INTEGRATIVE APPROACHES FOR BIOLOGICAL NETWORK INFERENCES

by

DONGCHUL KIM

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2014

To my wife.

ACKNOWLEDGEMENTS

I would like to thank my supervising professor Dr. Jean Gao for her invaluable advice in my doctoral studies. She constantly motivated and encouraged me to achieve better progress in my research. I also wish to thank my academic advisors Dr. Hao Che, Dr. Ramez Elmasri, and Dr. Manfred Huber for their interest in my research and for taking time to serve in my dissertation committee.

July 18, 2014

ABSTRACT

INTEGRATIVE APPROACHES FOR BIOLOGICAL NETWORK INFERENCES

Dongchul Kim, Ph.D.

The University of Texas at Arlington, 2014

Supervising Professor: Jean X. Gao

Inferring biological networks from high-throughput bioinformatics data is one of the most interesting areas in the systems biology research in order to elucidate cellular and physiological mechanisms. In this thesis, network inference methods are proposed to solve biological problems.

We first investigated how the exposure to low dose ionizing radiation (IR) affects the human body by observing the signaling pathway associated with Ataxia Telangiectasia mutated using Reverse Phase Protein Array and isogenic human Ataxia Telangiectasia cells under different amounts and durations of IR exposure. DNA damage-caused pathways are derived from learning Bayesian networks in integration with prior knowledge such as Protein-Protein Interactions and signaling pathways from well-known databases. The experimental results show which proteins are involved in signaling pathways under IR, how the inferred pathways are different under low and high doses of IR, and how the selected proteins regulate each other in the inferred pathways.

In network inference research, there are two issues to solve. First, depending on the structural or computational model of inference method, the performance tends to

be inconsistent due to innately different advantages and limitations of the methods. Second, sparse linear regression that is penalized by the regularization parameter and bootstrapping-based sparse linear regression methods were suggested as state of the art in recent related works for network inference. However, they are not effective for a small sample size data and also a true regulator could be missed if the target gene is strongly affected by an indirect regulator with high correlation or another true regulator. To solve the limitations of bootstrapping, a lasso-based random feature selection algorithm is proposed to achieve better performance.

In order to elucidate the overall relationships between gene expressions and genetic perturbations, we propose a network inference method to infer gene regulatory network where Single Nucleotide Polymorphism (SNP) is involved as a regulator of genes. In the most of the network inferences named as SNP-Gene Regulatory Network (SGRN) inference, pairs of SNP-gene are given by separately performing expression Quantitative Trait Loci (eQTL) mappings. A SGRN inference method without predefined eQTL information is proposed assuming a gene is regulated by a single SNP at most.

We also studied how a medicine can be customized to individual patients considering biological features of the patients, i.e., Personalized Medicine. Our goal is to predict drug sensitivity levels of cancer patients in order to provide an optimal drug to the patients avoiding a waste of time with ineffective treatments. For the classification of patients to the optimal drug, we employed Bayesian Network Classifier (BNC) that consists of two components, parameters and network structure. Since the networks of BNC represent the dependency of proteins, these multiple networks of BNCs for multiple drugs also provide important information of relationships between proteins in order to identify the biomarkers of a target cancer from the integration of the multiple networks.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

Introduction

Most of biological process are operated by interactions of a number of biological components such as proteins and genes. Many scientists have carried out the research on the discovery of biological networks such as gene regulatory network and signal transduction in order to gain the insight of diseases for the development of possible therapies. Over the last few decades, high-throughput technologies such as gene microarray and protein microarray have been developed and they enable us to more effectively infer what relationships are there between genes and proteins. In this thesis, we propose computational methods to infer gene regulatory networks and protein signaling pathways for four different problems and applications as follows.

In chapter 2, DNA damage-related pathway was inferred to analyze the effect of low dose IR exposure to human body. To construct the signaling pathway, protein expression levels are measured by using Reverse Phase Protein Array under different amounts and durations of IR exposure In order to verify which proteins could be involved in a DNA damage-caused pathway, only proteins that highly interact with each other under IR are selected by using correlation coefficient. Then we performed the pathway inference that is derived from learning Bayesian networks in combination with prior knowledge such as Protein-Protein Interactions (PPIs) and signaling pathways from well-known databases. Learning Bayesian networks is based on a score and search scheme that provides the highest-scored network structure given a score function, and the prior knowledge is included in the score function as a prior probability by using Dempster-Shafer theory (DST). In this way, the inferred network can

be more likely to be similar to already discovered pathways and consistent with confirmed PPIs for more reliable inference. The experimental results show which proteins are involved in signaling pathways under IR, how the inferred pathways are different under low and high doses of IR, and how the selected proteins regulate each other in the inferred pathways. As our contribution, overall results confirm that low dose IR could cause DNA damage and thereby induce and affect related signaling pathways such as apoptosis, cell cycle, and DNA repair.

In chapter 3, a novel network inference was proposed by integrating two methods, which have different type of criterion. Many inference methods have been developed by using a variety of computational models and approaches such as Bayesian networks, information theory, regression model with bootstrapping, and so on. However, there are two issues to solve. First, depending on the structural or computational model of inference method, the results tend to be inconsistent due to innately different advantages and limitations of the methods. Therefore the combination of dissimilar approaches is demanded as an alternative way in order to overcome the limitations of standalone methods through complementary integration. Second, sparse linear regression that is penalized by the regularization parameter (lasso) and bootstrapping-based sparse linear regression methods were suggested as state of the art in recent related works for network inference but they are not effective for a small sample size data and also a true regulator could be missed if the target gene is strongly affected by an indirect regulator with high correlation or another true regulator. We present two novel network inference methods based on the integration of three different criteria, (i) z-score to measure the variation of gene expression from knockout data, (ii) mutual information for the dependency between two genes, and (iii) linear regression-based feature selection. We proposed a lasso-based random feature selection algorithm (LARF) to achieve better performance overcoming the limitations of

bootstrapping as mentioned above. There are three contributions as follows. First, our z score-based method to measure gene expression variations from knockout data is more effective than similar criteria of related works. Second, we confirmed that the true regulator selection can be effectively improved by LARF. Lastly, we verified that an integrative approach can clearly outperform a single method when two different methods are effectively jointed. In the experiments, our method is validated by outperforming the state of the art on DREAM challenge data, and then LARF is applied to inferences of gene regulatory network associated with Psychiatric disorders.

In chapter 4, we proposed a network inference method to infer gene regulatory network integrating gene expression data with Single Nucleotide Polymorphism (SNP) in order to elucidate the overall relationships between gene expressions and genetic perturbations. In the most of the network inferences named as SNP-Gene Regulatory Network (SGRN) inference, pairs of SNP-gene are given by separately performing expression Quantitative Trait Loci (eQTL) mappings. In this chapter, we propose a SGRN inference method without pre-defined eQTL information assuming a gene is regulated by a single SNP at most. To evaluate the performance, the proposed method was applied to random data generated from synthetic networks and parameters. There are three contributions. First, the proposed method provides both the gene regulatory inference and eQTL identification. Second, the experimental results demonstrated that integration of multiple methods can produce competitive performances. Lastly, the proposed method was also applied to psychiatric disorder data in order to explore how the method works with real data.

In chapter 5, we proposed a novel learning structure method for Bayesian network classifier (BNC) that represents the conditional relationships between proteins. Based on the proposed BNC, we introduced Personalized Medicine (PM) that provides the optimal drug for a given patient. More precisely, BNC predicts the effect of

drug for the patient and the drug that is more likely to work better is recommended to the patient. BNC has attracted researchers' attention since the naive Bayes classifier is fast and simple but competitive to state-of-the-art algorithms in performance comparisons. BNC consists of two components, a structure of Bayesian network and conditional probabilities of each variable given parent variables in the network structure. In BNC studies, it is crucial to construct and estimate discriminative network and parameters for better performance. Over the last decade, the research has been focused on learning a structure that maximizes conditional likelihood (CL) for more discriminative classifier than maximum likelihood-based methods. However, finding the optimal structure is NP problem as searching space is exponential. Thus, only if CL function is decomposed, we may find the optimal structure in linear time. As it is known that there is no closed form of CL function which is decomposable into each variable, it is still challenge to build a network structure that maximizes CL. In this chapter, we proposed conditional mutual information-based scoring criterion, which is decomposable so that we can find the structure where CL is maximized. We performed the evaluation variety of benchmark data sets in order to demonstrate the performance of proposed classifier in comparison to state-of-the-art methods. After the proposed method was applied to lung cancer, the network structures of 23 BNCs are integrated to discover biomarkers.

CHAPTER 2

Effects of Low Dose Ionizing Radiation in DNA Damage-caused Pathways Inferred by using Reverse Phase Protein Array and Bayesian Networks

2.1 Introduction

Recently, the importance of understanding the biological effects of exposure to low dose Ionizing Radiation (IR) emerged because of not only occupational exposures to uranium miners, X-ray operators and astronauts but also from non-occupational exposures such as nuclear power plant accidents caused by natural disasters such as earthquakes and the aftermath of tsunamis [4, 5]. IR-caused damage to Deoxyribonucleic acid (DNA) constitutes a broad range of base damage and double strand breaks and induces the operation of relevant signaling pathways such as DNA repair, cell cycle control, and cell apoptosis [6]. In this chapter, we aim to investigate which signaling pathways are activated and how they operated under different doses (0 cGy, 4 cGy, 10 cGy, 50 cGy, 1 Gy, and 5 Gy) and time periods (1, 6, 24, 48, and 72 hours) in order to verify if our body is affected by low dose IR as well as high dose IR. More precisely our subgoals are to understand the effect of low dose IR. We aim (i) to verify which proteins could be involved in DNA damage-caused pathways and to discover candidate biomarkers associated with the response of IR exposure, (ii) to investigate how DNA damage-caused pathways are activated by IR and how the inferred pathways are different under low and high doses of IR, and (iii) to analyze how the expression levels of selected proteins are changed and regulated by each other on the inferred pathways.

To achieve this aim, we mainly analyzed how the activated signaling pathways are inferred by measuring how the protein expression level is different under low dose (4 cGy, 10 cGy) and high dose (1 Gy, 5 Gy) IR. To quantitatively measure the systemic responses of proteins in pathways, Reverse Phase Protein Array (RPPA) is used in conjunction with the Quantum dots (Qdot) nano-technology. RPPA was originally introduced by Liotta [7, 8] and it was designed for quantitatively profiling protein expression levels in a large number of biological samples. In RPPA, sample lysates are immobilized in a series of dilutions to generate dilution curves for quantitative measurements. It is able to use only a small amount (nanoliter) of sample while other protein arrays immobilize antibodies. After primary and secondary antibodies are probed, a signal is detected by Qdot assays. Qdot is a nano-metal fluorophore with a brighter and linear signal, which prevents photo-bleaching effect that often occurs in organic fluorophores. In addition, RPPA offers more accurate information to infer a signaling pathway with post-translational modifications (phosphorylation) not obtainable by gene microarray and PPI [9, 10]. We refer the readers to previous work[11] for more details.

Once DNA damage is caused by IR, the detection of DNA damage through the MRN complex (MRE11-RAD50-NBS1) initiates the response of DNA damage. Then the signal of DNA damage is transmitted to the ataxia-telangiectasia mutated (ATM), that passes the signal to various protective pathways [12]. Hence, isogenic human Ataxia Telangiectasia (A-T) cells are suitable to study of DNA damage response induced by IR because cellular phenotype of A-T cells shows the hypersensitivity to ionization radiation and defects in the ATM signal transduction. In fact, ATM is now widely used as a sensitive monitor of the activation of DNA damage responses after exposure to ionizing radiation [13]. In this way, A-T cells are used to study DNA damage responses possibly induced by low dose IR.

In order to observe how ATM related-pathways are operated by IR, the network inference is performed with proteins, which highly interact with each other by using the correlation coefficient. The network inference we use is based on learning Bayesian networks, in which an optimal structure is constructed by maximizing the probabilistic fitness of the network structure to a given data [14]. Learning structures of Bayesian networks has been explored over the last decade, which contains the development of searching and scoring schemes. To find the optimal structure maximizing the score function, we apply the searching method that is based on linear programming relaxation approach and originally developed by Jaakkola et al.[15] as an exact method [16] rather than a heuristic searching approach [17]. In addition, to overcome inherent limitations of biological data such as noise and limited number of samples, many methods have been proposed to integrate observed data with existing knowledge of interactions so as to increase more reliability and decrease false positive and false negative [18, 19]. To this end, we integrate RPPA with existing PPIs and pathways as prior knowledge using Dempster's combination rule and heuristic basic belief assignments of DST in learning Bayesian networks.

By analyzing the experiments, we can imply that low dose IR causes DNA damage and induces the DNA damage-caused signaling pathways such as the responses to high dose IR by observing the following results. First, the networks that consist of high correlated edges show the similarity in low and high dose IR data compared to non-treatment data (0 cGy). Secondly, the inferred pathways of low dose IR data (ATM+/-) are more similar to those of high dose IR data than non-treatment data. Lastly, when we observe ATM+ data only, the expression level of ATM in non-treatment data is discriminative with low dose IR as well as high dose.

Through the analysis in the experiments, we not only confirm that ATM, CHEK1, CHEK2, and TP53 conduct DNA damage-caused signaling but also pro-

vide the evidence that CDKN1A, PTEN, AKT1, BCL2, and GSK3B could have a significant role in DNA damage response. In addition, ATM-CHEK1/2-TP53-PTEN pathway is distinctly activated in both low and high dose IR and it seems that this pathway is related to regulation of CDKN1A and GSK3B that are associated with cell cycle. Although CASP8 and CASP3, which operate apoptosis, were not selected in the correlation test, the result shows that both CASP8 and CASP3 are gradually activated by the exposure to low and high doses of IR but not non-treatment.

2.2 Methods

2.2.1 Learning Bayesian Networks

Consider a finite set $V = \{X_1, X_2, \ldots, X_n\}$ of n discrete random variables for a given data set. These variables are represented by a set of nodes and can be connected by edges in Bayesian networks. A Bayesian network is a directed acyclic graph (DAG) which represents the conditional dependence between variables through oriented edges. The joint distribution with these conditional probability distributions is defined as follows:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\pi_i) \tag{2.1}$$

where π_i is a set of variables which are the parent nodes of X_i in network structure and we define a possible parent set of X_i as $\Pi_i = \{\pi_i^1, ..., \pi_i^{2^{n-1}}\}$. Learning Bayesian network is to find the structure that best fits a given data. In this paper, score and search scheme is used for learning Bayesian network. First, we measure the degree of fitness between estimated network and given data using score function [20], and then try other structures until the optimal structure which has the maximum score is

found. Given a network structure G and data D (samples of random variables), the score function K2 [21] is defined as follows:

$$Score(G:D) = \log(p(G)) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!\,)\right] \quad (2.2)$$

where r_i is the number of states for variable X_i and q_i is the number of possible configurations of a parent set of X_i. N_{ijk} is the number of instances where the variable X_i takes the k_{th} state and the parent nodes of X_i have the j_{th} ($j = 1, 2, \ldots, q_i$) configuration. N_{ij} is the total number of instance where the parent nodes of X_i have the j_{th} configuration. If the prior probability $\log(p(G))$ in (2) is fixed or ignored given a network structure, the second term of scoring function can be decomposable into each node i like $\sum_{i=1}^{n} S_i(\pi_i^*)$ where π_i^* is a selected parent set of X_i among Π_i and S_i is a decomposed second term of score function (2.2). The goal is to find $G^* = \{\pi_1^*, \ldots, \pi_n^*\}$ maximizing $\sum_{i=1}^{n} S_i(\pi_i)$. However, since graph G should be acyclic, each π_i cannot be selected independently without considering the parents of the other nodes. This is the most critical problem in learning Bayesian network.

2.2.2 Reducing Search Space

As the number of possible parent set of a node is 2^{n-1} and the total number of possible network structure is $n^{2^{n-1}}$ including cyclic structures, exponential searching space is another difficult problem. Even if we limit the number of possible parents to one node, the total number of possible structures will be n^n. Still searching space is exponential. For this reason, we prune away as many edges as possible from searching space using Mutual Information (MI) which is defined as

$$MI(X_i, X_j) = \sum_{X_i, X_j} p(X_i, X_j)\log\left(\frac{p(X_i, X_j)}{p(X_i)p(X_j)}\right), \quad (2.3)$$

After we build MI matrix in which each element MI_{ij} indicates MI value between X_i and X_j, we select only the edges whose MI_{ij} is higher than heuristic threshold.

Hence, we exclude unnecessary edges effectively so that the number of possible π_i can be reduced effectively. In addition, the number of parent of each node can be limited by given a priori number from 2^{n-1} to $\binom{n-1}{m}$ where m is the maximum number of parent nodes of each node. In our experiments, m is set to 4.

2.2.3 Cluster based Linear Programming Relaxation

To find the optimal structure that maximizes the score, we employed learning Bayesian networks based on linear programming relaxation [15]. First, the objective function is defined as

$$\max_{G} \sum_{i=1}^{n} S_i(\pi_i) \qquad s.t.\ \pi_i \in G \tag{2.4}$$

Constraint to be relaxed is defined as

$$\sum_{i \in C} I_C(\pi_i) \geq 1 \qquad s.t.\ C \in \mathcal{C} \tag{2.5}$$

where C is a cluster (or set) of nodes and $I_C(\pi_i)$ is an indicator function. If any node of π_i is in the cluster C, $I_C(\pi_i) = 0$. Otherwise, $I_C(\pi_i) = 1$. This constraint is from the fact that any subset of nodes in acyclic graph has at least one node whose parent is outside of the acyclic graph. So if selected π_i^* of all nodes satisfies the constraint for every possible clusters (all subset of nodes), G^* is a acyclic graph. With this constraint, dual problem can be defined as

$$\min \sum_{i=1}^{n} \max_{\pi_i \in \Pi_i} [S_i(\pi_i) + \sum_{C: i \in C} \lambda_C I_C(\pi_i)] - \sum_{C \in \mathcal{C}} \lambda_C \tag{2.6}$$

$$s.t. \lambda_C \geq 0, \forall C \subseteq V$$

where V is all subsets of nodes and λ_C is a dual variable for each cluster (each constraint). Since the number of λ_C is exponential, we initially set all λ_C to zero and $\mathcal{C}$ to $\emptyset$, and then we iteratively add a single cluster into $\mathcal{C}$ and optimize λ_C. In every iteration, the relaxation for a single constraint is performed by adding a cluster and

all dual variables (λ_C) is updated (optimized). Until dual value is equal to primal value, a new cluster is added in $\mathcal{C}$ in each iteration.

For more details about optimization of dual problem in learning Bayesian networks using cluster-based linear programming relaxation, we refer the reader to the literatures[15, 22]

2.2.4 Prior Knowledge

2.2.4.1 Prior Probability

For a more reliable network inference, RPPA is integrated with STRING PPI data [23] and KEGG pathway data [24] as prior knowledge using prior probability of score function (2). $p(E_{ij} = 1)$ is the probability of the edge that is oriented from node i to j. The probability that there is no directed edge between parent node i to child node j is $p(E_{ij} = 0) = 1 - p(E_{ij} = 1)$. Through this probability of a single directed edge rather than the probability for all cases between two nodes $(p(E_{ij} = 0, E_{ji} = 0), p(E_{ij} = 1, E_{ji} = 0), p(E_{ij} = 0, E_{ji} = 1), p(E_{ij} = 1, E_{ji} = 1))$, we can define a prior probability of a network structure G as

$$\log(p(G)) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \log(p(E_{ij} = e_G(i, j))) \tag{2.7}$$

where $e_G(i, j)$ is a binary indicator function for $E_{ij} \in E_G$, i.e., that the edge E_{ij} exists in structure G. $p(E_{ij})$ also allows the prior probability, $\log(p(G))$, to be decomposed in Bayesian scoring function as follows;

$$\begin{aligned} Score(G : D) = \sum_{i=1}^{n} \Bigg[& \sum_{h=1, h\neq i}^{n} \log(p(E_{hi} = e_G(h, i))) \\ & + \sum_{j=1}^{q_i} \left[\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right] \Bigg] \end{aligned} \tag{2.8}$$

2.2.4.2 Dempster-Shafer Theory (DST)

To combine different evidences about the existence of two interactive proteins in real signaling pathways, we adopt DST as a fusion method being able to integrate different types of biological data. In the DST model, the Frame of Discernment is defined as $\Theta = \{\theta_1, \ldots, \theta_n\}$ that consists of a finite discrete set of exhaustive and exclusive elements as hypotheses space. In our case, there are two discrete hypotheses, $\Theta = \{\theta_1, \theta_2\}$, where θ_1 represents that an edge exists between given two nodes, otherwise θ_2. The power set of Θ, denoted 2^Θ, is defined by all subsets of Θ, i.e.,$2^\Theta = \{\theta_1, \theta_2, \{\theta_1, \theta_2\}, \phi\}$. Basic Belief Assignment (BBA) is a mass function $m : 2^\Theta \rightarrow [0, 1]$, that encodes a support of an element A of 2^Θ satisfying $m(\phi) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. For example, $m(\theta_1)$ indicates the degree of belief that two nodes are connected, $m(\theta_1, \theta_2)$ represents the uncertainty of the connection of two nodes. In DST, uncertainty of two proteins' interaction in signaling pathway is represented by an interval rather than a single probability. The lower and upper bounds are called Belief and Plausibility respectively.

$$Bel_m(A) = \sum_{A' \subseteq A} m(A') \qquad Plaus_m(A) = \sum_{A' \cap A \neq \varnothing} m(A') \tag{2.9}$$

Bel is the degree of a belief to which the evidence supports A, whereas $Plaus$ is the degree of belief to which the evidence fails to refute A. Multiple evidences, i.e., m_1 and m_2 for a belief can be combined by Dempster's rule of combination defined as

$$(m_1 \oplus m_2)(A) = m_3(A) = \frac{\sum_{A_1 \cap A_2 = A} m_1(A_1) m_2(A_2)}{1 - \sum_{A_1 \cup A_2 = \varnothing} m_1(A_1) m_2(A_2)} \tag{2.10}$$

After we combine different evidence for an edge, the Smets pignistic probability transformation is used to calculate the probability of A from combined BBAs [25].

$$P\{A\} = \sum_{A' \in 2^\Omega} \frac{A' \cap A}{A'} m(A') \tag{2.11}$$

For example, if BBAs in Table 2.2 for two sources of evidence are given with $\alpha = 0.4$ and we assume that an edge E_{ij} exists in KEGG pathway but not in STRING PPI, we can calculate prior probability of E_{ij}, $p(E_{ij} = 1)$, as follows;

$$\begin{aligned} m_3(\theta_1) = & \quad \frac{m_1(\theta_1) \cdot m_2(\theta_1) + m_1(\theta_1) \cdot m_2(\theta_1, \theta_2) + m_1(\theta_1, \theta_2) \cdot m_2(\theta_1)}{1 - (m_1(\theta_1) \cdot m_2(\theta_2) + m_1(\theta_2) \cdot m_2(\theta_1))} \\ = & \quad \frac{0.4 \times 0 + 0.4 \times 0.6 + 0 \times 0.6}{1 - (0.4 \times 0.4 + 0 \times 0)} = 0.2857 \\ m_3(\theta_1, \theta_2) = & \quad \frac{m_1(\theta_1, \theta_2) \cdot m_2(\theta_1, \theta_2)}{1 - (m_1(\theta_1) \cdot m_2(\theta_2) + m_1(\theta_2) \cdot m_2(\theta_1))} \\ = & \quad \frac{0.6 \times 0.6}{1 - (0.4 \times 0.4 + 0 \times 0)} = 0.4286 \\ p\{\theta_1\} = & \quad m3(\theta_1) + \frac{1}{2} m3(\theta_1, \theta_2) = 0.2857 + 0.5 \times 0.4286 = 0.5 \end{aligned}$$

2.3 Experiments

2.3.1 Materials

2.3.1.1 Reverse Phase Protein Array

Quantum dot reverse-phase protein array is used to profile the dynamic responses under different doses of IR. ATM-deficient (ATM-) human fibroblasts were

Table 2.1: 67 antibodies of RPPA.

MTOR	CTNNB1	CHEK1	CDH1	MDM2	MAPK14	pMAPK14
pCHEK2	pATM	RB1	pRB1	MAP3K1	pSRC	PTEN
STAT3	CASP8	IGF1R	IRS1	GSK3B	pGSK3	pMDM2
pSTAT3	AKT1	pAKT1	CASP3	PRKDC	pPRKDC	EGFR
pEGFR	RELA	pRELA	NQO1	CDKN1A	CDKN1B	pPTEN
pMAP3K1	BCL2	pBCL2	CASP9	CDK4	pMAPK1	NFKBIA
MAPK8	KL	CDKN2A	TP53	pTP53	SMAD3	SRC
VIM	CLU	ATM	CHEK2	MAPK1	HSPB1	IGFBP3
pCHEK1	pSMAD3	H2AFX	pMTOR	pMAPK8	pIRS1(y896)	pIRS1(y1179)
pPRKDC(S2056)	pIGF1R(y1158.62.63)	pIGF1R(y1162.63)				

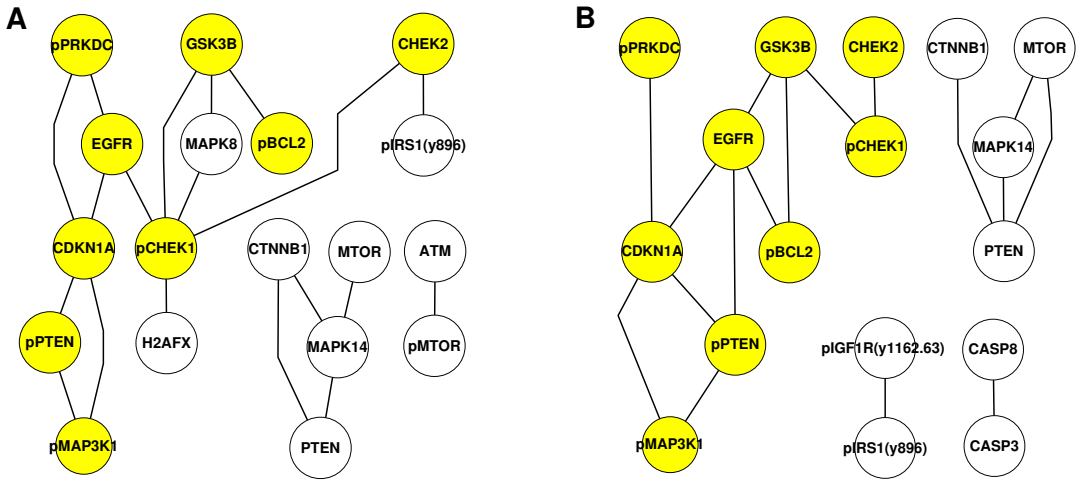


Figure 2.1: Relevance networks for (A) low and (B) high dose IR. The edge weight is correlation coefficient between connected two nodes. The network consists of only edges which weight is higher than 0.9 and excludes the isolated nodes.

isolated from a patient with A-T phenotype, and ATM-proficient (ATM+, clone YZ5) cells were those cells complemented with a wild-type ATM gene. The isogenic pair of ATM cells were treated with a series of IR doses (0 cGy, 4 cGy, 10 cGy, 50 cGy, 1 Gy, and 5 Gy); cell lysates were collected at different time points (1, 6, 24, 48, and 72 h), serially diluted and spotted on protein arrays in triplicates. To profile the dynamic responses of proteins in relevant signaling pathways, we use commercially available antibody sampler kits (Cell Signaling Technology, Inc.) including 67 antibodies (Table 2.1). For better signal detection, biotinylated secondary antibodies and streptoavidin conjugated Quantum Dot 655 (Invitrogen, Inc.) are used to amplify the signals. Signal readouts are the intensities of EC50 of each dilution curve. The measured intensities are corrected by intensities of total protein stain (SyproRuby, Invitrogen, Inc.) for protein loading, then normalized into zero to one [26]. In our experiments, we grouped data into three data sets, (i) 0 cGy (ii) 4 cGy and 10 cGy, and (iii) 1 Gy and 5 Gy as non-treatment, low dose IR, and high dose IR respectively.

