Imperial College London
Department of Electrical & Electronic Engineering

Stochastic Modeling and Inference of Large-scale Gene Regulatory Networks

Haseong Kim

April 2012

Supervised by Professor Erol Gelenbe

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Electrical & Electronic Engineering of
Imperial College London
and the Diploma of Imperial College London

Abstract

Gene regulatory networks (GRNs) consist of thousands of genes and proteins which are dynamically interacting with each other. Researchers have investigated how to uncover these unknown interactions by observing expressions of biological molecules with various statistical/mathematical methods. Once these regulatory structures are revealed, it is necessary to understand their dynamical behaviors since pathway activities could be changed by their given conditions. Therefore, both the regulatory structure estimation and dynamics modeling of GRNs are essential for biological research.

Generally, GRN dynamics are usually investigated via stochastic models since molecular interactions are basically discrete and stochastic processes. However, this stochastic nature requires heavy simulation time to find the steady-state solution of the GRNs where thousands of genes are involved. This large number of genes also causes difficulties such as dimensionality problem in estimating their regulatory structure.

This thesis mainly focuses on developing methodologies for large-scale GRN analyses. It includes applications of a stochastic process theory called G-networks and a reverse engineering technique for large-scale GRNs. Additionally a series of bioinformatics techniques was applied in brain tumor data to detect disease candidate genes along with their large-scale GRNs.

The proposed techniques such as stochastic modeling (bottom-up) and reverse engineering (top-down) could provide a systematic view of a complex system and an efficient guideline to identify candidate genes or pathways triggering a specific phenotype of a cell. As further work, the combinatorial use of the modeling and reverse engineering approaches would be helpful in obtaining a reliable mathematical model and even in developing a synthetic biological system.

I dedicate this thesis
to my father, **Kim, Il Ho** and my mother, **Choi, Hyun Soon**.

Acknowledgements

In the first place, I would like to give a big thanks to my supervisor, Professor Erol Gelenbe, for his constant support since when I started my PhD study. Thanks to his academic enthusiasm, I could start my degree in this precious college and he has generously provided me financial support which has been a very big help to keep up my research. Most of the applications in this thesis could be implemented as they are based on his strong theoretical understandings.

I am also thankful to my family for their encouragement and financial support, especially for my parents, Il-Ho Kim and Hyun-Soon Choi. They gave me everything they have throughout my whole life. Nothing can be compared with their love and sacrifice. I also would like to thank my brother, Shin-Chul Kim, and sister in raw, Sun-Mi Lee. They hold my back free during the whole period of my study.

Finally, it is a pleasure to acknowledge my colleagues at Intelligent Systems Networks for their helpful discussions. Also I am very lucky to be here in London with Korean society members at Imperial College London. My special thanks go to Dr. Kim, Soo-Hyun and Dr. Sohn, Keehoon. Dr. Kim constantly encouraged me to understand biological issues and helped me to live in London in many ways of supports. Dr. Sohn helped me to settle down in UK and has advised me from the very beginning of my PhD study.

Contents

Abstract **2**

Acknowledgements **4**

1 Introduction **11**
1.1 Motivation 11
1.2 Objectives 12
1.3 Thesis outline 13
1.4 Publications 13
1.5 Conference Talks 14
1.6 Statement of Originality 15

2 Background Theories **16**
2.1 Gene Regulatory Principles 16
2.2 Stochastic Gene Expression Modeling for GRNs 18
2.3 Gillespie Algorithm 25
2.4 Queuing Networks 27
2.4.1 Basic Queuing Theory 28
2.4.2 Queueing Networks 33
2.4.3 G-networks for GRNs 34

3 Stochastic GRN Modeling with Switch-like Gene Responses **38**
3.1 Introduction 38
3.2 Gene Expression Model 39
3.3 Hill Function for Protein Activation 40
3.4 Toggle Switch Gene Expression Model 41
3.5 ScbA-ScbR System 44
3.6 Conclusions 47

4 G-network based Abnormal Gene Expression Detection 50
4.1 Introduction . . . 50
4.2 Abnormal Gene Expression Detection . . . 51
4.3 Simple Gene Regulatory Networks using Stochastic Gene Expression Model . . . 53
4.4 Modeling Cell Cycle Gene Regulatory Networks in Budding Yeast 55
4.5 Conclusions . . . 59

5 G-network based Abnormal Pathway Detection 61
5.1 Introduction . . . 61
5.2 Abnormal Edge Detection . . . 62
5.3 Evaluation with Simulated Networks . . . 66
5.4 Application with p53 Networks . . . 68
5.5 Conclusions . . . 72

6 G-network based Two Layer Stochastic GRN Modeling 75
6.1 Introduction . . . 75
6.2 Two Layer Gene Regulatory Network Model . . . 76
6.2.1 Biological assumptions for gene regulation . . . 76
6.2.2 DNA layer modeling . . . 77
6.2.3 Protein layer modeling . . . 78
6.2.4 Bridging mRNA and protein layers . . . 82
6.3 Toy Example of Single Gene Expression . . . 83
6.4 Conclusions . . . 86

7 Reconstruction of Large-scale GRNs using Bayesian Model Averaging 87
7.1 Introduction . . . 87
7.2 Bayesian Model Averaging Networks (BMAnet) . . . 88
7.2.1 BMA for linear regression models . . . 88
7.2.2 Prior Update . . . 89
7.2.3 Two stage model selection . . . 90
7.2.4 Build a final GRN structure . . . 91
7.3 Simulated Network Comparison . . . 92
7.4 Human Brain Tumor GRNs . . . 94
7.5 Conclusions . . . 96

8 Mining of Disease Candidate Genes via Large-scale GRNs 97
8.1 Introduction . . . 97
8.2 Network Evaluation . . . 98
8.3 Disease Candidate Gene Detection . . . 100
8.4 Simulation Study . . . 100
8.5 Human Brain Tumor GRNs . . . 101
8.6 Conclusions . . . 106

9 Meta Analysis for Brain Tumorigenesis 107
9.1 Introduction . . . 107
9.2 Meta-analysis of Microarray Data . . . 108
9.3 Analysis of Brain Tumor Biopsy Samples and Validation of the Microarray . . . 112
9.4 Clustering and Gene Ontology Analysis . . . 113
9.5 Expression Pattern of Extracellular Matrix Genes . . . 116
9.6 Conclusions . . . 118

10 Large-scale GRN Analysis for Brain Tumorigenesis 120
10.1 Improving Abnormal Pathway Detection Algorithm . . . 120
10.2 Evaluation of the Pathway Detection Algorithm . . . 121
10.2.1 Example: 20 node gene network . . . 122
10.2.2 Performance comparison . . . 124
10.3 Abnormal Pathway Detection of Brain Tumor Data . . . 125
10.4 Conclusions . . . 128

11 Conclusions and Perspectives 131
11.1 Thesis Contributions . . . 131
11.2 Applications for Synthetic Biology . . . 132
11.3 Discussions and Future Works . . . 133

Bibliography 135

List of Tables

2.1 Parameters of the four gene network 25

4.1 Input rate λ^+ and steady-state probability q_i in the four gene network 56
4.2 Steady state probabilities in the yeast cell cycle GRN model . . . 59
4.3 Income rates in the yeast cell cycle GRN model 59

5.1 Average expression levels of simulated data and their p-values of t-tests 67
5.2 Assumed transition probabilities in the normal condition and their estimations in the abnormal condition 68
5.3 Assumed transition probabilities in the normal condition and their estimations from the cancer samples 73

7.1 Significantly over-represented GO terms of the top 100 regulatory genes 96

8.1 Significantly over-represented GO terms (Biological Process) with p-value < 0.01. 104

9.1 Microarray samples from GEO database 109
9.2 Differentially expressed probes in the three meta-analyses 110
9.3 Statistics of KAL1 in three meta-analyses 111
9.4 Twelve expression patterns of differential probes 111
9.5 Brain tumor samples from St. George's hospital 113
9.6 Differential expression of KAL1 in different tumor types 114
9.7 ANOVA result of KAL1 followd by Tukey's tests 114

10.1 The result of abnormal activity detection in the 20 gene network . 124
10.2 Performance comparison between abnormal pathway detection algorithms 125
10.3 Top 20 KEGG pathways sorted by PW 130

List of Figures

2.1 Seven biological processes for gene regulation 17

3.1 Assumptions for the stochastic gene expressions 40
3.2 The toggle switch system 42
3.3 The protein expression of the toggle switch model 43
3.4 ScbR/ScbA System 44
3.5 The expression profile of the ScbA-ScbR system 46
3.6 Switch like behavior of ScbR and ScbA 47
3.7 ScbA expression levels 48

4.1 Repressilatory gene regulatory network structure with four genes . 53
4.2 Cell cycle regulatory network structure in budding yeast 57
4.3 Cell cycle regulatory network structure with selected 13 genes . . 57
4.4 Expression profiles of selected 13 genes 58

5.1 Correlation between the abnormal transition probabilities 65
5.2 The simulated gene regulatory networks consisting of three genes 66
5.3 The protein expression profiles of the simulated network 67
5.4 The p53 network consisting of 28 genes 69
5.5 Average mRNA expression levels of genes in two datasets 70
5.6 The p53 network with the results of GSE12941 dataset analysis . . 71
5.7 The p53 network with the result of GSE6222 dataset analysis . . . 72

6.1 Transcription inhibition and active degradation models 84
6.2 Steady-state probabilities and average number of molecules 85

7.1 Numerically computed α_{Er} 92
7.2 In/out-degree distributions 93
7.3 The performance comparison 93
7.4 Estimated brain tumor gene networks 95

8.1 Comparison of 10 measurements of the three reverse engineering methods (BMAnet, Enet, GGM) along with the true networks (True). Please see the context for detailed description. 102
8.2 (a) In-degree and (b) out-degree distributions of nontumor, grade III, and IV tumor network structures. 103
8.3 Biplots of Rwr (a) non-tumor vs. grade III, (b) non-tumor vs. grade IV, (c) grade III vs. grade IV. In each plot, we divided genes in three groups, U_{AB}: genes having high Rwr in B group but low in A group, D_{AB}: genes having low Rwr in B group but high in A group, B_{AB}: genes having high Rwr values in both A and B group. (d) co-expression networks of Non-tumor, (e) grade III, and (f) grade IV samples. Light gray dots represent the known tumor related genes and dark gray nodes indicate the identified 169 genes 105

9.1 The number of differentially expressed probes 110
9.2 Expression profile of KAL1 in brain tumor tissues 114
9.3 GOC values in each cluster analysis 116
9.4 The GO tree for the KAL1 cluster with optimal cluster size 17 . . 117
9.5 Significantly expressed ECM genes identified in the three meta-analyses 118

10.1 Randomly generated network structure with 20 nodes 122
10.2 Expression profiles of the 20 genes for 60 second 123
10.3 Human non-metabolic gene networks containing 1136 genes from KEGG database 126
10.4 The human non-metabolic gene networks represented by two gene layers 127
10.5 Identified abnormal interactions using normal and high grade tumor data 129

1 Introduction

1.1 Motivation

Gene regulatory networks (GRNs) can play a key role to uncover the functions of genes and their genetic effect on a specific phenotype. Thanks to the advance of measuring technology such as microarrays, biologists easily capture the expression patterns of mRNA/protein under different conditions. Along with these bursts of biological data, various statistical/mathematical models have been introduced for the GRN analysis. These models can be used not only to infer with the unknown regulatory structures among genes but also to predict the dynamic behaviors of the system.

A cell where tens of thousands of genes simultaneously interact each other shows dynamical behaviors responding to its environmental conditions. Sometimes it is more important to know the quantitative changes of expression level among genes rather than their regulatory structure. For example, in an apoptosis modeling study of Neumann *et al.*, the cell death pathway that depends on caspase-8 and the survival pathway triggered by NF-κB are shown to be activated in parallel. Also they proposed the fate of the cell is determined by the delicate signal balance between c-FLIP$_{\mathrm{L}}$ and the initiator caspase [85]. So it is essential to monitor the dynamics of the whole system to understand the environmental impact on the phenotype of the system. The dynamics of a system can be investigated by modeling the system mathematically so that one can monitor how the system is affected by the model parameters (bottom-up approach).

Generally, stochastic modeling is applied to describe GRNs because of the discrete and stochastic nature of gene expression. The stochastic effects are more significant at a single molecule level rather than in general modeling approaches [94]. The basic assumption of the stochastic models is that the system is 'well stirred', that is each molecule have equal chance to be reacted in the system and every reactions are simulated by using the Gillespie algorithm to find the steady-state of the system [55]. However, simulating every single reaction in a large-scale

GRN gives rise to heavy computation times. So more efficient stochastic modeling techniques need to be applied to the complex biological system.

On the other hand, the stochastic modeling is only available when its regulatory structure is known. So inferring the regulatory structure should be carried out in advance of the dynamics modeling especially when the system includes unknown biological pathways. There are various mathematical/statistical studies known as "reverse engineering", which reveals the unknown interactions among biological molecules from mRNA expression microarray data (top-down approach) [42, 80, 90]. However, there are two challenges for the large-scale GRN inference. Firstly, the microarray data consist of at most hundreds of samples while the number of genes could be up to tens of thousands, which causes the dimensionality problem such as high variance and over-fitting. So most of linear modeling approaches such as ODE and regression employ regularization techniques [116, 60]. Secondly, the possibility of false positive interactions is getting higher as the number of genes increases. Therefore, reverse engineering methods should be developed to manage various sources of biological data such as DNA-protein, protein-protein binding information and known pathways in published papers to reduce the false positives in estimating the GRN structure.

1.2 Objectives

This thesis mainly discusses the techniques of large-scale GRN analysis especially for stochastic modeling and reverse engineering.

- In the stochastic modeling, to tackle the heavy computation time in finding a steady-state of stochastic models, we employed a G-network which is a special class of queuing networks and provides a closed form solution of the system in a steady-state. This analytical solution of G-networks enables us to extend the dynamics analysis to the large-scale GRNs. The proposed stochastic modeling study focuses on finding the abnormally behaving genes and pathways causing disease development.

- This thesis also aims to develop a new reverse engineering method which efficiently integrates various sources of biological data and can be available to the large number of genes. In this approach, Bayesian model average (BMA) technique was applied since its ensemble process and Bayesian ap-

proach can resolve the dimensionality problem and the integration issue, respectively.

- Beside these two main objectives, in order to analyze brain tumor gene expression data, a series of bioinformatics methods (Meta-analysis, clustering, Gene Ontology based evaluation) were performed and found candidate genes/pathways using the proposed large-scale GRN techniques.

1.3 Thesis outline

This thesis consists of mainly three subjects, stochastic modeling (Chapter 3, 4, 5, and 6), reverse engineering (Chapter 7 and 8), and their applications (Chapter 9 and 10). Specifically, in Chapter 2, biological and mathematical assumptions are reviewed and followed by stochastic process theory including the G-network theory. In Chapter 3, the conventional stochastic modeling technique, Gillespie algorithm, is used to describe the switch-like behavior of a gene. Then Chapter 4 and 5 represent the GRN modeling applications for detecting the abnormal genes/pathways using G-networks which tackle the limitation of the conventional stochastic modeling approach. In Chapter 6, two-layer model consisting of mRNA and protein layers is introduced as a more delicate model than the G-network model. Chapter 7 represents a new reverse engineering method for large-scale GRN inference and Chapter 8 shows how to identify disease candidate genes based on the estimated GRNs. In order to apply these approaches to real data, brain tumor mRNA expression data were collected from a public database and a series of bioinformatics techniques including Meta-analysis, clustering and Gene Ontology based evaluation were applied to the brain tumor data in Chapter 9. Then Chapter 10 shows the abnormal pathway detection technique is applicable to the large-scale brain tumor GRN with corresponding mRNA expression data. Chapter 11 will draw conclusions and discuss the perspectives of this thesis.

1.4 Publications

1. **H. Kim** and E. Gelenbe, Stochastic Gene Expression Modeling with Hill Function for Switch-like Gene Responses, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2012, 9(4): 973-979 (Chapter 3)

2. **H. Kim** and E. Gelenbe, Anomaly detection in gene expression via stochastic models of gene regulatory networks. *BMC Genomics* 2009, 10(Suppl 3): S26 (Chapter 4)

3. **H. Kim** and E. Gelenbe, Stochastic Gene Expression Model Base Gene Regulatory Networks. *Springer Proceedings in Physics*, 2010, 135:235-244 (Chapter 4)

4. **H. Kim**, R. Ataly and E.Gelenbe, G-Network Modelling Based Abnormal Pathway Detection in Gene Regulatory Networks. *Computer and Information Sciences II*, 2011, 257-263 (Chapter 5)

5. **H. Kim** and E. Gelenbe, G-networks based Two Layer Stochastic Modeling of Gene Regulatory Networks with Post-translational Processes, *Interdisciplinary Bio Central*, 2011, 3:8 (Chapter 6)

6. **H. Kim** and E.Gelenbe, Inference of large-scale gene regulatory networks via Bayesian model averaging, *IEEE Transactions on NanoBioscience* 2012 (Accepted) (Chapter 7)

7. **H. Kim**, et. al., Identifying Disease Candidate Genes via Large-scale Gene Network Analysis, *International Journal of Data Mining and Bioinformatics*, 2012 (Accepted) (Chapter 8)

8. **H. Kim**, et. al., Anosmin-1 promotes brain tumorigenesis by integrin-mediated signal pathways, (Submitted) (Chapter 9)

9. **H. Kim** and E.Gelenbe, Differentially Activated Pathways in Large-scale Gene Regulatory Networks (In preparation) (Chapter 10)

10. **H. Kim**, Modeling and analysis of gene regulatory networks based on the G-network, *Computer Journal* (In preparation)

1.5 Conference Talks

1. **H. Kim** and E. Gelenbe, Stochastic Gene Expression Modeling with Hill Function for Switch-like Gene Responses, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010, Hong Kong (Chapter 3)

2. **H. Kim** and E. Gelenbe, Steady-State Analysis of Gene Regulatory Networks via G-Networks, *The International Conference on Bioinformatics (InCoB)*, 2009, Singapore (Chapter 4)

3. **H. Kim**, R. Ataly and E.Gelenbe, G-network Modeling based Abnormal Pathway Detection in Gene Regulatory Networks, *International Symposium on Computer and Information Sciences (ISCIS)*, 2011, UK (Chapter 5)

4. **H. Kim** and E.Gelenbe, Inference of large-scale gene regulatory networks via Bayesian model averaging, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2011, USA (Chapter 7)

5. **H. Kim** and E.Gelenbe, Stochastic model based abnormality detection in a large-scale gene regulatory network, *European Research Consortium for Informatics and Mathematics (ERCIM)*, 2011, UK (Chapter 10)

1.6 Statement of Originality

This thesis is submitted for the degree of Doctor in Philosophy in the Department of Electrical and Electronic Engineering at Imperial College London. The research work reported herein was carried out, unless otherwise stated, by the author in the Department of Electrical and Electronic Engineering at Imperial College London between October 2008 and October 2011. No part of this thesis has been submitted in support of an application for a degree or qualification of this or any other university or educational establishment.

2 Background Theories

A model in this thesis represents a set of mathematical equations to describe various biochemical reactions for gene regulation principles. So in Section 2.1, we will review the background knowledge of the biological principles for the GRN modeling. Then the stochastic modeling (Section 2.2) will be presented along with a tool for a numerical solution, the Gillespie algorithm, followed by the G-network theory (Section 2.4.3) which is the basis of our GRN modeling.

2.1 Gene Regulatory Principles

The gene regulation is largely involved with the activities of various protein molecules such as transcription factors, repressors, and activators. Proteins act alone or together with other proteins to promote or block the expression of genes. Based on these protein activities, our study focuses on following seven gene regulation related mechanisms (Figure 2.1).

1. Transcription (Figure 2.1 (a))
   Activator proteins such as transcription factors enhance the interaction between an RNA polymerase and the promoter site of a gene by increasing their binding affinity. Some proteins bind to an enhancer which is a short region in a DNA strand to increase transcription levels. Apart from this transcription process, there is another DNA-protein interaction called repression that DNA-binding proteins (repressors) are attached to the operator of a gene in order to prevent an RNA polymerase binds to its target promoter and decrease the transcription rate of the gene.

2. Translation (Figure 2.1 (b))
   In prokaryotic cells, ribosomes bind successively to an mRNA as soon as it is accessible to the mRNA and produce corresponding proteins successively. These translation processes are continued until the mRNA is degraded by an RNAse-E. It is known that the expression pattern of proteins is different with

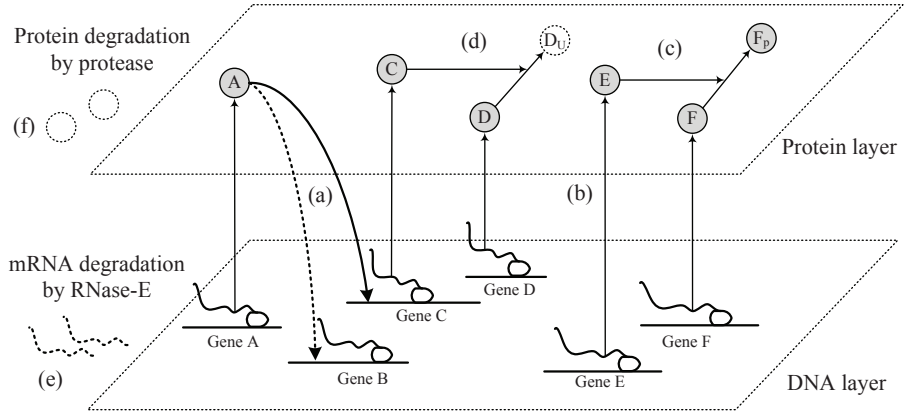


Figure 2.1: Seven biological processes for gene regulation. Gene A activates the transcription of gene C and represses gene B. Protein C add ubiquitin to protein D which will be degraded after the ubiquitination. Gene E produces kinase E which phosphorylates protein F and F is activated after this phosphorylation. [69]

that of mRNAs. Many studies have interpreted the variation in terms of a burst which takes place in brief periods of high expression intensity followed by long periods of low intensity [115, 25].

3. Phosphorylation (Figure 2.1 (c))
   Protein phosphorylation is usually involved in the positive regulation of gene expressions. It activates or inactivates many proteins by adding a phosphate (PO_4) group to the substrate proteins or other organic molecules. The phosphorylation is accomplished by a kinase which transfers phosphate groups from high-energy donor molecules. In our study, phosphorylation processes are assumed to be involved with only positive gene regulations.

4. Ubiquitin dependent protein degradation (Figure 2.1 (d))
   Ubiquitination is a post-translational modification process that labels proteins for degradation by attaching ubiquitin monomers. This process is performed through a cascade of three enzymes, E1, E2, and E3. E1 activates an ubiquitin which is then transferred to ubiquitin-conjugating enzyme E2. A substrate for the ubiquitination is recognized by ubiquitin ligase E3. The ubiquination is required the hundreds of ubiquitin ligases which is the complex protein of E2-E3 where E2 attaches the ubiquitin to the target protein. Then this tagged substrate is degraded into peptides by 26S proteasome.

5. mRNA degradation (Figure 2.1 (e))
   The lifetimes of prokaryotic mRNAs range from seconds to more than an hour. An mRNA is cleaved by RNase-E and then is degraded by a combination of ribonucleases including $3' - 5'$ exonucleases. RNase-E shares its binding site with ribosomes which are responsable for the protein production. In eukaryotic cells, mRNAs can survive from several minutes to days. The degradation of actively translated mRNAs is blocked by ribosomes and poly-A binding protein.

6. Protease dependent protein degradation (Figure 2.1 (f))
   Proteases can be any proteins that are responsible for protein breakdown. There are six groups of proteases which are found in lysosomes of animal cells and in vacuoles of yeast and plant cells. These proteases are involved in a multitude of physiological reactions from simple digestion of food proteins to highly-regulated systems such as apoptosis. In our model, the degradation effect of proteases is fixed as a constant while that of ubiquitination depends on the level of E3 enzyme.

7. Protein-protein interaction
   Many functional molecular processes including the phosphorylation and ubiquitination are based on a large number of protein complexes which are organized by their protein-protein interactions. For instance, inducers which bind to repressors and prevent the repressors' activity can be also modeled with this protein-protein interaction. These interactions could be modeled by a binary reaction channel with equation $A + B \rightarrow C$ where A, B, and C are all protein molecules. However, some cases such as $A + B \rightarrow C + D$ where more than two protein products are involved cannot be explained by the G-network approach. So, in our G-network approach, a customer in the G-networks represents a latent object containing gene expression information rather than the actual protein molecules so that the reaction can be simplified into a bi-fan structure, $A + B \rightarrow C$ and $A + B \rightarrow D$.

2.2 Stochastic Gene Expression Modeling for GRNs

As shown in the previous section of the biological principles, gene expression is considered as a set of interactions of discrete molecules so stochastic modeling is a natural way to describe the gene regulation processes.

1. Gene Activation

   Depending on cell growth conditions, there are several copies of partially replicated chromosomes [94]. Let n_1 and n_1^{max} be the number of active genes and a constant that denotes the maximum copies of genes (chromosomes), respectively. Each copy of the genes spontaneously switches ON and OFF with rate λ_1 and γ_1, respectively. The stationary distribution for the number of active genes is then binomial where the probability of being ON is $Pr_{ON} = \lambda_1/(\lambda_1 + \gamma_1)$. However, in this study, it is assumed that the changing rate of the number of active genes is zero. That is, $dn_1/dt = 0$.

2. Transcription

   The transcription is typically assumed to follow an exponential distribution [81]. Let RP be RNA polymerase and Pro be the corresponding promoter in a DNA strand. After the RNA polymerase binds to the promoter, it switches from a closed complex (RPc) to an open complex (RPo), $RP + Pro \leftrightarrow RPc \leftrightarrow RPo$. The initiation of transcription is only activated by the RNA polymerase-promoter open complex (RPo). The time interval between successive transcripts is considered to have an exponential distribution with $1/T_{avg}$ where T_{avg} is the average transcript initiation interval. Then, the probability of the transcription initiation reaction in a small time interval Δt is $1/T_{avg} \exp(-t \cdot 1/T_{avg})\Delta t$ where t is the time. Let λ_2 be the activation rate of RPo then $\lambda_2 = 1/T_{avg}$ and the production rate per unit time is $\lambda_2 \cdot n_1$. The transcription initiation rate, λ_2, of mRNA is assumed $0.0067 sec^{-1}$ (1 transcript per 2.5min) [115]. The reaction equation is

   $$RPo \xrightarrow{\lambda_2} RPo + R \tag{2.1}$$

   where R is mRNA. Note that we are only concerned with RNA polymerase-promoter open complex (RPo) in this transcription modeling. However, there will be more discussions about activator proteins such as transcription factors.

3. Translation

   In prokaryotic cells, ribosomes bind successively to an mRNA as soon as it is accessible behind the transcribing RNA polymerase. Multiple ribosomes space about 80 nucleotides [81]. Corresponding proteins are produced successively from the attached ribosomes. These translation processes are con-

tinued until the mRNA is degraded by an RNAse-E. An RNAse-E directly competes with ribosomes to bind corresponding promoters because these two types of molecules share their binding site of a promoter. So the competition leads to successful translation or degradation of mRNAs. Then the probability for the number of proteins, n_3, is $Pr(n_3 = n) = p^n(1-p)$ where p and $(1-p)$ are the probability that a ribosome and an RNase-E bind to the mRNA, respectively. Let T_D be an average time interval between successive competitions. Then translation initiation rate λ_3 is $\lambda_3 = 1/T_D$. Now, p can be obtained using protein burst size, b. Many studies have interpreted variations of proteins in terms of burst which takes place in brief periods of high expression intensity followed by long periods of low intensity [115, 25]. Let's assume the burst size is equal to an average number of proteins per one mRNA. Then the probability that an ribosome binds to an mRNA is computed by solving $b = p/(1-p)$ [81]. All the appropriate parameter values of the λ_3, γ_3, b, and p will be explained along with degradations of mRNA and proteins. Equation (2.2) depicts the chemical equation of the translation process.

$$R(t) \xrightarrow{\lambda_3} R(t + \Delta t) + P(t + \Delta t) \tag{2.2}$$

The time delay, Δt, is employed to describe post-transcriptional processes such as RNA splicing [16]. It is also known that the time delays induce stochastic oscillations. Bratsun *et al.* simulated stochastic models with delayed time of degradation and negative interactions processes [16] while McAdams and Arkin studied a delayed time for protein activations after their promoters are activated [81]. Δt is assumed 0.5 in this study.

4. Protein dimerization

   Proteins often act in a form of a dimer or an oligomer rather in isolation. It is known that the protein multimerization is an important feature in regulations of bio-molecules such as enzyme and transcription factors [77]. The dimer dissociation constant K_{D1} is varied from 0.1 to 10 where $K_{D1} = [P]^2/[PP] \approx k_{d1}/k_{a1}$; k_{a1} and k_{d1} are the dimer association and dissociation rate, respectively [20]. In this study, $K_{D1} = 0.1, k_{a1} = 1.0 \times 10^6 M^{-1} sec^{-1}$.

$$P + P \underset{k_{d1}}{\overset{k_{a1}}{\rightleftharpoons}} PP \tag{2.3}$$

   The protein concentration growth is affected by additional parameters: ini-

tial cell volume, and cell growth rate. The cell growth is fixed rate and the cell volume is considered $1.6 \times 10^{-15}L$. Let c_{a1} be the stochastic rate of the dimerization then c_{a1} can be converted from k_{a1}, $c_{a1} = (2 \times 10^6 M^{-1}sec^{-1})/(6.023 \times 10^{23}mol^{-1} \times 1.6 \times 10^{-15}L) \approx 0.002sec^{-1}$ (Note that if the reaction is second order, $c_{a1} = (2 \times k_{a1})/(n_A \times V)$ where n_A is Avogadro constant and V is a volume of a cell) [126]. With this dimer association rate and the dimer dissociation constant, we can obtain that the stochastic rate of the monomerization, c_{d1} is $0.0002sec^{-1}$.

5. Degradation of mRNA and protein
   Generally, the nth order reaction equation is defined by $(d[A])/dt = k[A]^n$ where k is reaction constant and $[A]$ is the concentration of substance A (mRNA or protein). When the order is one ($n = 1$), the decrease in the concentration of A over time is $\log([A]/[A]_0) = -kt$ where $[A]_0$ is the initial concentration of A at time $t = 0$. So the half-life of A is $T_{half} = \log(2)/k$. Then, the surviving mRNAs, n_2, in the population after transcription is blocked would be $n_2 = n_{2,0}p^{T_{half}/T_D}$. This is equal to $T_{half} = -(\log(2)/\log(p))T_D$ [81]. The average half-life of mRNAs is assumed by $300sec$ [20]. Therefore, the translation initiation rate $\lambda_3 = 1/T_D$ can be computed with appropriate p which is obtained by using the protein burst size b. Note that b is varied; 100 for an average E. coli gene [94], 40 for lacZ, and 5 for lacA [115]. In this study, b is assumed 20 so the probability that a ribosome binds to the mRNA, p, is 0.952. Then, the average time interval between successive competitions of ribosome and RNase-E, T_D, is $21.1168sec$. Therefore, the translation initiation rate λ_3 is $0.04736sec^{-1}$. The degradation rate of mRNA, γ_2, is $0.00231sec^{-1}$ (=$log(2)/300sec$). The average of half-life of proteins is known as from 15min to 120min [115]. It is known that dimers or oligomers are more stable than their monomeric components [20]. In our study, the half-life of monomer proteins and dimer proteins are assumed as 15min and 20min, respectively. So the degradation rates of monomers (γ_3) and dimers (γ_4) are $0.00077sec^{-1}$ and $0.000578sec^{-1}$.

6. Repressor-Operator Association
   A repressor is a DNA binding protein that inhibits the transcription of genes. The activation of an RNA polymerase is blocked when a repressor binds to

the corresponding gene's operator.

$$P + O \underset{k_{d2}}{\overset{k_{a2}}{\rightleftharpoons}} PO \tag{2.4}$$

where P, O, and PO represent a repressor protein, operator, and repressor-operator complex, respectively. k_{a2} and k_{d2} are the repressor-operator (RO) complex association and dissociation rates. $K_{D2}(= k_{d2}/k_{a2} = [P][O]/[PO])$ is the PO complex equilibrium constants where the rate of dissociation of the PO complex is significantly affected by changes in association rate constants. K_{D2} is varied from 1 to 1000 nM and $K_{D2} = 10$ [20]. The kinetics of association (k_{a2}) is known to be from $1 \times 10^8 M^{-1}sec^{-1}$ to $2 \times 10^9 M^{-1}sec^{-1}$ [56]. We use $1 \times 10^8 M^{-1}sec^{-1}$ so the converted association stochastic rate, c_{a2}, is $1 \times 10^8 M^{-1}sec^{-1}/(6.023 \times 10^{23} mol^{-1} \times 1.6 \times 10^{-15} L) = 0.104 sec^{-1}$ and the dissociation stochastic rate, c_{d2}, is 1.04 nM/sec.

7. Activator-Promoter/Enhancer Association
   An activator is a DNA binding protein that increases the RNA polymerase activity. This type of transcription factor binds to either enhancer or promoter regions of DNA strand. After the activators bind to the enhancer, the RNA polymerase is stabilized and activates the transcription. So, not only the transcription initiation rate λ_2 in Equation (2.1), the concentration of activators also affects to the mRNA production. So we have the following reaction equation that produces RNA polymerase-Promoter complex (RPo),

$$PP + Pro \underset{k_{d2}}{\overset{k_{a2}}{\rightleftharpoons}} RPo \tag{2.5}$$

   where PP can be either activators and RNA polymerases. We use the same kinetic constants with the repressor-operator association because this reaction is also a kind of DNA-protein associations.

8. Phosphorylation and Ubiquitination
   Phosphorylation is a process that adds a phosphate (PO_4) group to a protein. This process turns many enzymes' activity on and off, which may cause or prevent the mechanisms of diseases. Enzymes involved in this phosphorylation process are called kinases. The reversible process of phosphorylation

is dephosphorylation which removes the phosphate group from a phosphorylated protein. An enzyme called phosphatases performs this dephosphorylation. A protein activated by the phosphorylation has several possible phosphorylation sites on its amino acid residues such as serine, threonine, and tyrosine. Many modeling approaches simplify the expression processes including the multiple phosphorylation by introducing time delays to the model [81, 107]. However, we introduced a reaction equation to represent a phosphorylation process that a kinase phosphorylates an enzyme.

$$PP_{kinase} + PP \underset{k_{d3}}{\overset{k_{a3}}{\rightleftharpoons}} PP_{kinase} + PPp \tag{2.6}$$

where PPp represents the phosphorylated protein.

Ubiquitination is also a post-translational protein modification process that labels proteins to promote their degradation. Though the ubiquitination process requires E1, E2, and E3 enzymes, we simply present the process with an E2-E3 ligase and a corresponding protein ubiquitinated by the ligase.

$$PP_{ligase} + PP \underset{k_{d3}}{\overset{k_{a3}}{\rightleftharpoons}} PP_{kinase} + PPu \tag{2.7}$$

$$PPu \xrightarrow{\gamma_5} \phi \tag{2.8}$$

where PPu denotes the ubiquitinazed protein. (2.8) represents the degradation process of the ubiquitinated protein. In this study, all protein-protein association/disassociation rate are assumed to be the same $k_{a3} = k_{a1}$ and $k_{d3} = k_{d1}$. Also the degradation rate γ_5 is set to be the same as the dimer protein degradation rate, $\gamma_5 = \gamma_4$. In these post-translational processes, time delays can be also introduced to describe multiple phosphorylation/ubiquitinations.

9. Rate Constant Conversion

   Most of continuous kinetic models use deterministic rate constants, k, for the system reactions. In order to carry out a stochastic simulation, the rate constants, k, should be converted to stochastic rate constants, c, representing molecular reaction hazards [126]. The stochastic model represents their amount of species as integer value indicating the number of molecules of the species while the deterministic model use M (moles per liter, mole/L)

for the concentration of the species. In order to convert the concentration to numbers of molecules, the volume of the container, V (liters, L), is required. Then for a concentration of X of $[X]$ in a volume of V liters, there are $[X]V$ mole of X. Therefore, there are $n_A[X]V$ mole, where $n_A \simeq 6.023 \times 10^{23}$ is Avogadro's constant (The number of molecules in a mole). In this study, the volume of an E. coli cell is assumed $\pi/2 \times 10^{-15} \approx 1.6 \times 10^{-15} L$ [15]. For example, if the concentration of X is $10^{-5} M$ within an E. coli cell, the number of molecules is

$$n_A[X]V = 6.023 \times 10^{23} \times 10^{-5} \text{ mole/L} \times \frac{\pi}{2} \times 10^{-15} L = 9,461 \text{ mole}$$

For a zeroth order reaction

$$\emptyset \xrightarrow{k} X,$$

the deterministic rate constant is kM/s. So X is produced at a rate of $n_A kVMs^{-1}$ for a volume V. Then the stochastic rate law, c is $n_A kV$ mole/s.

