optimal network structure.

In a promising study, Zou and Conzen (2005) used a pre-determined threshold for estimating changes in the expression (up/down regulation) of individual genes. Genes that usually have either simultaneous or antecedent changes in expression when compared to their targets were considered as potential regulators. This consideration allowed them to restrict the number of possible regulators of each gene which subsequently reduces the search space. Their method was successful in reconstructing medium-scale networks from experimental microarray data containing 105 genes with improved performance. However, genes are expressed in an arbitrary pattern; the assignment of a pre-determined global threshold is likely

to reveal many irrelevant regulators and make the learning algorithm computationally expensive. Nevertheless, this study demonstrated that the proper exploitation of biological insight in restricting the number of potential regulators of a target gene is a promising way of addressing the current challenges associated with the DBN-based GRN models.

4.3 Background

This section briefly introduces the intricate regulation of different transcription factors during the cell division process. It also presents the basic algorithm for learning DBN from complete data.

4.3.1 Regulation of transcription factors in yeast cell cycle

The GRN in the yeast cell cycle has been revealed as a serial regulation of transcription factors (TFs), whereby transcriptional activators of one phase regulate a group of periodically expressed genes and the activators of the following phases (Bahler 2005). In budding yeast, there are 9 well-known TFs which form transcriptional complexes to regulate the proper progression of the phases during cell division cycle. These are G1-phase specific TFs (MBP1, SWI4, and SWI6) and G2/M phase TFs (FKH1, FKH2, NDD1, MCM1, SWI5 and ACE2). A complete transcriptional regulation among these TFs during the cell cycle of budding yeast has been demonstrated in Figure 2.7.

In the budding yeast, cells decide whether to commit cell division in a process called Start at the end of the G1 phase. The expression of several genes is activated by the two related TF complexes, namely SBF and MBF, during the late G1. The recently expressed genes promote the initiation of DNA replication and other events to facilitate the G1/S transition. The SBF complex consists of two protein components, Swi4p and Swi6p, and

regulates genes which function in budding, as well as membrane and cell-wall biosynthesis, while the MBF complex contains Swi6p and Mbp1p and regulates many genes involved in DNA replication and repair (Iyer et al. 2001, Simon et al. 2001). In late G1 phase, the MBF and SBF complexes activate the expression of the TFs Ndd1 which in turn activates the expression of the Fkh2p-Mcm1p complexes in G2 phase. During this phase, the transcription of MBF and SBF is switched off through the Clb1/2p-Cdk1p CDK complex (Tanay et al. 1993, Koch et al. 1996, Spellman et al. 1998), which is itself activated by a combination of events following transcriptional activation of SBF and MBF. The forkhead TFs, Fkh1p and Fkh2p regulates the transcription of genes that are required for the transition into G2/M phase (Carlsson and Mahlapuu 2002). However, the transcription of Fkh1 is activated by the SBF and MBF complexes in the late G1 phase. The other TF, Ndd1p also plays a positive regulatory role in transcription for the proper transition and progression of G2/M phase. The Ndd1p-Fkh2p-Mcm1p complex activates the transcription of SWI5 and ACE2 during G2/M phase. To complete one cycle, the swi5p-Ace2p complex activates the transcription of SWI4, or induces the transcription of the cyclin gene CLN3 which facilitate the M/G1 transition.

4.3.2 Learning DBN from data

Learning a DBN from data can be divided into four levels of increasing complexity. The first level is the simplest one where the structure of the network is known and the learning algorithm needs to estimate the network parameters from the complete data. The next level is relatively complex as the structure is unknown. In this level, the learning algorithm learns both the edges between the nodes and the sets of parameters from the complete data. The third level of complexity arises from the incomplete data where the learning algorithm needs to assign the missing values and then learn the parameters for the pre-specified network

structure. The most complex level is the learning of both the structure and the parameters of the network from incomplete data. In this thesis, we focus on learning the unknown structure of the GRNs from complete gene expression data.

Given a set of experimental dataset D, the learning algorithm learns two components, a Directed Acyclic Graph (G) and a set of conditional probabilities (θ) that better explains the data. The first component is the structure of the network which contains genes as network nodes and edges as direct relationships among the nodes. We assume that there is no edge between nodes within a time slice and the same network structure is unfolded over the consecutive time slices. The Second component is sets of parameters associated with each node which quantifies the intensity of a regulatory relationship (edge) between the nodes. Assume, **M** is the space of all possible models, the algorithm first finds a model $M^* \in$ **M,** that is most supported by the data D:

$$M^* = argmax_M\{P(M|D)\} \tag{4.1}$$

Having the best structure M^* and the data D, the algorithm estimates the parameters that best fits the structure:

$$\theta = argmax_\theta\{P(\theta|M^*, D)\} \tag{4.2}$$

If we apply Bayes' rule to Equation (4.1) we get:

$$P(M|D) \propto P(M)\, P(D|M) \tag{4.3}$$

where the marginal likelihood implies an integration over the whole parameter space:

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)\, d\theta \tag{4.4}$$

The integral in Equation (4.4) is analytically tractable in case of complete data and if the prior $P(\theta|M)$ and the likelihood $P(D|\theta, M)$ satisfy certain regularity conditions as discussed in (Heckerman 1994, 1995). The term $P(M)$ in equation 4.3 is the prior over structures. The simplest type of prior is the one which is uniform over the structures and can

be defined as in equation 4.5, given the set of all possible models **M** in the space:

$$P(M) = \frac{1}{|\boldsymbol{M}|} \tag{4.5}$$

where |**M**| denotes the number of possible models. A detailed discussion on different types of priors can be found in the comprehensive study of Heckerman (1995).

The other term $P(D|M)$ in equation (4.3) is the marginal likelihood which can be factorized as in equation 4.6, given that all the regularity conditions discussed in Heckerman (1994, 1995) are satisfied,

$$P(D|M) = \prod_{i=1}^{n} P(X_i, pa(X_i)|\mathrm{D}) \tag{4.6}$$

where node X_i and its parents pa(X_i) form a structure and $P(X_i, pa(X_i)|\mathrm{D})$ is the score of the structure given the data D.

One popular approach for finding the model that is best supported by the data is to compute a scoring function for all possible structures M∈**M** and choose the one that maximizes the score. One widely used scoring function is the Bayesian scoring metric (BSM), which is simply the log posterior probability of M given D:

$$BSM\ (M, D) = \log P(M|D) \tag{4.7}$$

By applying the Bayes rule, the scoring function in equation 4.7 can be rewritten as in 4.8,

$$BSM(M, D) = \log(M)\ logP\ (D|M) + c \tag{4.8}$$

where the constant *c* is the same for all structures and log P(M) is the log prior over structures. In case of uninformative prior, every structure is equally likely which means that *log P(M*) is the same for all possible structures. Hence both *logP(M)* and *c* can be safely ignored. Therefore, the problem becomes how to find the best marginal likelihood given the data D. The pseudo code in Table 4.1 shows the basic algorithm to search for the network structure G and parameters θ that maximize the marginal likelihood given the data D.

Table 4.1: Pseudo code of DBN structure learning

Input: Data D and a network G (V, E),

V = {1,…..,N} and E = {Φ} if uninformative prior

Output: Network G, Conditional Probabilities θ

step1: initialize θ

step2: for i = 1, 2……N

step 2.1: generate Q = PowerSet(V) except Φ

where card(Q) = 2^N

step2.2: for each X ∈ Q, Compute a BSM score,

step2.3: find the subset, X ∈ Q with max Score

step2.4: add edges in G from each element of X to i.

step3: end

One major problem associated with this algorithm is that the number of possible structures increases rapidly with the number of nodes as we can see in Table 4.2 which makes the exhaustive search impossible. The second problem is that because of the small number of samples in the available data, the algorithm finds many structures with high scoring posterior probability leading to a huge uncertainty about the best structure.

Table 4.2: Number of nodes vs. number of possible network structures (source Murphy 2001b)

number of nodes	number of possible structures
2	3
4	543
6	3.6×10^6
8	7.8×10^{11}
10	4.2×10^{18}

4.4 Methods

In this section, we propose a GRN model based on DBN, which utilizes biological domain knowledge to decompose the entire GRN into overlapping sub-networks and learn each sub-network individually. The framework of our GRN model is illustrated in Figure 4.1 and the component modules are discussed in the following subsections.

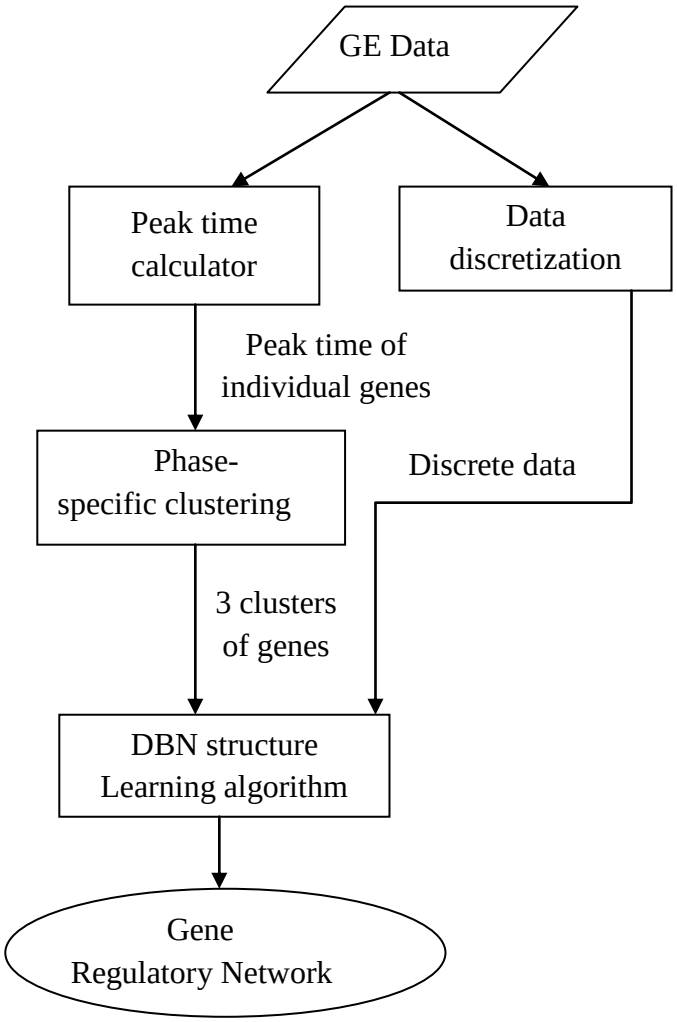


Figure 4.1: Framework of the proposed GRN model. The model groups genes into 3 different clusters corresponding to the phases of the cell cycle and learn the sub-networks from each cluster separately using the DBN structure learning algorithm.

4.4.1 Data discretization

The gene expression profiles store the fluctuation of transcription levels of genes as they go through different phases of cell division. In order to analyze such data with discrete models, we have discretized the expression profiles with a simple 2-state discretization method, illustrated in equation (4.9). The discretization method is chosen at this stage of our study to keep the model computationally feasible as the number of computations grows exponentially with the number of distinct values for each gene in the dataset. For all genes i = 1, …,N and time points t = 1, …, T,

$$X_{it} = \begin{cases} 2 & if \quad X_{it} \geq \overline{X_i} \\ 1 & if \quad X_{it} < \overline{X_i} \end{cases} \tag{4.9}$$

where $\overline{X_i}$ is the average expression level for gene i.

4.4.2 Peak time calculator of genes

One finding of the genome wide transcription measurements through the cell cycle is that a high proportion of CCR genes are periodically expressed. This implies that gene products which are required at a specific point in the cycle are transcribed into mRNA to produce protein and perform phase specific tasks. Upon completing their tasks, genes are inactivated by their repressors. As a consequence, the transcription patterns shift dramatically as cells transit from one phase of the cell cycle to another. Most of the known TFs exhibit this behavior and become maximally expressed just before they are required (Pramila et al. 2006).

de Lichtenberg et al. (2005) proposed a method for determining peak times of periodic genes. The peak time of a gene is the time of the cell cycle when the gene becomes functionally active. Initially, they have calculated a Fourier score (F_i) using equation (4.10)

which quantifies the periodicity of a given gene i.

$$F_i = \sqrt{\left(\sum_t \sin(wt).x_i(t)\right)^2 + \left(\sum_t \cos(wt).x_i(t)\right)^2} \quad (4.10)$$

where $x_i(t)$ is the expression level of gene i at time t. Based on the calculated score, de Lichtenberg et al. (2005) have ranked genes which are periodically expressed. Then, in order to compute the peak time of each periodically expressed gene, each expression profile has been approximated by a sine wave and the time of the peak expression for a gene is defined as the time where the sine wave attains its maximum.

4.4.3 Assigning genes in phase specific clusters

A promising approach to deal with the high dimensionality problem is to use some heuristics to restrict the number of regulators of a target gene. In this chapter, we implement this restriction with the assumption that genes which are at their peak during a specific phase of the cell cycle usually regulate each other. However, the transition of cell cycle phases is controlled by proper activation and repression of TFs between consecutive phases.

Depending on the peak time, each gene in our expression profile has been assigned to three groups corresponding to the biological phases of the cell cycle. The mapping of cell cycle phases to the timeline of the dataset has been drawn from the study of Pramila et al. (2006) and is shown in Figure 4.2. In this mapping, G_2 and M phases have not been identified separately and the period of the cell cycle is set to 58. Therefore, the three groups are G1, S and G2/M. We thus have assigned those genes into a group which are at their peak in the span of the same phase. To illustrate the process, assume an expression profile of 5 genes A-E. The peak time of these genes are 19, 23, 95, 68 and 48 respectively. According to the mapping in

Figure 4.2, three genes A, B and D constitute the G1 group whereas B and C are assigned to the S group. The fifth gene E is the member of the G2/M group. However, it should be noted that by this mapping gene B has been assigned to both the G1 and the S group. The reason is that the phases of the cell cycle are not uniformly separated in the above mapping.

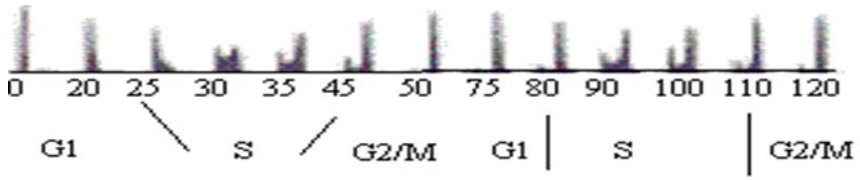


Figure 4.2: Mapping of the cell cycle phases to the timeline of the microarray experiment. [taken from Pramila et al. (2006)]

4.4.4 Inference of GRN with DBN algorithm

To explain the learning algorithm, we assume that the cardinality of one such group is n. Therefore, for a given target gene i, the number of potential regulators to consider is 2^n, as any combination of genes can be the candidate regulators. For a large n, the search space is grows super-exponentially as shown in Table 4.2. In order to deal with this problem, we have restricted the fan-in (the number of input arcs) of each node in the network to k ($\leq$n) with the assumption that there might be genes in the dataset which do not act as regulators. As a result of this constraint, the size of the search space drops to ${}^nC_k$.

Since a DBN has two components to learn, we have initialized the conditional probabilities of a gene in relation to its parents calculated from the observed data. In order to learn the parent set, a scoring function (Maximum likelihood or Bayesian) computes the score of a subset of regulators in conjunction to its target gene using the prior conditional probabilities. The set of regulators with the maximum score is chosen and connections are

constructed between each gene in the chosen subset of regulators and the target gene. If more than one subset scores the maximum, then the parent set is the union of those subsets.

4.5 Experiments and Results

To see how successfully the proposed GRN model can estimate the network topology and address the current challenges, we test the model through the analysis of two recently published experimental microarray datasets (Pramila et al. 2006). To verify the estimated networks, we follow the standard method which is widely used in the literature (Zou and Conzen 2005). We derive a network of known connections from various sources. Some of these sources report the result through in-vivo experiments; others publish in-silico results. On reconstructing the network from real experimental data, each estimated connection is verified against the known network. If a connection exists in both the estimated and the known network, it is considered as a correctly identified connection. If an estimated connection is identified in the opposite directions then we name the connection as misdirected. Finally, if a connection does not exist in the known network but it is estimated by the model, it is considered as an incorrectly identified connection.

4.5.1 Experimental data

We have analyzed two separate microarray gene expression data sets of the budding yeast. Each of these datasets contains expression profiles of 150 cell cycle regulated genes, which have been identified as periodically expressed by the recently published works of Pramila et al. (2006) and de Lichtenberg et al. (2005). The gene expression data are generated and published by Pramila et al. (2006) who have carried out two microarray experiments on cells of the budding yeast as they advance through different stages of the cell-cycle. They

used an alpha factor to induce cell synchronization and the resulting data sets have been named as alpha30 and alpha38 respectively. In order to examine the dynamic behavior of gene expression, they have sampled the microarray at an interval of 5 minutes ranging from t = 0 to t = 120. This length covers approximately two complete yeast cell cycles. However, data points 0, 10 and 105 have been discarded from both datasets for unsatisfactory hybridization leaving 22 samples to analyze. Finally, all the data drawn from these two experiments were processed using an error model in the Rosetta Resolver Version 3.2 Expression Data Analysis System.

In the study of de Lichtenberg et al. (2005), the authors ranked genes in the deceasing order of periodicity. From this published list, we have included 150 top ranked genes in our working datasets given that they have complete expression profiles in the experimental data of Pramila et al. (2006).

4.5.2 Experimental setup

Our experiments are conducted on a computer system with a dual core Intel processor (1.83 GHz) and 2 GB RAM, running windows XP (Professional). Among other available tools, we have chosen Bayes Net Toolbox (BNT) to construct the DBN, which is written in MATLAB and freely provided by (Murphy 2001a). The experiments are setup and run under the MATLAB environment with version 7.6.0.324 (R2008a).

4.5.3 Experimental results

Initially, to test the feasibility of our proposed GRN model, we analyze a small segment of the dataset, including 13 transcription factors that are known to be involved in cell cycle transcription of budding yeast. The dataset include G1-phase specific TFs (MBP1,

SWI4 and SWI6), S-phase specific TFs (HCM1, WHI5 and YOX1) and G2/M phase TFs (FKH1, FKH2, NDD1, YHP1, MCM1, SWI5 and ACE2). Among these TFs, some act as activators and others as repressors. For instance, MCM1 acts as an activator at the beginning of G1 phase and transcriptionally activates SBF (SWI4 & SWI6) complex. At the end of the G1 phase, the SBF complex activates the repressor YOX1, which subsequently represses MCM1. The known network topology of these TFs is extracted from various studies (Pramila et al. 2006, Simon et al. 2001) and is shown in Figure 4.3. Each node in the figure represents a gene and an arrow from gene i to gene j means a direct influence of i on j and a line indicates bidirectional influence. The figure is automatically generated by MATLAB program which receives an adjacency matrix of the estimated network from the model. Due to the image conversion limitations of the program, the figure lacks visual quality; however it is included here for the completeness of the results.

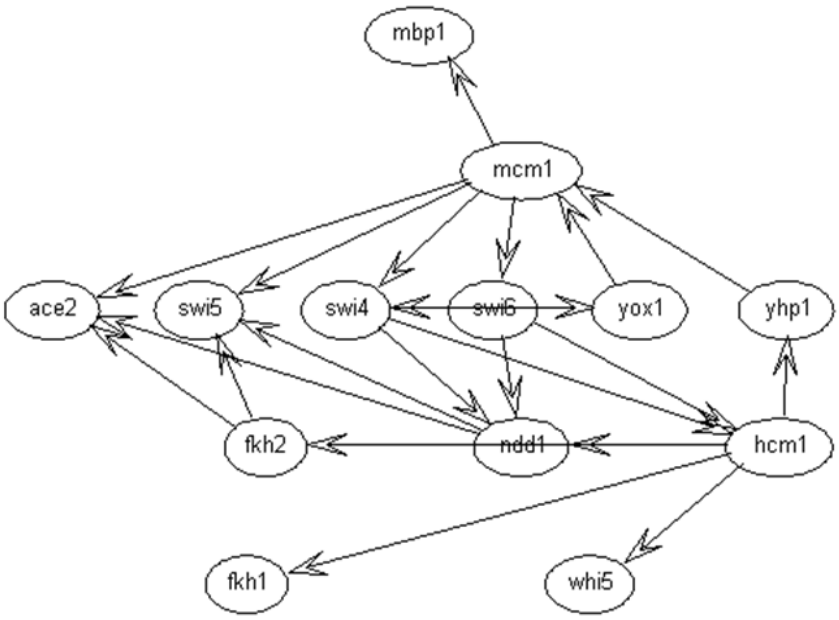


Figure 4.3: Target network of the yeast cell cycle TFs. Each node in the network represents a gene and an edge represents direct regulatory relationship.[generated by MATLAB].

In this experiment, our DBN based GRN model is able to identify only direct regulatory interactions among genes. However, in biological networks, genes are regulated by both direct and indirect influence of the regulators. Figure 4.4 shows the reconstructed network structure which is estimated by our proposed model.

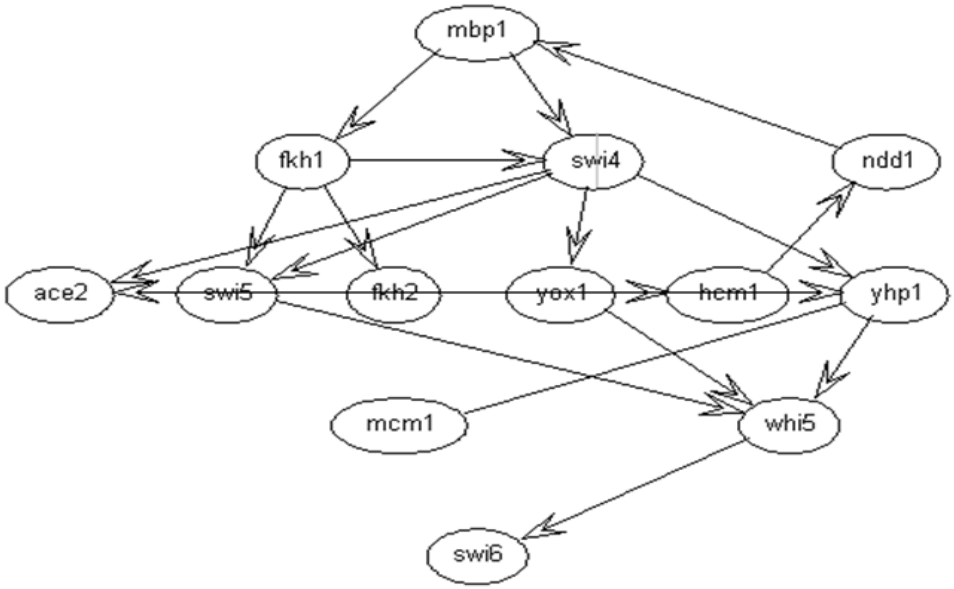


Figure 4.4: Estimated network structure of the yeast cell cycle TFs.
[generated by MATLAB]

Inspecting the constructed network, it can be found that only 4 direct interactions have been correctly identified by our model, whereas there are 22 known connections in the known network. However, if we apply the k-skip validation which assumes that k-genes are skipped in estimating the GRN between the regulator and the target gene (Chaturvedi et al. 2007), our estimated network is able to explain the regulatory interactions more accurately. For instance, in the known network, there is a direct influence of HCM1 on WHI5 whereas in the inferred network the interaction is identified as HCM1->YHP1->WHI5. With 1-skip validation, this is a true positive interaction. Since we are working at the transcription level and regulatory interactions take place after translation, it is sensible to apply the k-skip model for GRN

validation. Of the known 22 connections in Figure 4.4; our model is able to infer 15 connections (approx. 70%) using 1-skip validation. In addition, examining all the extracted connections with respect to the known roles of TFs, it is found that, in most cases, the prediction confirmed the prior knowledge of cell cycle regulation, which establishes the validity of our approach.

Finally, we apply our GRN model along with two existing DBN models (Murphy and Mian 1999, Zou and Conzen 2005) on two different experimental datasets including 150 genes. We compute the accuracy and the computation time of the models to compare their performances. The experimental results of these models on two different datasets are summarized in Table 4.3 and Table 4.4. In the tables, '*Total Identified Relationships*" shows the total number of predicted gene relationships. '*Correctly Identified Relationships*' specifies predicted relationships that have been established in yeast cell cycle regulation as a direct influence. '*Misdirected Relationships*' represents a gene relationship that is predicted to be in the reverse order of a known relationship. '*Accuracy*' is the percentage of correctly predicted relationships out of the total number of known regulator-target relationships. '*Computation Time*' is the running time of the DBN model.

Table 4.3: Dataset alpha38, includes transcription levels of 150 genes with a sampling interval of 5 minutes and a total of 22 time points.

Method	Total Identified Relationships	Correctly Identified Relationships	Misdirected Relationships	Accuracy (%)	Computation Time
GRN_{Murphy}	671	12	2	4	8 hrs 12 mins
GRN_{Zou}	222	3	0	0.33	6 hrs 53 mins
GRN_{Phase}	609	17	3	5.67	2 hrs 43 mins 31secs

Table 4.4: Dataset alpha30, includes transcription levels of 150 genes with a sampling interval of 5 minutes and a total of 22 time points.

Method	Total Identified Relationships	Correctly Identified Relationships	Misdirected Relationship	Accuracy (%)	Computation Time
GRN_{Murphy}	752	8	4	2.67	8 hrs 3 mins
GRN_{Zou}	228	4	0	1.2987	5 hrs 24 mins
GRN_{Phase}	786	15	11	5	2 hrs 40 min

It should be noted that the number of true connections learned by GRN_{Phase} has increased for both datasets in comparison to GRN_{Murphy} (Murphy and Mian 1999) and GRN_{Zou} (Zou and Conzen 2005). One reason for such an improvement is the application of biological domain knowledge in deciphering the problem space. However, the number of correctly identified relationships by all three models is very low. We can speculate some reasons behind such low accuracy. The Primary reason is the lack of known regulator-pair connections to verify the inferred interactions. Second, the level of noise present in the experimental data may have an adverse effect on the potency of the methods. Finally, since we have used a fraction of the genes of the complete GRN, the absence of some important genes may cause not only the loss of true connections but also the inference of false positive connections.

In conjunction with accuracy, the computational time of our proposed model, GRN_{Phase} has improved remarkably in comparison with GRN_{Murphy} and GRN_{Zou}. Of the three models compared, GRN_{Murphy} is the most computationally expensive. The reason is that the method checks all possible combinations of gene regulators for finding the optimal network structure. Though GRN_{Zou} cuts down the number of candidate regulators of a target gene

before learning, their predefined threshold for up/down regulation is ineffective to find most of the potential regulators. We speculate that the lack of biological foundation in defining the threshold is the main reason for such a failure. In contrast, GRN_{Phase} uses biological domain knowledge to restrict the number of candidate regulators of a target gene and is able to learn GRN with limited time which is twice as fast as that of GRN_{Zou}.

4.6 Conclusion

In this chapter, we have exploited some biological features of the cell cycle to achieve two goals. These are: 1) to improve the number of correctly predicted regulator-target relationships and 2) to reduce the computational time of the DBN structure learning algorithm. We have assumed that a gene is maximally expressed to affect and control the regulation of other genes and on completing certain tasks; it is in turn repressed by some other regulator genes. Thus the whole cell cycle is progressed systematically through the successive activation and inactivation of CCR genes. In order to employ this assumption in our study, we have calculated the peak times of individual genes which fall into one/more phases of the cell cycle. Therefore, genes that peak in the interval of the same phase of the cell cycle have been grouped together. Finally, we have applied the DBN algorithm within distinct phases and performed an exhaustive search within a group to find the regulator-target pairs. The reason for this exhaustive search is that genes are usually activated by regulators with earlier or simultaneous peak times but their repressors may be maximally expressed at later time points.

The phase-specific grouping of genes has discarded irrelevant regulators from the candidate parent list of a target gene and enhanced the probability of predicting true positive connections between genes. In addition, the reduction in search space has made the learning

algorithm computationally feasible. The run time of our model (GRN_{Phase}) is one-third in comparison to Murphy's DBN approach (GRN_{Murphy}) and half of GRN_{Zou}. In terms of accuracy, GRN_{Phase} does not outperform GRN_{Murphy} remarkably; however, the number of true positive connections has been increased with GRN_{Phase} for both the datasets and the accuracy could have been further improved by applying the k-skip model to validate the inferred interactions.

It is worth mentioning that both our expression profiles are complete, that is, there are no missing values or hidden variables. However, they do not represent the exact picture of microarray data. Most of the available microarray datasets have thousands of missing values. As we are computing the peak times of individual genes from their expression profiles, our proposed approach cannot be directly applied to incomplete datasets. Hence, it requires predicting the missing values from incomplete data by using algorithms such as nearest neighbor averaging.

In conclusion, our study in this chapter is a small step towards constructing the whole GRN of the yeast cell cycle. We extend our work in the following chapters to further improve the performance of the proposed GRN model by utilizing biological features of gene regulation and other sources of available biological data such as Protein-Protein Interaction (PPI) data.

CHAPTER 5

CO-REGULATION OF CO-EXPRESSED GENES[1]

The main focus of this chapter is to identify groups of co-expressed genes and their underlying co-regulation mechanism from the experimental gene expression data. A partitioning algorithm is employed to group genes into k optimal number of clusters with the intention that genes in the same cluster have similar expression patterns over time. The algorithm also finds a representative gene for each of these clusters, known as mediods. The purpose of this partitioning is to mitigate the number of target genes as well as their potential regulators in estimating the underlying co-regulation network. Nevertheless, in order to infer the complete network of gene regulation, we combine the network of co-expressed genes with the network estimated in chapter 4. Through the analysis of experimental microarray data, we demonstrate that the combined model is computationally more efficient and topologically accurate in inferring gene regulation networks compared to other existing models. To study the performance of the proposed model we analyze the same microarray datasets of the yeast cell cycle as previously used in chapter 4 with 50 more added genes.