2.3.2 Results

Each subsection presents the results of our subgoals; what proteins could be involved in a DNA damage-caused pathway, how a DNA damage-caused pathway is activated by IR and how the inferred pathways are different under low and high doses of IR, and how the expression levels of selected proteins are regulated by each other on the inferred pathways. In addition, how the prior knowledge is defined is presented here.

2.3.2.1 DNA Damage-associated Proteins

In order to infer the activated signaling pathways, we selected the proteins that are strongly interact with each other using the correlation coefficient. To do this, after we constructed a clique graph, in which edge weight represents correlation coefficient between two proteins, all edges were removed except the edges whose correlation was higher than the pre-defined threshold (0.9 in our experiment). We call this network *relevance* networks. As shown in Figure 2.1, there are two relevance networks for low and high dose IR data. It is assumed that highly interconnected components (yellow colored connected graph in Figure 2.1(b) of the relevance network for high dose IR are likely to be associated with DNA damage-caused signaling pathways. Not surprisingly it is shown that the 9 proteins selected in high dose IR appear in the relevance network of low dose IR and the selected protein set includes CHEK1 and CHEK2 that mediate the signals from ATM as it is known well. Also we need to note that the two network structures are similar which means that low dose IR affects the human body and induce related signaling pathways in the same manner as high dose IR. Additionally we selected three proteins, ATM, AKT1, TP53 that have relationships with the already selected 9 proteins or have a significant role in

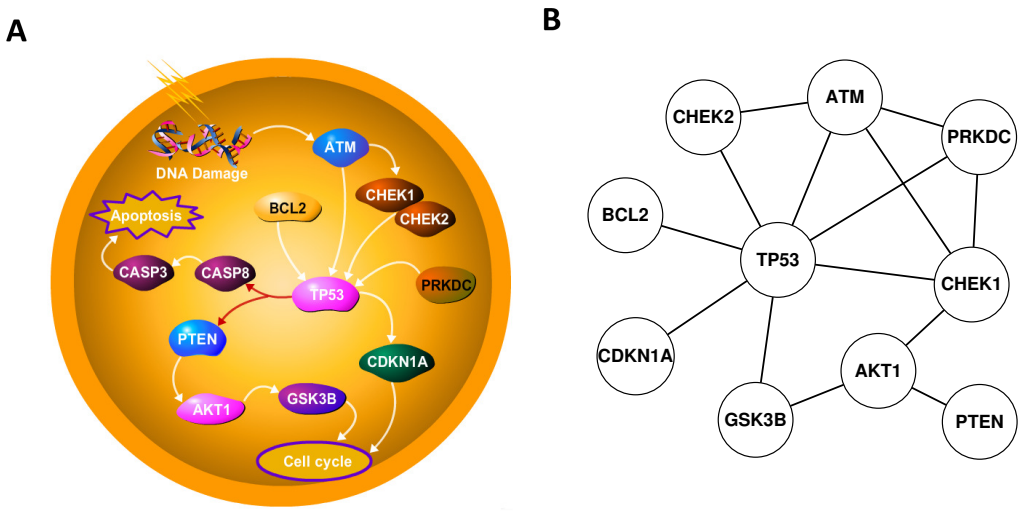


Figure 2.2: Prior knowledge for signaling pathway of selected proteins. (A) Signaling pathway from KEGG where red colored edges (TP53-CASP8 and TP53-PTEN) indicate indirect regulation. (B) PPIs from STRING PPI database.

cell cycle, apoptosis pathway. A total of 12 selected proteins are marked by bold text in Table 2.1.

2.3.2.2 Prior Knowledge

The prior knowledge we used is from the PPI and signaling pathway database, STRING and KEGG respectively. According to the signaling pathways in KEGG, our 67 proteins are involved in a total of 58 pathways. Among 58 pathways, we focus

Table 2.2: Basic belief assignment for evidence KEGG pathway and STRING PPI. θ_1 and θ_2 represent *connected* and *disconnected* respectively. In our experiment, α is set to 0.3.

Evidence		$m(\theta_1)$	$m(\theta_2)$	$m(\theta_1, \theta_2)$
KEGG	connected	α	0	1-α
	disconnected	0	α	1-α
STRING	connected	α	0	1-α
	disconnected	0	α	1-α

on only four pathways (cell cycle, p53, PI3K-AKT, and apoptosis) that are directly associated with 12 pre-selected proteins (Figure 2.2). As shown in Figure 2.2(a), the interactions between the selected proteins are retrieved from four selected pathways. CASP8-CASP3 associated with apoptosis are also included to display the relationship between selected proteins and apoptosis. Red edges, TP53-CASP8 and TP53-PTEN, indicate the interaction that requires additional mediation in the signaling pathway. The interactions between ATM, CHEK1/2, PRKDC, TP53, and CDKN1A are from the cell cycle pathway. TP53, CDKN1A, PTEN, and AKT1 interact in the PI3K-AKT pathway. For PPIs, we gained 14 interactions which have a high (experimental) confidence level in the STRING database. Since EGFR and MAP3K1 do not have any PPI with other proteins, they are excluded from prior knowledge for PPIs and pathways. The parameter α in DST is heuristically set to 0.3 since we have to set a balanced value, which means that the inferred network is too similar to the reference network (Figure 2.2(a)) if α is set closer to 1.

2.3.2.3 DNA Damage-caused Pathways

In Figure 2.3, the inferred pathways for non-treatment, low dose IR, high dose IR data are displayed, and the parameter alpha is set to 0 and 0.3 for each data so as to compare the inferred networks with and without prior knowledge of signaling pathways of 12 selected proteins. For example, Figure 2.3(a) and (b) is for non-treatment data with 0 and 0.3 for parameter alpha. The most discriminative feature between exposure to IR and non-exposure is the regulation between TP53 and PTEN. In both low and high dose IR, TP53 regulates PTEN but not in non-treatment. The fact that TP53 regulated by ATM or CHEK1/2 interacts with PTEN that is related to cell survival being consistent with [27] implies that DNA damage-caused pathway such as PTEN-dependent cell survival pathway could be more likely to be induced

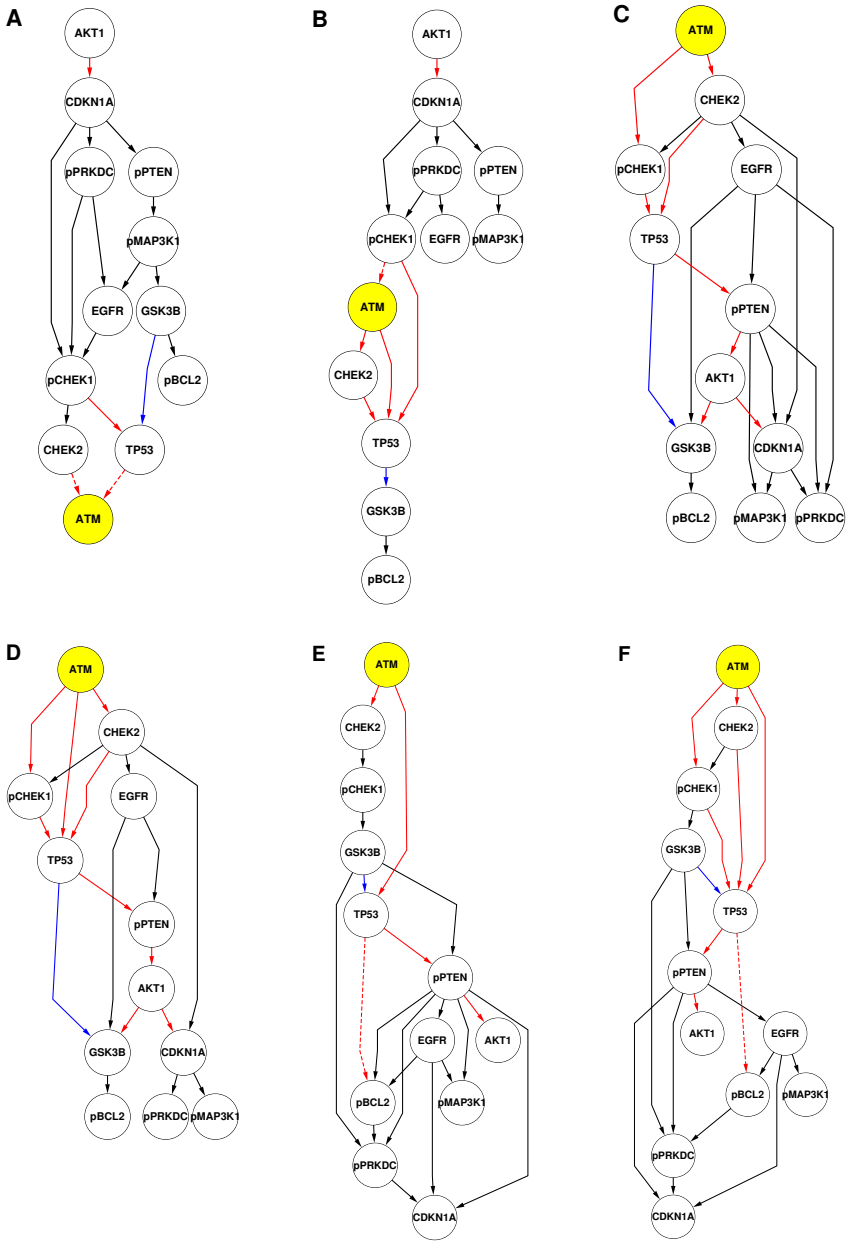


Figure 2.3: Inferred signaling pathways. (A) non-treatment without prior knowledge, (B) non-treatment with prior knowledge, (C) low dose IR without prior knowledge, and (D) low dose IR with prior knowledge, and (E) high dose IR without prior knowledge, (F) high dose IR with prior knowledge. Red edge indicates the regulation that exists in KEGG pathway. Dotted edge refers that the direction is reversed compared to the direction of edge in KEGG pathway. Blue colored edge (TP53-GSK3B) is referred by STRING PPIs

by low dose as well as high does compared to non-treatment. By extension, we note that ATM-TP53-PTEN-mediated signal in both low and high dose IR is transferred into AKT that has the important role of cell cycle progression and survival similarly suggested in related works [28]. The common feature of all the data sets is first that interactions between ATM, CHEK1/2, and TP53. The only difference between non-treatment and exposure to IR is what TP53 regulates after ATM or CHEK1/2-mediated signal. Another common feature we note is the edge between TP53 and GSK3B that is related to apoptosis as discussed in other studies [29]. In addition, it seems that BCL2 is also highly correlated with TP53 and GSK3B in the pathway as concluded in other works [30].

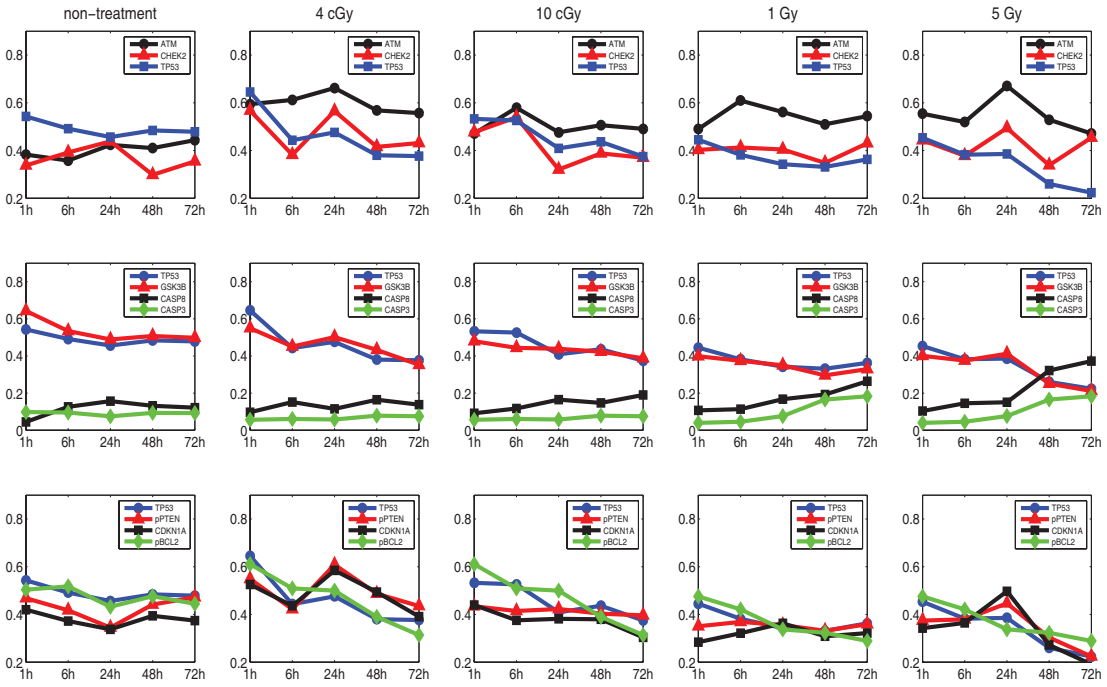


Figure 2.4: The changes of expression intensities of selected proteins in different time periods and doses of IR.

2.3.2.4 Regulatory in DNA Damage-caused Pathways

Even if an edge between two proteins commonly in both non-treatment and IR exposure data, we need to analyze how they regulate through the interaction. We need to especially see how ATM interacts with CHEK1/2 and TP53 in more quantitative regulation rather than just connectivity since the edges between them appear commonly in all the data sets. Figure 2.4 shows the changes of expression of selected proteins under different time periods and IR doses so that we can analyze how they regulate differently through the inferred pathways. The first row of Figure 2.4 is about three proteins, ATM, CHEK2, and TP53, which are directly affected by DNA damage. Since ATM is the important mediator to transmit DNA damage signal, not surprisingly, expressions of ATM in all dose levels are higher than in non-treatment (0 cGy). Note that ATM and TP53 mostly have a same trend of expression changes after being exposed to low/high dose IR for more than 6 hours. In other words, if ATM increases, TP53 increase or otherwise, TP53 decreases. As the synchronized trend is not shown in 0 cGy, we can imply that TP53 is regulated by ATM in both low and high dose IR. CHEK2 also has similar change of expression with TP53 and ATM as it is known that CHEK1/2 mediate the signal from ATM to TP53. In the second row of Figure 2.4, the expression of CASP8 and CASP3 are analyzed in order to see how apoptosis-associated proteins are regulated under different dose levels and time periods even though we didn't include these two proteins in the pathway inference. In both low and high dose IR, CASP8 and CASP3 are more activated when higher dose for a longer time is applied. So, although we need more investigation to understand how DNA damage-caused ATM-TP53 signal is transmitted to CASP8-CASP3 pathway through what mediators like studied in other works [31], it can be implied that the body responds to low dose IR activating apoptosis pathways. The

last row of Figure 2.4 is to verify the strong dependency of PTEN, CDKN1A, and BCL2 to TP53 while IR dose and exposure time period increases. As we mentioned in section 3.2.3, it has been confirmed that these three proteins associated with cell cycle control pathways are involved in the DNA damage-caused signaling pathways under both low and high dose IR.

2.4 Conclusion

In this chapter, we investigate how DNA damage caused by the exposure to low dose IR affects signaling pathways. To measure the expression intensities of protein, RPPA is used with 67 antibodies. Using relevance network, 12 proteins that are highly correlated to each other are selected as DNA damage-associated proteins. The signaling pathways of those proteins are inferred by leaning Bayesian networks and prior knowledge of PPIs and pathways are referred for more reliable inference. The pathways are inferred with three data sets based on different IR doses (non-treatment, low dose, high dose). In the inferred pathways both with and without prior knowledge, the networks for low and high dose IR have more similarity compared to the network for non-treatment. Especially discriminative feature of networks for low/high dose IR is ATM-CHEK1/2-TP53-PTEN-AKT1 pathway. The evidences that low dose IR can affect the body through DNA damage-caused signaling pathway are clearly provided by not only the similarity of networks for low/high dose IR but also quantitative analysis for expression intensities of ATM and CASP8/3. In addition the role of PTEN, CDKN1A, and BCL2 in DNA damage-caused signaling pathway could be studied to fully understand the mechanism of DNA damage response as a future work.

CHAPTER 3

Integrative Approach for Inference of Gene Regulatory Networks using Lasso-based Random Featuring

3.1 Introduction

In chapter 2, biological networks were inferred by using learning Bayesian networks as a standalone method. In chapter 3, we introduce an integration of two different methods in order to overcome limitations of standalone methods. Basically the inference method should be determined depending on both what kind of data such as gene expression, gene-Transcription Factor (TF) [35], or protein-protein interaction (PPI) [36] are used to infer and which type of network model, such as directed or indirected graph [37], we assume. In addition, we have to consider the case of data integration. Namely, not only individual data but also multiple data types together (i.e. integration of gene expression and gene-TF data [38]) can be used for more reliable inference [39, 40]. As an assumption in this work, we limit our inference methods for directed network with a single data type: gene expression data. In order to decipher regulatory interactions with gene microarray data, which provides the gene expression level regulated by the other genes directly or indirectly, the number of effective network inference methods have been proposed by employing a variety of computational and structural models based on boolean networks [41], Bayesian networks [42], information theory [43], regression model [44], and so on. Depending on the different approaches, however, the results tend to be irregular due to inherently different advantages and limitations of each of the inference solutions [45]. The results of the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project

[46] describe well the pros and cons of the different methods as well as how effectively they can work together when the advantages of all methods are integrated (but it doesn't mean any combination always outperforms any other standalone method). More specifically, we note that they conclude two points through the experiments that (i) there is a limit to a single criterion for continuous improvement of network inference research without the integration and (ii) specifically the bootstrapping (re-sampling) based regression method [47] is required to avoid overfitting in regression methods [46].

As the motivation of our first strategy to this end, we focus on an integration of Mutual Information (MI) and L_1 regularized linear regression referred to as lasso [48] but we exclude the learning Bayesian network in the integration. The learning structure of Bayesian networks is somewhat infeasible due to both the discretization problem of a small sample size data and the high cost of computational learning in large scale data. MI is an information theoretic criteria that has been conventionally used for learning large scale network structure [49]. Although MI based approaches such as CLR [50] and ARACNE [51] are limited to reconstructing only an indirected graph unlike linear regression and Bayesian networks, these methods have the popular advantages of computational simplicity and non-linear dependency enabler. In practice, the shortcoming of MI is that it is prone to fail in differentiation between indirect regulation and direct ones. For example, when the edge from G1 to G3 in Figure 3.1 is indirect regulation and the edge from G2 to G3 is direct, MI can recover feed-forward loops comparatively well but not cascades. Highly correlated indirect regulation edges of cascades (G1$\rightarrow$G3 in figure 3.1(b) are likely to be selected by MI. Lasso is also frequently used to select the regulators of a given target gene assuming sparseness of GRN in order to avoid the overfitting of the least-squares problem. In contrast to MI, indirect regulation edge in cascades could be pruned away by lasso in

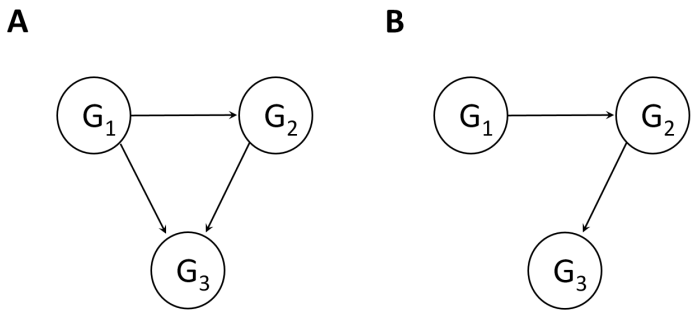


Figure 3.1: (A) Feed-forward loops and (B) Cascades. When G3 is a target gene, G1→G3 and G2→G3 of Cascades are indirect and direct regulations respectively. In MI-based methods, indirect regulations are likely to be selected incorrectly in Cascades. In regression based method, strong direct regulators are more likely to be selected than another direct regulator in Feed-forward loops.

which the objective function is penalized for sparsity by a regularization parameter, called the tuning parameter λ. However, a weakness of regression-based method is that only a strong direct regulator is more likely to be selected than another direct regulator in Feed-forward loops. Therefore, the integration of two methods is considered to deal with the trade-off. The motivation of our second strategy is that the property of knockout data allows us to measure statistical variations between wild-type gene expression and perturbed gene expression after knocking them out to provide the cause-effect information between those two genes. However, there is the limitation that the method is only applicable to gene knockout data.

In this chapter, we propose two methods, IMLARF and ISLARF. First, IMLARF indicates the integration of MI and LARF and consists of three steps. The first step of IMLARF is to build a matrix where each element is an edge score calculated by MI. In order to overcome the limitation of MI as mentioned above, the second step is to construct another edge score matrix using LARF, then the two edge score matrices are combined as the last step. In LARF, we regard a sparse linear regression as a feature selection since our goal is to identify the regulators that best predict

the expression level of target genes. The problem is that features selected by lasso tend to be overfitted to a given tuning parameter λ, and thus the *unstability* problem caused by this overfitting can be solved by using bootstrapping [52, 43] in which data is randomly re-sampled so that a more stable selection can be achieved. However, the limitation of re-sampling is that it could not be effective in the case of a small sample size. Another limitation of bootstrapping is that the true variable (regulator gene) is likely to be missed (false negative) when strong indirect or direct regulators exist. LARF is similar to bootstrapping but LARF selects variables among randomly pre-selected candidate features in each iteration over different tuning parameters of lasso optimization so that true features weakly correlated to the target gene could not be missed, excluding indirect or direct regulators from the feature set. The second method we propose is ISLARF, which integrates two criteria, ZS and LARF. ZS is the name of the criteria that uses the z-score of variation of the knocked out gene expression. Although ISLARF is available only to knockout data, the performance is highly superior to other z-score based similar methods with knockout data in related works. In the experimental evaluation, we validate the proposed method on a dataset from the DREAM3 challenge [53]. In addition, we explore the gene networks of Psychiatric disease with the related genes. The results shows that the proposed method significantly outperforms the state-of-the art [54, 55] and re-builds the known regulations of genes possibly associated with Psychiatric Disorders.

3.2 Methods

3.2.1 Problem Definition

We begin with a brief definition of problems and notations. The network we target is a directed graph that consists of n nodes and $n(n-1)$ edges representing

genes and regulations respectively. Given a matrix $\mathbf{X}\in\mathbb{R}^{N\times n}$ where N is number of samples, we denote the i-th column by a vector $\mathbf{x}_i$ indicating expression levels of i-th gene over N samples, and we also let $X = \{X_1, ..., X_n\}$ be a set of variables (genes, features, node, and variable are interchangeably used in this chapter). The goal of our work is to not only identify the regulators given a target gene but also to define the confidence level of regulation as a weight of the edge. In other words, we estimate the weight of all possible regulations, which are directed edges between all pairs of nodes $\{X_i \leftarrow X_j : i, j \in X\}$ in the network , then select only edges that have a higher weight than pre-defined threshold θ. As a final result, therefore, a weight matrix $\mathbf{W}\in\mathbb{R}^{n\times n}$ is returned by the inference method, and W_j^i represents a confidence level of the regulation when target gene i is connected to activator or suppressor gene j. In the following sections, we present how the edge weight is estimated by information theory, the LARF algorithm, and the z-score from knockout data.