Let's consider a first order reaction

$$X \xrightarrow{k} ?,$$

This reaction involves $[X]$ so the number of molecule X is required given a volume V. The number, $x = n_A[X]V$ mole. Then X decreases at rate kx molecules per second so the stochastic rate $c = k$ mole/s.

For a second order reaction

$$X + Y \xrightarrow{k} ?,$$

the deterministic rate law is $k[X][Y]Ms^{-1}$. The reaction proceeds at a rate of $n_A k[X][Y]V = kxy/(n_A V)$ mole/s where $x = n_A[X]V$, $y = n_A[Y]V$. Since the stochastic rate law is cxy mole/s, the stochastic rate, $c = k/n_A V$. On the other hand, in the case of dimerization reactions,

$$X + X \xrightarrow{k} ?,$$

the deterministic rate law is $k[X]^2$, so the concentration of X decreases at

Table 2.1: Parameters of the four gene network

Parameters		Values	References
Transcription initiation	λ_2	0.0067	[115]
Translation initiation	λ_3	0.0474	[94, 115]
mRNA degradation	γ_2	0.00023	[115]
Monomer degradation	γ_3	0.00077	[21, 115]
Dimer degradation	γ_4, γ_5	0.00058	[21, 115]
Protein-protein association	c_{a1}	0.0023	[21]
Protein-protein dissociation	c_{d1}	0.00023	[21]
DNA-protein association	c_{a2}	0.1038	[56]
DNA-protein dissociation	c_{d2}	1.04	[56]
Burst size	b	20	[94, 115]
Time delay	Δt	0.5	[17]
Cell growth rate and its cell volume are fixed.			

rate $n_A 2k[X]^2 V = 2kx^2/(n_A V)$ mole/s. The stochastic rate law is $cx(x-1)/2$ so that molecules of X are consumed at a rate of $cx(x-1)$ mole/s. For large x, $x(x-1)$ can be approximated by x^2 so $c = 2k/n_A V$. Table 2.1 shows the reaction rates used in our study along with their published papers.

2.3 Gillespie Algorithm

The Gillespie algorithm [54] is one of the commonly used numerical methods to find the possible solution of the stochastic equations. Let's consider a well-stirred system of molecules of N chemical species $\{S_1, ..., S_N\}$, which interact through M chemical reactions $\{R_1, ..., R_M\}$. Let $X_i(t)$ denote the number of molecules of species S_i in the system at time t. The goal of the Gillespie algorithm is to estimate the state vector $\mathbf{X} \equiv (X_1(t), ..., X_N(t))$, given the system initial state $\mathbf{X}(t_0) = \mathbf{x}_0$. Each reaction R_j is characterized mathematically by two quantities. The first is state-change vector $\mathbf{v}_j \equiv (v_{1j}, ..., v_{Nj})$ where v_{ij} is the change in the S_i molecular population caused by one R_j reaction. If the system is in state $\mathbf{x}$ and R_j occurs than the next state of the system is $\mathbf{x} + \mathbf{v}_j$. The other characterizing quantity for R_j is its propensity function a_j, which is defined so that

$$a_j(x)dt = Pr\{\text{a reaction } R_j \text{will occur in } [t, t+dt) | \mathbf{X}(t) = \mathbf{x}\} \tag{2.9}$$

If R_j is the unimolecular reaction $X_1 \xrightarrow{c_j} P$ where P is product(s), the probability that one of x_1 will undergo the R_j reaction in the next dt is $x_1 c_j dt$. So the propensity function is $a_j(\mathbf{x}) = c_j x_1$. If R_j is a bimolecular reaction, $X_1 + X_2 \xrightarrow{c_j} P$, the probability that one of the $x_1 x_2$ pairs reacts R_j in the next dt is $x_1 x_2 c_j dt$. So the propensity function is $a_j(\mathbf{x}) = c_j x_1 x_2$. Our interest is to infer the probability

$$P(\mathbf{x}, t|\mathbf{x}_0, t_0) = Pr\{\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(t_0) = \mathbf{x}_0\} \tag{2.10}$$

The time-evolution equation for the $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$ can be derived and we obtain following the chemical master equation (CME):

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(\mathbf{x} - \mathbf{v}_j)P(\mathbf{x} - \mathbf{v}_j, t|\mathbf{x}_0, t_0) - a_j(\mathbf{x}, t|\mathbf{x}_0, t_0)] \tag{2.11}$$

The CME (Equation (2.11)) can be rewritten

$$\frac{d}{dt}P(x, t|x_0, 0) = a(x + v)P(x + v, t|x_0, 0) - a(x)P(x, t|x_0, 0) \tag{2.12}$$

where state-change, $v = -1$. Because $P(x + v, t|x_0, 0) = 0$, we obtain

$$P(x, t|x_0, 0) = \frac{x_0!}{x!(x_0 - x)!}e^{-ctx}(1 - e^{ct})^{(x_0 - x)}(x = 0, ..., x_0) \tag{2.13}$$

The CME is actually a set of coupled ODEs, with one equation for every possible combination of reactant molecules. In some simple cases, CME can be solved analytically but it is difficult in other cases to obtain the solutions. The Gillespie algorithm enables us to obtain the simulated trajectories of $\mathbf{X}(t)$ rather than solving the CME equations by using analytical or numerical methods which give us the probability density function of $\mathbf{X}(t)$. Let's define a new probability function $p(\tau, j|\mathbf{x}, t)$, which is defined as follows.

$$p(\tau, j|\mathbf{x}, t)d\tau = Pr\{\text{The next reaction} R_j \text{will occur in } [t + \tau, t + \tau + d\tau)|\mathbf{X}(t) = \mathbf{x}\} \tag{2.14}$$

This function is the joint probability density function of the two random variables, a time to the next reaction (τ) and an index of the next reaction (j), given that the system is currently in state $\mathbf{x}$. The derived formula for Equation (2.14) is

as follows,

$$p(\tau, j|\mathbf{x}, t) = a_j(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau), \tag{2.15}$$

where

$$a_0(\mathbf{x}) = \sum_{i=1}^{M} a_i(\mathbf{x})$$

Equation (2.15) implies that τ is an exponential random variable with mean $1/a_0(\mathbf{x})$, while j is a statistically independent integer random variable with point probabilities $a_j(\mathbf{x})/a_0(\mathbf{x})$. In order to generate samples of τ and j two random numbers r_1 and r_2 are drawn from the uniform distribution in the unit interval, and take

$$\tau = \frac{1}{a_0(\mathbf{x})} ln\left(\frac{1}{r_1}\right) \tag{2.16}$$

$$j = \text{the smallest integer satisfying } \sum_{i=1}^{j} a_{i(\mathbf{x})} > r_2 a_0(\mathbf{x}) \tag{2.17}$$

Then we can use stochastic simulation algorithm (SSA) to obtain the numerical realization of the process $\mathbf{X}(t)$:

1. Initialize the time $t = t_0$ and the system's state $\mathbf{x} = \mathbf{x}_0$
2. With the system in state $\mathbf{x}$ at time t, evaluate all the $a_j(\mathbf{x})$ and their sum $a_0(\mathbf{x})$
3. Generate values for τ and j using Equations (2.16) and (2.17)
4. Effect the next reaction by replacing $t \longleftarrow t + \tau$ and $\mathbf{x} \longleftarrow \mathbf{x} + \mathbf{v}_j$
5. Record ($\mathbf{x}$,t) as desired.
6. Return to Step 1, or else end the simulation

2.4 Queuing Networks

This section presents the G-network theory which is used in the GRN modeling of this thesis. The G-Network is a probabilistic queuing network having special "customers" which include positive and negative customers, signals and triggers [47, 48]. It was originally developed as a model of stochastic neuronal networks

[53] with "negative and positive signals or spikes" which represent inhibition and excitation. This theory was firstly applied to GRNs with simplifying assumptions concerning gene expression in [2]. In terms of GRNs, a queue is a "place" in which mRNAs/proteins are stored, and the mRNAs/proteins can be considered to be "customers" of the G-Networks. The positive and negative customers are interpreted as the protein activities such as transcription factors, inducers and repressors. This section starts with the basic theory of queuing system including queuing network and followed by the G-network theory.

2.4.1 Basic Queuing Theory

A queueing system is characterized by an input process, a service and queue disciplines. A commonly used notation for a queueing system is as follows:

Arrival process / Service distribution / Number of servers / Waiting room

There are several possible characters used in the first two positions. D is deterministic, M is Markovian-Poisson for the arrival process or Exponential for the service time. G represents General and GI represents General and Independent for the arrival process. $Geom$ is Geometric. The fourth position is used for the number of buffer places in addition to the number of servers which is in the third position. If the waiting room is unlimited, the fourth position is not used.

Steady-state probability

In a single server queue ($G/G/1$), λ is the mean arrival rate of the arbitrary arrivaling customers and μ is the mean service rate of the server. If $\mu > \lambda$ then the queue is stable, which means it will eventually reach the empty state. Let U be defined as the proportion of time the queue is busy then $U = \lambda/\mu$. It can be proved as follows: The average number of customers within time period L is λL. The average number of customers that has been served during L is $\mu U L$. Since L is large and the queue is stable so $\lambda L = \mu U L$. Hence $U = \lambda/\mu$

Let $X(t)$ denote the number of customers in a system at time t and define $P_n, n \geq 0$ by

$$P_n = \lim_{t \to \infty} Pr\{X(t) = n\} \tag{2.18}$$

That is, p_n is the long-run probability (steady-state probability) that there will be n customers in the system. So P_0 is the probability that the system is empty. For

$G/G/1$ queue we have,

$$P_0 = 1 - U = 1 - \lambda/\mu \tag{2.19}$$

Lindley's equation & Little's theorem

Let's suppose a queueing system where customers enter the system at a random time and wait to get service, after being served they leave the system. We will represent S_n and W_n are the service and waiting time of the nth customer, respectively, and the instance of arrival and departure of the customer will be denoted as a_n and d_n. Than we have following equations.

$$\begin{aligned} d_n &= a_n + W_n + S_n \\ W_{n+1} &= \begin{cases} 0 & \text{if } a_{n+1} \geq d_n \\ d_n - a_{n+1} & \text{if } a_{n+1} < d_n \end{cases} \\ W_{n+1} &= \begin{cases} 0 & \text{if } W_n + S_n - (a_{n+1} - a_n) \leq 0 \\ W_n + S_n - (a_{n+1} - a_n) & \text{if } W_n + S_n - (a_{n+1} - a_n) > 0 \end{cases} \end{aligned}$$

Let $A_{n+1} = a_{n+1} - a_n$ be the inter arrival time (Exponential). Then the Lindley's Equation is

$$W_{n+1} = \begin{cases} 0 & \text{if } W_n + S_n - A_{n+1} \leq 0 \\ W_n + S_n - A_{n+1} & \text{if } W_n + S_n - A_{n+1} > 0 \end{cases}$$

If $S_n > A_{n+1}$ for all n then $W_n \to +\infty$

Now, we consider a figure depicting arrival and departure processes in time space T. If a customer arrives, the curve increases one vertically and if a customer departs, the curve decreases one. Each rectangle enclosed by the curve begins at some point a_i and ends at some point $d_{D(i)}$ where $d_{D(i)} = \text{inf}\{d_l : d_l > a_i\}$ and $K(d_l) = K(a_i)$ where $K(t)$ is the number of customers at time t and 'inf' means the smallest one (time). Let $E[K]$ is the average number of customers in

the system and $E[\Theta]$ is the average spending time in the system in $[0, T]$,

$$\begin{aligned} E[K] &= \frac{1}{T}\int_0^T K(t)dt \\ &= \frac{1}{T}\sum_{i=1}^{r}(d_{D(i)} - a_i) \\ &= \frac{1}{T}\sum_{i=1}^{r}(d_i - a_i) \\ E[\Theta] &= \frac{1}{N}\sum_{i=1}^{r}(d_i - a_i) \end{aligned}$$

where N is the total number of arrival(departure).
Then, the total spend time in the system is

$$\begin{aligned} TE[K] &= NE[\Theta] \\ E[K] &= \frac{N}{T}E[\Theta] \\ E[K] &= \text{(arrival rate)} \times E[\Theta] \end{aligned}$$

This equation is known as Little's theorem which tells us that the average number of customers in a queue is the arrival rate times the average time that a customer spends in the queue. This simple formula is valid for all general queueing system.

$M/M/1$ Queue

$M/M/1$ queue is the simplest Markovian queue which has only a single server and infinite buffer. The two 'M's refer to the fact that both the inter-arrival and the service distributions are exponential with parameter $1/\lambda$ and μ, respectively. Let P_n, for $n = 0, 1, ...$ be the steady-state probability that the queue length is n. Then for each $n \geq 0$, the rate at which the process enters state n equals the rate at which it leaves state n. That is the rate at which the process leaves state 0, λP_0, is equal to the rate at which the process enters state 0, μP_1. So we have the following equation.

$$\lambda P_0 = \mu P_1 \tag{2.20}$$

Let's suppose state n. The process can leave this state either by an arrival or a departure. So the process will leave this state at a rate of $\lambda + \mu$. Since the

proportion of time the process is P_n, the rate at which the process leaves state n is $(\lambda + \mu)P_n$. On the other hand, state n can be entered either from state $n-1$ via an arrival or from state $n+1$ via a departure. Hence, the rate which the process enters state n is $\lambda P_{n-1} + \mu P_{n+1}$. So the second equation can be obtained.

$$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1} \tag{2.21}$$

These two equations are known as balance equations. The solution of the balance equations are as follows,

$$P_0 = 1 - \frac{\lambda}{\mu} \tag{2.22}$$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \tag{2.23}$$

where $\rho = \frac{\lambda}{\mu} \leq 1$. Let L be a random number representing the number of customers in the system. Then the average number of customers is,

$$\begin{aligned} E[L] &= \sum_{n=0}^{\infty} nP_n \\ &= \sum_{n=0}^{\infty} n\rho^n(1-\rho) \\ &= \rho(1-\rho) + 2\rho^2(1-\rho) + ... + (n-1)\rho^{(n-1)}(1-\rho) + n\rho^n(1-\rho) \\ &= \rho + \rho^2 + \rho^3 + ... + \rho^n \\ &= \frac{\rho}{1-\rho} \end{aligned} \tag{2.24}$$

Alternatively, this average number of customers can be obtained using probability generating function.

$$\begin{aligned} \lambda P_0 &= \mu P_1 \\ \lambda P_1 &= \mu P_2 \\ &... \\ \lambda P_n &= \mu P_{n+1} \end{aligned}$$

Let us multiply above equations by z^n and summing up both sides will gives

$$\begin{aligned}
\lambda P_0 z^0 &= \mu P_1 z^1 / z \\
\lambda P_1 z^1 &= \mu P_2 z^2 / z \\
&... \\
\lambda P_n z^n &= \mu P_{n+1} z^{n+1} / z
\end{aligned} \tag{2.25}$$

$$\begin{aligned}
\Rightarrow \lambda \sum_{i=0}^{n} P_i z^i &= \mu / z \sum_{i=0}^{n} P_{i+1} z^{i+1} \\
\Rightarrow \lambda \sum_{i=0}^{\infty} P_i z^i &= \mu / z \sum_{i=0}^{\infty} (P_i z^i - P_0 z^0) \\
\lambda \Pi(z) &= \mu / z (\Pi(z) - P_0) \\
\rho \Pi(z) &= \frac{\Pi(z) - P_0}{z}
\end{aligned} \tag{2.26}$$

Let z approach 1 (from below) and simple algebraic manipulation of (2.26) gives

$$\begin{aligned}
P_0 &= 1 - \rho \\
\Rightarrow \Pi(z) &= \frac{1 - \rho}{1 - \rho z}
\end{aligned}$$

After taking derivative and substituting $z = 1$ we obtain

$$E[Q] = \Pi^{(1)}(1) = \frac{\rho}{1 - \rho}$$

Let W be the amount of time a customer spends in the system. Then

$$E[W] = \frac{E[L]}{\lambda} \tag{2.27}$$

$$= \frac{1}{\mu - \lambda} \tag{2.28}$$

Then average number of customers in a queue, $E[L_Q]$, and the average amount of

time a customer wait in a queue, $E[W_Q]$, are as follows,

$$\begin{aligned} E[W_Q] &= E[W] - E[S] \\ &= E[W] - \frac{1}{\mu} \\ &= \frac{\lambda}{\mu(\mu - \lambda)} \\ E[L_Q] &= \lambda E[W_Q] \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)} \end{aligned} \tag{2.29}$$

where $E[S]$ is the average amount of time a customer spend in service.

2.4.2 Queueing Networks

A queueing network consists of queues interconnected by edges which customers move along. There are two groups of queueing network models, open queueing networks and closed queueing networks but we will only review the open queueing networks.

Tandem system

Let's consider a tandem queuing system (feed-forward network) consisting of two stable $M/M/1$ queues in series. Customers arrive at queue 1 with a Poisson rate λ. After being served by queue 1 they then join the queue in front of queue 2. Each server serves one customer at a time with rate μ_i where $i = 1, 2$. Let (n, m) be the state of the server pair. The balance equations are

$$\begin{aligned} \lambda P_{0,0} &= \mu_2 P_{0,1} \\ (\lambda + \mu_1) P_{n,0} &= \mu_2 P_{n,1} + \lambda P_{n-1,0} \\ (\lambda + \mu_2) P_{0,m} &= \mu_2 P_{0,m+1} + \mu_1 P_{1,m-1} \\ (\lambda + \mu_1 + \mu_2) P_{n,m} &= \mu_2 P_{n,m+1} + \mu_1 P_{n+1,m-1} + \lambda P_{n-1,m} \end{aligned}$$

Rather than directly solve these balance equations, a solution is guessed and verify that it corresponds to the equations. The solution of these set of equations is

$$Pn, m = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right) \tag{2.30}$$

It can be proven to be correct by filling this solution into the following equation

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P_{i,j} = 1$$

Jackson networks

Jackson networks allow for feed-back loops in addition to the feed-forward networks. Suppose a queuing network consists of M queues where each of which arrival rate is λ_i and queue length dependent service rate is $\mu_i(n_i)$ where n_i is its number of customers. Let p_{ij} be the probability that a customer moves from ith queue to jth queue. Then total arrival rate of jth queue is

$$\lambda_i = \lambda p_{0,j} + \sum_{j=1}^{M} \lambda_j p_{i,j} \tag{2.31}$$

where λ is the total external arrival rate of the system and $p_{0,j}$ is its proportion for the jth queue. The departure probability of a customer is $p_{j,0} = \sum_{i=0}^{M} p_{j,i}$. Even thought the arrival processes of this Jackson networks are not Poisson processes, the joint probability, $\pi(\mathbf{n})$, of the system still has a product-form as follows,

$$\begin{aligned}\pi(\mathbf{n}) &= \prod_{i=1}^{M} P_{n_i} \\ &= \prod_{i=1}^{M} C_i \left(\frac{\lambda_i^{n_i}}{\prod_{k=1}^{n_i} \mu_i(k)} \right)\end{aligned} \tag{2.32}$$

where C_i is determined by the normalization condition that $\sum_{n_i=0}^{\infty} C_i \left(\frac{\lambda_i^{n_i}}{\prod_{k=1}^{n_i} \mu_i(k)} \right) = 1$.

2.4.3 G-networks for GRNs

This section describes the G-network theory which is the basis of this thesis. The G-network is a class of probabilistic queuing network models with special agents such as positive and negative customers, signals, triggers, and resets [47, 48, 52, 51]. The use of negative customers enables us to model negative gene regulatory interactions. The G-Networks were firstly applied to GRNs with simplifying assumptions concerning gene expression in [2] and more generalized model for GRNs was introduced in [50]. In terms of GRNs, a node in the G-networks is

a "place" in which mRNAs are stored, and a customer is considered an mRNA molecule. Note that the customers or signals of the G-Network can be any biological molecules but, in our study, we focus on behaviours of mRNAs because the mRNA expression is one of the commonly available biological data.

Let $X_i(t)$ be an integer-valued random variable which represents a quantity (mRNA) of the gene i at time t. If the $X_i(t)$ is zero, the gene i cannot interact with other genes. If the ith gene interacts with other genes, the following events occur:

- With probability p^+_{ij}, gene i activates gene j; when this happens, $X_i(t)$ is depleted by 1 and $X_j(t)$ is increased by 1
- With probability p^-_{ij}, gene i inhibits gene j; when this happens, both $X_i(t)$ and $X_j(t)$ are depleted by 1
- With probability p_{ijl} gene i joins with gene j to act upon gene l in excitatory mode, as a result of which both $X_i(t)$ and $X_j(t)$ are reduced by 1, while $X_l(t)$ is increased by 1. However, our models, in this thesis, do not include this type of interaction.
- With probability d_i, which is defined as follows,

$$d_i + \sum_{j=1}^{n} \left(p^+_{ij} + p^-_{ij} + \sum_{l=1}^{n} p_{ijl} \right) = 1$$

  the signal of gene i exits the system so $X_i(t)$ is depleted by 1. $p_{ijl} = 0$ in this thesis.

Let λ^+_i and λ^-_i be the positive and negative packet input rates to the ith node, respectively, and μ_i is the packet firing rate (service rate) of the ith node. The network state is represented by the n-vector of non-negative integers $\mathbf{x} = \{x_1, ..., x_n\}$, and we also define the vectors of non-negative integers $\mathbf{x}^+_i = \{x_1, ..., x_i+1, ..., x_n\}$, $\mathbf{x}^-_i = \{x_1, ..., x_i - 1, ..., x_n\}$, $\mathbf{x}^{+-}_{ij} = \{x_1, ..., x_i + 1, x_j - 1, ..., x_n\}$, Now let the random process $\mathbf{X}(t) = \{X_1(t), ..., X_n(t)\}$ be defined with $X_i(t)$. If $P(\mathbf{x}, t)$ is the probability that $\mathbf{X}(t)$ takes the value $\mathbf{x}$ at time t, then the balanced equation of the G-networks is:

$$
\begin{aligned}
P(\mathbf{x}, t+\Delta t) = \sum_{i=1}^{n} \Big[& (\lambda_i^+ \Delta t + o(\Delta t)) P(\mathbf{x}_i^-, t) I(\mathbf{x}_i > 0) + (\lambda_i^- \Delta t + o(\Delta t)) P(\mathbf{x}_i^+, t) \\
& + \sum_{j=1}^{n} \Big\{ (\mu_i p_{ij}^+ \Delta t + o(\Delta)) P(\mathbf{x}_{ij}^{+-}, t) I(\mathbf{x}_j > 0) \\
& + (\mu_i p_{ij}^- \Delta t + o(\Delta)) P(\mathbf{x}_{ij}^{++}, t) + (\mu_i p_{ij}^- \Delta t + o(\Delta)) P(\mathbf{x}_i^+, t) I(\mathbf{x}_j = 0) \Big\} \\
& + (\mu_i d_i \Delta t + o(\Delta t)) P(\mathbf{x}_i^+, t) + (1 - (\lambda_i^+ + \lambda_i^- + \mu_i) \Delta t + o(\Delta t)) P(\mathbf{x}, t) \Big]
\end{aligned}
\tag{2.33}
$$

where $I(C)$ is 1 if C is true and 0 otherwise, and $o(\Delta t) \to 0$ as $\Delta t \to 0$. The first term describes the increment of the ith queue length by its positive packets and the second term is explaining the decrement of the ith queue length by negative packets. The third term is the probability that a positive packet moves from the ith node to the j node. The fourth and fifth terms describe a negative packet coming from the ith node decreases the jth queue length. The sixth term is the probability that a packet in the ith node acts its function but nothing is happened except the packet is disappeared. The last term depicts the probability that nothing is happened.

In order to find a solution of this equation, let's divide (2.33) by Δt and introduce the equilibrium probability distribution of the system $P(\mathbf{x}) = lim_{t\to\infty} P(\mathbf{x}, t)$ then we obtain following dynamic behaviour,

$$
\begin{aligned}
\frac{\partial P(\mathbf{x})}{\partial t} = \sum_{i=1}^{n} \Big[& \lambda_i^+ P(\mathbf{x}_i^-) I(\mathbf{x}_i) > 0) + (\lambda_i^- P(\mathbf{x}_i^+) + \mu_i d_i) P(\mathbf{k}_i^+) \\
& + \sum_{j=1}^{n} \Big\{ \mu_i p_{ij}^+ P(\mathbf{x}_{ij}^{+-}) I(\mathbf{x}_j > 0) + \mu_i p_{ij}^- P(\mathbf{x}_{ij}^{++}) + \mu_i p_{ij}^- P(\mathbf{x}_i^+) I(\mathbf{x}_j = 0)) \Big\} \\
& - (\lambda_i^+ + \lambda_i^- + \mu_i) P(\mathbf{x}) I(\mathbf{x}_i > 0) \Big]
\end{aligned}
\tag{2.34}
$$

Now, let's assume the following probability that gene i is expressed in steady-state,

$$
q_i = \frac{\Lambda_i^+}{\mu_i + \Lambda_i^-} \tag{2.35}
$$

$$
\begin{aligned}
\text{where} \quad \Lambda_i^+ &= \lambda_i^+ + \sum_{j=1}^{n} q_j \mu_j p_{ji}^+ \\
\Lambda_i^- &= \mu_i + \lambda_i^- + \sum_{j=1}^{n} q_j \mu_j p_{ji}^-
\end{aligned}
\tag{2.36}
$$

Then we can show that the following product form solution satisfies Equation (2.33) in steady-state [48].

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} q_i^{x_i}(1 - q_i) \tag{2.37}$$

where $q_i < 1$. That is by substituting Equations (2.36) and (2.37) into Equation (2.34) whose dynamics is zero, Equation (2.33) can be proven to satisfy its equality [48]. Although queueing theory is a common computational tool, The G-Networks are an essential departure from queueing theory; in particular conventional queues could not be possibly applied to GRNs because the notion of inhibition does not exist in queueing theory but was introduced by the G-Network theory.

3 Stochastic GRN Modeling with Switch-like Gene Responses

Based on the biological assumptions presented in Section 2.1, this chapter reviews the conventional stochastic modeling approach with Gillespie algorithm [54] which is generally used to find a possible solution of stochastic reaction equations. Additionally, we will show this stochastic model can efficiently explain the switch-like behavior of a gene by employing a Hill function. This chapter provides the understandings of detailed stochastic gene expression processes and suggests an efficient way to explain a binary response of a gene.

3.1 Introduction

The gene expression is the sum of processes that results in a specific level of mRNAs and proteins. So the understanding of underlying gene expression mechanisms can be a good starting point to solve many biological problems in systems biology. For example, gene regulatory networks (GRNs) have become main tools to uncover the relationships between genes, whose mechanisms are based on the gene expression changes [42, 91]. Though there are many studies of gene expression using mathematical and statistical models, there are still many controversial issues about how such models can provide decisive insight into gene expression processes. In several recent studies, gene expression is described by discrete and probabilistic processes rather than as continuous and deterministic concentration changes [81, 94]. Also, dynamics of GRNs is investigated using a stochastic modeling approach [50] which enables us to detect abnormal gene behaviors in steady-states [67].

The stochasticity known as ”gene expression noise” can be either disadvantageous for reduction or beneficial in a sense of evolution [75]. But recently the stochastic nature of biological processes has been identified as playing important functional roles in a cell such as physiological regulation mechanisms and evo-

lutionary adaptation [34]. ODE based deterministic models do not easily explain these noise oriented cellular functions. So, many ODE based studies added a noise term to their ODE in order to describe various biological processes such as protein folding [121], oscillatory expressions [28], and network dynamics [1]. For example, Navozhay *et. al.* determined how a negative feedback regulation affects to the gene expression of TetR-based synthetic transcriptional networks [86]. They used stochastic simulations for their gene regulatory model, but ODEs are also employed to describe the dose-responses of the negative reaction which shows S-shape expression profiles in their synthetic experiment.

Stochastic simulation models are also widely used to describe the stochastic gene expression mechanisms [81, 94]. The Gillespie algorithm is generally used in most of the stochastic models to find a possible solution of their stochastic reaction equations [54]. In stochastic models, however, switch-like behaviour of a gene in is only achieved by gene expression cascade consisting of two more genes, which means the model requires additional set of parameters. It could be a disadvantage of simulation based approach because of the computation time. The situation could be worse when corresponding parameter values are not known. To tackle this problem, we modified the Gillespie algorithm by adding the Hill function which describes protein activation. Our approach is applied to the toggle switch model [43] and the ScbA-ScbR antibiotic metabolite production system of microorganisms [82] to simulate their expression data which show switch-like and oscillatory behavior.

3.2 Gene Expression Model

In this study, eight processes of the gene expression are assumed while the expression mechanism is conventionally explained by three stage models (promoter activation, transcription, and translation) [104]. For each stage of the eight processes, its corresponding reaction rate parameters are derived from published information. Figure 3.1 depicts the assumptions of our model. Let n_1, n_2, and n_3 represent the copy number of activated gene (DNA), mRNA, and protein, respectively. The stationary distribution of n_1 is binomial with ON (λ_1) and OFF (γ_1) rates. mRNAs are initiated by a RNA polymerase which bind to DNA string with average time average interval T_{avg}. So this transcription initiation process has an exponential distribution with $\lambda_2 = 1/T_{avg}$. Then proteins are produced by ribosomes attached to an mRNA until this mRNA is degraded by an RNase-E. But because the RNase-

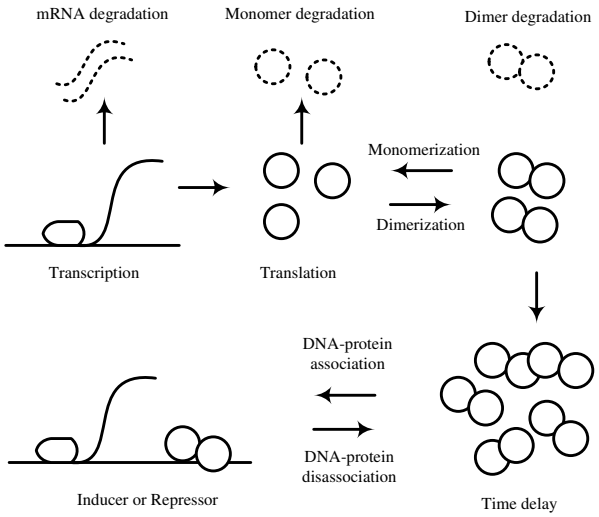


Figure 3.1: Assumptions for the stochastic gene expressions

E shares the binding site with ribosome, the degradation rate of mRNAs can be obtained using the fact that the average number of proteins per one mRNA (protein bursting size b) is equal to the average of a negative binormial distribution. The other processes such as protein-protein/protein-DNA binding and phosphorylation/ubiquitination are also considered and their parameters (Table 2.1) are obtained in the basis of published literatures. Please see Section 2.2 for more detailed stochastic modeling for each process.

3.3 Hill Function for Protein Activation

In an ordinary differential equation (ODE) model, the net change of the mRNA concentration is generally described by a simple rate equation, $\frac{dn_2}{dt} = \lambda_2 \cdot g([P]) - \gamma_2 \cdot n_2$ where $g([P])$ is the Hill function of protein P (transcription factor) [32]. This Hill function $g(x)$ is generally defined as follows,

$$g(x) = \frac{x^n}{K^n + x^n} \tag{3.1}$$

where K is the threshold for the regulatory influence of x on a target gene, and n is a steepness parameter. This function increases from 0 to 1 as x goes ∞ and the inhibitory function is $1 - g(x)$. In our study, this Hill function is used to compute the propensity of protein activation in the Gillespie algorithm.

The goal of the Gillespie algorithm is to estimate the state vector $\mathbf{X}(t) = (X_1(t), ..., X_N(t))$, given the system initial state $\mathbf{X}(t_0) = \mathbf{x}_0$ where $X_i(t)$ denotes the number of molecules of species S_i in the system at time t. Consider M chemical reactions $\{R_1, ..., R_M\}$. Each reaction R_j is characterized by two quantities, state-change vector $\mathbf{v}_j = (v_{1j}, ..., v_{Nj})$ and propensity function a_j. v_{ij} is the change of the ith molecular population caused by one R_j reaction. If the system is in state $\mathbf{x}$ and R_j occurs than the next state of the system will be $\mathbf{x} + \mathbf{v}_j$. The propensity function a_j can be defined so that $a_j(x)dt = Pr\{R_j \text{ will occur in } [t, t + dt) | \mathbf{X}(t) = \mathbf{x}\}$ and

$$a_j(x) = c_j[x] \tag{3.2}$$

where $[x]$ the concentration of the involved species x and c_j is the kinetics of the reaction R_j. We now add the Hill function (3.1) to (3.2) to describe the switch-like behavors of a gene as follows.

$$a_j(x) = c_j[x] \cdot \frac{[x]^n}{K^n + [x]^n} \tag{3.3}$$

For example, let's assume a dimer protein PP_A represses the expression of gene B by binding to the operator of gene B.

$$Pro_B + PP_A \xrightarrow{c_{a2}} Pro_B PP_A \tag{3.4}$$

Then, the probability that PP_A will undergo this reaction in the next dt is $[Pro_B][PP_A]c_{a2}dt$. So the propensity function is $a_j([PP_A], [Pro_B]) = c_{a2}[Pro_B][PP_A]$. But if the gene B is known to have the switch-like response, its propensity function is redefined as follows

$$a_j(PP_A, Pro_B) = c_{a2} \cdot [Pro_B] \cdot [PP_A] \cdot \frac{[PP_A]^n}{K^n + [PP_A]^n} \tag{3.5}$$

In this study, K and n are assumed 50 and 7.

3.4 Toggle Switch Gene Expression Model

The toggle switch system, proposed by [43], consists of two repressor molecules and corresponding two inducers. In their model, a switch is achieved by controlling the levels of inducers. In this study, however, we focus on the effect of the Hill functions on the behavior of the two genes without their inducers. Figure 3.2 shows

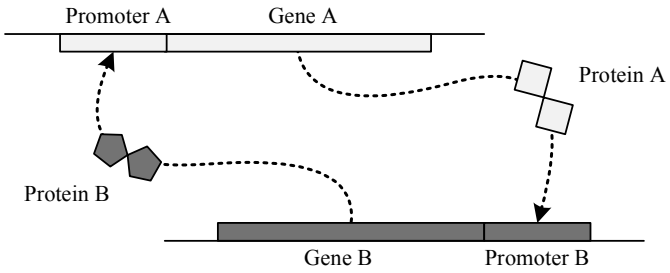


Figure 3.2: The toggle switch system. As repressors, two genes, A and B, are inhibited by each other. Each gene has five reactions as described in Equations (3.6) $\sim$ (3.10)

our toggle switch system. Each promoter is inhibited by the repressor which is transcribed by the opposing promoter.

$$Pro_i(t) + TF_i(t) \xrightarrow{\lambda_2} Pro_i(t) + TF_i(t) + R_i(t) \tag{3.6}$$

$$R_i(t) \xrightarrow{\lambda_3} R_i(t + \Delta t) + P_i(t + \Delta t) \tag{3.7}$$

$$P_i(t) + P_i(t) \underset{c_{d1}}{\overset{c_{a1}}{\rightleftharpoons}} PP_i(t) \tag{3.8}$$

$$Pro_i(t) + PP_j(t) \underset{c_{d2}}{\overset{c_{a2}}{\rightleftharpoons}} Pro_iPP_j(t) \tag{3.9}$$

$$\begin{aligned} R_i(t) &\xrightarrow{\gamma_2} \phi(t) \\ P_i(t) &\xrightarrow{\gamma_3} \phi(t) \\ PP_i(t) &\xrightarrow{\gamma_4} \phi(t) \end{aligned} \tag{3.10}$$

where $i, j \in \{A, B\}$. Equation (3.6) shows the transcription process where TF_i is a transcription factor. Equations (3.7) and (3.8) are for the translation and dimeriztion processes, respectively. Equation (3.9) represents the repression process of protein A binding to the promoter B. So the expression of protein B can be the reporter of the repressor-inducer binding activity. In order to simulate this system and to compare the expression trajectory of each molecule, we used the Gillespie algorithm which was described in Chapter 2.