---

[1] This chapter presents the results reported in a published journal paper (Shermin and Orgun, 2010) and a conference paper (Shermin and Orgun, 2009b).

5.1 Introduction

Over the past few years, several lines of evidence suggest that similar patterns in gene expression profiles signify relationships between genes (Eisen et al. 1998, Marcotte et al. 1999). By definition, genes having similar expression pattern across a set of samples, are termed as co-expressed, and genes which are regulated by common transcription factors (TFs) are known as co-regulated. In genomic studies, the analysis of microarray gene expression data facilitates the exploration of co-expression patterns and locates groups of co- transcribed genes. There are several points of interest in the identification of co-expressed genes. Firstly, several studies (Eisen et al. 1998, Spellman et al. 1998) suggested that many functionally related genes are co-expressed. Hence, grouping genes with similar expression patterns could reveal the function of previously uncharacterized genes. Secondly, co-expression may reveal insight into gene regulation. For instance, if a TF regulates the expression of two genes, then we might expect the genes to be co-expressed. Therefore, there is likely to be a relationship between co-expression and co-regulation. This chapter focuses on exploring this likely relationship and investigating its contribution towards estimating transcriptional circuitry.

In eukaryotes, the transcriptional regulatory mechanism underlying co-regulation of multiple genes is exceedingly complex (Brazhnik et al. 2002). Nevertheless, the coordinated regulation of the expression of co-expressed genes can take place at different levels, such as transcription (Mootha et al. 2004), or translation (Enriquez et al. 1999, Di Liegro et al. 2000). In this chapter, we focus on the coordinated regulation of genes at the transcription level.

In general, gene expression data obtained by high-throughput microarray experiments is organized in coherent groups of genes using different statistical methods such as hierarchical clustering, self-organizing maps, *K*-means clustering, principle component

analysis etc. Several good reviews of the popular clustering algorithms for gene expression data can be found in Sherlock et al. (2000) and Sharan et al. (2002). Most of these approaches identify clusters of co-expressed genes that demonstrate similar expression patterns over time. In this chapter, we employ a partitioning algorithm known as Partitioning Around the Mediods (PAM) to identify groups of co-expressed genes. Initially, the algorithm computes k optimal number of clusters from the dataset and then groups the genes into the k clusters. Each of these k clusters has a representative gene, known as mediods, which can be defined as the center point of the cluster and whose average dissimilarity to all the genes in the cluster is minimal. Finally, a DBN structure learning algorithm is applied among the mediod genes to discover the coordinated regulation of the co-expressed genes. For the discovery of the entire gene regulation mechanism, the co-regulation network of co-expressed genes is merged with the network of individual gene as discussed in chapter 4.

5.2 Related Work

Cluster analysis is the most popular exploratory technique for pattern discovery and to identify groups of genes with similar expression patterns as well as identifying regulatory factors (Tavazoie et al. 1999, Heyer et al. 1999, Geiss et al. 2003, Ohler et al. 2001, Wolfsberg et al. 1999, Jelinsky et al. 2000). Different variations of cluster analysis such as hierarchical (Pietrokovski et al. 1996, Hughes et al. 2000) and Bayesian (Qin et al. 2003) have been also adapted to discover co-regulated genes and their transcription factor binding sites. However, the use of cluster analysis to identify co-expressed genes or biological function has its own limitations. In particular, clustering of biological data always returns clusters independent of biological relevance. Most importantly, microarray data can be quite noisy, and cluster analysis of such data may find patterns in noise.

Yeung et al. (2004) studied the effectiveness of cluster analysis in finding co-regulated genes from co-expressed genes. They identified several factors that affect the likelihood of finding co-regulated genes. These are: 1) the number of microarray experiments in the datasets, 2) the clustering algorithm used and 3) the diversity of experiments in a microarray dataset. They concluded that the ability to identify co-regulated genes from clustering results, is strongly dependent on the number of microarray experiments used in the analysis

Yu et al. (2003) employed several data sources to create an extensive map of the transcriptional regulatory network, comprising 180 TFs with their respective target genes. By integrating this network with gene expression data, they found that genes regulated by the same TF tend to be co-expressed and the degree of co-expression is proportional to the number of TFs they share. To investigate how co-regulation corresponds to co-expression, Zhang et al. (2004) retrieved regulator-target gene pairs from the Yeast Promoter Database and investigated the expression profiles of target genes with the same TF. Their observation suggested that a regulator can be functional over a certain span of time during cell development and therefore genes may be partially co-regulated, such as during a specific phase of the cell cycle.

Veerla et al. (2006) introduced a new clustering method, known as promoter clustering, which groups the promoters with respect to their high scoring motif content. Their study showed that promoter clustering greatly improves the identification of shared transcription factor binding sites (TFBS) in co-expressed genes.

Noort et al. (2003) claimed that the knowledge of gene co-expression contributes weakly in predicting functional interactions. To be able to predict gene functions, the authors combined conserved co-expression with homology data. In a more recent study, Waveren et al. (2008) validated the previously known fact that the energy producing genes in eukaryotic

cells are co-expressed at the transcription level. Nevertheless, their finding suggested that there might be an intricate mechanism of co-regulation of these co-expressed genes at the mRNA level.

From the literature of these related works we conclude that the identification of co-expressed genes may contribute to the estimation of gene regulation at the transcription level.

5.3 Background

In this section, we explore the relationships among the genes by plotting their expression profiles and observing the patterns of their expression. The expression profiles are taken from the same dataset as used in chapter 2.

5.3.1 Co-expression of Co-regulated genes in yeast cell cycle

Gene expression data generated by the high-throughput technologies such as cDNA microarray provides a rich source for the identification of co-expressed genes. The study of Yu et al. (2003) suggested four different types of temporal relationships between co-expressed genes. These are: correlated, time shifted, inverted and inverted time-shifted. In order to investigate co-expression among the genes in our experimental dataset, we have plotted the expression profiles of target genes which have the same regulator as shown in Figure 5.1. The regulation association among the target genes and corresponding TF(s) is derived from the Yeastract database (Teixeira et al. 2006, Monteiro et al. 2008). For instance, REB1 (YBR049C) is a TF of the yeast cell cycle which regulates the expression of two other CCR genes, CDC9 (YDL164C) and PGK1 (YCR012W). As illustrated in Figure 5.1A, the target genes show similar expression over time and are assumed to be co-expressed. Time shifted relationship is exhibited by G1/S phase TFs, SWI4 (YER111C) and SWI6 (YLR182W),

which are regulated by the same regulator STE12 (YHR084W). The temporal relationship of these co-expressed genes is illustrated in Figure 5.1B.

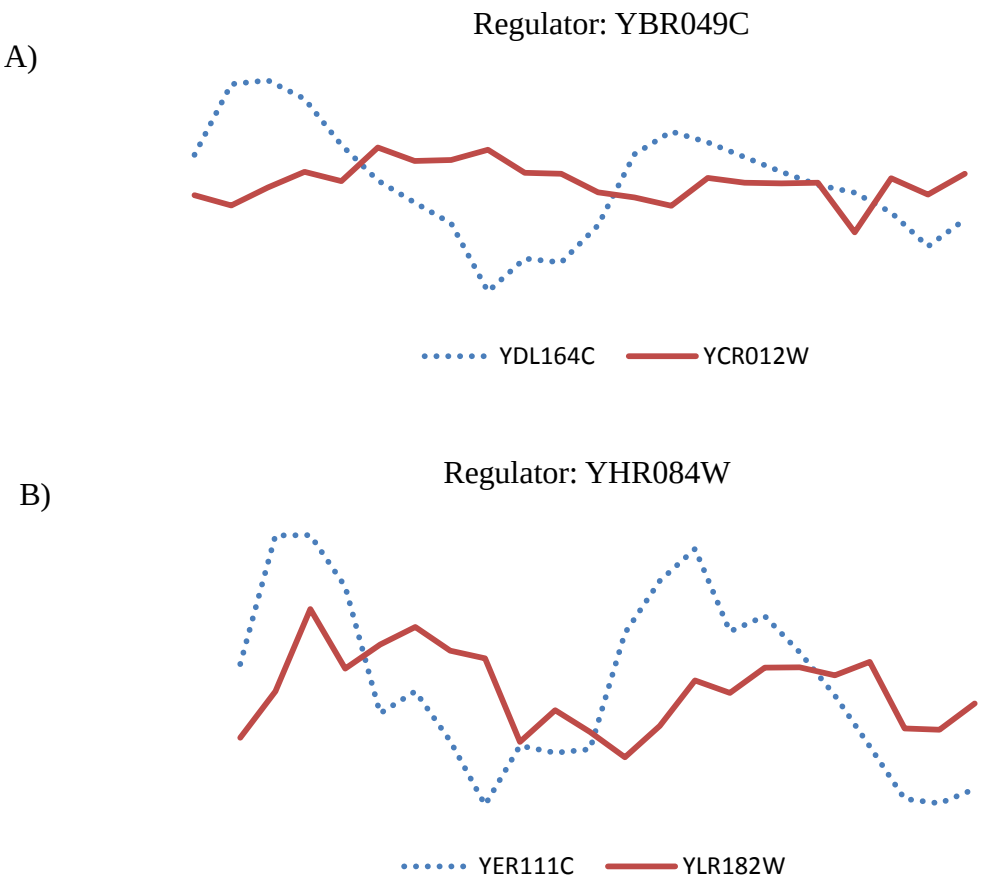


Figure 5.1: Expression profiles of co-regulated genes. A) Co-expressed B) Time Shifted Co-expressed

In a complex regulation network, a target gene can be regulated by multiple TFs as well as multiple genes can share a set of regulators. When genes share common regulators, they are likely to show similar expression patterns. For example, gene YHP1 (YDR451C), HCM1 (YCR065W) and GIN4 (YDR507C) share regulators SWI6 (YLR182W) and MBP1 (YDL056W) during the cell division process in budding yeast. We have plotted the expression profiles of these co-regulated genes in Figure 5.2, which exhibits partial co-expression over time.

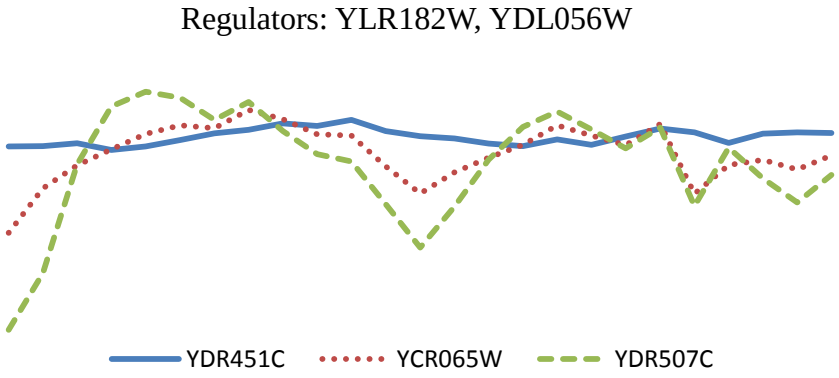


Figure 5.2: Expression profiles of co-regulated genes groups.

5.4 Methods

In this chapter, we address the high dimensionality problem associated with the gene regulatory network (GRN) reconstruction from experimental data with the perception of co-expressed and co-regulated genes. Although clustering is a widely used technique for the identification of co-expressed genes, we have employed a partitioning algorithm to group genes as co-expressed. Finally, a simple DBN structure learning algorithm is applied to estimate the co-regulation of these co-expressed genes. In order to infer a complete GRN model, we integrate the GRN (GRN_{phase}) as discussed in chapter 4 with the network of co-expressed genes. The framework of the GRN model, that estimates co-regulation of co-expressed genes, is shown in Figure 5.3. As the figure shows, the proposed model is a discrete model and the data is discretized with the same method as discussed in chapter 4.

5.4.1 Partitioning Around the Mediods (PAM)

Partitioning Around the Mediods (PAM) is a partitioning algorithm which clusters the data around an optimal number of representative objects called centrotypes or mediods. A mediod is the object of the cluster for which the average dissimilarity matrix to all other

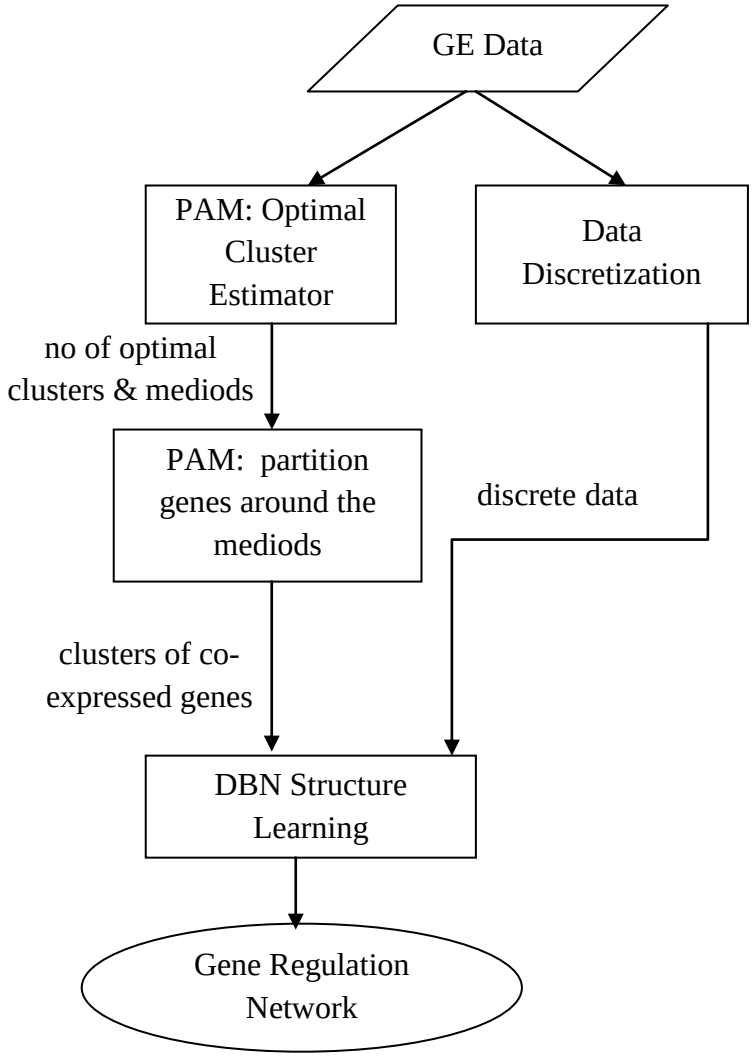


Figure 5.3: Framework of the proposed GRN model.

objects in it becomes minimal. PAM has a few distinct advantages when compared to other well-known k-means clustering algorithms (R core Team 2009). These are: 1) the algorithm computes the dissimilarity matrix from the given dataset and clusters genes based on thismeasurement; 2) the algorithm shows more robustness because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances in placing genes into different clusters and 3) it computes an optimal number of clusters from the data. In practice, PAM is embedded in many statistical analysis systems, such as SAS, R, S+.

The algorithm's main phase consists of finding the k optimal number of clusters in the dataset. After finding an initial set of mediods for each cluster, each object of the data is grouped with the nearest mediod. That is, object i is placed into cluster A when mediod m_A is nearer than any other mediods m_B, that is, d (i, m_A) ≤ d (i ,m_B) for all B = 1,….,k, where d(i, j) is the dissimilarity measure between objects i and j. The k representative objects should minimize the objective function in (5.1), which is the sum of the dissimilarities of all objects to the mediod

$$Objective\ function = \sum_i d(i, m_i) \tag{5.1}$$

The algorithm proceeds in two steps:

1. BUILD-step: This step sequentially selects k "centrally located" objects, to be used as initial mediods.
2. SWAP-step: If the objective function can be reduced by interchanging a selected object with an unselected object, then the swap operation is carried out. This is continued till the objective function can no longer be improved.

5.4.2 Identification of co-expressed genes with PAM

The Partitioning Around the Mediods (PAM) cluster algorithm is applied on the gene expression data to find groups of co-expressed genes which have similar expression patterns. The clustering process is achieved in two steps: 1) use the algorithm to predict the k optimal number of clusters into which the data will be split and 2) group the dataset into k clusters in a way that is similar to the k-means clustering algorithm.

Since patterns showing by the gene expression are random in nature, it is challenging to define an appropriate similarity measure for clustering. In order to deal with the random behavior of gene expression we have transformed each gene profile using the function in

(5.2). With this transformation, the expression profile of each gene is represented in a uniform pattern, where $X_i(t) = 2$ means over-expression; that is, the expression changes between the consecutive time points (t, and t+1) is over the average expression change for gene i. Similarly, $X_i(t) = 1$ signifies down-expression; that is, the expression changes between time point t and t+1 is below the average expression change.

$$X_i(t) = \begin{cases} 2 \quad if \quad abs(X_i(t+1) - X_i(t)) \geq \overline{X_i} \\ 1 \quad if \quad abs(X_i(t+1) - X_i(t)) < \overline{X_i} \end{cases} \tag{5.2}$$

where $X_i(t)$ is the expression level of gene i at time t and $\overline{X_i} = avg(abs(X_i(t+1) - X_i(t)))$.

5.4.3 Inference of regulation among groups of co-expressed genes

In order to identify the activators and repressors of co-expressed genes, we apply the dynamic Bayesian network (DBN) learning algorithm among the mediods of the clusters. Assume that there are n genes in the dataset, grouped into k optimal number of non-overlapping clusters. Each cluster is represented by a mediod gene, m_i, where i = 1..k. Therefore, regulatory connections have been sought among these k mediod genes using the DBN-based GRN model as described in section 4.4.4.

To demonstrate the learning process, assume that, the model identifies m_j and m_k as the potential regulators of a mediod gene, m_i. In the learned GRN, the member genes of the clusters represented by m_j and m_k are considered as the co-regulators of the co-expressed genes which are represented by m_i.

A similar DBN algorithm is applied within the genes of each cluster to find regulation among co-expressed genes.

5.4.4 Merge networks learned by the models

Given n genes in the dataset, the identified causal connections among genes are placed into an n×n matrix, where each row corresponds to a regulator gene. If a gene j is regulated by a gene i , then the cell [i,j] in the matrix contains 1, otherwise 0.

Assume, A and B are two such matrices learned by two different GRN models. In order to merge the networks, we used a Bit-wise OR function given in (5.3)

$$Merged\_Network = BITOR\ (A, B) \tag{5.3}$$

5.5 Experiments and Results

To investigate how effectively our proposed DBN-based model can infer co-regulation of co-expressed gene and to what extent the current challenges can be addressed, we run experiments on two recently published real microarray datasets (Pramila et al. 2006). Initially, we apply the PAM algorithm on the experimental microarray data to find groups of co-expressed genes. The algorithm identifies 52 clusters of the 200 genes and some of these clusters contain one gene only; that is, these genes do not share any similarity in their expression pattern with others. From the resulting clusters of these genes, we find that the partitioning algorithm assigns genes to different clusters which are believed to be co-expressed. For instance, the two cell cycles regulated TFs HCM1 and YHP1 are believed to be co-expressed and co-regulated by the MBF complex (SWI6 and MPB1). However, our PAM algorithm clusters these TFs into two separate clusters. We speculate, the lack of biological relevance in the algorithm is a key factor for such partitioning. Moreover, the experimental drawbacks such as precision in designing the microarray experiments and the inherent noise in the data affect the computation of the dissimilarity function and so the partitioning of the genes.

5.5.1 Experimental data

We analyze two separate microarray gene expression data sets of budding yeast as previously used in chapter 4 with 50 more added genes. The newly added genes are chosen in the same way as discussed in section 4.5.1. This addition of new genes results in a total of 200 cell cycle regulated genes over 22 time points in each of the datasets. We increase the size of the data set in this analysis to investigate the performance of the models in the larger context.

5.5.2 Experimental setup

The experiments are conducted on a computer system with dual core Intel processor (1.83 GHz) and 2 GB RAM, running windows XP (Professional). The free statistical software, R (R Core Team 2009) has been used to implement the PAM clustering algorithm which partitions the datasets into some optimal number of groups. Together with R, we have used Bayesian Net Toolbox (BNT) which is written in MATLAB and freely provided by Murphy (2001a) to construct DBNs. The experiments are setup and run under the MATLAB environment with version 7.6.0.324(R2008a).

5.5.3 Experimental results

In this section, we evaluate a GRN model ($GRN_{2\text{-stage}}$) that learns two regulation networks of overlapping connections in two stages. At the first step, the model finds regulation of individual genes and in the following step, it estimates coordinated regulation of co-expressed genes. Finally, the two overlapping networks are merged together to infer a final GRN that represents the data best. For the verification of the estimated network, we use the same method as discussed in section 4.5.

As stated earlier, the pitfall of working with real data is that we know a very small fraction of the connections in the target network. Hence, in the initial experiment, we apply our 2-stage GRN model to extract a small scale network of 13 known TFs only. These include G1-phase specific TFs (MBP1, SWI4, SWI6), S-phase specific TFs (HCM1, WHI5, YOX1) and G2/M phase TFs (FKH1, FKH2, NDD1, YHP1, MCM1, SWI5 and ACE2). The known target network topology of these TFs is extracted from various sources (Simon et al. 2001, Teixeira et al. 2006, and Monteiro et al. 2008) and is shown in Figure 5.4. As discussed in section 4.5.3, the figure is automatically generated by the MATLAB program.

In the known network, SWI4 is activated by MCM1 during the G1 phase. It then transcriptionally triggers S-phase specific TF YOX1, which subsequently represses MCM1. As a result of this serial regulation, the generation of SWI4 slides down. With the application of our 2-stage GRN approach, we are able to identify this small fragment of regulatory program accurately. Then again, the G2/M phase TF, FKH2 activates the co-expressed genes ACE2 and SWI5 during M-phase to assist with Chromatid separation. Our proposed model can learn these connections correctly as well. The topology of the estimated network is shown in Figure 5.5.

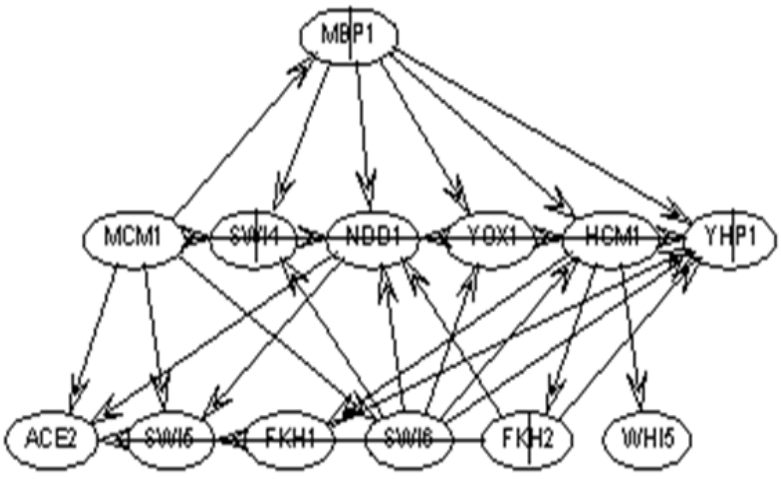


Figure 5.4: The known network structure of the 13 cell cycle TFs. [generated by MATLAB]

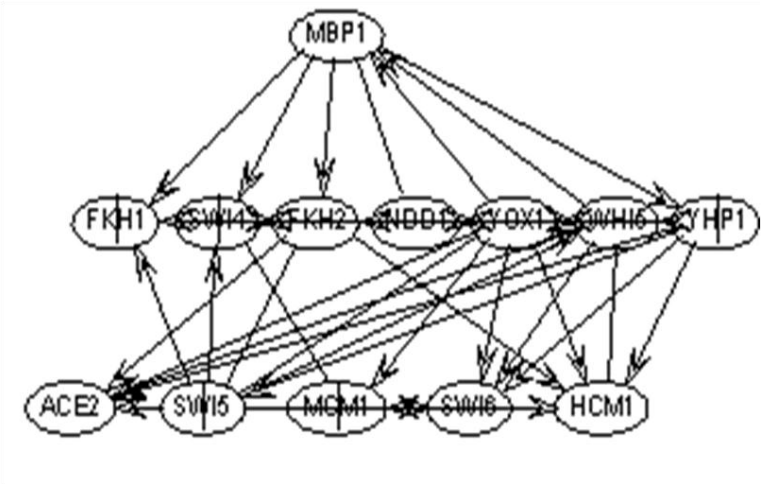


Figure 5.5: The estimated network of the 13 known cell cycle TFs.
[generated by MATLAB]

Of the known 46 true interactions in the target network in Figure 5.4; our model is able to infer 13 connections (28% approx). Moreover, inspecting all the identified connections with respect to the known roles of TFs, it is found that, in most cases, the prediction confirmed the prior knowledge of cell cycle regulation, which establishes the validity of our approach.

Finally, we validate our proposed 2-stage approach in analyzing two large microarrray datasets. Each of these datasets consists of 200 genes over 22 time points. In order to verify the potency of the approach with respect to the current challenges, we compute two criteria, accuracy and computation time for each datasets. 'Accuracy' is the percentage of correctly identified relationships out of the total number of known regulator-target relationships and 'Computation time' is the run time of the whole approach. The experimental results of our proposed 2-stage approach ($GRN_{2\text{-}stage}$) together with other existing DBN-based GRN models (GRN_{Murphy}, GRN_{Zou}, and GRN_{Phase}) on the two datasets have been summarized in Table 5.1 and Table 5.2 respectively. In the tables, 'Correctly Identified Relationships' specifies predicted relationships that have been established in yeast cell cycle regulation as a direct

influence and ‘Total identified relationships‘ shows the total number of predicted gene relationships.

Table 5.1: Dataset alpha30, includes transcription levels of 200 genes with a sampling interval of 5 minutes and a total of 22 time points.

Method	Total Identified Relationships	Correctly Identified Relationships	Accuracy (%)	Computation Time
GRN_{Murphy}	1267	13	1.823	38 hrs 10 mins 19 secs
GRN_{Zou}	288	7	0.982	27 hrs 18 mins 32 secs
GRN_{Phase}	727	21	2.945	14 hrs 15 mins 58 secs
$GRN_{Co\text{-}expressed}$	1139	21	2.945	13 mins 14 secs
$GRN_{2\text{-}stage}$	1782	37	5.189	14 hrs 29 mins 12 secs

Table 5.2: Dataset alpha38, includes transcription level of 200 genes with a sampling interval of 5 minutes and a total of 22 time points.

Method	Total Identified relationships	Correctly Identified Relationships	Accuracy (%)	Computation Time
GRN_{Murphy}	1063	15	2.104	37 hrs 27 mins 11 secs
GRN_{Zou}	277	9	1.262	26 hrs 43 mins 3 secs
GRN_{Phase}	1065	30	4.2076	14 hrs 41 mins 35 secs
$GRN_{Co\text{-}expressed}$	552	30	4.2076	13 mins 16sec
$GRN_{2\text{-}stage}$	1618	52	7.293	14 hrs 54 mins 51 secs

By inspecting the reconstructed networks, it is found that the number of true relationships predicted by our 2-stage approach has increased in comparison with the existing GRN models. The second challenge, the excessive computational cost of inferring GRN has

been addressed remarkably by $GRN_{2\text{-stage}}$ in comparison with GRN_{Murphy} and GRN_{Zou}. We believe that the use of biological knowledge in finding regulators of individual and co-expressed genes is the key to the success of our new 2-stage approach. However, due to the merging of two regulation networks, the number of total relationships inferred by $GRN_{2\text{-stage}}$ is higher compared to the other methods.

5.6 Conclusions

In this chapter, we have applied a partitioning algorithm, Partition Around the Mediods (PAM), on the microarray gene expression data to identify groups of co-expressed genes. The partitioning algorithm groups genes into k optimal number of clusters and then each cluster is represented by a mediods gene. In order to infer the underlying co-regulation mechanism of these co-expressed genes, we have applied the DBN learning algorithm among the mediods genes.

The two main goals of our study in this chapter are: 1) to improve the number of correctly predicted regulator-target relationships and 2) to reduce the computation time of the DBN algorithm. The proposed model has been evaluated through the analysis of two separate datasets of yeast cell cycle along with some existing DBN-based GRN models. Each of these experimental datasets contains 200 cell cycle regulated genes over 22 time points. The partitioning of the genes around mediods and the notion of finding regulation among the mediods has drastically reduced the search space; hence the computation time of the model has significantly improved in comparison to the existing models. The low computation cost of the model has encouraged us to combine the proposed model with the model discussed in chapter 4 to obtain a more complete picture of gene regulation in the yeast cell cycle. The ensemble of these models has also increased the number of true regulator-target gene pairs.

However, the number of total relationships estimated by the combined model is considerably high in compare to the other models. We can speculate some reasons for this high number. Firstly, the partitioning algorithm barely can identify groups of co-expressed genes correctly due to the inherent noise and imprecision of the microarray experiments. This shortcoming has led to the distribution of co-expressed genes into different clusters. Secondly, while applying the DBN algorithm to estimate regulation among the mediods, a good number of irrelevant connections have been inferred due to the incorrect partitioning of the co-expressed genes.