3.2.2 Overview

3.2.2.1 IMLARF and ISLARF

The first method we propose, IMLARF, consists of three steps. Figure 3.2 describes the overview of the proposed method. First, a symmetric edge weight matrix M is calculated by mutual information assuming that, if two genes have a higher mutual dependency, they are more likely to be in the regulation relationship. Second, another edge weight matrix F is produced by the LARF algorithm that consistently gives higher weight to the true edge from regulator to target gene. Lastly, the two weight matrices are combined by their entry-wise product $M \circ F = \{M_j^i \cdot F_j^i | i, j = 1, ..., n\}$. The second method, ISLARF, is similar to IMLARF but using z-score matrix, S, is used instead of MI matrix. If S_j^i has higher value, gene i is more likely

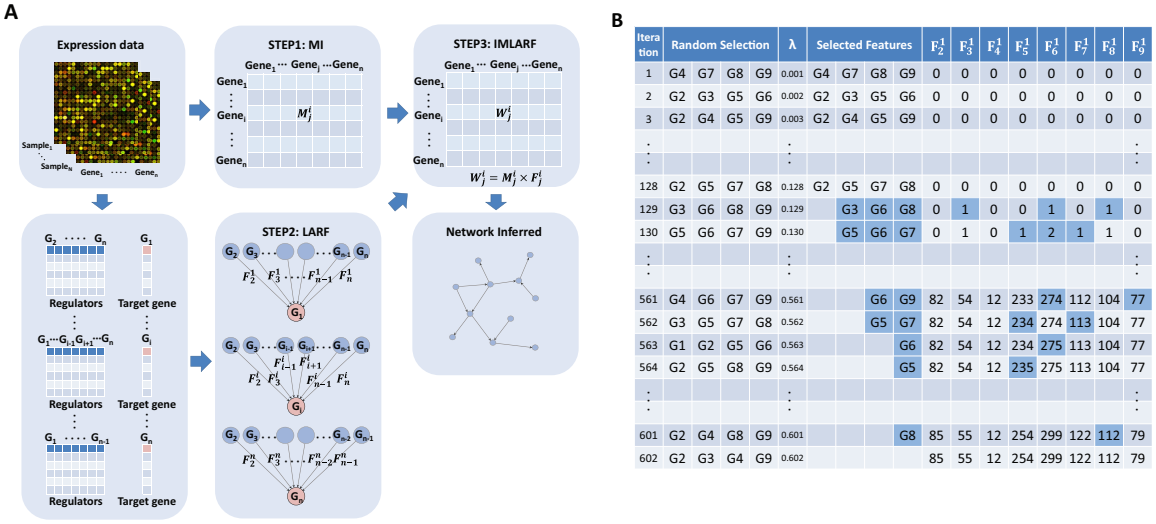
to be regulated by gene j. So in the last step S is combined with F by their entry-wise product $S \circ F$

3.2.3 Information Theoretic Approach

3.2.3.1 Mutual Information Matrix

The dependency of two genes, X_i and X_j, can be measured by MI defined as

$$I(X_i, X_j) = \sum_{X_i, X_j} p(X_i, X_j) \log \frac{p(X_i, X_j)}{p(X_i)p(X_j)}), \tag{3.1}$$



Iteration	Random Selection				λ	Selected Features				F_2^1	F_3^1	F_4^1	F_5^1	F_6^1	F_7^1	F_8^1	F_9^1
1	G4	G7	G8	G9	0.001	G4	G7	G8	G9	0	0	0	0	0	0	0	0
2	G2	G3	G5	G6	0.002	G2	G3	G5	G6	0	0	0	0	0	0	0	0
3	G2	G4	G5	G9	0.003	G2	G4	G5	G9	0	0	0	0	0	0	0	0
⋮					⋮												⋮
128	G2	G5	G7	G8	0.128	G2	G5	G7	G8	0	0	0	0	0	0	0	0
129	G3	G6	G8	G9	0.129		G3	G6	G8	0	1	0	0	1	0	1	0
130	G5	G6	G7	G9	0.130		G5	G6	G7	0	1	0	1	2	1	1	0
⋮					⋮												⋮
561	G4	G6	G7	G9	0.561			G6	G9	82	54	12	233	274	112	104	77
562	G3	G5	G7	G8	0.562			G5	G7	82	54	12	234	274	113	104	77
563	G1	G2	G5	G6	0.563				G6	82	54	12	234	275	113	104	77
564	G2	G5	G8	G9	0.564				G5	82	54	12	235	275	113	104	77
⋮					⋮												⋮
601	G2	G4	G8	G9	0.601				G8	85	55	12	254	299	122	112	79
602	G2	G3	G4	G9	0.602					85	55	12	254	299	122	112	79

Figure 3.2: (A) Overview of IMLARF. The algorithm consists of three steps, the construction of matrix (i) M and (ii) F and (iii) pairwise product of M and F. In ISLARF, the matrix M in step 1 is simply replaced with the matrix S (Section 2.4). (B) An example of procedures of LARF. It shows how the row vector F^1 of frequency matrix F given target gene G1 and 8 other candidate regulators (G2∼G9). By a predefined α, four random features are selected among eight genes in each iteration. In the beginning, F^1 is not increased and four random features are selected without sparsity since λ is not increased enough yet. The more λ is increased, the more the number of selected features (blue-colored cells) is decreased. If no feature is selected due to a highly increased λ, the iteration and frequency measure is finished.

The strength of MI is the ability to measure non-linear dependencies of genes, but the limitation in practice is that the discretization of gene expression is required to calculate the probability of X_i and X_j. Instead, if we assume the Gaussian distribution of gene expression, MI can be computed with its original continuous values by using Gaussian mutual information [56] defined as

$$I(X_i, X_j) = -\frac{1}{2} \log \frac{|cov(X_i, X_j)|}{|cov(X_i, X_i)||cov(X_j, X_j)|}, \tag{3.2}$$

where $cov(X_i)$ is the covariance matrix of variable X_i, and $|cov|$ is the determinant of covariance matrix. The reader is referred to [57] for more details. We build MI matrix in which each element M_j^i indicates the dependency between X_i and X_j which means that X_i and X_j are independent if $M_j^i = 0$ or M_j^i is relatively lower than other edges. Networks with the edges whose M_j^i are higher than the heuristic threshold are referred to as relevance networks. Two critical limitations of relevance networks, however, are that firstly, MI does not provide the direction of edges due to $M_j^i = M_i^j$, and secondly, the high co-expression and indirect regulation may cause false positives.

3.2.4 Statistical Approach

3.2.4.1 Z-score and gene knockout data

We note that knockout data implies cause-effect information. The gene expression level after the perturbation of another certain gene provides the chance to observe if the gene is downstream of the perturbed gene. For example, if the variation between wild type of gene j (X_j^{wt}) and gene j expression measured after gene i is knocked out is high, gene j is likely to be regulated by gene i. The variation matrix D is defined as

$$D_j^i = X_j^{-i} - X_j^{wt} \tag{3.3}$$

$$S_i^j = \left| \frac{D_j^i - \mu_{D_j}}{\sigma_{D_j}} \right| \tag{3.4}$$

where X_j^{-i} is the expression level of gene j after knocking gene i out, and μ_{D_j} and σ_{D_j} is mean and standard deviation of j-th column vector D_j of variation matrix D respectively. As the z-score of D_j^i over D_j is the weight of regulation edge $Gi \rightarrow Gj$, the z-score of D_j^i is equivalent to S_i^j of edge weight matrix S. The limitation of this criterion is the availability only in knockout data.

Algorithm 1 LARF algorithm

procedure LARF(X, α, r,*stepsize*,t)
  for $i \leftarrow 1, n$ **do**
    for $h \leftarrow 1, t$ **do**
      $\lambda \leftarrow stepsize$
      repeat
        $X_{random} \leftarrow RandomFeatures(X^{\setminus i}, (n-1) \times \alpha)$
        $X' \leftarrow RandomSamples(X, N \times r)$
        $X_{selected} \leftarrow Lasso(X'_i, X'_{random}, \lambda)$
        if $0 < |X_{selected}| < n \times \alpha$ **then**
          $F^i_{X_{selected}} \leftarrow F^i_{X_{selected}} + 1$
        end if
        $\lambda \leftarrow \lambda + stepsize$
      until $X_{selected} = \emptyset$
    end for
    $F^i \leftarrow Normalization(F^i)$
  end for
  return F
end procedure

3.2.5 LARF Algorithm

The third approach for complementary integration of inference methods is based on L_1-regularized linear regression (lasso) defined as

$$argmin_{\beta}||X_i - X^{\setminus i} \cdot \beta_i||_2^2 + \lambda||\beta_i||_1 \tag{3.5}$$

where coefficient column vector β_i represents regulation relationships between the target gene i and others. More precisely, after β_i is optimized to minimize the objective function (3.5), then if the j-th element of β_i is zero, gene j does not regulate gene i, otherwise it does. The optimization is performed for each target gene i, $i \in X$. Coefficient matrix $B = \{\beta_1, ..., \beta_n\}^T$ is equivalent to adjacency matrix where non-zero B_{ij} is the regulation edge from regulator gene j to target gene i. The tuning parameter λ in lasso is used to enforce network sparsity, so the number of selected (non-zero coefficient) variables varies with different λ. In our works, we regard variable selection of lasso as a feature selection to predict a target gene's expression level.

To overcome the overfitting problem and the strong indirect regulation problem, lasso is iteratively performed over different λ with randomly pre-defined candidate features rather than random samples like bootstrapping. More precisely, the basic idea of LARF is that lasso is iteratively performed with only randomly selected candidate features while increasing the tuning parameter, then giving weight to each feature by counting how many times each feature is selected in the iterations. We predefine the fraction of the number of all possible features as a parameter α ($0 < \alpha < 1$) for the candidate features. For example, when the number of all possible regulators is n=100, α=0.2 means that only 20 random candidate genes are used in a single iteration of lasso. After random featuring, random sampling is performed with parameter r which decides how many samples are used from the original data. For instance, when the original sample size is N=200 and r=0.7, only 140 random samples are used in each

iteration of lasso. With randomly (uniform distribution) selected features and samples by parameter α, we iteratively run lasso over increasing tuning parameter λ until lasso does not select any features due to a certain high λ. In each iteration, random candidate features and samples are re-defined again. Tuning parameter starts from zero and increases by the parameter *stepsize* that should be small enough, (e.g 0.001). Otherwise, both re-featuring and re-sampling will be biased. For each iteration, the frequency matrix F is updated. The i-th row of F is the frequency of feature selection for target gene i (F_i^i is supposed to be zero). For example, Figure 3.2(b) describes how the F^i is measured. After finishing the iterations (repeat in line 5), we iteratively perform t times (t=10 in our experiments) of the process from line 5 to 13 again, and then i-th row vector of the frequency matrix is normalized by

$$F_j^i = \frac{(F_j^i - max(F^i)}{max(F^i) - min(F_{-i}^i)}, \tag{3.6}$$

where

$$F_{-i}^i = \{F_j^i, j = 1, ..., i-1, i+1, ...n\}, \tag{3.7}$$

and $max(F^i)$ and $max(F_{-i}^i)$ is maximum value of i-th row vector of F and minimum of F_{-i}^i.

3.3 Results

We first evaluated the performance of IMLARF and ISLARF on synthetic simulation data as compared to the state of the art, and then explored the inferred networks with real gene microarray data for psychiatric disorders. The synthetic, non-linear expression data is from DREAM3 *In Silico* Network challenge in which the data is created with the subnetworks of well-known reference networks for *Yeast*. To assess the edge weight matrix W elicited by proposed methods, first the matrix is converted to an edge list sorted by the confidence levels (weight), then the top

k confidence level edges are selected to measure the accuracy criteria, such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The receiver operating characteristic (ROC) curves as a parametric curve were traced over different $k = 1, ..., n(n-1)$ to examine the tradeoff between the true positive rate (TPR) and false positive rate (FPR). The criteria to represent the performance are defined as following:

- TPR=TP/(TP+FN)
- FPR=FP/(FP+TN)
- AUROC: the area under ROC curve.

We compared our method to each standalone method without integrations and also other well known the state of the art methods. The abbreviations of algorithms are listed below:

- MI: edge is scored by mutual information
- ZS: relative variation from wild type is measured by z-score.
- LARF: lasso based random featuring and sampling.
- IMLARF: integration of MI and LARF
- ISLARF: integration of ZS and LARF
- ZDR: top rank in DREAM 3 [54]
- GENIE3: top rank in DREAM 4 [55]
- TIGRESS: top rank in DREAM 5 [52]

3.3.1 Evaluation on the DREAM3 Benchmarks

3.3.1.1 Materials

The data for DREAM3 *In Silico* Network challenge consists of three differently sized networks, (10, 50, and 100 genes), and there are five gold-standard networks for

A

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	0	0.52	1	0.35	0.94	0	0.57	0.53	0.85	0.15
G2	0.20	0	0.66	0.58	0	0.22	0.09	0.73	1	0.25
G3	0.95	0.92	0	0	0.91	0.60	0.73	0.64	0.30	1
G4	0.79	0.76	0	0	0.61	1	0.48	0.75	0.67	0.24
G5	0.91	0.14	0.77	0.21	0	0.65	1	0.10	0	0.39
G6	0.03	0.35	0.27	0.92	0.80	0	0	0.04	1	0.47
G7	0.08	0.04	0.00	0.29	0.65	0.08	0	1	0	0.85
G8	0.07	0.66	0.47	0.58	0.10	0	1.00	0	0.33	0.17
G9	0.36	1	0.15	0.63	0	0.96	0.04	0.40	0	0.35
G10	0	0.38	0.95	0.17	0.40	1	0.84	0.27	0.30	0

B

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	0	0.64	0.92	1	0.67	0.12	0.20	0.04	0.02	0
G2	1	0	0.79	0.60	0.32	0.01	0.16	0.08	0.84	0
G3	0.83	0.79	0	0.11	1	0.55	0.26	0	0.58	0.13
G4	0.96	0.48	0.25	0	0.33	1	0.08	0	0.20	0.21
G5	1	0.20	0.89	0.20	0	0	0.87	0.10	0.15	0.33
G6	0.76	0.21	0.80	1	0.17	0	0.49	0	0.45	0.52
G7	0.41	0	0.05	0.01	0.93	0.56	0	0.92	0.03	1
G8	0.70	0.23	0.51	0.09	0.36	0	1	0	0.21	0.72
G9	0.24	1	0.91	0.98	0.54	0.11	0.53	0	0	0.73
G10	0.16	0	0.35	0.87	0.65	0.85	1	0.36	0.62	0

C

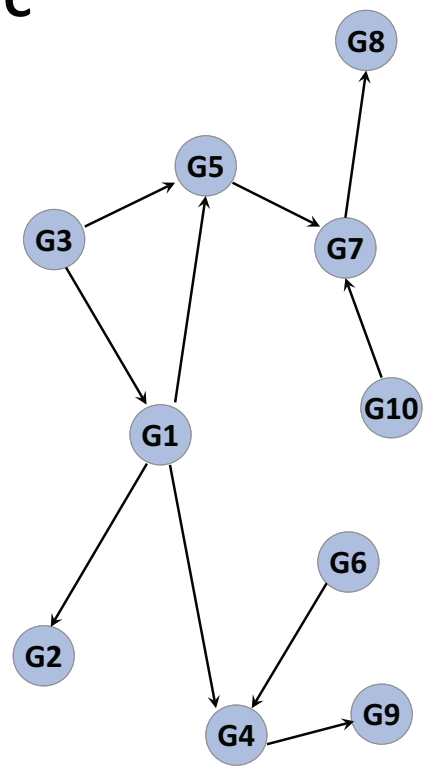


Figure 3.3: (A) The result of LARF with only random sampling (B) The result of LARF with only random featuring. (C) True network of 10-gene Yeast1 in DREAM3

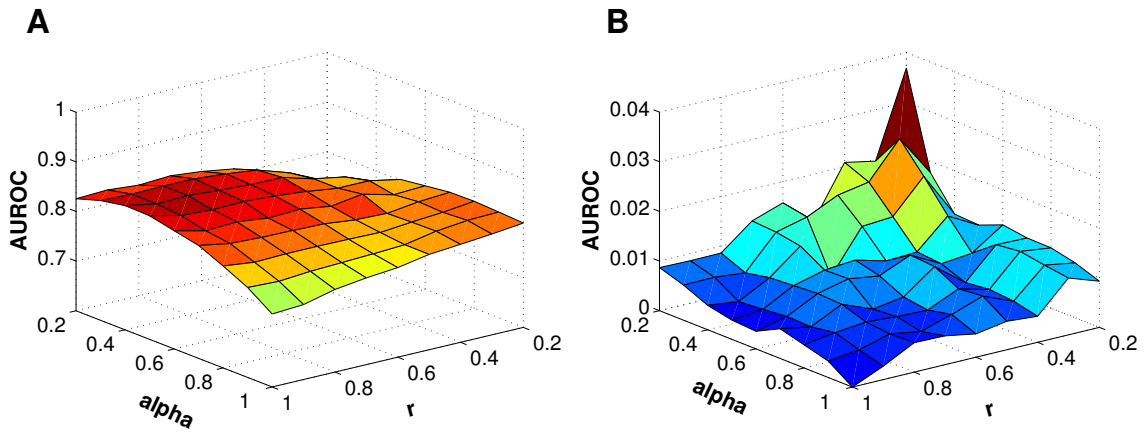


Figure 3.4: (A) Mean and (B) standard deviation of AUROC with different parameters and 10 iterations of experiments for 50-gene Yeast1 network.

each size (total of 15 networks). The five networks are named Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3. From each true network, three different data types (knockdown, knockout perturbations, and time series data) are provided, and the knockdown and knockout data includes a single wild type sample. In our experiments, only knockout data is used and 10-gene, 50-gene, 100-gene of Yeast1 networks are mainly tested.

3.3.1.2 Random sampling vs Random featuring

To evaluate how much more effectively LARF selects true edges than random sampling, we compared them with 10-gene Yeast1 network in Figure 3.3. Figure 3.3(a) is the result of LARF with only random sampling (α=1, r=0.5) and 3b is with only random featuring (α=0.5, r=1). The normalized edge score is the average of 10 experiments and yellow colored cells indicate true edges. In Figure 3.3(a), though G2's true regulator is G1, G2←G3 is relatively higher than G2←G1 probably because of indirect regulation from G3 to G2 through G1. In Figure 3.3(b), G2←G1 is correctly estimated as true edge by random featuring. Similarly two true edges (G4←G1 and G5←G1) are inferred with the highest weight in random featuring but random sampling gives only 0.79 and 0.91 to two true edges (G4←G1 and G5←G1) due to another true edges (G4←G6 and G5←G3) have strong direct regulation (1 and 0.99).

3.3.1.3 Setting Parameters

Before we compare our methods to other methods, we explored the optimal parameters that give the best results. As described in Figure 3.4, the mean and standard deviation of AUROC are measured after LARF are 10 times performed over different parameters, α and r, for 50-gene Yeast1 network. The range of parameter is 0.2∼1 due to too small number of feature and sample in 10-gene network data.

Table 3.1: AUROC of standalone and integrative methods. In the case of LARF-based methods, mean and deviation are measured after each method is performed 10 times for Yeast1 network of DREAM3. The integration of more than two methods is simply done by entry-wise product of edge score matrix. In TIGRESS-TF, the list of TF is provided as TIGRESS is designed for DREAM5 challenge data in which TF is given. Asterisk(*)-marked methods require knockout data.

Method	10-gene	50-gene	100-gene
GENIE3	0.9175	0.8427	0.8631
TIGRESS	0.7044 ± 0.0056	0.8179 ± 0.0025	0.7690 ± 0.0023
TIGRESS-TF	0.8154 ± 0.0037	0.9006 ± 0.0010	0.8777 ± 0.0009
MI	0.9312	0.8329	0.8586
LARF	0.9250 ± 0.0154	**0.8489** ± 0.0038	0.8610 ± 0.0039
IMLARF	**0.9425** ± 0.0047	0.8487 ± 0.0032	**0.8701** ± 0.0012
ZDR*	0.8975	0.9223	0.8876
ZS*	0.9725	0.9204	0.8870
ZS*+MI	0.9775	0.8931	0.8925
ISLARF*	**0.9892** ± 0.0021	**0.9301** ± 0.0049	**0.9065** ± 0.0029

The best result (0.8501±0.0049) is recorded with α=0.4 and r=1 for 50-gene Yeast1 data. This indicates that the random sampling rate does not necessarily need to be applied to avoid overfitting once random featuring is applied. In addition, the figure also shows that the AUROC can be decreased with high standard deviation if both parameters are too small. If the sample size is small (N=10), the deviation is quite high in low α and r though AUROC is high. As the best result for 10-gene and 100-gene Yeast1 data, 0.925±0.0125 and 0.8611±0.0046 were achieved with α=0.5, r=1 and α=0.4, r=1 respectively. It also shows random sampling could not make an improvement in both small and large sample sizes. Therefore we applied fixed parameters α=0.5, r=1 to all data sets in the experiments.

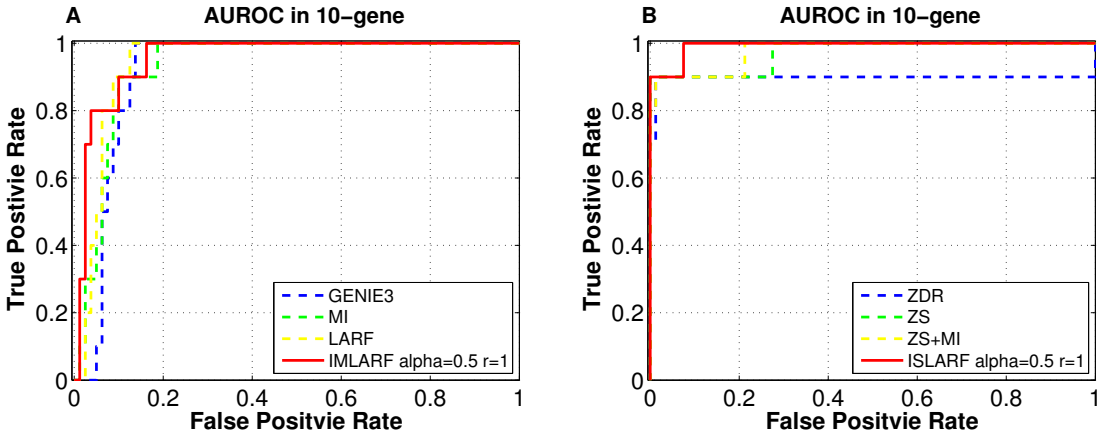


Figure 3.5: ROC of the methods (A) without and (B) with gene deletion information in 10-gene network

3.3.1.4 Effect of integration and performance comparisons

Table 3.1 presents the performance of integrative approaches compared to a single method. The integration of MI and LARF outperforms standalone MI and LARF except 50-gene. Similarly the performance of ISLARF is better than other integration such as ZS+MI and standalone ZS. If knockout data is not available, IMLARF will be the best method as ZS is not applicable. Since ZDR is based on knockout data, the result shows that ZDR is quite better than other methods such as IMLARF except in a small size network. In Figure 3.5, the AUROC for proposed methods and the state of the art methods with 10-gene Yeast1 data are plotted after only a single experiment. Overall results show that ISLARF is the best method if knockout data is available, otherwise IMLARF is superior to other methods.

3.3.2 Inference of GRN for Psychiatric Disorders

In this section, the proposed method is applied to real gene expression data for psychiatric disorders. Through the experiments, we evaluate how the method constructs the network and explore what potential biomarkers of Psychiatric disor-

ders are in the inferred networks. Psychiatric disorders data that are provided from the Stanley Medical Research Institute (SMRI) consist of gene expression data of 25833 genes and 131 samples (43 controls and 88 cases) including bipolar disorder, schizophrenia, major depression as three major psychiatric diseases.

To select genes possibly associated with psychiatric disorders, two statistical tests, t-test and z-test [58], are performed. In Figure 3.6(a), all genes are plotted by using p-value of t-test for y-axis and z-test value for x-axis, and the plot shows that two tests shows similar results in linear patterns. From these two tests, we selected 1407 genes as cut-off values are set to $-log_{10}(0.01)$ and ± 2.326 for t-test (y-axis) and z-test (x-axis). To find a module of genes that may interact to each other in Psychiatric disorders, we initially built a correlation matrix whose element of ith row and jth column is absolute value of correlation between expressions of ith and jth genes, and then clustering is performed to the estimated correlation matrix as shown in Figure 3.6(b). Based on the result of clustering, we manually set 8 groups of genes (yellow squares).

To analyze the relationship between clusters, first, IMLARF was applied to all 1407 genes with setting θ to 0.2. Figure 3.7 shows only the two largest components of the inferred network where node color indicates a cluster number after small components of the network are removed from the figure. The result is consistent with the correlation matrix in Figure 3.6(b) showing the features as follows: (i) cluster 3, 6, and 8 in the network strongly and exclusively interact to each other, (ii) cluster 2, 4, and 5 are complicatedly interacting together, (iii) cluster 7 is widespread over the whole network. To observe the strong regulation of the network, we inferred network with all the genes again after setting θ to 0.4. As a result, we displayed the second largest component in the inferred network in Figure 3.8(a). Most nodes of the network are genes of cluster 3 implying that cluster 3 is most exclusively and strongly interact-

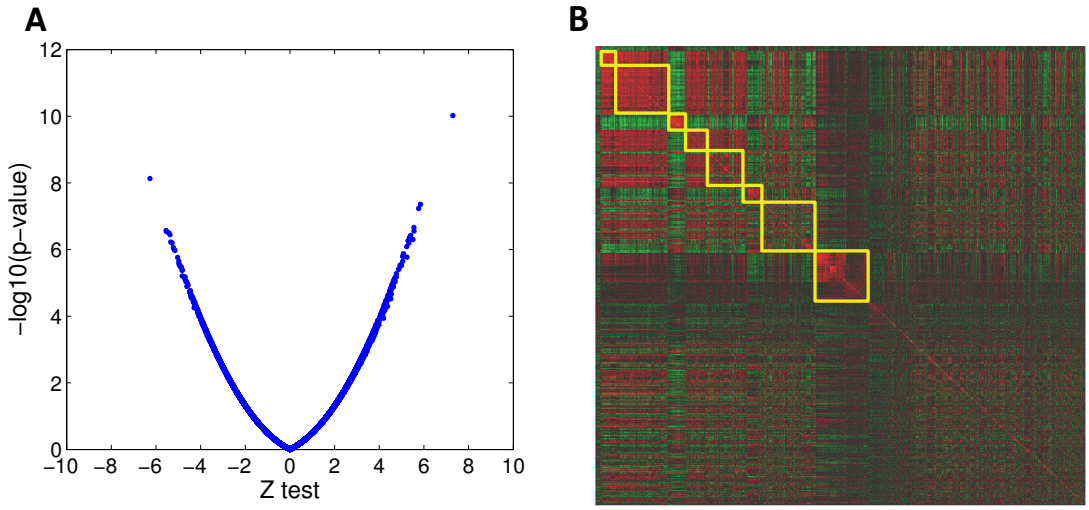


Figure 3.6: (A) t-test and z-test (B) clustered correlation matrix and 8 clusters (yellow squares)

ing within the cluster. It is noted that 7 genes, DAO [59], PRDX6 [60], KCNN3 [61], TCF7L2 [62], RFX4 [63], FYN [64], and B3GAT2 [65] (yellow-colored nodes) , relevant to psychiatric disorders are involved and interestingly these genes except B3GAT2 constitute a connected subgraph. Blue-colored nodes indicate the genes that have more than two connection to yellow nodes supposing that these genes are likely to be susceptible to psychiatric disorders (In this chapter we call yellow and blue gene reference gene and susceptible gene respectively. We define a gene as a reference gene if a gene appears with a psychiatric disease in the title of related literatures). There are 4 genes, SOX9, HEPH, AQP1, and SDC3 as susceptible genes, and it was already reported that SDC3 has a weak association with schizophrenia in related GWAS [66]. Figure 3.8(b) is the inferred network for cluster 7, and a total of 8 genes known as psychiatric disorder-related genes in related literatures are found as following: TEF[67], NR1D1[68], KIF13A[69], ADCYAP1R1[70], MDGA1[71], GNAZ[72], CNR1[73], and DCLK1[74]. Additionally we defined 5 genes, ZBTB20, MAP7, ZBTB16, ANK2, and MRAP2, as susceptible genes, and surprisingly ZBTB20[75], MAP7[76], ZBTB16[77],

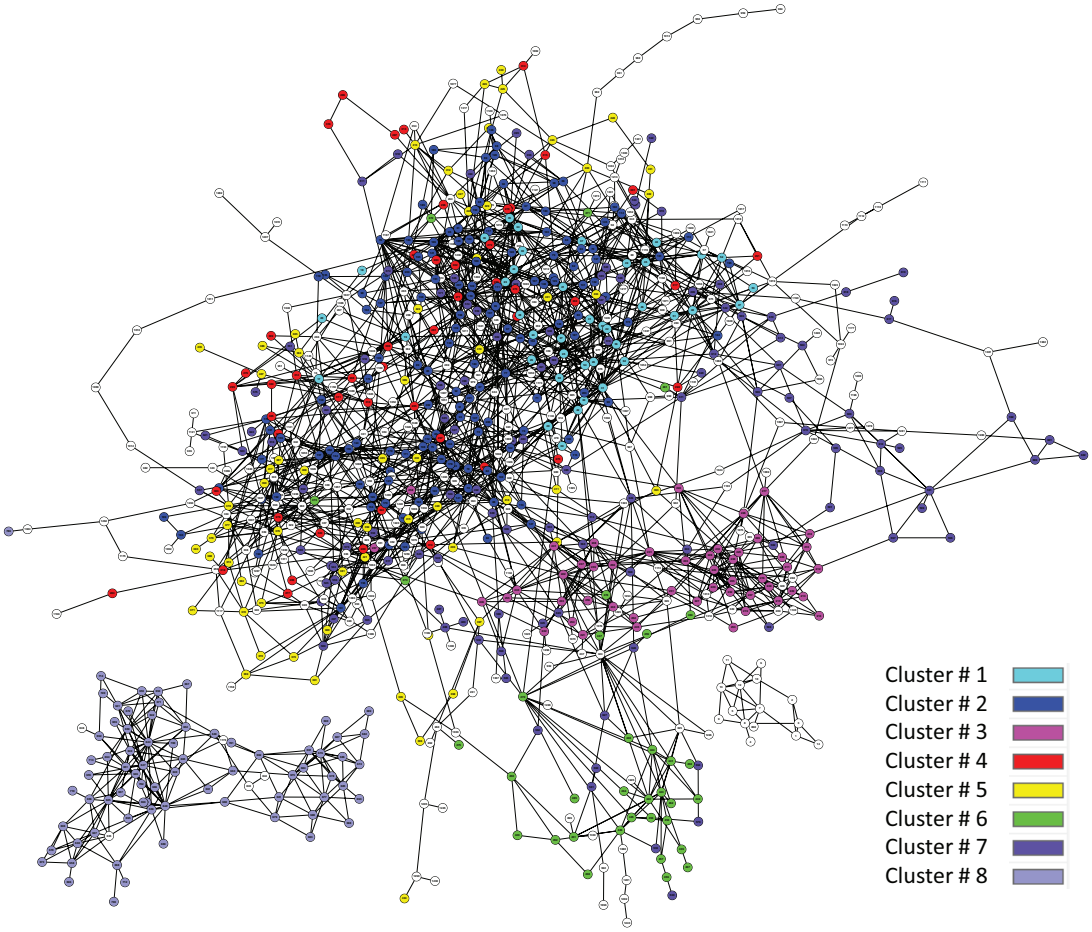


Figure 3.7: Large components of network inferred with 1407 genes

ANK2[78] was also reported as schizophrenia disorder-associated genes in SNP and CNV-based studies. So we imply that it is worth to investigate the genes that have only an edge to reference gene as candidate genes associated with psychiatric disorder. In addition, reference genes in the network tend to interact with each other directly or indirectly though susceptible genes but they are not widely spread implying they may work together or may be co-regulated by another unknown biomarker.