Figure 3.3 shows our simulated protein expression profiles of gene A (gray) and B (black) when the Hill function is not applied (upper) and is applied (lower). In both cases, the system state switches depending on the concentrations of protein A and B. However, the system with the Hill function does not change its states as

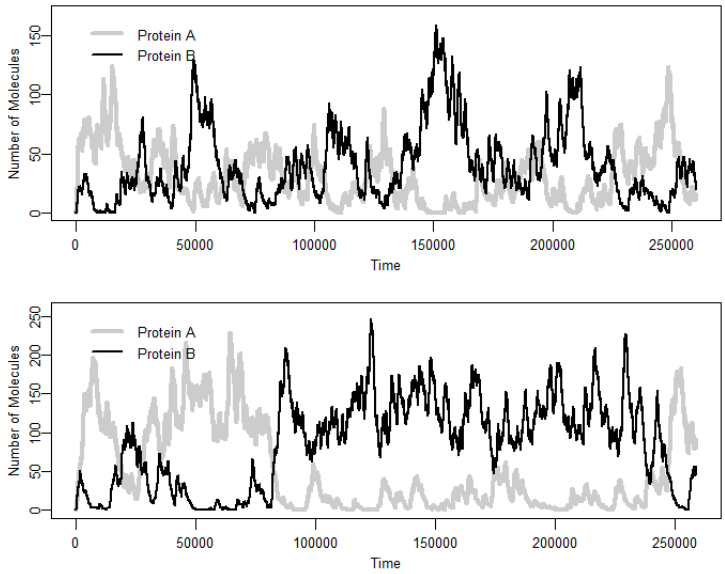


Figure 3.3: The protein expression of the toggle switch model without (upper) and with (lower) the Hill functions. In both cases, expressions are changed due to the stochastic noise but the state switching frequency of the system with the Hill function is not as much as that of the system without the Hill function

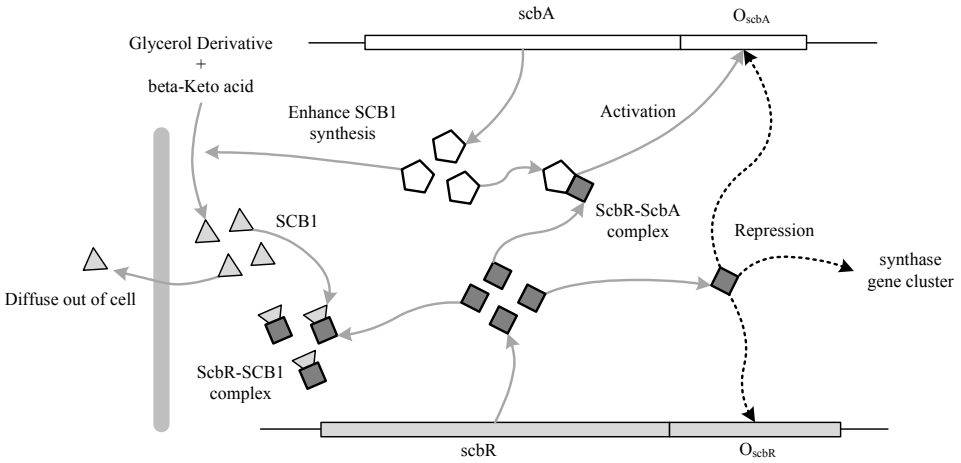


Figure 3.4: ScbR/ScbA System

much as those of a system without the Hill function. It is mainly because the Hill function prevents fluctuations of the protein expressions if the signal strength of an activator or an inhibitor is below its threshold.

3.5 ScbA-ScbR System

Many microorganisms produce variety of secondary metabolites. Most of these metabolites have antibiotic activities to provide an advantage in their survival, and some of them are commercially valuable. Antibiotics, however, are often toxic even to their producers so some microorganisms make signaling molecules to control the production of the antibiotics [82]. SCB1 is known that its addition to agar culture develops localized antibiotic production in *S. coelicolor* [113]. In Figure 3.4, a gene pair, scbA and scbR, is involved in the regulation of the synthesis of SCB1 [82]. These two genes encode a cytoplasmic receptor protein (ScbR) and an amplifier protein (ScbA), respectively. ScbR can directly bind to the promoter of cryptic type I polyketide synthase gene cluster (cpk) and represses its expression. cpk is a group of genes required for the biosynthesis of a given secondary metabolite and its genes are found to be the principle target of the extra-cellular regulatory molecule, SCB1 [112].

However, the production of this gene cluster has not been determined. Sarika Mehra *et. al.* [82] studied their hypothesis of the formation of a ScbA-ScbR complex using mathematical models and they showed their system exhibits bistability

and the states of the system are switched in response to a threshold SCB1. In this system, cells are initially OFF steady-state where ScbR represses the expression of cpk genes. If the concentration of SCB1 is over a certain criterion, the state is turned ON which means SCB1 inhibits the activity of ScbR so cpk genes are expressed. We simulate this system having the Hill type reaction in binding the SCB1 and ScbR molecules which are observed as the reporters of the system.

Following equations represent the reactions of our simulated system (Pol: RNA polymerase, SR: ScbR, SA: ScbA, SB: SCB1). Reactions for ScbR are shown in Equations (3.11) - (3.15) and Equations (3.16) - (3.20) are for ScbA. ScbR activates SCB1 from the inactive form of SCB1 (P^*_{SB}) in Equation (3.21). The Hill function is applied in Equation (3.22) for the ScbR-SCB1 binding reaction. Repressions of ScbR are depicted in Equations (3.25) and (3.26).

$$Pro_{SR}(t) + TFPol_{SR}(t) \xrightarrow{\lambda_2} Pro_{SR}(t) + TFPol_{SR}(t) + R_{SR}(t) \tag{3.11}$$

$$TF_{SR}(t) + Pol_{SR}(t) \underset{c_{d2}}{\overset{c_{a2}}{\rightleftharpoons}} TFPol_{SR}(t) \tag{3.12}$$

$$R_{SR}(t) \xrightarrow{\lambda_3} R_{SR}(t + \Delta t) + P_{SR}(t + \Delta t) \tag{3.13}$$

$$R_{SR}(t) \xrightarrow{\gamma_2} \phi(t) \tag{3.14}$$

$$P_{SR}(t) \xrightarrow{\gamma_3} \phi(t) \tag{3.15}$$

$$Pro_{SA}(t) + TFPol_{SA}(t) \xrightarrow{\lambda_2} Pro_{SA}(t) + TFPol_{SA}(t) + R_{SA}(t) \tag{3.16}$$

$$P_{SA}P_{SR}(t) + Pol_{SA}(t) \underset{c_{d2}}{\overset{c_{a2}}{\rightleftharpoons}} TFPol_{SA}(t) \tag{3.17}$$

$$R_{SA}(t) \xrightarrow{\lambda_3} R_{SA}(t + \Delta t) + P_{SA}(t + \Delta t) \tag{3.18}$$

$$R_{SA}(t) \xrightarrow{\gamma_2} \phi(t) \tag{3.19}$$

$$P_{SA}(t) \xrightarrow{\gamma_3} \phi(t) \tag{3.20}$$

$$P_{SA}(t) + P^*_{SB} \xrightarrow{c_{a3}} P_{SA}(t) + P_{SB}(t) \tag{3.21}$$

$$P_{SB}(t) + P_{SR}(t) \underset{c_{d1}}{\overset{c_{a1}}{\rightleftharpoons}} P_{SB}P_{SR} \tag{3.22}$$

$$P_{SB}(t) \xrightarrow{\gamma_3} \phi(t) \tag{3.23}$$

$$P_{SB}P_{SR}(t) \xrightarrow{\gamma_4} \phi(t) \tag{3.24}$$

$$P_{SR}(t) + Rpo_{SR}(t) \underset{c_{d2}}{\overset{c_{a2}}{\rightleftharpoons}} P_{SR}Rpo_{SR} \tag{3.25}$$

$$P_{SR}(t) + Rpo_{SA}(t) \underset{c_{d2}}{\overset{c_{a2}}{\rightleftharpoons}} P_{SR}Rpo_{SA} \tag{3.26}$$

$$P_{SA}(t) + P_{SR}(t) \underset{c_{d1}}{\overset{c_{a1}}{\rightleftharpoons}} P_{SA}P_{SR} \tag{3.27}$$

$$P_{SA}P_{SR}(t) \xrightarrow{\gamma_4} \phi(t) \tag{3.28}$$

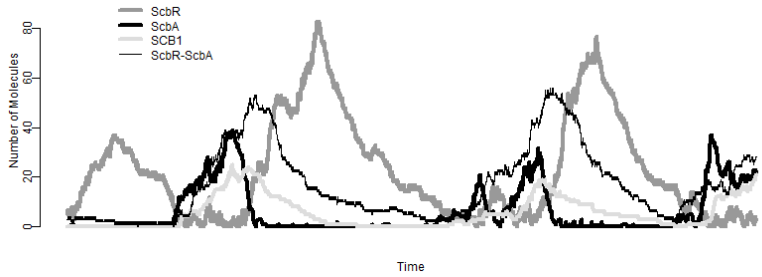


Figure 3.5: Oscillatory expression profiles engendered by the negative auto-regulatory feedback of ScbR (thick and dark gray). As ScbR level decreases, ScbA (thick and black) expression increases, which causes SCB1 (thick and light gray) level growth. On the other hand, the repression effect of ScbR on itself and ScbA is recovered again as ScbR expression increases.

Though, the parameters in Table 2.1, say "$C1$", are derived from published information to reduce the parameter space, it is necessary to find a set of appropriate parameter values ("$C2$") so as to this Scb-ScR system shows oscillatory expression pattern triggered by the negative feedback of ScbR. The empirically found parameter values, $C2$, include $c_{a1} = 0.0005, c_{d1} = 0.0001, c_{a2} = 0.1, c_{d2} = 0.1, c_{a3} = 0.000001$, and $c_{d3} = 0.00008$. All the other parameters are the same as Table 2.1. Note that the system behavior is largely depending on the parameter values. So, as further work, it is necessary to perform sensitivity analysis in order to evaluate how robust these empirically chosen values are.

In Figure 3.5, the system clearly shows the oscillatory expression caused by the negative auto-regulatory feedback of ScbR [96]. Also Figure 3.6 illustrates the switch-like behavior of the system. The levels of ScbA (Right panel) and ScbR (Left panel) decreases and increases, respectively, as SCB1 concentration rises. The Hill function applied system (rectangle) shows the S-shape curve of each molecules depending on the level of SCB1 in both conditions of $C1$ (black solid lines) and $C2$ (gray solid lines). In these two different conditions, the system shows the similar switch-like expression patterns because of the Hill function. So we expect this switch-like behavior would also be observed in the other set of parameter conditions as long as the Hill function is embedded in the system.

Additionally, we looked into the sensitivity of parameter c_{a3} in Equation (3.21) where the Hill function is applied. In Figure 3.7, the mean values of ScbA expression in the system with the Hill function are distributed between 0 and 100 for all

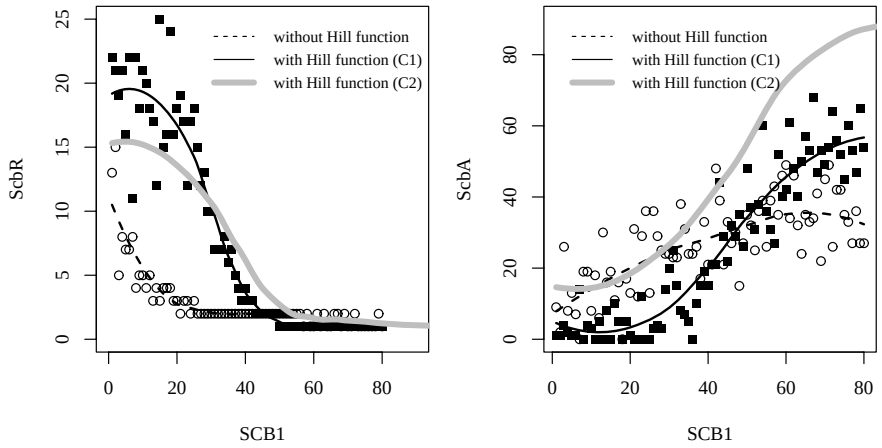


Figure 3.6: Switch like behavior of ScbR (Left) and ScbA (Right) where solid line with rectangle and dotted line with circle represent the corresponding molecules with and without the Hill function, respectively in the condition $C1$. The X and Y axes of each plot represent the number of the corresponding molecules. Gray line depicts the switch like behavior of each molecule with the condition $C2$. The lines are predicted values using the local regression method (Loess).

c_{a3} values while the non Hill function applied system shows sudden expression burst up to 200 when $c_{a3} \geq 0.001$. Note that, in condition $C2$, the ScbA-ScbR system has both oscillatory and switch-like behavior within 100 protein level. So the Hill function makes a parameter less sensitive which could be an advantage in modeling with stable gene expression. However, becoming low sensitive parameter could be also a drawback in finding optimal parameters, which will be discussed in the next section.

3.6 Conclusions

In ODE models, the Hill function is commonly used to describe the chemical reactions of metabolites and the switch-like gene responses. In this study, we propose a delicate stochastic gene expression modelling method with the Hill function which can successfully characterize the switch-like gene behavior. Several important biological properties including protein dimerization, time delay of transcription, and DNA-protein, protein-protein interactions are considered and their stochastic rate parameters are estimated based on published information.

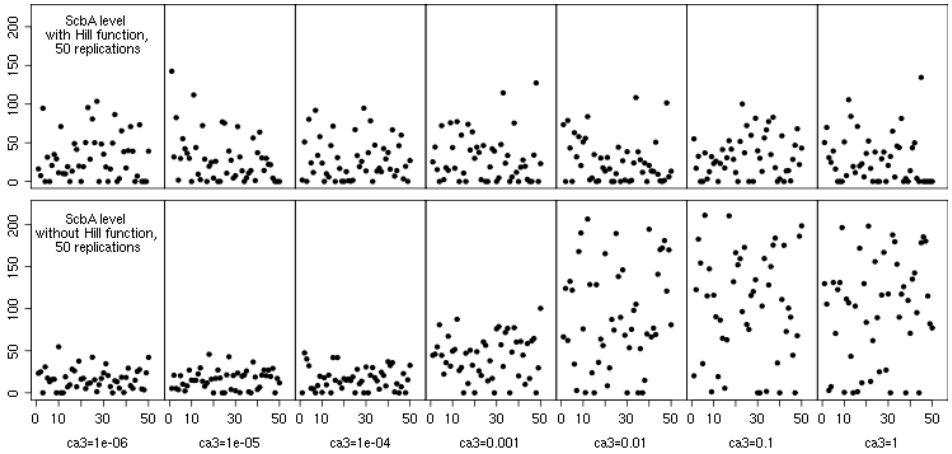


Figure 3.7: ScbA expression levels by varying the value of c_{a3} in Equation (3.21) in condition $C2$. There are 50 replications for each value of c_{a3} and each dot represents the mean of ScbA expression for 50,000 seconds. Upper and lower panels show the ScbA level with and without the Hill function, respectively.

Our toggle switch system with the Hill function shows robust expression patterns against the internal/external perturbations as its Hill function makes the system stable under the condition below a certain criterion. The resistance of our body to external factors causing a disease might be an example of this robust gene response. Though the original paper of the ScbA-ScbR experiment shows the three steady-states (OFF, ON, and bi-stable) along with their possible parameter spaces, our study shows only switch-like behaviors of the system with fixed parameters as this study focuses on the effect of the Hill function.

Though our model shows successful switch-like behavior, The Hill function could be a drawback of stochastic models when we search for optimal parameter values because the system with the Hill function becomes less sensitive. For example, in Figure 3.7, we observed that the system is stable with $c_{a3} \geq 0.001$ while the Hill function applied system provide no evidence of this information. So the Hill function should be carefully applied to the system we interested in. Otherwise, we can apply the Hill function to the pathway which is less important but need to be stable expression.

It has been important to be able to express analytical solutions for complex stochastic networks so that the computational effort and potential numerical errors can be reduced as much as possible. Related work with an engineering flavour

[49] has aimed at obtaining useful analytical solutions for genetic algorithms, so that closed form expressions can be used to replace lengthy Monte Carlo simulations. Other related work on obtaining analytical solutions to complex multiparticle stochastic models relates to the interactions and motions of molecules [45, 46]. Such analytical results also have the advantage of avoiding the need to truncate very large or infinite systems of equations that describe the related Markov processes, thus avoiding the inevitable truncation errors when an infinite or very large system of equations is replaced by a smaller number of equations. Additionally, by taking the advantages of these analytical solutions, we will extend the previous G-network based GRN model with more delicate biological assumptions, which will be introduced in the next chapter.

4 G-network based Abnormal Gene Expression Detection

Though previous chapter successfully describes the gene expression process using stochastic modeling approach, the numerical way that solves the reaction equations limits its availability in a genome scale analysis. In this chapter, complex GRN processes are simplified and then we applied G-networks whose customers contain mRNA expression information and move around nodes in the networks. Since G-networks have an analytic solution of the stochastic reaction equations, it could be used to model large-scale GRNs. Based on this technique, this chapter presents an abnormal gene expression detection method while Chapter 5 will describe more improved technique which is for abnormal pathway detection. As the first G-network application of our thesis, this chapter focuses on to demonstrate the availability of G-networks in modeling real biological GRNs where the various kinetics of biological processes are mapped to the parameters of the G-networks.

4.1 Introduction

Identifying the key features and dynamics of gene regulatory networks (GRNs) is an important step towards understanding behaviors of biological systems. Thanks to the development of high-throughput technology, the amount of microarray gene expression data has greatly increased, and numerous mathematical models attempt to explain gene regulations using gene networks [42, 101]. Once the network structure of a system is inferred, its long-run behavior needs to be investigated. However, most methods focus on the inference of a network structure which only provides the snapshot of given data. Probabilistic Boolean networks (PBNs) provide a way to find the steady-state of GRNs [105], but PBNs are limited by the computational complexity of both their reverse engineering algorithm and their steady-state probability estimation [29].

In [50], a new approach to the steady-state analysis of GRNs based on G-

Network theory [47, 48] is proposed, while G-Networks were firstly applied to GRNs with simplifying assumptions concerning gene expression in [2]. However, the G-Network approach also exhibits specific difficulties caused by the large number of parameters. Thus, in this study, we reduce the number of model parameters on the basis of biological assumptions and focus on estimating two parameters in particular: the total input rate which means the mRNA initiation rate and steady-state probability that a gene is expressed (activated). In this way, abnormal gene expression patterns are identified in a system where, in a given condition, the abnormal expression is defined as the unexpected different over/under expression level from that of a normal condition. For the evaluation of the proposed method, we used a simulation dataset generated by using the stochastic gene expression models processed with the widely accepted Gillespie algorithm [55, 17]. Also a real biological dataset obtained from the cell cycle of the budding yeast [92] is examined using the proposed method.

There are two essential novelties in this study. First, our approach enables us to obtain the steady-state of GRNs with only polynomial computational complexity thanks to the product form solution of G-Networks. The computational cost due to large memory space and non-polynomial computational complexity are basic limitations in conventional methods such as PBN. Also our method can provide new measures such as the input rate and the steady-state probability to detect differentially expressed genes in microarray analysis.

4.2 Abnormal Gene Expression Detection

Please see Section 2.4.3 for the detailed GRN modeling via G-networks. Our aim is to estimate the input rate (λ_i^+) of the ith gene and its steady-state probability, (q_i) in Equation (2.36). Let's define $f_i^+(q_j) = \sum_{j=1}^{n} q_j \mu_j p_{ji}^+$ and $f_i^-(q_j) = \sum_{j=1}^{n} q_j \mu_j p_{ji}^+$ where the probabilities, p_{ij}^+ and p_{ij}^-, are set to be one for the simplicity. Then q_i can be rewritten as follows

$$q_i = \frac{\lambda_i^+ + f_i^+}{\mu_i + \lambda_i^- + f_i^-} \tag{4.1}$$

We fix the μ_i as the out-degrees of gene i and the degradation rate of mRNA, λ_i^-, as a constant $0.00578 sec^{-1}$ [115]. Eventually, we need to estimate only two parameters, the input rate, λ_i^+, and the steady-state probability, q_i. Let λ_i^{lower} is the lower bound of the λ_i^+, which is larger than zero. The lower bound is regarded

as an initial transcription rate without any external input. In this study, we use $\lambda_i^{lower} = 0.0025$ [57, 115]. The upper bound of the input rate, λ_i^{upper} , is obtained by assuming all positive effects from other genes to the ith gene are zero and the other genes giving negative effects are fully activated. That is

$$\lambda_i^{upper} = q_i^* \big(\mu_i + \lambda_i^- + f_i^-(q_j^*)\big)$$

where $\lambda_i^- = 0.00578$ and $q_i^* = q_j^* = 1$. As equations of q_i for all genes cannot be solved one after the other, an iterative method is used to find the optimal solution of q_i. Let $q_i^{(0)}$ is the initial value of q_i. Then $q_i^{(0)}$ can be obtained as follows,

$$\begin{aligned} q_i^{(0)} &= E_j\big[q_{ij}^{(0)}\big] \\ &= E_j\big[x_{ij}/\max(x_{ij})\big] \end{aligned}$$

where x_{ij} is the observed mRNAs expression level of the ith gene at the jth time point and $\max(x_{ij})$ is the maximum value among all observations of the ith gene. Let λ_{il} be an input rate between the lower and the upper bounds ($\lambda_i^{lower} \leq \lambda_{il} \leq \lambda_i^{upper}$). For an input λ_{il}, the steady-state probability q_{il} can be obtained numerically by solving Equation (4.1) with the $q_i^{(0)}$. Once the steady-state probability is determined, the log-likelihood (say $\log L_{il}$) of the given model can be computed by using Equation (2.37) which is the same form of the likelihood of geometric distribution. It is known that the log-likelihood of geometric distribution is convex so we choose appropriate value $\lambda_i^{(0)}$ which maximizes the log-likelihood function.

$$\lambda_i^{(0)} = \arg\max_{\lambda_{il}}(\log L_{il}) \tag{4.2}$$

$$\text{where } \log L_{il} = \log\left(\prod_{i=1}^{n} q_{il}^{x_i}(1-q_{il})\right)$$

Now we can compute $q_i^{(1)}$ by substituting λ_i^+ to $\lambda_i^{(0)}$ in Equation (4.1) and, in the same way of $\lambda_i^{(0)}$, $\lambda_i^{(1)}$ can be obtained using $q_i^{(1)}$. The optimal q_i and λ_i^+ can be eventually found by updating these value iteratively until $d^{(t)} < \delta$ where $d^{(t)}$ is the difference between $q_i^{(t)}$ and $q_i^{(t-1)}$ at tth iteration. In this study, δ is 0.0001.

4.3 Simple Gene Regulatory Networks using Stochastic Gene Expression Model

In order to assess our G-Network model, we construct a simple GRN structure and generate the expression data using a synthetic stochastic gene expression model [81, 94]. This stochastic gene expression model has several biological features including protein dimerization [98] and time delay for protein signalling [81].

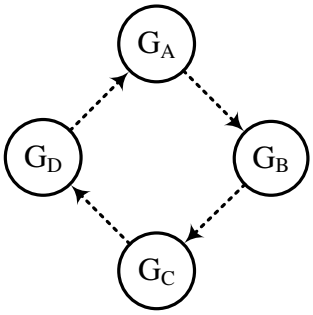


Figure 4.1: Repressilatory gene regulatory network structure with four genes. Each gene represses the transcription of its neighbor gene. This repressilator network with time delay has been shown to oscillate both theoretically and experimentally [35, 127]

Figure 4.1 shows the repressilator network structure whose gene has following gene expression principles: the number of proteins per cell chases the number of mRNAs which in turn follows the number of active genes [94]. Reaction equations (4.3) $\sim$ (4.9) give the corresponding processes (RPo: RNA polymerase open complex, Pro: promoter, R: mRNA, P: protein monomer, PP: protein dimer, 0:

degradation, t: time, and Δt: time increment):

$$\mathrm{RPo}_i(t) + \mathrm{Pro}_i(t) \xrightarrow{\lambda_2} \mathrm{RPo}_i(t) + \mathrm{Pro}_i(t) + \mathrm{R}_i(t) \tag{4.3}$$

$$\mathrm{R}_i(t) \xrightarrow{\lambda_3} \mathrm{R}_i(t) + \mathrm{P}_i(t) \tag{4.4}$$

$$\mathrm{P}_i(t) + \mathrm{P}_j(t) \xrightarrow{c_{a1}} \mathrm{PP}_{ij}(t) \tag{4.5}$$

$$\mathrm{PP}_{ij}(t) \xrightarrow{c_{d1}} \mathrm{P}_i(t) + \mathrm{P}_j(t) \tag{4.6}$$

$$\mathrm{Pro}_l(t) + \mathrm{PP}_{ij}(t) \xrightarrow{c_{a2}} \mathrm{Pro}_l\mathrm{PP}_{ij}(t + \Delta t) \tag{4.7}$$

$$\mathrm{Pro}_l\mathrm{PP}_{ij}(t) + \mathrm{PP}_{mn}(t) \xrightarrow{c_{d2}} \mathrm{Pro}_l(t + \Delta t) + \mathrm{PP}_{ij}(t + \Delta t) + \mathrm{PP}_{mn}(t + \Delta t) \tag{4.8}$$

$$\begin{aligned} \mathrm{R}_i(t) &\xrightarrow{\gamma_2} 0(t) \\ \mathrm{P}_i(t) &\xrightarrow{\gamma_3} 0(t) \\ \mathrm{PP}_{ij}(t) &\xrightarrow{\gamma_4} 0(t) \end{aligned} \tag{4.9}$$

where $i, j \in \{A, B, C, D\}$ in Figure 4.1. In addition, we assume that proteins such as transcription factors and repressors require accumulation time, Δt, for their activation [81, 17]. In order to generate gene expression data, modified Gillespie algorithm is applied [55, 17]. The cell growth rate and cell volume are fixed, and only five cells are considered. Detailed parameters are summarized in Table 2.1 with their references.

The transcription process in Equation (4.3) follows an exponential distribution with transcription initiation rate λ_2 [115]. The translation processes are given in Equation (4.4) and include direct competition between the ribosome binding site and the RNAse-E binding site which degrade the mRNAs. Thus the translation process follows a geometric distribution with probability p and busting size $b = p(1-p)$ [81, 115]. T_D is the average time interval between successive competitions, and the number of surviving mRNAs n_2 in the population after transcription is blocked with $n_2 = n_{2,0}p^{t/T_D}$. This is equal to $T_{half} = -(log(2)/log(p))T_D$ [81]. Thus the translation initiation rate, $\lambda_3 = 1/T_D$, can be computed. The protein dimer association and disassociation rates are k_{a2} and k_{d2}, respectively, as shown in Equations (4.5) and (4.6) [21]. We also consider the DNA-protein association and disassociation rates (k_{a1} and k_{d2} in Equations (4.7) and (4.8), respectively) [56]. The degradation rate of mRNA and of proteins are obtained by using the half-life of each molecule (4.9) [21, 115]. Please see Section 2.2 for detailed stochastic modeling description.

We generated three sets of expression data (Say dataset 1, 2, and 3) where

each dataset has two groups, normal and case groups. These two groups were obtained with the same parameter values except for the transcription initiation rate of G_A, $0.0012sec^{-1}$, which is the half of the transcription rate in the normal group, $0.0025sec^{-1}$. Both groups were simulated during 3,000 seconds. In order to compare these two groups, we performed not only the G-Network analysis but also the t-test which is widely used to find differentially expressed genes in the microarray analysis. Datasets 1 and 2 consist of 50 samples each of which was drawn from the whole data points. The third dataset consists of 500 samples which were randomly chosen from the original observations.

Table 4.1 summarizes the results of the three datasets. For Dataset 1, the expression of G_A is significantly different (p-value of t-test < 0.01) while the difference of the G_A expression in Dataset 2 is not significant (p-value=0.202). On the other hand, in both datasets, q_A and λ_A^+ have the lowest values among the four nodes. These two measures show the consistent results in the all datasets including large sample dataset (Dataset 3). The ratios (case/normal) also show that the q_A and λ_A^+, in the case group, are smaller than one while the other ratios stay around one. In Dataset 3, the p-value of t-test of G_B is significant along with that of G_A because the expression of G_A directly affects the expression of G_B. However, G_B is not the causal gene in this network. Our G-Network analysis reveals that only G_A has lower q and λ^+ values than other nodes including G_B. All these results concur with our simulation condition where the transcription rate of G_A is one half of the normal one.

4.4 Modeling Cell Cycle Gene Regulatory Networks in Budding Yeast

The cell cycle regulated transcription and its overall controls have been studied in detail for budding yeast [12]. Recent developments in high-throughput microarray techniques help to reveal many of yeast genes controlling the cell cycle [38] which consists of four distinct phases: Gap1 (G1), Synthesis (S), Gap2 (G2), and Mitosis (M). The cells grow during their G1 and G2 phases and their DNA is replicated during the S phase. In the M phase, cell growth stops and the cell is divided into two daughter cells that include nuclear division. Many genes are involved with specific cell cycle phases, but the number of key regulators that are responsible for the control of the cell cycle process is much smaller. Thus, based on published

Table 4.1: Input rate λ^+ and steady-state probability q_i in the four gene network

		Normal		Case				t-test
		q	λ^+	q	λ^+	q_c/q_a	λ_c^+/λ_a^+	p-value
Dataset 1 50 Samples	G_A	0.512	0.765	**0.296**	**0.465**	**0.57**	**0.60**	0.000
	G_B	0.517	0.785	0.595	0.775	1.15	0.98	0.123
	G_C	0.502	0.765	0.546	0.875	1.08	1.14	0.311
	G_D	0.487	0.735	0.563	0.875	1.15	1.19	0.127
Dataset 2 50 Samples	G_A	0.445	0.675	**0.369**	**0.565**	**0.82**	**0.83**	0.202
	G_B	0.423	0.615	0.556	0.765	1.31	1.24	0.016
	G_C	0.472	0.675	0.432	0.675	0.91	1.00	0.439
	G_D	0.510	0.755	0.525	0.755	1.02	1.00	0.661
Dataset 3 500 Samples	G_A	0.474	0.725	**0.319**	**0.495**	**0.67**	**0.68**	0.000
	G_B	0.503	0.745	0.584	0.775	1.16	1.04	0.000
	G_C	0.460	0.695	0.443	0.705	0.96	1.01	0.304
	G_D	0.521	0.765	0.541	0.785	1.03	1.02	0.122

information, we can build a cell cycle GRN with the key regulators in budding yeast as shown in Figure 4.2, although the relationships that contribute to the true regulatory network structure of the cell cycle still remain uncertain. Therefore we simplify the cell cycle network structure by selecting 13 key regulatory genes (the gray circles in Figure 4.2) and connect the genes without regard to the transcriptional and post-transcriptional processes. Figure 4.3 shows the reconstructed regulatory network structure.

The activity of cyclin-dependent kinases (CDKs) plays an important role in controlling periodic events during cell cycle. Some studies of cell cycle with high-throughput technologies have suggested alternative regulation models of periodic transcription [38]. D. Olando *et., al.* [92] measured the transcription levels of cell cycle related genes with the use of Yeast 2.0 oligonucleotide array and determined the manner in which transcription factor networks contribute to CDKs and to global regulation of the cell-cycle transcription process. This microarray dataset is also used in our study with the cell cycle network structure of Figure 4.3. The dataset consists of two groups: one group is obtained from wild-type (WT) cells and the other is from cyclin-mutant (CM) cells which are disrupted for all S-phase and mitotic cyclins (mutate clb1, 2, 3, 4, 5, and 6).

The microarray data consist of a total of 30 data points taken over 270 minutes. We subdivided it into five states (groups), each consisting of 6 data points. The expression levels were transformed by taking the natural logarithm. Figure 4.4 de-

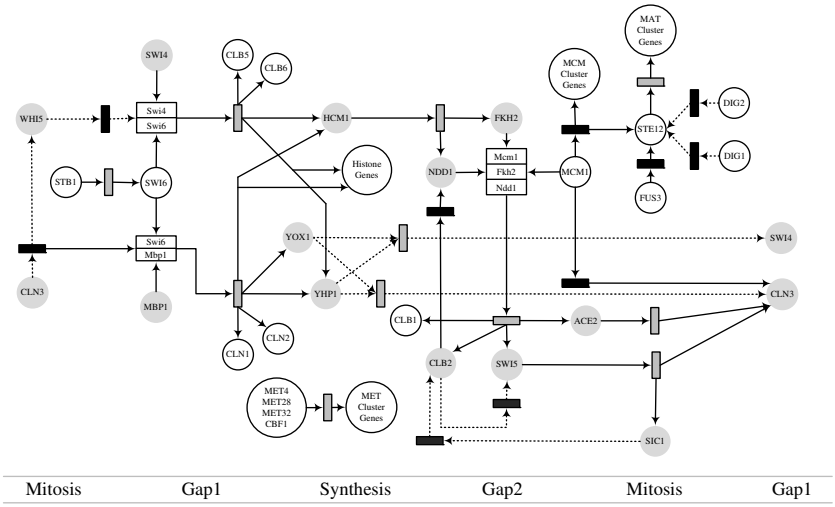


Figure 4.2: Cell cycle regulatory network structure in budding yeast. The genes are represented by circles. Complex molecules consisted of two more proteins are represented by white rectangles. The gray and black boxes are the transcription and post-transcription processes, respectively. The activations are depicted by the solid lines and the inhibitions or repressions are shown by the dashed lines. Only the gray circles are used in the G-network model.

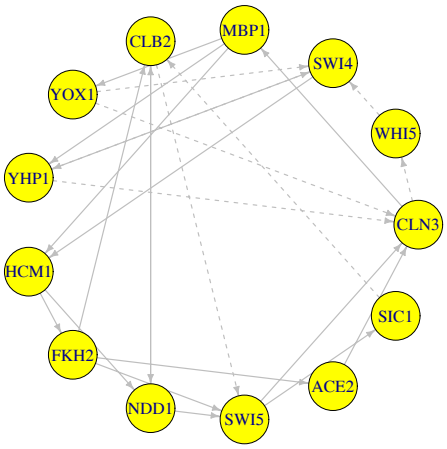


Figure 4.3: Cell cycle regulatory network structure with selected 13 genes. In terms of G-networks, each node represents a queue where customers stored. Customers move around through the edges. Solid and dashed lines are the positive and negative interactions, respectively.

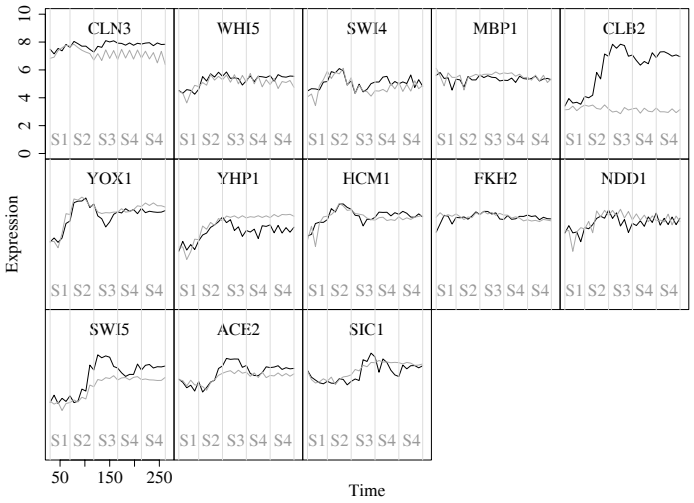


Figure 4.4: Expression profiles of selected 13 genes. Black and gray lines represent the wild-type and clb-mutant groups' expression levels, respectively.

picts the transformed expression profiles of the 13 genes with 5 states. The black and gray solid lines are the expression profiles from WT and CM cells, respectively, and S1, S2, ... , S5 represent the five states. It is obvious that the profiles of CLB2 are different between WT and CM cells because the CM dataset is designed to monitor the cell cycle processes without the clb cyclines.

Table 4.2 summarizes the steady-state probabilities of 13 genes in the cell cycle GRN. All genes have similar steady-state probabilities in the WT and CM cell groups except for CLB2 in the CM group, which has a lower steady-state probability than the elements of the WT group: as shown in Table 3, the ratio of CM/WT is smaller than one (bold letter). This smaller probability can be explained by considering the experimental design of the CM dataset which is obtained without clb cyclines. Also, the original study of this dataset suggested alternative cell cycle regulatory pathways in [92] which had revealed that over 70% of the cell cycle related genes were expressed periodically without the clb cyclines. In our results, the steady-state probabilities of the CM group are consistent with that of the WT group. These results draw the same conclusion as the original study, i.e. that the steady-state of the 12 genes does not entirely depend on the expression of CLB2. Table 4.3 shows the estimated total input rate of the 13 genes. This result also

Table 4.2: Steady state probabilities in the yeast cell cycle GRN model.