As in chapter 4, we have verified the entire estimated network and investigated how the computation time and the accuracy of prediction have been improved. Although sub-networks analyses might provide insight of individual clusters and the estimated network within the cluster, In this chapter, our main focus is to reduce the problem space using some heuristics and therefore subnetworks analyses have not been conducted. First, our proposed model reduces the problem space by clustering genes into groups and then, finds the regulation network among the mediods of the clusters. We have also investigated the effect of these two step reductions in the problem space on the computation time and the overall prediction accuracy of the model.

In conclusion, our study in this chapter is a small step towards constructing the whole GRN of the yeast cell cycle. However, we can extend our work in different ways. Firstly, in our analysis; we have used the mean of the expression change to discretize the time series, and to define the similarity matrix in identifying the groups of co-expressed genes. This measure may have a diverse effect on the identification of groups of co-expressed genes because of the outliers in the data. The incorporation of more biological driven knowledge can improve the performance of the partitioning algorithm. Secondly, we can redefine the scoring

function, which is the key to finding the regulator-target relationships. Finally, we plan to utilize other sources of biological data such as Protein-Protein Interaction (PPI) data to restrict the number of potential regulators of a target gene in our future models.

CHAPTER 6

TRANSCRIPTIONAL REGULATION FROM MULTI-SOURCE DATA[1]

This chapter focuses mostly on addressing the data scarcity problem associated with the computational reconstruction of gene regulatory networks (GRN). The inadequate amount of experimental conditions compared to the number of random variables is considered to be one of the major obstacles in discovering transcriptional regulation with high accuracy. In addition, the high complexity in the gene regulation mechanism makes the application of available statistical and machine learning methods infeasible. In this chapter, we study a model based on dynamic Bayesian networks to predict transcriptional regulation by integrating transcription factor binding site data (TFBS) and protein-protein interaction (PPI) data with gene expression data. The knowledge of genetic interactions between proteins and the presence of transcription factors binding site at the promoter region of a gene are utilized to restrict the number of potential regulators of each target gene. We demonstrate the effectiveness of combining multiple data sources in predicting transcriptional regulation through the analysis of yeast cell cycle data.

---

[1] This chapter presents the results of a conference paper (Shermin et al. 2011) published in the Proc. of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine.

6.1 Introduction

In the post-genomic era, a diverse collection of high-throughput biological databases are available which provide a rich source for the study of the underlying cellular processes. These databases include gene expression profiles, protein-protein interaction (PPI), transcription factor binding site (TFBS), pathways, yeast two-hybrid experiments etc. In a living cell, PPI occurs when two or more proteins bind together to accomplish a certain biological function. For instance, the DNA replication during the S-phase of the cell division cycle are carried out by large molecular machines that are built from a large number of protein components organized by their PPI. Therefore, PPIs are at the core of the entire interaction network of any living cell. The TFBS database stores the transcription factors (TF) and their corresponding DNA sequences that they interact with. This data provides evidence about potential gene regulation and is a relevant source for the study of gene expression at the transcriptional level.

When estimating a network from gene expression data alone, a common problem is that the number of time points in the data, is limited compared to the number of random variables in the network which makes the estimation task difficult. In this chapter, we address this challenge by incorporating multiple sources of biological data with the microarray gene expression data.

The basic structure learning algorithm of DBN usually explores every combination of genes as potential regulators of a target gene. Therefore, the search space grows exponentially with the number of genes included in the dataset. In addition, because of the inherent noise and imprecision of microarray data, the inference algorithms can hardly find the true regulator of a target gene. As discussed in the previous chapters, one way of addressing these fundamental problems is to narrow down the number of potential regulators of a target gene

by leveraging domain knowledge. In this chapter, we use the two sources of biological data, PPI data and TFBS data, to restrict the number of potential regulators of a target gene. Of the two types of known interactions between proteins/genes, we assume that the genetic interactions could carry some biological evidence of one interactor being regulated by the other. Then again, TFs which have at least one binding site in the promoter region of a target gene are considered as potential regulators along with the genetic interactors of the gene. In summary, we decompose the entire network into several sub-networks; each of these sub-networks contains a specific gene and its potential regulators. Finally, the DBN structure learning algorithm is applied to learn the candidate regulators of the target gene from the corresponding biologically driven sub-network.

The incorporation of these biological data proves to be quite complementary to expression data, since it dramatically decreases the amount of expression data needed to discover regulatory networks by DBN models. This chapter aims to investigate the effectiveness of integrating other sources of biological data in reconstructing GRN from expression data. In addition, the proposed model is capable of predicting missing values using the nearest neighbor averaging algorithm. Therefore, in contrast to study in the previous chapters, we test the model on a popular dataset of the yeast cell cycle which has a reasonable amount of missing values in it.

6.2. Related Work

Though most of the available reverse engineering approaches rely solely on the temporal gene expression data only, an increasing number of recent works have paid attention on incorporating multiple sources of data for reconstructing GRN.

Hartemink et al. (2002) introduced a new notion in the process of reconstructing

GRNs by combining multiple sources of experimental data. In their study, they used the location data to influence prior of the model and learnt the structure of the network from the gene expression data. They discovered a network of thirty three genes which is involved in yeast pheromone response and is consistent with the current understanding regarding this regulatory network.

Tamada et al. (2003) studied a statistical method for estimating gene networks and detecting promoter elements simultaneously. The authors made an assumption that genes which are regulated by a common TF, may share a consensus motif in their promoter regions of the DNA sequences. In an iterative process, their method detects consensus motifs based on the structure of the estimated network, and then re-estimates the network using the result of the motif detection until the network becomes stable. In another study, Segal et al. (2003) incorporated both gene expression and promoter sequence data to estimate the GRN. Their method estimated motif profile from the promoter sequence data and identified regulatory modules in a set of experiments through a common motif profile. Then, the authors used the EM algorithm to refine the motif profile by adding or deleting motifs so as to best explain the expression data as a function of the regulatory modules.

Nariai et al. (2004) proposed a statistical method for inferring transcriptional regulation from microarray data and used PPI data to refine the estimated GRN. In a recent study, Zhang et al. (2007) used dynamic Bayesian network (DBN) with structural Expectation Maximization (SEM-DBN), to model GRNs from both gene expression data and transcriptional factor binding site data. They used the binding site data to introduce the prior knowledge to SEM-DBN model and microarray expression data for structure likelihood. Werhli, and Husmeier (2007) studied a Bayesian approach where they systematically incorporated multiple sources of biological data to estimate a prior distribution over network

structures in the form of a Gibbs distribution. Zhu et al. (2008) demonstrated that the integration of diverse ranges of molecular data, including genotypic, gene expression, TFBS and PPI data, can enhance the predictive power of the GRNs. In a most recent study, Zhang et al. (2010) studied an integrative framework to infer gene regulatory modules from the cell cycle of cancer cells by incorporating multiple sources of biological data, including gene expression profiles, gene ontology, and molecular interaction. They identified network motifs from the molecular interaction data and applied a recurrent neural network model to examine the relationship between TFs and their target gene groups.

We conclude this literature review of related works with the finding that GRNs can be reconstructed from a single source of data such as gene expression data; however, incorporation of other sources of experimental data is believed to be effective in improving the estimation accuracy of the reconstructed networks.

6.3 Background

In this section, we describe the two sources of molecular data that are used in the proposed model. This includes PPI data and the TFBS data. This section also discusses the feasibility of applying these data in reconstructing GRNs at the transcription level.

6.3.1 Protein-protein interaction data

Experimentally identified PPI data have been extracted from the BioGRID database (release 3.1.72) (Breitkreutz et al. 2008). BioGRID is an online interaction repository that compiles data through intensive curation effort. The database searches publications that report raw protein and genetic interactions from a range of organisms such as yeast, human etc. The repository is continually updating to new releases by including the newly reported

interactions. In the database, individual interactions are recorded as binary relationships between two proteins or genes. The database stores both physical and genetic interactions of proteins. The physical interactions include relationships such as the direct physical binding of two proteins, co-existence in a stable complex etc. On the other hand, the genetic interaction refers to the relationship between two proteins or genes where over-expression or deletion of one gene/protein has an impact on the other gene/protein. While searching the BioGRID database with the TF, MBP1 which is involved in the regulation of cell cycle progression from G1 to S phase, 95 unique interactions have been identified. Of these interactions, there are 26 physical interactions and 69 genetic interactions. The transcription co-factor SWI6 has physical interaction with MPB1 in the PPI network which establishes the known fact that these two TFs form a complex to regulate the transcription at the G1/S transition. The other well-known transcriptional activator, SWI4 has genetic interaction with MBP1 in the PPI network. In the yeast cell cycle, SWI4 is an established activator which forms a complex with SWI6 and regulates the expression of genes at the G1/S transition. Since there is no physical interaction between SWI4 and MBP1 in the PPI network, we have assumed that SWI4 can be a potential regulator of MBP1. Figure 6.1 shows the genetic interaction network among the 13 known TFs of yeast cell cycle which is extracted from the BioGRID database. This includes G1-phase specific TFs (MBP1, SWI4, and SWI6), S-phase specific TFs (HCM1, WHI5, YOX1) and G2/M phase TFs (FKH1, FKH2, NDD1, YHP1, MCM1, SWI5 and ACE2). It is worth mentioning that there are a few TFs in the figure which have no genetic interaction with the others such as WHI5 in the PPI network. However, this does not imply that WHI5 has no regulatory contribution at the transcription level. As discussed in section 4.5.3, the figure in 6.1 is generated by the MATLAB program.

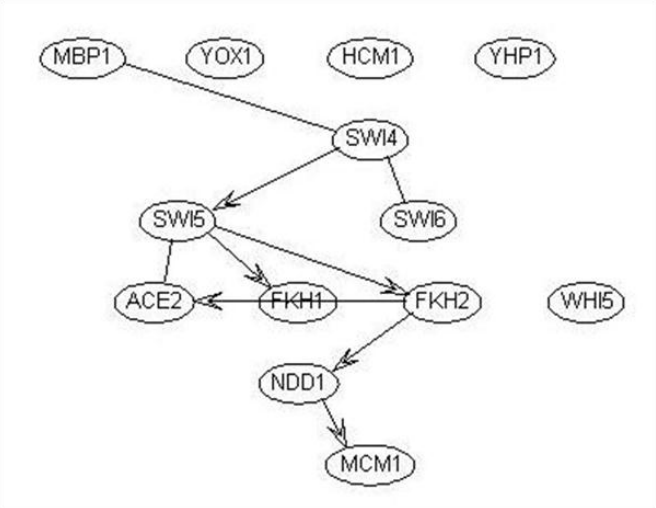


Figure 6.1: Genetic interaction network among the 13 known TFs of yeast cell cycle. The network is extracted from the online interaction repository, known as BioGRID. [generated by MATLAB]

6.3.2 Transcription factor binding site data

The potential regulatory associations between TFs and their target genes in yeast have been extracted from the YEASTRACT database (Monteiro et al. 2008, Teixeir et al. 2006). The database contains regulatory associations between the yeast genes, which are denominated as "documented" and "potential" associations. Despite the fact that the existence of the TF binding site in the promoter region of a gene does not necessarily make it a target of the corresponding TF, we investigate the potential regulatory association and consider a TF as a potential regulator of a target gene if the former has at least one binding site in the promoter region of the latter. While searching the YEASTRACT database with the TF MBP1 as a target gene, the search generates 34 potential regulators and the transcriptional activator, SWI4 is one them. Therefore, both sources of biological data suggest that there is possibly a regulatory effect of SWI4 on the expression of MBP1.

Figure 6.2 shows the potential regulation network among the 13 known TFs of yeast

cell cycle extracted from the YEASTRACT database. Every gene in the figure has one or more regulators, for example WHI5 has three potential regulators which are SWI4, FKH2, and FKH1. Therefore, we can assume that while combining the two potential networks together, all the genes have at least one regulator.

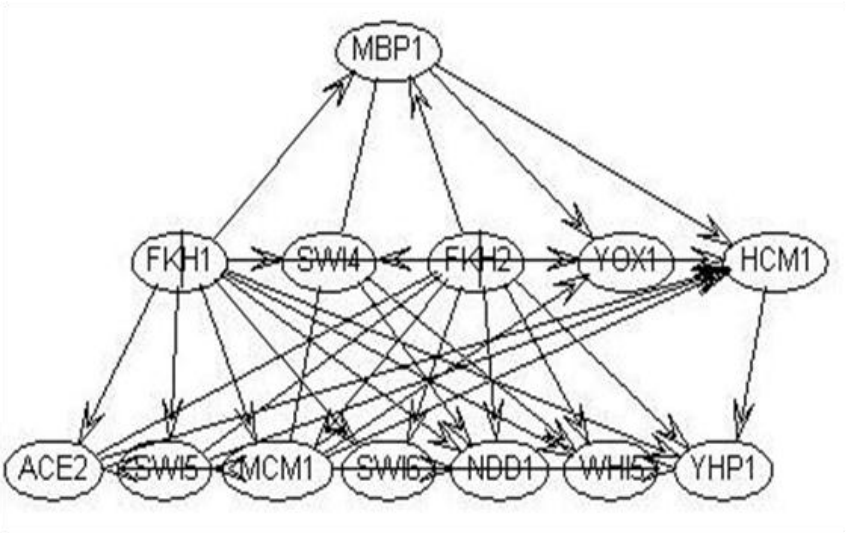


Figure 6.2: Potential regulatory association among the 13 known TFs of yeast cell cycle extracted from the YEASTRACT database. [generated by MATLAB]

6.4 Methods

In this chapter, our goal is to incorporate other sources of biological data in estimating the structure of transcriptional regulation of yeast cell cycle from microarray gene expression data. The microarray dataset contains gene expression levels of thousands of genes over distinct phases of the cell cycle. Given this massive dataset, the structure learning of gene regulation network using the algorithm described in Table 4.1, chapter 4, becomes computationally expensive as the size of Q (the number of potential regulators) grows exponentially with N, the number of genes in the dataset. The algorithm considers each gene in the dataset as a potential regulator of a target gene. However, biological networks are mostly scale-free (Han 2008); they have a few highly connected nodes in the network. In the

context of GRNs, these nodes represent the transcriptions factors and regulate the expression of the majority of the genes. This fact suggests that not all the genes in the dataset are TFs; hence they have no regulatory affect on other genes.

In this chapter, we utilize this feature of gene regulation networks by extracting a set of potential TFs R(i), where card(R(i)) $\leq$ N and R(i) $\in$ N for each gene i by integrating different layers of biological data. This utilization of multi-sources biological data in estimating transcriptional regulation confers an integrative framework of the GRN model as shown in Figure 6.3. The following subsections discuss the methodologies required for the extraction of R(i) and learning the regulation network with DBN. We also discuss the methods adapted in the model to impute missing values in the dataset.

6.4.1 Impute missing values

An illustrious fact about all available microarray data is that they contain missing values. Since the model proposed is only applicable to complete data, we estimate the missing values using nearest neighbor averaging algorithm. To find the k nearest neighbors, the algorithm computes the euclidean distance of a gene with missing values to its neighbors. The search for k nearest neighbors is confined to the columns for which that gene has no missing values. Each candidate neighbor may also miss some of the expression values used to calculate the distance. In cases where the candidate neighbor gene has missing values, an average distance is computed from the non-missing expression values of the neighboring genes. Having found the k nearest neighbors of a gene, the missing expression values are imputed by averaging those of its neighbors. This algorithm fails for predicting missing values of those genes that have all the neighbors missing.

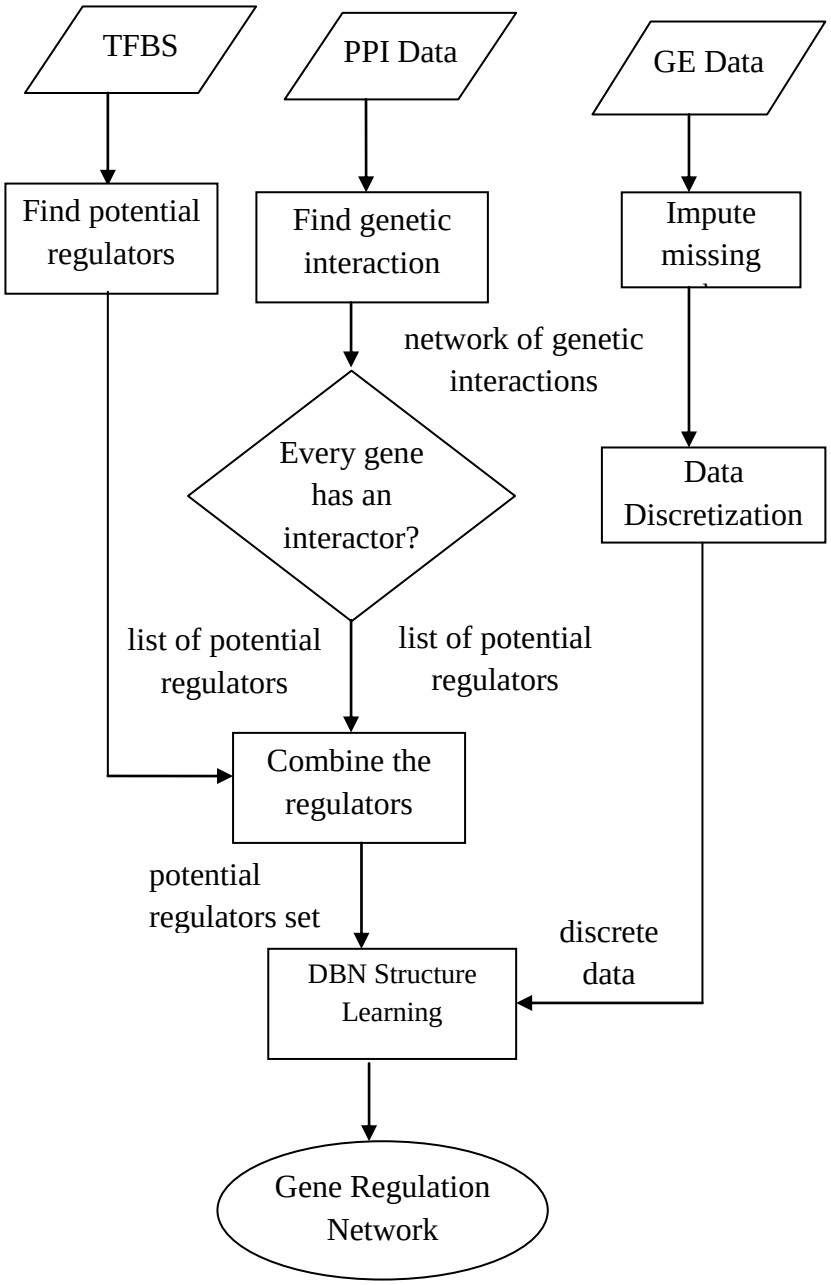


Figure 6.3: Framework of the proposed GRN model. The model incorporates two sources of biological data in the learning process.

6.4.2 Extraction of potential regulators of genes

As mentioned in section 6.3, two sources of biological data are used to extract the potential regulators of a target gene. These are TFBS and PPI data. We assume that the

interactors may have a regulatory affect on each other, although, none of these two sources of interaction data show any direct evidence of gene regulation. This assumption has significantly reduced the number potential regulators of a target gene, without sacrificing accuracy, which consequently contributed towards a scalable model for estimating GRN.

For each gene i in the dataset D, we extract a regulator set $R_{BS}(i)$ from the binding site data where each $r \in R_{BS}(i)$ has at least one binding site in the promoter region of i. As described in Section 6.3, there are two types of interactions between proteins published in the BioGRID database (Breitkreutz et al. 2008). In this study, we assume the genetic interaction between two proteins as a possible regulatory interaction. For each gene another potential regulator set $R_{PPI}(i)$ has been extracted, where each $s \in R_{PPI}(i)$ is a genetic interactor of i and $s \in N$. Finally the two potential regulator sets are combined to generate a potential regulator set R(i), where $R(i) = R_{BS}(i) \cup R_{PPI}(i)$.

6.4.3 Learning the structure of GRN

Given that there are N genes in the dataset D and for any gene i, the regulator list contains m potential regulators, that is Card(R (i)) = m. As any combination of the potential regulators may regulate the expression of the target gene i, the total number of potential regulators to consider is 2^m. However, for a large m, the search space is enormous. To deal with this dimensionality problem, we further restrict the fan-in (the number of input edges) of each node in the network to k (**<<**m). As a consequence, the size of the search space is reduced to ${}^{m}C_{k}$.

These two levels of restrictions on the number of potential regulators of a target gene has facilitated the structure learning algorithm described in Table 4.1 to be applicable on a

dataset containing hundreds of genes.

6.5 Experiments and Results

To demonstrate the effectiveness of integrating multi-sources data at the different level of the proposed model, we apply it on a popular experimental dataset. The detailed experimental setup including data and the results are presented in the following subsections.

6.5.1 Experimental data

In this chapter, we have analyzed the real time-course dataset of yeast cell cycle (Spellman et al. 1998). We choose to work with this dataset because of its incompleteness; that is there are missing values in it. The dataset contains gene-expression measurements of the mRNA levels of 6178 Open reading frames using four different cell synchronization methods: cdc15, cdc28, alpha factor and elutriation with 24, 17, 18 and 14 time points respectively. Though cdc15 dataset has the maximum number of time points, we have chosen the alpha-factor dataset because of the fewer missing values in it. As discussed in section 6.4.1, the nearest neighbor averaging algorithm is used to impute the missing values in the dataset. However, the cost of this imputation can be excessive for a large number of genes in the dataset with missing values.

6.5.2 Experimental setup

The experiments are conducted on a computer system with dual core Intel processor (1.83 GHz) and 2 GB RAM, running windows XP (Professional). To estimate the missing values in the dataset, we have used the knn.impute function of the free statistical software, R (R Core Team 2009). Together with R, we have also used Bayesian Net Toolbox (BNT)

which is written in MATLAB and freely provided by Murphy (2001a) to construct DBNs. The experiments are setup and run under the MATLAB environment with version 7.6.0.324 (R2008a) and R 2.13.1.

6.5.3 Experimental results

We run two separate experiments in a similar fashion as described in the previous chapters to illustrate the effectiveness of the proposed GRN model. The first experiment involves the estimation of a small-scale network including 13 TF only. In experiment 2, we compare the performance of the proposed model with the models discussed in chapters 4 and 5 through the analysis of the benchmark dataset containing 250 cell cycle regulated genes over 18 time points. We increase the size of the dataset by 50 more genes in compared to the previous chapter to test the GRN models in estimating relatively large-scale networks. The known regulation network among these genes is extracted from multiple sources (Simon et al. 2001, Teixeira et al. 2006, Monteiro et al. 2008, KEGG 2000, Kanehisa et al. 2006, Kanehisa et al. 2008). For the verification of the estimated networks, we use the same method as discussed in section 4.5. The detailed description of the experiments and results are discussed in the following subsections.

6.5.3.1 Experiment 1

In experiment1, the model is tested on a small segment of the dataset, including 13 TFs that are known to be involved in cell cycle transcription of budding yeast. The target network among these TFs is shown in Figure 6.4. In the figure, an arrow from gene i to gene j means a direct influence of i on j and a line indicates bidirectional influence.

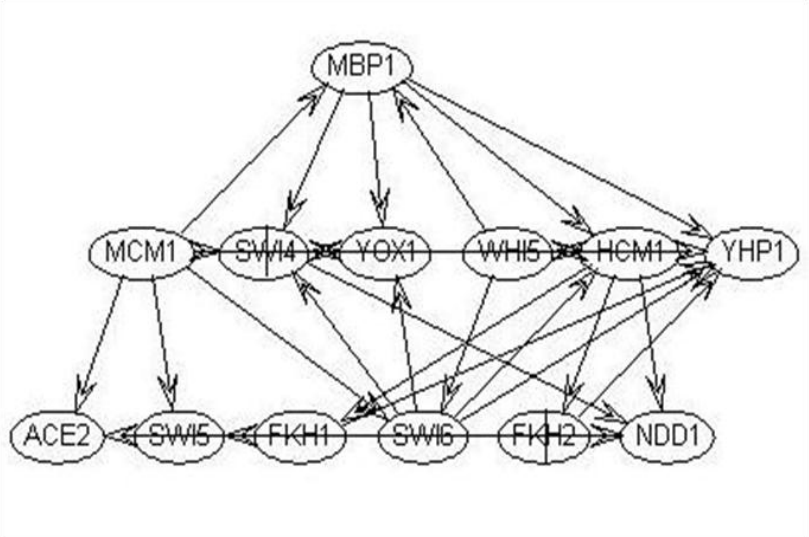


Figure 6.4: Known network of the 13 known TFs in the yeast cell cycle.
[generated by MATLAB]

In estimating the transcriptional regulation, the model first extracts the potential regulation network among these TFs from multiple data sources as discussed in section 6.3. The derived network structure among the TFs is shown in Figure 6.5. At the next step, the model applies DBN learning algorithm to refine the potential regulation network by observing the available gene expression data. Figure 6.6 illustrates the final network structure estimated by our proposed model.

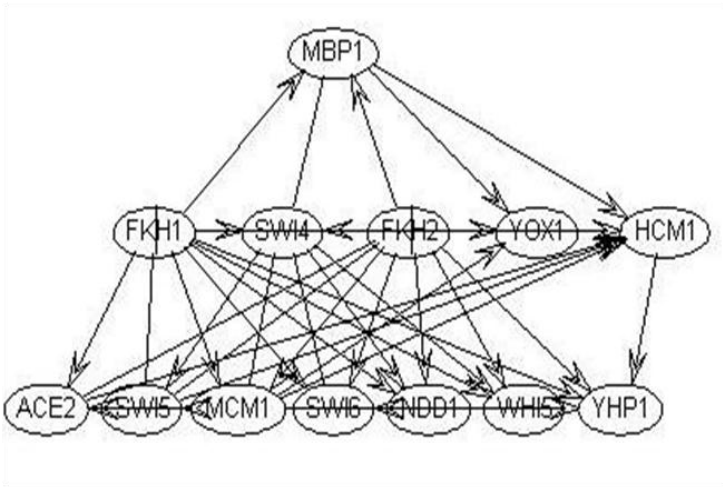


Figure 6.5: Derived network among the 13 known TFs in the yeast cell cycle.
[generated by MATLAB]

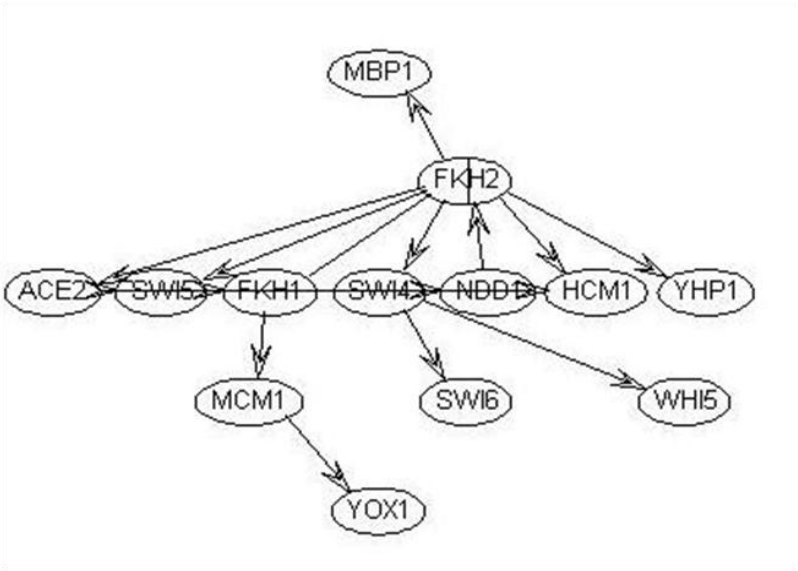


Figure 6.6: Estimated network among the 13 known TFs in the yeast cell cycle. [generated by MATLAB]

Inspecting the estimated network, it can be found that only 8 direct interactions have been correctly identified by our model of the 43 known relationships. However, the incorporation of multiple sources data at the different layer of the model has narrowed down the search space briskly and the model has estimated only 20 connections as true positive. The G1-phase specific TF, MBP1 is a regulator of the other TF SWI4 in the target network. This regulatory relationship has been correctly derived in the potential network structure from different data sources. However, our model fails to infer this relationship as true positive by observing the gene expression data. We speculate that the inherent noise in the available data is the key factor to such a wrong estimation. Nevertheless, the correct derivation of key relationships in the potential network inspires us to integrate multiple sources of biological data in reconstructing GRN.

6.5.3.1 Experiment 2

In experiment 2, we apply our proposed GRN model ($GRN_{Multi\text{-}sources}$) together with the other models (GRN_{Phase} and $GRN_{Co\text{-}expressed}$) on a large dataset. We also include two

existing DBN-based GRN models (GRN_{Murphy}, GRN_{Zou}) in this analysis. Two performance evaluation criteria, accuracy and the computation time are computed for evaluating the performance of the models. The experimental results of applying these models have been summarized in Table 6.1. In the table, 'Total identified relationships' shows the total number of predicted TF-gene relationships. The 'Correctly Identified Relationships' specifies predicted relationships that are established in the yeast cell cycle regulation. 'Accuracy' is the percentage of correctly identified relationships out of the total number of known regulator-target relationships 'Computation Time' is the running time of the analysis.

Table 6.1: The dataset includes transcription levels of 250 genes with a sampling interval of 7 minutes and a total of 18 time points.