The network inference result for the combination of cluster 4 and 5 is shown in Figure 3.8(c) consisting of two components. There are 10 reference genes such as DLG4[79]], MIF[80], SLC6A5[81], GAD1[82], GAD2[83], GOT2[84], RGS9[85],

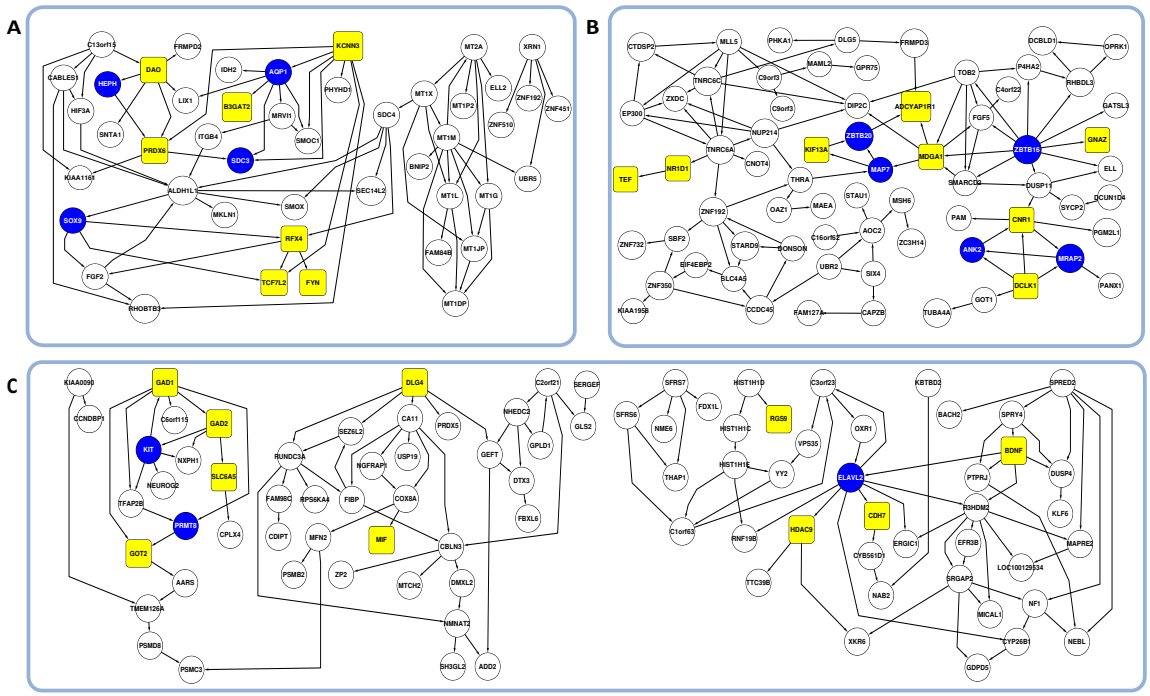


Figure 3.8: Inferred gene regulatory networks for (A) Cluster 3 (B) Cluster 7 (C) Cluster 4 and 5. Yellow-colored nodes indicate the genes known as Psychiatric disorder genes in the literatures. Blue-colored nodes are the genes that are connected to more than two yellow genes.

HDAC9[86], CDH7[87], and BDNF[88], and 3 susceptible genes such as PRMT8, KIT, and ELAVL2. It is noted that ELAVL2 has connections to three reference nodes and was reported as schizophrenia-related gene in recent GWAS [89].

3.4 Discussion

The difference between ZS and z-score of [54] is in whether the absolute value of variation D_j^i is taken before z-scoring or original value of D_j^i is used. In our method, we simply calculate the z-score to measure how many deviations the observed variation is above or below while the absolute value of variation $|D_j^i|$ is used for z-score. Since we want to know how much the variation of a gene is higher than another target gene after knockout of the source gene, the use of D_j^i rather than $|D_j^i|$ is more

reasonable and it is not guaranteed to select high-variant genes if absolute value of D_j^i is used. Since random featuring and random sampling are performed in iterations of lasso, the computational time is significantly increased especially in finding optimal parameters. In implementation, the stepsize, therefore, should be set to a reasonably small value, and parallel processing (i.e. *parfor* in matlab) can reduce the processing time in practice (In our case, eight local cores are used). As a future work, we can integrate TF information additionally in the inference so that we can get more reliable results, and then also apply our method to DREAM5 challenge data for comparison to TIGRESS that utilizes TF information.

3.5 Conclusion

We presented two integrative approaches for gene regulatory network inference combining two different algorithms. First, IMLARF that we proposed is based on the integration of MI and LARF, which is a novel regression-based random featuring, to overcome the limitation of random sampling and MI. Secondly, ISLARF is the combination of LARF and ZS that is based on the z-score of variation of expression after the candidate regulator is knocked out. Both integrative methods outperform the standalone methods and the selected state ofthe art techniques on DREAM3 challenge data. In application to inference of the gene regulation associated with psychiatric disorders, we applied IMLARF to gene expression data and inferred the interactions between genes reported known as psychiatric disorder-associated genes and susceptible genes defined by inferred networks.

CHAPTER 4

Inference of Gene Regulatory Networks by Integrating SNP and Gene Expression

4.1 Introduction

In the previous chapter, we introduced new methods to construct biological networks with only a single type of biological data. In this chapter, a new network inference method is proposed considering the integration of two different data types that have regulatory relationships. In order to understand more accurate causal relationships between a complex disease and genetic variations, we need to consider how the genotypic perturbations affect expression phenotypes that are potentially associated with a target disease. In other words, it is more crucial to look at the overall mechanisms considering a series of three factors, which include genetic variations, altering gene regulations, and caused diseases rather than partial mappings between them. Therefore it is important to evaluate how genetic perturbations affect genes on regulatory networks that are associated with a target disease phenotype. In practice, when biological networks are inferred with high throughput data, we have to consider not only the relationships among genes but also how genetic factors such as Single Nucleotide Polymorphism (SNP) and Copy Number Variation (CNV) can affect genes in Gene Regulatory Network (GRN). Over the last decade, research for mapping genotype to expression phenotype or disease phenotype such as expression Quantitative Trait Loci (eQTL) study and Genome Wide Association Study have been actively performed [90]. However, we are now required to do a network-based analysis with genotype data and gene expression because it is more effective in discovering underlying biological process from genotype to phenotype. In doing so, the

analysis of SNP-Gene Regulatory Networks (SGRN) will provide more definite relationships of genotypic causes and phenotypic effects so that it will facilitate prognosis and drug designs for therapies.

In this chapter we propose a SGRN inference method. In order to identify regulatory interactions among genes, quite a number of network inference methods have been developed by using gene expression data such as gene microarray. Those methods can be generally classified into different theoretical categories: Boolean networks [41, 91], Mutual Information [43, 92], Bayesian networks (BN) [42, 93], and Regression [44, 94]. As each method has its own advantages and limitations under different assumptions and network models such as acyclic or cyclic network and directed or undirected network, there should be trade-offs in inferences given a different target network structure and applications [46]. For example, the MI-based approach is very simple and fast so that it can build a large scale network (e.g. genome wide scale) but it cannot estimate direction of edges. It produces worse performance than other approaches in detecting linear cascading structures [46]. The BN-based inference is limited to imply only acyclic network with high computational cost while the regression-based approach supports both directed and cyclic network, which are assumed in SGRN. In addition to directed network model, it should be considered that SGRN is different from conventional GRN inference. In SGRN inference, a gene can be regulated by SNPs as well as other genes, but SNPs are assumed not be regulated by other SNPs. That is, a SNP cannot be a child node in the network.

Recently, a number of approaches have been suggested to infer SGRNs integrating genetic variation and gene expression data. Kim et al. [95] considered genetic perturbations, gene expression, disease phenotypes together to find the causal genes to a disease. The electric circuit approach and heuristic search were used to infer SGRN where causal genes are mapped to SNP in the preliminary step before network

inference. Keurentjes et al. [96] built a snp-gene network associated with a particular phenotype, but this method also performed eQTL mapping (SNP-gene) to define the candidate regulator genes before genetic network construction. In addition, Kim and Xing [97] used lasso regression considering the case that a SNP is weakly associated with highly-correlated multiple traits rather than a single trait. Chen et al. [98] focused on identifying which pathway among those already known pathways was more likely to be affected by changes of genotype and gene expression rather than inferring a new pathway. The related works we especially noted are the methods that are based on Structural Equation Modeling (SEM) [99, 100, 101, 102]. SEM allows us to not only incorporate eQTL information to gene expression in a single model but also identify eQTL simultaneously. However, Logsdon and Mezey [101] assumed that every gene has at least one eQTL, and eQTL mapping was performed by preprocessing but not in a network inference step . Cai et al. [102] introduced sparsity-aware maximum likelihood (SML), which can be potentially extended for eQTL identification. However, SNP-gene pairs were still given in evaluations and implementations of the SML algorithm.

In this chapter, we proposed a novel method to infer SGRN where both eQTL identification as well as SGRN inference are performed simultaneously given a set of gene expression and genotype data without assuming eQTLs are known. The proposed method is based on SEM and multiple steps of edge filtering such as elastic net regression, and iterative adaptive lasso. Basically SEM is a regression-based model which is likely to select as many variables causing an overfitting, so the sparsity is enforced by lasso (l_1-regularized least square estimation) considering the sparsity of biological network. Initial weights of edges are estimated by ridge regression [103] and elastic net regression [104], and then the second step is to identify final eQTLs from candidate SNPs selected in the first steps. In the last step, the final network is

constructed by iterative adaptive lasso. The first two steps are to fix SNPs before selecting genes. In the third step, edges are selected by iteratively giving more penalties to the edge whose weight is relatively low until network structure is converged.

To evaluate the method, we explore the performance with a simulated data set that is generated from random networks with different number of samples, nodes, and expected number of edges per node. The result shows that the method can achieve a high detection rate of true edges with low false discovery rate without eQTL information. In addition, to explore the performance in real expression phenotype and SNP data, the method was applied to the psychiatric disorder data. After genes and SNPs were selected from related GWAS, it was tested how the method identify true positive edges between genes and SNPs without eQTL information.

4.2 Method

4.2.1 Problem Definitions

We define the problem and notations here. Let $Y \in \mathbb{R}^{M_g \times N}$ denote the matrix of gene expression levels of M_g genes and N samples where a row vector $\mathbf{y}_i = \{y_{i1}, ..., y_{iN}\}$ is observed expression levels of ith gene. X is $M_s \times N$ matrix to denote genotypes of individuals where $x_{ij} \in \{1, 2, 3\}$ represents the number of minor alleles of ith SNP of jth sample as an element of matrix X supposing that the number of minor alleles should be zero, one, or two in real data. So, x_{ij} represents a relative quantity of minor alleles of samples. As a gene can be regulated by other genes and genetic variations (SNPs), we define SEM as

$$\mathbf{y}_i = \mathbf{b}_i Y + \mathbf{f}_i X + \mu_i + \varepsilon_i, \tag{4.1}$$

where $\mathbf{b}_i$ denotes ith row vector of square matrix $B \in \mathbb{R}^{M_g \times M_g}$; $\mathbf{f}_i$ denotes ith row vector of square matrix $F \in \mathbb{R}^{M_g \times M_s}$; μ_i is a model bias; and ε_i is a residual modeled

as zero-mean Gaussian with a variance σ^2. As we assume there is no self-regulation (self-loop edge), $b_{ii} = 0, \forall i = 1, ..., M_g$ where b_{ii} denotes ith element of $\mathbf{b}_i$. The parameters of $\mathbf{b}_i$ and $\mathbf{f}_i$ decide the network structure defining the weight of regulation from every possible genes and SNPs to a target gene i. For example, if there is no regulation relationship (directed edge) from jth gene to ith gene, b_{ij} is set to zero. Similarly f_{ij} has non-zero value as a weight of regulation from jth SNP to ith gene if jth SNP is identified as an eQTL for ith gene. It is assumed that each gene has at least one eQTL but it is unknown which SNP among a given set of SNPs is an eQTL for a target gene. Our goal in this model is to find B and F that best fit to observed gene expression and genotype data. To make the problem simpler, we remove μ_i from (4.1) by applying mean centering for row vectors $\mathbf{y}_i$ and $\mathbf{x}_i$ to have zero mean. The goal is to find $\mathbf{b}_i$ and $\mathbf{f}_i$ that minimize a residual ε_i, so (4.1) can be expressed in a least square minimization problem as:

$$\underset{\mathbf{b}_i, \mathbf{f}_i}{\operatorname{argmin}} ||\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X||_2^2 \tag{4.2}$$

However, regression tends to select as many genes and SNPs as possible to explain the expression level of target gene i. To avoid the over-fitting, sparse regression methods such as ridge regression, elastic net, and lasso are used.

4.2.2 The algorithm

The method we propose is based on l_1-regularized linear regression known as lasso [105] that yields a sparsity of variable selection. The algorithm consists of 3 steps, (i) *elastic net*, (ii) *lasso*, and (iii) *iterative adaptive lasso*. The first two steps are to decide F where SNPs are selected but their coefficients can be changed in third step. Then, B is finalized by iterative adaptive lasso in the last step.

4.2.2.1 Ridge regression (Step 1-1)

In ridge regression, the coefficient values of irrelevant SNPs and genes to a target gene shrink to zero (but not exactly zero) while those of eQTLs and regulator genes of a target gene tend to be higher. Ridge regression of (4.2) is defined as

$$\underset{\mathbf{b}_i,\mathbf{f}_i}{\operatorname{argmin}}||\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X||_2^2 + \lambda_1||\mathbf{b}_i||_2^2 + \lambda_2||\mathbf{f}_i||_2^2 \tag{4.3}$$

Given penalty weights, λ_1 and λ_2, the optimal $\mathbf{b}_i$ and $\mathbf{f}_i$ can be obtained by closed form solution given by

$$\mathbf{f}_i = (\mathbf{y}_i - \mathbf{b}_i Y)X^T(XX^T + \lambda_2 I)^{-1}, \tag{4.4}$$

$$\mathbf{b}_i = (\mathbf{y}_i - \mathbf{f}_i X)Y^T(YY^T + \lambda_1 I)^{-1}. \tag{4.5}$$

Replacing (5) for $\mathbf{b}_i$ in (4) yields

$$\mathbf{f}_i = \mathbf{y}_i S_1(XS_1 + \lambda_2 I)^{-1}, \tag{4.6}$$

where

$$S_1 = X^T - Y^T(YY^T + \lambda_1 I)^{-1}YX^T. \tag{4.7}$$

After calculating $\mathbf{f}_i$ first in (4.6), and then (4.5) can be solved. In this manner, matrices B and F are estimated by computing each $\mathbf{b}_i$ and $\mathbf{f}_i$, $i = 1, ..., M_g$. Parameter λ_1 and λ_2 that decide the degree of sparsity of B and F are determined by K-fold cross-validation. K is set to 5 in our experiments.

4.2.2.2 Elastic net (Step 1-2)

Note that zero weighted coefficient cannot be recovered back to non-zero in adaptive lasso of step 3. Therefore, in order to carefully remain only SNPs that are

more likely to be true eQTLs in $\mathbf{f}_i$, we give l_1-norm penalty to only $\mathbf{f}_i$ but not $\mathbf{b}_i$ using elastic net defined as

$$\underset{\mathbf{b}_i,\mathbf{f}_i}{\operatorname{argmin}}||\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X||_2^2 + \lambda_1||\mathbf{b}_i||_2^2 + \lambda_2||\mathbf{f}_i||_1 \tag{4.8}$$

As the objective function is convex, which guarantees a convergence, f_{ij} can be optimized by using coordinate descent iteration given parameters, λ_1 and λ_2. To find the optimal $\mathbf{f}_i$, the derivative of (4.8) with respect to f_{ij} is considered as follows:

$$\mathbf{f}_i X X_j^T - \mathbf{y}_i X_j^T + \mathbf{b}_i Y X_j^T + \lambda_2 \partial_{f_{ij}}||\mathbf{f}_i||_1 \tag{4.9}$$

Since the derivative of (8) with respect to $\mathbf{b}_i$ is same as (4.5), $\mathbf{b}_i$ in (4.9) is substituted with (5), and then (4.9) is simplified to

$$(\mathbf{f}_{i(-j)} X_{(-j)} - \mathbf{y}_i) S_2 + f_{ij} \mathbf{x}_j S_2 - \lambda_2 \partial_{f_{ij}}||\mathbf{f}_i||_1, \tag{4.10}$$

where

$$S_2 = (Y^T (Y Y^T + \lambda_1 I)^{-1} Y - I)\mathbf{x}_j^T, \tag{4.11}$$

$\mathbf{f}_{i(-j)}$ indicates row vector $\mathbf{f}_i$ whose jth element is removed, $X_{(-j)}$ denotes matrix X whose jth row is removed, and $\mathbf{x}_j$ is jth row vector of X. After defining $C_j = (\mathbf{f}_{i(-j)} X_{(-j)} - \mathbf{y}_i) S_2$ and $a_j = \mathbf{x}_j S_2$ in (4.10), the update rule in the coordinate descent algorithm is written as

$$f_{ij} = \begin{cases} (-C_j - \lambda_2)/a_j & \text{if } C_j < \text{-}\lambda_2, \\ 0 & \text{if } C_j \leq |\lambda_2|, \\ (-C_j + \lambda_2)/a_j & \text{if } C_j > \lambda_2. \end{cases} \tag{4.12}$$

Algorithm 2 describes the procedures to solve (4.8) in step 1-2. If f_{ij} is non-zero, jth SNP is a candidate eQTL for ith gene.

4.2.2.3 Lasso (Step 2)

In order to finalize a SNP (a single non-zero f_{ij} of $\mathbf{f}_i$) for each gene i, we apply Lasso to combined matrix of Y and X as follows:

$$||\mathbf{y}_i - \mathbf{h}_i Z||_2^2 + \lambda ||\mathbf{h}_i||_1 \tag{4.13}$$

where

$$Z^T = [Y_{(-i)}^T, X_{(-k_i{}^*)}^T]. \tag{4.14}$$

$k_i{}^*$ denotes indices of low vectors where $f_{ij} = 0, j \in k_i{}^*$. So, $X_{(-k_i{}^*)}$ is a matrix X whose $k_i{}^*$ rows are removed. If the number of rows of $X_{(-k_i{}^*)}$ is greater than pre-defined heuristic number N_k (i.e. 5 in our experiments), only top N_k highest f_{ij} of absolute values of $\mathbf{f}_i$ but not all non-zero f_{ij} are selected for $X_{(-k_i{}^*)}$. In step 2, we iteratively estimate $\mathbf{h}_i$, decreasing λ from a high value that lets $\mathbf{h}_i$ have a zero vector. Regardless of elements of $\mathbf{h}_i$ for $Y_{(-i)}$, we note only which element of $\mathbf{h}_i$ for $X_{(-k_i{}^*)}$ has a non-zero value first assuming that the corresponding candidate SNP to h_{ij} is more likely to regulate a target gene i if h_{ij} for a row vector of $X_{(-k_i{}^*)}$ has non-zero value earlier than other elements of $\mathbf{h}_i$ during λ decreases.

4.2.2.4 Adaptive Lasso (subroutine of step 3)

Adaptive lasso is defined as

$$\underset{\mathbf{b}_i, \mathbf{f}_i}{\operatorname{argmin}} ||\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X||_2^2 + \lambda_1 ||\mathbf{b}_i||_{1,\mathbf{w}_i^b} + \lambda_2 ||\mathbf{f}_i||_{1,\mathbf{w}_i^f}, \tag{4.15}$$

where

$$||\mathbf{b}_i||_{1,\mathbf{w}_i^b} = \sum_j^N |b_{ij} \cdot w_{ij}^b|, \quad ||\mathbf{f}_i||_{1,\mathbf{w}_i^f} = \sum_j^N |f_{ij} \cdot w_{ij}^f|. \tag{4.16}$$

In (4.16), penalty weights, vector $\mathbf{w}_i^b$ and $\mathbf{w}_i^f$, are defined as

$$w_{ij}^b = (\hat{b}_{ij})^{-\alpha}, w_{ij}^f = (\hat{f}_{ij})^{-\beta}, \forall j = \{1, ..., M_g\} \tag{4.17}$$

Algorithm 2 Optimization for *elastic net* in step 1-2

procedure ELASTIC(Y, X, $\hat{\lambda}_1$, $\hat{\lambda}_2$, i, ε) ▷ $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are optimal parameters estimated by cross validation
  while $err > \varepsilon$ **do**
    $\mathbf{b}_i^{old} = \mathbf{b}_i$,$\mathbf{f}_i^{old} = \mathbf{f}_i$
    for $j \leftarrow 1, M_s$ **do**
      Update f_{ij} via (12)
    end for
    Update $\mathbf{b}_i$ via (5)
    $err = ||\mathbf{b}_i^{old} - \mathbf{b}_i||_2 + ||\mathbf{f}_i^{old} - \mathbf{f}_i||_2$
  end while
  return $\mathbf{b}_i$ and $\mathbf{f}_i$
end procedure

where $\hat{b}_{ij}$ and $\hat{f}_{ij}$ are estimated in step 2 that yields a sparsity to $\mathbf{f}_i$ but not $\mathbf{b}_i$. Zero coefficient of $\hat{\mathbf{f}}_i$ in step 2 is not considered as an eQTL for gene i. So, zero $\hat{f}_{ij}$ yields zero w_{ij}^f in (4.17), and then if w_{ij}^f is zero, f_{ij} will never have non-zero value in adaptive lasso of step 3 (4.16). The parameter α and β decide how much previous estimation such as $\hat{b}_{ij}$ or $\hat{f}_{ij}$ is reflected to next estimation of b_{ij} or f_{ij}. Therefore, f_{ij} that has smaller penalty weight w_{ij}^f is more likely to have non-zero value. In addition, we consider a special case that α and β is set to zero supposing that (i) we do not give a penalty weight to b_{ij} or f_{ij} by setting w_{ij}^b or w_{ij}^f to 1 if $\hat{b}_{ij}$ or $\hat{f}_{ij}$ is non-zero and (ii) we do not estimate elements of $\mathbf{b}_i$ or $\mathbf{f}_i$ by setting w_{ij}^b or w_{ij}^f to infinity if $\hat{b}_{ij}$ or $\hat{f}_{ij}$ is zero. The solution is similar to step 2 in which either $\mathbf{b}_i$ or $\mathbf{f}_i$ is optimized by coordinate

Algorithm 3 Optimization for adaptive lasso as a subroutine of step 3

procedure ADAPTIVE LASSO(Y,X,$\hat{\lambda}_1$,$\hat{\lambda}_2$,i,α,β,$\hat{\mathbf{b}}_i$,$\hat{\mathbf{f}}_i$) ▷ $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are optimal parameters preliminary estimated by cross validation

  Compute $\mathbf{w}_i^b$ and $\mathbf{w}_i^f$ ($w_{ij}^b = (\hat{b}_{ij})^{-\alpha}$,$w_{ij}^f = (\hat{f}_{ij})^{-\beta}$)

  while $err > \varepsilon$ **do**

    $\mathbf{b}_i^{old} = \mathbf{b}_i$,$\mathbf{f}_i^{old} = \mathbf{f}_i$

    for $j \leftarrow 1, M_g$ **do**

      Update b_{ij} via (19)

    end for

    for $j \leftarrow 1, M_s$ **do**

      Update f_{ij} via (20)

    end for

    $err = ||\mathbf{b}_i^{old} - \mathbf{b}_i||_2 + ||\mathbf{f}_i^{old} - \mathbf{f}_i||_2$

  end while

  return $\mathbf{b}_i$ and $\mathbf{f}_i$

end procedure

descent algorithm but it is applied to solve both $\mathbf{b}_i$ and $\mathbf{f}_i$ in step 3. Derivative of (4.15) with respect to b_{ij} yields

$$\mathbf{b}_i Y \mathbf{y}_j^T - \mathbf{y}_i \mathbf{y}_j^T + \mathbf{f}_i X \mathbf{y}_j^T + \lambda_1 \partial_{\mathbf{b}_{ij}} ||\mathbf{b}_i||_{1,\mathbf{w}_i^b}$$

$$= b_{ij} \mathbf{y}_j \mathbf{y}_j^T + (\mathbf{b}_{i(-j)} Y_{(-j)} - \mathbf{y}_i + \mathbf{f}_i X) \mathbf{y}_j^T + \lambda_1 \partial_{b_{ij}} ||\mathbf{b}_i||_{1,\mathbf{w}_i^b}$$

where $\mathbf{b}_{i(-j)}$ indicates row vector $\mathbf{b}_i$ whose jth element is removed, and $Y_{(-j)}$ denotes matrix Y whose jth row is removed. After setting $C_j^b = (\mathbf{b}_{i(-j)}Y_{(-j)} - \mathbf{y}_i + \mathbf{f}_i X)\mathbf{y}_j^T$ and $a_j^b = \mathbf{y}_j\mathbf{y}_j^T$, the update rule for b_{ij} is as follows:

$$b_{ij} = \begin{cases} (-C_j^b - w_{ij}^b \cdot \lambda_1)/a_j^b & \text{if } C_j^b < \text{-}w_{ij}^b \cdot \lambda_1, \\ 0 & \text{if } C_j^b \leq |w_{ij}^b \cdot \lambda_1|, \\ (-C_j^b + w_{ij}^b \cdot \lambda_1)/a_j^b & \text{if } C_j^b > w_{ij}^b \cdot \lambda_1. \end{cases}$$

We can also estimate f_{ij} in similar way. After define $C_j^f = (\mathbf{f}_{i(-j)}X_{(-j)} - \mathbf{y}_i + \mathbf{b}_i Y)\mathbf{x}_j^T$ and $a_j^f = \mathbf{x}_j\mathbf{x}_j^T$, the update rule for f_{ij} is given as

$$f_{ij} = \begin{cases} (-C_j^f - w_{ij}^f \cdot \lambda_2)/a_j^f & \text{if } C_j^f < \text{-}w_{ij}^f \cdot \lambda_2, \\ 0 & \text{if } C_j^f \leq |w_{ij}^f \cdot \lambda_2|, \\ (-C_j^f + w_{ij}^f \cdot \lambda_2)/a_j^f & \text{if } C_j^f > w_{ij}^f \cdot \lambda_2. \end{cases}$$

When $\mathbf{b}_i$ and $\mathbf{f}_i$ are updated, updated single element b_{ij} or f_{ij} immediately affect to updating the next elements. In addition, updating order of elements can be changed since convex objective function is converged in any order of elements to update. Algorithm 3 shows the optimization procedure of adaptive lasso.