State	Cells	CLN3	WHI5	SWI4	MBP1	CLB2	YOX1	YHP1	HCM1	FKH2	NDD1	SWI5	ACE2	SIC1
	WT	0.880	0.813	0.829	0.839	0.784	0.99	0.803	0.843	0.855	0.836	0.799	0.99	0.99
S1	CM	0.878	0.814	0.818	0.848	0.770	0.99	0.802	0.842	0.864	0.839	0.787	0.99	0.99
	C/W	0.998	1.001	0.987	1.011	**0.981**	1.00	0.999	0.999	1.011	1.004	0.986	1.00	1.00
	WT	0.882	0.845	0.845	0.840	0.847	0.99	0.850	0.870	0.863	0.863	0.825	0.99	0.99
S2	CM	0.876	0.837	0.846	0.847	0.769	0.99	0.853	0.873	0.865	0.861	0.807	0.99	0.99
	C/W	0.994	0.990	1.000	1.008	**0.909**	1.00	1.004	1.004	1.002	0.998	0.978	1.00	1.00
	WT	0.890	0.840	0.826	0.846	0.886	0.99	0.844	0.855	0.863	0.854	0.871	0.99	0.99
S3	CM	0.880	0.846	0.820	0.849	0.751	0.99	0.863	0.863	0.869	0.870	0.840	0.99	0.99
	C/W	0.989	1.008	0.993	1.003	**0.847**	1.00	1.022	1.010	1.007	1.019	0.964	1.00	1.00
	WT	0.890	0.841	0.837	0.845	0.866	0.99	0.839	0.870	0.862	0.853	0.857	0.99	0.99
S4	CM	0.879	0.835	0.821	0.849	0.757	0.99	0.864	0.864	0.859	0.863	0.845	0.99	0.99
	C/W	0.988	0.993	0.982	1.005	**0.874**	1.00	1.029	0.994	0.996	1.012	0.986	1.00	1.00
	WT	0.891	0.850	0.837	0.846	0.877	0.99	0.839	0.869	0.862	0.856	0.865	0.99	0.99
S5	CM	0.869	0.830	0.823	0.842	0.756	0.99	0.862	0.862	0.857	0.861	0.845	0.99	0.99
	C/W	0.976	0.977	0.983	0.995	**0.862**	1.00	1.027	0.991	0.994	1.006	0.976	1.00	1.00

Table 4.3: Income rates in the yeast cell cycle GRN model.

State	Cells	CLN3	WHI5	SWI4	MBP1	CLB2	YOX1	YHP1	HCM1	FKH2	NDD1	SWI5	ACE2	SIC1
	WT	4.127	2.248	5.309	0.763	1.278	2.006	0.914	0.995	0.884	1.015	1.783	1.006	1.006
S1	CM	4.127	2.248	5.238	0.793	1.217	2.006	0.914	0.995	0.914	1.036	1.702	1.006	1.006
	C/W	1.000	1.000	0.987	1.040	**0.953**	1.000	1.000	1.000	1.034	1.020	0.955	1.000	1.000
	WT	4.187	2.339	5.521	0.763	1.430	2.006	0.995	1.036	0.854	0.995	1.945	1.006	1.006
S2	CM	4.187	2.309	5.521	0.793	1.187	2.006	0.995	1.036	0.854	1.056	1.743	1.006	1.006
	C/W	1.000	0.987	1.000	1.040	**0.830**	1.000	1.000	1.000	1.000	1.061	0.896	1.000	1.000
	WT	4.187	2.339	5.379	0.763	1.551	2.006	0.995	1.015	0.884	0.955	2.187	1.006	1.006
S3	CM	4.187	2.339	5.379	0.793	1.127	2.006	1.036	1.036	0.884	1.096	1.824	1.006	1.006
	C/W	1.000	1.000	1.000	1.040	**0.726**	1.000	1.041	1.020	1.000	1.148	0.834	1.000	1.000
	WT	4.187	2.339	5.450	0.763	1.490	2.006	0.975	1.036	0.854	0.955	2.106	1.006	1.006
S4	CM	4.187	2.309	5.379	0.793	1.157	2.006	1.036	1.036	0.854	1.076	1.864	1.006	1.006
	C/W	1.000	0.987	0.987	1.040	**0.776**	1.000	1.062	1.000	1.000	1.127	0.885	1.000	1.000
	WT	4.187	2.369	5.450	0.763	1.521	2.006	0.975	1.036	0.854	0.955	2.147	1.006	1.006
S5	CM	4.127	2.278	5.379	0.793	1.157	2.006	1.036	1.036	0.854	1.076	1.864	1.006	1.006
	C/W	0.986	0.962	0.987	1.040	**0.761**	1.000	1.062	1.000	1.000	1.127	0.868	1.000	1.000

shows that only the input rates of CLB2 decrease in the CM group.

4.5 Conclusions

In this chapter, two model parameters of G-networks, the steady-state probability, q_i, and the total input rate, λ_i^+, are estimated by determining the boundary of λ_i^+ and using the grid search. Simulated gene expression data consisted of two groups (normal and case) are examined for the evaluation of the proposed method. The two groups of the data are exactly the same except that the transcription initiation rate of the case group. We observed that the G-network based modeling is able to detect the abnormally expressed genes, while the t-test possibly produces false positives. In the real data analysis, we observed that the steady-state probability of

CLB2 is lower than that of the other agents, which might suggest that the key genes of cell cycle regulation can be expressed without the clb cyclines as the original experiment study concluded.

However, our real data analysis only suggests the same information which is shown by the data 4.4. In order to improve the availability of our proposed model, it is necessary to find which interaction is crucial to the system robustness via sensitivity analysis. In terms of data reliability, the unchanged steady-state probabilities in all the five states may be a problematic because the cell cycle has four phases (G1, S, G2, M) and expressions of genes involved with a specific phase are expected to be different from those in other phases. Our model does not currently include simultaneous interactions of three or more nodes. However, this is not a limitation of the model, since it suffices to include chain representations of dependencies in the G-Network model as has been done for neuronal networks [53] to cover excitatory and inhibitory effects that involve three or more nodes, and in fact random chains of nodes of any length. Although, in this study, the transition probabilities that gene i affects to gene j, p_{ij}^{+} and p_{ij}^{-} in Equation (4.1), are fixed at the value one, we expect that the conventional reverse engineering GRN methods using the "Ensemble" method [71] can provide these probabilities more accurately for an improved steady-state analysis of GRNs.

5 G-network based Abnormal Pathway Detection

In the previous chapter, we introduced the abnormal gene expression detection technique based on G-networks. However, the previous approach fixed its customer transition probabilities, which possibly causes deterministic and unrealistic results. In this chapter, we will focus on to estimate the transition probabilities and compare them to their initial assumptions so that we can identify the abnormal pathway activities between two different conditions. The proposed method will be assessed by simulation data and the mRNA expression data of tumor patients. Our results suggest that the proposed method can provide a different conclusion of conventional gene expression analyses finding differentially expressed genes.

5.1 Introduction

A fundamental problem of biology is to understand complex gene regulatory networks (GRNs). Thus various mathematical and statistical models have been introduced for inference of GRNs [42, 131, 80, 90, 63]. Along with these reverse engineering approaches, the development of microarray techniques has accelerated the discovery of regulatory patterns among genes, leading to the construction of GRNs from data. Based on such networks, over-represented biological processes or pathways of a group of genes are identified by mapping them onto the gene ontology (GO) terms or on regulatory structures [6, 125, 114]. These pathway analyses provide the annotations and functional insight of the group of genes which are usually determined by conventional statistical tests such as the t-test and ANOVA.

However, these differentially expressed gene (DEG) derived analyses are limited in detecting defective pathways since they only observe the amount of expression of a gene itself rather than considering the flows of expression signals that communicate with neighboring genes. For example, in an apoptosis modeling study,

Neumann *et al.* suggested that the cell death pathways that depend on caspase-8 and survival pathways triggered by NF-κB are activated in parallel, and that the fate of the cell is determined by the delicate signal balance between c-FLIP$_{\mathrm{L}}$ and the initiator caspase [85]. Therefore, the flow of a gene's expression signal may be more important than the DEG detection in determining or predicting the fate of a cell using pre-apoptotic expression data. Moreover, this observation tells us that for a reliable prediction, it is necessary to find the steady-state of a system where various signals are being activated and where a subtle change of key molecules determines the status of the system.

Thus this study aims to detect the abnormal pathways rather than DEGs in GRNs by modelling them using G-Networks [47, 48]. The parameters of the G-network model are first deduced from gene expression signals of normal samples with the assumption of uniform transition probabilities. Then the transition probabilities in abnormal samples were estimated so that they minimize the difference between the observed and the predicted steady-state probabilities with constraints. Then permutation tests were performed to determine the statistical robustness of the estimated transition probabilities. We evaluated this approach using a simulated example based on stochastic gene expression models [68] and a real example where the p53 network is represented by 30 genes from the KEGG database [66], whose gene expression data is obtained from the NCBI Gene Expression Omnibus (GEO) database [4]. The proposed method revealed the abnormal pathways that are activated in cancer status, which corresponds to published evidences but cannot be found with conventional DEG approaches.

5.2 Abnormal Edge Detection

The same way as the previous chapter, our model is based on G-networks where each gene is represented by a node and customers are containing mRNA gene expression information. The detailed G-network theory is introduced in Section 2.4.3. The packets in the G-network represent latent objects containing the gene expression signal, and we assume that the number of packets is proportional to the mRNA expression levels which are actually observable data. We also assume that the mRNA levels that are measured are the observations of the steady-state. A simplified model where $p_{ijl} = 0$ of the G-network model [50] is used so that only first order effects between genes are included, rather than second order effects that can also be considered by the theory.

If we denote the *average mRNA level* of the ith gene by $\bar{x}_i$, then from the product form solution of the G-network model we know that $\bar{x}_i = q_i/(1-q_i)$ from Equation (2.37). Therefore the steady-state probability that there is at least one mRNA of the ith gene is

$$q_i = \frac{\bar{x}_i}{\bar{x}_i + 1} \tag{5.1}$$

The key task of our study is to identify the transition probabilities, p^+_{ji} and p^-_{ji} (Equation (2.36)), that a customer goes from the jth node to the ith node in abnormal conditions and compare them with the corresponding initial values in normal conditions.

Estimate rate parameters in normal conditions

Our G-network model in this chapter uses the same notation as that of the previous chapter so λ^+_i and λ^-_i are the positive and negative customer input rates, respectively, and μ_i is the service rate (firing rate). To determine the G-network parameters using normal data we use Equation (2.36) where $\mathbf{p}_j = \{p^+_{ji}, p^-_{ji}, p_{jli}\}$, and λ^+_i, λ^-_i, μ_i are all unknown. Note that $p_{ijl} = 0$ in our study. In normal conditions, we need following two assumptions

- The transition probabilities, $\mathbf{p}_j$, have uniform values; All elements in $\mathbf{p}_j$ have $1/(D^{out}_j + 1)$ where D^{out}_j is the out-degrees of gene j
- The firing rate, μ_i, is proportional to the out-degree distribution of the ith gene; $\mu_i = c \cdot D^{out}_i$ where c is a scaling parameter.

Therefore, for each gene, we have two unknown parameters, λ^+_i and λ^-_i, and one unknown global parameter, c. From Equation (2.36), we explicitly denote the functional dependencies of the q_i via the functions f_i as follows:

$$q_i = f_i(\mathbf{q}_j, \mathbf{p}_j; \lambda^+_i, \lambda^-_i, c \cdot D^{out}_i), \tag{5.2}$$

where $\mathbf{q}_j = \{q_j\}$ and $j \in \{\text{indices of genes that affect to the } i\text{th gene }\}$. Now, the parameters can be chosen as follows:

- Set the initial values $\lambda^+_i = 0.0062 sec^{-1}$ and $\lambda^-_i = 0.002 sec^{-1}$ from [115].
- Choose the parameter c as the value $\hat{c}$ which minimises the squared error below with the above inital values of λ^+_i and λ^-_i, and the observed gene

expression values of Equation (5.1)

$$\sum_i \left(q_i - f_i(\mathbf{q}_j, \mathbf{p}_j, \lambda_i^+, \lambda_i^-; c \cdot D_i^{out})\right)^2 \tag{5.3}$$

- For each gene, substitute $\hat{\mu}_i = \hat{c} \cdot D_i^{out}$ and find the input rate $\hat{\lambda}_i^+$ to the value which minimises

$$\left(q_i - f_i(\mathbf{q}_j, \mathbf{p}_j, \lambda_i^-, \hat{\mu}_i; \lambda_i^+)\right)^2 \tag{5.4}$$

  with the constraint $0 \leq \hat{\lambda}_i^+ \leq \mu_i + \lambda_i^- + \Lambda_i^- - \Lambda_i^+$ which comes from the fact that $0 \leq \lambda_i^+$ and $0 \leq q_i \leq 1$.

- Find $\hat{\lambda}_i^-$ to the value which yields the same value of the corresponding q_i.

Determine transition probabilities in abnormal conditions

Now our problem is to find the transition probabilities, $\hat{\mathbf{p}}_j$, from abnormal data under the assumption that all the other parameters are the same with that of the normal data. Let q_i' be the observed steady-state probability of the ith gene in an abnormal condition, and let $\mathbf{p}_j'$ be the set of transition probabilities in the same condition. We find the probabilities by minimizing the following sum of squared error with a constraint $0 \leq \mathbf{p}_j' \leq 1$:

$$\sum_i (q_i' - f_i(\mathbf{q}_{ij}', \hat{\lambda}_i^+, \hat{\lambda}_i^-, \hat{\boldsymbol{\mu}}_{ij}; \mathbf{p}_j'))^2 \tag{5.5}$$

In all the optimization steps to minimize Equation (5.5), we use the Barzilai-Borwein spectral methods in BB package of R [120] which is closely related to the well known Conjugate-Gradient method. However since such methods yield local minima, we proceed via an iterative search:

1. We first generate H solutions $\{\mathbf{p}_j'^{(1)}, \ \ldots \ , \mathbf{p}_j'^{(H)}\}$ starting with $H = 2000$ randomly selected initial values ranging from 0 to 1.

2. Among these we select a subset of size $0.05H$ which yields the smallest error (Equation (5.5)).

3. We draw H random samples again but now with values in the range of the values of the previous best 5%.

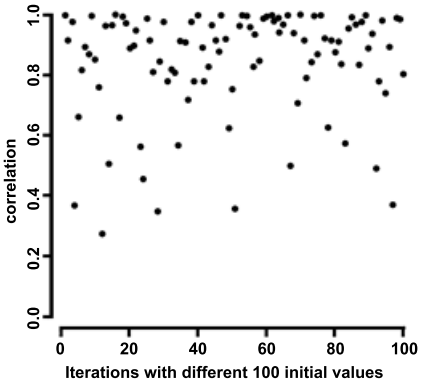


Figure 5.1: Correlation between the abnormal transition probabilities obtained via the randomized optimization procedure and the purely random search

4. Go to the 2nd step and repeat these iterations until the sum of all differences between the estimated probabilities and its previous value does not bigger than 0.01.

This approach has also been compared with a direct random search for the local minimum, and we observe the strong correlation between results obtained by both approaches shown in Figure 5.1.

Permutation test of the transition probabilities

When the estimated $\hat{p}'_{ij}$ differs from its normal value p_{ij}, it is necessary to test whether the difference is statistically significant or not. The null hypothesis of this test could be $\hat{p}'_{ji} \leq p_{ji}$ if $\hat{p}'_{ji} > p_{ji}$ (or $\hat{p}'_{ji} \geq p_{ji}$ if $\hat{p}'_{ji} < p_{ji}$). To perform the test, its samples were shuffled at random and divided into normal and abnormal groups with the same sizes of the original groups. Then the proposed method is applied in the same way as the original data. Let M be the number of permutations and $p_{ij}^{(m)}$ be the estimated transition probability of the mth permutation. Then the empirical p-value of $\hat{p}'_{ij}$ can be obtained as follows,

$$p\text{-value of } \hat{p}'_{ij} = \begin{cases} \frac{1}{M}\sum_{m=1}^{M} I(\hat{p}'_{ij} \leq p_{ij}^{(m)}) & \text{if } p'_{ij} > p_{ij} \\ \frac{1}{M}\sum_{m=1}^{M} I(\hat{p}'_{ij} \geq p_{ij}^{(m)}) & \text{if } p'_{ij} < p_{ij} \end{cases}$$

where $I(C)$ is the indicator function. Thus if the p-value is less than a criterion α then the null hypothesis is rejected. We used $\alpha = 0.001$.

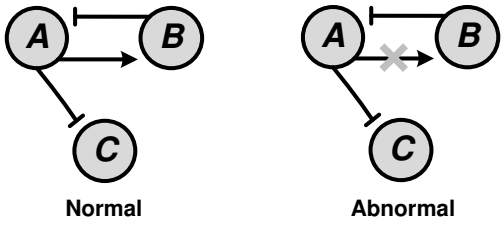


Figure 5.2: The simulated gene regulatory networks consisting of three genes

5.3 Evaluation with Simulated Networks

In order to evaluate our proposed method, we built an artificial network consisting of three genes. The simulated mRNA expression data were obtained by the stochastic gene expression modelling [68] where most of the reaction rate parameters were taken from experimental studies. Let A, B, and C be the three genes whose interactions are i) A activates B, ii) B inhibits A, and iii) A inhibits C (Figure 5.2). The normal data are simulated for $60,000$ sec with normal reaction rate parameters while the abnormal data are obtained with the same reaction rates as the normal data except that the transcription rate of B by A is one third of the normal. Each gene has following reaction equations,

$$DNA_i + G_i \xrightarrow{\lambda_1} DNA_i + G_i \tag{5.6}$$

$$G_i \xrightarrow{\lambda_2} G_i + G_j \tag{5.7}$$

$$G_i + G_j \xrightarrow{\lambda_2} G_i \tag{5.8}$$

$$G_i \xrightarrow{\lambda_3} \phi \tag{5.9}$$

where $i, j \in \{A, B, C\}$. Equation (5.6) to (5.9) represent the initiation, activation, inhibition and degradation processes, respectively (DNA_i: DNA, G_i: mRNA/protein molecules, $\lambda_1 = 1$, $\lambda_2 = 0.25$, $\lambda_3 = 0.3$). Gillespie algorithm was used to simulate these reaction equations [54]. For a more detailed description of the stochastic gene expression model, see Section 2.2. Figure 5.3 shows the protein expression profiles of the three genes. Though the profiles are protein expressions for the clear representation, our model uses mRNA expression levels as the observed data.

In each group (normal and abnormal), 10 samples were randomly chosen and t-test was performed to evaluate the significance of their mean difference. Table 5.1

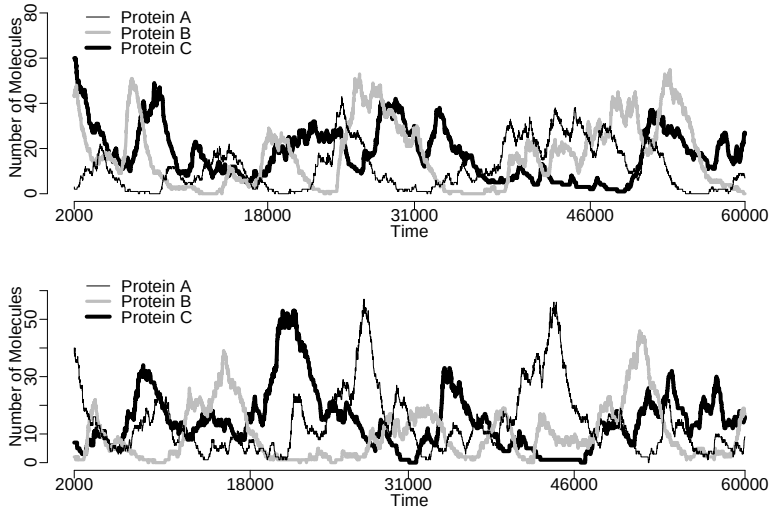


Figure 5.3: The protein expression profiles of the simulated network. The expressions of normal genes (upper) and abnormal genes (lower). The abnormal transcription rate of B is set as one third of the normal rate, which causes the up regulation of A and down regulation of C.

Table 5.1: Average expression levels of the randomly chosen 10 samples and their p-values of t-tests.

Genes	Normal Avg. Exp.	Abnormal Avg. Exp.	p-values of t-tests
A	2.5	3.4	0.118
B	18.9	15.7	0.014
C	13.9	10.6	0.005

shows the average expression levels and p-values of the t-tests. In the abnormal condition, the expression of B is lower than the normal level as the transcription of B is meant to be interrupted. Thus A is over-expressed and C's expression is lowered as the inhibition effect of A decreases. From the p-values of Table 5.1, we could conclude that B and C cause the abnormality of the system but our approach (Table 5.2) can draw a different conclusion that the pathways between A and B are significantly inactivated (p^+_{AB} and p^+_{BA} decreased) while A still induces the expression of C (p^-_{AC} is not significant), which shows better agreement with the simulated conditions than the p-values of the t-test.

Table 5.2: Assumed transition probabilities in the normal condition and their estimations in the abnormal condition. p-values are the empirical p-values of the permutation test with 1000 iterations.

Pathways	Normal Trans. Prob.	Abnormal Trans. Prob.	Empirical p-values
p^{+}_{AB} (A $\rightarrow$ B)	0.2	0.162	0.004
p^{-}_{AC} (A $\rightarrow$ C)	0.2	0.140	0.394
p^{-}_{BA} (B $\rightarrow$ A)	0.25	< 0.001	0.006

5.4 Application with p53 Networks

In order to assess our approach using real experimental data, we built a p53 network which is a well studied system in human cells and whose most important feature is tumor suppression when DNA is damaged. The regulatory structure of the p53 pathway with 28 genes was constructed on the basis of the KEGG database [66] (Figure 5.4), and we used two microarray mRNA expression datasets from GEO [4]. The first dataset (GSE12941) consists of 10 non-tumor liver tissue and 10 hepatocellular carcinoma (HCC) samples. The second (GSE6222) is a dataset for the study of liver cancer progression in HCC. In this dataset, there are 2 normal and 10 HCC samples. Before applying the proposed method, the data were normalized and scaled with mean 3 and variance 1 because the average number of mRNAs of a gene in a single cell is assumed to be approximately 3 (The mRNA transcription and degradation rates are assumed to be $0.0062sec^{-1}$ and $0.002sec^{-1}$ from [115] so that the average is $0.0062/0.002 \approx 3$).

Figure 5.5 shows the average expression levels of genes in each dataset, and their p-values of t-tests which can detect DEGs. Since the interest of this study is the only 28 genes, multiple testing corrections were not applied to these DEG tests. The two datasets share 9 significant DEGs with 0.05 significance level while GSE12941 has 4 more DEGs. This similarity can be confirmed by observing their expression patterns (Figure 5.5 top and middle). However, interpreting the DEGs even when we know their regulatory structure is still challenging. Table 5.3 shows the packet transition probabilities estimated from the proposed method and their empirical p-values by the permutation test. These results along with the p-values of the t-test for DEGs are summarized together in Figure 5.6 and Figure 5.7 (see the caption of Figure 5.6 for their detailed description).

From the analysis, We drew following four conclusions:

- Under non-stressed conditions, p53 stays at low levels since it is degraded

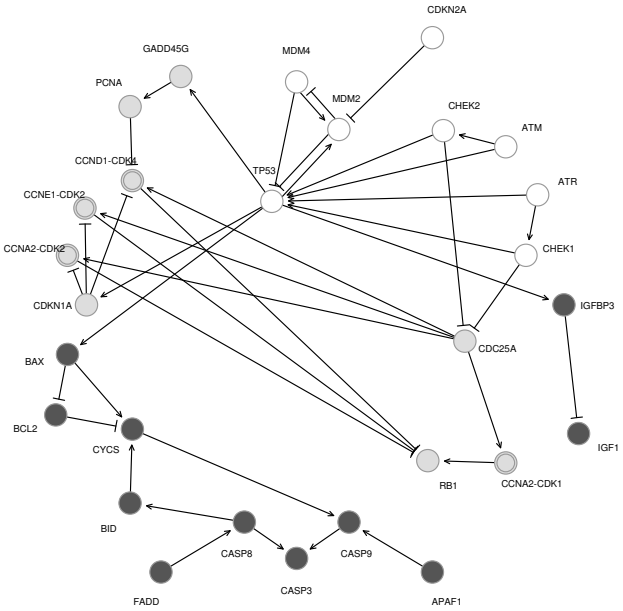


Figure 5.4: The p53 network consisting of 28 genes. Gray, white, and black nodes indicate p53 upstream, cell cycle, and apoptosis related genes, respectively. Double lined circles depict cyclin complex

rapidly by MDM2/MDM4 while the interactions between p53 and MDM2/MDM4 are blocked by ATM/ATR in stressed conditions [61, 19]. In our results, though the inhibition effect from MDM2/MDM4 to p53 decreases as expected, pathways between ATM/ATR and p53 are not significantly activated so p53 seems not working properly for its function of tumor suppression.

- ATM/ATR-CHEK1/CHEK2 pathways are well known to be activated when the DNA is damaged [19]. Despite the apparent lack of significance of ATM and ATR in their t-tests, their pathway activities through CHEK1/CHEK2, CDC25A, CCNA2-CDK1 to RB1 were clearly revealed in both datasets. RB1, as a tumor suppressor gene, prevents cells from replicating damaged DNA by suppressing the cell cycle [123]. However, the activation effect from CCNA2-CDK1 to RB1 decreases, which might be a reason for the tumor development.

- One of significant pathways in both datasets is TP53-IGFBP3-IGF1 [102]. IGF1 can cause many cellular effects including cell proliferation and anti-apoptotic processes [93]. This gene expression is inhibited by IGFBP3

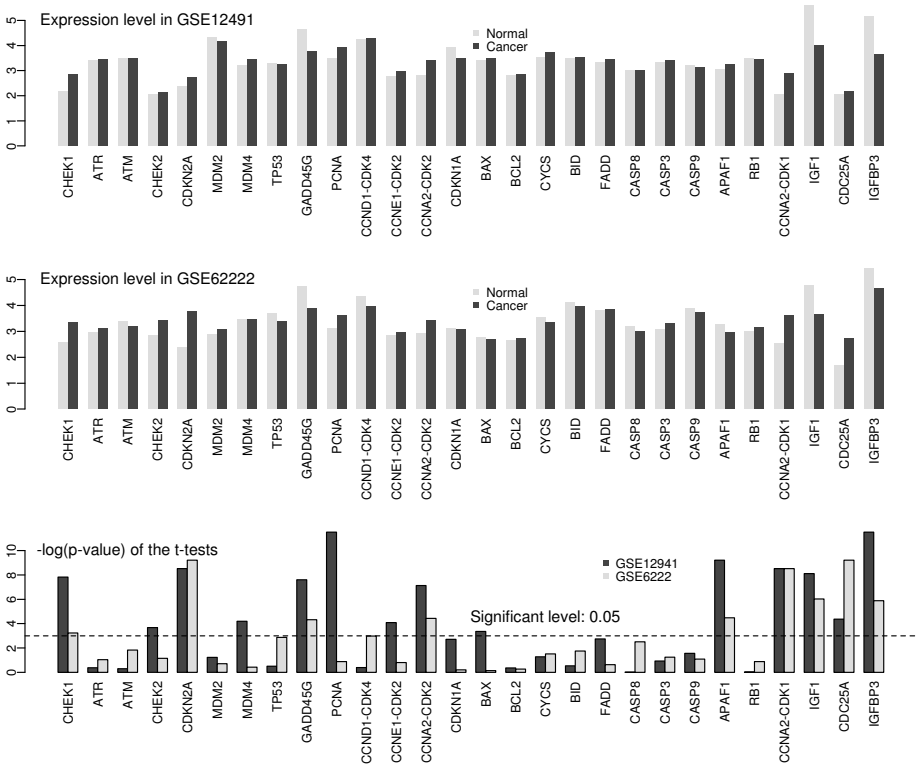


Figure 5.5: Average mRNA expression levels of genes in two datasets, GSE12941 (top) and GSE6222 (middle). The p-values of the t-tests are depicted in the bottom panel where y axis represents negative natural logarithms of the p-values. Note that multiple testing corrections are not applied in these t-tests.

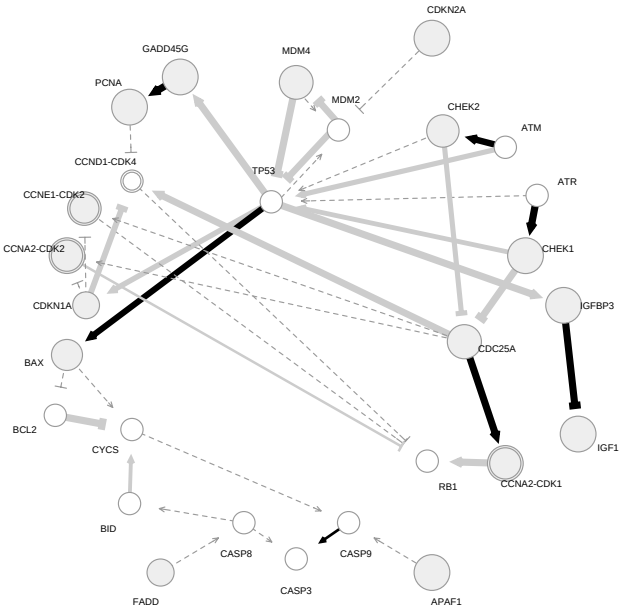


Figure 5.6: The p53 network with the results of GSE12941 dataset analysis. The solid line represents significantly activated (black) or inactivated (grey) pathways while the dashed line indicates non-significant pathways. Wider lines represent more significant pathways. The grey nodes are the significant DEGs from the t-test. The radius of a node is larger if its DEG is more significant with a 0.05 significant level. White nodes indicate non-significant genes.

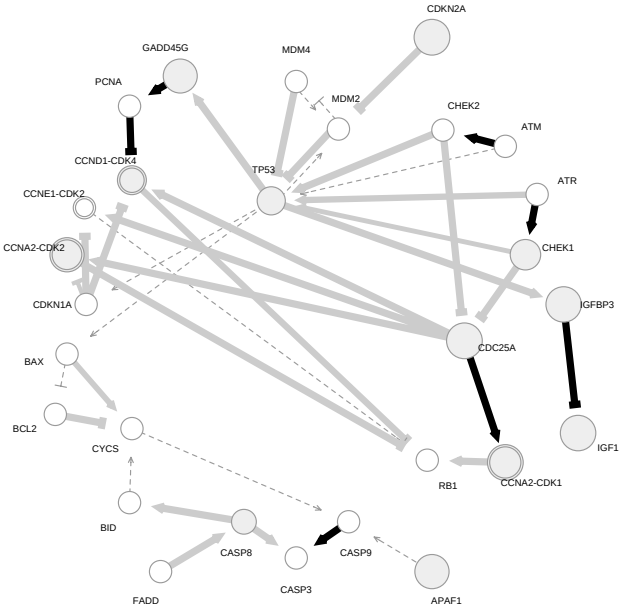


Figure 5.7: The p53 network with the results of GSE6222 dataset analysis. Please see the legend of Figure 5.6 for graphical interpretation. Though some pathway activations are different from Figure 5.6, ATR-CHEK1-TP53, ATM-CHEK2-TP53, TP53-IGFBP3-IGF1, CHEK2-CDC25A-CCNA2-CDC1, and caspase pathways show similar pathway activation.

which is also reported that its expression induces apoptosis in an independent manner of IGF1 [102]. That is, if IGFBP is up-regulated by TP53, cells go to apoptosis. However, their t-tests show that they are all significantly down-regulated and our results also indicate that their pathways are not activated.

- In both datasets, caspase (CASP3, CASP8, and CASP9) mediated pathways which are well known to trigger cell apoptosis are not activated or significantly inactivated in tumor samples. These results seem to show that our method properly reflects the properties of the tumor samples.

5.5 Conclusions

We have proposed a new approach for detecting abnormal pathways in GRNs based on G-network modeling which has a computationally tractable steady-state

Table 5.3: Assumed transition probabilities in the normal condition and their estimations from the cancer samples. p-values represent the empirical p-values of the permutation test.

		GSE6222		GSE12941	
Pathways	Normal	Cancer	p-value	Cancer	p-value
ATR→CHEK1	0.333	0.650	0.014	1.000	0.000
ATM→CHEK2	0.333	0.578	0.028	0.519	0.011
CDKN2A→MDM2	0.500	0.361	0.170	0.562	0.115
MDM4→MDM2	0.333	0.276	0.358	0.276	0.308
TP53→MDM2	0.167	0.277	0.024	0.199	0.186
MDM2→MDM4	0.333	0.329	0.428	0.249	0.018
CHEK1→TP53	0.333	0.428	0.171	0.528	0.064
ATR→TP53	0.333	0.114	0.000	0.000	0.000
ATM→TP53	0.333	0.141	0.025	0.256	0.084
CHEK2→TP53	0.333	0.450	0.115	0.466	0.162
MDM2→TP53	0.333	0.452	0.111	0.412	0.129
MDM4→TP53	0.333	0.559	0.039	0.370	0.364
TP53→GADD45G	0.167	0.118	0.049	0.102	0.001
GADD45G→PCNA	0.500	0.766	0.029	0.708	0.000
PCNA→CCND1-CDK4	0.500	0.714	0.014	0.263	0.006
CDKN1A→CCND1-CDK4	0.250	0.347	0.064	0.239	0.448
CDC25A→CCND1-CDK4	0.200	0.187	0.296	0.000	0.000
CDKN1A→CCNE1-CDK2	0.250	0.353	0.122	0.000	0.000
CDC25A→CCNE1-CDK2	0.200	0.145	0.120	0.000	0.000
CDKN1A→CCNA2-CDK2	0.250	0.213	0.332	0.000	0.000
CDC25A→CCNA2-CDK2	0.200	0.156	0.186	0.000	0.000
TP53→CDKN1A	0.167	0.147	0.563	0.078	0.027
TP53→BAX	0.167	0.155	0.445	0.185	0.011
BAX→BCL2	0.333	0.343	0.595	0.331	0.473
BAX→CYCS	0.333	0.293	0.373	0.410	0.121
BCL2→CYCS	0.500	0.620	0.066	0.514	0.445
BID→CYCS	0.500	0.490	0.511	0.484	0.439
CASP8→BID	0.333	0.320	0.259	0.343	0.343
FADD→CASP8	0.500	0.379	0.156	0.471	0.358
CASP8→CASP3	0.333	0.396	0.164	0.387	0.102
CASP9→CASP3	0.500	0.547	0.320	0.473	0.336
CYCS→CASP9	0.500	0.531	0.342	0.381	0.052
APAF1→CASP9	0.500	0.435	0.193	0.485	0.389
CCND1-CDK4→RB1	0.500	0.410	0.119	0.593	0.111
CCNE1-CDK2→RB1	0.500	0.435	0.223	0.545	0.247
CCNA2-CDK2→RB1	0.500	0.534	0.348	0.540	0.263
CCNA2-CDK1→RB1	0.500	0.435	0.197	0.458	0.325
CDC25A→CCNA2-CDK1	0.200	0.303	0.140	1.000	0.000
IGFBP3→IGF1	0.500	0.712	0.018	0.816	0.000
CHEK1→CDC25A	0.333	0.000	0.000	0.000	0.000
CHEK2→CDC25A	0.333	0.000	0.000	0.000	0.000
TP53→IGFBP3	0.167	0.144	0.096	0.057	0.000

solution and therefore does not require a large number of samples from time-dependent data. Moreover this analytic solution allows the extension of the approach to very large-scale GRN systems which will be shown in Chapter 10. Our method provides an effective way to describe the flows of gene expression signals including negative or inhibitory effects. In the p53 network example with mRNA expression data, we showed that the interpretations of the abnormal path-

ways somewhat agree with the biological conditions of the cancer samples.

6 G-network based Two Layer Stochastic GRN Modeling

In Chapter 4 and 5, we have introduced the GRN model based on G-network theory. As Chapter 3 shows, the gene expression process is a sum of many complicated mRNA and protein interactions but the G-network models oversimplified the sophisticated gene expression processes by assuming that customers contain all the gene expression information. So, in this chapter, an extended version of the G-network modeling will be introduced by dividing the simplified G-network model into two layers, DNA and protein layers. This extended GRN model shows a possibility that the G-network approach can describe more detailed and complicated GRN processes.