Method	Total Identified relationships	Correctly Identified Relationships	Accuracy (%)	Computation Time
GRN_{Murphy}	436	12	1.548	3 days 17 hrs 37 mins 9 secs
GRN_{Zou}	250	0	0	3 days 2 hrs 17 mins 23 secs
GRN_{Phase}	1835	18	2.323	22 hrs 29 mins
$GRN_{Co\text{-}xpressed}$	2309	17	2.193	17 hrs 59 mins 55 secs
$GRN_{Multi\text{-}sources}$	551	106	13.678	1 min

The first two rows in Table 6.1 represent the networks estimated by two existing DBN-based GRN models, GRN_{Murphy} (Murphy and Mian 1999) and GRN_{Zou} (Zou and Conzen 2005). The third and fourth row represent the models ($GRN_{Co\text{-}expressed}$, $GRN_{Multi\text{-}sources}$) discussed in chapters 4 and 5 of this thesis. The bottom row represents the network estimated by the model ($GRN_{Multi\text{-}sources}$) proposed in this chapter. The result shows that the integration of binding site and PPI data has a drastic effect on the computation time, which learns the

network of 250 genes within a minute whereas the existing models take 4 days, 3 days, 1 day and 17 hrs respectively.

In the target network, there are 775 established yeast cell cycle relationships extracted from multiple sources (Simon et al. 2001, Teixeira et al. 2006, Monteiro et al. 2008, KEGG 2000, Kanehisa et al. 2006, Kanehisa et al. 2008). Our proposed model, $GRN_{Multi\text{-}sources}$ identifies a total of 551 connections in the predicted network and 106 of them are correct relationships, which gives approximately 14% accuracy of the model. In contrast, most of the existing models show very low accuracy with much higher computation time. In particular, when GRN_{Zou} is applied on this benchmark dataset, the method fails to predict any true connections.

6.6 Conclusions

In this chapter, we have studied a GRN model through the analysis of real microarray data of yeast cell cycle. We have shown how the integration of multi-data sources such as TFBS and PPI data can contribute to the prediction of transcriptional regulation. The experimental results show that the integration of biological domain knowledge has removed extraneous genes from the potential regulator sets. As a consequence, the model is able to estimate the transcriptional regulation with relatively higher accuracy in a very short time. This outcome also establishes the proposed model as a competitive method for the global analysis of transcriptional regulation

Despite the fact that the integration of multi-data sources enriches the estimation of transcriptional regulation, it has some pitfalls. The PPI data does not provide any regulatory evidence between them. Likewise, genes having binding sites of other genes in their promoter region may or may not have any regulatory connections. Therefore, such restriction on the

search space may exclude some true regulators from the potential regulator set. Most importantly, both PPI and TFBS data are susceptible to noise as they are collected through different biological experiments. These limitations pose a critical challenge on incorporating available sources of experimental data in the reconstruction of GRNs.

The GRN model proposed in this chapter predicts the missing values with nearest averaging algorithm and then estimates the regulatory network form the complete data. Therefore, the model is not directly applicable on incomplete dataset. In the future, we can incorporate methods in our model such as structural expectation maximization, which can handle the missing values in the data. We also plan to extend our work by using informative priors in the DBN structure learning which might further improve the accuracy of the model.

CHAPTER 7

PERFORMANCE ANALYSIS OF THE GRN MODELS

The study of gene regulatory network (GRN) models is a major area of research in systems and computational biology and the construction of network models is among the most important challenges in these disciplines. In general, the GRN models characterize quantitative knowledge concerning gene regulation which is hidden in the underlying data. However, knowledge about the underlying biological structures from which these data originate is often incomplete or unavailable. This lack of knowledge poses another critical issue which concerns the validation of the network models. The validation of GRN models can be approached from different perspectives. In recent years, the simulated datasets from DREAM (Dialogue for Reverse Engineering Assessments and Methods) project are gaining interest for validating GRN models. Therefore, we validate our proposed GRN models through the analysis of both simulated and experimental data in this chapter. We compare the performance of the network models in terms of the benchmark criteria of precision and recall together with the run-time complexity and three other statistical measures. To investigate the significance of the models in reconstructing gene regulation from available data, we also compare them against synthetically generated random networks. After thorough experimental validation, we confer that the network model as discussed in chapter 6 exhibits significantly higher performance over the other models. Although the computation time of the models discussed in chapters 4 and 5 has been significantly improved, they identify a low number of true positives in the estimated network.

7.1 Introduction

Despite the huge amount of post-genomic data generated from microarray experiments and tens of computational methods proposed for reverse engineering GRN from such data, knowledge about the underlying biological structures which generate these data is often incomplete or unavailable. This limited knowledge makes it challenging to validate the interactions reconstructed from experimental data alone. Yet, microarray datasets typically have large number of genes with very few samples. As mentioned in the previous chapters, this insufficient data poses another challenge to the traditional statistical or machine learning methods for inferring GRN accurately. Due to these limitations of real experimental data, the use of simulated data for evaluating GRN models is gaining interest. In general, simulated gene expression data is derived from a synthetic network which consists of a topology that determines the structure of the network and a qualitative model for each of the interactions between the genes. The simulated data has the advantage that the true structure is fully known which makes the model validation task attainable. The main disadvantage is that this data is often dissimilar to the experimental data and the validation of model performance based solely on simulated data may be biased.

Nevertheless, it has become a popular practice in genomics research to use both forms of data for the purpose of validating models (Li and Chan 2008, Noman et al. 2007, Li et al. 2011). Initially, the models are tested on relatively large simulated datasets, containing up to 100 genes, and then they are applied on small experimental datasets including tens of genes only. Some of these studies also injected different levels of noise in the simulated data to approximate the true characteristics of the experimental data. Despite all the efforts and the high accessibility of simulated data, our main interest lies in discovering GRN from real experimental data.

In the following sections, we evaluate the performance of the DBN-based GRN models that have been discussed in the previous three chapters in conjunction with two existing models. The benchmark criteria, precision and recall have been calculated along with the computation time for each model. In addition, several other measures such as F-measure, Negative Predictive Value (NPV) and specificity are computed for the statistical verification of the models. In the initial experiment, the models have been validated through the analysis of simulated data that are generated using the GeneNetWeaver from DREAM (Dialogue for Reverse Engineering Assessments and Methods) project with some induced noise. Secondly, we compare the performance of the estimated networks with synthetically generated random networks. Finally, we evaluate the performance of the models through the analysis of experimental data of the yeast cell cycle with missing values in it. Therefore through the various experiments, we evaluate both the performance and the robustness of the GRN models to noisy and incomplete data.

7.2 Existing DBN-based Models for Comparison

As mentioned in chapter 4, a good number of DBN-based GRN models have been proposed to develop dynamic models of gene interaction from time course data. Among these models, we choose two models for the purpose of performance analysis. The first model (GRN_{Murphy}) is the earlier study of Murphy and Mian (1999) that employed the classic DBN structure learning algorithm, namely REVEAL, for estimating the GRN. Given an unknown structure and complete data, the algorithm finds the parent set for each node independently. For a digitized node (2 possible values), there are 2^n parents which can be arranged in a lattice and the problem is to find the highest score in the lattice. In order to deal with the exponential growth of the lattice with n, Zou and Conzen (2005) studied a model (GRN_{Zou}) that reduces

the number of possible parent sets using biological interpretation of microarray data. In case of transcriptional regulation, these parent sets represent any combination of transcription factors that regulate the expression of a target gene. They used a pre-determined threshold for estimating changes in the expression (up/down regulation) of individual genes. Genes that usually have either simultaneous or antecedent changes in expression when compared to their targets were considered as potential Transcription Factors (TF). This consideration allowed them to restrict possible regulators of each gene thus reducing the size of the lattice.

As described in last three chapters, our proposed DBN-based GRN models also utilize biological domain knowledge to reduce the size of the lattice and then employ the REVEAL algorithm to find the structure of the GRN. Therefore, we compare our proposed models with the above mentioned approaches (GRN_{Murphy} and GRN_{Zou}) for several reasons: (1) they employ a similar principle in learning the structure of the GRN, (2) they have been tested in the same domain that is, in the analysis of yeast cell cycle gene expression data, and (3) The comparative analysis allows us to quantify the effectiveness of incorporating biological domain knowledge in reconstructing GRN with DBN.

7.3 Model Evaluation Criteria

We use two benchmark metrics, namely precision (P) and recall (R) to evaluate the performance of our GRN models. "Precision" measures the proportion of actual positives which are correctly identified by the model. "Recall" measures the proportion of actual positives which are known in the true network. In addition to these benchmark metrics, we also compute the computation time of each model. Four factors contribute in the calculation of precision and recall. These are: True Positive (T_p), a connection that exists both in the true network and the estimated network), True Negative (T_n), a connection that does not exist in

either network), False Positive (F_p), a connection that exists only in estimated network) and False Negative (F_n), a connection that exists only in the true network). These factors are computed using the same method as discussed in section 4.5. Given these factors, the benchmark criteria can be defined as in equation 7.1 and equation 7.2.

$$Precision(p) = \frac{T_p}{T_p + F_p} \tag{7.1}$$

$$Recall(c) = \frac{T_p}{T_p + F_n} \tag{7.2}$$

In association with the precision and the recall, we compute three measures to statistically test the different GRN models. These are F-measure, Negative Predictive Value and Specificity. The F-measure is the harmonic mean of the benchmark metrics and can be computed as in equation 7.3.

$$F - measure = 2. \frac{recall.precision}{recall + precision} \tag{7.3}$$

The NPV measures the proportion of negative connections which are correctly identified as negative by the model and is defined as in equation 7.4. A high NPV for a GRN model means that when the model estimates a negative connection, it is most likely correct in its estimation.

$$Negative\ Predictive\ Value\ (NPV) = \frac{T_n}{T_n + F_n} \tag{7.4}$$

The Specificity is a statistical measure which quantifies the proportion of negatives which are correctly identified by the model and can be computed as in equation 7.5. A theoretical optimal system can achieve 100% specificity.

$$Specificity = \frac{T_n}{T_n + F_p} \tag{7.5}$$

7.4 Experimental Setup

The experiments are conducted on a computer system with Intel® Core ™ i5 CPU 760 @ 2.8 GHz and 4 GB RAM, running Windows 7 (Professional). The free statistical software, R version 2.15.0 (R Core Team 2009) is used to impute missing values in the real experimental dataset and find groups of co-expressed genes. Similar to the previous chapters, we use Bayesian Net Toolbox (Murphy 2001a) for constructing DBN and run the experiments under MATLAB environment with version 7.11.0. (R2010b).

7.5 Analysis of Simulated Data

The simulated data and the respective synthetic networks are generated using the Java application GeneNetWeaver (GNW) (Marbach et al. 2009). This network generator has been used as part of the DREAM (Dialogue for Reverse Engineering Assessments and Methods) initiative (Prill et al. 2010). GNW builds synthetic networks by specifying a biologically relevant topology and implementing a model to generate simulated data. GNW has the option to either grow the initial topology from a seed node or a randomly selected node in the source gene network (budding yeast in this thesis). Then the network grows by progressively adding a randomly selected neighboring node till the desired size is reached. Finally, each model can be used to generate simulated time course gene expression data.

We test the performance of our proposed GRN models ($GRN_{Co\text{-}xpressed}$ and $GRN_{Multi\text{-}sources}$) along with two existing models (GRN_{Kevin} and GRN_{Zou}). We leave out the GRN model (GRN_{Phase}) discussed in chapter 4 for the analysis of simulated data. This is because GRN_{Phase} is applicable particularly on genes that exhibit periodicity in their gene expression. Most importantly, GRN_{Phase} clusters genes based on very specific knowledge of the biological process (yeast cell cycle in this thesis) that generates the data. As GNW generates networks

by choosing genes arbitrarily from all the yeast genes, the proposed model lacks in finding a common biological insight for clustering genes into groups.

Each of the above mentioned models was applied to five independent networks of size 20, 50 and 100 genes. These sub-networks are extracted from the high-dimensional yeast GRN with 4441 nodes and 12873 edges. Since we have selected the yeast cell cycle as the biological domain in this thesis, the most well known transcription factors such as MBP1, SWI6, and SWI4 have been set as seed nodes for extracting the sub-networks. A model consisting of ordinary and stochastic differential equations with Gaussian noise has been generated for each synthetic network. Then each GNW-generated network-model was used to simulate time-series datasets with a total of 21 time points, where t_max = 100.

A complete comparison of the four GRN models through the analysis of 5 different datasets of size 20, 50 and 100 are shown in figures 7.1, 7.2 and 7.3 respectively and the corresponding data is listed in tables 7.1, 7.2 and 7.3. It is noteworthy that the last two criteria, NPV and Specificity show high values for all models in all the three tables. This implies that all the GRN models under study are capable of estimating true negative connections with high accuracy (95% approx). Therefore, we have excluded them in the subsequent figures considering that these measures are not significantly distinguishing model performance. For the network size of 20 genes, we have set MBP1 as the seed node arbitrarily and used the greedy method to choose the neighboring genes. This selection implies that each simulated dataset contains the time-series expression data of 20 genes starting from MBP1 and the included genes vary across the datasets. For the network of size 50, we have selected SWI6 as the seed node, which is again a random selection except the fact that it is a well-known TF in the yeast cell cycle gene regulation. In a similar fashion, we have selected SWI4 as the seed node for generating the networks of size 100.

Table 7.1: Comparison of performance among the four different DBN-based GRN models. The first two GRN models are existing works and the latter two have been proposed in this thesis. The models have been applied on 5 different simulated datasets of 20 genes over 21 time points. For each dataset, the best performing model in respect to the benchmark criteria, precision and recall has been highlighted in bold.

Dataset	Methods	Total identified relationships	Correctly identified relationships	Misdirected relationships	Precision (%)	Recall (%)	F-measure	Negative Predictive Value	Specificity (%)
1	GRN_{Murphy}	39	1	2	2.70	5.88	3.70	0.95	95.09
	GRN_{Zou}	36	2	1	5.71	11.76	7.69	0.95	95.43
	$GRN_{Co\text{-}expressed}$	80	7	2	10.45	41.18	16.67	0.97	96.62
	$GRN_{Multi\text{-}sources}$	**21**	**16**	**0**	**76.19**	**94.12**	**84.21**	**1.00**	**99.71**
2	GRN_{Murphy}	40	5	2	13.16	26.32	17.54	0.96	95.67
	GRN_{Zou}	38	0	2	0.00	0.00	0.00	0.94	94.15
	$GRN_{Co\text{-}expressed}$	83	5	3	7.25	26.32	11.36	0.95	95.21
	$GRN_{Multi\text{-}sources}$	**21**	**19**	**0**	**90.48**	**100.00**	**95.00**	**1.00**	**100.00**
3	GRN_{Murphy}	39	3	2	8.57	15.79	11.11	0.95	95.09
	GRN_{Zou}	36	3	2	8.82	15.79	11.32	0.95	95.11
	$GRN_{Co\text{-}expressed}$	98	2	4	2.50	10.53	4.04	0.94	93.95
	$GRN_{Multi\text{-}sources}$	**24**	**16**	**0**	**66.67**	**84.21**	**74.42**	**0.99**	**99.11**
4	GRN_{Murphy}	39	5	1	13.89	26.32	18.18	0.96	95.69
	GRN_{Zou}	39	4	2	10.81	21.05	14.29	0.95	95.37
	$GRN_{Co\text{-}expressed}$	96	1	1	1.30	5.26	2.08	0.94	93.66
	$GRN_{Multi\text{-}sources}$	**23**	**15**	**0**	**65.22**	**78.95**	**71.43**	**0.99**	**98.82**
5	GRN_{Murphy}	38	3	2	9.38	15.79	11.76	0.95	95.14
	GRN_{Zou}	31	0	1	0.00	0.00	0.00	0.94	94.26
	$GRN_{Co\text{-}expressed}$	89	2	4	3.03	10.53	4.71	0.94	94.24
	$GRN_{Multi\text{-}sources}$	**20**	**17**	**0**	**85.00**	**89.47**	**87.18**	**0.99**	**99.41**

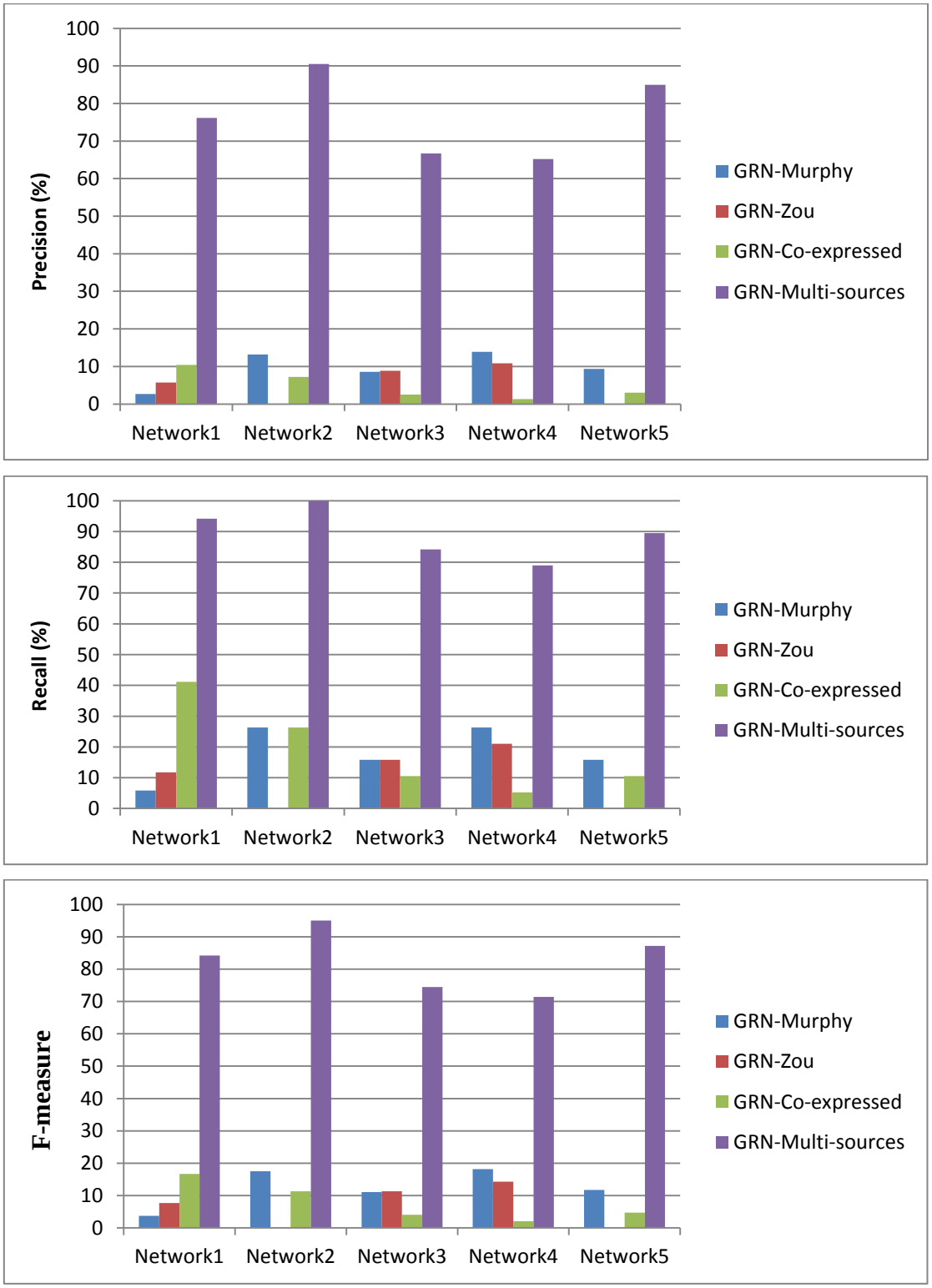


Figure 7.1: Comparison of performance in terms of precision (P), recall (R) and F-measure among the four DBN-based GRN models through the analysis of 5 different simulated datasets including 20 genes.

Table 7.2: Comparison of performance among the four different DBN-based GRN models. The first two GRN models are existing works and the latter two have been proposed in this thesis. The models have been applied on 5 different simulated datasets of 50 genes over 21 time points. For each dataset, the best performing model in respect to the benchmark criteria, precision and recall has been highlighted in bold.

Dataset	Methods	Total identified relationships	Correctly identified relationships	Misdirected relationships	Precision (%)	Recall (%)	F-measure	Negative Predictive Value	Specificity (%)
1	GRN_{Murphy}	100	5	3	5.26	4.76	5.00	0.96	95.56
	GRN_{Zou}	102	3	6	3.13	2.86	2.99	0.95	95.46
	$GRN_{Co\text{-}expressed}$	248	2	5	0.96	1.90	1.28	0.95	95.18
	$GRN_{Multi\text{-}sources}$	**80**	**28**	**7**	**38.36**	**26.67**	**31.46**	**0.97**	**96.61**
2	GRN_{Murphy}	96	2	2	2.13	1.87	1.99	0.95	95.33
	GRN_{Zou}	93	8	2	8.79	7.48	8.08	0.96	95.60
	$GRN_{Co\text{-}expressed}$	206	7	7	4.43	6.54	5.28	0.95	95.42
	$GRN_{Multi\text{-}sources}$	**74**	**26**	**3**	**36.62**	**24.30**	**29.21**	**0.96**	**96.43**
3	GRN_{Murphy}	96	9	5	10.23	8.26	9.14	0.96	95.56
	GRN_{Zou}	98	3	5	3.23	2.75	2.97	0.95	95.28
	$GRN_{Co\text{-}expressed}$	214	13	6	7.39	11.93	9.12	0.96	95.57
	$GRN_{Multi\text{-}sources}$	**78**	**28**	**4**	**38.36**	**25.69**	**30.77**	**0.96**	**96.43**
4	GRN_{Murphy}	100	0	4	0.00	0.00	0.00	0.95	94.78
	GRN_{Zou}	100	2	4	2.08	1.71	1.88	0.95	94.86
	$GRN_{Co\text{-}expressed}$	236	28	11	14.97	23.93	18.42	0.96	95.85
	$GRN_{Multi\text{-}sources}$	**89**	**32**	**3**	**37.65**	**27.35**	**31.68**	**0.96**	**96.22**
5	GRN_{Murphy}	100	5	3	5.43	4.63	5.00	0.95	95.42
	GRN_{Zou}	100	2	3	2.06	1.85	1.95	0.95	95.28
	$GRN_{Co\text{-}expressed}$	239	5	9	2.62	4.63	3.34	0.95	95.21
	$GRN_{Multi\text{-}sources}$	**77**	**26**	**6**	**36.62**	**24.07**	**29.05**	**0.96**	**96.39**

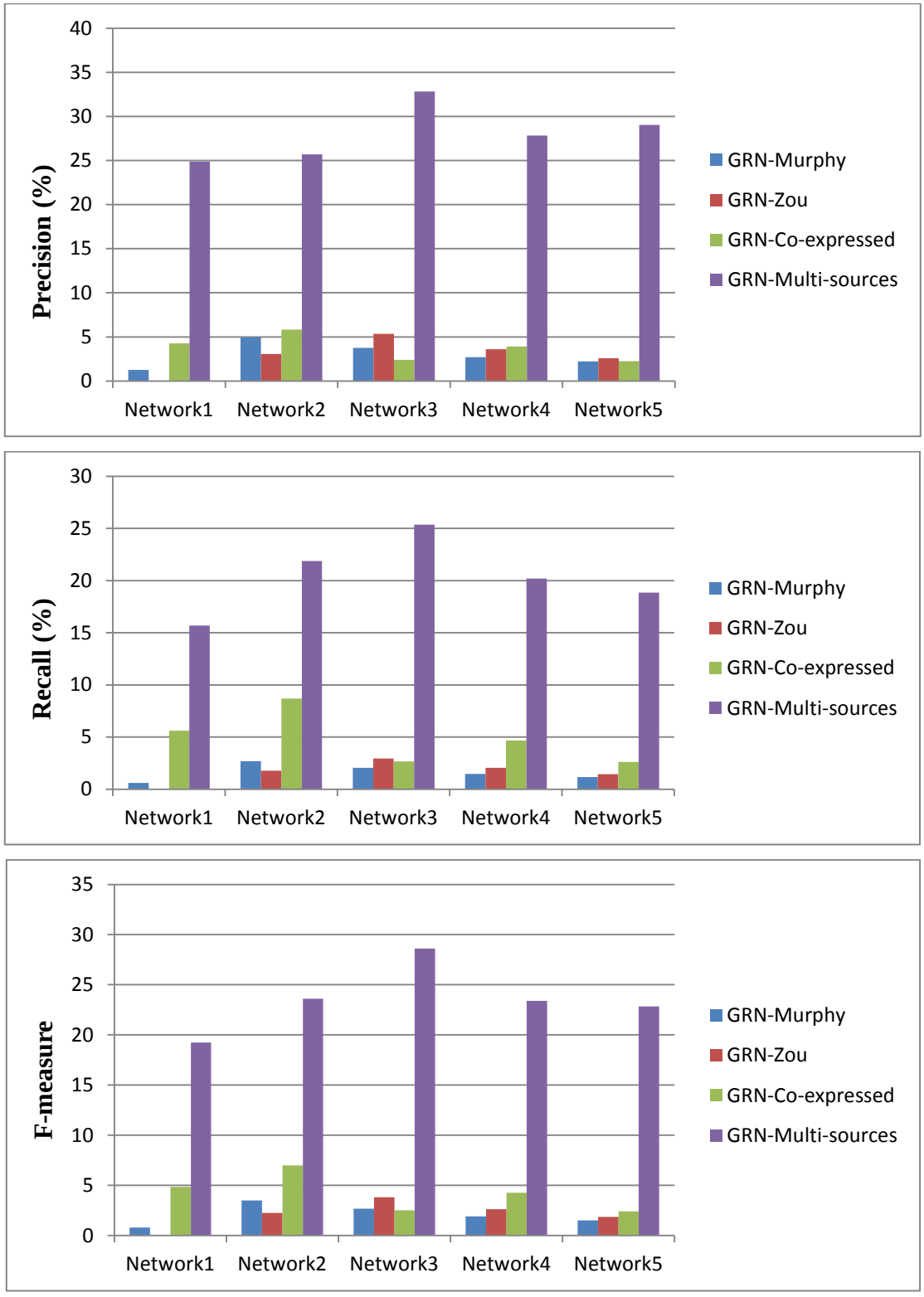


Figure 7.2: Comparison of performance in terms of precision (P), recall (R) and F-measure among the four DBN-based GRN models through the analysis of 5 different simulated datasets including 50 genes.

Table 7.3: Comparison of performance among the four different DBN-based GRN models. The first two GRN models are existing works and the latter two have been proposed in this thesis. The models have been applied on 5 different simulated datasets of 100 genes over 21 time points. For each dataset, the best performing model in respect to the benchmark criteria, precision and recall has been highlighted in bold.

Dataset	Methods	Total identified relationships	Correctly identified relationships	Misdirected relationships	Precision (%)	Recall (%)	F-measure	Negative Predictive Value	Specificity (%)
1	GRN_{Murphy}	168	2	7	1.25	0.59	0.80	0.96	96.43
	GRN_{Zou}	166	0	5	0.00	0.00	0.00	0.96	96.49
	$GRN_{Co\text{-}expressed}$	566	41	15	4.26	5.62	4.85	0.97	96.50
	$GRN_{Multi\text{-}sources}$	**222**	**53**	**6**	**24.88**	**15.68**	**19.24**	**0.97**	**96.95**
2	GRN_{Murphy}	191	9	6	4.95	2.69	3.49	0.97	96.54
	GRN_{Zou}	196	6	7	3.06	1.78	2.25	0.97	96.54
	$GRN_{Co\text{-}expressed}$	592	29	13	5.84	8.68	6.98	0.97	96.64
	$GRN_{Multi\text{-}sources}$	**294**	**73**	**7**	**25.70**	**21.86**	**23.62**	**0.97**	**97.19**
3	GRN_{Murphy}	194	7	5	3.74	2.06	2.66	0.96	96.46
	GRN_{Zou}	193	10	6	5.35	2.95	3.80	0.96	96.49
	$GRN_{Co\text{-}expressed}$	451	9	11	2.39	2.65	2.52	0.96	96.41
	$GRN_{Multi\text{-}sources}$	**272**	**86**	**5**	**32.82**	**25.37**	**28.62**	**0.97**	**97.28**
4	GRN_{Murphy}	196	5	9	2.70	1.46	1.90	0.96	96.41
	GRN_{Zou}	195	7	6	3.59	2.05	2.61	0.97	96.51
	$GRN_{Co\text{-}expressed}$	496	16	14	3.91	4.68	4.26	0.96	96.44
	$GRN_{Multi\text{-}sources}$	**258**	**69**	**8**	**27.82**	**20.18**	**23.39**	**0.97**	**97.07**
5	GRN_{Murphy}	192	4	8	2.21	1.16	1.52	0.96	96.36
	GRN_{Zou}	200	5	7	2.59	1.45	1.86	0.96	96.37
	$GRN_{Co\text{-}expressed}$	479	9	13	2.24	2.61	2.41	0.96	96.33
	$GRN_{Multi\text{-}sources}$	**233**	**65**	**7**	**29.02**	**18.84**	**22.85**	**0.97**	**97.00**

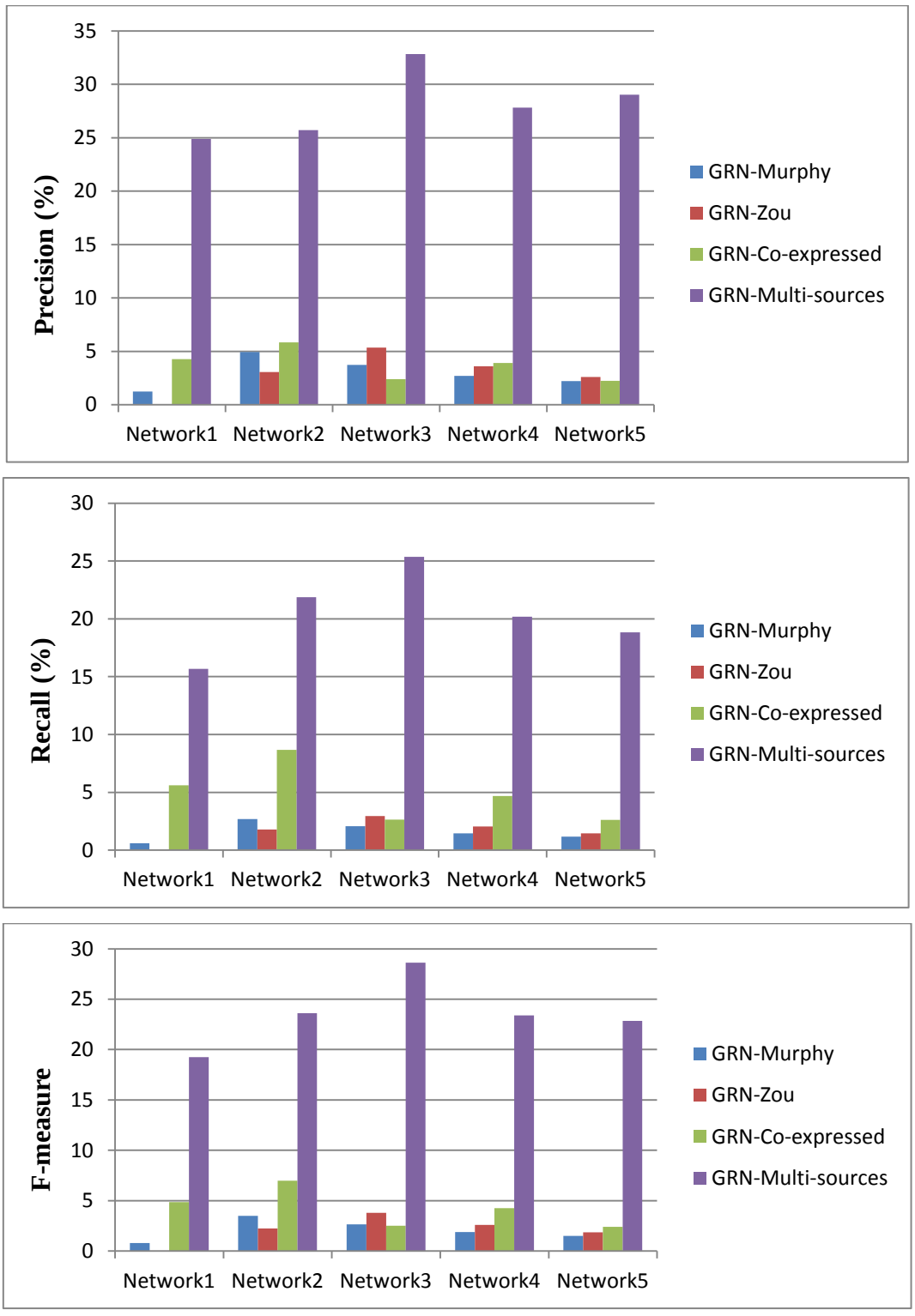


Figure 7.3: Comparison of performance in terms of precision (P), recall (R) and F-measure among the four DBN-based GRN models through the analysis of 5 different simulated datasets including 100 genes.