4.2.2.5 Iterative Adaptive Lasso (Step 3)

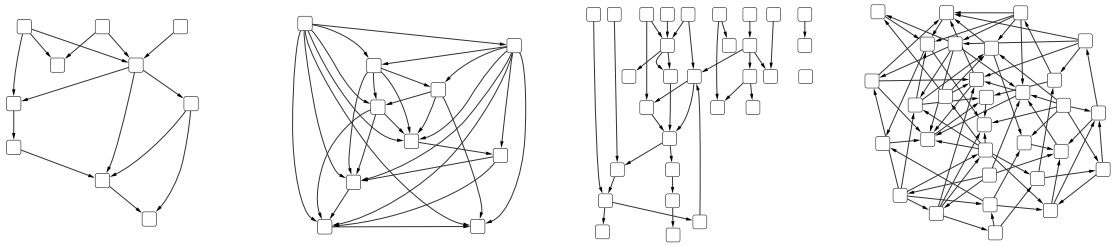
Even if $\mathbf{b}_i$ and $\mathbf{f}_i$ are estimated in step 1 and 2, there should be still many false positive edges yet. The primary goal of step 1 and 2 is to carefully get rid of only edges that are more unlikely to be true positive edges. So, instead of simply applying adaptive lasso, we developed iterative adaptive lasso to improve the performance of naive adaptive lasso. The motivation of iterative adaptive lasso is that the coefficient value of the variable considerably depends on the value of α and β which are fixed to 1 and 0.5 in [101, 102] respectively. In iterative adaptive lasso, adaptive lasso is

Algorithm 4 Iterative Adaptive Lasso in step 3

1: **procedure** ITERATIVE ADAPTIVE LASSO(Y,X,$\hat{B}$,$\hat{F}$) ▷ $Ne(B)$ denote the number of non-zero elements in B and F
2: $[B^R,F^R] = Ridge(Y,\ X,\ \hat{\Lambda}_1^R,\ \hat{F})$ in (3)
3: $\alpha = 1,\ \beta = 1$
4: **for** $i \leftarrow 1, M_g$ **do**
5: $[\mathbf{b}_i,\ \mathbf{f}_i] = AdaptiveLasso(Y,\ X,\ \lambda_1 = 0.001,\ \lambda_2 = 0,\ i,\ \alpha,\ \beta,\ \mathbf{b}_i^R,\ \mathbf{f}_i^R)$ in (15)
6: **end for**
7: **while** $N_e(B)$ are decreased by increased α **do**
8: **while** $N_e(B)$ are decreased **do**
9: **for** $i \leftarrow 1, M_g$ **do**
10: $[\mathbf{b}_i,\ \mathbf{f}_i] = AdaptiveLasso(Y,\ X,\ \hat{\lambda}_1,\ \lambda_2 = 0,\ i,\ \alpha,\ \beta,\ \mathbf{b}_i,\ \mathbf{f}_i)$ in (15)
11: **end for**
12: **end while**
13: $\alpha = \alpha + 1$
14: **end while**
15: **return** B and F
16: **end procedure**

iteratively applied incrementally changing α and β until there is no more change in the total number of selected edges of B and F so that more coefficients of irrelevant variables can be shrunk to zero.

Algorithm 4 presents a detailed procedure of iterative adaptive lasso. $\hat{B}$ and $\hat{F}$ estimated in step 2 are used as arguments. On line 2, B and F are initialized by Ridge regression. Λ_1^R is a vector of optimal parameters of λ_1 for B^R in (3) but there is no penalty to F^R ($\Lambda_2^R = 0$). For F^R we estimate only non-zero elements of $\hat{F}$ that



(a) M=10, E_g=1 (b) M=10, E_g=3 (c) M=30, E_g=1 (d) M=30, E_g=3

Figure 4.1: Example of simulated networks with different parameter settings. M and E_g indicate the number of genes and expected number of edges per node respectively.

is estimated in step 2. Again, B and F are initialized by adaptive lasso in order that elements of B are updated by weights of B^R. In this initialization, b_{ij} that has a small value can shrink to zero. Based on updated B and F, Λ_1 (a vector of λ_1 for B on line 9) is estimated again by cross validation of adaptive lasso before line 6 starts. Initially the second *while* loop updates B until no change in $Ne(B)$. Once the second *while* loop is terminated, α is increased, and then the second loop is performed again. If the second *while* loop is terminated without any change of $Ne(B)$, the first *while* loop is terminated.

4.3 Results

4.3.1 Simulation Studies

To evaluate the proposed method, we first perform simulations based on randomly generated acyclic networks. The simulation settings are similar to [101, 102]. M denotes the number of genes and SNPs and is set to 10, 20, and 30. $M \times N$ matrix B is initialized to zero matrix where N is a sample size, then elements of B are randomly selected as directed edges. The selected b_{ij} has random coefficient value uniformly distributed over 0.5$\sim$1 or -0.5$\sim$-1. Since we consider a single eQTL per gene (E_s=1), a single element (f_{ii}) is selected from each row vector ($\mathbf{f}_i$). So, F is a

diagonal matrix. x_{ij} is randomly set as 1, 2, or 3 with the probabilities 0.25, 0.5, and 0.25, respectively. Y is generated by calculating $Y = (I - B)^{-1}(FX + E)$ where E_{ij} is generated from Gaussian distribution with zero mean and variance 0.01. The number of samples for each network size is N=100, 200, 300, 400, 500. The number of edges per gene on average is set to E_g=1, 2, and 3. Given data Y and X, performances of predicting B and F are evaluated by comparing true network and inferred network.

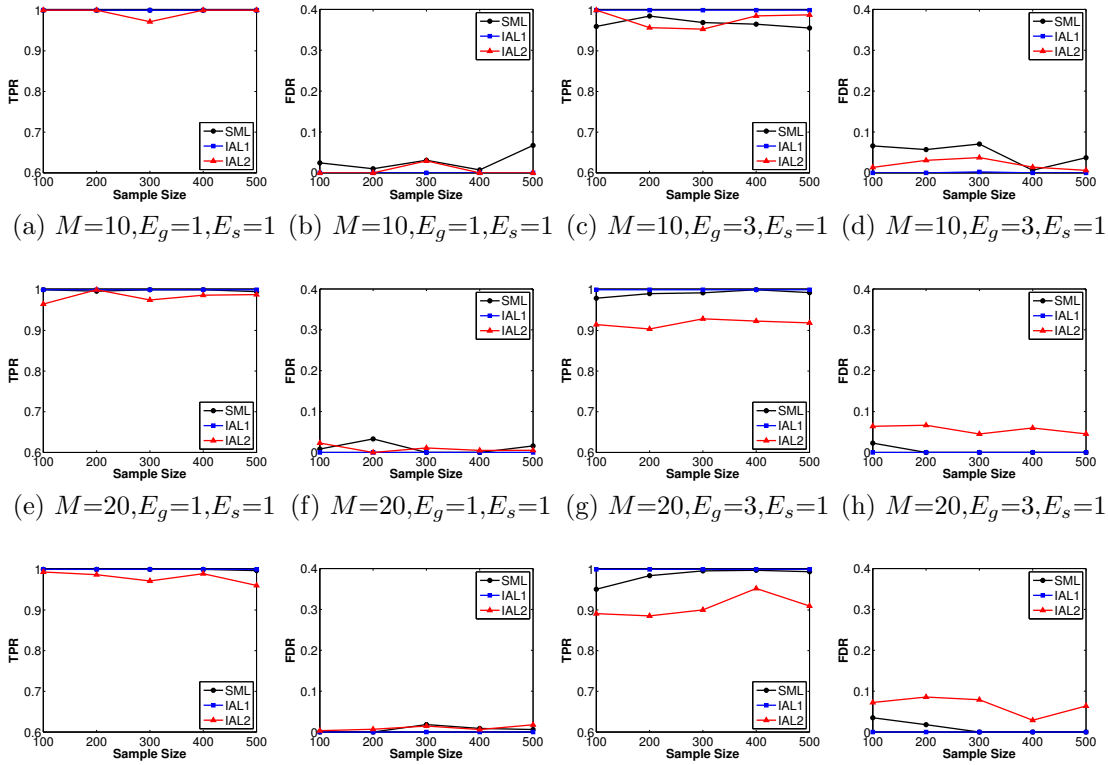
Figure 4.1 displays the examples of networks, where SNP nodes are excluded. For the evaluation, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) edges are counted to measure the accuracy criteria such as True Positive Rate (TPR) and False Discovery Rate (FDR) that are defined as

- TPR=TP/(TP+FN)
- FDR=FP/(TP+FP)

In order to evaluate our method, IAL is compared to SML [102]. As SML infers only B with known non-zero element indices of F, we consider two versions of IAL, IAL without eQTL information and IAL with eQTL information where step 1 and 2 are skipped and only step 3 is performed with non-zero element index of $\mathbf{f}_i$. SML is tested by using the code the author implemented in [102]. The abbreviations of algorithms to compare in Figure 4.2 and Table 3.1 are listed below:

- SML: Sparsity-aware Maximum Likelihood algorithm with eQTL information [102]
- IAL1: IAL with eQTL information
- IAL2: IAL without eQTL information

Ten replicate simulations are performed and each simulation has a different topology. The results of the different settings (M and E_g) are displayed in Figure 4.2. It is shown that IAL1 is superior to SML in all data sets regardless of sample size. We also note that TPR of IAL2 is higher than 0.9 and FDR is less than 0.1 on average



(a) M=10,E_g=1,E_s=1 (b) M=10,E_g=1,E_s=1 (c) M=10,E_g=3,E_s=1 (d) M=10,E_g=3,E_s=1

(e) M=20,E_g=1,E_s=1 (f) M=20,E_g=1,E_s=1 (g) M=20,E_g=3,E_s=1 (h) M=20,E_g=3,E_s=1

(i) M=30,E_g=1,E_s=1 (j) M=30,E_g=1,E_s=1 (k) M=30,E_g=3,E_s=1 (l) M=30,E_g=3,E_s=1

Figure 4.2: True positive rate and false discovery rate under different numbers of edges and nodes.

in any sample size. It validates that the proposed IAL works very effectively when eQTL is known. In addition, the performance of IAL1 is consistent in different sample sizes while the performance of SML tends to be decreased with small sample size and complicate network (E_g=3). In network inference, it is known that the performance of inference is very sensitive to the network size and density. In the inference of densely connected and large networks, the computational cost will exponentially increase and the FDR may increase because there are more possible variables that may explain a target node better than true regulators. IAL1 performed consistently in all three different network size while the performance of SML is affected by the network size in dense networks (E_g=3). However, IAL2 shows consistent TPRs and FDRs in all

Table 4.1: TPR and FDR of SML, IAL1, and IAL2

N	M	TPR			FDR		
		SML	IAL1	IAL2	SML	IAL1	IAL2
100	10	0.9888	1.0000	0.9742	0.0860	0	0.0104
	20	0.9980	1.0000	0.9448	0.0503	0	0.0292
	30	0.9951	1.0000	0.8936	0.0364	0	0.0754
500	10	0.9967	1.0000	1.0000	0.0704	0	0
	20	0.9850	1.0000	0.9436	0.0400	0	0.0369
	30	1.0000	1.0000	0.9128	0.0016	0	0.0562

Expected number of edges per node is E_g=2 and 10 replicates of random network are used. N and M indicate the number of samples and genes respectively.

three different network sizes when the network density is normal (E_g=1) while TPR of IAL2 in Figure 4.1(g) and (k) is lower than (c) and also FDR increases in Table 3.1 when the network size increases in more dense networks (E_g=2). The result shows that the performance is better in sparse networks (E_g=1) than dense networks (E_g=3) because a complicate structure is more likely to cause false positive edges because of indirect regulations. For example, TPRs in Figure 4.1(a), (e), and (i) are much better than (c), (g), and (k). Similarly FDR is quite increased with E_g=3 in Figure 4.1(d), (h), and (l) compared to the case of $E_g = 1$ in Figure 4.1(b), (f), and (j).

Overall results imply that the proposed IAL1 works perfectly with known F in any network size and density. It means that the performance of IAL2 is significantly affected by false positive inference of F in step 1 and 2 because of unknown F. More precisely $\mathbf{b}_i$ without sparsity in step 2 is more likely to have false positive non-zero elements even though a number of candidate elements of $\mathbf{b}_i$ are filtered in step 1. Therefore the selection of non-zero element of $\mathbf{b}_i$ in IAL2 is the most critical part since IAL1 is able to correctly infer B only if F is given as eQTL information.

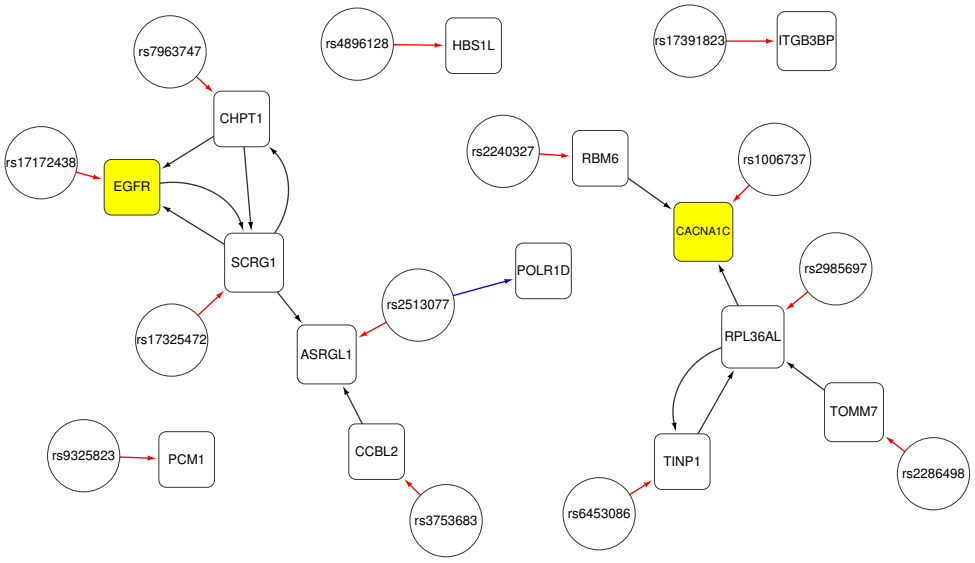


Figure 4.3: The inferred SGRN with 14 pairs of gene and SNP selected from [1, 2, 3]

4.3.2 Experiments with Psychiatric Disorder Data

In this section, the proposed method is applied to real gene expression and genotype data for psychiatric disorder. In the application to real data, we explore the performance of GRN inferences and eQTL identifications through the inferred networks. As far as we know, the proposed method is the first solution to provide both GRN inference and eQTL identification. Thus, the performance comparison with other methods was not performed. The psychiatric disorder data consists of gene expression data of 25833 genes and 852963 SNPs for 131 samples, which were measured from human brain. Since we focus on the network inference but not gene selection, the network construction is performed with a pre-defined set of genes and SNPs that are selected by preliminary test of multiple sets of genes and eQTLs based on related GWAS for psychiatric disorders. The result of SGRN inference is displayed in Figure 4.3 where two yellow colored genes, EGFR and CACNA1C, are selected from [2, 3] and the rest of two pairs are from [1]. In applying IAL2 to the data, the

weights of α and β are set to 0.5 instead of 1. Otherwise, $N_e(\mathbf{f}_i)$ tend to be zero. The reason for this is that gene variables are more correlated their eQTLs because generally eQTLs are independently selected to other genes. In Figure 4.3, SNP and gene are distinguished by node shape, and a red edge indicates a correct edge from eQTL to corresponding gene. A blue edge represents false positive eQTL mapping. For eQTL identification, one false positive edge appears and thirteen true positive edges are detected (TPR=0.9286, FDR=0.0714).

4.4 Discussion

The most difficult part in network inference is to identify directions of edges. In the adjacency matrix B, both B_{ij} and B_{ji} could have a high coefficient value. In this case, regression-based methods tend to show better performance than MI-based methods because candidate edges are evaluated together in regression-based methods but each edge is independently evaluated to other edges in MI-based methods. Despite of the advantage, the regression-based method needs to be integrated with other methods that can provide a different information of structure. Another issue to improve in IAL is the computational cost to estimate two different λs per each row. Intuitively, a searched optimal λ per each row of B and F should provide a better result but it causes a high computation cost. Lastly, we also assumed that a gene has at least a single eQTL given a set of genes and SNPs, but multiple eQTL should be considered and a gene may not have any eQTL in practice. Thus, the multiple eQTL of a gene is a future work in SGRN inference.

4.5 Conclusion

In this chapter, we proposed a novel network inference method that provides both eQTL identification and network construction of both genes and SNPs. In order to understand gene regulatory mechanisms for a target disease phenotype, the regulatory network inference needs to consider effect of genetic variation and expression phenotype together but not only gene expression data. To achieve the high quality of reliable inference with better TPR and FDR, three different regression skills are integrated. Ridge regression and elastic net are used to remove more likely false positive edges and select eQTL as preliminary steps, and then the finial network is estimated by iterative adaptive lasso removing more false positive edges between genes. Through the experiments with synthetic data, it was demonstrated that IAL1 outperforms SML in SGRN inference and also IAL2 performs eQTL identification effectively. The method was also applied to psychiatric disorder data. Using the genes and eQTLs selected from GWAS of psychiatric disorder, we explored the ability of eQTL identification through inferred SGRN.

CHAPTER 5

Learning Structure of Bayesian Network Classifier and Application to Personalized Medicine and Biomarker Identification for Lung Cancer

5.1 Introduction

While the proposed methods were to integrate the methods or data in previous chapters, we discover biomarkers based on integration of multiple networks, which are inferred for the application, Personalized Medicine (PM) in this chpater.

5.1.1 Personalized Medicine

The goal of PM is to provide a medicine that is customized to individual patients considering biological features of the patients [106]. In this chapter, our goal is to predict drug sensitivity levels of cancer patients in order to provide an optimal drug to the patients avoiding a waste of time with ineffective treatments. It is assumed that different drug sensitivity of cancer patients are derived from different biomolecular characteristics of the patient's cancer. In other words, proteomic profiling can provide important pathophysiologic cues regarding responses to chemotherapies since medicinal effect is closely relevant to signal transduction pathways of a cancer [107]. For biological profiling of patients, quantitative patterns of protein expression are measured by using a protein microarray.

To quantitatively measure the expression level of proteins, Reverse Phase Protein Array (RPPA) is used in conjunction with the quantum dots (Qdot) nanotechnology. RPPA was originally introduced by Liotta [108] and is designed for quantitative measurement of protein expression. In RPPA, sample lysates are im-

mobilized in series of dilutions to generate dilution curves with only small amount (nanoliter) of the sample. After primary and secondary antibodies are probed, signal is detected by Qdot assays. Qdot is a nano-metal fluorophore that provides more bright and enhanced linear signal. In addition, RPPA is able to measure posttranslational modifications for more accurate pathophysiologic information in a signaling pathway [109].

Figure 5.1 describes the framework of PM. In the first step, a classifier is trained by RPPA and drug sensitivity data where proteins are considered as attributes and a drug is as class label. In the second step, RPPA of a patient's sample is tested by the trained classifier, then in the last step, the classifier predicts high or low as a drug sensitivity of the given test sample. Based on the result of the predictions for multiple drugs, a set of drugs that are more likely to have low sensitivity is recommended. For the classification, we employed Bayesian Network Classifier (BNC) that consists of two components, (i) parameters and (ii) network structure. Since the networks of BNC represents the dependency of proteins, these multiple networks of BNCs for multiple drugs also provide important information of relationships between proteins in order to identify the biomarkers of a target cancer from the multiple networks. In the following subsection, we present the background of BNC and propose a new method to build an improved BNC in the prediction.

5.1.2 Bayesian Network Classifier

BNC is one of the most popular models for classification algorithms. The basic model of BNC is Naive Bayes classifier (NB) [110], which is based on a simple form of Bayes theorem and relatively competitive to state-of-the-art classifiers such as Support Vector Machine (SVM) [111] and k-nearest neighbors (kNN) [112] in the

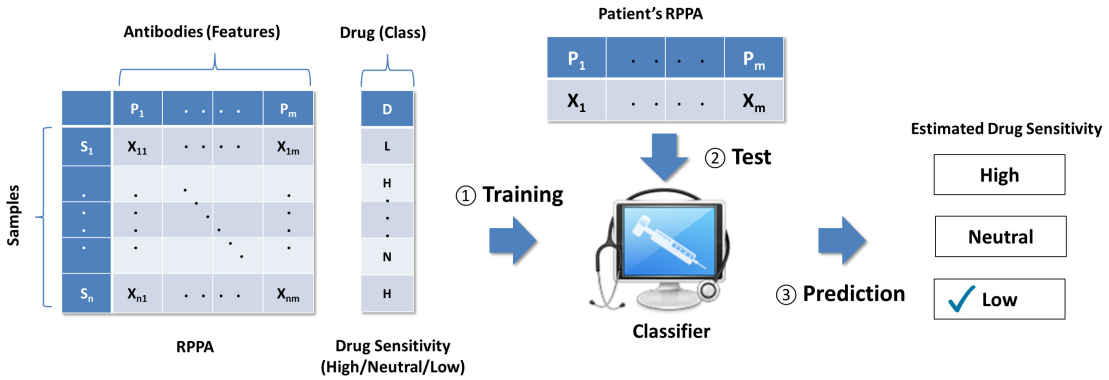


Figure 5.1: The framework of personalized medicine

performance comparison [113]. These attractive potentials of NB have been noted and BNC has been constantly improved by researchers.

NB is based on the assumption that all attributes (attribute, variable, and node are interchangeably used in this chapter) are conditionally independent to each other given a class label. To more intuitively understand the conditional independence between attributes, NB is represented as a Bayesian Network in Figure 5.3 where child nodes are conditionally independent to each other given a class node. However, the assumption is impractical in real applications [114, 115], where some attributes are more likely to be related to each other. Therefore, it is important to consider the dependencies between attributes and these additional information could be utilized to improve classification accuracy. In other words, NB can be enhanced by augmenting edges which are shown as dotted arrows in Figure 5.4. Section 5.2.4 presents how NB is modified by these additional edges in detail.

Basically building a BNC consists of three steps, (i) learning a structure of Bayesian Network (the process to define edges is referred to as *learning structure* of BNC), (ii) estimating parameters (i.e. conditional probability tables), and then (iii)

predicting the class label that maximizes the posterior probability of Bayes' theorem given a test instance as described in Section 5.2.2.

During past years, many methods for learning parameters have been suggested, and they assumed that discriminative parameters can lead to better prediction accuracy than the Maximum Likelihood (ML)-based parameter estimation [116, 117, 118, 119]. However, it is worth to note that Pernkopf's experimental analysis [120] implies that the discriminative parameter may not be effective if the discriminative structure is already sufficient to classify. Although it is true that the discriminative parameter learning is crucial in BNC research, we focus on only learning structure in this chapter.

As an pioneering work in the BNC studies, Friedman et al. [121] showed that an *unrestricted* Bayesian network (Fig. 2), which is constructed by conventional learning Bayesian network method [122], is not discriminative in their experiments. It was also demonstrated that additional edges between attribute nodes in NB can be a more discriminative structure. Since the added edges forms a tree structure as assumed, the network structure they proposed is called Tree Augmented Naive Bayes (TAN). The algorithm to build a TAN structure is based on Maximum Spanning Tree (MST) that maximizes likelihood of Bayes theorem. In classification, however, a higher likelihood does not guarantee a lower error. Instead, Conditional Likelihood (CL) of the class labels given the features can be used to build more discriminative structure for a lower error [123]. In other words, the structure that maximizes CL is more likely to be discriminative than likelihood. Nevertheless, CL has been only approximately or heuristically estimated by researchers as it has been known that there is no known closed form that is decomposable into each variable. Grossman et al. [124] suggested the method in which greedy algorithm is employed to search the structure that maximizes CL. Although the experiments results confirm that

CL-based structure is more likely to reduce error in prediction than likelihood, the heuristic searching and scoring algorithm are computationally too expensive to be applied to a large number of samples and attributes. Carvalho et al. [125] proposed the score criteria for the approximation of CL and it was empirically proved by the experimental results that CL-based discriminative structure learning of BNC could outperform state-of-the-art classifiers.