6.1 Introduction

One of the fundamental problems of systems biology is to understand a complex biological system such as a gene regulatory network (GRN). Various mathematical and statistical models have been introduced for the inference of the GRNs [42, 80, 90, 63]. Along with these reverse engineering approaches, the development of bioengineering techniques accelerates to uncover the regulatory mechanisms between genes for the GRN construction. Once a GRN is built, it is essential to monitor its cellular behaviour in response to external stresses or environmental changes. Especially, studying the steady-state effect of the external perturbation to the GRNs allows us to predict the long-term behaviours of a biological system [106].

Recently, probabilistic Boolean Networks (PBNs) are proposed to investigate the steady-state of GRNs [105], but their model is not detailed enough to explain the delicate cellular processes and, in spite of the simplified (binarized) signal, they still suffer from their heavy computing time [29]. A new approach to the steady-state analysis of GRNs based on G-Network theory [47, 48] is proposed in

[50]. The G-Network is a probabilistic queuing network having special customers which include positive and negative customers, signals and triggers [47, 48]. In contrast to a normal queueing theory, G-networks introduced negative customers which can describe the inhibitory effect in GRNs.

In our previous work [67] which is also based on G-networks, GRNs are analyzed to monitor their steady-state behaviours and to detect anomaly expressed genes/pathways responding to external perturbations. However, these models cannot account for the effects of post-translational processes such as phosphorylation and ubiquitination which play an important role in gene regulation. Mitogen-activated protein (MAP) kinase pathway in mammalian systems is an example of non-transcriptional and cascade signaling pathways [79]. F. Markowetz *et. al.* proposed a nested effect model (NEM) that can infer these non-transcriptional structure of innate immune response (IIR) in Drosophila melanogaster [14, 78, 118].

Our proposed model consists of two layers one of which is a DNA (mRNA) layer and the other is a protein layer. The mRNA expression on the DNA layer is described by a single queueing system and the protein layer is modeled by G-networks. The translation process and transcription factor activities are responsible for the transmission of gene regulatory information between the two layers. We will show a simple numerical example to help the understanding of the proposed model. Also two different regulatory network models responding to external signals are compared with each other in terms of their long-run behaviours.

6.2 Two Layer Gene Regulatory Network Model

6.2.1 Biological assumptions for gene regulation

The gene regulation is largely involved with the activities of various protein molecules such as transcription factors, repressors, and activators. Proteins act alone or together with other proteins to promote or block the expression of genes. Based on these protein activities, our study focuses on seven gene regulation related mechanisms; Transcription, Translation, Phosphorylation, Ubiquitin dependent protein degradation, Protease dependent protein degradation, mRNA degradation, and Protein-protein interaction (Figure 2.1). Please see Section 2.1 for the details of these processes.

6.2.2 DNA layer modeling

Let's imagine a cell where a gene produces its mRNAs. Then, the cell, as a queue, could be considered as a $M/M/\infty$ system of the queueing theory where its input and output processes are Poisson (M) and ∞ indicates infinite number of servers. The number of mRNAs in the cell is affected by a input (transcription) rate (λ_i) which depends on the level of its transcription factors. The initial transcription rate of a gene with its transcription factors is assumed to be a constant (e.g. $\lambda_i = 0.0067sec^{-1}$, one mRNA per 2.5min [115, 58]). The degradation rate of an mRNA is also fixed as μ_i which is obtained on the basis of the half-life of an mRNA molecule. However, the total amount of degrade mRNAs depends on the current number of mRNA, n. So the average number of mRNA molecules can be computed by modifying birth-death process whose input and output rates are dependent on the current state of the system [5].

Let p_{i_n} be the probability that a cell has n molecules of the ith mRNA and $\mu_{i_n} = n\mu_i$ where $1 \leq n \leq \infty$. Then the following balance equations are obtained.

$$\begin{aligned} p_{i_1} &= \frac{\lambda_{i_1}}{\mu_{i_1}} p_{i_0} = \frac{\lambda_i}{\mu_i} p_{i_0} \\ p_{i_2} &= \frac{\lambda_{i_2}}{\mu_{i_2}} p_{i_1} = \frac{\lambda_i^2}{2\mu_i^2} p_{i_0} \\ &\ldots \\ p_{i_n} &= \frac{\lambda_i}{\mu_{i_n}} p_{i_{n-1}} = \frac{1}{n!} \left(\frac{\lambda_i}{\mu_i}\right)^n p_{i_0} \end{aligned} \tag{6.1}$$

Let $\frac{\lambda_i}{\mu_i} = \rho_i, 0 < \rho_i < 1$. Since the sum of all probabilities is one, p_{i_0} is

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} \rho_i^n p_{i_0} &= 1 \\ p_{i_0} &= e^{-\rho_i} \end{aligned} \tag{6.2}$$

Therefore, the steay-state probability that there are n mRNAs in a cell is given by as follows,

$$p_{i_n} = \frac{\rho_i^n}{n!} e^{-\rho_i} \tag{6.3}$$

Then the average number of mRNAs (LR_i) in a cell is

$$\begin{aligned} LR_i &= \sum_{n=0}^{\infty} np_{i_n} \\ &= \sum_{n=0}^{\infty} n\frac{\rho_i^n}{n!}e^{-\rho_i} \\ &= \rho_i \end{aligned} \tag{6.4}$$

6.2.3 Protein layer modeling

Though the number of mRNAs and proteins can be directly/indirectly affected by each other, it is more realistic to model mRNA and protein systems separately. It is because an mRNA only provides itself as a template of corresponding protein production and many modeling approaches assume that the degradation of the mRNA naturally depends on the population size of the mRNA. Therefore, apart from the mRNA system modeling, protein expression is independently described by G-network theory [47, 48]. A G-Network is a stochastic model having positive and negative agents and signals. Note that the agents and signals represent proteins in this protein layer. Our model is specially focusing on the number of "activated" proteins rather than protein monomers which is the primary product of translation. Monomer proteins could be activated by multimorization or phosphorylation.

(P-I) Activated protein input by translation process

In one mRNA molecule, the probability that the number of proteins (n_3) produced per one burst is $p^{n_3}(1-p)$ where p and $(1-p)$ are the probabilities that a ribosome and an RNase-E bind to the mRNA, respectively. Let's assume the burst size is equal to the average number of proteins per one mRNA then, $b = p/(1-p)$. Note that b is varied; 100 for an average E. coli gene [94], 40 for lacZ, and 5 for lacA [115]. Let T_D be the average time interval between successive competitions of ribosome and RNase-E. Then the surviving mRNAs, n_2, in the population after its transcription is blocked would be $n_2 = n_{2_0}p^{T_{half}/T_D}$ which is equal to $T_{half} = -(\log(2)/\log(p))T_D$ [81].

$$T_D = -\frac{\log(b) - \log(1+b)}{\log(2)}T_{half} \tag{6.5}$$

Therefore, if we know the average half-life of mRNAs and a burst size, the translation initiation rate for a protein monomer, $\lambda_{mono,i} = L_{R,i}/T_D$, can

be determined. The geometric distribution of the above protein expression is the discrete version of gamma distribution (sum of exponential distribution) in [95, 25] which include the effect of protein delusion by cell cycle. So our assumed protein expression can be considered a gamma distribution with its average frequency of expression burst per cell cycle is set to one. We assume the burst size, b, is 7.8 and the half-life of mRNA, T_{half}, is 300 sec.

Some proteins are activated in a form of a dimer or an oligomer rather in monomer. It is known that the protein multimerization is an important feature in regulations of bio-molecules such as enzymes or transcription factors [77]. The dimer dissociation constant K_D is varied from 0.1 to 10 where $K_D = [P]^2/[PP]$; $[P]$ and $[PP]$ are the monomer and dimer protein concentrations, respectively [20]. In this study, we assume the input rate of a activated protein is

$$\lambda_i^+ = 10 \cdot \lambda_{mono,i}^2 \tag{6.6}$$

(P-II) Activated protein input by phosphorylation

Consider protein A which is activated by a phosphorylation process where kinase B is responsible for the phosphorylation. Then, the reaction equation will be $A_p + B \xrightarrow{\lambda_{p,i}} A + B_p$ where A_p and B_p are phosphorylated A and B, respectively, and $\lambda_{p,i}$ is the phosphorylation rate. However, our protein expression model only concerns about the activated proteins so the reaction can be re-written as $A_p \xrightarrow{\lambda_{p,i}} B_p$. That is, when this reaction occurs, A_p and B_p decreases and increases by 1, respectively. We assume $\lambda_i^+ = 0.0023$ which is a protein-protein association rate in [21]

(P-III) Activated protein output by ubiquitination

As we mentioned in Section 2.1, an ubiquitinaton target protein is recognized by E3 enzymes which is bound with E2 enzymes together. This process can be expressed by $A_u + B \xrightarrow{\lambda_i^-} A + B_u$ which is similar with the case of phosphorylation. Then the reduced form of this equation will be $A_u \xrightarrow{\mu_{p,i}} B_u$. The negative customers of G-networks are used to model this active degradation process so λ_i^- is a negative customer input rate. We use $\lambda_i^- = 0.0023$ the same as the phosphorylation rate.

(P-IV) Activated protein output by protease dependent degradation

The 1st order reaction equation is defined by $(d[A])/dt = k[A]$ where k is

reaction constant and $[A]$ is the concentration of protein A. Then the decrement of the concentration of A over time is $\log([A]/[A]_0) = -kt$ where $[A]_0$ is the initial concentration of A at time $t = 0$. Therefore, if we know the half-life of a protein, the reaction (degradation) rate can be obtained by $k = \log(2)/T_{half}$. The average of half-life of proteins is known as from 15min to 120min [115]. We assume the half-life of activated proteins is 30min so the activated protein output rate by protease dependent degradation is $k = 0.000385$.

(P-V) Protein-protein interaction

In our model, two types of protein reaction channels are considered [45]. (i) Protein P_i turns into protein P_j, whose reaction equation is $P_i \rightarrow P_j$ which is the same form as that of phosphorylation and ubiquitination processes. (ii) Protein complex P_l is synthesized by binding two proteins P_i and P_j together. In this case, the reaction equation is $P_i + P_j \rightarrow P_l$. Note that the case involving more than two proteins can be modeled by considering the reactants, P_i and P_j, are protein complexes which are previously formed by the same reaction.

Based on the these mathematical assumptions, we can build the balance equation of the protein expression which is the same form of G-networks [50]. Let's consider a random process $\mathbf{X}(t) = [X_1(t), ..., X_n(t)]$ where $X_i(t)$ be an integer-valued random variable representing the number of ith protein molecules at time t ($t \geq 0$) and an n-vector of non-negative integers $\mathbf{x} = [x_1, ..., x_n]$. If $Pr(\mathbf{x}, t)$ is the probability that $\mathbf{X}(t)$ takes $\mathbf{x}$ at time t, then the probability that $\mathbf{X}(t)$ have $\mathbf{x}$ at

time $t + \Delta t$ is defined as

$$
\begin{aligned}
P(\mathbf{x}, t+\Delta t) = \sum_{i=1}^{n} \Big[& (\lambda_i^+ \Delta t + o(\Delta t)) Pr(\mathbf{x}_i^-, t) I(\mathbf{x}_i > 0) \\
& + (\lambda_i^- \Delta t + o(\Delta t)) Pr(\mathbf{x}_i^+, t) \\
& + \sum_{j=1}^{n} \Big\{ (p_{ij}^+ \mu_i \Delta t + o(\Delta)) Pr(\mathbf{x}_{ij}^{+-}, t) I(\mathbf{x}_j > 0) \\
& + (p_{ij}^- \mu_i \Delta t + o(\Delta)) Pr(\mathbf{x}_{ij}^{++}, t) \\
& + (p_{ij}^- \mu_i \Delta t + o(\Delta)) Pr(\mathbf{x}_i^+, t) I(\mathbf{x}_j = 0) \Big\} \\
& + (d_i \mu_i \Delta t + o(\Delta t)) Pr(\mathbf{x}_i^+, t) \\
& + (1 - (\lambda_i^+ + \lambda_i^- + \mu_i) \Delta t + o(\Delta t)) P(\mathbf{x}, t) \Big]
\end{aligned}
\tag{6.7}
$$

where $\mathbf{x}_i^+ (\mathbf{x}_i^-)$ is a vector that the value of ith element is $x_i + 1$ ($x_i - 1$); $I(C)$ is indicator function which is 1 if the condition, C, is true or 0 other wise; $o(\Delta t) \to 0$ as $\Delta t \to 0$; λ_i^+ and λ_i^- are the positive and negative input rates of the ith protein, respectively. $d_i \times r_i$ is the protein degradation rate (0.000385) where d_i is the probability that a customer (protein) leaves the system and μ_i is a server activation rate that the ith protein affects to the other protein or disappear. The first term describes the increment of the activated proteins by its translation and the second term explains the active degradation by ubiquitination. p_{ij}^+ (p_{ij}^-) is the probability that protein i phosphorylates (ubiquitinates) protein j. So the third term is the probability that protein i phosphorylates protein j in time Δt. The fourth and fifth terms describe the ubiquitination process where protein i ubiquitinases protein j with probability p_{ij}^- so both of the i and j are depleted by 1 if this event is happened. The sixth term is the probability that the protein i is degraded by the protease dependent (non-ubiquinization) protein degradation (nothing is happened except that protein i lose one molecule). The last term represents nothing happened. The solution of this balance equation is shown in Equation (2.36) and the steady-state probability that there are x_i molecules of activated protein i in a cell is

$$
\begin{aligned}
\lim_{t \to \infty} p(X_i = x_i, t) &= p(x_i) \\
&= q_i^{x_i} (1 - q_i)
\end{aligned}
\tag{6.8}
$$

Therefore, not only the protein expression from a single mRNA but also the expression considering our biological assumptions including post-translation processes can be explained by a geometric distribution. The average number of activated protein i is $LP_i = q_i/(1-q_i)$ ($0 \leq LP_i < \infty$, $0 \leq q_i < 1$). Note that the probabilities, p_{ij}^+ and p_{ij}^- are all set to $1/n_i^{out}$ where $1/n_i^{out}$ is the out degree of the ith protein.

6.2.4 Bridging mRNA and protein layers

Since the number of mRNAs depends on the concentration of its transcription factors (activated proteins), the input rate of an mRNA can be a function of its transcription factors. Consider the ith mRNA is transcribed by jth transcription factor. Then the input rate of the ith mRNA is $\lambda_i = f(LP_j)$. f can take various forms depending on the types of ith gene response. In this study, we define f as similar to the Michaelis-Menten description of enzyme kinetics,

$$f^+ = \frac{\alpha(LP_j/K)^h}{1+(LP_j/K)^h}, \quad f^- = \frac{\alpha}{1+(LP_j/K)^h} \tag{6.9}$$

where f^+ and f^- indicate activation and repression function, respectively; K is the Michaelis constant; h is a Hill coefficient; α is a scaling parameter. We assume $h = 4$ (tetramerized protein) and $K = b(= 7.8)$ (Average number of protein molecules per one transcription). The scaling parameter $\alpha = 0.0133$ is obtained under the condition of $K = LP_j$ so that the transcription rate with the effect of Equation (6.9) is to be the same as $1/T_{init}$ which is the initial transcription rate in a single gene expression model ($T_{init} = 150sec$ in this study). It is because the initial transcription rate assumes that the level of its corresponding transcription factors is sufficient enough.

The average number of protein j, LP_j, is depending on its average number of mRNA, LR_j, as described in assumption (P-I). Let $\rho_i^{(0)}$ be the initial value of ρ_i. Then we can compute $LP_i^{(0)}$ using corresponding input and output rates of ith protein. If protein j is the transcription factor of ith gene, $\rho_i^{(1)}$ can be recalculated based on the jth protein level $LP_j^{(0)}$ which is computed by the same way as $LP_i^{(0)}$. We can find the optimal average numbers (including ρ_i and q_i) of mRNAs and proteins by iterating these steps until the following join steady-state probability of

protein expression is converged.

$$p(\mathbf{x}) = \prod_{i=1}^{N} q_i^{x_i}(1 - q_i) \tag{6.10}$$

where $\mathbf{x} = \{x_1, x_2, ..., x_N\}$. That is, if $p(\mathbf{x})^{(k)}$ is the joint probability which is computed using corresponding kth parameters, then the iteration is continued until $|p(\mathbf{x})^{(k)} - p(\mathbf{x})^{(k-1)}| < c$ where c is stopping criterion.

6.3 Toy Example of Single Gene Expression

In this section, we will describe a single gene system given its parameter values except p_{ij}^{+} and p_{ij}^{-}. Then by inferring these transition probabilities, we will try to interpret the status of the system. Let's assume there is a gene A which produces mRNA and corresponding protein. The transcription rate of mRNA is $\lambda_A = 0.0067 sec^{-1}$ by assuming one mRNA is produced every 150 sec. The mRNA half-life is set to 300 sec so the degradation rate of the mRNA is $\mu_A = 0.00231 (= log(2)/300 sec)$. Therefore, the average number of mRNA in a cell is $LR_A = \rho_A = 2.9$ (Equation 6.4). The steady-state probability that at least one mRNA in a cell is $0.945 (= 1 - e^{-\rho_A})$ (Equation (6.2)).

Now, we need to compute the average number of protein A, LP_A, in a cell. However, a previous experimental study [25] implies the LP_A of our model should be the same as the burst size $b (= 7.8)$ when we assume there is no interaction with other proteins and the burst frequency per cell cycle is fixed one. So, the average number of protein A is $q_A/(1 - q_A) = b = 7.8$. In order to describe the actual processes in protein layer, we need to know the unknown parameters in the Equation (2.36) such as μ_i, p_{ji}^{+} and p_{ji}^{-}. The translation initiation rate $\lambda_A^{+} = 0.0308$ is obtained by using Equation (6.6) and the ubiquitination rate is $\lambda_A^{-} = 0.0023$ which is caused by some unknown enzymes. Then by the Equation (2.36), $\mu_A = 0.0327$. The degradation rate of this protein molecule is $d_A \times \mu_A = 0.000385$ by assuming the half-life of a protein is 1800 sec. Therefore, $d_A = 0.1177$ ($\sum_{j \in S(i)} (p_{ij}^{+} + p_{ij}^{-}) > 0$) which means this system should have some interactions with other genes when the given parameters are assumed.

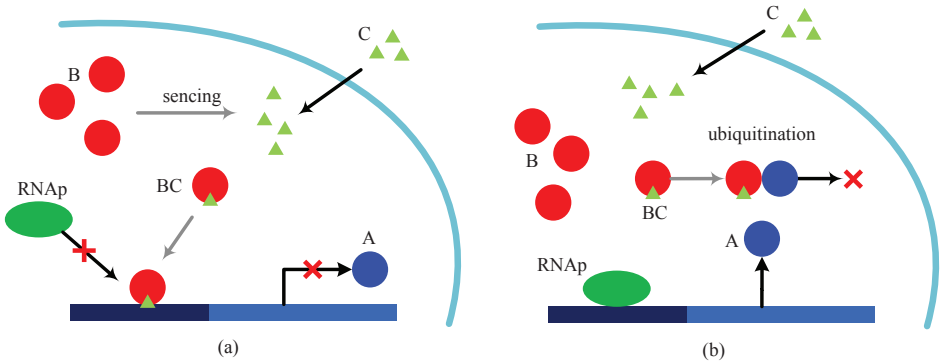


Figure 6.1: (a) Transcription inhibition model (M_{ti}). B senses the intracelluar level of a small molecule, C, and the complex BC represses the expression of A by binding to A's operator. (b) Active degradation model (M_{ad}). The protein A is degraded by the protein complex BC which acts as an ubiquitin ligase.

Active degradation for a fast response to a stress

In a GRN, the efficient way of achieving a fast response to a external stress is known to use the active degradation of proteins rather than repressing their transcription [108]. For example, 50% of LexA in E. coli is degraded within 1 minute after its related protein RecA is activated, which looks to be hard to reduce the concentration of LexA by repressing the transcription of LexA in that short time [74]. Here, we show two different types of systems, transcription inhibition model (M_{ti}) and active degradation model (M_{ad}), responding to an extracellular signal. In model M_{ti} (Figure 6.1 (a)), protein B senses the level of signal C and their protein complex BC represses the expression of gene A by binding its operator site. On the other hand, model M_{ad} (Figure 6.1 (b)) has an active degradation process in responding to the signal C. Protein complex BC acts as an ubiquitin ligase which is responsible to the active degradation of protein A.

These two systems are modeled using our proposed stochastic method and compared in terms of the response efficiency by measuring the steady-state probability (SSP) and the average number of molecules (ANM). Figure 6.2 shows the SSPs and ANMs with respect to the input rates of external signal C which is varied from 0 to 0.1.

In Figure 6.2 (c) and (f), the SSP and ANM of the complex BC are shown as almost the same in both systems, which means there is no difference between

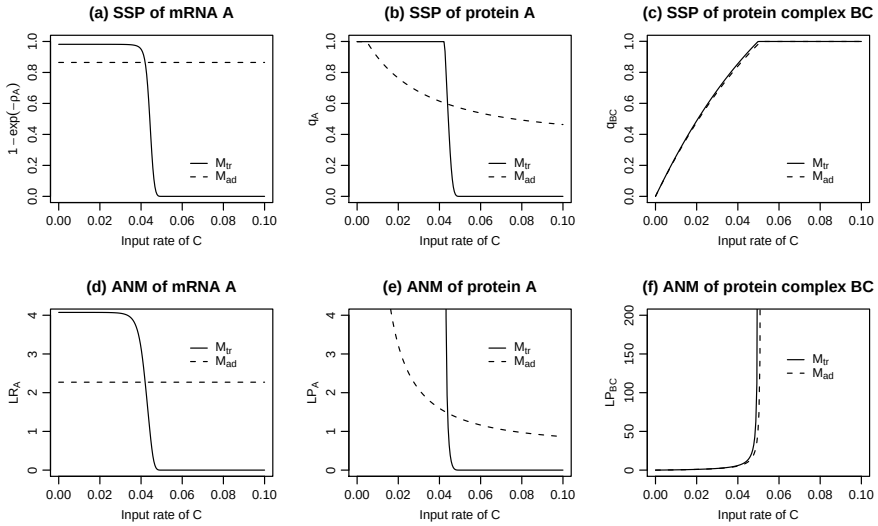


Figure 6.2: (a)-(c) Steady-state probabilities (SSP) that each queue is not empty and (d)-(f) average number of molecules (ANM) in both M_{ti} (solid line) and M_{ti} (dashed line).

these two systems in detecting the internal signal C. The reason that complex BC has the steep increment of ANM in (f) is because it is the expected numbers of molecules in the steady-states of the systems. Figure 6.2 (b) and (c) show that M_{ad} has a faster response system (reacts in the low level of C) than M_{ti} which needs an enough level of C (≈ 0.05) to turn ON its response mechanism. This results are consistent with the previous study [25] which shows the active degradation systems are responsible for the fast responses.

However, if the level of C is kept above 0.05 in M_{ti}, protein A is completely removed whereas M_{ad} has a possibility that some of protein A still stay in a cell even when the concentration of C is high enough. It might be because the mRNA of the protein A in M_{ad} keep producing its protein molecules even though the molecules are degraded quickly (Figure 6.2 (a)). Therefore, we can conclude that M_{ad} is more sensitive and faster response system than M_{ti} but M_{ti} could be a more reliable system in responding to the external signal when the signal is strong enough.

6.4 Conclusions

In this study, we proposed a new modeling approach to analyze the steady-state of GRNs which include post translation processes such as phosphorylation and ubiquitination. The analytic solution of G-networks and queuing theory enable us to explore the long run behaviour of GRNs without the expensive computation of steady-state analysis. Also the negative signals of G-networks provide a successful way to describe the active degradation of protein molecules by ubiquitination, which is not possible in conventional queueing system. In the stress response example, our method shows that the active degradation system responses more quickly to external signals than the transcription repression system. However, the transcription repression system might be more appropriate to wipe out a target protein completely.

7 Reconstruction of Large-scale GRNs using Bayesian Model Averaging

Since the stochastic modeling techniques, in previous chapters, require well defined regulatory relationships among biological molecules, it is essential to infer the regulatory structure among unknown molecules by observing their expressions. In this chapter, therefore, we investigate a reverse engineering technique which enables us to build large-scale GRNs. In addition, this approach is able to integrate various sources of biological information for reliable GRN structure estimation.

7.1 Introduction

A gene regulatory network (GRN) plays a key role in uncovering various biological problems, for example, it can identify transcription factors of a specific disease marker genes [27] and investigate evolutional clues in developmental processes which affect certain disease development [37]. Thus various mathematical/statistical methods are introduced to find gene regulatory networks [42, 131, 80, 90, 63] which usually infer the causalities among genes based on their expression intensities.

However, there are two challenges for genome-wide scale GRN construction. Firstly, microarray data used in the GRN inference consist of at most hundreds of samples (n) while the number of genes (p) could be up to tens of thousands. So models based on ordinary differential equations or linear regression with the ordinary least square are employing regularization approaches where the ordinary least square estimation with a penalty makes the solution unique and shrinks predictors towards zero [116, 60]. H. Zou and T. Hastie introduced Elastic-net (Enet) which imposes on a combination of Lasso and Ridge penalties, which is useful for the case that $n << p$ even when its predictors are highly correlated [130, 40]. Graphical Gaussian model (GGM) with Stein-type shrinkage estimator [89, 103] is also available for the large-scale GRN construction. GGM has been shown it

outperformed Mutual information and Pearson correlation based algorithms [109]. Along with this dimensionality problem, various sources of heterogeneous biological data need an appropriate integration method [129, 62] to enhance the network inference which primarily uses mRNA expression data. Bayesian networks (BNs) have been used to combine different types of data by updating their prior probability [84, 11]. However, the BNs with discretized data suffer from information loss and that of continuous data is also computationally intractable.

Here we focus on to tackle these two problems together, which is not only one of the important issues in Systems biology but also difficult to achieve due to the large-scale and complexity of a living system. As a similar approach, J. Zhu *et. al.* integrated transcription factor binding site and protein-protein interaction information along with gene expression data to construct large-scale GRNs [129]. However, their approach possibly loses gene expression information because of data discretization and the maximum in-degree of their network structure is limited by three in each node. In our previous study [70], a large-scale GRN is built using regression model based ensemble approach. In this study, the similar strategy is applied but we apply Bayesian model averaging (BMA) technique which enables us not only to merge appropriate locally fitted models but also to integrate our prior knowledge in obtaining the posterior probabilities of our interest final model.

Our method shows better sensitivities than Enet and GGM in simulation study with a fixed specificity. Also, the result of brain tumor grade dependent three large-scale networks draws a conclusion that the genes having large variance of connectivity among the three networks are mainly involved with the regulatory and immune related gene ontology (GO), which is a different insight of the differentially expressed (DE) gene approach.

7.2 Bayesian Model Averaging Networks (BMAnet)

7.2.1 BMA for linear regression models

Suppose a standardized gene expression dataset $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ where $\mathbf{x}_i$ is length n vector and the lth regression model, M_{il}, with K in-degrees

$$\mathbf{x}_i = \sum_{k=1}^{K} b_{ki}\mathbf{x}_j + e_i \tag{7.1}$$

where e_i is an error term and b_{ki} is a coefficient representing the effect of gene k on the ith gene. Let θ_{ji} be the true coefficient. Then our aim is to find $p(\Theta|X)$ where $\Theta = (\theta_1, ..., \theta_i)$ and $\theta_i = (\theta_{1i}, ..., \theta_{ji})(i, j = \{1, ..., p\})$. By Bayes' rule and the law of total probability,

$$p(\theta_{ji}|X) = \sum_{l=1}^{L} p(\theta_{ji}|X, M_{il})p(M_{il}|X) \tag{7.2}$$

$$\text{where } p(M_{il}|X) = \frac{p(X|M_{il})p(M_{il})}{\sum_{h=1}^{L} p(X|M_{ih})p(M_{ih})} \tag{7.3}$$

and L is the number of all possible models which include the jth gene as one of their predictor variables. These equations mean that the full posterior distribution of θ_{ji} is a weighted average of its posterior distributions, $p(\theta_{ji}|X, M_{il})$. The weight is the posterior model probability, $p(M_l|X)$ in Equation (7.3). Raftery A.E. obtained the posterior model probability approximation which is called Bayesian information criteria (BIC) of the model M_{il} [97]. BIC of M_{il} is $n \log(1 - R^2) + K \log n$ where R^2 is the coefficient of determinant of the model M_{il}. In our study, K is limited by 3. So, in an exhaustive searching, the maximum number of models $L = \sum_{k=1}^{K} \binom{n}{k}$. Though the maximum in-degree is fixed as three, our ensemble approach results that a node can have more than three in-degrees which will be shown in Result section.

7.2.2 Prior Update

BMA models possibly include false positives since the full posterior probability, $p(\theta_{ji}|X)$, depends on the model goodness of fits (BICs in this study) which are somtimes good enough even when the model has insignificant variables along with highly significant variables. To minimize this defective effect, each selected model is weighted using the model priors, $p(M_{il})$, which is getting steeply higher value as its variables are all significant. To this end, we define the prior distribution having the form of Gibbs distribution which is successfully used in other gene network integration studies [65, 124].

Let $E_1(M_{il})$ be the 'energy' function of M_{il}. This energy can be considered as a gap between observed and expected degree that the model M_{il} is true.

$$E_1(M_{il}) = \frac{1}{K} \sum_{k=1}^{K} (1 - I(\text{pv}_k < \alpha)) \tag{7.4}$$

where $I(C)$ is an indicator function which is 1 when C is true or 0 other wise. K is the total in-degrees and pv_k is the p-value of the t-test for the coefficient of the kth predicted variable in M_{il}. Equation (7.4) explains that the energy is zero when the coefficients of the model M_{il} are all significant with a criterion α while it increases as the model contains insignificant coefficients. So the prior distribution having the form of a Gibbs distribution is

$$P(M_{il}|\beta) = \frac{e^{-\beta E_1(M_{il})}}{\sum_{h=1}^{L'} e^{-\beta E_1(M_{ih})}} \tag{7.5}$$

where L' is the total number of selected regression models and β is a hyper-parameter which indicates the strength of the influence of the prior knowledge.

Also, other biological sources can be taken to enhance the biological significance of the inferred GRNs. In our study, DNA-protein binding affinity [36] is used to update the prior information. Let S_i be a set of genes whose expression affect the ith gene which is the response variable of M_{il}. Then the corresponding energy function is defined as

$$E_2(M_{il}) = \frac{1}{K} \sum_{k=1}^{K} (1 - I(\mathbf{x}_k \in S_i)) \tag{7.6}$$

where $\mathbf{x}_k$ is the kth predictive variable of M_{il}. So the prior probability distribution with the two energy functions, E_1 and E_2, is

$$P(M_{il}|\beta_1, \beta_2) = \frac{e^{-(\beta_1 E_1(M_{il}) + \beta_2 E_2(M_{il}))}}{\sum_{h=1}^{L'} e^{-(\beta_1 E_1(M_{ih}) + \beta_2 E_2(M_{ih}))}} \tag{7.7}$$

In the same way, other biological information can be integrated for the prior update. We used $\beta_1 = \beta_2 = 5$.

7.2.3 Two stage model selection

In our ensemble approach, the large number of genes not only causes the massive computation time but also detects too many possibly true models even when their maximum in-degree is limited by $K = k_1$. So we firstly construct a preliminary GRN structure, say G_1, with $K = k_1$. For model selection, an ordinary BMA application uses Occam's window [76] which is a generalization of the Occam's razor. Occam's window selects (i) models that are much less likely than the most likely model and (ii) models that have more likely sub models nested within them.

In our application, however, only (ii) strategy is applied since best m models are selected in each set of models having k predictors via forward variable probing instead of (i). So total $m \times K$ regression models are selected for one target gene. In the second stage, the whole process is repeated with $K = k_2$. But genes directly connected to a target gene in G_1 are only considered as the candidates of a new GRN structure, G_2. This study uses $k_1 = 3, k_2 = 10, m = 10$.

7.2.4 Build a final GRN structure

From [97], the posterior probability that the effect, θ_{ji}, is none zero is

$$P(\theta_{ji} \neq 0|X) = \sum_{l=1}^{L'} p(M_{il}|X) I(j \in \Psi(M_{il})) \tag{7.8}$$

where $\Psi(M_{il})$ is a set of predictors in the model M_{il}. In order to determine the final GRN structure, we introduce a measure called edge ratio, Er, which can be used as a criterion to prune the edges whose probability of Equation (7.8) is likely to be zero.

$$Er = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(P(\theta_{ji} \neq 0|X) \neq 0)}{n(n-1)/2} \tag{7.9}$$

where the denominator is the number of all possible undirected edges and the numerator indicates the number of undirected edges whose effect sizes are non-zero. It is known that the in-degree distribution of a network is scale free, $ck^{-\gamma}$ where k is in-degrees, c is a normalization constant, and γ is known as $2 < \gamma < 3$ in a biological network [3] ($\gamma = 2.5$ in this study). Therefore, Er can be expressed as a function of the total number of genes, n, and its criterion α_{Er} can be obtained in a numerical way using the following equation,

$$\alpha_{Er} = \frac{2c \sum_{k=1}^{n-1} k^{1-\gamma}}{n-1} \tag{7.10}$$

The final network structure is determined by deleting the edges having the smallest probability of Equation (7.8) until $Er \leq \alpha_{Er}$. We use $\alpha_{Er} = 0.002$ (Figure 7.1) which is in agreement with $Er \approx 0.002$ observed in DNA-protein/protein-protein networks from BIOGRID [110]. The posterior mean approximation of the effect size which can be used to identify the sign of regulatory relationship between genes

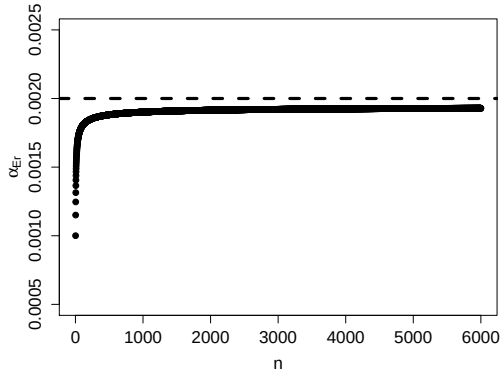


Figure 7.1: Numerically computed α_{Er} in Equation (7.10). This study uses $\alpha_{Er} = 0.002$ (dashed line)

was introduced in [97]

$$E(\theta_{ji}|X, \theta_{ji} \neq 0) \approx \sum_{l=1}^{L'} b_{ji} p(M_{il}|X) I(j \in \Psi(M_{il})) \tag{7.11}$$

7.3 Simulated Network Comparison

To evaluate the proposed method, we generated large-scale gene expression data as follows. Firstly a large-scale network structure whose in-degree distribution follows scale-free [3] was obtained using R package "igraph". This network structure can be converted into a covariance matrix [44] with appropriate edge weights which are randomly chosen among $\{-1, -0.5, 0.5, 1\}$. Then the expression data were generated from the multivariate normal distribution with the obtained covariance matrix using R package "mvtnorm". We then employed following three measures to compare the proposed Bayesian model averaging network (BMAnet) and BMAnet with second stage (BMAnet2) to the other two methods, elastic-net (Enet) [130] and Gaussian graphical model with shrinkage estimator (GGM) [89].

$$\begin{aligned} \text{Sensitivity} &= \frac{TP}{TP + FN} \\ \text{Specificity} &= \frac{TN}{FP + TN} \\ \text{Positive predictive value}(PPV) &= \frac{TP}{TP + FP} \end{aligned} \tag{7.12}$$

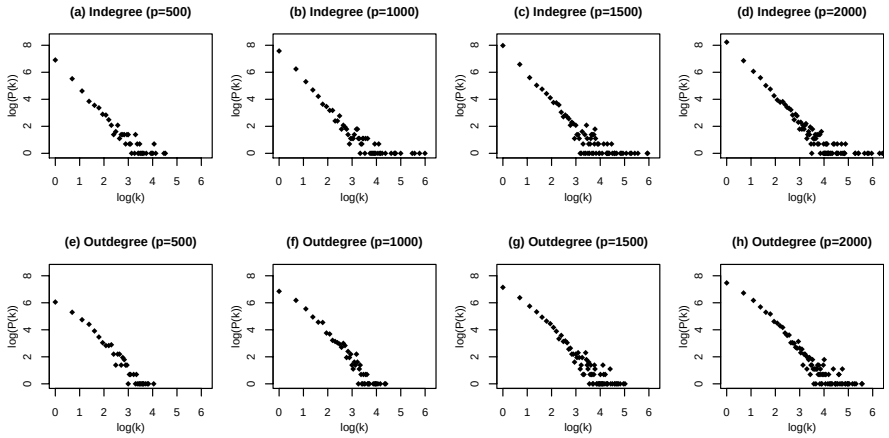


Figure 7.2: In/out-degree distributions of 500, 1000, 1500, and 2000 node networks. x-axis is the logarithm of in/out-degrees and y-axis represents the logarithm of the proportion of corresponding in/out-degrees.