In the above tables, we record three additional factors together with model evaluation criteria. These factors are: "Total identified relationships" which is the total number of regulator-gene relationships estimated by the model, "Correctly identified relationships" is the true positive relationships and "Misdirected relationships" represents false positives that have reverse relationships in the true network. In this analysis, we exclude run-time complexity of the models as performance evaluation criteria because of the small size of the networks.

As shown in the figures (Figure 7.1, 7.2 and 7.3), the GRN model ($GRN_{Multi\text{-}sources}$) proposed in chapter 6 of this thesis exhibits significantly improved performance in terms of precision, recall and F-measure for all the datasets. We speculate that the incorporation of biological knowledge of gene regulation contributes towards such improvements. The other model, $GRN_{Co\text{-}expressed}$ shows inconsistent performance across the datasets. This is because the model solely depends on the data to find clusters of co-expressed genes without including any biological relevance. As a result it is sometimes unable to determine the groups of co-expressed genes successfully which consequently makes the model unpredictable. In addition, the model estimates a high number of regulatory relationships compared to the other models as it finds co-regulation of co-expressed genes. The two existing models (GRN_{Murphy} and GRN_{Zou}) show consistently poor performance for all data sizes. Though the model GRN_{Zou} utilizes biological interpretation to generate a preprocessed network, their static and predefined cut-offs for determining up and down-regulation of individual genes has a varied effect on its performance.

7.6 Analysis of Experimental Data

To evaluate the performance of our proposed GRN models in analyzing real experimental data, we have applied them on the benchmark dataset of the yeast cell cycle

(Spellman et al. 1998). We test the models in three different steps. At the first step, we assess the models through the analysis of small scale networks including 19 genes only. In the next step, the proposed models are compared with synthetically generated random networks to study their feasibility. Finally, we evaluate the performance of the models through the analysis of large-datasets including 300 genes.

7.6.1 Inference of small-scale networks

In this section, we analyze a small dataset containing 19 cell cycle regulated yeast genes. These are MPS1, BUB1, BUB3, MAD1, MAD2, MAD3, CDH1, CDC27, CDC20, CDC14, NET1, MOB1, DBF2, PDS1, ESP1, SMC1, MCD1, SMC3 and IRR1. We choose to work on this particular dataset so that it allows us to compare our models with other GRN models which were also tested on this dataset. In this analysis, we include two models based on PMDL (predictive minimum description length). The first model, $GRN_{Chaitankar}$ has been recently studied in the paper of Chaitankar et al. (2010) and the other model GRN_{Wentao} has been proposed by Wentao et al. (2006). The target network among these 19 genes has been extracted from the KEGG pathway database (KEGG 2000, Kanehisa et al. 2006, Kanehisa et al. 2008) as shown in Figure 7.4.

Chaitankar et al. (2010) proposed a new algorithm which incorporates mutual information, conditional mutual information and predictive minimum description length (PMDL) to estimate the regulatory interactions between genes. They inferred a total of 30 edges, of which nine are correctly identified as shown in Figure 7.5. Another PMDL based model proposed by Wentao et al. (2006) identified a total of nine edges, of which only one is correctly inferred. As explained in section 4.5.3, the figures, 7.4 and 7.5 are generated by the Matlab program and lack visual quality.

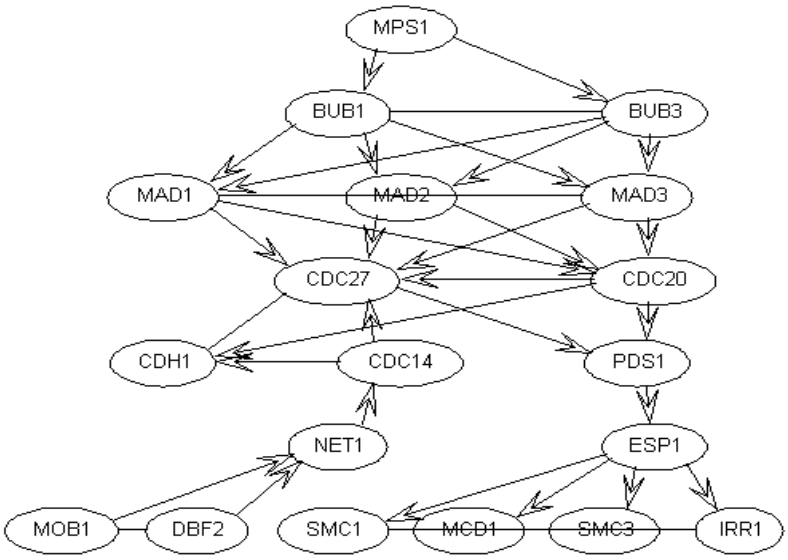


Figure 7.4: Target network among the 19 CCR genes extracted from the KEGG pathway database. In the network, each node represents a gene and an edge shows direct relationships among two genes. [generated by Matlab]

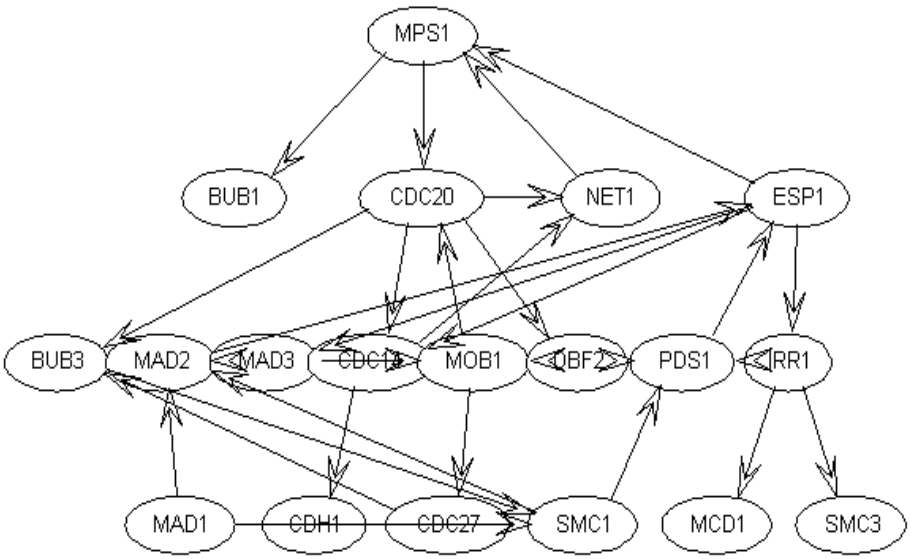


Figure 7.5: Network estimated by the PMDL-based GRN model proposed in Chaitankar et al. (2010). The model estimates a total of 30 edges, of which 9 are true positives. [generated by Matlab].

To investigate how successfully DBN-based GRN models can estimate gene regulation in compare to other models, we analyze the same dataset including the same 19

genes with our proposed models (GRN_{Phase}, $GRN_{Co\text{-}expressed}$, and $GRN_{Multi\text{-}sources}$) together with the existing models (GRN_{Murphy} and GRN_{Zou}). The models estimated a total of 31, 48, 40, 26 and 25 regulatory relationships respectively. The first two models infer only 2 true positive connections, whereas the last two models estimate 3 connections correctly. The other model, $GRN_{Multi\text{-}sources}$ identifies a maximum of 17 true positive connections in the estimated network as shown in Figure 7.6.

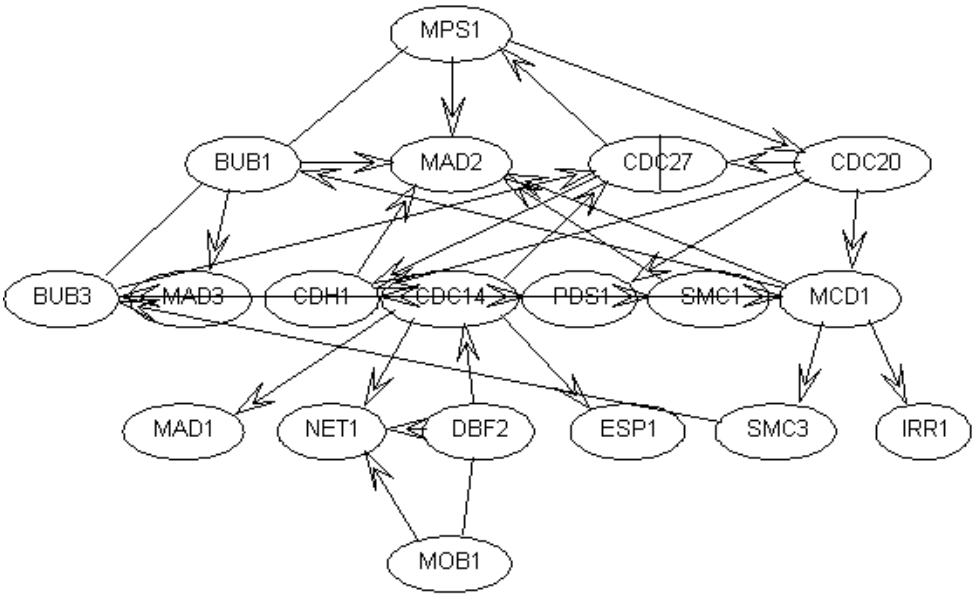


Figure 7.6: Network estimated by $GRN_{Multi\text{-}sources}$ as discussed in chapter 6. There are 40 connections in the network, of which 17 are identified as true positives. [generated by Matlab].

Next, we compare the two PMDL based models together with the 5 DBN-based GRN models in terms of the benchmark criteria, precision and recall as well as the statistical measure of model accuracy, F-measure, as shown in Figure 7.7. Similar to the previous section, we do not include the measures of true negative rate, NPV and Specificity, in this performance analysis as all the models demonstrate very high accuracy in estimating true negative connections. Figure 7.7 shows that the DBN-based model, $GRN_{Multi\text{-}sources}$ exhibits

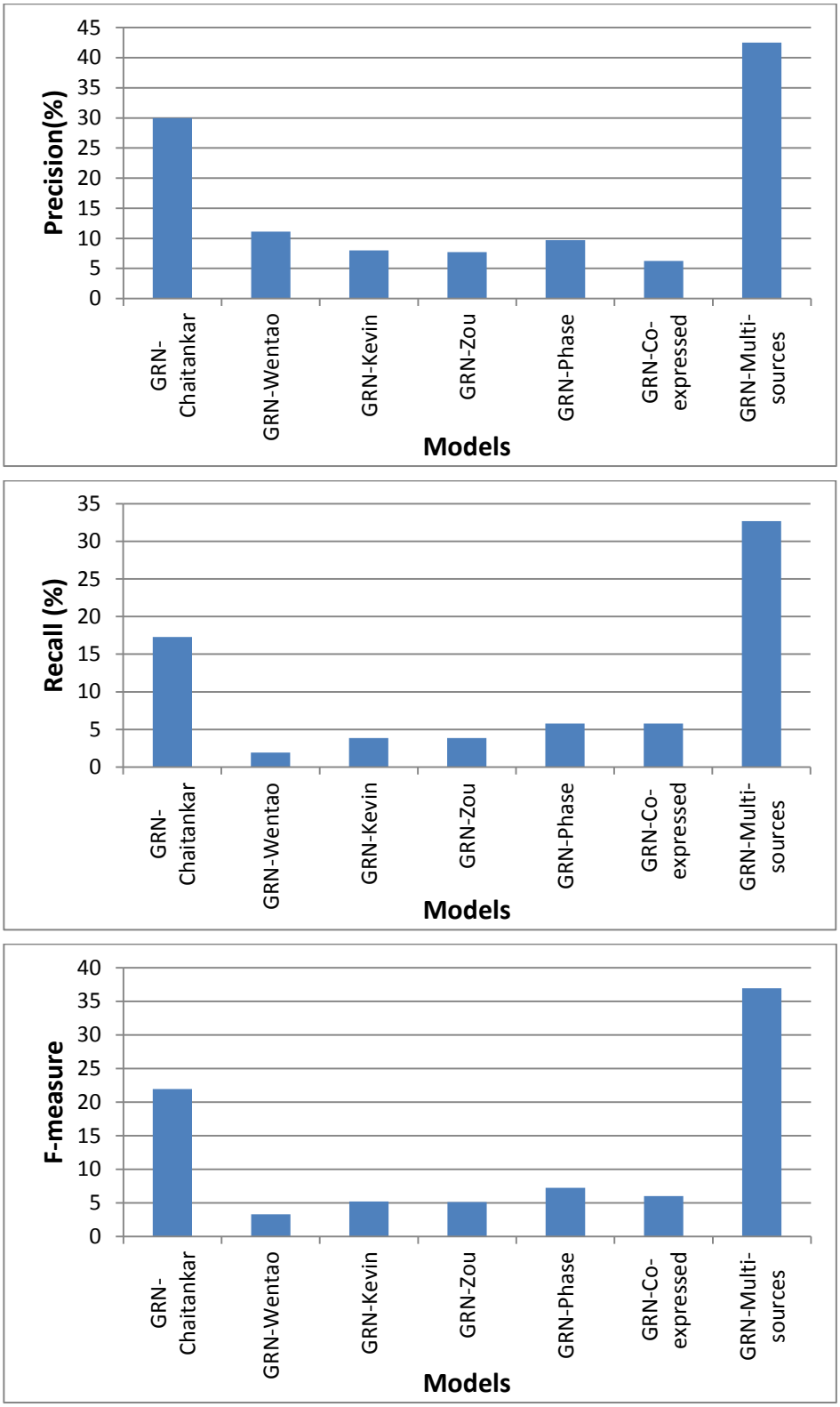


Figure 7.7: Comparison of the performance of 7 GRN models in terms of precision (P), recall(R) and F-measure. The first two models are based on PMDL and the latter models are DBN-based.

significantly better performance (42.5% precision 35% recall and an F-measure of 36.96) over the other models. The second best performing model is the PMDL-based model with precision of 30%, recall of 19% and an F-measure of 21.95. The other models exhibit inconsistent performance in terms of all three criteria. For instance, GRN_{Wentao} shows a precision of 10%, whereas the recall and F-measure of the network is the lowest of all the models. Overall, $GRN_{Co\text{-}expressed}$ exhibits poor performance in terms of both recall and precision as well as F-measure.

We conclude this analysis of small scale networks with the finding that the DBN-based model ($GRN_{Multi\text{-}sources}$) proposed in chapter 6 of this thesis outperforms all other models in estimating gene regulation from experimental data, which also confirms our previous findings of analyzing simulated data.

7.6.2 Comparison with synthetically generated random networks

To further validate the GRN models (GRN_{Phase}, $GRN_{Co\text{-}expressed}$ and $GRN_{Multi\text{-}sources}$), we evaluate the performance of the models by comparing the estimated GRNs with synthetically generated random networks. The models have been applied on the cell cycle datasets (Spellman et al. 1998) of varying sizes including 20, 40, 60, 80 and 100 genes and the respective target networks are extracted from multiple sources (Simon et al. 2001, Teixeira et al. 2006, Monteiro et al. 2008, KEGG 2000, Kanehisa et al. 2006, Kanehisa et al.2008). The random networks are generated using the algorithm described in the paper of Wentao et al. (2006) and the parameters (nodes and edges) are set according to the estimated GRN. The algorithm for random networks has been run 1000 times each for a specific size of the network and an average precision and recall are computed.

At first, we estimate networks of varying dimensions with the GRN model, GRN_{Phase}

from the experimental data. The precision, recall and F-measure of the estimated network against the random network with a varying network size are plotted in Figures 7.8a 7.8b and 7.8c respectively. In Figure 7.8, it is observed that GRN_{Phase} estimates networks with slightly higher precision and recall for small scale networks in comparison to the synthetically generated random networks. The F-measure of the model, GRN_{Phase} also demonstrates a similar performance over the random networks. Nevertheless, the performance of the estimated network declines in terms of all three criteria as the dimension of the network grows and is not significantly distinguishable over the randomly generated network.

In the next experiment, we analyze the same sets of data with the GRN model ($GRN_{Co\text{-}expressed}$) that finds co-regulation of co-expressed genes as discussed in chapter 5. We compare and plot the precision, recall and F-measure of the estimated networks of varying dimensions with corresponding random networks as shown in Figure 7.9a, 7.9b and 7.9c respectively. In the Figure 7.9, the GRN model, $GRN_{Co\text{-}expressed}$ exhibits inconsistent performance in comparison to the randomly generated networks. Although the model estimates small scale networks with a higher precision, recall and F-measure, it shows irregular performance as the network size grows. In some cases, it performs even poorer than the random networks. For instance, for the network dimension of 60, the precision of the model has declined in comparison to the random network and as the network size grows the performance of the model reaches the level of the random networks.

To evaluate the performance of our last GRN model $GRN_{Multi\text{-}sources}$ (discussed in chapter 6), we analyze the same sets of data including same number of genes. The precision of the estimated networks against the random networks with varying network dimension is plotted in Figure 7.10a, the recall is shown in Figure 7.10b and the F-measure is in 7.10c. According to Figure 7.10, $GRN_{Multi\text{-}sources}$ estimates networks with significantly improved

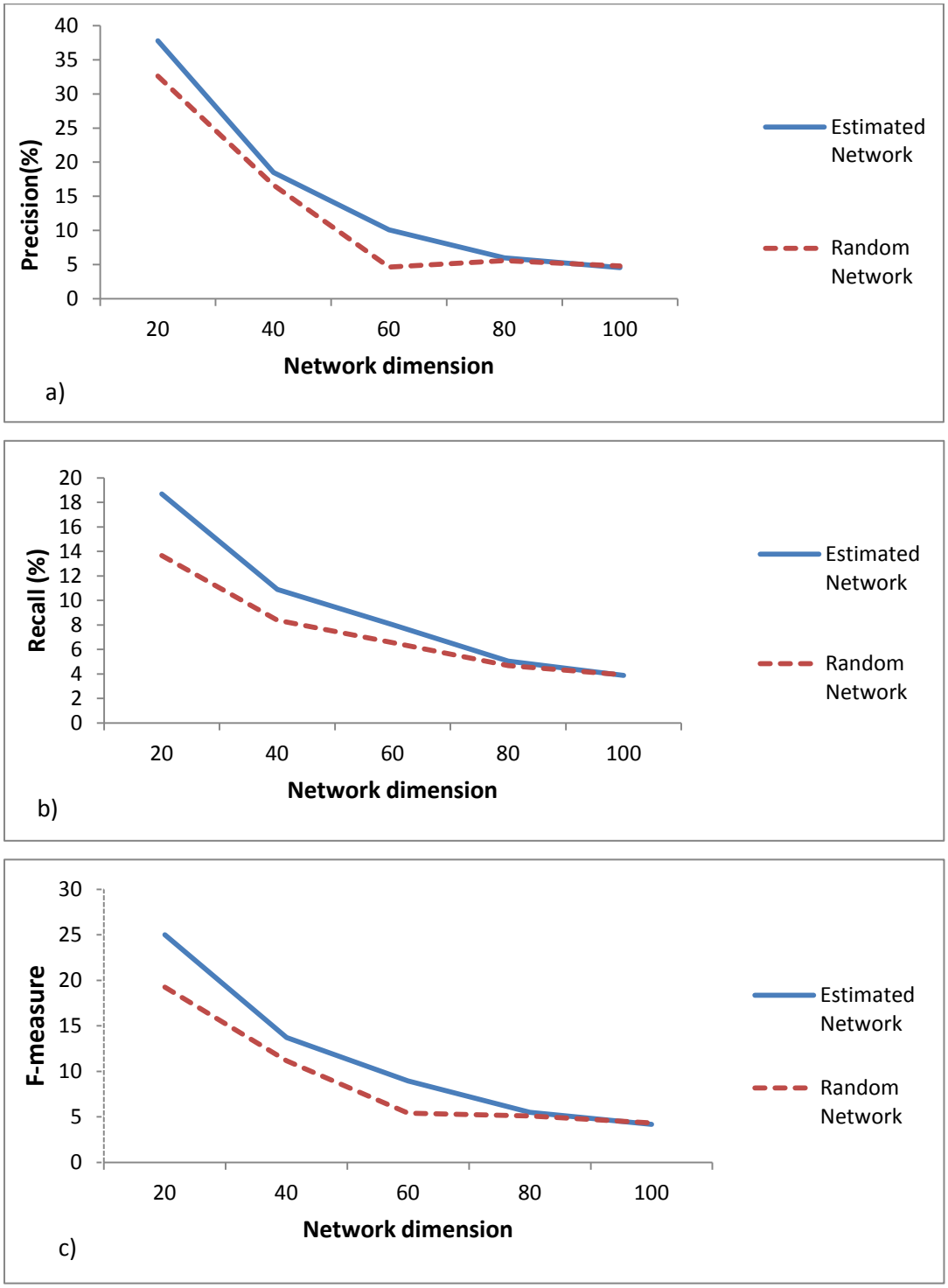


Figure 7.8: Performance evaluation of GRN_{Phase} (discussed in chapter 4) against synthetically generated random networks. The model estimates networks of different dimensions, including 20, 40, 60, 80 and 100 genes. For each network dimension, 1000 random networks are generated and the average precision and recall are calculated: a) precision, b) recall and c) F-measure.

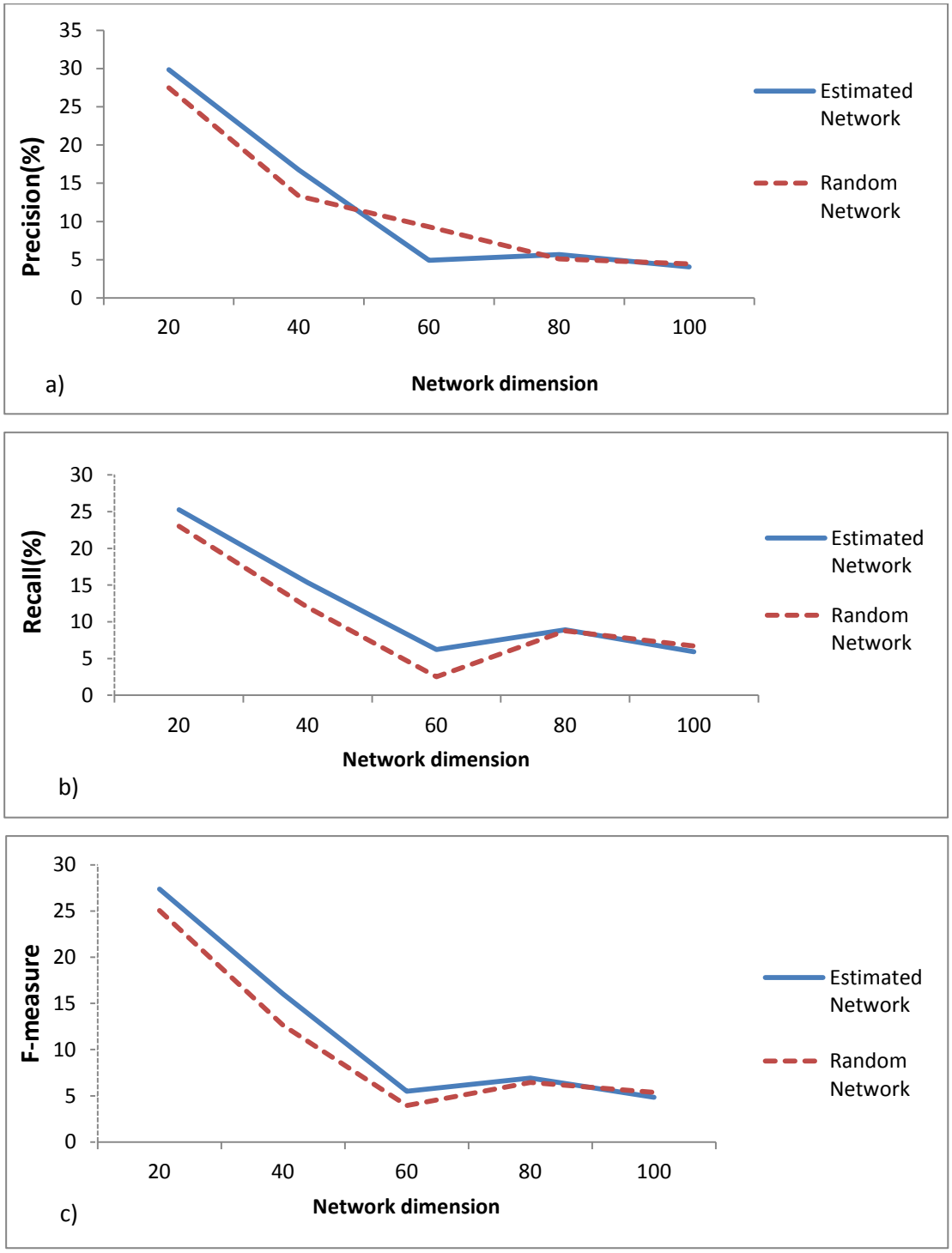


Figure 7.9: Performance evaluation of $GRN_{Co\text{-}expressed}$ (discussed in chapter 5) against synthetically generated random networks. The model estimates networks of different dimensions, including 20, 40, 60, 80 and 100 genes. For each network dimension, 1000 random networks are generated and the average precision and recall are calculated: a) precision, b) recall and c) F-measure.