In this chapter, we introduce a learning structure method based on Conditional Mutual Information (CMI). As our main contribution, it is proved that CL can be decomposable when the structure is TAN. The decomposed criteria is expressed by CMI. In our work, CMI we propose is referred to as Discriminative CMI (DCMI) in order to distinguish it from Friedman's CMI, which is derived from the decomposed likelihood but not CL. In addition, DCMI-based TAN structure is improved by removing edges. It means that the final structure could be a partial 1-tree but not TAN structure. In order to demonstrate the superior performance of the proposed methods to state-of-the-art methods, we performed the experiments with twenty five different well-known benchmark data sets. The results confirms that DCMI-based method achieves significant improvements on classification accuracy.

5.2 Preliminaries

5.2.1 Bayesian Network

Bayesian network is Directed Acyclic Graph (DAG) that encodes the joint probability distribution over a set of random variables $\mathbf{X} = \{X_1, ..., X_n\}$ where n is the number of variables. In this chapter we assume all variables are discrete or discretized as a preprocessing. Bayesian network is defined by a pair $B = (G, \Theta)$. The first component G is a network structure where nodes and directed edges represent variables

and dependencies between variables respectively. In a given structure G, a set of parent variables of X_i is denoted by Π_{X_i}, and X_i has an element $x_{ik} \in \{x_{i1}, ..., x_{ir_i}\}$ where x_{ik} is the kth state of the variable X_i and r_i is the number of possible states of X_i. The second component Θ is a set of parameters for local conditional probability distributions representing the probability of a state of the variable X_i given states of the variable's parents Π_{X_i}. Parameters are defined as

$$P_B(X_i = x_{ik} | \Pi_{X_i} = \pi_{ij}) = \theta_{ijk} \tag{5.1}$$

where $\pi_{ij} \in \{\pi_{i1}, ..., \pi_{iq_i}\}$ is the jth *parent configuration* (the states of parents) of Π_{X_i} and q_i is the number of possible parent configuration given Π_{X_i} ($q_i = \Pi_{X_h \in \Pi_{X_i}} r_h$). The structure of Bayesian network defines a joint probability distribution over X given by the product of local distributions as following:

$$P_B(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | \Pi_{X_i}). \tag{5.2}$$

More precisely the local conditional distributions in the left-hand side pertaining the joint distribution encodes the network structure. These local conditional independences can be defined as that each variable is conditionally independent of its non-descendants given the state of its parents. For example, X_i is conditionally independent of its non-descendants given its parents Π_{X_i} in G. These property of Bayesian Network reduce the number of parameter and the computational cost of posterior probabilities.

5.2.2 Bayesian Network Classifier

BNC is a probabilistic classifier based on Bayes' theorem and Bayesian Networks. A set of random variables is defined as $\mathbf{X} = \{X_1, ..., X_n, C\}$ where C is a class

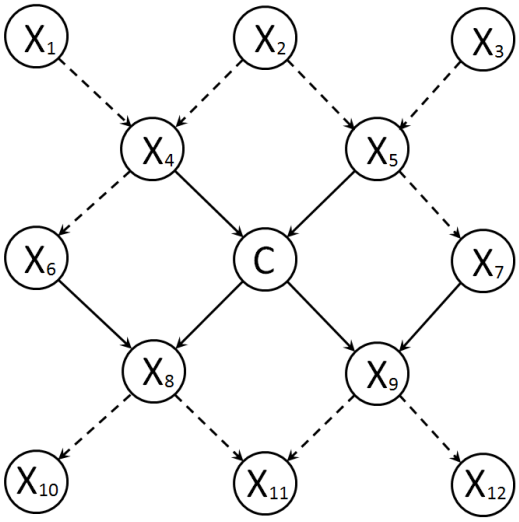


Figure 5.2: Example of unrestricted Bayesian networks. The solid edges indicate the edges between the class node and the Markov Blanket of the class node. Only the Markov Blanket of the class node provides the information about the class node in Bayesian Network Classifier.

variable. The goal of BNC is to predict c ($c \in C$) that maximizes $P_B(c|x_1^{'}, ..., x_n^*)$ given a test instance $(x_1^*, ..., x_n^*)$. Thus, BNC is defined as

$$\arg\max_{c \in C} P_B(C|X_1, \ldots, X_n) = \arg\max_{c \in C} \frac{P_B(C, X_1, \ldots, X_n)}{P_B(X_1, \ldots, X_n)}. \tag{5.3}$$

Since $P_B(X_1, \ldots, X_n)$ is independent to the decision of class C, the normalization constant $P_B(X_1, \ldots, X_n)$ can be canceled from (5.3) in the classifier. Then, BNC can be redefined as

$$\arg\max_{C} P_B(C|X_1, \ldots, X_n) = \arg\max_{C} P_B(C, X_1, \ldots, X_n). \tag{5.4}$$

Using the joint probability (5.2) of Bayesian Network, we can re-express (5.4) as

$$\arg\max_{C} P_B(C|\Pi_C) \prod_{i=1}^{n} P_B(X_i|\Pi_{X_i}). \tag{5.5}$$

5.2.3 Generative Model vs Discriminative Model

Let us consider a Bayesian network structure in Figure 5.2. In order to predict the class node C given the values of other nodes, we can define the BNC equation (5.5) as following:

$$\arg\max_{C} P(C|\Pi_C) \prod_{i=1}^{12} P(X_i|\Pi_{X_i}) \tag{5.6}$$

$$\begin{aligned} = \arg\max_{C} & P(C|X_4, X_5)\cdot P(X_1)\cdot P(X_2)\cdot P(X_3)\cdot \\ & P(X_4|X_1, X_2)\cdot P(X_5|X_2, X_3)\cdot P(X_6|X_4)\cdot \\ & P(X_7|X_5)\cdot P(X_8|X_6, C)\cdot P(X_9|C, X_7)\cdot P(X_{10}|X_8)\cdot \\ & P(X_{11}|X_8, X_9)\cdot P(X_{12}|X_9). \end{aligned} \tag{5.7}$$

Since the local conditional distributions that do not include C are constant, we can cancel all those local distributions from the classifier (5.6) as given by

$$\arg\max_{C} P(C|X_4, X_5)\cdot P(X_8|X_6, C)\cdot P(X_9|C, X_7). \tag{5.8}$$

In this case, the classifier uses only the Markov blanket of C because discarded variables are irrelevant to C. It means that the performance of BNC depends on the structure of BN. However, the most of algorithms for learning BN structure is to build a network, which is referred to as the *generative* model for the representation of a given data but not the *discriminative* model for classifications. In fact, Friedman [121] verified these limitations of generative model with the empirical results where *unrestricted* Bayesian network structures as a generative model cannot provide an effective structure for BNC in 25 data set from UCI Machine Learning repository [126].

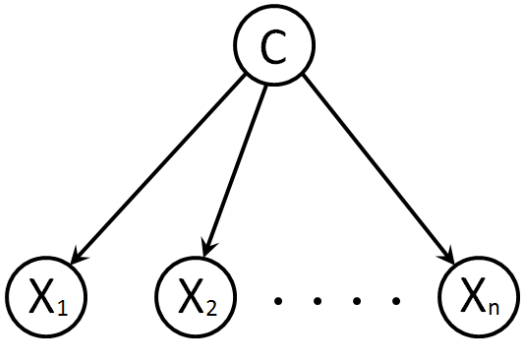


Figure 5.3: The network structure of the Naive Bayes Classifier

5.2.4 Naive Bayes Classifier and Tree Augmented Naive Bayes

In order to overcome the limitation of the unrestricted structure-based BNC (Figure 5.1), alternatively Naive Bayes Classifier (NBC) [110] has been applied in the variety of applications. NBC provides the competitive performance to the state of the art in classifications. NBC is a special case of BNCs assuming all variables are conditionally independent to each other variable given C. NBC has a simple Bayesian network structure shown in Figure 5.3 as a restricted structure where all variables have information for C. Based on the structure, BNC (5.5) can be re-written as

$$\arg\max_{C} P(C|X_1, \ldots, X_n) = \arg\max_{C} P(C) \prod_{i=1}^{n} P(X_i|C). \tag{5.9}$$

However, this conditional independence assumption is quite unreasonable since there could be correlations between variables (attributes) in real world applications. Friedman et al. [121] suggested Tree Augmented Naive Bayes (TAN) where extra edges between variables are added to the structure of NBC. More precisely the augmented edges between variables are restricted to a tree structure in TAN. For example, the structure with dotted edges in Figure 5.4 is supposed to be a tree in TAN. Each variable has at most two parents nodes (i.e. class variable as a default parent and one other variable). Only class variable has no parent node. To build the maximum likelihood TAN struc-

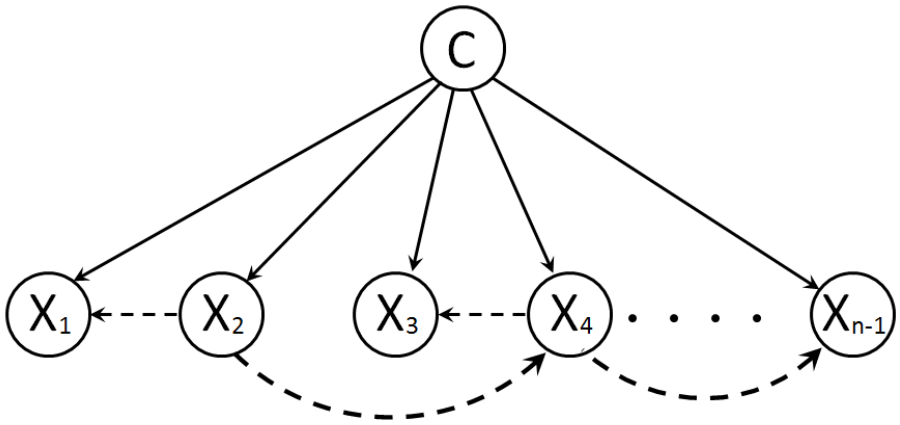


Figure 5.4: Example of TAN structure. The dotted edges compose a tree structure.

ture, the first step is to construct a Maximum weighted Spanning Tree (MST) where the weight of each undirected edge between two variables, X_i and X_j ($X_i, X_j \in \mathbf{X}$), is computed by Conditional Mutual Information (CMI), $I(X_i; X_j|C)$. The second step is to decide the direction of MST by setting the orientation of all edges to be outward from a root, and then add a class variable to the tree and add edges from class variable to all variables. It is summarized how CMI-based MST provides maximum likelihood TAN structure as following.

The goal of maximum likelihood TAN is to find the TAN structure that maximizes likelihood of the structure given data. The likelihood function (L) and its logarithm (LL) are given by

$$L(B|D) = \prod_{t}^{N} P(C^t) \prod_{i}^{n} P(X_i^t|\Pi_i^t, C^t), \tag{5.10}$$

$$LL(B|D) = \sum_{t}^{N} \log P(C^t) + \sum_{t}^{N} \sum_{i}^{n} \log P(X_i^t|\Pi_i^t, C^t) \tag{5.11}$$

where N is the number of instance. Since the first term of the right hand side in (5.10) is constant with respect to Π_i, the goal is to find Π_i that maximizes

$\sum_t^N \sum_i^n \log P(X_i^t|\Pi_i^t, C^t)$ given i. Therefore, the objective function can be redefined as

$$\arg\max_{\Pi_i} \sum_t^N \sum_i^n \log P(X_i^t|\Pi_i^t, C^t) \tag{5.12}$$

$$= \arg\max_{\Pi_i} N \sum_i^n P(X_i, \Pi_i, C) \log P(X_i|\Pi_i, C) \tag{5.13}$$

$$= \arg\max_{\Pi_i} -N \sum_i^n H(X_i|\Pi_i, C) \tag{5.14}$$

$$= \arg\max_{\Pi_i} -N \sum_i^n H(X_i) + N \sum_i^n I(X_i; \Pi_i, C). \tag{5.15}$$

Since the first term of the right hand side in (5.14) and N is constant with respect to Π_i, we can restate (5.14) as following:

$$\arg\max_{\Pi_i} \sum_i^n I(X_i; \Pi_i, C) \tag{5.16}$$

$$\arg\max_{\Pi_i} = \sum_i^n I(X_i; C) + I(X_i; \Pi_i|C) \tag{5.17}$$

$$\propto \arg\max_{\Pi_i} \sum_i^n I(X_i; \Pi_i|C) \tag{5.18}$$

5.3 Methods

5.3.1 Discriminative Structure Learning

5.3.1.1 Conditional Log-Likelihood Score Function

Our goal is to build a Bayesian network that maximizes CL by augmenting edges to the structure of NBC. We assume that edges augmented in NBC compose a tree structure like TAN. CL is reformed to a entropy equation as follows:

$$CL = \prod_{t=1}^{N} P(c^t | x_1^t, \ldots, x_n^t) \tag{5.19}$$

$$CLL = \sum_{t=1}^{N} \log P(c^t | x_1^t, \ldots, x_n^t) \tag{5.20}$$

$$= N \cdot \sum_{C,X} P(C, X_1, \ldots, X_n) \log P(C | X_1, \ldots, X_n)$$

$$= -N \cdot H(C | X_1, \ldots, X_n) \tag{5.21}$$

where H is entropy.

Proposition 1. Given the order of nodes, conditional log-likelihood for Bayesian network classifier can be decomposed into each node as follows:

$$-H(X_i | C, X_{i+1}, \ldots, X_n) + H(X_i | X_{i+1}, \ldots, X_n) \tag{5.22}$$

Given a TAN structure we can define the topological order of node in which node X_i can have parent node X_j only if i is earlier than j in the order ($X_i \prec X_j$). Now using following conditional entropy rule given by

$$H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z). \tag{5.23}$$

CLL (5.20) can be re-expressed as

$$-H(X_1 | C, X_2, \ldots, X_n) + H(X_1 | X_2, \ldots, X_n)$$
$$- H(C | X_2, \ldots, X_n)$$

where constant N is cancelled. Note that the third term is same form as $H(C|X_1, \ldots, X_m)$. Similarly we can expand $-H(C|X_2, \ldots, X_n)$ to

$$-H(X_2|C, X_3, \ldots, X_n) + H(X_2|C, X_3, \ldots, X_n) - H(C|X_3, \ldots, X_n).$$

Therefore we can iteratively decompose CLL into each node in the order:

$$\begin{aligned} & - H(C|X_1, \ldots, X_n) && (5.24) \\ = & - H(X_1|C, X_2, \ldots, X_n) + H(X_1|X_2, \ldots, X_n) \\ & - H(X_2|C, X_3, \ldots, X_n) + H(X_2|X_3, \ldots, X_n) \\ & \qquad \vdots \\ & - H(X_{n-1}|C, X_n) + H(X_{n-1}|X_n) \\ & - H(C|X_n) && (5.25) \end{aligned}$$

In Bayesian network, the node X_n of the last term is not decomposed since at least one node does not have parent. Now we can generalize (5.23) with two entropy terms for each node as follows.

$$\sum_{i}^{n-1} [H(X_i|X^{\setminus 1,\ldots,i}) - H(X_i|C, X^{\setminus 1,\ldots,i})] - H(C|X_n) \qquad (5.26)$$

Before we discuss more about learning TAN structure for BNC using (5.25), consider CLL of NB structure first.

Lemma 1. conditional log-likelihood for NB is equivalent to

$$\sum_{i}^{n-1} [-H(X_i|C) - H(X_i)] - H(C|X_n) \qquad (5.27)$$

The first term of (5.24) can be re-expressed in

$$\begin{aligned}
&\sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{P(C, X^{\setminus 1,\ldots,i})}{P(X_i, C, X^{\setminus 1,\ldots,i})} \\
=&\sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{P(X^{\setminus 1,\ldots,i}|C)P(C)}{P(X_i, X^{\setminus 1,\ldots,i}|C)P(C)} \\
=&\sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{P(X^{\setminus 1,\ldots,i}|C)P(C)}{P(X_i|C)P(X^{\setminus 1,\ldots,i}|C)P(C)} \qquad (5.28) \\
=&\sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{P(C)}{P(X_i, C)} \\
=&- H(X_i|C) \qquad (5.29)
\end{aligned}$$

where Xi and $X^{\setminus 1,\ldots,i}$ is conditionally independent given C in (5.27). Similarly, The second term of (5.24) can be given by

$$\begin{aligned}
&- \sum_{X} P(X_i, X^{\setminus 1,\ldots,i}) log \frac{P(X^{\setminus 1,\ldots,i})}{P(X_i, X^{\setminus 1,\ldots,i})} \\
=&- \sum_{X} P(X_i, X^{\setminus 1,\ldots,i}) log \sum_{C} \frac{P(X^{\setminus 1,\ldots,i}, C)}{P(X_i, X^{\setminus 1,\ldots,i}, C)} \\
=&- \sum_{X} P(X_i, X^{\setminus 1,\ldots,i}) log \sum_{C} \frac{P(X^{\setminus 1,\ldots,i}|C)P(C)}{P(X_i, X^{\setminus 1,\ldots,i}|C)P(C)} \\
=&- \sum_{X} P(X_i, X^{\setminus 1,\ldots,i}) log \sum_{C} \frac{P(X^{\setminus 1,\ldots,i}|C)P(C)}{P(X_i|C)P(X^{\setminus 1,\ldots,i}|C)P(C)} \\
=&- \sum_{X} P(X_i, X^{\setminus 1,\ldots,i}) log \sum_{C} \frac{P(C)}{P(X_i, C)} \\
=&- \sum_{X} P(X_i, X^{\setminus 1,\ldots,i}) log \frac{1}{P(X_i)} \\
=&H(X_i) \qquad (5.30)
\end{aligned}$$

As a result, it is reasonable that the attributes that have a lower entropy given C contribute more in classification of NB. In TAN structure we have to consider the

topological order of node in (5.23). (5.24) is re-written with the parent of node i as follows.

$$\sum_{i}^{n-1}[H(X_i|\Pi_i, X^{\setminus 1,\ldots,i,\Pi_i}) - H(X_i|C, \Pi_i, X^{\setminus 1,\ldots,i,\Pi_i})] \\ -H(C|X_n) \tag{5.31}$$

Theorem 1. If the structure of BN is limited to TAN, CLL can be decomposed as follows.

$$\sum_{i}^{n-1} I(X_i, C|\Pi_i) - H(C|X_n) \tag{5.32}$$

Recall TAN structure is supposed to have the node order and a node can have only a single parent node, and then we show the first term of (5.30) can be re-defined as following:

$$\begin{aligned} & - H(X_i|C, \Pi_i, X^{\setminus 1,\ldots,i,\Pi_i}) \\ =& \sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{P(C, \Pi_i, X^{\setminus 1,\ldots,i,\Pi_i})}{P(X_i, C, \Pi_i, X^{\setminus 1,\ldots,i,\Pi_i})} \\ =& \sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{1}{P(X_i|C, \Pi_i, X^{\setminus 1,\ldots,i,\Pi_i})} \qquad (5.33) \\ =& \sum_{C,X} P(X_i, C, X^{\setminus 1,\ldots,i}) log \frac{1}{P(X_i|C, \Pi_i)} \qquad (5.34) \\ =& - H(X_i|C, \Pi_i) \qquad (5.35) \end{aligned}$$

where $X^{\setminus 1,\ldots,i,\Pi_i}$ is canceled from (5.32) by using Markov blanket rule. Similarly the first term of (5.30) is re-expressed as

$$H(X_i|\Pi_i, X^{\setminus 1,\ldots,i,\Pi_i}) \tag{5.36}$$

$$= \sum_X P(X_i, X^{\setminus 1,\ldots,i}) log \frac{1}{P(X_i|\Pi_i, X^{\setminus 1,\ldots,i,\Pi_i})} \tag{5.37}$$

$$= \sum_X P(X_i, X^{\setminus 1,\ldots,i}) log \frac{1}{P(X_i|\Pi_i)} \tag{5.38}$$

$$= H(X_i|\Pi_i). \tag{5.39}$$

Therefore (5.24) of TAN structure is give by

$$\sum_i^{n-1} [-H(X_i|C, \Pi_i) + H(X_i|\Pi_i)] - H(C|X_n) \tag{5.40}$$

$$= \sum_i^{n-1} I(X_i, C|\Pi_i) - H(C|X_n). \tag{5.41}$$

Therefore, the decomposed score function to construct a TAN structure that maximizes CLL is defined as

$$\arg\max_{\Pi_i} \sum_i^{n-1} I(X_i, C|\Pi_i) - H(C|X_n). \tag{5.42}$$

$I(X_i; C|\Pi_i)$ changes depending on candidate parent node X_j. In order to construct TAN (selection of edges), first we define adjacency matrix W where each element is defined as $w_{ij} = I(X_i; C|X_j)$ and represents a score of edge, $X_i \leftarrow X_j$. However, we noted that $I(X_i; C|\Pi_i)$ tends to be higher than $I(X_j; C|\Pi_j)$ if $I(X_i; C)$ is higher than $I(X_j; C)$. In this case, $w_{i\Pi_i}$ is more likely to be selected than $w_{j\Pi_j}$ even if Π_i decreases $I(X_i; C|\Pi_i)$ and Π_j increases $I(X_j; C|\Pi_j)$. So, each row of W, W_i, is subtracted by $I(X_i; C)$ (i.e. $w_{ij} = I(X_i; C|\Pi_i) - I(X_i; C)$, which is equivalent to a decomposed of (5.41) - (5.26)). Interestingly $I(X_i; C|\Pi_i) - I(X_i; C)$ is equivalent to *interaction information* that represents how much Π_i can increase $I(X_i; C)$. In other

words, if $I(X_i; C|\Pi_i) - I(X_i; C)$ is negative, Π_i is not helpful to make BNC more discriminative supposing that X_i is *good* enough to explain C. As matrix W is now symmetric ($I(X_i; C|X_j) - I(X_i; C) = I(X_j; C|X_i) - I(X_j; C)$), we modify the scoring criterion more to give a preference to an edge between w_{ij} and w_{ji}. To this end, we let X_j has an edge $X_i \leftarrow X_j$ if $I(X_i; C)$ is higher than $I(X_j; C)$. Otherwise, X_i has an edge $X_j \leftarrow X_i$. Thus, the final DCMI is defined by

$$w_{ij} = I(X_i; C|X_j) - I(X_i; C) + I(X_j; C) \tag{5.43}$$

$$= I(X_j; C|X_i) \tag{5.44}$$

Basically DCMI is based on maximizing CLL but $X_i \leftarrow X_j$ is more likely to be selected if X_i do not have enough information about C but X_j has. Once W is calculated, a directed tree is constructed by using simple order-based method (OLDT) as defined in Algorithm 5.

Algorithm 5 Order-based Learning Directed Tree

procedure OLDT(W) ▷ W is a matrix weighted by DCMI
    Initialize $A = n \times n$ *zero matrix* ▷ A is an adjacency matrix
    $i, j = max_{i \in X, j \in X, i \neq j}\ w_{ij}$
    $a_{ij} = 1$, $S = \{i, j\}$
    for $t \leftarrow 3, n-1$ **do**
        $i, j = max_{i \in X \setminus S, j \in S}\ w_{ij}$
        $a_{ij} = 1$, $S = S \cup \{i\}$
    end for
    return A
end procedure

In order to additionally improve the DCMI-based learning structure, we remove edges that may enhance the performance except the edges between attributes and class label using Classification Rate (CR) [120, 127] defined as

$$CR = \frac{1}{N}\sum_{t=1}^{N} I(BNC(x_1^t, \ldots, x_n^t), c^t) \tag{5.45}$$

where N is the number of instances and $x_1^t, \ldots, x_n^t$ is the t-th instance in training data. $BNC(x_1, \ldots, x_n - 1)$ returns an estimated $\hat{c}$ as a result of prediction. c is the correct class label. $I(\hat{c}, c)$ is an indicator function that returns 1 if $\hat{c} = c$. After building the DCMI-based structure, an edge that increases CR is iteratively removed until there is no more edges to remove for better CR. CR-based Removing Edge algorithm (CRE) is described in Algorithm 6.

Algorithm 6 CR-based Removing Edge algorithm

procedure CRE(A)
  while $CR(A)$ increases **do**
    $i, j = max_{i,j,a_{ij}\neq 0}\ CR(A_{a_{ij}=0})$
    if $CR(A) < CR(A_{a_{ij}=0})$ **then**
      $A = A_{a_{ij}=0}$
    end if
  end while
  return A
end procedure

Table 5.1: Data sets used in the experiments

	Dataset	Attribute	Class	Train	Test
1	Australian	14	2	690	CV-5
2	Breast	10	2	683	CV-5
3	Chess	36	2	2130	1066
4	Cleve	13	2	296	CV-5
5	Corral	6	2	128	CV-5
6	Crx	15	2	653	CV-5
7	Diabetes	8	2	768	CV-5
8	Flare	10	2	1066	CV-5
9	German	20	2	1000	CV-5
10	Glass	9	7	214	CV-5
11	Glass2	9	2	163	CV-5
12	Heart	13	2	270	CV-5
13	Hepatitis	19	2	80	CV-5
14	Iris	4	3	150	CV-5
15	Letter	16	26	15000	5000
16	Lymphography	18	4	148	CV-5
17	Mofn-3-7-10	10	2	300	1024
18	Pima	8	2	768	CV-5
19	Satimage	36	6	4435	2000
20	Segment	19	7	1540	770
21	Shuttle-small	9	7	3866	1934
22	Soybean-large	35	19	562	CV-5
23	Vehicle	18	4	846	CV-5
24	Vote	16	2	435	CV-5
25	Waveform-21	21	3	300	4700

5.4 Results

5.4.1 DCMI with UCI Benchmark Data

We evaluated the classification performance of DCMI comparing to other BNC criteria and state-of-the-art classifiers. We performed our evaluation on 23 UCI benchmark data sets [126] and two synthetic data sets (Corral and Monfn-3-7-10) that most of the related works have used [121, 128]. The description of the data sets is presented in Table 5.2. Since it is assumed that variables are discrete, all continuous-valued at-

tributes were discretized by using the supervised entropy-based method [129] and instances with missing values were removed from the data sets. For parameter estimation, Ordinary Frequency Estimation (OFE), which is equivalent to ML-based estimation, is used, and zero probabilities of the conditional probability tables are replaced with a small value 0.00001. Holdout test and 5-fold cross validation are performed for large and small size data sets respectively.