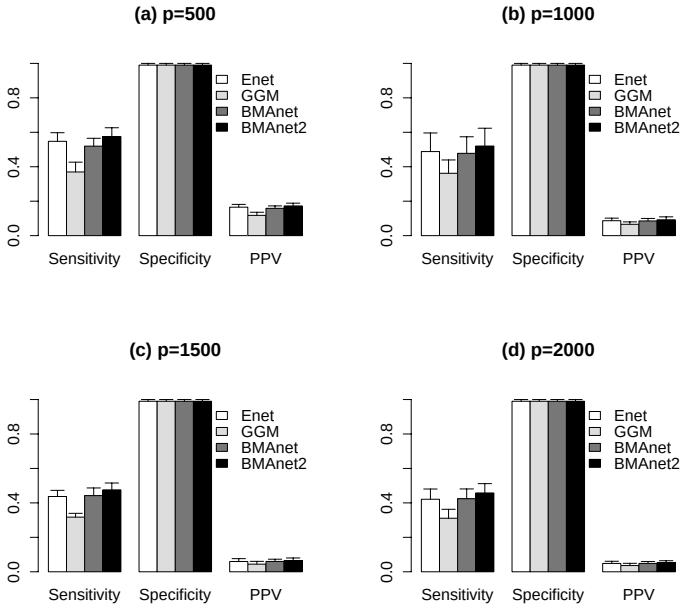


Figure 7.3: The performance comparison among BMAnet (dark gray), BMAnet2 (black), Enet (white), and GGM (gray).

where TP, FP, FN, and TN are the number of true positive, false positive, false negative, and true negative, respectively. In this study, we fixed specificity as 0.99 since the sparsity of a large-scale GRN structure causes relatively good specificity regardless of sensitivity.

We generated 20 datasets consisting of 5 replications for each of 500, 1000, 1500, and 2000 node networks. There are 100 samples in each dataset. Figure 7.2 shows the scale-free in/out-degree distributions of the simulated network structures. In Figure 7.3, BMAnet2 shows the best sensitivity in all four cases while that of Enet is slightly lower than BMAnet2. BMAnet is not good as much as BMAnet2 which means two stage approach clearly improves the performance of BMAnet. In all cases, PPVs are less than 0.2 and they tend to be getting lower as the number of genes increases, which is mainly because of the large amount of false positives in large-scale GRNs.

7.4 Human Brain Tumor GRNs

Our method is applied to human brain tumor dataset GSE4290 which is collected from Gene Expression Omnibus (GEO). This dataset has total 180 samples which consist of 23 non-tumor, 76 grade III tumor (26 astrocytomas + 50 oligodendrogliomas), and 81 grade IV tumor (glioblastomas) samples so we build three network structures for each tumor grade. Note that as the expression of this dataset was not measured in time course, the estimated networks are more likely to be "co-expression networks" rather than "gene regulatory networks". Transcription factor-promoter binding affinities estimated from multiple sources of biological data [36] are used as prior knowledge. Figure 7.4 shows the estimated three networks where edges are omitted and the degree distribution of the grade IV tumor network structure. In each network structure, nodes are placed more closely to the center as they have higher degrees. Black colored nodes are top 100 nodes (say "top 100 regulatory genes") which are highly connected to other genes in grade IV tumor while their degrees are low in non tumor samples. In order to predict the functional properties of these 100 regulatory genes, we obtained the over-represented GO terms of this group of genes using GOstat [6] which performs hyper-geometric test evaluating statistical significance of functional and molecular mechanisms of the interested genes. Table 7.1 shows most of the 100 regulatory genes are related with regulations of various biological processes. Some of them are also involved with cell development and immune system. In this result, only 5 genes are common between the top 100 regulatory genes and top 100 DE genes which are colored in white in Figure 7.4. That is detecting these regulatory genes could give different information of disease development from that of conventional DEG approach.

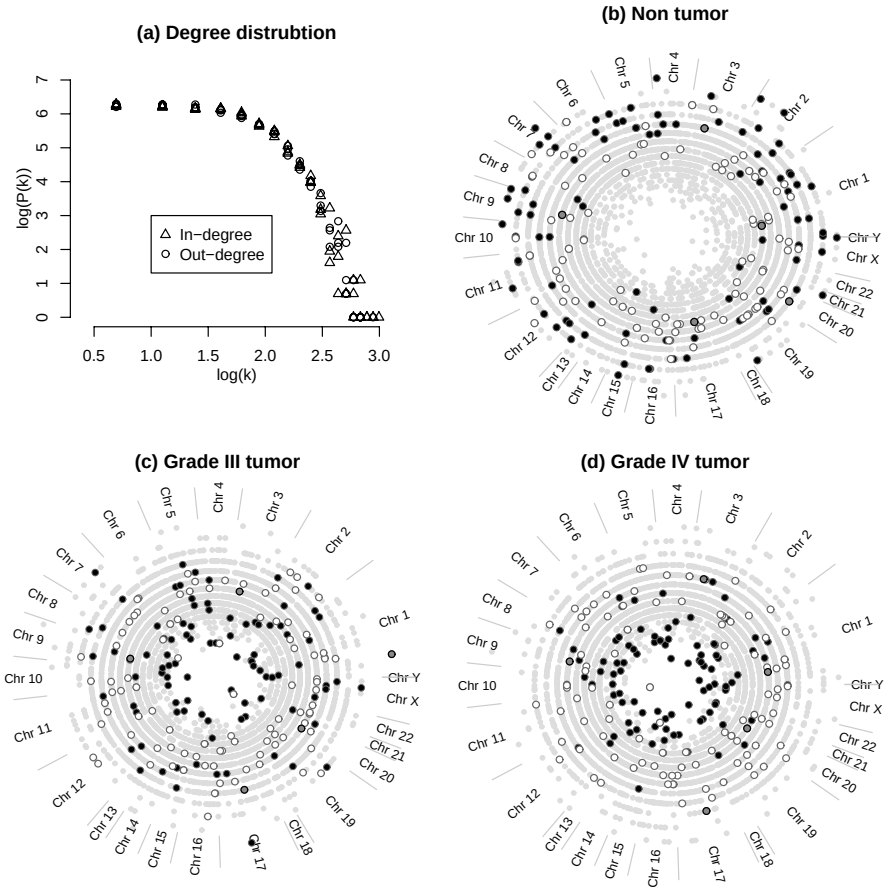


Figure 7.4: (a) Degree distribution of grade IV tumor network structure. Estimated network structures for (b) non-tumor (c) grade III tumor and (d) grade IV tumor. In each network structure, top 100 regulatory genes and top 100 DE genes are colored in black and white, respectively. Gray nodes indicate they belong to both top 100 regulatory genes and top 100 DE genes. As a node has higher degree it is more close to the center. Edges are omitted for clear representation.

Table 7.1: Significantly over-represented GO terms of the top 100 regulatory genes.

GO terms	GO term names	p-values
GO:0050789	regulation of physiological process	0.00442
GO:0048522	positive regulation of cellular physiological process	0.00442
GO:0048518	positive regulation of physiological process	0.00442
GO:0065007	biological regulation	0.00442
GO:0002376	immune system process	0.00823
GO:0048468	cell development	0.0108
GO:0050794	regulation of cellular process	0.0108
GO:0048869	cellular developmental process	0.0108
GO:0030154	cell differentiation	0.0108

7.5 Conclusions

In this study, we proposed a large-scale GRN construction method via BMA technique which is usually employed to take into account uncertainty of model selection problems. In this sense, BMA could be one of the best solutions for explaining GRNs where various molecules such as proteins alternatively interact with each other depending on their environmental conditions. Moreover, the Bayesian approach enables us to integrate different sources of biological information such as the protein-DNA binding affinity which is used in our tumor network analysis. Though our method has longer running time then Enet and VAR, it took approximately 15 hours for 4283 genes, which might be practically reasonable for large-scale networks. Via further investigation of the regulatory genes, it will be possible to identify the key regulatory genes for a development process of a certain disease.

8 Mining of Disease Candidate Genes via Large-scale GRNs

In the previous section, we showed a new reverse engineering technique which is available for large-scale GRN construction. The estimated GRNs provided a systematic view of each tumor grade (normal, grade III, and grade IV) but, for investigation of this disease, we need a tool to identify disease candidate genes in the basis of the estimated GRNs. So, in this section, various network evaluation measures are used for the comparison of estimated networks and one of the measures called random walk with restart (Rwr) is utilized to search for the disease candidate genes.

In the simulation study, our reverse engineering method (BMAnet) introduced in the previous section shows better performance than Elastic-net and Gaussian graphical models in terms of the network evaluation quantities though some topological measures are varying among the three reverse engineering methods. In the real data analysis, brain tumor networks constructed in the previous section were analyzed and we identified 169 brain tumor related candidate genes some of which were found to be related with “wound”, “apoptosis”, and “cell death” processes.

8.1 Introduction

Rapid growth of high-throughput biological data has led to the development of tools for gene regulatory networks (GRN) analysis which is capable of capturing not only the evidences of physical molecular interactions but also the perspective of their functional roles in a complex system. However, one of the main issues in systems biology is the network evaluation. Most of the reverse engineering techniques have been evaluated in terms of their sensitivity and specificity based on simulation data. However, when it comes to ’large-scale GRNs’, topological properties could be important as much as the individual component connectivity evaluation. It is mainly due to the meaningful genes could be linked to each other

as a module (motif) in the sparse and large-scale GRNs [24, 83, 9]). So number of graph alignment techniques based on graph theory have been introduced [39, 10, 73, 128] along with structural measurements [3].

The aim of our study is to find disease candidate genes using large-scale GRNs estimated from various source of biological information including mRNA expression. To this end, we build large-scale GRNs using BMAnet which is proposed in the previous section. Then the estimated networks are evaluated by 10 different quantities and one of the quantities called "random walk with restart (Rwr)" is utilized to identify the disease candidate genes. This Rwr is able to measure the distances from a gene to a set of genes that are known to be related with the target disease.

8.2 Network Evaluation

A network structure estimated by a reverse engineering method has been commonly evaluated in terms of sensitivity and specificity by comparing it to a reference network structure. But we used following 10 measures (In Figure 8.1) including (a) total number of edges and (b) maximum degree of a network graph.

(c) Difference of degree (Dd): The degree of a node means the number of connections to the other nodes. Let $d_g(i)$ be the degree of the ith gene in a network structure g ($i = \{1, 2, ..., n\}, g = \{A, B\}$). Then

$$\mathrm{Dd} = \frac{1}{n}\sum_{i=1}^{n}\{d_A(i) - d_B(i)\}^2 \tag{8.1}$$

(d) Sensitivity (Sens), (e) specificity (Spec), and (f) positive predictive value (Ppv) are defined as fallows,

$$\mathrm{Sens} = \frac{TP}{TP + FN} \tag{8.2}$$

$$\mathrm{Spec} = \frac{TN}{FP + TN} \tag{8.3}$$

$$\mathrm{Ppv} = \frac{TP}{TP + FP} \tag{8.4}$$

where TP, FP, FN, and TN are the number of true positive, false positive, false negative, and true negative, respectively.

(g) Similarity (Sim): It measures the similarity of two networks, A and B. Let E_g be the set of edges of a network structure g ($g = \{A, B\}$) then Sim can be defined

$$\text{Sim} = \frac{N(E_A \cap E_B)}{N(E_A \cup E_B)} \tag{8.5}$$

where $N(S)$ indicates the number of elements of a set S. Sim has the value 1 when both networks are the same, and decreases as two networks share less edges [128].

(h) Clustering coefficient (Cc): (Global) Clustering coefficient is a ratio between the number of connections existing among three nodes in its neighbering triplet structure and the maximal number of edges that can exist among them. If we denote the clustering coefficient of the ith node in a network g by $C_g(i)$ then Cc can be defined as

$$\text{Cc} = \frac{1}{n}\sum_{i=1}^{n}\{C_A(i) - C_B(i)\}^2 \tag{8.6}$$

(i) Average shortest path length (Asp): Shortest path length is the number of edges between two nodes in a (undirected) network so Asp is the mean distance of two nodes when these two are reachable.

$$\text{Asp} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} l(i,j) \tag{8.7}$$

where $l(i,j)$ is the shortest length between two nodes, $i, j(= 1, 2, ..., n)$. $l(i,j) = 0$ when $i = j$ or the i node cannot reach to the jth node.

(j) Random walk with restart (Rwr): Random walk on an undirected graph can be used to capture the relationship of a group of nodes in the given network structure [72]. As a variant of the discrete-time random walk process, random walk with restart allows the walk to start again from the initial node s with probability r in every time step. Let $\mathbf{p}^t$ be a length n vector where the ith element is the probability that the walker is at node i at time t. Then the next probability of Rwr is defined as follows.

$$\mathbf{p}^{t+1} = (1 + r)\mathbf{A}\mathbf{p}^t + r\mathbf{p}^0 \tag{8.8}$$

where $\mathbf{A}$ is the column-normalized adjacency matrix of the graph and $\mathbf{p}^0$ is the initial probability vector which includes a set of non-zero elements (a group of genes that we are interested in, denoted by Z). The steady state probability vector of $\mathbf{p}^t$ can be obtained by iterating Equation (8.8) until the difference between $\mathbf{p}^t$ and $\mathbf{p}^{t+1}$ is less than a criterion (10^{-6} in this study). Therefore, the Rwr of the group of genes is

$$\mathrm{Rwr} = \sum_{j \in Z} \mathbf{p}_j \tag{8.9}$$

where Z is the set of genes which are known to be related to a target disease.

8.3 Disease Candidate Gene Detection

Conventionally, disease candidate genes have been found using differentially expressed gene (DEG) approaches such as t-test or ANOVA. But in this research, we search for the candidate genes that closely linked to a group of disease related genes by measuring the distance from the group of genes to the candidate gene. Given the set of genes Z which are disease related genes, the distance between Z and the ith gene (Rwr_i) can be defined as

$$\mathrm{Rwr}_i = \sum_{j \in Z'} \mathbf{p}_j \tag{8.10}$$

where $Z' = \{Z, i\}$ and p_j is the jth element of vector $\mathbf{p}$. This Rwr has high value if the genes in Z are closely connected to the target gene i. Note that as an alternative of Rwr, Asp in Equation (8.7) could be used as a similarity measure among a group of genes but it is known that Rwr outperforms Asp which does not consider the topology of the group members [72].

8.4 Simulation Study

To evaluate the proposed Bayesian model averaging network (BMAnet), we generated large-scale gene expression data as follows. Firstly large-scale networks whose in-degree distribution follows scale-free [3] was obtained using R package "igraph". Each of these network structures was converted into a covariance matrix [44] with appropriate edge weights which were randomly chosen among

$\{-2, -1, -0.5, 0.5, 1, 2\}$. Then the expression data were generated from the multivariate normal distribution using R package "mvtnorm". In this way, we prepared 40 datasets consisting of 20 replications for 1000 and 2000 nodes, respectively. Each dataset has 50 samples. We used these simulated datasets to estimate their network structure using three reverse engineering methods, BMAnet, elastic-net (Enet) [130] and Gaussian graphical model with shrinkage estimator (GGM) [89]. In addition to the 8 quantities described in Method section, two more quantities, the number of edges and max degree, were measured from the inferred networks.

On the other hand, in order to evaluate Rwr used to detect disease candidate genes, we inserted a module network consisting of 10 nodes into each simulated network structure. A node in the module network is reachable directly or indirectly to any of the other 9 nodes. These 10 nodes are considered as the disease related genes (Z in Equation (8.10)). So high Rwr means that a walker tends to be stay in any of the 10 nodes with a high probability, which reflects the 10 nodes are closely linked together.

Figure 8.1 shows the mean and standard deviation of the 10 quantities including the total number of edges and max degrees in each of 1000 and 2000 node datasets. Note that we fixed the specificity by 0.99 in estimating network structures. In Figure 8.1, (a) the total number of edges shows similar amongst the three methods but is getting larger as the number of node increases despite the tight control of specificity (0.99). (b) The max degrees of the true network structures reached around 300 to 400 with large variances. Though BMAnet limited the maximum in-degree of a model by three, its estimated network structures have more than 50 degrees, which is almost the same as Enet. (c) In the difference of degrees, Enet shows the most closest to the true networks. However, BMAnet has the best performance in terms of (d) sensitivity, (f) positive predictive value and (g) similarity. (h) The clustering coefficient of BMAnet is also the most similar to that of true networks. (i) Average shortest path values of BMAnet and Enet are comparable to that of the true networks. But variance of BMAnet is smaller than the others. (j) Random walk with restart (The higher, the better) of BMAnet is relatively lower than Enet and GGM but BMAnet shows the most robust results.

8.5 Human Brain Tumor GRNs

In order to apply our approach to real data, human brain tumor dataset GSE4290 was collected from Gene Expression Omnibus (GEO) and its large-scale GRNs

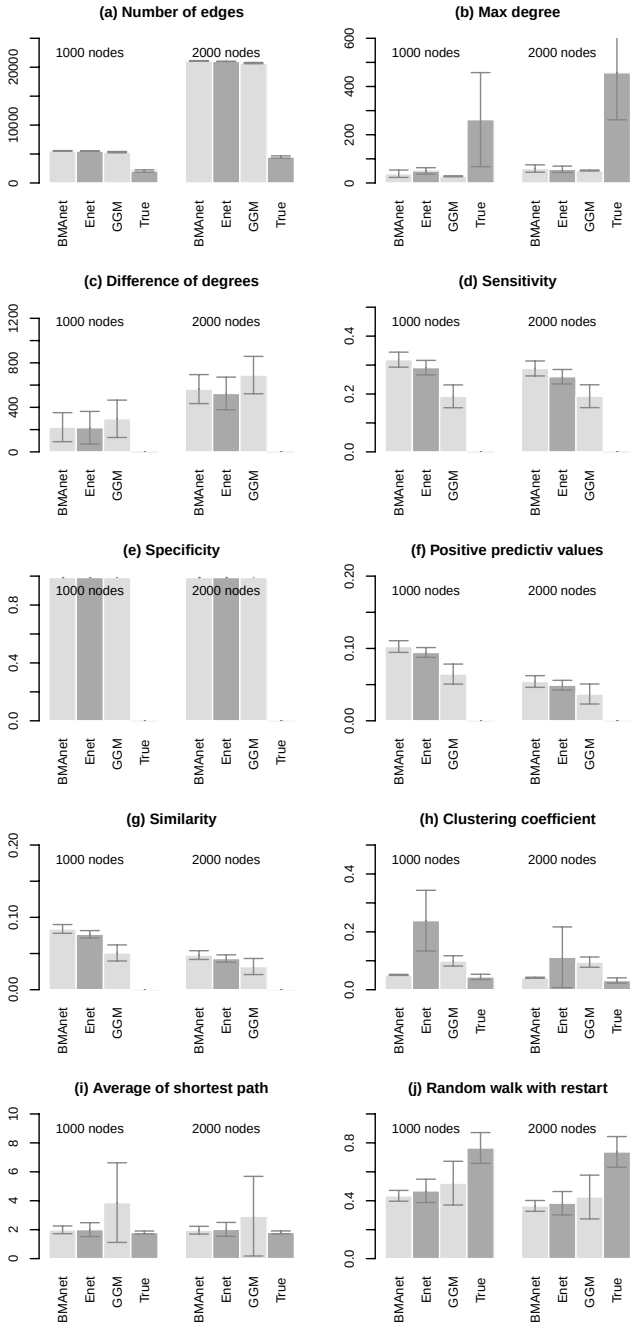


Figure 8.1: Comparison of 10 measurements of the three reverse engineering methods (BMAnet, Enet, GGM) along with the true networks (True). Please see the context for detailed description.

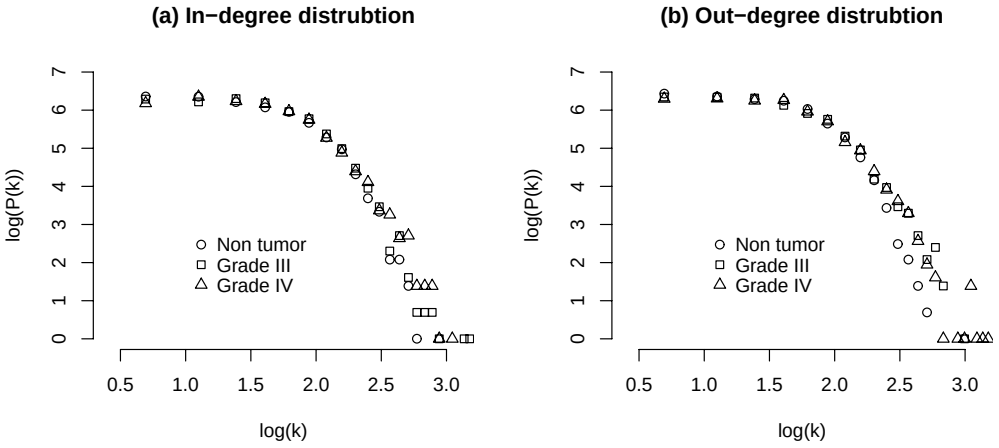


Figure 8.2: (a) In-degree and (b) out-degree distributions of nontumor, grade III, and IV tumor network structures.

were constructed. This dataset has total 107 samples which consist of 23 non-tumor, 76 grade III tumor (26 astrocytomas + 50 oligodendrogliomas), and 81 grade IV tumor (glioblastomas) samples. We normalized the expression data of each grade and selected 4422 genes based on the significant DEGs from ANOVA. Then BMAnet was applied to build the network structures for each of the three groups. Note that the samples of this dataset were not measured in time course, the estimated networks are more likely to be "co-expression networks" rather than "gene regulatory networks". Along with the expression datasets, DNA-protein binding affinity [36] is used to update the prior information of BMAnet.

Figure 8.2 shows the in/out-degree distributions of the estimated networks. Though the distributions are not exactly following the power-law distribution, they show a reasonable range of in/out-degrees covering more than 20 (e^3) despite a model in BMAnet has the maximum three in-degree limitations.

In order to find candidate genes affecting brain tumor, we found 107 tumor related genes from NCBI Gene database. Then the Rwr_i of Equation (8.10) were measured for all 4422 genes and identified 169 genes that are up-regulated in grade III and IV samples while they are not expressed in non-tumor samples. Figure (8.3) (a) $\sim$ (c) show the biplots of Rwr values among the three tumor grades; (a) non-tumor vs. grade III, (b) non-tumor vs. grade IV, (c) grade III vs. grade IV. In each plot, we divided four groups of genes based on their expression patterns, for example, U_{N3} (D_{N3}) indicates the genes having higher (lower) Rwr values in grade III while it is low (high) in non-tumor group. Genes in B_{N3} group have

Table 8.1: Significantly over-represented GO terms (Biological Process) with p-value < 0.01.

GO term names	p-values
response to wounding	0.00005
negative regulation of gene expression	0.0003
negative regulation of transcription	0.0004
inflammatory response	0.0007
negative regulation of nucleobase-containing compound metabolic proc.	0.0012
negative regulation of nitrogen compound metabolic process	0.0014
positive regulation of macromolecule metabolic process	0.002
negative regulation of macromolecule biosynthetic process	0.0022
regulation of apoptotic process	0.0024
regulation of programmed cell death	0.0026
regulation of cell death	0.0028
negative regulation of cellular biosynthetic process	0.0028
response to hypoxia	0.0032
negative regulation of biosynthetic process	0.0034
regulation of transcription from RNA polymerase III promoter	0.0039
response to oxygen levels	0.041
wound healing	0.045
negative regulation of macromolecule metabolic process	0.054
positive regulation of apoptotic process	0.066
positive regulation of programmed cell death	0.069
positive regulation of cell death	0.072
positive regulation of nitrogen compound metabolic process	0.093
regulation of cell cycle	0.093

high Rwr values in both grade III and non-tumor groups. Therefore, our interested genes in this study are belonging to $\{(U_{N3} \cup U_{N4}) \cap B_{34}\}$

In Figure (8.3), (d), (e) and (f) show the non-tumor, grade III, and grade IV gene networks, respectively. Light gray dots indicate the known 107 tumor related genes and dark gray ones denote the identified 169 genes via our approach. As we expected, the 169 genes are more closely linked with the known 107 genes in the grade III and IV networks than the non-tumor network. In order to estimate the functional properties of these 169 genes, we obtained the over-represented GO terms of these genes using DAVID [31] which performs Fisher exact test for evaluating the association between two independent groups (observed proportion in a group vs. background proportion in the other group). Table 8.1 shows GO terms whose p-values of the Exact test are less than an empirically chosen criterion (0.1). It shows that many of the 169 genes are related with regulations of "wounding", "cell death", "cell cycle", and "apoptosis".

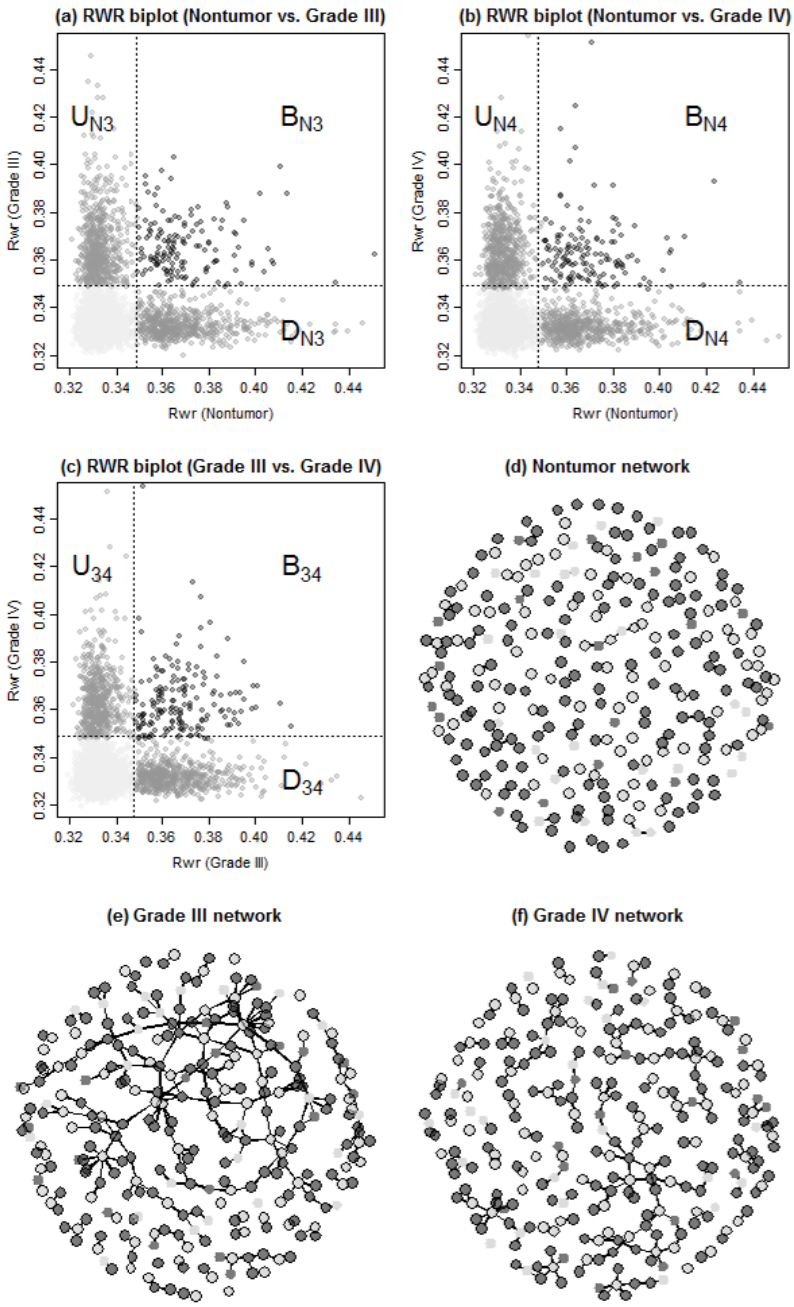


Figure 8.3: Biplots of Rwr (a) non-tumor vs. grade III, (b) non-tumor vs. grade IV, (c) grade III vs. grade IV. In each plot, we divided genes in three groups, U_{AB}: genes having high Rwr in B group but low in A group, D_{AB}: genes having low Rwr in B group but high in A group, B_{AB}: genes having high Rwr values in both A and B group. (d) co-expression networks of Non-tumor, (e) grade III, and (f) grade IV samples. Light gray dots represent the known tumor related genes and dark gray nodes indicate the identified 169 genes

8.6 Conclusions

In this study, we have shown network evaluation and disease candidate gene detection techniques. In order to evaluate networks estimated by BMAnet, 10 quantities of the networks were introduced and the simulation study reveals that BMAnet showed better or similar performance than Enet and GGM. The disease candidate genes were found by one of the network evaluation measures, Rwr_i, which is a distance from the ith gene to a group of genes known to be related with the target disease. Utilizing Rwr is not a new approach but we hope we can identify more reliable candidate genes by cooperating with a proper large-scale reverse engineering method. For example, in the brain tumor analysis, ANOVA and FDR adjustment detected too many DEGs (For example, 3023 out of 4422 genes have less FDR values than 0.01. These genes are indicated with black colored edges in Figure 8.3 (d), (e), and (f)) so it is not possible to pick up proper candidates without the estimated network structures.

Via further investigation of the candidate genes, we might be able to identify the disease triggering pathways which will provide more variety of drug targets in industry. But it is also necessary to understand more detailed regulatory relationships among the candidate genes including their proteins/metabolites' dynamic behaviors.

9 Meta Analysis for Brain Tumorigenesis

So far, this thesis has introduced the stochastic modeling and reverse engineering techniques. In order to extend these techniques to a large-scale real dataset, this chapter analyzes brain tumor expression data using a series of bioinformatics approaches including a meta-analysis. These analyses are aiming to select an appropriate set of genes which possibly cause tumorigenesis. The selected genes from the meta-analysis will be used for the abnormal pathway detection algorithm in Chapter 10. Along with the gene selection, this chapter focuses on investigating the function of a specific gene (KAL1) which could play an important role in brain tumorigenesis. Due to the difficulty of brain tissue sampling especially for normal samples, this study was started from the meta-analysis of the publically available datasets than gene ontology based clustering analysis was followed to infer the biological meaning of the clustered genes. These series of bioinformatics techniques would be useful to find biological functions of candidate genes in disease research.

9.1 Introduction

Anosmin-1 protein, encoded by KAL1 gene, is an extracellular matrix (ECM) associated protein which plays essential roles for the establishment of olfactory and gonadotrophin-releasing hormone neurons during early brain development [30, 88, 22]. There is, however, little comprehension of its role in the developed brain. Since reactivation of developmental signal pathways often takes a part in tumorigenesis, we investigate if anosmin-1-mediated cellular mechanisms associate with brain tumor formation.

Recent proliferation of expression databases enables us to collect the majority of gene expression samples from the public databases such as Gene Expression Omnibus (GEO) [33]. Along with these databases, meta-analysis provides a solution to integrate and analyze the heterogeneously collected datasets. We classify our

collected datasets from the GEO database into three groups (normal, low, and high grades) based on their histological characteristics. A rank-product based meta-analysis [18] is used to find differentially expressed probes among the different tumor grades. Then, we perform gene ontology (GO) based clustering analysis which reveals the genes having similar expression pattern with KAL1 as well as their functional annotations. In addition to the public datasets, we have 43 microarray profiles from brain biopsy samples of patients recruited at St. George's Hospital Neurosurgery department and their analyses reveals KAL1 is differentially expressed not only between different tumour grades but also different tissue types.

9.2 Meta-analysis of Microarray Data

To obtain a global gene expression profile of different grades of human brain tumors, we first investigated publicly available genome wide microarray data by using meta-analysis which allows the integration and analysis of heterogeneous datasets. We chose the rank-product method, known to outperform other meta-analysis approaches (13, 14). We have collected data samples from 10 independent studies in GEO database (Table 9.1). In order to categorize the data according to their histological characteristics and WHO grading system, we carefully screened the tissue types in each study and reviewed the description of the sample collection. To minimize the heterogeneity in data sampling, we chose datasets obtained only from Affymatrix GeneChip U133A (GPL96), and Gene Chip Robust Multi-array Average (GCRMA) normalization was applied. The data were classified into three groups normal (N) brain, low grade (L) tumor (WHO grade I and II) and high grade (H) tumor (WHO grade III and IV). In total, we analyzed 461 datasets consisting of 19273 probes and firstly conducted the meta-analysis on each pair of the three groups - normal vs. low grade (NL); low grade vs. high grade (LH); and normal vs. high grade (NH) - under the null hypothesis that the paired groups have same mean expression values.

The meta-analysis allows us to address a common hypothesis from combined results of several different studies. The rank-product method, one of the most efficient techniques among meta-analysis approaches [18], simply ranks the ratio of a genes expression levels among samples and then tests the significant of its statistic which is a product of the corresponding ranks. This approach has higher sensitivity and selectivity than the conventional t-test based methods [64]. Gene Chip Robust

Table 9.1: Microarray samples (total 461) from GEO database used in the meta-analyses

Tumor group	GEO Series ID	num. of samples
Normal (53)	GSE12649	34 (adult postmortem brain)
	GSE5389	11 (adult postmortem brain)
	GSE5490	8 (adult postmortem brain)
Low (32)	GSE12907	21 (grade I astrocytoma, pilocytic)
	GSE3185	3 (grade I & II astrocytoma)
	GSE4780	5 (grade I & II meningioma)
	GSE8692	3 (oligodendroglioma, anaplastic mixed glioma)
High (376)	GSE13041	191 (Glioblastoma Multiforme)
	GSE4412	85 (Glioblastoma Multiforme, anaplastic astrocytoma, anaplastic mixed oligo-astrocytoma, anaplastic oligodendro-glioma)
	GSE4271	100 (grade III & IV astrocytoma)
	Total	461

Multi-array Average (GCRMA) normalization is applied to the raw datasets of the three groups (Normal (N), Low (L), and High (H)) and the meta-analysis is performed in each pair (NL, LH, and NH) to find differentially expressed genes. Suppose the case that normal and low group comparison (NL).

Method

Let $X_{i,j,k}$ be the logarithm of the ith probe expression of the jth sample in the kth group ($i = \{1, ..., p\}, j = \{1, ..., n\}, k = \{N, L, H\}$). Then we can compute the ratio of the ith probe between corresponding expression of the jth sample in low group and jth sample in normal group, $Y_{i,m} = X_{i,j,N} - X_{i,j,L}, m = \{1, , M(= n_N n_L)\}$ where n_N and n_L are the total number of samples in normal and low groups, respectively. After ranking the $Y_{i,m}$ among M comparisons, the rank product statistic RP_i of the ith probe can be computed, $RP_i = \left(\sum_{m=1}^{M} R_{i,m}\right)^{1/M}$ where $R_{i,m}$ is the rank of $Y_{i,m}$. In order to generate the null distribution for our hypothesis that the expression distributions of normal and low groups are not the same, the probes in each sample are permuted randomly regardless of their groups. Then the permuted rank product statistics RP_i^* are computed in the same way as RP_i. Now the empirical p-values of RP_i can be obtained as follows

$$pval_i = \frac{\sum_{r=1}^{N_p} I(RP_{i,r} < RP_i)}{N_p} \tag{9.1}$$

where N_p is the total number of permutations and $I(C)$ is the indicate function which is 1 if C is true or 0 other wise. If the ith probe is more highly expressed in

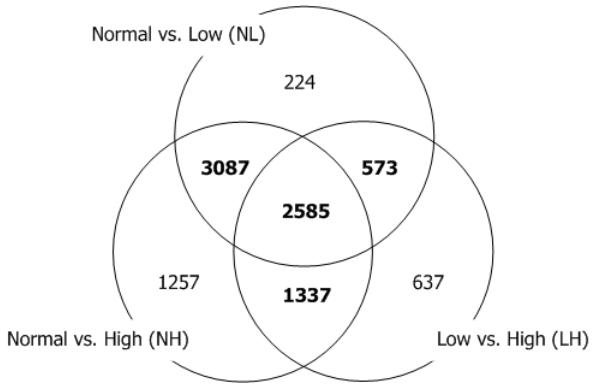


Figure 9.1: A diagram showing the number of differentially expressed probes among the three tests (NL, NH, and LH).