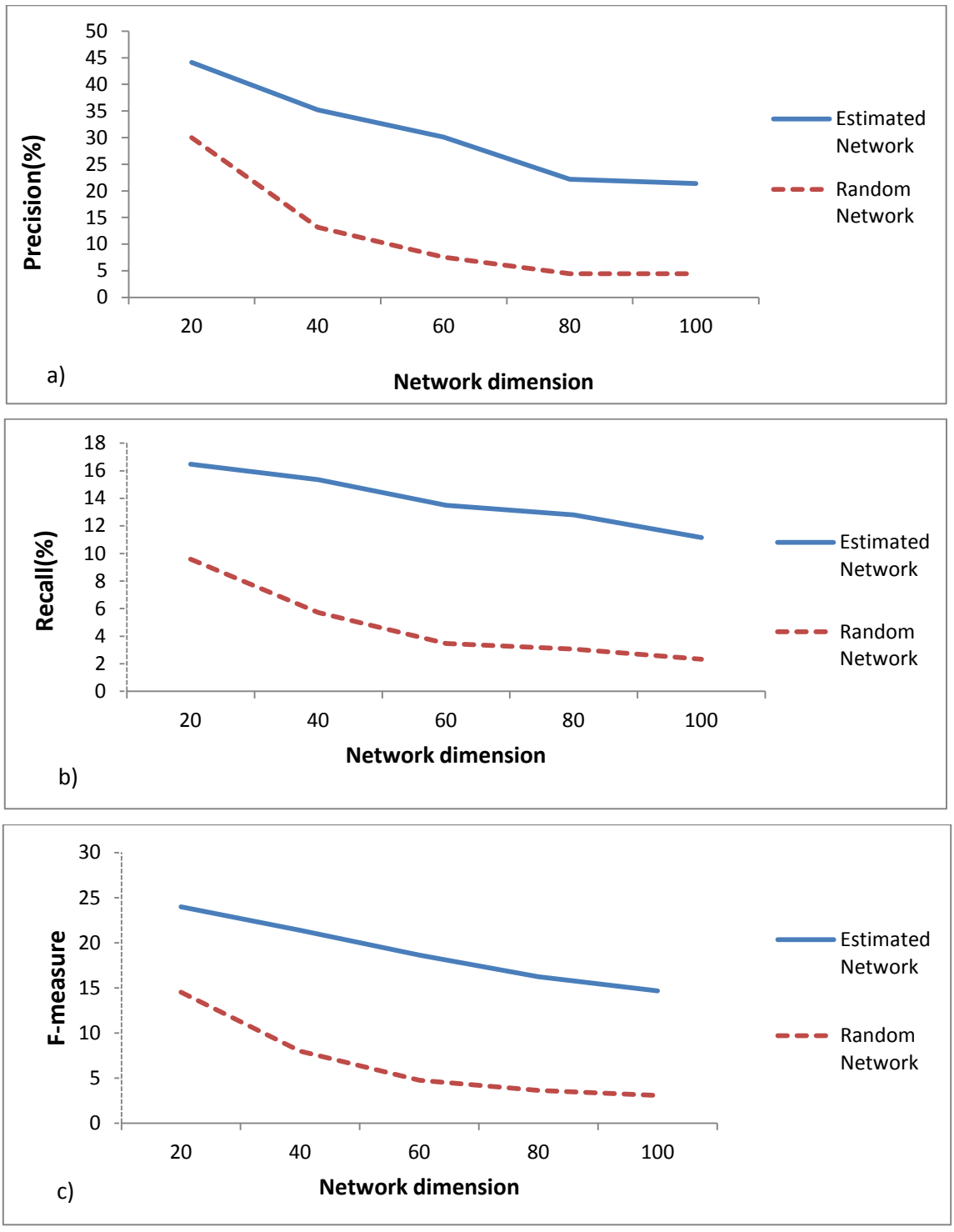


Figure 7.10: Performance evaluation of $GRN_{Multi\text{-}sources}$ (discussed in chapter 6) against synthetically generated random networks. The model estimates networks of different dimensions, including 20, 40, 60, 80 and 100 genes. For each network dimension, 1000 random networks are generated and the average precision and recall are calculated: a) precision b) recall and c) F-measure.

performance in comparison with the synthetically generated random networks. The precision and recall of the randomly generated networks plunge sharply with the increasing number of nodes in the network whereas the performance of $GRN_{Multi\text{-}sources}$ declines steadily. For instance, the precision of the estimated network for size 20 is 44%, whereas the random networks have an average value of 30%. With the addition of 20 new genes in the dataset, the precision of the random networks falls to 13% which is as high as 60% decline in value. The precision of $GRN_{Multi\text{-}sources}$ also drops to 35% which is 10% decline and shows similar fall as the network size grows. The performance of the random network declines even drastically with further increments of the network dimension. While comparing the performance of the estimated networks over the synthetically generated random networks in terms of F-measure, we observe a similar pattern. The F-measure of the estimated network demonstrates a steady fall whereas it shows a rapid decline for the random networks as the size grows.

To summarize the performance of the GRN models against the synthetically generated random networks, we confer that the two models, GRN_{Phase} and $GRN_{Co\text{-}expressed}$ exhibit improved performance for the analysis of small-scale networks, however the models suffer as the dimension of the network grows. The other model, $GRN_{Multi\text{-}sources}$ outperforms the other two models and shows steady performance with the increasing number of genes in the dataset.

7.6.3 Inference of large-scale networks

Finally, we validate the performance of the DBN-based models (GRN_{Murphy}, GRN_{Zou}, GRN_{Phase}, $GRN_{Co\text{-}expressed}$, and $GRN_{Multi\text{-}sources}$) through the analysis of large-scale experimental data containing 300 genes. The working data is extracted from the yeast cell cycle datasets of Spellman et al. (1998) and the known networks are extracted from multiple sources (Simon et al., 2001; Teixeira et al., 2006; Monteiro et al., 2008, KEGG 2000, Kanehisa et al. 2006,

Kanehisa et al. 2008). The complete performance comparison of the models in analyzing large scale experimental data is depicted in Figure 7.11 and the corresponding data is given in Table 7.4. Similar to the previous analyses in sections 7.5 and 7.6, we exclude the measures of true negative rate, NPV and Specificity in Figure 7.11 because all the models demonstrate very high accuracy in estimating true negative connections.

Table 7.4: Comparison of performance among the five different DBN-based GRN models. The first two GRN models are existing works and the latter three have been proposed in this thesis. The models have been applied on an experimental dataset of 300 genes over 18 time points. The best performing model in respect to the benchmark criteria, precision and recall has been highlighted in bold.

Methods	**Total identified relationships**	**Correctly identified relationships**	**Misdirected relationships**	**Precision (%)**	**Recall (%)**	**F-measure**	**Negative Predictive Value**	**Specificity (%)**	**Computation Time**
GRN_{Murphy}	-	-	-	-	-	-	-	-	stopped after 7 days
GRN_{Zou}	-	-	-	-	-	-	-	-	stopped after 7 days
GRN_{Phase}	2379	12	23	0.51	1.36	0.74	0.99	99.00	3 days 11 hrs 3mins
$GRN_{Co\text{-}expressed}$	2791	14	15	0.50	1.59	0.76	0.99	99.01	3days 17 hrs 2 mins
$GRN_{Multi\text{-}sources}$	**678**	**104**	**11**	**15.59**	**11.82**	**13.45**	**0.99**	**99.12**	**2 mins**

As shown in the table, we include the run-time complexity of each model as performance evaluation criteria for this large dataset. The existing two models, GRN_{Murphy} and GRN_{Zou} is unable to reconstruct the network in a week. We set 7 days as a cut-off time and

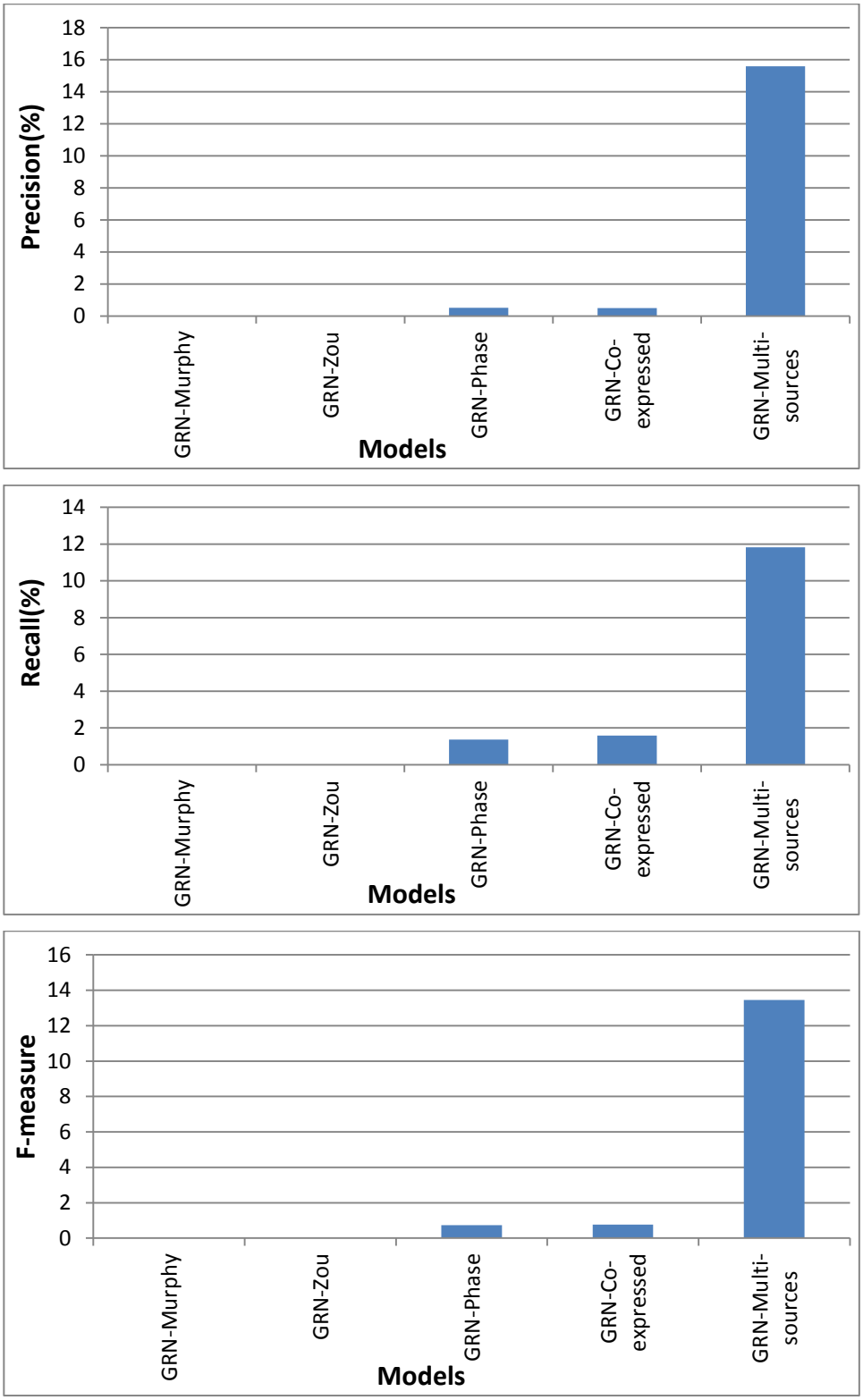


Figure 7.11: Comparison of the performance of 5 GRN models in terms of precision (P) and recall (R). The first two models are existing works and the latter models are proposed in this thesis.

stopped running the experiments after that. The other two models, GRN_{Phase} and $GRN_{Co-expressed}$ take more than 3 days to estimate the respective networks, whereas $GRN_{Multi-sources}$ infers the network in only two minutes. This is a significant achievement in terms of computation time for the reconstruction of large scale networks.

Nevertheless, each of these models suffers from the so-called curse of dimensionality problem as the search space grows exponentially with the number of potential regulators. Our proposed models deal with this high dimensionality problem by incorporating biological domain knowledge. They generate a preprocessed network by reducing the number of potential regulators for each gene which consequently restricts the growth of the search space; hence the models are capable of analyzing relatively large datasets.

In Figure 7.11, the precision and recall of GRN_{Murphy} and GRN_{Zou} are set to 0 as the models fail to estimate the networks within the cut-off time of 7 days. As a result, the F-measure of the respective models is also 0. Of the 880 known relationships in the known network, the models (GRN_{Phase} , $GRN_{Co-expressed}$, $GRN_{Multi-souces}$) identify 12 , 14 and 104 true positive connections respectively. Though, $GRN_{Multi-sources}$ estimates the maximum number of true positives and outperforms the other models, the recall of the model is just 11%. The models (GRN_{Phase}, $GRN_{Co-expressed}$, and $GRN_{Multi-sources}$) infer a total of 2378, 2791 and 678 relationships in their respective estimated networks. Since the number of false positive edges is inversely proportional to the precision of the model, the first two models have a precision which is as low as 0.5%, whereas $GRN_{Multi-sources}$ soars to 15%. Given that, F-measure is the harmonic mean of recall and precision, a similar performance is expected. The F-measure of GRN_{Phase} and $GRN_{Co-expressed}$ is computed as low as less than 1 whereas $GRN_{Multi-sources}$ records above 13.

To summarize the findings of analyzing the large experimental dataset with the GRN

models, we conclude that $GRN_{Multi\text{-}sources}$ outperforms all the model both in terms of benchmark criteria and the computation time of the models. This result is highly aligned with our previous findings as stated in the earlier sections.

7.7 Conclusion

In this chapter, we have validated the GRN models proposed in chapters 4, 5 and 6 of this thesis together with two existing models. The DBN-based GRN models are also compared against two PMDL-based models to investigate their relative performance for the analysis of small-scale networks. We have chosen PMDL-based models for two reasons: 1) the model is relatively recent study and 2) it has been applied in the domain of yeast cell cycle. For the purpose of model validation, we used both simulated and experimental datasets. The simulated data is from DREAM4 initiative which is considered as the benchmark data for assessing GRN models by the scientific community. To complete our validation process, we also compare the performance of the models with synthetically generated random networks.

Over the course of different form of validations, we found that $GRN_{Multi\text{-}sources}$ as discussed in chapter 6, exhibits significantly higher performance over the other models. The model reduces the problem space by incorporating other sources of biological data prior to estimating the network. This early reduction has significantly slowed down the exponential growth of the search space which has extraordinarily improved the run time of the model. Likewise, the removal of extraneous genes from the potential regulator list for each target gene has reduced the data requirements for the DBN learning algorithm. As a consequence, the model is capable of identifying a high number of true positive connections even from the noisy experimental data. The other two models, GRN_{Phase} and $GRN_{Co\text{-}expressed}$ show poor performances, particularly in terms of precision and recall. As described in chapter 4,

GRN_{Phase} clusters genes into groups based on their peak time and finds regulation within each group. Since the peak time of each gene is solely calculated from the expression profile, the model often fails to group them correctly. Similarly, we conjecture two major reasons for the poor performance of $GRN_{Co\text{-}expressed}$. These are: 1) the use of single source data and 2) clustering of genes without any biological relevance. Nevertheless, both models have been tested on relatively large datasets and they are capable of estimating networks within a reasonable amount of time.

In summarizing the obtained results through the analyses of different types and sizes of datasets, we find that both the precision and recall of the $GRN_{Multi\text{-}sources}$ demonstrate significant improvement. In analyzing the synthetic data of size 20, 50 and 100, the model shows 300-500% improvement over the existing models. The other two models, GRN_{Phase} and $GRN_{Co\text{-}expressed}$ show a similar or slightly better precision and recall. In the analysis of real microarray data, the model exhibits even better performance. The existing models fail to estimate the network in a week whereas $GRN_{Multi\text{-}sources}$ learns the network in couple of minutes with higher precision (13%) and recall (15%). This is a remarkable achievement of the model in constructing a large-scale network from noisy microarray data. The other two models discussed in this thesis, GRN_{Phase} and $GRN_{Co\text{-}expressed}$ estimate networks in a reasonable amount of time with a very low number of true positives. While comparing the precision and recall of the GRN models proposed in this thesis, we find $GRN_{Multi\text{-}sources}$ demonstrate over 1000% improvement over GRN_{Phase} and $GRN_{Co\text{-}expressed}$.

Since this chapter focuses on evaluating the performance of the proposed GRN models through the analysis of different data, an obvious question may arise; that is whether the model performance is data-dependent or not. None of the GRN models proposed in this thesis are data-dependent; that is the performance of the models is not directly dependent on the

gene expression data used for reconstructing the network. However, the performance of the models can vary due to the noise, incompleteness and other factors in the dataset. Though our models show steady performance on the analysis of both noisy and incomplete data, we have not performed any thorough analysis to quantify the robustness of the models. It is also worth mentioning that the performance of the models may vary in cases where we choose to include different set of genes in the data as shown in Table 7.1, 7.2 and 7.3. In these tables, for a given network size, we have generated 5 different datasets including different genes and the performance of the models slightly varies from dataset to dataset. Finally, all the experiments in this chapter confirm that the performance of the GRN models varies with the network dimension. Then again, the number of samples in the dataset may also contribute to the variation in model performance which is commonly known as stability analysis of the model. We leave this analysis as future work.

We conclude this chapter with the observation that the utilization of various biological knowledge extracted from available data sources, plays an important role in the construction of successful GRN model.

CHAPTER 8

SCALABILITY ANALYSIS

This chapter focuses on the scalability analysis of the DBN-based GRN models discussed in this thesis. In constructing gene regulatory networks (GRN), most of the computational models suffer from the curse of dimensionality; that is, their performance deteriorates with the increase in the network dimension. We consider a model as scalable if it maintains its level of performance when tested on larger datasets. The scalability of the models is investigated through the analysis of two experimental datasets of the yeast cell cycle. We start the analysis with a medium-scale network of 50 genes and increase the size of the network in steps of 50 new genes. Five such successive growths in network size result in six different datasets including 50, 100, 150, 200, 250 and 300 genes respectively. To evaluate the performance of the GRN models, we compute the two benchmark criteria, precision and recall in conjunction with the computation time.

Through comprehensive analysis, we demonstrate that $GRN_{Multi\text{-}sources}$ maintains its performance level in terms of both precision and recall when tested on larger datasets. Most importantly, the model estimates large-scale networks within minutes whereas other models fail to estimate them within a week. The models, GRN_{Phase} and $GRN_{Co\text{-}expressed}$ are capable of estimating networks including 300 genes within acceptable time; however the precision and recall suffers drastically as the network size grows. The two existing models GRN_{Murphy} and GRN_{Zou} show high sensitivity to the dimension of the network.

8.1 Introduction

In the post-genomic era, much attention has been given to the quest of developing computational methods for reconstructing GRNs. The performance of a computational model depends on the interrelationship between the sample size, data dimensionality, and model complexity (Wang 2008). The benchmark criteria, namely, precision and recall have been widely accepted by the scientific community for evaluating the performance of the model. However, the precision of most of the available models tends to deteriorate as the size of the network grows. This phenomenon is commonly known as the 'curse of dimensionality' (Duda et al. 2001) and leads to another decisive factor for the successful reconstruction of GRN which is the scalability of the model. In general, scalability is defined as a characteristic of a model that maintains its level of performance or efficiency when tested by larger operational demands. For the GRNs, a model is designated as scalable if it can estimates networks of different sizes (from tens to thousands genes) by preserving a steady performance level.

The biological data generated from the high-throughput technologies are characteristically high dimensional. To illustrate their high dimensionality, consider the simplest eukaryotic cell of budding yeast. There are more than 6000 protein-coding genes in the yeast cell and 800 of them have been identified as cell cycle regulated (Spellman et al. 1998). This implies that the GRN in the yeast cell cycle includes at least 800 genes. For a successful reconstruction of gene regulation in the yeast cell cycle, the computational models must have the capability to estimate networks containing hundreds of genes. At present, the performance of most of the models has been tested on small to medium scale networks including tens of genes only. Though a high value of precision/recall signifies the effectiveness of the model, the performance may deteriorate as the size of the network grows. Therefore, scalability analysis has become an integral part for investigating the effectiveness

of the GRN models. In this chapter, we perform the scalability analysis of the proposed GRN models through a series of experiments on two separate sets of experimental data (as used in chapters 4 and 5) with varying network dimension.

8.2 Some Existing Scalable Approaches

There is a growing body of literature that focuses on the development of scalable approaches for reconstructing GRN from high-dimensional microarray data. Some of these approaches incorporate other sources of biological data such as gene knockout data to address the dimensionality problem; others integrate modified algorithms to deal with the high computation time of the original models. In this section, we briefly review some scalable methods with a particular interest on the network size on which they have been tested on.

Rogers and Girolami (2005) studied a Sparse Bayesian regression algorithm for estimating interactions between genes. By using data from gene knockout experiments they decomposed the entire network into smaller sub-networks, each of which corresponds to one specific gene and its potential regulators. Finally, they applied the marginal likelihood maximization algorithm on individual sub-networks. The model has been tested on simulated data containing 30 genes only. The authors claimed that the model can be applied on large-scale networks as the independent sub-networks can easily be computed in parallel over several machines.

Huang et al. (2007) introduced two algorithms, a modified information-theory based Bayesian network algorithm and a modified association rule algorithm for estimating large-scale networks from microarray data. They assessed the scalability of their algorithms through the analysis of simulated datasets of varying sizes including 10, 50 and 200 genes. The experimental results showed that both algorithms estimated networks of 200 genes with 25%

accuracy. However, they have not discussed the computation time of their proposed algorithms which leaves a question on the further scalability of the techniques. In a comparative study, Sirbu et al. (2010) investigated the performance and scalability of different evolutionary algorithms (EA) in reconstructing GRN from microarray data. They performed scalability analysis of three methods on four different simulated datasets including 10, 20, 30 and 50 genes respectively. The authors concluded that two of the EA-based methods showed reasonable performance for networks including up to 30 genes and failed to estimate the larger network including 50 genes only. Tan et al. (2010) studied a scalable approach that uses a control policy to discard genes which are less important in the control of GRN in a preprocessing step. They computed a scoring function for each gene which indicates the relevance of the gene in the regulatory control. The genes which have a score less than a predefined threshold are ignored and removed from the dataset prior to modeling. To investigate the scalability of the method, the authors have analyzed both simulated and experimental data. Two different synthetic networks have been modeled, each containing 8 genes. The analysis of simulated data concluded that the selection of the predefined threshold has a huge impact on the performance of the approach.

8.3 Experimental Setup

The experiments are conducted on a computer system with Intel® Core ™ i5 CPU 760 @ 2.8 GHz and 4 GB RAM, running Windows 7 (Professional). The free statistical software, R version 2.15.0 (R Core Team 2009), is used to impute missing values in the real experimental dataset and find groups of co-expressed genes. Together with R, we have also used Bayesian Net Toolbox (BNT) which is written in MATLAB and freely provided by Murphy (2001a) to construct Dynamic Bayesian Networks (DBN). The experiments are setup

and run under the MATLAB environment with version 7.11.0. (R2010b)

8.4 Scalability Analysis of the Proposed GRN Models

Through the scalability analysis, we examine the effect of the curse of dimensionality on the performance of the GRN models as the network size grows. In this section, we perform the scalability analysis of the GRN models, GRN_{Phase}, $GRN_{Co\text{-}expressed}$ and $GRN_{Multi\text{-}sources}$ as discussed in chapter 4, chapter 5 and chapter 6 of this thesis. The two existing DBN-based methods, GRN_{Murphy} and GRN_{Zou} are also included in the analysis. As discussed earlier, the DBN structure learning algorithm shows exponential growth in computation time with the number of genes in the dataset. We speculate experiments with a large number of genes would become infeasible on our desktop PC. Hence, we set 7 days as a cut-off time for running an individual experiment.

Since the key focus of this thesis is to discover GRN using biological domain knowledge, we apply the aforementioned methods on two experimental datasets of the yeast cell cycle (Pramila et al. 2006). The same datasets have been previously used in chapters 4 and 5 which are named as alpha30 and alpha38. Each of them contains the expression levels of 4775 yeast genes over 22 time points. We start the analysis with a medium size network of 50 genes and add 50 more genes to the dataset as we progress. Unlike the models discussed in section 8.2, we choose to start with a medium-scale network as the effect of the curse of dimensionality is negligible for small-scale networks. We continue adding genes until any of the models reaches the cut-off time of 7 days; that is, the model fails to estimate the network within the cut-off time. Following this process, we come up with six datasets including 50, 100, 150, 200, 250 and 300 genes respectively.

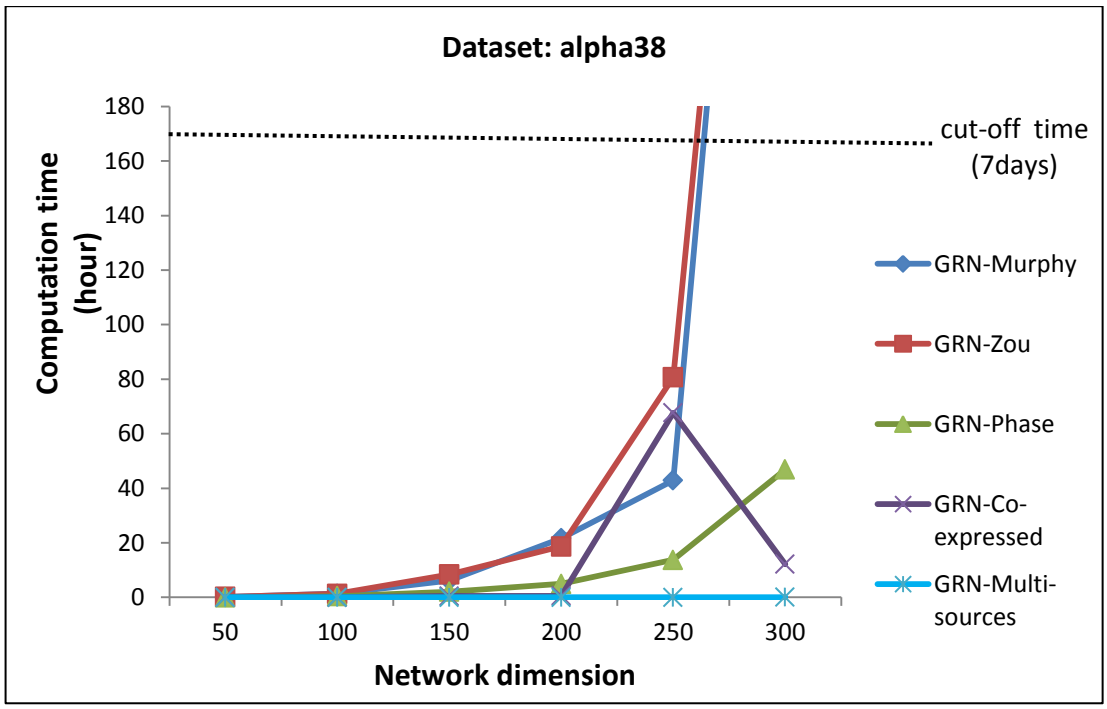
To compare the performance of the models with the increasing number of genes in the

dataset, we compute the benchmark criteria of precision and recall together with computation time. Since F-measure is the harmonic mean of the recall and precision and show similar performance level with precision, we exclude it from this analysis. In the following subsections, we demonstrate the scalability of the GRN models from the perspective of these three criteria.

8.4.1 Computation time

To investigate the effect of the curse of dimensionality on the computation time of the models with increasing network size, we plot them as shown in Figure 8.1. For both the datasets, the computation time of the first four models shows steady growth for medium-scale networks including up to 200 genes. For larger networks, GRN_{Murpy} and GRN_{Zou} exhibit sharp increase in computation time and they fail to learn networks of size 300 within the cuff-off time of 7 days. In contrast, GRN_{Phase} shows a gradual increase in the computation time for large-scale networks and can handle networks of 300 genes within 3 days. We speculate that the computation time of GRN_{Phase} would rise sharply for further increase of the network dimension. This is due to the fact that the model divides genes in three overlapping clusters corresponding to the phases of the cell cycle. As we add more genes in the dataset, the size of individual clusters grows which consequently increases the computation time of the model by factors that are exponential to the size of the individual clusters. $GRN_{Co\text{-}expressed}$ shows a slow increase in computation time for dataset alpha30 as the network size grows. However, for dataset alpha38, it exhibits interesting pattern. The computation time increases sharply for the network size of 250 and it falls as we add 50 more genes. This is because the model divides genes into k optimal number of clusters by observing the data and the computation time of the model is exponential in k.