Table 5.2: Accuracy of Bayesian network classifieirs

	Dataset	NB	CMI	CR
1	Australian	86.60±0.24	80.72±0.90	83.09±0.97
2	Breast	97.67±0.09	96.58±0.34	97.26±0.34
3	Chess	87.31±0.82	92.46±1.21	95.75±0.80
4	Cleve	83.15±0.80	80.49±1.46	81.20±1.30
5	Corral	86.24±1.52	98.48±1.02	92.34±2.66
6	Crx	86.23±0.34	81.03±0.88	82.52±0.96
7	Diabetes	78.34±0.35	78.07±0.53	77.29±0.78
8	Flare	79.10±0.25	81.59±0.42	81.29±0.41
9	German	76.17±0.60	73.34±0.96	74.46±0.85
10	Glass	77.34±1.32	77.89±1.52	78.95±1.58
11	Glass2	88.54±1.12	87.47±1.46	86.71±2.05
12	Heart	83.24±0.60	78.78±1.38	81.67±1.17
13	Hepatitis	94.38±0.95	90.00±2.15	95.81±1.02
14	Iris	95.03±0.46	94.73±1.22	95.13±1.23
15	Letter	74.64±0.51	87.21±0.44	87.69±0.71
16	Lymphography	80.35±1.53	79.86±2.15	78.94±2.43
17	Mofn-3-7-10	89.70±1.13	92.35±1.51	93.27±1.67
18	Pima	78.16±0.54	78.18±0.48	77.33±1.02
19	Satimage	82.80±0.70	87.16±0.72	88.38±0.42
20	Segment	93.55±0.89	94.71±0.90	95.59±0.93
21	Shuttle-small	92.56±0.83	94.76±0.43	94.72±0.45
22	Soybean-large	93.47±0.51	90.11±1.11	90.94±0.60
23	Vehicle	67.48±0.80	76.57±0.74	75.05±1.37
24	vote	90.21±0.40	92.61±0.80	93.91±0.60
25	Waveform-21	80.45±0.75	72.19±0.80	74.73±0.88
	Average	84.91	85.49	86.16

Table 5.3: Accuracy of Bayesian network classifieirs

	Dataset	SVM	DCMI	DCMI+CR
1	Australian	84.72±0.45	86.40±0.43	86.77±0.47
2	Breast	97.77±0.12	96.35±0.39	96.32±0.35
3	Chess	95.54±0.81	93.86±0.64	94.46±0.69
4	Cleve	82.00±1.02	80.19±1.19	80.28±1.26
5	Corral	86.83±2.17	98.04±1.17	98.55±1.12
6	Crx	84.61±0.72	86.21±0.49	86.42±0.56
7	Diabetes	77.13±0.62	77.49±0.67	77.82±0.71
8	Flare	80.93±0.31	80.57±0.39	80.59±0.51
9	German	76.10±0.65	75.37±0.64	75.29±0.56
10	Glass	69.28±1.36	79.09±1.79	79.44±1.24
11	Glass2	81.29±1.12	88.33±1.37	88.02±1.57
12	Heart	84.50±0.62	80.70±1.44	80.61±1.41
13	Hepatitis	88.50±2.32	92.25±2.24	92.25±2.24
14	Iris	96.27±0.55	94.83±1.08	94.70±0.93
15	Letter	97.28±0.23	85.44±0.46	85.60±0.46
16	Lymphography	79.44±2.02	77.67±1.88	77.53±2.03
17	Mofn-3-7-10	100.00±0.00	93.18±1.41	93.51±1.50
18	Pima	77.75±0.41	77.68±0.80	77.92±0.71
19	Satimage	75.93±0.86	86.04±0.60	86.13±0.57
20	Segment	95.71±0.83	95.47±0.59	95.57±0.69
21	Shuttle-small	94.12±0.56	94.48±0.45	94.44±0.45
22	Soybean-large	87.13±0.95	93.47±0.51	93.51±0.51
23	Vehicle	75.19±1.00	75.31±0.74	75.39±0.72
24	vote	95.97±0.48	95.43±0.45	95.33±0.59
25	Waveform-21	62.78±6.56	76.41±1.51	76.45±1.47
	Average	85.07	86.41	86.52

The abbreviations and brief descriptions for compared methods are as follows:

- **NB**: naive Bayes Classifier
- **CMI**: W in OLDT is calculated by CMI, $I(Xi;\Pi_i|C)$ [121]
- **CR**: In for loop of OLDT, W is iteratively re-calculated by CR [116] where w_{ij} indicates how much CR can increases when w_{ij} is augmented as a new edge

- **SVM**: Support vector machine with radial basis function kernel and gamma 0.07.
- **DCMI**: OLDT where W is calculated by DCMI.
- **DCMI+CR**: After TAN+DCMI, some edges are removed by CRE algorithm.

Table 5.4: Comparison of classifiers using the one-sided paired t-test

	CMI	CR	SVM	DCMI	DCMI+CR
NB	<0.01 ⇑	<0.01 ⇑	0.3145 ↑	<0.01 ⇑	<0.01 ⇑
CMI		<0.01 ⇑	0.0497 ⇐	<0.01 ⇑	<0.01 ⇑
CR			<0.01 ⇐	0.0114 ⇑	<0.01 ⇑
SVM				<0.01 ⇑	<0.01 ⇑
DCMI					<0.01 ⇑

The test for each data set was performed 20 times and same train and test data were used in all methods. The accuracy of classification on average with standard deviation are arranged in Table 5.1. In overall results, DCMI is better than other methods on average and the result also shows that removing edges can improve DCMI. Not surprisingly NB had the worst performance. In some data set, NB results in the best accuracies implying attributes of the data may be independent to each other. Table 5.3 shows the results of one-sided paired t-test with significance level 0.05 where the number is p-value, direction of arrow indicates better method, and double arrow means statistical significance at level p <0.05. Overall results of t-test shows that DCMI and DCMI+CR significantly outperforms other methods.

Table 5.5: 55 antibodies of selected proteins

pSrc(Y527)	p53	pGSK3	pmTOR	GSK3
pIRS1(Y1179)	p38	pp38	CyclinD3	NQO1
pIRS1(Y896)	p16	pJNK	pPTEN	CDK4
pIGF1R(Y1158-1162)	Src	RAF1	pRAF1	Bcl2
pIGF1R(Y1162-1163)	p27	pp53	Hsp27	IKBa
pEGFR(Y1173)	p21	sClu	IGF1R	MDM2
E-Cadherin	AKT	pBcl2	pNFkBp65	pERK
pSrc(Y416)	JNK	pAKT	Vimentin	PTEN
CyclinB1	Rb	pMDM2	NFkBp65	IRS1
b-Catenin	ERK	Stat3	b-Actin	EGFR
gH2AX	pRb	pIKBa	pStat3	mTOR

5.4.2 DCMI with Lung Cancer Data

5.4.2.1 Personalized Medicine

In order to biologically profile lung cancers, RPPA is used to measure protein expression levels of 75 cell lines with 55 antibodies (Table 5.4). For drug sensitivity test, 23 durgs were used with different set of cell lines. On average, 43 cell lines are used for a single drug. As a preprocessing, the drug sensitivities were discretized into 2 states (High or Low) by K-means clustering algorithm and protein expression level of RPPA is discretized by minimum entropy based discretization method [129]. Instead of all 55 antibodies, proteins were pre-selected by using DCMI network. To build DCMI network, we first calculated DCMI of all possible edges between attributes and select only edges that have a high value like top 1%. Then, only attributes that are not isolated in the DCMI network were selected. To evaluate the performance, the prediction accuracy was measured by using leave-one-out estimation due to limited number of instances.

The results, given in Table 5.5, show the classification accuracies for each drug. The best and worst accuracy of DCMI is 96.08% in Paclitaxel/Carboplatin

Table 5.6: Accuracy of Bayesian network classifieirs

	Drug Name	NB	CR	SVM	DCMI
1	8-aminoadenosine	86.67	86.67	91.11	88.89
2	8-Cl-adenosine	91.11	95.56	95.56	93.33
3	Carboplatin	82.22	82.22	82.22	84.44
4	Chloroquine	88.64	90.91	90.91	95.45
5	Cisplatin	88.37	86.05	83.82	88.37
6	Cyclopamine	77.78	80.00	77.78	82.22
7	Diazonamide	78.05	70.73	82.93	75.61
8	Docetaxel	92.68	95.12	95.12	90.24
9	Doxorubicin	73.91	76.09	54.35	80.43
10	Erlotinib	97.67	97.67	90.70	90.70
11	Etoposide	88.37	88.37	88.37	88.37
12	Gefitinib	95.00	95.00	87.50	95.00
13	Gemcitabine	88.64	90.91	84.09	90.91
14	Gemcitabine/Cisplatin	80.95	80.95	85.71	90.48
15	Irinotecan	77.50	75.00	72.50	72.50
16	Paclitaxel	87.23	87.23	82.98	87.23
17	Paclitaxel/Carboplatin	96.08	96.08	90.20	96.08
18	Peloruside A	85.71	88.10	95.24	90.48
19	Pemetrexed	86.36	86.36	90.91	84.09
20	Pemetrexed/Cisplatin	83.33	83.33	76.19	83.33
21	Smac Mimetic	94.87	94.87	100.00	89.74
22	Sorafenib	95.74	89.36	89.36	89.36
23	Vinorelbine	88.37	86.05	88.37	90.70
	Average	87.19	87.07	85.90	87.74

and 72.50% in Irinotecan respectively. On average, DCMI outperformed other methods with 87.74% accuracy.

5.4.2.2 Biomarker Indentification

In order to discover the biomarkers of lung cancer, we integrate the network structures of DCMI classifiers for all drugs into a single network. It is assumed that lung cancer-associated nodes and edges are more likely to be involved in the integrated network. To this end, all instances are used to estimate parameters as

training data given a drug. Figure 5.6 displays a integrated network where all edges and nodes of networks for 23 classifiers are included except class nodes. We note four interconnected proteins, NF-kB, Stat3, beta-Catenin, and IkBa as biomarkers for lung cancer. From Figure 5.6, it is implied that four proteins are conditionally dependent with many other proteins given drugs. In other words, if the selected proteins are associated with effect of drugs, they could be more related to the biological process of lung cancer. To make the integrated network simpler, we displayed the edges that are used by more than 2 classifiers in Figure 5.5. Still there are four biomarkers we note and specially b-catenin is centered by being connected to most of other proteins.

A number of literatures support that NF-kB are related to lung cancer. Basically NF-kB is a transcription factor that is involved variety of biological process such as regulation of immune response and inflammation. Along with known functions, NF-B is constitutively active in most cancers being reported that NF-kB could have a major role in oncogenesis [130, 131]. Especially it was reported that activation of NF-kB mediates apoptosis by inducing some members of the anti-apoptotic Bcl-2 family [132].

Stat3 is signal transducer and activator of transcription that is emerged as potential therapeutic target for cancers [133] since persistent activation of Stat has been reported in a number of cancers [134]. Especially it has been demonstrated that constitutive Stat3 DNA-binding activity are in multiple non-small cell lung cancer cell lines [135]. In Figure 5.6, we also note that Stat3 is connected to CDK4 whose expression level was significantly higher in lung cancer tissues than normal tissues [136].

Many studies reported that activation of the Wnt/-catenin signaling pathway plays an important role in lung cancer [137, 138] Especially B-catenin has attracted as effective drug target for lung cancer as it has been shown that reduced B-catenin

treated non-small-cell lung cancer well [139, 140] and also recently it was reported that tumor metastatsis can be led by Wnt/-catenin signaling [141]. It has reported that silencing of IBa can activate NF-B that is a potential key mediator of many cancer types including EGFR-mutant lung cancer [142, 143].

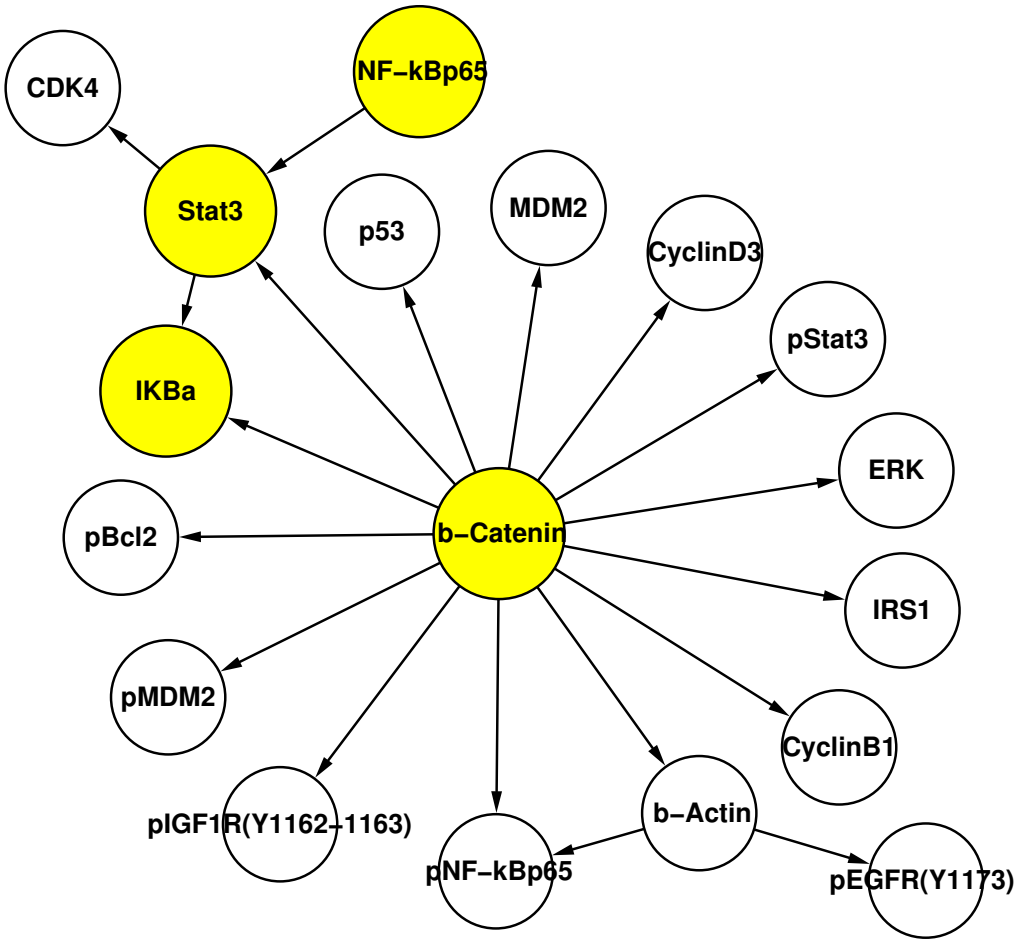


Figure 5.5: The integrated network of Bayesian network classifier for 23 drugs

5.5 Conclusion

We propose a new decomposable criterion for discriminative learning structure. The new criterion, DCMI, is based on maximizing CLL under the assumption of TAN structure. As mutual information-based criterion we proposed is decomposable, learning a discriminative tree structure can be performed in linear time. The performance of DCMI is compared with state-of-the-art methods on 25 benchmarking data sets. DCMI and DCMI with CR performed better than NB, LL-scored TAN, CR-scored TAN, and SVM with RBF kernel. The significant performance was statistically assessed by t-test. In the application to classification for personalized medicine, it was demonstrated that DCMI with pre-selected proteins effectively predicted the drug sensitivities so that optimal drugs could be recommended to lung cancer cancer patients. In addition, we explored that the network structure of classifier can provide important information of dependencies of proteins given drugs in order to discover biomarkers of lung cancer. Our future work include that extending DCMI to discriminative parameters and applying DCMI to feature selection for BNC with more antibodies and samples.

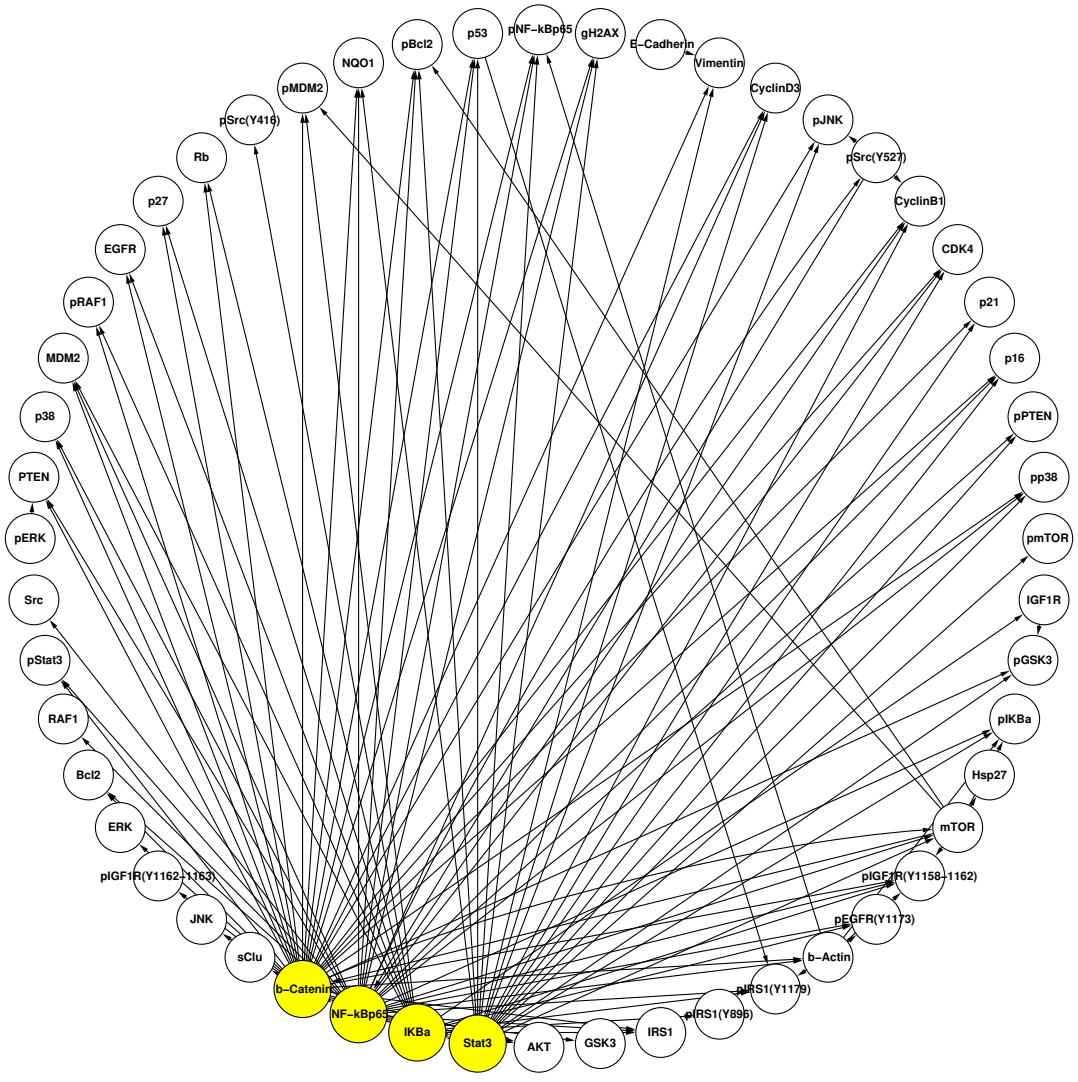


Figure 5.6: The integrated network of Bayesian network classifier for 23 drugs

REFERENCES

[1] C. Liu, L. Cheng, J. A. Badner, D. Zhang, D. W. Craig, M. Redman, and E. S. Gershon, "Whole genome association mapping of gene expression in the human prefrontal cortex," *Molecular psychiatry*, vol. 15, no. 8, pp. 779–784, 2010.

[2] P. Sklar, J. W. Smoller, J. Fan, M. A. R. Ferreira, R. H. Perlis, K. Chambert, V. L. Nimgaonkar, M. B. McQueen, S. V. Faraone, A. Kirby, P. I. W. de Bakker, M. N. Ogdie, M. E. Thase, G. S. Sachs, K. Todd-Brown, S. B. Gabriel, C. Sougnez, C. Gates, B. Blumenstiel, M. Defelice, K. G. Ardlie, J. Franklin, W. J. Muir, K. A. McGhee, D. J. MacIntyre, A. McLean, M. VanBeck, A. McQuillin, N. J. Bass, M. Robinson, J. Lawrence, A. Anjorin, D. Curtis, E. M. Scolnick, M. J. Daly, D. H. Blackwood, H. M. Gurling, and S. M. Purcell, "Whole-genome association study of bipolar disorder," *Molecular Psychiatry*, vol. 13, pp. 558–569, 2008.

[3] M. A. R. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, J. W. Smoller, D. Grozeva, J. Stone, I. Nikolov, K. Chambert, M. L. Hamshere, V. L. Nimgaonkar, V. Moskvina, M. E. Thase, S. Caesar, G. S. Sachs, J. Franklin, K. Gordon-Smith, K. G. Ardlie, S. B. Gabriel, C. Fraser, B. Blumenstiel, M. Defelice, G. Breen, M. Gill, D. W. Morris, A. Elkin, W. J. Muir, K. A. McGhee, R. Williamson, D. J. MacIntyre, A. W. MacLean, D. S. Clair, M. Robinson, M. V. Beck, A. C. P. Pereira, R. Kandaswamy, A. McQuillin, D. A. Collier, N. J. Bass, A. H. Young, J. Lawrence, I. N. Ferrier, A. Anjorin, A. Farmer, D. Curtis, E. M. Scolnick, P. McGuffin, M. J. Daly, A. P. Corvin, P. A. Hol-

mans, D. H. Blackwood, H. M. Gurling, M. J. Owen, S. M. Purcell, P. Sklar, and N. Craddock, "Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder," *Nature Genetics*, vol. 40, pp. 1056–1058, 2008.

[4] A. L. Fachin, S. S. Mello, P. Sandrin-Garcia, C. M. Junta, T. Ghilardi-Netto, E. A. Donadi, G. A. d. S. Passos, and E. T. Sakamoto-Hojo, "Gene expression profiles in radiation workers occupationally exposed to ionizing radiation:," *Journal of Radiation Research*, vol. 50, no. 1, pp. 61–71, 2009.

[5] M. Hatch, E. Ron, A. Bouville, L. Zablotska, and G. Howe, "The chernobyl disaster: Cancer following the accident at the chernobyl nuclear power plant," *Epidemiologic Reviews*, vol. 27, no. 1, pp. 56–66, 2005.

[6] H. Fakir, W. Hofmann, W. Y. Tan, and R. K. Sachs, "Triggering-response model for radiation-induced bystander effects," *Radiation Research*, vol. 171, no. 3, pp. 320–331, Feb. 2009.

[7] L. A. Liotta, V. Espina, A. I. Mehta, V. Calvert, K. Rosenblatt, D. Geho, P. J. Munson, L. Young, J. Wulfkuhle, and E. F. Petricoin III, "Protein microarrays: Meeting analytical challenges for clinical applications," *Cancer Cell*, vol. 3, no. 4, pp. 317–325, Apr. 2003.

[8] B. Wilson, L. A. Liotta, and E. P. III, "Mointoring proteins and protein networks using reverse phase protein arrays," *Disease Markers*, vol. 28, pp. 225–232, 2010.

[9] B. Spurrier, S. Ramalingam, and S. Nishizuka, "Reverse-phase protein lysate microarrays for cell signaling analysis," *Nat. Protocols*, vol. 3, no. 11, pp. 1796–1808, Oct. 2008.

[10] C. Mueller, L. A. Liotta, and V. Espina, "Reverse phase protein microarrays advance to use in clinical trials." *Molecular oncology*, vol. 4, no. 6, pp. 461–481, 2010.

[11] X. Wang, Y. Dong, A. Jiwani, Y. Zou, J. Pastor, M. Kuro-O, A. Habib, M. Ruan, D. Boothman, and C. Yang, "Improved protein arrays for quantitative systems analysis of the dynamics of signaling pathway interactions." *Proteome Science*, vol. 9, pp. 53–, 2011.

[12] L. Mullenders, M. Atkinson, H. Paretzke, L. Sabatier, and S. Bouffler, "Assessing cancer risks of low-dose radiation," *Nat Rev Cancer*, vol. 9, no. 8, pp. 596–604, Aug. 2009.

[13] C. J. Bakkenist and M. B. Kastan, "Dna damage activates atm through intermolecular autophosphorylation and dimer dissociation," *Nature*, vol. 421, no. 6922, pp. 499–506, Jan. 2003.

[14] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 717–729, 8 2010.

[15] T. Jaakkola, D. Sontag, A. Globerson, and M. Meila, "Learning bayesian network structure using lp relaxations," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 358–365.

[16] C. P. De Campos, Z. Zeng, and Q. Ji, "Structure learning of bayesian networks using constraints," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 113–120.

[17] D.-C. Kim, X. Wang, C.-R. Yang, and J. Gao, "Learning biological network using mutual information and conditional independence," *BMC Bioinformatics*, vol. 11, no. Suppl 3, p. S9, 2010.

[18] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via bayesian networks," *Journal of Bioinformatics & Computational Biology*, vol. 2, no. 1, pp. 77–98, Mar. 2004.

[19] D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. de Atauri, J. D. Aitchison, L. Hood, A. F. Siegel, and H. Bolouri, "A data integration methodology for systems biology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 48, pp. 17 296–17 301, Nov. 2005.

[20] L. de Campos, "A scoring function for learning bayesian networks based on mutual information and conditional independence tests," *J. Mach. Learn. Res.*, vol. 7, pp. 2149–2187, 2006.

[21] G. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[22] D. Sontag, A. Globerson, and T. S. Jaakkola, "Clusters and coarse partitions in lp relaxations," in *Advances in Neural Information Processing Systems*, 2009, pp. 1537–1544.

[23] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, *et al.*, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D561–D568, 2011.

[24] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "Kegg for representation and analysis of molecular networks involving diseases and drugs," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D355–D360, 2010.

[25] J. Dezert, F. Smarandache, and M. Daniel, "The generalized pignistic transformation," *arXiv preprint cs/0409007*, 2004.

[26] Y. B. Kim, C.-R. Yang, and J. Gao, "Functional proteomic pattern identification under low dose ionizing radiation," *Artificial Intelligence in Medicine*, vol. 49, no. 3, pp. 177–185, July 2010.

[27] V. Stambolic, D. MacPherson, D. Sas, Y. Lin, B. Snow, Y. Jang, S. Benchimol, and T. Mak, "Regulation of pten transcription by p53," *Molecular Cell*, vol. 8, no. 2, pp. 317 – 325, 2001.

[28] J.-S. Kim, X. Xu, H. Li, D. Solomon, W. S. Lane, T. Jin, and T. Waldman, "Mechanistic analysis of a dna damage-induced, pten-dependent size checkpoint in human cells," *Molecular and Cellular Biology*, vol. 31, no. 13, pp. 2756–2771, 2011.

[29] P. Watcharasit, G. N. Bijur, J. W. Zmijewski, L. Song, A. Zmijewska, X. Chen, G. V. W. Johnson, and R. S. Jope, "Direct, activating interaction between glycogen synthase kinase 3 beta and p53 after dna damage," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7951–7955, 2002.