Table 9.2: Differentially expressed probes in the three meta-analyses. 33.5% (=(3213+3257)/19273), 26.6% (=(3008+2124)/19273), and 42.8% (=(4706+3560)/19273) probes are differentially expressed with 0.001 significance level in NL, LH, and NH tests, respectively. The alternative hypothesis is shown in brackets.

Tests	Num. of significant probes [alternative hypothesis]		Total no. DE probes (genes)	% DE probes (DE probes/total probes)
Normal vs. Low (NL)	3213 [$N < L$]	3256 [$N > L$]	6470 (4864)	33.5% (6470/19273)
Low vs. High (LH)	3008 [$L < H$]	2124 [$L > H$]	5132 (4034)	26.6% (5132/19273)
Normal vs. High (NH)	4706 [$N < H$]	3560 [$N > H$]	8266 (6072)	42.8% (8266/19273)

low group than normal group, its rank-product statistic tends to be small and the empirical p-value closes to zero. In our study, N_p is 5000.

Result

The results show 33.5%, 26.6% and 42.8% of the total probes are differentially expressed (i.e. up- or down-regulated) in NL, LH and NH comparison, respectively, with 0.001 significance level. Total 9701 probes showed differential expression and the number of probes from each of the test results are shown in Figure 9.1 and Table 9.2.

Notably, in all the three tests, there was strong evidence (empirical p-value=0, which means there is no case (evidence) supporting the null hypothesis of the corresponding test during 5000 permutations) that KAL1 expression is upregulated in both low and high grade tumors compared to normal control, and its expression is more elevated in high grade tumors compared to low grade (Table 9.3).

Table 9.3: Statistics of KAL1 in three meta-analyses. The lower the rank-product value is, the more significantly different the KAL1 expression is in each test. Alternative hypothesis is shown in brackets.

Tests	Rank-product statistic (empirical p-value) [alternative hypothesis]	
Normal vs. Low (NL)	2637.756 (0) [$N < L$]	9955.508 (0.993) [$N > L$]
Low vs. High (LH)	3992.405 (0) [$L < H$]	5779.429 (0.072) [$L > H$]
Normal vs. High (NH)	2375.587 (0) [$N < H$]	12815.553 (1.000) [$N > H$]

Table 9.4: Twelve expression patterns of differential probes. The expression patterns of 7582 probes in the meta-analysis. Total 12 patterns and the probe numbers are shown. The $+$ or $-$ sign represents up- or down-regulation, respectively, in comparison to their base group denoted as B. (e.g. 2806 probes in pattern 1 are significantly over-expressed in the low group by the NL test and, at the same time, they are over-expressed in the high group by the NH test.) KAL1 belongs to pattern 1, 5 and 9 (shown in bold).

Pattern	Normal	Low	High	No. probes
1	**B**	+	+	**2806**
2	B	+	-	54
3	B	-	+	45
4	B	-	-	2767
5	-	**B**	+	**592**
6	-	B	-	1098
7	+	B	+	1063
8	+	B	-	405
9	-	-	**B**	**1629**
10	-	+	B	759
11	+	-	B	740
12	+	+	B	794

To further validate this finding and to pursue a clustering analysis which will be shown later, we focused on the 7582 differentially expressed probes which are commonly found in at least two meta-analysis tests (Figure 9.1). Each of three tests has four possible expression patterns so total 12 patterns and their probes are shown in Table 9.4. KAL1 belongs to three groups showing the expression pattern of $N < L < H$, which indicats a positive correlation of KAL1 expression with neoplastic progression towards high grade brain tumors.

9.3 Analysis of Brain Tumor Biopsy Samples and Validation of the Microarray

In addition to the public datasets, Dr. Soo-Hyun Kim generated microarray profiles from 42 brain tumor biopsy samples obtained at St. George's Hospital Neurosurgery department (Table 9.5). The tumor type and World Health Organization (WHO) grading for each sample were confirmed by two independent histopathologists. The cohort included 6 types of tumors - glioblastoma (11 cases), astrocytoma (4 cases), meningioma (15 cases), ependymoma (1 case), hemangioblasoma (1 case) and adenocarcinoma (10 cases). We grouped the 32 primary brain tumors (excluding 10 metastatic adenocarcinomas) into low (total 19 cases) and high (total 13 cases) grades based on histopathology diagnosis and conducted t-test with a null hypothesis that KAL1 level is same between two groups. Our unpaired two sample t-test results indicated that KAL1 is significantly upregulated in high grade tumors with p-value=4.81×10^{-6} (FDR=0.256).

However, in this t-test, 11 out of 13 high grade samples were glioblastoma and 15 out of 19 low grade samples were meningioma. Therefore, we asked whether there is any difference in KAL1 expression according to the tissue types. The mean expression value of KAL1 for each group is shown in Table 9.6. One-way ANOVA under a null hypothesis that expression of KAL1 among different tumor types is same gave a p-value of 2.20×10^{-8} of its F-test (FDR=8.95×10^{-7}). A further analysis by Tukey Kramer method [41] as a post hoc test showed that KAL1 is significantly upregulated in high grade primary brain tumors (glioblastoma and astrocytoma) compared to low grade primary brain tumors (meningioma) and secondary tumors (metastatic adenocarcinoma) (Table 9.7).

Next, Dr Soo-Hyun Kim validated these microarray results by directly assessing KAL1 mRNA in the tissue samples. Total RNA was extracted from formalin-fixed paraffin embedded (FFPE) tissue samples and real time quantitative RT-PCR was performed to generate the KAL1 mRNA concentration in 20 samples (8 high grade, 12 low grade). Welch's t-test (alternative: high $>$ low) confirmed that high grade tumors exhibited higher KAL1 expression compared to low grade (t-statistic=2.21, p-value=0.031) (Fig. 9.2 and Table 9.5). These results thus confirm our in silico data obtained from meta-analyses.

Table 9.5: Brain tumor samples from St. George's Hospital. The patients, types of tumors and KAL1 expression levels assessed by microarray and qRT-PCR are shown. KAL1 mRNA level determined by qRT-PCR is indicated as the concentration ratio normalized to β-actin; NA, not determined due to limited tissue availability.

Sample ID	Sex	Age	KAL1 level by microarray	KAL1 level by qRT-PCR	Type of Tumour
524	M	34	9.57	7.17×10^{-4}	
427	M	59	11.79	2.98×10^{-4}	
433	F	70	11.88	1.77×10^{-4}	
409	M	57	11.13	2.67×10^{-5}	
426	F	61	10.97	8.31×10^{-5}	
417	M	56	11.48	4.87×10^{-5}	Glioblastoma (WHO grade IV)
463	F	65	9.88	4.15×10^{-5}	
018	M	79	10.14	NA	
461	M	16	11.26	NA	
369	F	65	10.42	NA	
460	F	58	9.62	NA	
366	M	40	9.96	1.77×10^{-4}	Anaplastic ependymoma (WHO grade III)
140	F	27	11.08	NA	Anaplastic Astrocytoma (WHO grade III)
079	F	40	12.50	NA	Astrocytoma (WHO grade II)
403	M	19	10.28	NA	Gemistocytic Astrocytoma (WHO grade II)
360	M	49	3.94	7.73×10^{-5}	
501	F	71	8.95	7.72×10^{-5}	
549	F	66	6.55	5.96×10^{-5}	
048	F	78	5.73	3.39×10^{-5}	
521	F	88	7.77	3.28×10^{-5}	
514	F	46	7.56	2.74×10^{-5}	
169	M	66	8.00	1.90×10^{-5}	
167	F	35	6.99	9.98×10^{-6}	Meningioma (WHO grade I)
456	M	52	7.08	4.14×10^{-6}	
384	M	49	5.35	3.89×10^{-6}	
476	F	65	7.57	NA	
455	F	60	8.89	NA	
367	F	37	3.92	NA	
394	F	62	7.23	NA	
502	M	72	3.18	NA	
075	M	20	11.34	NA	Pilocytic Astrocytoma (WHO grade I)
537	F	71	6.76	2.64×10^{-5}	Haemangioblastoma (WHO grade I)
449	M	70	6.10	2.68×10^{-4}	
408	M	67	4.66	4.64×10^{-5}	
210	F	69	7.91	1.44×10^{-5}	
153	F	58	9.91	4.20×10^{-5}	
536	M	73	5.33	1.43×10^{-5}	Metastatic adenocarcinoma
166	F	56	6.14	NA	
475	M	60	8.13	NA	
398	F	73	8.78	NA	
211	F	65	9.22	NA	
438	M	61	6.99	NA	Metastatic large cell carcinoma

9.4 Clustering and Gene Ontology Analysis

One way to predict the functional properties of KAL1 in brain tumorigenesis is to review Gene Ontology (GO) of genes whose expression patterns are similar

Table 9.6: Differential expression of KAL1 in different tumor types. Microarray mean expression value of KAL1 gene in each tumor type is shown. Standard deviation in brackets.

	Meningioma	Astrocytoma	Glioblastoma	Adenocarcinoma
Mean (Std.dev.)	7.32 (1.75)	11.30 (0.85)	10.74 (0.85)	6.58 (1.79)

Table 9.7: ANOVA result of KAL1 followd by Tukey's tests. Each value represents the mean expression difference between two tumor types (column vs. row). p-value is shown in the bracket and only the significant results (p-value < 0.01) are in bold.

Tissue	Adenocarcinoma	Astrocytoma	Glioblastoma	Meningioma
Adenocarcinoma	-	**3.98** (4.4×10^{-4})	**3.42** (5.0×10^{-5})	-0.74 (0.633)
Astrocytoma	-	-	0.56 (0.92)	**-4.72** (1.6×10^{-5})
Glioblastoma	-	-	-	**-4.16** (2.0×10^{-6})
Meningioma	-	-	-	-

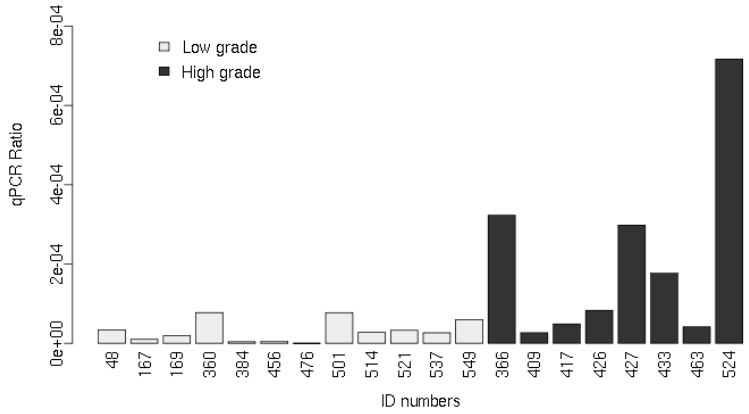


Figure 9.2: Expression profile of KAL1 in brain tumor tissues. KAL1 mRNA concentration in tumor biopsy samples, grouped as low and high grade, is assessed by qRT-PCR and normalized to β-actin.

with KAL1 (say "KAL1 cluster"). This information can be obtained using GOstat [7] which performs hyper-geometric test evaluating statistical significance of functional and molecular mechanisms of genes of interest. However, GOstat analysis of 592 probes in pattern 5 including KAL1 (Table 9.4) gave 349 over-represented GO terms with 0.01 criterion of the hyper-geometric test. Since these 349 GO terms are too many to characterize the gene group, we divided the whole 7582 probes into appropriate number of groups by applying the Partitioning Around Medoids (PAM) method [99] which is known to be more robust than k-means [119]. But

PAM requires optimal cluster size which is generally determined by the goodness of fit of each model such as silhouette [100].

Recently, GO based clustering validation and annotation approaches have been developed and provide biologically relevant interpretations of the clustered genes [7, 13]. We also applied the GO information to the PAM method to find optimal cluster size by introducing a new clustering validation measure called GO closeness (GOC). In our study, the over-represented GO terms of each cluster are firstly identified by hyper-geometric test and build a GO tree where all the GO terms are belonging to. Then we count the over-represented GO terms that are directly connected to each other in the GO tree. GOC is a ratio between the observed count and the expected count of the directly connected GO terms. Larger GOC means the GO terms are more closely located in the GO tree so it could be a guide line to group genes in terms of their biological functions.

Method

The new clustering validation measure, GOC, tells us how closely the over-represented GO terms are located in their GO tree where its very top leafs are the over-represented GO terms and linked down to all their ancestors until the root term. So, high GOC score of a cluster means that its functions in terms of significantly over-represented GO terms are closely related with each other. Note that irrespective of how many GO terms are over-presented, a cluster has only one GO tree because every GO term is eventually connected to one root term. In our case, the root is GO:0008150 (biological process). Let $C_{j,k}$ be the jth cluster in k size clustering analysis and $G_{j,k}$ be its GO tree. In order to compute GOC, we need to count the number of cases, $N_{j,k}$, that two over-represented GO terms are directly connected to each other in their GO tree. Then the GOC measure can be defined as follows,

$$GOC_{j,k} = \frac{N_{j,k}}{E(N_{j,k})} \tag{9.2}$$

where $E(N_{j,k})$ is the expected number of $N_{j,k}$. $E(N_{j,k})$ can be obtained by randomly permuting all the probes in all clusters without changing the number of probes in each cluster and counting the cases $(N_{j,k}^{(i)})$ that two over-represented GO terms are directly connected to each other. Then $E(N_{j,k})$ is

$$E(N_{j,k}) = \frac{1}{M_p} \sum_{i=1}^{M_p} N_{j,k}^{(i)} \tag{9.3}$$

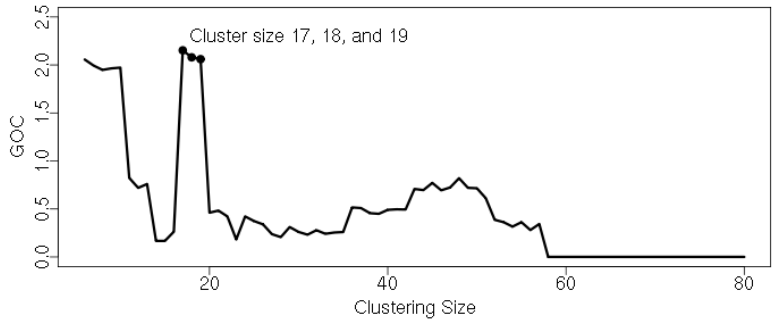


Figure 9.3: GOC values in each cluster analysis. The best cluster sizes are 17, 18, and 19.

where M_p is the total number of permutations (M_p = 500 in this study).

Result

We applied the PAM clustering to the 7582 probes selected from the meta-analysis with k cluster size ($k = 5, 6, ..., 80$). In each clustering analysis, significantly over-represented GO terms were identified using GOstat with criterion 0.01. Figure 9.3 shows the GOC values against the cluster sizes. The best cluster sizes were 17, 18, and 19 each of which identically has 364 probes and 29 over-represented GO terms.

In this KAL1 cluster, four key GO terms were primarily associated - neuronal development, morphogenesis, microtubule depolymerization and energy-coupled proton transport (Fig. 9.4). These results not only validate our Gene Ontology analysis as it confirmed the known functions of KAL1 (i.e. neuronal development and morphogenesis), but also suggest so far unrevealed function of KAL1 gene potentially involving microtubule formation and proton mobility.

9.5 Expression Pattern of Extracellular Matrix Genes

It has been suggested that extracellular matrix (ECM) in the brain has a unique composition, determining the characteristic invasion pattern of brain tumor cells [8]. Moreover, heparan sulfate proteoglycans are known to be required for the function of anosmin-1 [23, 59], thus we examined the expression pattern of several other ECM- and membrane-associated molecules to identify any common expression pattern. From the three meta-analyses, we reviewed the differential expression

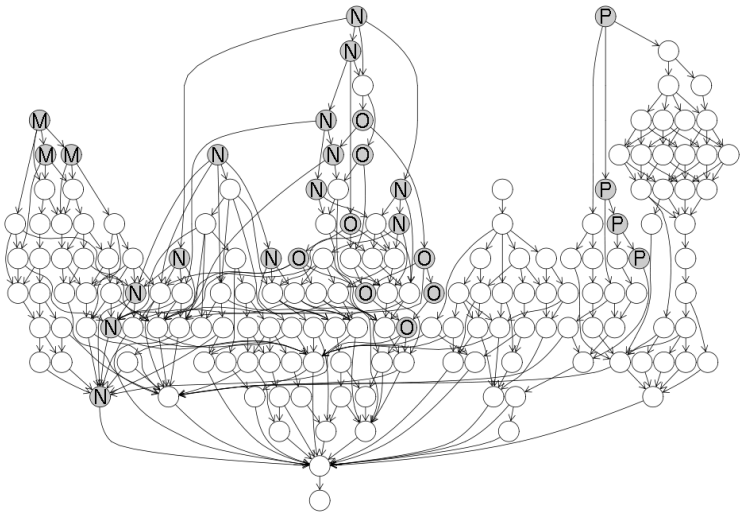


Figure 9.4: The GO tree for the KAL1 cluster with optimal cluster size 17. The 29 GO terms associated with KAL1 cluster can be grouped into 4 key terms - neuron development (GO:0010721, GO:0010975, GO:0016043, GO:0031175, GO:0031345, GO:0045664, GO:0050768, GO:0050770, GO:0050771, GO:0051128, GO:0051129, GO:0051961, GO:0051964), morphogenesis (GO:0000902, GO:0010769, GO:0022604, GO:0032989, GO:0032990, GO:0048667, GO:0048812, GO:0048858), microtubule formation (GO:0007026, GO:0031111, GO:0031114) and proton transport (GO:0006818, GO:0015985, GO:0015986, GO:0015992). Gray nodes represent each group - M for microtubule, N for neuron development, O for morphogenesis and P for proton transport.

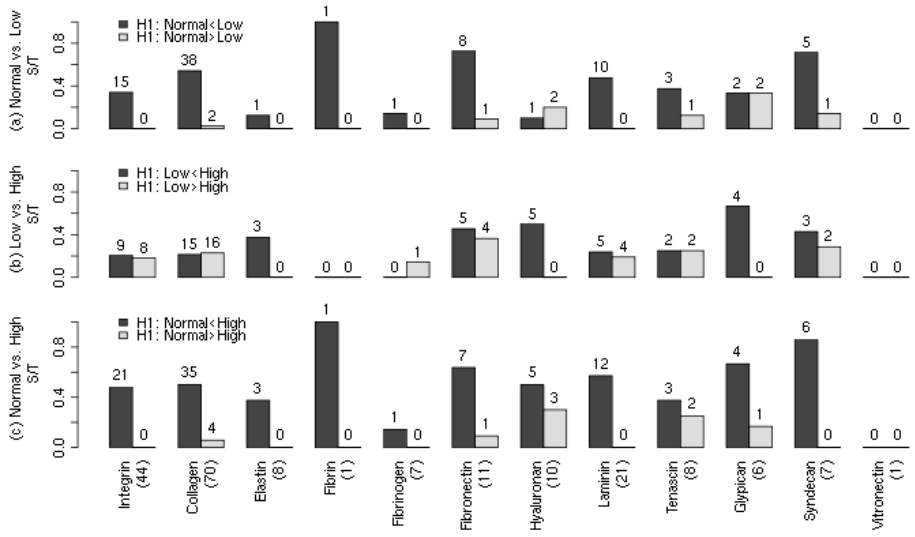


Figure 9.5: Significantly expressed ECM genes identified in the three meta-analyses with significant level 0.05. The number on each bar indicates the number of significant probes (S) in the corresponding gene family as indicated in the x-axis. The total number of probes (T) in each gene family is denoted in brackets below the gene name. The y-axis represents the value of S/T ratio in each gene family.

of integrins, collagen, elastin, fibrin, fibrinogen, fibronectin, hyaluronan, laminin, tenascin, vitronectin, glypican and syndecan, with 0.05 criterion (Fig. 9.5).

In NL test, most gene families except hyalurnonan, glypican, and vitronectin, are significantly increased in low grade tumors compared to normal tissue, while the LH test indicate that these genes maintain similarly high, if not further increased levels of expression in high grade tumors. In NH test, all gene families except vitronectin show significant up-regulation in high grade tumors. These results categorically demonstrate that there is a consistent and overall increase of ECM proteins and cell adhesion molecules during neoplastic progression. Interestingly, vitronectin showed a very low expression level throughout all samples, indicating that vitronectin is not abundantly present in brain tissues in general.

9.6 Conclusions

This study investigates if the cellular mechanisms of brain development related anosmin-1 associate with brain tumor formation. We performed meta-analysis and

statistical gene ontology analysis with publicly available microarray datasets as well as expression data obtained from a patient cohort, which indicated that KAL1 mRNA is significantly up regulated corresponding to the tumor progression when compared to the normal brain. Both in these public and hospital data analyses, Dr. Soo-Hyun Kim drew an conclusion that increased anosmin-1 expression may also module the interaction of tumor cells with ECM as it inhibited cell adhesion on the fibronectin matrix. These results suggest that the reactivation of the developmental pathways mediated by anosmin-1 contributes to neoplastic progression in the brain.

10 Large-scale GRN Analysis for Brain Tumorigenesis

Chapter 5 introduced the stochastic modeling approach to detect abnormal pathways of GRNs. However, the performance of this algorithm was not evaluated and it was only applied to a small network. So, in this chapter, we will improve the algorithm and evaluate it. For the evaluation of the proposed algorithm, stochastic simulation algorithm which is introduced in Chapter 2 is employed to generate simulation data. For a large-scale network application, 1136 genes are taken from the result of the brain tumor analysis in Chapter 9 and their abnormal pathway activations are identified by the stochastic modeling of a human gene regulatory network structure. Though the reverse engineering method in the previous chapter provides an efficient way to build a network structure, many of the estimated edges are still remained as a hypothesis not confirmed by experiments. So the human gene network structure used in this chapter is built only based on KEGG pathway information [87].

10.1 Improving Abnormal Pathway Detection Algorithm

In this section, we modify the abnormal pathway detection algorithm introduced in Chapter 5 and compare its performance with the previous algorithm. In the previous version, it estimates one more transition probability (Equation (2.36)) from one steady-state observation, which may cause low estimation accuracy. So instead of the one steady-state observation which is the mean of the gene expression values of all samples, all the measured values are taken to be used in an optimization algorithm with simple bounds on variables.

Let's revisit the solution (2.36) of our network model,

$$q_i = \frac{\lambda_i^+ + \sum_{j \in I} q_j \mu_j p_{ji}^+}{\mu_i + \lambda_i^- + \sum_{j \in I} q_j \mu_j p_{ji}^-} \tag{10.1}$$

where I is the set of genes that affects the ith gene. The previous algorithm firstly finds λ_i^+, λ_i^-, and μ_i by fixing the transition probabilities, $p_{ij} = \{p_{ij}^+, p_{ij}^-\}$, with the steady-state probabilities, $q_1, ..., q_i, ..., q_j$, of normal observations, a_i which is the average of mRNA level of the ith gene, $a_i = 1/m \sum_{l=1}^{m} x_{il}$ where m is the total number of samples. Then the algorithm tries to estimate p_{ij} with the inferred parameters and the steady-state probabilities of abnormal data, a_i'. However, this estimation could be only reliable when the number of unknown transition probabilities is one because we have only one observed data, a_i'. In order to improve this process, the new algorithm uses the whole observed data, x_{il}', instead of their average a_i'.

Let q_{il}' be the steady-state probability of the ith gene in the lth sample in *abnormal* conditions and $\mathbf{p}_j'$ be the corresponding transition probabilities where $j \in \{$ indices of genes that affect to the ith gene $\}$. Our aim is to estimate the probabilities $\mathbf{p}_j'$ from the observation $\mathbf{q}_{jl}' = \{q_{jl}'\}$ under the assumption that all the other parameters are the same with those of normal conditions ($\hat{\lambda}_i^+, \hat{\lambda}_i^-$, and $\hat{\mu}_i$). Then our problem becomes to minimize following function of $\mathbf{p}_j'$

$$\sum_l \left(q_{il}' - f_i(\mathbf{q}_{jl}', \hat{\lambda}_i^+, \hat{\lambda}_i^-, \hat{\mu}_i; \mathbf{p}_j') \right)^2 \tag{10.2}$$

with constraints $0 \leq \mathbf{p}_j'$ and $\sum_j \mathbf{p}_j' \leq 1$. To solve this problem, we used a R software "optim" with "L-RBFS-B" algorithm which solves linear/non-linear optimization problem with bound constraints. The significance of these estimated $\mathbf{p}_j'$ will be obtained using the same permutation tests introduced in Chapter 5.

10.2 Evaluation of the Pathway Detection Algorithm

This section describes the evaluation processes of the pathway detection algorithm including the details of simulation data generation. We firstly built a scale-free network structure using R package "igraph". Gillespie algorithm which is one of the well known stochastic simulation algorithm is applied to build appropriate gene expression data from the given network structure. Three values of sensitivity (SENS), specificity (SPEC), and positive predictive value (PPV) were measured for the comparison of the performance between the improved algorithm and the previous one.

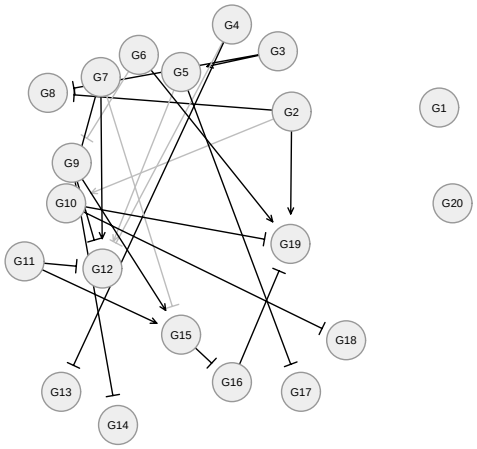


Figure 10.1: Randomly generated network structure with 20 nodes. In abnormal conditions, gray colored pathways are less activated than normal conditions.

10.2.1 Example: 20 node gene network

Figure 10.1 shows a randomly generated scale-free network structure consisting of 20 genes. This simulation assumed that each gene affects the other genes directly, which means it has straightforward gene regulatory processes. Their reaction equations are shown in Equation (10.3) $\sim$ (10.6).

$$DNA_i \xrightarrow{\lambda} DNA_i + G_i \tag{10.3}$$

$$G_i \xrightarrow{ac} G_i + G_j \tag{10.4}$$

$$G_i + G_j \xrightarrow{dc} G_i \tag{10.5}$$

$$G_i \xrightarrow{\mu} \phi \tag{10.6}$$

where DNA_i and G_i indicate the copy numbers of the ith DNA and gene, respectively, and $i, j = \{1, 2, ..., 20\}$. The default values of reaction constants were $\lambda = 1, \mu = 0.3, ac = 0.25$, and $dc = 0.075$ throughout this simulation study. Equation (10.3) shows the ith gene is expressed by its corresponding DNA. Equation (10.4) and (10.5) represent the excitation and inhibition of the ith gene that affects the jth gene. Equation (10.6) is for the degradation of the ith gene. Gillespie algorithm was applied to generate the expression data from the the reaction equations. Figure 10.2 shows the expression profile of each gene. Because the five gray colored pathways ($G6 \dashrightarrow G9$, $G2 \rightarrow G10$, $G4 \dashrightarrow G12$, $G5 \rightarrow G12$, and

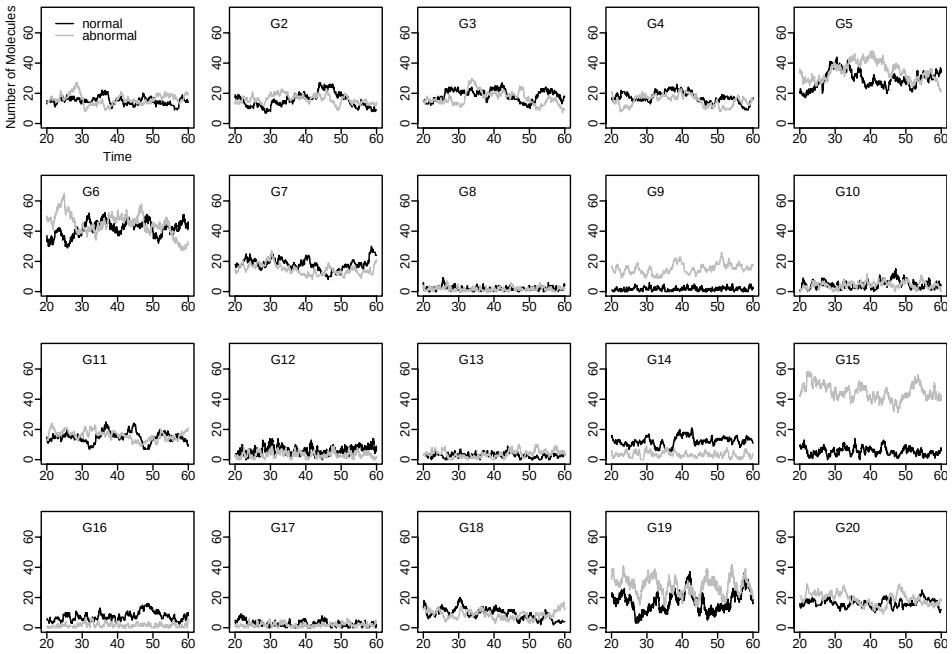


Figure 10.2: Expression profiles of the 20 genes for 60 second. Black and gray lines indicate the expression levels in the normal and abnormal conditions, respectively. In the abnormal condition, the reaction rates of the five gray edges in Figure 10.1 are zero (knockout).

$G7 \dashrightarrow G15$) of the network in Figure 10.1 are set to be inactivated in the abnormal condition, the expression patterns of $G9$, $G10$, and $G15$ are different from that of the normal condition. In particular, $G15$ expression considerably bursted out because the inhibitory effect from $G7$ is weakened and the activity of $G9$ is enhanced. This over-expressed effect of $G15$ resulted in $G16$ and $G19$ being down and up regulated, respectively.

The Gillespie algorithm generated approximately 12,000 time points for 60 seconds (data from 0 to 10 seconds were excluded as burning period) in each of normal and abnormal conditions. Among these time points we randomly selected 30 points to make the expression dataset and identified the abnormal pathways using the proposed algorithm where the number of permutations was 500. There are total 24 edges and 5 abnormal edges in this network. With a criterion $\alpha = 0.001$, the sensitivity is $5/5 = 1$, the specificity is $13/19 = 0.684$, and the positive predictive value is $5/(5 + 6) = 0.454$. The detailed results are represented in Table 10.1.

Table 10.1: The result of abnormal activity detection in the 20 gene network. First column shows the edges where solid and dashed arrows are excitatory and inhibitory effects, respectively. It also represents the transition probabilities (Equation (2.36)), the mean probabilities of the permuted probabilities and the empirical p-values (Equation (5.2)) of the permutation test. The true abnormal edges are shown in bold text.

Edges	Normal prob.	Mean perm prob.	p-values of
$G3 \rightarrow G5$	0.333	0.361	0.004
$G5 \rightarrow G6$	0.250	0.252	0.310
$G2 \rightarrow G8$	0.250	0.295	0.364
$G3 \dashrightarrow G8$	0.333	0.057	0.068
$G6 \dashrightarrow G9$	**0.333**	**0.000**	**0.000**
$G2 \rightarrow G10$	**0.250**	**0.080**	**0.000**
$G7 \dashrightarrow G10$	0.250	0.000	0.000
$G4 \dashrightarrow G12$	**0.333**	**0.000**	**0.000**
$G5 \rightarrow G12$	**0.250**	**0.000**	**0.000**
$G7 \rightarrow G12$	0.250	0.097	0.024
$G9 \dashrightarrow G12$	0.250	0.097	0.136
$G11 \dashrightarrow G12$	0.333	0.000	0.000
$G4 \dashrightarrow G13$	0.333	0.001	0.000
$G9 \dashrightarrow G14$	0.250	0.966	0.000
$G7 \dashrightarrow G15$	**0.250**	**0.000**	**0.000**
$G9 \rightarrow G15$	0.250	1.000	0.000
$G11 \rightarrow G15$	0.333	0.246	0.082
$G15 \dashrightarrow G16$	0.500	0.953	0.000
$G5 \dashrightarrow G17$	0.250	0.138	0.126
$G10 \dashrightarrow G18$	0.333	0.379	0.318
$G2 \rightarrow G19$	0.250	0.177	0.296
$G6 \rightarrow G19$	0.333	0.411	0.310
$G10 \dashrightarrow G19$	0.333	0.113	0.132
$G16 \dashrightarrow G19$	0.500	0.575	0.358

10.2.2 Performance comparison

Total 15 datasets were generated and five of them were prepared with 20 gene networks having 5 abnormal edges. Another five datasets were simulated using 100 gene networks with 5 abnormal edges and the other five have 100 genes with 10 abnormal edges. Each dataset consisted of normal and abnormal data whose reaction rates were the same as the previous 20 node example. The three evaluation values (SENS, SPEC, and ACC) were measured for the performance comparison. In this result, the improved algorithm shows better performances than the original one when its network has 20 nodes. However, as the number of genes increases, PPV dramatically decreases though all Sens and Spec have approximately 0.7 and

Table 10.2: Performance comparison between improved abnormal pathway detection algorithm and the previous one. "New" and "Org" represent the improved and original algorithms, respectively.

			New			Org		
# genes	# edges	# true abnormal edges	Sens	Spec	PPV	Sens	Spec	PPV
20	112	5	0.6	0.727	0.500	0.8	0.636	0.500
20	140	5	0.6	0.866	0.600	0.2	0.666	0.166
20	273	5	1.0	0.411	0.200	0.4	0.470	0.100
20	98	5	1.0	0.333	0.454	0.8	0.777	0.666
20	105	5	1.0	0.900	0.833	1.0	0.900	0.833
	Mean		0.84	0.647	0.517	0.64	0.690	0.450
100	994	5	1.0	0.656	0.096	0.6	0.854	0.130
100	735	5	0.8	0.600	0.090	0.4	0.770	0.080
100	630	5	0.6	0.717	0.111	0.6	0.835	0.176
100	1309	5	0.8	0.576	0.049	0.6	0.489	0.031
100	1393	5	0.6	0.520	0.031	0.4	0.835	0.058
	Mean		0.76	0.614	0.075	0.52	0.756	0.095
100	938	10	0.3	0.580	0.054	0.1	0.782	0.035
100	1141	10	0.8	0.673	0.137	0.4	0.790	0.111
100	1190	10	0.6	0.612	0.088	0.5	0.606	0.073
100	889	10	1.0	0.572	0.166	0.6	0.623	0.120
100	1134	10	0.7	0.565	0.095	0.5	0.611	0.078
	Mean		0.68	0.600	0.108	0.420	0.683	0.083

0.6, which is mainly because the number of edges grows much faster than the number of genes. On the other hand, it reveals that the method seems to show robust performance against the number of abnormal edges.

10.3 Abnormal Pathway Detection of Brain Tumor Data

Now we apply the new algorithm to the brain tumor datasets which are introduced in Chapter 9. As mentioned before, this study used known pathway information to construct the network structure instead of estimating it using reverse engineering techniques in Chapter 7. So human non-metabolic pathways were taken from KEGG database and 1339 gene network structure was constructed. Among the 1339 genes, we again took 1136 genes which appeared in the 7582 genes selected from Meta-analysis in Chapter 9.