A)



B)

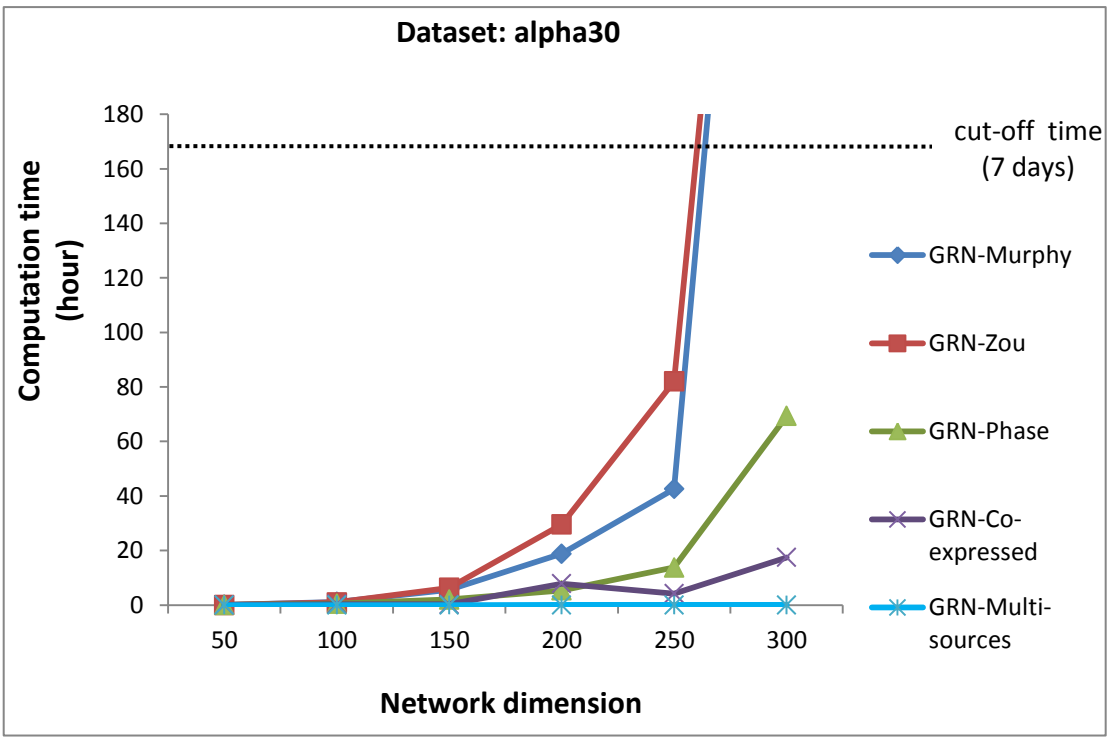


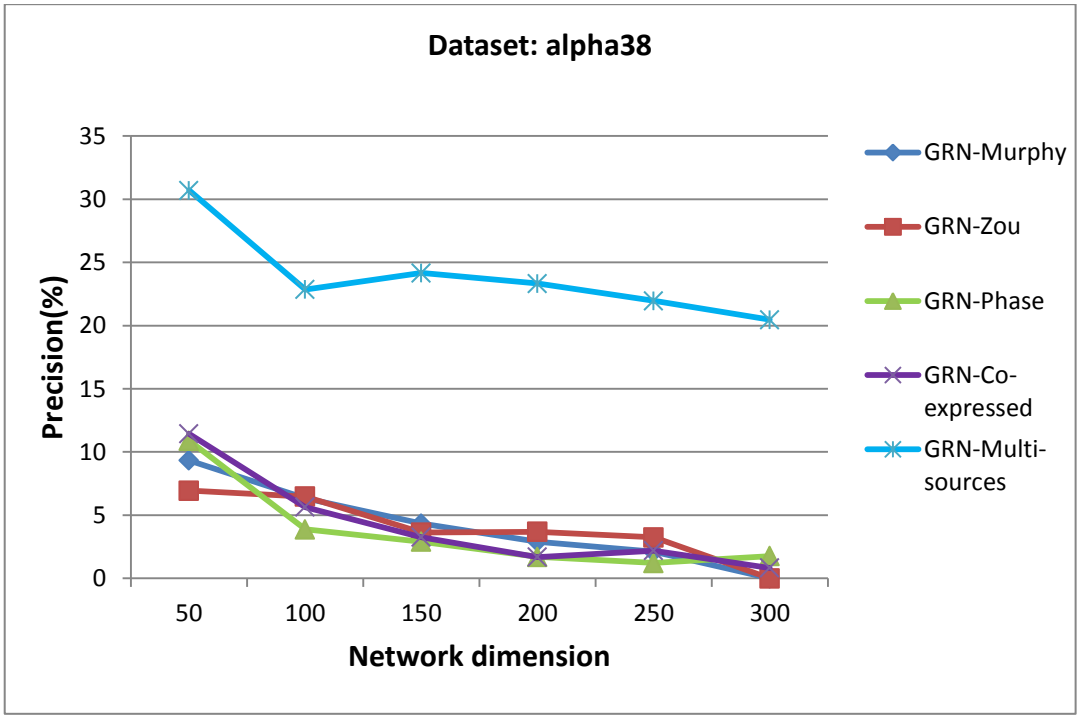
Figure 8.1: Comparison of the computation time of five DBN-based GRN models. The models are applied on six different datasets including 50, 100, 150, 200, 250 and 300 genes respectively. The dotted line shows the cut-off time for running an experiment. A) Dataset alpha38 and B) Dataset alpha30.

As shown in Figure 8.1, $GRN_{Multi\text{-}sources}$ estimates networks of 300 genes within 2 mins. This is a significantly improved performance compared to the other models. We speculate that the model is scalable to networks with even thousands of genes. This is because the size of the potential regulator set for a target gene increases at a much slower rate than the dimension of the network. Then again, the model decomposes the entire network into smaller sub-networks by incorporating PPI and TFBS data, where each sub-network consists of a specific gene and its potential regulators. Therefore, the DBN structure learning algorithm can easily be applied on individual sub-networks in parallel across several machines and the entire network can be estimated in an acceptable time.

8.4.2 Precision

Although computational feasibility is an important criterion for a successful reconstruction of GRN, biologists are more interested on the precision of the estimated network. To see how the precision of the GRN models is affected with the varying network dimension, we plot them as shown in Figure 8.2. Among the five models, $GRN_{Multi\text{-}sources}$ shows significantly higher precision for both the datasets. Although the precision of the models decreases gradually as the network size grows, $GRN_{Multi\text{-}sources}$ preserves a reasonable level which is approximately 20% for larger networks including 300 genes. In contrast, the other models show consistently poor precision for both the datasets. For larger networks of size 300, the precision of the models reaches as low as 1% (approx) which can be easily achieved with randomly generated networks as we discussed in chapter 7. Therefore, we conclude this analysis with the findings that the two existing models (GRN_{Murphy} and GRN_{Zou}) along with GRN_{Phase} and $GRN_{Co\text{-}expressed}$ severely suffer from the curse of dimensionality, although the latter two models preserve a slightly better precision level.

A)



B)

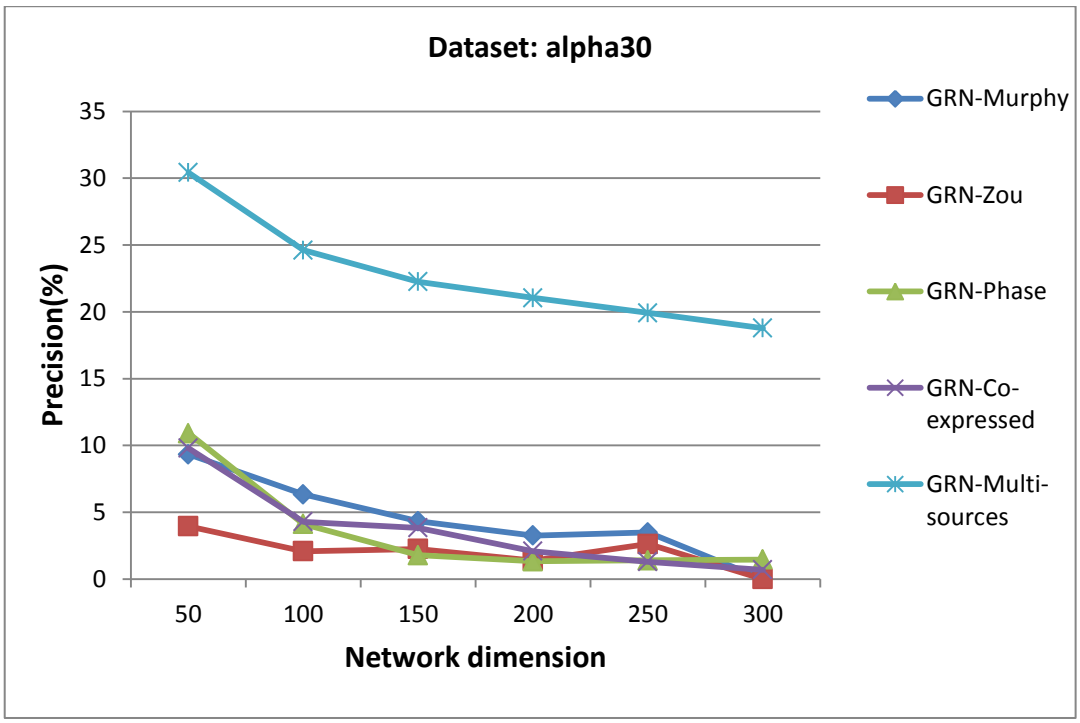


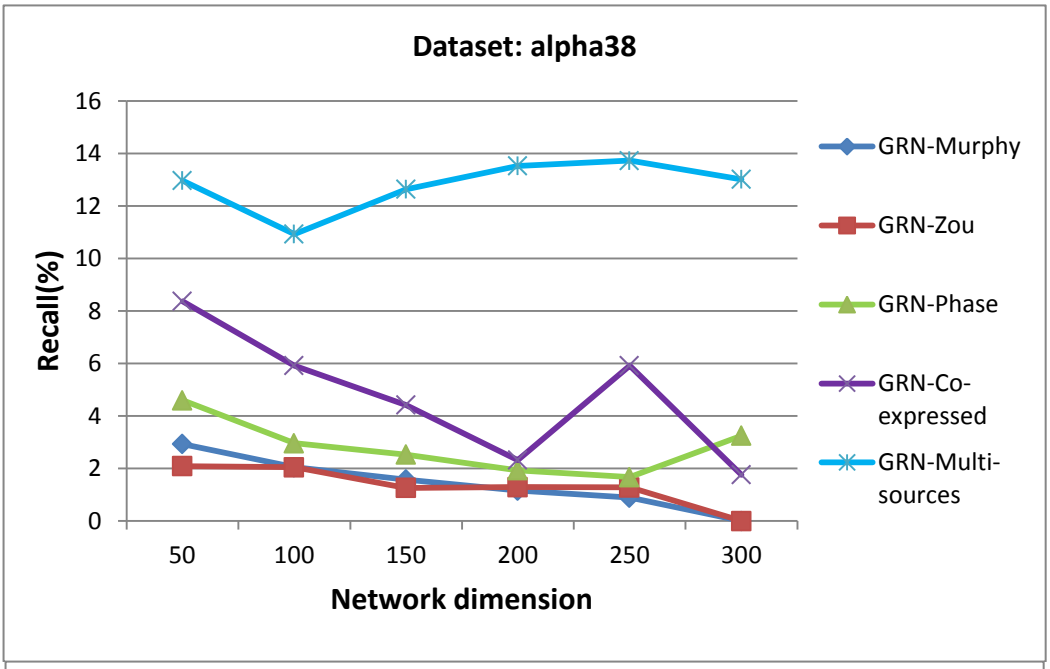
Figure 8.2: Comparison of the precision of five DBN-based GRN models. The models are applied on six different datasets including 50, 100, 150, 200, 250 and 300 genes respectively. Precision of the estimated networks is set to zero in cases where the models fail to estimate the networks within the cut-off time of 7 days. A) Dataset alpha38 and B) Dataset alpha30.

8.4.3 Recall

The recall of an estimated network represents the ratio of correctly identified connections out of the total known connections in the true network. From a biological perspective, a high recall value of the estimated network is a critical criterion for the successful reconstruction of GRN. To examine the effect of the curse of dimensionality on the recall of the DBN-based GRN models, we plot them in Figure 8.3. The existing two models, GRN_{Murphy} and GRN_{Zou} shows consistently low recall value for both datasets. In particular, GRN_{Zou} estimates a very low number of true positive connections as the network size grows, whereas GRN_{Murphy} maintains a steady performance. The other two models, GRN_{Phase} and $GRN_{Co\text{-}expressed}$ maintain better performance in comparison to the existing models, although the recall gradually falls with the increasing network size. For dataset alpha38, $GRN_{Co\text{-}expressed}$ exhibits a sharp increase in the recall value for a network of size 250 and plunges again with the addition of 50 more genes. This is because the performance of $GRN_{Co\text{-}expressed}$ partially depends on the identification of the k optimal number of clusters. In contrast, $GRN_{Multi\text{-}sources}$ shows an interesting pattern in the recall value for both the datasets. The recall of the model gradually falls for medium-scale networks and it increases slowly for large-scale networks. Nevertheless, $GRN_{Multi\text{-}sources}$ exhibits a higher recall in comparison to the other models and maintains a steady level with the growing network size. The complete experimental results of the analysis of the six different datasets including a varying number of genes are summarized in the tables given below. Table 8.1 presents the results of the analysis of dataset alpha30 and Table 8.2 shows the analysis of dataset alpha38. In the tables, we list the precision, recall and computation time of each of the DBN-based GRN models for individual network sizes. We also present the number of true positives estimated by each of these models. The results in both tables confirm that $GRN_{Multi\text{-}sources}$ is the best performing model and preserves steady

performance level when tested on larger datasets. This finding motivates us to perform further analysis of the model.

A)



B)

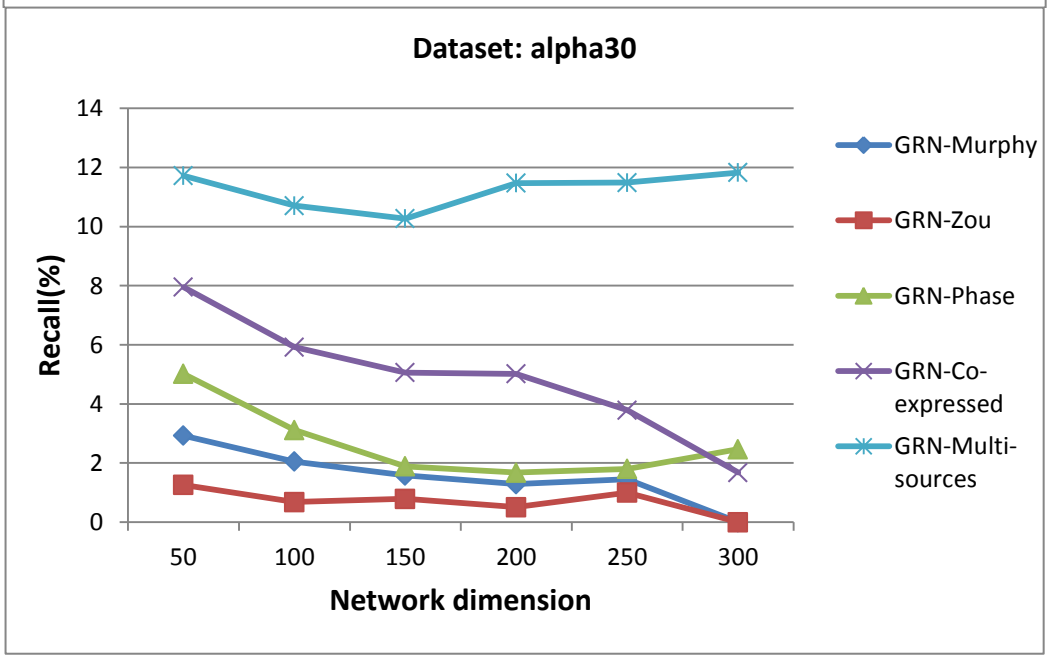


Figure 8.3: Comparison of the recall of five DBN-based GRN models. The models are applied on six different datasets including 50, 100, 150, 200, 250 and 300 genes respectively. Recall of the estimated networks is set to zero in cases where the models fail to estimate the networks within the cut-off time of 7 days. A) Dataset alpha38 and B) Dataset alpha30.

Table 8.1: Dataset alpha30. Comparison of performance among the five different DBN-based GRN models. The first two GRN models are existing works and the latter three have been proposed in this thesis. The models are applied on 6 different datasets of varying size including 50, 100, 150, 200, 250 and 300 genes over 22 time points. For each dataset, the best performing model has been highlighted in bold.

Network size	Methods	Number of true positives	Precision (%)	Recall (%)	Computation time
50	GRN_{Murphy}	7	9.33	2.93	5m 15s
	GRN_{Zou}	3	3.95	1.26	3m 22s
	GRN_{Phase}	12	10.91	5.02	1m 39s
	$GRN_{Co\text{-}expressed}$	19	9.8	7.95	2m 30s
	$GRN_{Multi\text{-}sources}$	**28**	**30.43**	**11.72**	**5s**
100	GRN_{Murphy}	9	6.34	2.05	1h 9m 25s
	GRN_{Zou}	3	2.08	0.68	59m 5s
	GRN_{Phase}	14	4.12	3.12	29m 33s
	$GRN_{Co\text{-}expressed}$	26	4.29	5.92	19m 36s
	$GRN_{Multi\text{-}sources}$	**47**	**24.61**	**10.71**	**16s**
150	GRN_{Murphy}	10	4.33	1.58	5h 34m 27s
	GRN_{Zou}	5	2.25	0.79	6h 17m 29s
	GRN_{Phase}	12	1.79	1.89	2h 2m
	$GRN_{Co\text{-}expressed}$	32	3.83	5.06	18m
	$GRN_{Multi\text{-}sources}$	**65**	**22.26**	**10.27**	**21s**
200	GRN_{Murphy}	10	3.26	1.29	18h 49m 29s
	GRN_{Zou}	4	1.39	0.51	1d 5h 34m
	GRN_{Phase}	13	1.33	1.68	5h 22m 29s
	$GRN_{Co\text{-}expressed}$	39	2.08	5.02	7h 50m 44s
	$GRN_{Multi\text{-}sources}$	**89**	**21.04**	**11.47**	**28s**
250	GRN_{Murphy}	13	3.49	1.45	1d 18h 34m
	GRN_{Zou}	9	2.62	1.00	3d 10h 1m
	GRN_{Phase}	17	1.4	1.8	13h 44m 52s
	$GRN_{Co\text{-}expressed}$	34	1.3	3.79	4h 12m 18s
	$GRN_{Multi\text{-}sources}$	**103**	**19.92**	**11.49**	**35s**
300	GRN_{Murphy}	-	-	-	-
	GRN_{Zou}	-	-	-	-
	GRN_{Phase}	25	1.46	2.47	2d 21h 20m
	$GRN_{Co\text{-}expressed}$	17	0.69	1.68	17h 29m 40s
	$GRN_{Multi\text{-}sources}$	**120**	**18.78**	**11.83**	**2m 7s**

Table 8.2: Dataset alpha38. Comparison of performance among the five different DBN-based GRN models. The first two GRN models are existing works and the latter three have been proposed in this thesis. The models are applied on 6 different datasets of varying size including 50, 100, 150, 200, 250 and 300 genes over 22 time points. For each dataset, the best performing model has been highlighted in bold.

Network size	Methods	Number of true positives	Precision (%)	Recall (%)	Computation time
50	GRN_{Murphy}	7	9.33	2.93	6m 52s
	GRN_{Zou}	5	6.94	2.09	10m 14s
	GRN_{Phase}	11	10.89	4.60	57s
	$GRN_{Co\text{-}expressed}$	20	11.43	8.37	14s
	$GRN_{Multi\text{-}sources}$	**31**	**30.7**	**12.97**	**4s**
100	GRN_{Murphy}	9	6.34	2.05	1h 16m 43s
	GRN_{Zou}	9	6.47	2.05	1h 16m 39s
	GRN_{Phase}	13	3.88	2.96	24m 22s
	$GRN_{Co\text{-}expressed}$	26	5.64	5.92	59s
	$GRN_{Multi\text{-}sources}$	**48**	**22.86**	**10.93**	**10s**
150	GRN_{Murphy}	10	4.33	1.58	6h 10m 6s
	GRN_{Zou}	8	3.62	1.26	8h 23m 41s
	GRN_{Phase}	16	2.9	2.53	2h 2m 45s
	$GRN_{Co\text{-}expressed}$	28	3.26	4.42	26m 54s
	$GRN_{Multi\text{-}sources}$	**80**	**24.17**	**12.64**	**22s**
200	GRN_{Murphy}	9	2.90	1.16	21h 39m 24s
	GRN_{Zou}	10	3.69	1.29	18h 41m 16s
	GRN_{Phase}	15	1.7	1.93	4h 55m 57s
	$GRN_{Co\text{-}expressed}$	18	1.69	2.32	29m 56s
	$GRN_{Multi\text{-}sources}$	**105**	**23.33**	**13.53**	**29s**
250	GRN_{Murphy}	8	2.1	0.89	1d 18h 55m
	GRN_{Zou}	11	3.24	1.28	3d 8h 43m
	GRN_{Phase}	15	1.22	1.67	13h 11m 22s
	$GRN_{Co\text{-}expressed}$	53	2.19	5.92	2d 19h 42m
	$GRN_{Multi\text{-}sources}$	**123**	**21.96**	**13.73**	**32s**
300	GRN_{Murphy}	-	-	-	-
	GRN_{Zou}	-	-	-	-
	GRN_{Phase}	33	1.76	3.25	1d 22h 52m
	$GRN_{Co\text{-}expressed}$	18	0.82	1.76	12h 7m 24s
	$GRN_{Multi\text{-}sources}$	**132**	**20.47**	**13.02**	**1m 49s**

8.5. Extended Scalability Analysis of $GRN_{Multi\text{-}sources}$

As we demonstrate in the previous section, $GRN_{Multi\text{-}sources}$ shows minimal sensitivity to the curse of dimensionality in comparison to the other DBN-based GRN models. In particular, the computational time of the model has dropped to a surprisingly low level. This result encourages us to apply the model on a dataset including up to 1000 genes. We exclude the other DBN-based GRN models from this analysis because they have shown poor performance in estimating large-scale networks. Since validation is a major challenge for analyzing experimental data, we choose simulated data from the DREAM4 project. A thorough description of the simulator and data generation process can be found in section 7.5 of chapter 7. From the yeast network, we extract 3 sub-networks of varying dimensions including 100, 500 and 1000 genes over 21 time points, where t_max = 100. The main focus of this analysis is to see how $GRN_{Multi\text{-}sources}$ scales to large-scale networks including up to 1000 genes. Therefore, we extract only one sub-network for each dimension. The experimental results of analyzing these medium to large scale networks are summarized in Table 8.3. Similar to section 8.4, we compute precision, recall and computation time of the estimated networks for each network size.

Table 8.3: Comparison of the performance of $GRN_{Multi\text{-}sources}$ in analyzing medium to large-scale networks including 100, 500 and 1000 genes respectively. The synthetic data are extracted from the DREAM4 project; each contains 21 time points.

Network size	Number of true positives	Precision (%)	Recall (%)	Computation time
100	86	25.37	32.20	6s
500	243	12.03	16.23	2m 34s
1000	307	10.67	10.67	6m 18s

Much to our surprise, the model estimates a network of 1000 genes in less than

7 mins which is a significant achievement in addressing the dimensionality problem of GRN. The increase in computation time as the network size grows in graphically presented in Figure 8.4. To compare the precision and recall of the estimated networks with the growing number of genes in the dataset, we plot them in Figure 8.5. The model estimates a network of 100 genes with a reasonably higher precision and recall value. However, as we increase the network size to 500, both of them show about 50% drop in values. With the further growth of the network size, the recall of the model drops slightly (10%), whereas the precision shows almost 60% fall in value. We observe these sharp drops because of the huge increase in the network dimension which is from 500 to 1000. We speculate that the small sample size (21 time points) in comparison to the number of genes (1000 genes) in the dataset is the main reason for such performance of the model. The effect of insufficient data is especially observable on the precision of the model as it dramatically increases the number of false positives with the increase in the network size.

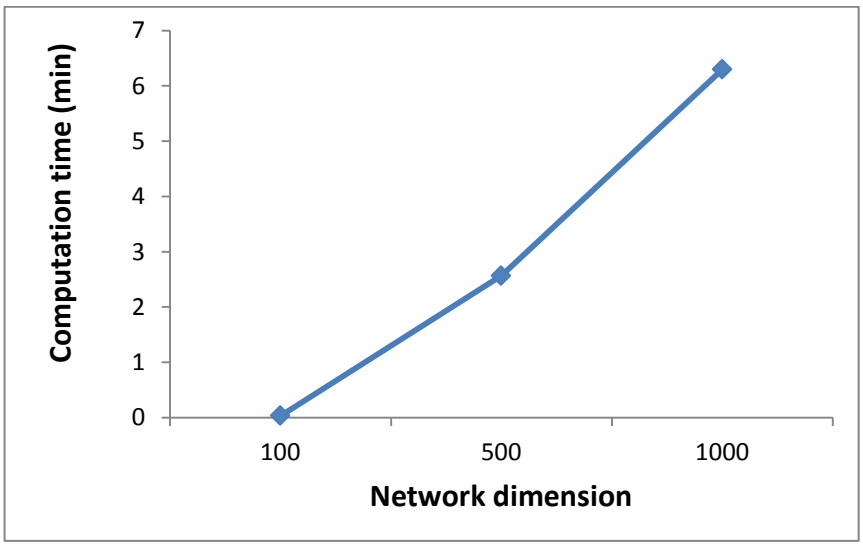


Figure 8.4: Computation time of $GRN_{Multi\text{-}sources}$ with varying network sizes including 100, 500 and 1000 genes.

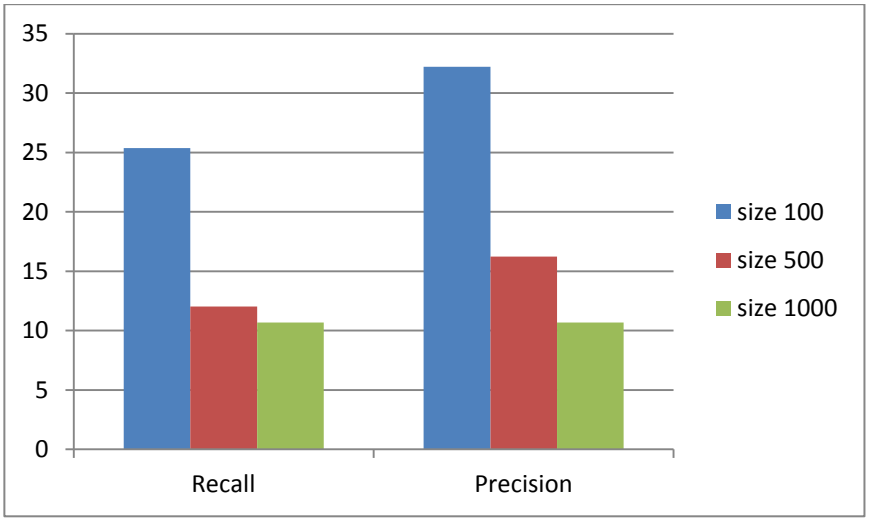


Figure 8.5: Comparison of the precision and recall of $GRN_{Multi\text{-}sources}$ with varying network dimensions including 100, 500 and 1000 genes.

8.6 Conclusion

In this chapter, we have discussed the scalability analysis of the DBN-based GRN models and studied the effect of curse of dimensionality on the performance of the models. Our analysis has demonstrated that $GRN_{Multi\text{-}sources}$ outperforms the other DBN-based GRN models and maintains scalability even for large-scale networks including up to 1000 genes. As discussed in chapter 7, we speculate that the incorporation of other sources of biological data in reducing the problem space is the key to such an achievement. Nevertheless, the model shows a dramatic fall in precision and recall as we increase the network dimension by hundreds of genes. We hypothesize that the small number of samples in the dataset is at the core of such deterioration in performance. In the future, we plan to perform stability analysis to study the effect of the sample size on the performance of the model.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

Since the completion of sequencing the genomes of various organisms, especially the human genome, a new field of research, namely functional genomics has emerged. This field primarily concerns with patterns of gene expression during various conditions. The emergence of this new field has accompanied a number of technological advancements such as microarrays for measuring the patterns of gene expression on a genome-wide scale. The outcome of such a development is the generation of a large amount of genomic data which are expected to be utilized to answer a wide range of biological questions. For instance, the data can provide insight in disease diagnosis; that is, which genes cause a particular disease. Therefore, the key goals of functional genomics research are to explain what the genes do and study when, where and how they are expressed as an orchestrated system. In this thesis, we have focused on the discovery of this orchestrated system, commonly known as gene regulatory networks (GRN), with the application of dynamic Bayesian networks (DBN).

The computational reconstruction of GRN from data faces a range of challenges. Some of these challenges have been highlighted in chapter 1 of this thesis, whereas chapter 3 includes a thorough discussion of the existing computational methods and their associated challenges. In this thesis, we mainly focus on three major challenges: 1) the high computational time of GRN reconstruction 2) the low number of true positives in the estimated network and 3) low scalability; that is, the performance of the model deteriorates as the network size grows. The first challenge arises from the high dimensionality of the

biological data and the data scarcity contributes significantly towards the second challenge. As discussed in chapters 4, 5 and 6 of this thesis, we have addressed these challenges by exploiting biological domain knowledge in the GRN reconstruction process. In chapters 7 and 8, we have quantified to what extent the chosen challenges have been addressed by such exploitation in comparison with the classic DBN-based GRN model. In the following sections, we summarize our contributions in addressing those challenges and propose several promising future research directions.

9.1 Summary of Contributions

This thesis has made several contributions in addressing the current challenges of computational reconstruction of GRN from microarray gene expression data. Most remarkably, the computational time of the computational methods (DBN in this thesis) has reduced to a surprisingly low level, whereas the number of true positives in the estimated network has improved significantly. This improvement would potentially help biologists to explain the observed gene expression phenomenon on a larger scale such as interactions of all the cell-cycle regulated genes during the cell division process. The particular contributions are listed as follows.

1. The DBN-based GRN model (GRN_{Phase}) as explained in chapter 4 addresses the high computation time of the reconstruction process. We have illustrated a novel way of incorporating biological domain knowledge of the cellular process under study in decomposing the entire problem into several sub-problems. As a consequence of this decomposition, the computation time of the model has been reduced significantly. At the same time, the utilization of biologically relevant knowledge marginally increases the number of true positives in the estimated network.

2. Most of the available GRN models identify the regulators for individual genes in the process of building up the entire network. However, it is equally important to find co-regulation of co-expressed genes. We have discussed a GRN model ($GRN_{Co\text{-}expressed}$) based on DBN in chapter 5 which finds groups of co-expressed genes and their corresponding co-regulators. The partitioning of the entire network into k modules (groups of co-expressed genes) and finding regulation among these modules have reduced the computation time of the model significantly. The model has also shown a slight increase in the number of true positives in the estimated network compared to the two other exiting models. Gaining precise knowledge of co-expressed and co-regulated genes is vital as it can assist biologists to predict the characteristics of unknown genes.
3. We have illustrated another DBN-based GRN model ($GRN_{Multi\text{-}sources}$) in chapter 6 that incorporates two sources of biological data. These are protein-protein interaction (PPI) data and transcription factor binding site (TFBS) data. A good number of models have already incorporated such data in reconstructing GRN from gene expression analysis. The novelty of our model is that we have considered the genetic interactions as possible regulatory relationships among two given genes, whereas most of the existing models incorporate the physical interactions which specify the physical associations between proteins. With the utilization of these two data sources, we have removed extraneous genes from the potential regulators list for each target gene. As a consequence, the computation time of the model has been reduced to a remarkably low level with a significant increase in the number of true positives in the estimated network.
4. To validate a GRN model, most of the studies analyze mainly simulated data including up to 100 genes and then apply the model on experimental data containing tens of genes only. In contrast, in this thesis, we have validated the performance of our proposed GRN

models through the analysis of three different experimental datasets including up to 300 genes. Two of these datasets are complete, whereas the other has missing values in it. The analysis of both noisy and incomplete data has also demonstrated the robustness of our proposed model to some extent. Through different experiments, we have shown that our proposed GRN models exhibit consistent performance for all the three datasets. It is worth to mention that the model proposed in chapter 6 demonstrates 300-500% improvement over the existing models in analyzing simulated data from the DREAM4 project whereas the models in chapters 4 and 5 shows similar or slightly better precision and recall. This improvement in performance is even more remarkable in analyzing real microarray data including hundreds of genes.