[30] P. Ngok-Ngam, P. Watcharasit, A. Thiantanawat, and J. Satayavivad, "Pharmacological inhibition of gsk3 attenuates dna damage-induced apoptosis via reduction of p53 mitochondrial translocation and bax oligomerization in neuroblastoma sh-sy5y cells," *Cellular & Molecular Biology Letters*, vol. 18, no. 1, pp. 58–74, 2013.

[31] H.-F. Ding, Y.-L. Lin, G. McGill, P. Juo, H. Zhu, J. Blenis, J. Yuan, and D. E. Fisher, "Essential role for caspase-8 in transcription-independent apoptosis triggered by p53," *Journal of Biological Chemistry*, vol. 275, no. 49, pp. 38 905–38 911, 2000.

[32] J. D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis," *Nat Rev Genet*, vol. 7, pp. 200–210, 2006.

[33] P. B. Madhamshettiwar, S. R. Maetschke, M. J. Davis, A. Reverter, and M. A. Ragan, "Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets," *Genome medicine*, vol. 4, no. 5, pp. 1–16, 2012.

[34] X. Wang and O. Gotoh, "Inference of cancer-specific gene regulatory networks using soft computing rules," *Gene Regulation and Systems Biology*, vol. 4, pp. 19–34, 03 2010.

[35] K. Chen and N. Rajewsky, "The evolution of gene regulation by transcription factors and microrna," *Nature Reviews Genetics*, vol. 8, pp. 93–103, 2007.

[36] L. Bonetta, "Protein-protein interactions: Interactome under construction," *Nature*, vol. 468, pp. 851–854, 2010.

[37] A. Tresch, T. Beissbarth, H. Sultmann, R. Kuner, A. Poustka, and A. Buness, "Discrimination of direct and indirect interactions in a network of regulatory effects," *Journal of Computational Biology*, vol. 14, pp. 1217–1228, 2007.

[38] C. Cheng, K.-K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, *et al.*, "Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data," *PLoS computational biology*, vol. 7, no. 11, p. e1002190, 2011.

[39] M. Hecker, S. Lambeck, S. Toepfer, E. v. Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models-a review." *Biosystems*, vol. 96, pp. 86–103, 2009.

[40] S. Zhang, Q. Li, J. Liu, and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules," *Bioinformatics*, vol. 27, no. 13, pp. i401–i409, 2011.

[41] J. Liang and J. Han, "Stochastic boolean networks: An efficient approach to modeling gene regulatory networks," *BMC Systems Biology*, vol. 6, no. 1, p. 113, 2012.

[42] N. Xuan, M. Chetty, R. Coppel, and P. Wangikar, "Gene regulatory network modeling via global optimization of high-order dynamic bayesian network," *BMC Bioinformatics*, vol. 13, no. 1, p. 131, 2012.

[43] R. de Matos Simoes and F. Emmert-Streib, "Bagging statistical network inference from large-scale gene expression data," *PLoS ONE*, vol. 7, no. 3, p. e33624, 03 2012.

[44] G. Geeven, R. E. van Kesteren, A. B. Smit, and M. C. M. de Gunst, "Identification of context-specific gene regulatory networks with gemulagene expression modeling using lasso," *Bioinformatics*, vol. 28, no. 2, pp. 214–221, 2012.

[45] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, pp. 717–729, 2010.

[46] D. Marbach, J. C. Costello, R. Kuffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, and G. Stolovitzky, "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, pp. 796–804, 2012.

[47] N. Meinshausen and P. Buhlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[48] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[49] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, "Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks," *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 12 182–12 186, 2000.

[50] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol*, vol. 5, no. 1, p. e8, 01 2007.

[51] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." *BMC bioinformatics*, vol. 7 Suppl 1, 2006.

[52] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "Tigress: Trustful inference of gene regulation using stability selection." *BMC systems biology*, vol. 6, p. 145, 2012.

[53] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky, "Towards a rigorous assessment of systems biology models: The dream3 challenges," *PLoS ONE*, vol. 5, no. 2, p. e9202, 02 2010.

[54] A. Pinna, N. Soranzo, and A. de la Fuente, "From knockouts to networks: Establishing direct cause-effect relationships through graph analysis," *PLoS ONE*, vol. 5, no. 10, p. e12912, 10 2010.

[55] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, no. 9, p. e12776, 09 2010.

[56] A. Gretton, R. Herbrich, and A. Hyvrinen, "Kernel methods for measuring independence," *Journal of Machine Learning Research*, vol. 6, pp. 2075–2129, 2005.

[57] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[58] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using z score transformation," *The Journal of molecular diagnostics*, vol. 5, no. 2, pp. 73–81, 2003.

[59] E. Sacchetti, C. Scassellati, A. Minelli, P. Valsecchi, C. Bonvicini, P. Pasqualetti, A. Galluzzo, R. Pioli, and M. Gennarelli, "Schizophrenia susceptibility and nmda-receptor mediated signalling: an association study involving 32 tagsnps of dao, daoa, ppp3cc, and dtnbp1 genes," *BMC Medical Genetics*, vol. 14, no. 1, p. 33, 2013.

[60] D. Martins-de Souza, W. Gattaz, A. Schmitt, J. Novello, S. Marangoni, C. Turck, and E. Dias-Neto, "Proteome analysis of schizophrenia patients wernicke's area reveals an energy metabolism dysregulation," *BMC Psychiatry*, vol. 9, no. 1, p. 17, 2009.

[61] S. Grube, M. F. Gerchen, B. Adamcio, L. A. Pardo, S. Martin, D. Malzahn, S. Papiol, M. Begemann, K. Ribbe, H. Friedrichs, K. A. Radyushkin, M. Mller, F. Benseler, J. Riggert, P. Falkai, H. Bickebller, K.-A. Nave, N. Brose, W. Sthmer, and H. Ehrenreich, "A cag repeat polymorphism of kcnn3 predicts sk3 channel function and cognitive performance in schizophrenia," *EMBO Molecular Medicine*, vol. 3, no. 6, pp. 309–319, 2011.

[62] A. Alkelai, L. Greenbaum, S. Lupoli, Y. Kohn, K. Sarner-Kanyas, E. Ben-Asher, D. Lancet, F. Macciardi, and B. Lerer, "Association of the type 2 diabetes mellitus susceptibility gene, tcf7l2, with schizophrenia in an arab-israeli family sample," *PLoS ONE*, vol. 7, no. 1, p. e29228, 2012.

[63] B. Glaser, G. Kirov, N. J. Bray, E. Green, M. C. O'Donovan, N. Craddock, and M. J. A. Owen, "Identification of a potential bipolar risk haplotype in the gene encoding the winged-helix transcription factor rfx4," *Mol Psychiatry*, vol. 10, pp. 920–927, 2005.

[64] L. Wu, Y. Huang, J. Li, H. Zhao, H. Du, Q. Jin, X. Zhao, H. Ma, and G. Zhu, "Association study of the fyn gene with schizophrenia in the chinese-han population," *Psychiatric genetics*, vol. 23, pp. 39–40, 2013.

[65] A. K. Kahler, S. Djurovic, L. M. Rimol, A. A. Brown, L. Athanasiu, E. G. Jonsson, T. Hansen, O. Gustafsson, H. Hall, I. Giegling, P. Muglia, S. Cichon, M. Rietschel, O. P. Pietilainen, L. Peltonen, E. Bramon, D. Collier, D. S. Clair, E. Sigurdsson, H. Petursson, D. Rujescu, I. Melle, T. Werge, V. M. Steen, A. M. Dale, R. T. Matthews, I. Agartz, and O. A. Andreassen, "Candidate gene analysis of the human natural killer-1 carbohydrate pathway and perineuronal nets in schizophrenia: B3gat2 is associated with disease risk and cortical surface area," *Biological Psychiatry*, vol. 69, no. 1, pp. 90 – 96, 2011.

[66] T. Moons, S. Claes, G. J. Martens, J. Peuskens, K. M. V. Loo, J. E. V. Schijndel, M. D. Hert, and R. van Winkel, "Clock genes and body composition in patients with schizophrenia under treatment with antipsychotic drugs," *Schizophrenia Research*, vol. 125, pp. 187 – 193, 2011.

[67] P. Hua, W. Liu, S.-H. Kuo, Y. Zhao, L. Chen, N. Zhang, C. Wang, S. Guo, L. Wang, H. Xiao, *et al.*, "Association of tef polymorphism with depression in parkinson disease," *Movement Disorders*, vol. 27, no. 13, pp. 1694–1697, 2012.

[68] M. Manchia, A. Squassina, D. Congiu, C. Chillotti, R. Ardau, G. Severino, and M. Del Zompo, "Interacting genes in lithium prophylaxis: preliminary results of an exploratory analysis on the role of dgkh and nr1d1 gene polymorphisms in

199 sardinian bipolar patients," *Neuroscience letters*, vol. 467, no. 2, pp. 67–71, 2009.

[69] S. Jamain, H. Quach, M. Fellous, and T. Bourgeron, "Identification of the human kif13a gene homologous to drosophila kinesin-73 and candidate for schizophrenia," *Genomics*, vol. 74, no. 1, pp. 36–44, 2001.

[70] M. Uddin, S.-C. Chang, C. Zhang, K. Ressler, K. B. Mercer, S. Galea, K. M. Keyes, K. A. McLaughlin, D. E. Wildman, A. E. Aiello, *et al.*, "Adcyap1r1 genotype, posttraumatic stress disorder, and depression among women exposed to childhood maltreatment," *Depression and anxiety*, 2012.

[71] J. Li, J. Liu, G. Feng, T. Li, Q. Zhao, Y. Li, Z. Hu, L. Zheng, Z. Zeng, L. He, *et al.*, "The¡ i¿ mdga¡/i¿ 1 gene confers risk to schizophrenia and bipolar disorder," *Schizophrenia research*, vol. 125, no. 2, pp. 194–200, 2011.

[72] T. Saito, D. F. Papolos, D. Chernak, M. H. Rapaport, J. R. Kelsoe, and H. M. Lachman, "Analysis of gnaz gene polymorphism in bipolar affective disorder," *American journal of medical genetics*, vol. 88, no. 4, pp. 324–328, 1999.

[73] R. Schennach, P. Zill, M. Obermeier, D. Hauer, S. Dehning, A. Cerovecki, M. Opgen-Rhein, R. Musil, I. Spellmann, J. Matz, *et al.*, "The cnr1 gene in depression and schizophrenia-is there an association with early improvement and response?" *Psychiatry research*, vol. 196, no. 1, p. 160, 2012.

[74] B. Håvik, F. A. Degenhardt, S. Johansson, C. P. Fernandes, A. Hinney, A. Scherag, H. Lybæk, S. Djurovic, A. Christoforou, K. M. Ersland, *et al.*, "Dclk1 variants are associated across schizophrenia and attention deficit/hyperactivity disorder," *PloS one*, vol. 7, no. 4, p. e35424, 2012.

[75] M. Ikeda, Y. Tomita, A. Mouri, M. Koga, T. Okochi, R. Yoshimura, Y. Yamanouchi, Y. Kinoshita, R. Hashimoto, H. J. Williams, *et al.*, "Identification of novel candidate genes for treatment response to risperidone and suscepti-

bility for schizophrenia: integrated analysis among pharmacogenomics, mouse expression, and genetic case-control association approaches," *Biological psychiatry*, vol. 67, no. 3, pp. 263–269, 2010.

[76] F. Torri, A. Akelai, S. Lupoli, M. Sironi, D. Amann-Zalcenstein, M. Fumagalli, C. Dal Fiume, E. Ben-Asher, K. Kanyas, R. Cagliani, *et al.*, "Fine mapping of ahi1 as a schizophrenia susceptibility gene: from association to evolutionary evidence," *The FASEB journal*, vol. 24, no. 8, pp. 3066–3082, 2010.

[77] J. Sun, P. Jia, A. H. Fanous, E. van den Oord, X. Chen, B. P. Riley, R. L. Amdur, K. S. Kendler, and Z. Zhao, "Schizophrenia gene networks and pathways and their applications for novel candidate gene selection," *PLoS One*, vol. 5, no. 6, p. e11351, 2010.

[78] G. Costain, A. C. Lionel, D. Merico, P. Forsythe, K. Russell, C. Lowther, T. Yuen, J. Husted, D. J. Stavropoulos, M. Speevak, *et al.*, "Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays," *Human molecular genetics*, p. ddt297, 2013.

[79] S. Balan, K. Yamada, E. Hattori, Y. Iwayama, T. Toyota, T. Ohnishi, M. Maekawa, M. Toyoshima, Y. Iwata, K. Suzuki, *et al.*, "Population-specific haplotype association of the postsynaptic density gene dlg4 with schizophrenia, in family-based association studies," *PloS one*, vol. 8, no. 7, p. e70302, 2013.

[80] L. de la Fontaine, M. J. Schwarz, M. Riedel, S. Dehning, A. Douhet, I. Spellmann, N. Kleindienst, P. Zill, H. Plischke, R. Gruber, *et al.*, "Investigating disease susceptibility and the negative correlation of schizophrenia and rheumatoid arthritis focusing on mif and cd14 gene polymorphisms," *Psychiatry research*, vol. 144, no. 1, pp. 39–47, 2006.

[81] X. Deng, N. Sagata, N. Takeuchi, M. Tanaka, H. Ninomiya, N. Iwata, N. Ozaki, H. Shibata, and Y. Fukumaki, "Association study of polymorphisms in the

neutral amino acid transporter genes slc1a4, slc1a5 and the glycine transporter genes slc6a5, slc6a9 with schizophrenia," *BMC psychiatry*, vol. 8, no. 1, p. 58, 2008.

[82] R. Bharadwaj, Y. Jiang, W. Mao, M. Jakovcevski, A. Dincer, W. Krueger, K. Garbett, C. Whittle, J. S. Tushir, J. Liu, *et al.*, "Conserved chromosome 2q31 conformations are associated with transcriptional regulation of gad1 gaba synthesis enzyme and altered in prefrontal cortex of subjects with schizophrenia," *The Journal of Neuroscience*, vol. 33, no. 29, pp. 11 839–11 851, 2013.

[83] S. Arai, H. Shibata, M. Sakai, H. Ninomiya, N. Iwata, N. Ozaki, and Y. Fukumaki, "Association analysis of the glutamic acid decarboxylase 2 and the glutamine synthetase genes (gad2, glul) with schizophrenia," *Psychiatric genetics*, vol. 19, no. 1, pp. 6–13, 2009.

[84] S.-J. Tsai, C.-J. Hong, Y.-J. Liou, and D.-L. Liao, "Association study of got2 genetic polymorphisms and schizophrenia." *Psychiatric genetics*, vol. 17, p. 314, 2007.

[85] R. Herrmann, B. Lee, and V. Y. Arshavsky, "Rgs9 knockout causes a short delay in light responses of on-bipolar cells," *PloS one*, vol. 6, no. 11, p. e27573, 2011.

[86] B. Lang, T. M. A. Alrahbeni, D. St Clair, D. H. Blackwood, *et al.*, "Hdac9 is implicated in schizophrenia and expressed specifically in post-mitotic neurons but not in adult neural stem cells," *American journal of stem cells*, vol. 1, no. 1, p. 31, 2012.

[87] P. Soronen, H. Ollila, M. Antila, K. Silander, O. Palo, T. Kieseppä, J. Lönnqvist, L. Peltonen, A. Tuulio-Henriksson, T. Partonen, *et al.*, "Replication of gwas of bipolar disorder: association of snps near cdh7 with bipolar

disorder and visual processing," *Molecular psychiatry*, vol. 15, no. 1, pp. 4–6, 2010.

[88] Y.-H. Chang, S.-Y. Lee, S.-L. Chen, N.-S. Tzeng, T.-Y. Wang, I. Hui Lee, P. See Chen, S.-Y. Huang, Y. Kuang Yang, H.-C. Ko, *et al.*, "Genetic variants of the bdnf and drd3 genes in bipolar disorder comorbid with anxiety disorder," *Journal of affective disorders*, 2013.

[89] K. Yamada, Y. Iwayama, E. Hattori, K. Iwamoto, T. Toyota, T. Ohnishi, H. Ohba, M. Maekawa, T. Kato, and T. Yoshikawa, "Genome-wide association study of schizophrenia in japanese population," *PloS one*, vol. 6, no. 6, p. e20468, 2011.

[90] A. C. Nica and E. T. Dermitzakis, "Expression quantitative trait loci: present and future," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620, 2013.

[91] B. Vasic, V. Ravanmehr, and A. Krishnan, "An information theoretic approach to constructing robust boolean gene regulatory networks," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 9, no. 1, pp. 52 –65, jan.-feb. 2012.

[92] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring connectivity of genetic regulatory networks using information-theoretic criteria," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 5, no. 2, pp. 262–274, Apr. 2008.

[93] D.-C. Kim, X. Wang, C.-R. Yang, and J. Gao, "Learning biological network using mutual information and conditional independence," *BMC Bioinformatics*, vol. 11, no. Suppl 3, p. S9, 2010.

[94] 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*, vol. 2, pp. 254–261, 2005.

[95] Y.-A. Kim, S. Wuchty, and T. M. Przytycka, "Identifying causal genes and dysregulated pathways in complex diseases," *PLoS Comput Biol*, vol. 7, no. 3, p. e1001095, 03 2011.

[96] J. J. B. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen, "Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci," *Proceedings of the National Academy of Sciences*, vol. 104, no. 5, pp. 1708–1713, 2007.

[97] S. Kim and E. P. Xing, "Statistical estimation of correlated genome associations to a quantitative trait network," *PLoS Genet*, vol. 5, no. 8, p. e1000587, 08 2009.

[98] L. Chen, L. Zhang, Y. Zhao, L. Xu, Y. Shang, Q. Wang, W. Li, H. Wang, and X. Li, "Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing snps and pathways," *Bioinformatics*, vol. 25, no. 2, pp. 237–242, 2009.

[99] M. Xiong, J. Li, and X. Fang, "Identification of genetic networks," *Genetics*, vol. 166, no. 2, pp. 1037–1052, 2004.

[100] B. Liu, A. de la Fuente, and I. Hoeschele, "Gene network inference via structural equation modeling in genetical genomics experiments," *Genetics*, vol. 178, no. 3, pp. 1763–1776, 2008.

[101] B. A. Logsdon and J. Mezey, "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001014, 2010.

[102] X. Cai, J. A. Bazerque, and G. B. Giannakis, "Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations," *PLoS Comput Biol*, vol. 9, no. 5, p. e1003068, 2013.

[103] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[104] T. Zou, Hui AU Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301–320, 2005.

[105] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[106] R. Conti, D. L. Veenstra, K. Armstrong, L. J. Lesko, and S. D. Grosse, "Personalized medicine and genomics: challenges and opportunities in assessing effectiveness, cost-effectiveness, and future research priorities," *Medical Decision Making*, vol. 30, no. 3, pp. 328–340, 2010.

[107] S. M. Kornblau, R. Tibes, Y. H. Qiu, W. Chen, H. M. Kantarjian, M. Andreeff, K. R. Coombes, and G. B. Mills, "Functional proteomic profiling of aml predicts response and survival," *Blood*, vol. 113, no. 1, pp. 154–164, 2009.

[108] L. A. Liotta, V. Espina, A. I. Mehta, V. Calvert, K. Rosenblatt, D. Geho, P. J. Munson, L. Young, J. Wulfkuhle, and E. F. Petricoin III, "Protein microarrays: meeting analytical challenges for clinical applications," *Cancer cell*, vol. 3, no. 4, pp. 317–325, 2003.

[109] X. Wang, Y. Dong, A. J. Jiwani, Y. Zou, J. Pastor, M. Kuro-o, A. A. Habib, M. Ruan, D. A. Boothman, and C.-R. Yang, "Improved protein arrays for quantitative systems analysis of the dynamics of signaling pathway interactions," *Proteome science*, vol. 9, no. 1, pp. 1–13, 2011.

[110] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* John Wiley & Sons Inc, June 1973.

[111] V. N. Vapnik, *The nature of statistical learning theory.* New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[112] B. V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* Mc Graw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, 1991.

[113] J. Huang, J. Lu, and C. X. Ling, "Comparing naive bayes, decision trees, and svm with auc and accuracy," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on.* IEEE, 2003, pp. 553–556.

[114] J. Rodriguez, A. Perez, D. Arteta, D. Tejedor, and J. Lozano, "Using multidimensional bayesian network classifiers to assist the treatment of multiple sclerosis," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 1705 –1715, nov. 2012.

[115] V. R. R. Mohanty and M. R. Patra, "Classification of web services using bayesian network," *Journal of Software Engineering and Applications*, vol. 5, no. 4, pp. 291–296, 2012.

[116] F. Pernkopf, M. Wohlmayr, and S. Tschiatschek, "Maximum margin bayesian network classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 3, pp. 521 –532, march 2012.

[117] F. Pernkopf and M. Wohlmayr, "On discriminative parameter learning of bayesian network classifiers," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ser. ECML PKDD '09, 2009, pp. 221–237.

[118] T. Roos, H. Wettig, P. Grnwald, P. Myllymki, and H. Tirri, "On discriminative bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, no. 3, pp. 267–296, 2005.

[119] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative parameter learning for bayesian networks," in *Proceedings of the 25th international conference on Machine learning*, ser. ICML '08, New York, NY, USA, 2008, pp. 1016–1023.

[120] F. Pernkopf and J. Bilmes, "Discriminative versus generative parameter and structure learning of bayesian network classifiers," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05, New York, NY, USA, 2005, pp. 657–664.

[121] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Science*, vol. 163, pp. 131–163, 1997.

[122] R. E. Neapolitan, *Learning Bayesian Networks*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2003.

[123] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[124] D. Grossman and P. Domingos, "Learning bayesian network classifiers by maximizing conditional likelihood," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 46.

[125] A. M. Carvalho, T. Roos, A. L. Oliveira, and P. Myllymäki, "Discriminative learning of bayesian networks via factorized conditional log-likelihood," *The Journal of Machine Learning Research*, vol. 12, pp. 2181–2210, 2011.

[126] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

[127] E. Keogh and M. Pazzani, "Learning augmented bayesian classifiers: A comparison of distribution-based and classification-based approaches," in *Proceedings*

of the seventh international workshop on artificial intelligence and statistics. Citeseer, 1999, pp. 225–230.

[128] F. Pernkopf and J. A. Bilmes, "Efficient heuristics for discriminative structure learning of bayesian network classifiers," *The Journal of Machine Learning Research*, vol. 11, pp. 2323–2360, 2010.

[129] Fayyad and Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," 1993, pp. 1022–1027.

[130] X. Dolcet, D. Llobet, J. Pallares, and X. Matias-Guiu, "Nf-kb in development and progression of human cancer," *Virchows archiv*, vol. 446, no. 5, pp. 475–482, 2005.

[131] M. Chaturvedi, B. Sung, V. Yadav, R. Kannappan, and B. Aggarwal, "Nf-κb addiction and its role in cancer:one size does not fit all," *Oncogene*, vol. 30, no. 14, pp. 1615–1630, 2011.

[132] J.-U. Lee, R. Hosotani, M. Wada, R. Doi, T. Kosiba, K. Fujimoto, Y. Miyamoto, S. Tsuji, S. Nakajima, Y. Nishimura, *et al.*, "Role of bcl-2 family proteins (bax, bcl-2 and bcl-x) on cellular susceptibility to radiation in pancreatic cancer cells," *European Journal of Cancer*, vol. 35, no. 9, pp. 1374–1380, 1999.

[133] H. Yu and R. Jove, "The stats of cancernew molecular targets come of age," *Nature Reviews Cancer*, vol. 4, no. 2, pp. 97–105, 2004.

[134] R. Buettner, L. B. Mora, and R. Jove, "Activated stat signaling in human tumors provides novel molecular targets for therapeutic intervention," *Clinical cancer research*, vol. 8, no. 4, pp. 945–954, 2002.

[135] L. Song, J. Turkson, J. G. Karras, R. Jove, and E. B. Haura, "Activation of stat3 by receptor tyrosine kinases and cytokines regulates survival in human non-small cell carcinoma cells," *Oncogene*, vol. 22, no. 27, pp. 4150–4165, 2003.

[136] A. Wu, B. Wu, J. Guo, W. Luo, D. Wu, H. Yang, Y. Zhen, X. Yu, H. Wang, Y. Zhou, *et al.*, "Elevated expression of cdk4 in lung cancer," *J Transl Med*, vol. 9, no. 38, p. 38, 2011.

[137] N. Sunaga, T. Kohno, F. T. Kolligs, E. R. Fearon, R. Saito, and J. Yokota, "Constitutive activation of the Wnt signaling pathway by CTNNB1 mutations in a subset of human lung adenocarcinoma," *Genes Chromosomes & Cancer*, vol. 30, pp. 316–321, 2001.

[138] Y. Teng, X. Wang, Y. Wang, and D. Ma, "Wnt/β-catenin signaling regulates cancer stem cells in lung cancer A549 cells," *Biochemical and Biophysical Research Communications*, vol. 392, pp. 373–379, 2010.

[139] J. M. Retera, M. P. Leers, M. A. Sulzer, and P. H. Theunissen, "The expression of beta-catenin in non-small-cell lung cancer: a clinicopathological study," *Journal of Clinical Pathology*, vol. 51, pp. 891–894, 1998.

[140] H. H. Luu, R. Zhang, R. C. Haydon, E. Rayburn, Q. Kang, W. Si, J. K. Park, H. Wang, Y. Peng, W. Jiang, and T.-C. He, "Wnt /β-Catenin Signaling Pathway as Novel Cancer Drug Targets," *Current Cancer Drug Targets*, vol. 4, pp. 653–671, 2004.

[141] D. X. Nguyen, A. C. Chiang, X. H.-F. Zhang, J. Y. Kim, M. G. Kris, M. Ladanyi, W. L. Gerald, and J. Massagu, "WNT/TCF Signaling through LEF1 and HOXB9 Mediates Lung Adenocarcinoma Metastasis," *Cell*, vol. 138, pp. 51–62, 2009.

[142] W. Chen, Z. Li, L. Bai, and Y. Lin, "Nf-kappab in lung cancer, a carcinogenesis mediator and a prevention and therapy target." *Frontiers in bioscience (Landmark edition)*, vol. 16, pp. 1172–1185, 2010.

[143] T. G. Bivona, H. Hieronymus, J. Parker, K. Chang, M. Taron, R. Rosell, P. Moonsamy, K. Dahlman, V. A. Miller, C. Costa, *et al.*, "Fas and nf-[kgr]

b signalling modulate dependence of lung cancers on mutant egfr," *Nature*, vol. 471, no. 7339, pp. 523–526, 2011.

BIOGRAPHICAL STATEMENT

Dongchul Kim received his B.S. in the computer science from Kyungsung University, Korea in 2003 and in international trade from Hankuk University of foreign studies, Korea in 2001. He received his M.S. and Ph.D. degrees from The University of Texas at Arlington in 2007 and 2014, respectively, all in Computer Science. His current research interest is in the area of machine learning and pattern recognition with applications to bioinformatics and biomedical problems.