Figure 10.3 shows the human gene network with 1136 genes which are involved in 122 KEGG pathways such as cell cycle and p53 signaling pathways. However, this network plot has a limitation to know which genes are associated with which KEGG pathways. So we represent our results with a different network structure shown in Figure 10.4. Top layer of the figure represent the 122 KEGG pathways each of which is related with the genes on the middle layer. Basically the genes in the middle layer are the same with that of bottom layer so all the directed edges

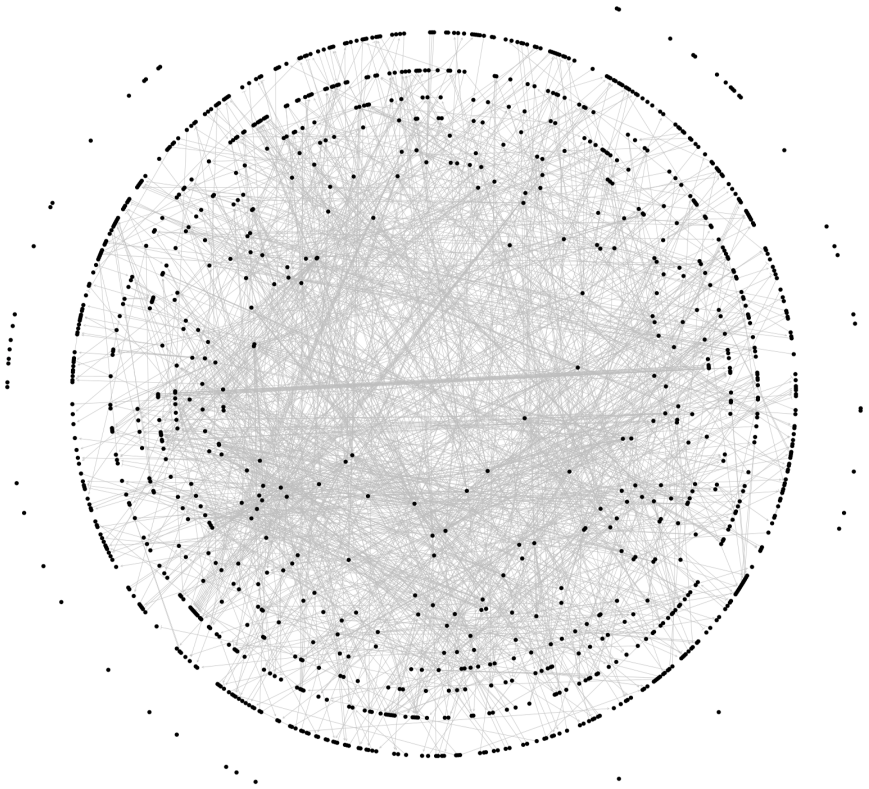


Figure 10.3: Human non-metabolic gene network containing 1136 genes from KEGG database. Genes closed to the center indicate that they have higher degrees than the other genes which are located in the outside of the network structure. The genes are placed in the order of the chromosomal position (in clockwise order from 9 o'clock). Note that this figure is only for the visualization purposes.

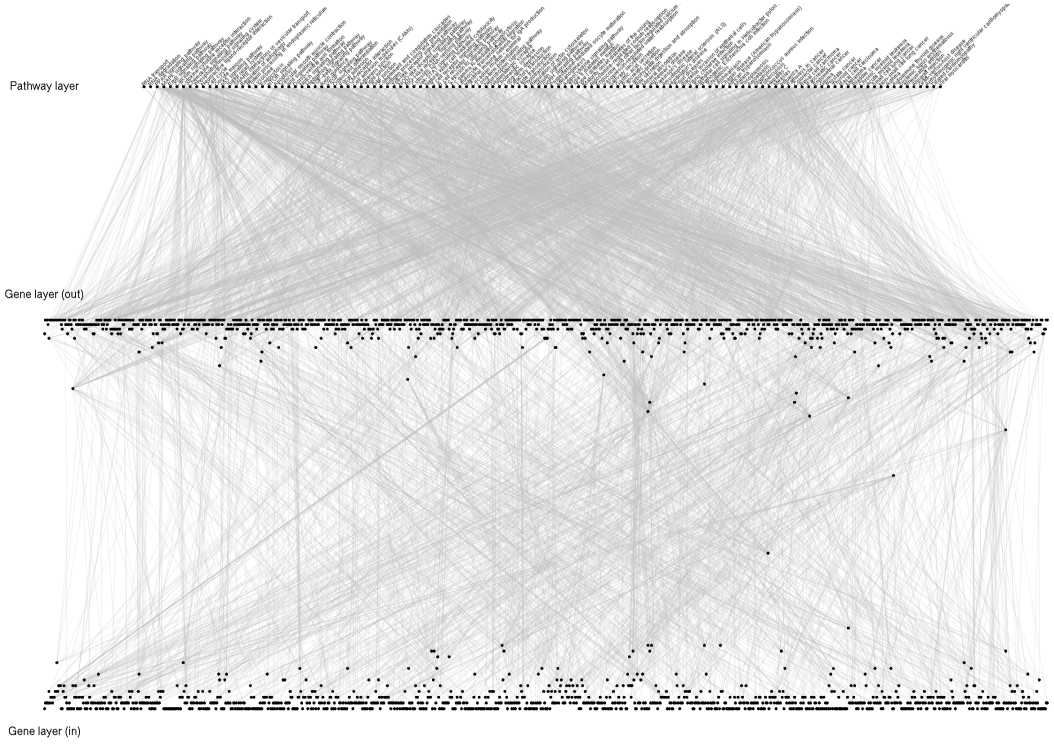


Figure 10.4: The human non-metabolic gene network with 1136 genes and their involved 122 pathways from KEGG database. Top layer represents the 122 KEGG pathways each of which is connected to its related genes on the middle layer. The middle and bottom layers contain the identical genes but the directed edges are linked from the genes in the middle layer to those of the bottom layer. So, in the middle layer, if a gene has higher out-degrees than other genes, the gene is located at lower position than the others. On the other hand, in the bottom layer, genes having higher in-degrees are depicted higher position than the others. The genes are placed in the order of chromosomal order from the left to right.

are started from the middle layer to the bottom layer. Please see the caption of the figure for more detailed description.

We applied the algorithm to the normal and high grade brain tumor data and found significantly activated or inactivated pathways which are shown in Figure 10.5. Though this figure provides a systematic view of the system, it is not clear which KEGG pathways are involved in the abnormal pathways identified by the

proposed algorithm. So a new measure IKp_i is introduced as follows. Let Kp_i be the ith KEGG pathway and Ap denote an abnormal pathway detected by our method. In order to determine which Kp is more involved in the detected abnormal pathways, we introduced a measure called "Importance of Kp_i" (IKp_i) as follows,

$$IKp_i = \frac{\text{\# of genes } \in \{Ap \cap Kp_i\}}{\text{\# of genes } \in Kp_i} \tag{10.7}$$

That is the importance of Kp_i is a proportion of the number of genes which belong to significantly activated/inactivated pathways in the ith KEGG pathway and total number of genes in the ith KEGG pathway. So a KEGG pathway having higher IKp means it could be the best biological process to describe the data we used. Based on this IKp, we sorted the KEGG pathways and it turned out that most of highly ranked pathways are related with cancer or brain diseases (Bold text in Table 10.3). Note that a highly ranked KEGG pathway is depicted larger and darker than the others in Figure 10.5.

10.4 Conclusions

In this chapter, it is shown that the improved abnormal pathway detection algorithm outperforms the previous one in terms of sensitivity and computation time. In the real large-scale network example with brain tumor data and KEGG human network structures, we showed that the significantly activated/inactivated pathways are related with KEGG cancer pathways such as Bladder cancer, Melanoma, and Thyroid cancer. Though the algorithm suffers from the high false positive rate (= 1-specificity) especially when the number of genes is large, our approach is shown to be available to a large-scale network which is difficult to be modeled with conventional stochastic modeling approaches. Further researches for having low false positives and mining biologically meaningful evidences from the results will be required for a successful application.

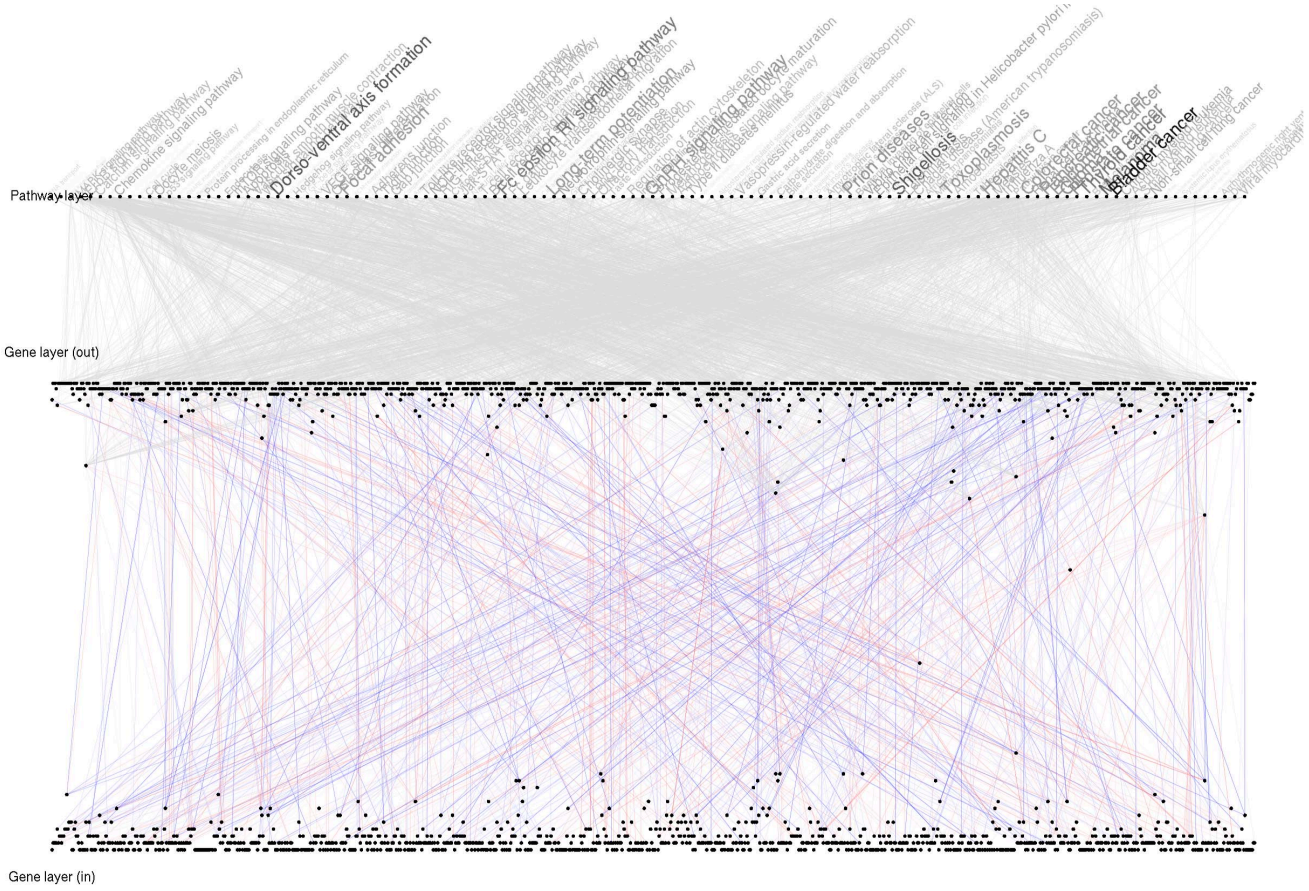


Figure 10.5: Identified abnormal interactions using normal and high grade tumor data. Red and blue lines represent significantly activated and inactivated pathways, respectively. At the top layer, KEGG pathways which are involved in more number of significantly detected abnormal pathways are shown bigger and darker than the others.

Table 10.3: Top 20 KEGG pathways sorted by IKp (second column). The third column represents the number of involved genes in the corresponding KEGG pathway. 11 (Bold text) out of 20 are all related with cancer or brain disease.

Pathway names	PW	Number of genes
Bladder cancer	6.683	13
Dorso-ventral axis formation	5.933	6
Shigellosis	5.864	18
Melanoma	5.747	17
Thyroid cancer	5.225	8
Fc epsilon RI signaling pathway	5.183	26
Glioma	4.638	28
Prostate cancer	4.614	35
Prion diseases	4.463	11
GnRH signaling pathway	4.457	29
Pancreatic cancer	4.422	33
Toxoplasmosis	4.390	31
Hepatitis C	4.350	44
Long-term potentiation	4.300	21
Endometrial cancer	4.290	21
Focal adhesion	4.266	52
Colorectal cancer	4.242	21
Neurotrophin signaling pathway	4.137	57
Vibrio cholerae infection	4.100	8
Epithelial cell signaling in Helicobacter pylori infection	4.100	23

11 Conclusions and Perspectives

Thanks to the development of measurement devices and technology in biology, a vast amount of biological data is being generated everyday. This advancement pushes researchers toward the genome wide view of a system and appropriate tools which can handle the large amount of data are required. In this dissertation, we have shown statistical methods available for large-scale gene regulatory networks. In particular, G-networks in the stochastic modeling approaches provide a closed form solution in the steady-states of a system so we can model large-scale GRNs and detect abnormal genes/pathways of the networks. Additionally, we proposed a new reverse engineering approach to estimate large-scale GRN structure and showed a gene mining approach to identify disease candidate genes based on the large-scale GRNs. Brain tumor datasets preprocessed by conventional bioinformatics approaches were successfully analyzed by the proposed large-scale techniques and found reliable Gene Ontology evidences for the identified genes. In the remainder of this chapter, we will summarize the thesis contributions and discuss the directions for future work.

11.1 Thesis Contributions

This thesis can contribute to further biological researches in terms of following three directions.

- **Dynamics analysis of a large-scale GRN using the probabilistic modeling (Bottom-up) approach** - Since the nature of the gene regulation processes is based on discrete and probabilistic, probabilistic models are more appropriate than ODE models for describing biological systems. However, the usage of the conventional probabilistic modeling approaches with the Gillespie algorithm is limited by their computational difficulty when the number of involved molecules is large. Our G-network based modeling approach enables us to model the large-scale GRNs and to investigate the dynamical behaviors of the network structure with respect to their given

conditions (Chapter 4, 5). As an example, we showed the brain tumor patients have different activated/inactivated pathways depending on their tumor grades (Chapter 10).

- **Reverse engineering (Top-down) for the large-scale GRN construction** - As the GRNs play an important role in identifying the cellular functions of genes, it is essential to estimate the regulatory structures among molecules by observing their expression patterns. Here we proposed a new reverse engineering method which can efficiently be applicable to the large-scale dataset and integrate various sources of data. This approach can provide functional evidences of the regulatory molecules (Chapter 7). Also it could be a good initial structure for the system dynamics investigation via the stochastic modeling approaches.

- **Framework for GRN analysis with large and heterogeneous biological data** - Along with the rapid growth of heterogeneous biological data, numerous analysis methods have been developed in the fields of bioinformatics and systems biology. So, a successful GRN analysis could be achieved by applying these methods to appropriate biological problems and datasets. In this sense, our series of analyses ranging from raw data preprocess to the dynamics modeling can provide the guideline of a proper GRN analysis. More specifically, starting from raw datasets of brain tumor, we performed Meta-analysis with data preprocessing and gene ontology based clustering to identify differentially expressed genes and to infer their functions in the target disease (Chapter 9). Based on these processed gene expression values along with other publically available data, the proposed reverse engineering technique build gene networks (Chapter 7) and found the disease candidate genes (Chapter 8). Then finally the regulatory structure can be used to model its dynamics which can detect abnormal gene/pathway activities (Chapter 4 and 5).

11.2 Applications for Synthetic Biology

Synthetic biology as an emerging field of biotechnology with a concept of "engineering of living systems" has been developed in the last decade and shows promising solutions to health, energy and environmental problems which might be difficult to overcome with the traditional approaches. Though synthetic biology

requires expertise from diverse disciplines, modeling and inference are becoming important features for designing a new synthetic system as the system is getting complex and more biological parts are available [117, 122]. Therefore the techniques in our thesis could be useful for synthetic biology researches in terms of following two perspectives.

- **Iterative design of a synthetic system** - Building a successful synthetic system is accompanied with good mathematical models which can fully predict the behavior of the system. In a simple system, finding appropriate model parameters would be enough for explaining the system's behavior. However, as many biological devices are available, reverse engineering methods need to be used not only in determining which combination of the devices leads to the optimal model but also in probing undesirable interactions among involved molecules. Currently, a synthetic system is designed by iteratively performing the mathematical model optimization and in vivo experiments [26]. So, a combinatorial approach of our bottom-up and top-down techniques can be useful to design a robust and reliable synthetic system.

- **Long-term monitoring tool for a system** - As synthetic biology develops, ethical and safety issues are emerging in public society. However, the advance of the technologies is inevitable because of their promising products for clean energy and medical solutions. So it is essential to know how much an introduced circuit affects to the behavior of a cell throughout its generations. Answering this question via stochastic modeling approaches is difficult due to the complexity and large-scale of a living system. Our approaches which predict the long-run behavior of the complex system and infer regulatory structures from observations can be used to reveal the behavioral and evolutional properties of the synthetic system. These assessment and monitoring methods would be essential for the safe development of synthetic biology.

11.3 Discussions and Future Works

This thesis has shown the possibilities that modeling and reverse engineering approaches can provide useful information explaining the diseases- related genes and pathways. However, we still have some more issues to improve for future work as follows.

- Our stochastic modeling approach has three types of unknown parameters such as input, output rates of a signal, steady-state probabilities and signal transition probabilities. However, estimating these parameters with limited observations would be difficult in terms of the global optimum. So it would be necessary to know the actual generation/degradation rate of mRNA and protein under a specific condition via real experiments. Alternatively, it is possible to estimate this dynamics information of molecules from their static data by developing a new method where additional knowledge can be used together.

- Though Chapter 6 shows our model can be extended by dividing two layers, it is not a completely analytical approach because it needs a numerical optimization process in synchronizing the states of DNA and protein layers. Therefore, it will be necessary to evaluate the optimization process as well as to find the analytical solution by modeling the two layers together. On the other hand, along with these DNA and protein layers, our model can be further extended by adding a metabolic layer where various metabolites are conveniently observable but hard to be analyzed. This model should be able to explain reverse reactions for the metabolic layer modeling.

- Many network modules in a living cell shows oscillatory behaviors which play an important role in various dynamic cellular processes such as the circadian clock and the cell cycle. However, this thesis has focused on explaining the steady-state behavior of a system. The G-network approach is not applicable for investigating the oscillatory expressions of a system. However, a large-scale system could be narrowed down into small network modules by the G-network modeling and the detailed dynamics of the identified modules could be addressed by the conventional stochastic modeling technique such as the Gillespie algorithm. Also an analytical study for finding the solution of the non steady-state system such as a quasi-stable system was introduced by N. Strelkowa and M. Barahona [111].

- Our stochastic model detecting the abnormal activities of genes/pathways needs to be assessed in their robustness. For example, if one edge were removed or added to the p53 network then would it give the same conclusions? Under an assumption that no network model in biology is perfect, it is essential to evaluate this question. Once this structural robustness analysis has

been done, sensitivity analysis of the parameters or genes could provide the key molecules responsible for the phenotype.

- The proposed reverse engineering method provides an efficient way to summarize various biological evidences using the Gibbs distribution prior. However, this approach is simply a weighted sum of the information regardless of how the related components systematically interact with each other. Obtaining the weight of each prior by considering the biological relationships of the data would be one possible approach for the better estimation. As a model is getting more complex, it would be becoming harder to apply the model to large-scale datasets. However, we can take an advantage of the advancement in computer hardware and software tools such as parallel computing with GPUs.

- Finally, our applications are rather focused on finding a way which works for the real biological data. In order to be a reliable and practical framework of GRN analysis, we should have more real applications reflecting the genuine meanings of biological problems. Also, as we discussed in the previous subsection, cooperating with the study of bioengineered real organisms in synthetic biology would be helpful to address more complex biological problems.

Bibliography

[1] P. Ao. Global view of bionetwork dynamics: adaptive landscape. *Journal of Genetics and Genomics*, 36(2):63–73, 2009.

[2] A. Arazi, E. Ben-Jacob, and U. Yechiali. Bridging genetic networks and queueing theory. *Physica A: Statistical Mechanics and its Applications*, 332:585–616, 2004.

[3] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[4] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic acids research*, 35(Database issue):D760, 2007.

[5] D.I.M.C.D. BDD and N. Lenaugasse. Stationary Queueing Models with Aspects of Customer Impatience and Retrial Behaviour.

[6] T. Beissbarth and T. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, page 881, 2004.

[7] T. Beißbarth and T.P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[8] A.C. Bellail, S.B. Hunter, D.J. Brat, C. Tan, and E.G. Van Meir. Microregional extracellular matrix heterogeneity in brain modulates glioma cell invasion. *The international journal of biochemistry & cell biology*, 36(6):1046–1069, 2004.

[9] J. Berg and M. Lässig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41):14689, 2004.

[10] J. Berg and M. Lässig. Cross-species analysis of biological networks by bayesian alignment. *Proceedings of the National Academy of Sciences*, 103(29):10967–10972, 2006.

[11] A. Bernard and A.J. Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. In *Pac Symp Biocomput*, volume 10, pages 459–470. Citeseer, 2005.

[12] J. Bloom and F.R. Cross. Multiple levels of cyclin specificity in cell-cycle control. *Nature Reviews Molecular Cell Biology*, 8(2):149–160, 2007.

[13] N. Bolshakova, F. Azuaje, and P. Cunningham. Incorporating biological domain knowledge into cluster validity assessment. *Applications of Evolutionary Computing*, pages 13–22, 2006.

[14] M. Boutros, H. Agaisse, and N. Perrimon. Sequential activation of signaling pathways during innate immune responses in Drosophila. *Developmental Cell*, 3(5):711–722, 2002.

[15] J.M. Bower and H. Bolouri. *Computational modeling of genetic and biochemical networks*. The MIT Press, 2004.

[16] D. Bratsun, D. Volfson, L.S. Tsimring, and J. Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences*, 102(41):14593, 2005.

[17] D. Bratsun, D. Volfson, L.S. Tsimring, and J. Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences*, 102(41):14593–14598, 2005.

[18] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1):83–92, 2004.

[19] C.J. Brown, S. Lain, C.S. Verma, A.R. Fersht, and D.P. Lane. Awakening guardian angels: drugging the p53 pathway. *Nature Reviews Cancer*, 9(12):862–873, 2009.

[20] N.E. Buchler, U. Gerland, and T. Hwa. Nonlinear protein degradation and the function of genetic circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9559, 2005.

[21] N.E. Buchler, U. Gerland, and T. Hwa. Nonlinear protein degradation and the function of genetic circuits. *Proceedings of the National Academy of Sciences*, 102(27):9559–9564, 2005.

[22] H.E. Bülow, K.L. Berry, L.H. Topper, E. Peles, and O. Hobert. Heparan sulfate proteoglycan-dependent induction of axon branching and axon misrouting by the kallmann syndrome gene kal-1. *Proceedings of the National Academy of Sciences*, 99(9):6346, 2002.

[23] H.E. Bülow and O. Hobert. Differential sulfations and epimerization define heparan sulfate specificity in nervous system development. *Neuron*, 41(5):723–736, 2004.

[24] Z. Burda, A. Krzywicki, OC Martin, and M. Zagorski. Motifs emerge from function in model gene regulatory networks. *Proceedings of the National Academy of Sciences*, 108(42):17263–17268, 2011.

[25] L. Cai, N. Friedman, and X.S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[26] I. Cantone, L. Marucci, F. Iorio, M.A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. Di Bernardo, D. Di Bernardo, and M.P. Cosma. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172–181, 2009.

[27] M.S. Carro, W.K. Lim, M.J. Alvarez, R.J. Bollo, X. Zhao, E.Y. Snyder, E.P. Sulman, S.L. Anne, F. Doetsch, H. Colman, et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, 2009.

[28] J. Chahine, R.J. Oliveira, V.B.P. Leite, and J. Wang. Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proceedings of the National Academy of Sciences*, 104(37):14646, 2007.

[29] W.K. Ching, S. Zhang, M.K. Ng, and T. Akutsu. An approximation method for solving the steady-state probability distribution of probabilistic Boolean networks. *Bioinformatics*, 23(12):1511, 2007.

[30] C. Choy and S.H. Kim. Biological actions and interactions of anosmin-1. 2010.

[31] B.T.S. Da Wei Huang, R.A. Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2008.

[32] H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

[33] R. Edgar, M. Domrachev, and A.E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207, 2002.

[34] A. Eldar and M.B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.

[35] M.B. Elowitz, S. Leibler, et al. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[36] J. Ernst, H.L. Plasterer, I. Simon, and Z. Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome research*, 20(4):526, 2010.

[37] D.H. Erwin and E.H. Davidson. The evolution of hierarchical gene regulatory networks. *Nature Reviews Genetics*, 10(2):141–148, 2009.

[38] G.R. Fink, P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[39] J. Flannick, A. Novak, B.S. Srinivasan, H.H. McAdams, and S. Batzoglou. Graemlin: general and robust alignment of multiple large interaction networks. *Genome research*, 16(9):1169–1181, 2006.

[40] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[41] J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *Computers, IEEE Transactions on*, 100(9):881–890, 1974.

[42] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[43] T.S. Gardner, C.R. Cantor, and J.J. Collins. Construction of a genetic toggle switch inEscherichia coli. *Nature*, 403:339–342, 2000.

[44] D. Geiger and D. Heckerman. Learning gaussian networks. Technical report, Technical report, Microsoft Re, 1994.

[45] E. Gelenb e. Network of interacting synthetic molecules in equilibrium. *Proc. Royal Society A (Mathematical and Physical Sciences)*, 464:2219–2228, 2008.

[46] E. Gelenb e. Search in unknown random environments. *Physical Review E*, 82(6):061112, 2010.

[47] E. Gelenbe. Product-form queueing networks with negative and positive customers. *Journal of Applied Probability*, pages 656–663, 1991.

[48] E. Gelenbe. G-networks with triggered customer movement. *Journal of Applied Probability*, pages 742–748, 1993.

[49] E. Gelenbe. Genetic algorithms with analytical solution. *Robotics and Autonomous Systems*, 22:59–64, 1997.

[50] E. Gelenbe. Steady-state solution of probabilistic gene regulatory networks. *J. Theor. Biol Phys Rev E*, 76:031903, 2007.

[51] E. Gelenbe and J.M. Fourneau. G-networks with resets. *Performance Evaluation*, 49(1-4):179–191, 2002.

[52] E. Gelenbe and A. Labed. G-networks with multiple classes of signals and positive customers. *European journal of operational research*, 108(2):293–305, 1998.

[53] E. Gelenbe and S. Timotheou. Random neural networks with synchronized interactions. *Neural Computation*, 20(9):2308–2324, 2008.

[54] D.T. Gillespie. Stochastic simulation of chemical kinetics. 2007.

[55] D.T. Gillespie et al. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[56] DV Goeddel, DG Yansura, and MH Caruthers. Binding of synthetic lactose operator DNAs to lactose repressors. *Proceedings of the National Academy of Sciences*, 74(8):3292–3296, 1977.

[57] I. Golding, J. Paulsson, S.M. Zawilski, and E.C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[58] I. Golding, J. Paulsson, S.M. Zawilski, and E.C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[59] D. González-Martínez, S.H. Kim, Y. Hu, S. Guimond, J. Schofield, P. Winyard, G.B. Vannelli, J. Turnbull, and P.M. Bouloux. Anosmin-1 modulates fibroblast growth factor receptor 1 signaling in human gonadotropin-releasing hormone olfactory neuroblasts through a heparan sulfate-dependent mechanism. *The Journal of neuroscience*, 24(46):10384, 2004.

[60] M. Gustafsson, M. Hornquist, and A. Lombardi. Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(3):261, 2005.

[61] Y. Haupt, R. Maya, A. Kazaz, and M. Oren. Mdm2 promotes the rapid degradation of p53. *Nature Reviews Cancer*, 1997.

[62] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models–A review. *Biosystems*, 96(1):86–103, 2009.

[63] O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, D.S. Charnock-Jones, et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, 24(7):932, 2008.

[64] F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374, 2008.

[65] S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using Bayesian networks and non-

parametric regression. In *Pacific Symposium on Biocomputing*, volume 7, pages 175–186, 2002.

[66] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27, 2000.

[67] H. Kim and E. Gelenbe. Anomaly detection in gene expression via stochastic models of gene regulatory networks. *BMC Genomics*, 10(Suppl):S26, 2009.

[68] H. Kim and E. Gelenbe. Stochastic Gene Expression Modeling with Hill Function for Switch-like Gene Responses. *In Proceedings of International Conference on Bioinformatics Biomedicine*, 2010.

[69] H. Kim and E. Gelenbe. G-network based two layer stochastic modeling of gene regulatory networks with post-translational processes. *Interdisciplinary Bio Central*, 3:8–12, 2011.

[70] H. Kim, J. Lee, and T. Park. Inference of Large Scale Gene Regulatory Networks Using Regressionbased Approach. *Journal of Bioinformatics and Computational Biology*, 7(4):717–35, 2009.

[71] Haseong Kim, Jae Lee, and Taesung Park. Inference of Large Scale Gene Regulatory Networks Using Regressionbased Approach. *Journal of Bioinformatics and Computational Biology*, page Inpress, 2009.

[72] S. Köhler, S. Bauer, D. Horn, and P.N. Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[73] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, 7(50):1341–1354, 2010.

[74] J.W. Little and M. Gellert. The SOS regulatory system: control of its state by the level of RecA protease. *Journal of Molecular Biology*, 167(4):791–808, 1983.

[75] H. Maamar, A. Raj, and D. Dubnau. Noise in gene expression determines cell fate in Bacillus subtilis. *Science*, 317(5837):526, 2007.

[76] D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

[77] N.J. Marianayagam, M. Sunde, and J.M. Matthews. The power of two: protein dimerization in biology. *Trends in Biochemical Sciences*, 29(11):618–625, 2004.

[78] F. Markowetz, D. Kostka, O.G. Troyanskaya, and R. Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305, 2007.

[79] CJ Marshall et al. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell*, 80(2):179–186, 1995.

[80] S. Martin, Z. Zhang, A. Martino, and J.L. Faulon. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, 23(7):866, 2007.

[81] H.H. McAdams and A. Arkin. Stochastic mechanisms in gene expression, 1997.

[82] S. Mehra, S. Charaniya, E. Takano, and W.S. Hu. A Bistable Gene Switch for Antibiotic Biosynthesis: The Butyrolactone Regulon in Streptomyces coelicolor. *PLoS One*, 3(7), 2008.

[83] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science's STKE*, 298(5594):824, 2002.

[84] N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. In *Pacific Symposium on Biocomputing 2004: Hawaii, USA, 6-10 January 2004*, page 336. World Scientific Pub Co Inc, 2003.

[85] L. Neumann, C. Pforr, J. Beaudouin, A. Pappa, N. Fricker, P.H. Krammer, I.N. Lavrik, and R. Eils. Dynamics within the cd95 death-inducing signaling complex decide life and death of cells. *Molecular systems biology*, 6(1), 2010.

[86] D. Nevozhay, R.M. Adams, K.F. Murphy, K. Josić, and G. Balázsi. Negative autoregulation linearizes the dose–response and suppresses the heterogeneity of gene expression. *Proceedings of the National Academy of Sciences*, 106(13):5123, 2009.

[87] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29, 1999.

[88] K. Okubo, F. Sakai, E.L. Lau, G. Yoshizaki, Y. Takeuchi, K. Naruse, K. Aida, and Y. Nagahama. Forebrain gonadotropin-releasing hormone neuronal development: insights from transgenic medaka and the relevance to x-linked kallmann syndrome. *Endocrinology*, 147(3):1076, 2006.

[89] R. Opgen-Rhein and K. Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(1):37, 2007.

[90] R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC bioinformatics*, 8(Suppl 2):S3, 2007.

[91] R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC bioinformatics*, 8(Suppl 2):S3, 2007.

[92] D.A. Orlando, C.Y. Lin, A. Bernard, J.Y. Wang, J.E.S. Socolar, E.S. Iversen, A.J. Hartemink, and S.B. Haase. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, 453(7197):944–947, 2008.

[93] M. Párrizas, A.R. Saltiel, and D. LeRoith. Insulin-like growth factor 1 inhibits apoptosis using the phosphatidylinositol 3-kinase and mitogen-activated protein kinase pathways. *Journal of Biological Chemistry*, 272(1):154, 1997.

[94] J. Paulsson. Models of stochastic gene expression. *Physics of life reviews*, 2(2):157–175, 2005.

[95] J. Paulsson, O.G. Berg, and M. Ehrenberg. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97(13):7148, 2000.

[96] T.J. Perkins and P.S. Swain. Strategies for cellular decision-making. *Molecular Systems Biology*, 5(1), 2009.

[97] A.E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–163, 1995.

[98] A. Ribeiro, R. Zhu, and S.A. Kauffman. A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology*, 13(9):1630–1639, 2006.

[99] P.J. Rousseeuw and L. Kaufman. Finding groups in data, 1990.

[100] P.J. Rousseeuw, A.M. Leroy, and J. Wiley. *Robust regression and outlier detection*, volume 3. Wiley Online Library, 1987.

[101] J. Schafer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[102] L.J. Schedlich and L.D. Graham. Role of insulin-like growth factor binding protein-3 in breast cancer cell growth. *Microscopy research and technique*, 59(1):12–22, 2002.

[103] J. Schfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005.

[104] V. Shahrezaei and P.S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256, 2008.

[105] I. Shmulevich, I. Gluhovsky, R.F. Hashimoto, E.R. Dougherty, and W. Zhang. Steady-state analysis of genetic regulatory networks modelled by probabilistic Boolean networks. *Comparative and Functional Genomics*, 4(6):601–608, 2003.

[106] I. Shmulevich, I. Gluhovsky, R.F. Hashimoto, E.R. Dougherty, and W. Zhang. Steady-state analysis of genetic regulatory networks modelled

by probabilistic Boolean networks. *Comparative and Functional Genomics*, 4(6):601–608, 2003.

[107] P. Smolen, D.A. Baxter, and J.H. Byrne. Modeling circadian oscillations with interlocking positive and negative feedback loops. *Journal of Neuroscience*, 21(17):6644, 2001.

[108] K. Sneppen, S. Krishna, and S. Semsey. Simplified Models of Biological Networks. *Annual review of biophysics*, 2010.

[109] N. Soranzo, G. Bianconi, and C. Altafini. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, 23(13):1640, 2007.

[110] C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535, 2006.

[111] N. Strelkowa and M. Barahona. Switchable genetic oscillator operating in quasi-stable mode. *Journal of The Royal Society Interface*, 7(48):1071–1082, 2010.

[112] E. Takano, H. Kinoshita, V. Mersinias, G. Bucca, G. Hotchkiss, T. Nihira, C.P. Smith, M. Bibb, W. Wohlleben, and K. Chater. A bacterial hormone(the SCB 1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I polyketide biosynthetic gene cluster of Streptomyces coelicolor. *Molecular microbiology*, 56(2):465–479, 2005.

[113] E. Takano, T. Nihira, Y. Hara, J.J. Jones, C.J.L. Gershater, Y. Yamada, and M. Bibb. Purification and Structural Determination of SCB1, a γ-Butyrolactone That Elicits Antibiotic Production inStreptomyces coelicolor A3 (2). *Journal of Biological Chemistry*, 275(15):11010, 2000.

[114] A.L. Tarca, S. Draghici, P. Khatri, S.S. Hassan, P. Mittal, J. Kim, C.J. Kim, J.P. Kusanovic, and R. Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75, 2009.

[115] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614, 2001.

[116] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[117] M. Tigges, T.T. Marquez-Lago, J. Stelling, and M. Fussenegger. A tunable synthetic mammalian oscillator. *Nature*, 457(7227):309–312, 2009.

[118] A. Tresch and F. Markowetz. Structure learning in nested effects models. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

[119] M. Van der Laan, K. Pollard, and J. Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

[120] R. Varadhan and P.D. Gilbert. Bb: An r package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32(4):1–26, 2009.

[121] J. Wang, L. Xu, and E. Wang. Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. *Proceedings of the National Academy of Sciences*, 105(34):12271, 2008.

[122] W. Weber and M. Fussenegger. Molecular diversity–the toolbox for synthetic gene switches and networks. *Current Opinion in Chemical Biology*, 2011.

[123] R.A. Weinberg. The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–330, 1995.

[124] A.V. Werhli and D. Husmeier. Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *Journal of bioinformatics and computational biology*, 6(3):543–572, 2008.

[125] T. Werner. Bioinformatics applications for pathway analysis of microarray data. *Current opinion in biotechnology*, 19(1):50–54, 2008.

[126] D.J. Wilkinson. *Stochastic modelling for systems biology*. Chapman & Hall/CRC, 2006.

[127] F.X. Wu. Stability and bifurcation of ring-structured genetic regulatory networks with time delays. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, (99):1–9, 2011.

[128] R. Xulvi-Brunet and H. Li. Co-expression networks: graph properties and topological comparisons. *Bioinformatics*, 26(2):205–214, 2010.

[129] J. Zhu, B. Zhang, E.N. Smith, B. Drees, R.B. Brem, L. Kruglyak, R.E. Bumgarner, and E.E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, 40(7):854–861, 2008.

[130] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[131] M. Zou and S.D. Conzen. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71, 2005.