5. One of the noteworthy contributions of this thesis is the scalability analysis of the proposed GRN models. In the literature, a number of scalable approaches have been reported. However, most of them have performed scalability analysis including up to 50 genes only where there are at least 800 genes in the GRN of the unicellular yeast cell cycle. In this thesis, we have performed scalability analysis by applying the models on both simulated and experimental data. The model proposed in chapter 6 of this thesis showed high scalability on a series of simulated datasets including up to 1000 genes. The other two models as proposed in chapter 4 and chapter 5 have shown poor performance with the growth of the network size.

9.2 Future Work

Despite significant improvement in the computation time of the GRN models, the number of true positives in the estimated network is still low. We think, there are several open research issues that need further investigation.

1. The use of prior knowledge about the network structure or the cellular process under study has been shown to be effective in the scope of this thesis. However, in the proposed GRN models, we have utilized prior knowledge to narrow down the search space for finding the structure that fits the data best. In other words, the prior knowledge has been used in a pre-processing step. To learn the topology of the network, the DBN structure learning algorithm uses a prior probability which is uninformative-that is, it is a uniform prior, where every network structure is equally likely. A possible future work may consider the use of prior knowledge as an informative prior, which is not uniformly distributed over each possible network structure. We expect that the integration of informative prior in structure learning would improve the precision of the estimated network.
2. The GRN models proposed in this thesis, especially $GRN_{Multi\text{-}sources}$, have shown significant improvement in addressing the high dimensionality problem of GRN reconstruction. The model ($GRN_{Multi\text{-}sources}$) has also estimated large-scale networks including 300 genes with a reasonable precision level of 20% (approx.). However, this precision level is still low from a biological point of view. We speculate that the imprecision and the lack of accuracy in the collecting the PPI and TFBS data are the key factors to such a low level of precision. Future work may look for incorporating other sources of biological data such as Chip-Chip data to enrich the set of potential regulators for each target gene.
3. Throughout the thesis, we have claimed that the low number of samples in the experimental data is one of the major challenges in reconstructing GRN with a high level of precision. Future work may investigate this claim through the analysis of simulated data of varying sample sizes as the size of the network grows. Such a study may also provide some level of quantification for model stability.

4. In this thesis, we have chosen DBN as our preferred model to demonstrate the effectiveness of incorporating biological domain knowledge in reconstructing GRN. However, the reconstruction accuracy of the DBN-based GRN models is still low. We observe that the inherent challenges that have been discussed in chapter 3 contribute significantly for such low accuracy. We hypothesize that the principles of our GRN models can be used in conjunction with any other computational methods such as S-System. Therefore, in a further study, it would be interesting to investigate this hypothesis as well as the potency of other available models in reconstructing GRN from experimental data.
5. The GRN models that have been discussed in this thesis are discrete models. We have used the mean of the gene expression to discretize the time series, which is exceptionally susceptible to the presence of outliers in the noisy microarray data. Thus, it would be more effective to apply discretization methods which take into account gene-gene relationships such as correlation.

9.3 Final Remarks

This thesis has utilized biological domain knowledge in reconstructing GRNs from the gene expression data. The problem of GRN reconstruction presents significant challenges to existing computational methods. The models proposed in this thesis have shown substantial progress on addressing some of these challenges, in particular on developing a scalable model. Therefore, we expect that the work in this thesis will serve as a foundation for further advances.

BIBLIOGRAPHY

Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression pattern under the Boolean network model. In *Pacific Symposium on Biocomputing*, pp. 17-28.

Amon, A., Tyers, M., Futcher, B., and Nasmyth K. (1993). Mechanisms that help the yeast cell cycle clock tick: G2 cyclins transcriptionally activate G2 cyclins and repress G1 cyclins. *Cell*, 74: 993–1007.

Aris, V.M., Cody, M.J., Cheng, J., Dermody, J.J., Soteropoulos, P., Recce, M., and Tolias, P.P. (2004). Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer. *BMC Bioinformatics*, 5:185.

Baeck, T., Fogel, D. B., and Michalewicz, Z., (2000). *Evolutionary Computation 1: Basic Algorithms Operators*. Institute of Physics Publishing Bristol and Philadelphia.

Bahler, J. (2005). Cell-Cycle Control of Gene Expression in Budding and Fission Yeast. *Annual Review of Genetics*, 39: 69–94.

Baldi, P., and Hatfield, G. W. (2002). *DNA Microarrays and Gene expression: from experiemnts to data analysis and modelling*. Cambridge university press.

Banerjee, N., and Zhang, M.Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Research*, 31:7024–7031.

Bar-Joseph, Z., Gerber, G.K., Lee, T. I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., and Gifford, D.K. (2003). *Nature Biotechnology*, 21:337–1342.

Beer, M.A., and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, 117:185–198.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 22:

757-62.

Bindea, G., Mlecnik, B., Hackl, H, Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman1, W., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 25(8): 1091-1093.

Blasi, M.F., Casorelli, I., Colosimo, A., Blasi, F. S., Bignami, M., and Giuliani, A. (2005). A recursive network approach can identify constitutive. *Physica A 348:349-370.*

Brazhnik, P., de la Fuente, A., and Mendes, P (2002). Gene networks: how to put the function in genomics. *Trends in Biotechnology*, 20:467–472.

Breitkreutz, B.J., Stark C., and Tyers, M. (2003). The GRID: the General Repository for Interaction Datasets. *Genome Biology*, 4:R23

Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bähler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Research,* 36: D637-640.

Carlsson, P., and Mahlapuu, M. (2002). Forkhead transcription factors: key players in development and metabolism. *Dev. Biol,* 250:1 – 23.

Chaitankar, V., Ghosh Preetam, P., Edward, J., Gong, P., Deng, Y., and Zhang, C. (2010). A novel gene network inference algorithm using predictive minimum description length approach. *BMC Systems Biolog,* 4(Suppl 1):S7.

Chang, L.W., Nagarajan R, Magee, J.A., Milbrandt, J., and Stormo, G.D. (2006). A systematic model to predict transcriptional regulatory mechanisms based on over representation of transcription factor binding profiles. Genome Research, 16:405-13.

Chaturvedi, I., Sakharkar, M.K., and Rajapakse, J.C. (2007). Validation of gene regulatory networks from protein-protein interaction data: application to cell-cycle regulation. *Lecture Notes in Computer Science (Pattern Recognition in Bioinformatics)*, 4774: 300-310.

Chen, C.C., and Zhong, S. (2008). Inferring gene regulatory networks by thermodynamic modeling. *BMC Genomics*, 9 Suppl. 2:S19.

Chen, T., He H., and Church, G.M. (1999). Modeling gene expression with differential

equations. *Pacific Symposium on Biocomputing*, 4**:**29-40.

Cho, K.H., Choo, S.M., Jung, S.H., Kim, J.R., Choi, H.S., and Kim, J. (2007). Reverse engineering of gene regulatory networks. *IET System Biology*, 1(3):149–163.

Cinquemani, E., Milias-Argeitis, A., and Summers, S. (2008). Stochastic dynamics of genetic networks: modelling and parameter identification. *Bioinformatics*, 24:2748-54.

Crick, F. H. (1970). Central dogma of molecular biology. *Nature*, 227:561-563.

Curto, R., Voit, E.O., and Sorribas A. *(1997.)* Validation and steady-state analysis of a power-law model of purine metabolism in man. *Biochemical Journal*, 324:761-775.

de Brevern, A.G., Hazout, S., and Malpertuy, A. (2004): Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, 5:114.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. Journal of Computational *Biology*, 9(1):67-103.

de Lichtenberg, U., Jensen, L.J., Fausboll, A., Jensen, T.S., Bork, P. and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, 21:1164-1171.

Dennis, G. Jr., Sherman, B.T., Hosack DA, Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4:R60.

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics, 16(8):707-726.

Di Bernardo, D, Gardner, T.S., and Collins J.J. (2004). Robust identification of large genetic networks. In *Pacific Symposium on Biocomputing*, 9:486-97.

Di Liegro, C.M., Bellafiore, M., Izquierdo, J.M., Rantanen, A., and Cuezva, J.M. (2000). 3'-untranslated regions of oxidative phosphorylation mRNAs function in vivo as enhancers of translation. *Biochemical Journal*, 1:109–115.

Dojer, N., Gambin, A., Mizera, A., Wilczyński, B., and Tiuryn, J. (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, 7: 249-259.

Dougherty, E.R. (2001). Small sample issues for microarray-based classification. *Comp.*

Funct. Genomics, 2(1):28–34.

Duda, R.O., Hart, P.E., and Stork, D.G. (2001). *Pattern Classification*. Wiley, New York, 3rd edition.

Duin, R.P.W. (2000). Classifiers in almost empty spaces. *In 15th International Conference on Pattern Recognition*, Barcelona, Spain.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–14868.

Enriquez, J.A., Fernandez-Silva, P., and Montoya, J. (1999). Autonomous regulation in mammalian mitochondrial DNA transcription. *Biol Chem.*, 380:737–747.

Friedman N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*,
303(5659): 799-805.

Friedman, N. (1998). The Bayesian structural EM algorithm. *Proceeding of the Fourteenth conference on Uncertainty in artificial intelligence*, pp.129-138. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601-620.

Friedman, N., Murphy, K., and Russell, S. (1998) 'Learning the Structure of Dynamic Probabilistic Networks', In *Proceedings of fourteenth conference on uncertainty in artificial intelligence (UAI)*, pp. 139-147.

Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. (2002). A Structural EM Algorithm for Phylogenetic Inference. *Journal of Computational Biology*, 9(2): 331-353.

Gao, F., Foat, B.C., and Bussemaker, H.J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC *Bioinformatics*, 5:31-40.

Gao, S., and Wang, X. (2011). Quantitative utilization of prior biological knowledge in the Bayesian network modeling of gene expression data. *BMC Bioinformatics*, 12:359.

Gardner, T.S., di Bernardo, D., Lorenzo, D., Collins, J.J. (2003). Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science*,

301(5629):102-105.

Geiss, G.K., Carter, V.S.,, He, Y., Kwieciszewski, B.K.,, Holzman, T., Korth, M.J., Lazaro, C.A., Fausto, N., Bumgarner, R.E., and Katze, M.G. (2003): Gene expression profiling of the cellular transcriptional network regulated by alpha/beta interferon and its partial attenuation by the hepatitis C virus nonstructural 5A protein. *Journal of Virology*, 77(11):6367-75.

Gershon, D. (2002). Dealing with the data deluge. *Nature,* 416(6883):889–91.

Han, J.D. (2008). Understanding biological functions through molecular networks. *Cell Res*earch, 18:224-237.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. (2002). Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models. *Paificc. Symosium. on Biocomputing*, 437- 449.

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., and Botstein, D. (1999). Imputing Missing Data for Gene Expression Arrays. Technical Report, Division of Biostatistics, Stanford University.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R.(2009). Gene regulatory network inference: data integration in dynamic models-a review. Biosystems, 96(1):86-103.

Heckerman, D. (1994). Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research, Redmond, Washington.

Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington.

Hernminger, B.M., Saelim, B., and Sullivan, P.F. (2007). Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. *J Am Soc Inf Sci Tec*, 58:2341-52.

Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999). Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research*, 9: 1106-1115.

Ho, S.Y., Hsieh, C.H., and Yu, F.C. (2007*)*. An intelligent two-stage evolutionary algorithm for dynamic pathway identification from gene expression profiles. *IEEE/ACM Trans*

Comput Biol Bioinform, 4:648-60.

Huang, S. (1999). Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.*, 77:469-80.

Huang, Z., Li, J., Su, H., Watts, G.S, Chen, H. (2007). Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining. *Decision Support Systems*, 43(4): 1207-1225.

Hughes, J.D., Estep, P.W., Tarazoie, S., and Church, G.M. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296:1205–1214.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19:2271-2282.

Hwang, D., Schmitt, W.A., and Stephanopoulos, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, 18(9):1184-93.

Ideker, T., Thorsson, V., and Karp, R. (2000). Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design. *Pacific Symposium on Biocomputing*, 5:302-313.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, 31: 370–377.

Iyer, V.R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P.O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409:533–38.

Jagath, C.R., and Piyushkumar A.M. (2011). Stability of building gene regulatory networks with sparse autoregressive models. BMC Bioinformatics, 12(Suppl 13):S17.

Jain, A., Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell*, 19(2):153–58.

Jelinsky, S.A., Estep, P., Church, G.M., and Samson, L.D. (2000). Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4

links base excision repair with proteasomes. *Molecular Cell Biology*, 20(21):8157-67.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36: D480–D484.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:D354–357.

Karp, R. M., Stoughton, R., and Yeung, K. Y. (1999). Algorithms for choosing differential gene expression experiments. Proceedings of the third annual international conference on Computational molecular *biology (RECOMB '99),* pp. 208-217, New York, USA.

KEGG (2000). Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28:27–30.

Kepes, F. (2007). Biological networks. Vol 3 , World Scientific publishing.

Kikuchi, S., Tominaga, D., Arita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19*:*643-50.

Kim, H., Lee, J., Park, T. (2007). Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics*, 8:8-37.

Kim, S., Imoto, S. and Miyano, S. (2003). Dynamic Bayesian Networks and Nonparametric Regression for Nonlinear Modelling of Gene Networks from Time Series Expression Data. *Computational System Biology*, 75:57-65.

Kimura, S., Ide, K., and Kashihara, A. (2005). Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics*, 21:1154-63.

Klebanov, L., and Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biol Direct.,* 2: 9.

Koch, C., Schleiffer, A., Ammerer, G., and Nasmyth, K. (1996). Switching transcription on and off during the yeast cell cycle: Cln/Cdc28 kinases activate bound transcription factor SBF (Swi4/Swi6) at start, whereas Clb/Cdc28 kinases displace it from the promoter in G2. *Genes Dev,* 10:129–41.

Laubenbacher, R., and Stigler, B. (2004). A computational algebra approach to the reverse

engineering of gene regulatory networks. *Journal of Theoretical Biology*, 229:523–537.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z.,Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M.,Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel E., Gifford, D.K., and Young, R.A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804.

Lee, W.P., and Tzou, W.S. (2009). Computational methods for discovering gene networks from expression data. *Brief Bioinform*, 10(4): 408-423.

Lee, W.P., and Yang, K.C. (2008). A clustering-based approach for inferring recurrent neural networks as gene regulatory networks. *Neurocomputing*, 71:600-10.

Li, H., Wang, N., Gong, P., Perkins, E.J. and Zhang, C. (2011). Learning the tructure of gene regulatory networks from time series gene expression data. *BMC Genomics*, 12(Suppl 5):S13.

Li, P., Zhang, C., Perkins, E., J., Gong, P., and Deng, Y. (2007). Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics*, 8(Suppl 7): S13.

Liang, S., Fuhrman, S., and Somogyi, R. (1998). REVEAL, a General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

Linhart, C., Halperin, Y., Shamir, R. (2008). Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Research*, 18:1180-1189.

Loots, G.G., Chain, P.S., Mabery, S., Rasley, A., Garcia, E., and Ovcharenko, I. (2006). Array2BIO: from microarray expression data to functional annotation of co-regulated genes. *BMC Bioinformatics*, 7:307.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. J*ournal of Computational Biology*, 16:229.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (1999). A

combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86.

Mehra, S., Hu, W.S., and Karypis, G. (2004). A Boolean algorithm for reconstructing the structure of regulatory networks. *Metab Eng* ,6:326-39.

Michalewicz, Z., (1994). Evolutionary Computation Techniques for Nonlinear Programming Problems. *International Transactions in Operational Research*, 2:223-240.

Mobini, R., Anderson, B.A., and Erjefalt, J. (2009). A module-based analytical strategy to identify novel disease genes shows an inhibitory role of interleukin 7 Receptor in allergic inflammation. *BMC SYS Bio*, 3:19.

Monteiro, P.T., Teixeira, M.C., d'Orey, S., Tenreiro, S., Mira, N.P., Pais, H., Francisco, A.P., Carvalho, A.M., Lourenço, A.B., Sá-Correia, I., Oliveira, A.L., and Freitas, A.T. (2008). YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in *Saccharomyces cerevisiae***.** *Nucl. Acids Research*, 36:D132-D136.

Mootha, V.K., Handschin, C., Arlow, D., Xie, X., St Pierre, J., Sihag, S., Yang, W., Altshuler, D., Puigserver, P., Patterson, N., Willy, P.J., Schulman, I.G., Heyman, R.A., Lander, E.S., Spiegelman, B.M. (2004). Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc Natl Acad Sci U S A.* 101:6570–6575.

Murphy, K.P., and Milan, S. (1999). Modelling gene expression data using dynamic Bayesian networks. Technical report, MIT artificial intelligence laboratory.

Murphy, K. P. (2001a). Bayes Net Toolbox for Matlab. Computing Science and Statistic.

Murphy, K. P. (2001b). An introduction to graphical models. Technical report, MIT, Artificial Inteligence Laboratory.

Nam, H., Lee, K., and Lee, D. (2009). Identification of temporal association rules from time-series microarray data sets. *BMC Bioinformatics*, 10(Suppl 3): S6.

Nariai, N., Kim, S., Imoto, S., and Miyano, S. (2004). Using Protein-Protein Interactions for Refining Gene Networks Estimated from Microarray Data by Bayesian Networks. *Pacific Symposium on Biocomputing*, 336-347.

Nikitin, A., Egorov, S., Daraselia, N., Mazo, I. (2003). Pathway studio—the analysis and

navigation of molecular networks. *Bioinformatics*, 19:155-57.

Noman, N., Iba, H. (2007). Inferring gene regulatory networks using differential evolution with local search heuristics. *IEEE/ACM Trans Comput Biol Bioinform.*, 4:634-47.

Noort, V.V., Snel, B and Huynen, M.A. (2003). Predicting gene function by conserved co-expression. *Trends in Genetics*, 19 : 238-242.

Nguyen, X., Madhu C., Ross C., and Pramod P.W. (2012). Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network. *BMC Bioinformatics*, 13:131.

Ohler, U., Niemann, H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, 17(2):56-60.

Okuda, S., Yamada, T., Hamajima M, Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, *36:W423-W426.*

Ong, I., Glasner, J.D. and Page, D. (2002). Modelling Regulatory Pathways in E. coli from Time Series Expression Profiles. *Bioinformatics*, 18(S1):S241-S248.

Ota, K., Goto, S., and Kanehisa, M. (2004).Comparative analysis of transcriptional regulation in eukaryotic cell cycles. *Proceedings of the Fourth International Workshop on Bioinformatics and Systems Biology*, 4:26-27.

Ott, S., Imoto, S., and Miyano.S. (2004). Finding optimal models for small gene networks. *Pacific Symposium on Biocomputing*, 9:557–567.

Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alché-Buc, F. (2003). Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19(Suppl 2) : 138-48.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, 24, 3836–3845).

Piétu, G., Mariage-Samson, R., Fayein, N., Matingou, C., Eveno, E., Houlgatte, R., Decraene, C., Vandenbrouck, Y., Tahi, F., Devignes, M., Wirkner, U., Ansorge, W., Cox, D., Nagase, T., Nomura, N., and Auffray, C. (1999). The Genexpress IMAGE Knowledge Base of the Human Brain Transcriptome: A Prototype Integrated Resource for Functional

and Computational Genomics. *Genome Res*earch, 9:775-792.

Pilpel, Y., Sudarsanam, P., and Church, G.M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29:153–159.

Posekany, A., Felsenstein, K., and Sykacek, P. (2011). Biological Assessment of Robust Noise Models in Microarray Data Analysis. *Bioinformatics*, 27: 807-814.

Pramila, T., Shawna, M., William, S.N. and Linda, L.B., (2006). The Forkhead Transcription Factor HCm1 Regulates Chromosome Segregation Genes and Fills the S-phase Gap in the Transcriptional Circuitry of the Cell Cycle. *Genes & Development*, 20:2266-2278.

Prill, R.J., Marbach, D., Saez-Rodriguez, J., Sorger, P.K., Alexopoulos, L.G., Xue, X., Clarke, N.D., Altan-Bonnet, G., and Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, 5(2):e9202.

Qian, J., Lin, J, Luscombe, N.M., Yu, H., and Gerstein, M. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19:1917–1926.

Qin, Z.S., McCue, L.A., Thompson, W., Mayerhofer, L., Lawrence, C.E., and Liu, J.S. (2003). Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nature Biotechnology*, 21:435 – 439.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria (http://www.r-project.org/).

Rogers, S., and Girolami, M. (2005). A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, 21(14):3131-3137.

Roth, C.M. (2002). Quantifying Gene Expression. *Curr. Issues Mol. Biol.*, 4: 93-100.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 16 : 939–945.

Savageau, M.A. (1991). Biochemical Systems Theory: operational differences among variant representations and their significance. *J. Theor. Biol.*, 151:509-530.

Savageau, M.A. (1998). Rules for the evolution of gene circuitry. *Pacific Symposium on*

Biocomputing, 3:54-65.

Schlitt, T., and Brazma, A. (2004). Current approaches to gene regulatory network. *Science*, 303(5659)**:**799-805.

Scott, M.S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D.Y., and Hallett, M. (2005). Identifying regulatory subnetworks for a set of genes. *Mol. Cell Proteomics*, 4:683–692.

Segal, E., Yelensky, R., and Koller, D. (2003). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *ISMB* (Supplement of Bioinformatics): 273-282.

Sharan, R., Elkon, R., Shamir, R. (2000). Cluster analysis and its applications to gene expression data. *In Ernst Schering Workshop on Bioinformatics and Genome Analysis*, (38):83-108. Berlin, Springer-Verlag

Shatkay, H., Edwards, S., Wilbur, W. J., and Boguski, M. (2000). Genes, Themes, and Microarrays: Using Information Retrieval for Large-Scale Gene Analysis. *ISMB*, pp. 317-328

Sherlock, G. (2000). Analysis of large-scale gene expression data. *Curr Opin Immunol*, 12:201-205.

Shermin, A., and Orgun, A.M. (2009a). Using dynamic bayesian networks to infer gene regulatory networks from expression profiles. *Proc. of the 24th Annual ACM Symposium on Applied Computing (SAC 2009),* pp.799-803. Honolulu, USA.

Shermin, A., and Orgun, M.A. (2009b). A 2-Stage Approach for Inferring Gene Regulatory Networks Using Dynamic Bayesian Networks. *Proc of the 3rd IEEE International Conference on Bioinformatics & Biomedicine (BIBM09), pp.*166-169. Washington DC, USA.

Shermin, A., and Orgun, M.A. (2010). Analysis of microarray data to infer transcription regulation in the yeast cell cycle. *I. J. Functional Informatics and Personalised Medicine*, 3(1): 73-88.

Shermin, A., Jamil, H., and Orgun, M.A. (2011). A scalable approach for inferring transcriptional regulation in the yeast cell cycle.

Shmulevich, I., Dougherty, E.R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean

networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261-274.

Sima, C., Hua, J., and Jung, S. (2009). Inference of Gene Regulatory Networks Using Time-Series Data: A Survey. *Curr Genomics*, 10(6): 416–429.

Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2001). Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle. *Cell*, 106:697–708.

Sirbu, A., Ruskin, H. J. and Crane, M. (2010). Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC Bioinformatics*, 11:59.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Sacccharomyces Cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9:3273–3297.

Spieth, C., Hassis, N and Streichert, F. (2006). Comparing Mathematical Models on the Problem of Network Inference. *Proceedings of the 8th annual conference on Genetic and evolutionary computation (GECCO*), 12:279-286.

Spieth, C., Streichert, F., Speer, N., nad Zell, A. (2005). Inferring Regulatory Systems with Noisy Pathway Information. *German Conference on Bioinformatics*, pp.193-203.

Streib, F.E, Dehmer, M., Bakır, G. H., and M¨uhlh¨auser, M. (2005). Influence of Noise on the Inference of Dynamic Bayesian Networks from Short Time Series. *World Academy of Science, Engineering and Technology*, 10: 70-74.

Styczynski, M. P., and Stephanopoulos, G. (2005). Overview of computational methods for the inference of gene regulatory networks. *Comp. Chem. Eng*, 29:519–534.

Sui, H.S.J., Mortimer, J.R., Arenillas D.J., Brumm, J., Walsh, C.J., Kennedy, B.P., and Wasserman, W.W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*, 33:3154-64.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Estimating gene networks from gene expression data by combining BN model with promoter element detection. *Bioinformatics*, 19 (suppl 2): ii227-ii236.

Tan, M., Alhajj, R., and Polat, F. (2010). Scalable approach for effective control of gene regulatory networks. *Artificial Intelligence in Medicine*, 48(1): 51-59.

Tanay, A., and Shamir, R., (2001). Computational expansion of genetic networks. *Bioinformatics*, 17(suppl 1): 270-278.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. *Nat Genet*, 22:281-285.

Teixeira M.C., Monteiro P., Jain P., Tenreiro S., Fernandes A.R., Mira N.P., Alenquer M., Freitas A.T., Oliveira A.L. and Sá-Correia, I. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucl. Acids Res.*, 34: D446-D451. Oxford University Press.

Teng, L., and Chan, L. (2008). Discovering Distinct Patterns in Gene Expression Profiles. *Journal of Integrative Bioinformatics*, 5(2):105.

Tofigh, A., Sjlund, E. Höglund, M., and Lagergren, J. (2011). A Global Structural EM Algorithmfor a Model of Cancer Progression. *Neural Information Processing Systems (NIPS)*, pp. 163-171.

Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *PNAS*, 99 : 14031-14036.

Veerla, S., and Höglund, M. (2006). Analysis of promoter regions of co-expressed genes identified bymicroarray analysis. *BMC Bioinformatics*, 7:384.

Vohradsky, J. (2001) . *Neural network model of gene expression*. *FASEB Journal*, 15:846-54.

Wang, W., Cherry, J.M., Nochomovitz, Y., Jolly, E., Botstein, D.,and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl. Acad. Sci. USA*, 102 : 1998–2003.

Wang, Y., Miller, D.J., and Clarke, R. (2008). Approaches to working in high-dimensional data spaces: gene expression microarrays. *British Journal of Cancer*, 98:1023 – 1028.

Watson, J. D. and Crick, F. H. (1958). On protein synthesis. The Symposia of the Society for Experimental Biology, 12:138-163.

Watson, J.D., Baker, T.A, Bell, S.P, Gann, A, Levine, M. and Losick. R. (2008). Molecular Biology of the Gene. CSHL Press, Pearson Education, Sixth edition.

Waveren, C.V., and Moraes, C.T. (2008). Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. *BMC Genomics*, 9: 18.

Wentao, Z., Serpedin, E., Dougherty E. R. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, 22(17):2129–2135.

Werhli, A.V., and Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1): Article 15.

Wolfsberg, T.G., Gabrielian, A.E., Campbell, M.J., Cho, R.J., Spouge, J.L., and Landsman, D. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome Res.*, 9: 775-792.

Wyrick, J.J., and Young, R.A. (2002). Deciphering gene expression regulatory networks. *Curr Opin Genet Dev*, 12:130-136.

Xenarios, I., and Eisenberg, D. (2001). Protein interaction databases. *Curr Opin Biotechnol*., 12:334-9.

Xiao, Y. (2009). A tutorial on analysis and simulation of Boolean gene regulatory network models. *Curr Genomics*, 10(7):511-525.

Xu, R., Venayagamoorthy, G.K., Wunsch, D.C. (2007). Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization. *Neural Networks*, 20:917-27.

Yeung N., Cline, M.S., Kuchinsky, A., Smoot, M.E., and Bader, G.D. (2008). Exploring biological networks with Cytoscape software. *Curr Protoc Bioinformatics*, Chapter 8: Unit 8.13.

Yeung, K.Y., Medvedovic, M., and Bumgarner, R.E. (2004). From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biology*, 5:R48.

Yu, H., Luscombe, M. N., Qian, J. and Gerstein, M., (2003). Genome Analysis of gene expression relationships in transcriptional regulatory network. *Trends Genet*, 19(8):422-427.

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A., and Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18) : 3594-3603.

Zhang, M.Q. (1999). Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.*,9(8):681-8.

Zhang, Y., Deng, Z., Jiang, H., and Jia, P. (2007). Inferring gene regulatory networks from multiple data sources via a dynamic Bayesian network with structural EM. *Lecture Notes in Computer Science*, 4544: 204-214.

Zhang, Y., Xuan, J., de los Reyes, B.G., Clarke, R., and Ressom, H.W. (2010). Reconstruction of Gene Regulatory Modules in Cancer Cell Cycle by Multi-Source Data Integration. *PLoS ONE*, 5(4): e10268.

Zhang, Y., Zha, H., Wang, J.Z., and Chu, C.H. (2004). Gene Co-regulation vs. Co-expression. Technical Report, The Pennsylvania State University,USA.

Zheng, J., Wu, J., and Sun, Z. (2003). An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Research*, 31:1995-2005.

Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E., and Schadt, E.F. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics,* 40:854 - 861.

Zou, M. and Conzen, S. D. (2005). A New Dynamic Bayesian network (DBN) Approach forIdentifying Gene Regulatory Networks from Time Course Microarray Data. B*ioinformatics*, 21:71–79.