DETECTING CANCER-RELATED GENES AND GENE-GENE INTERACTIONS BY MACHINE LEARNING METHODS

By

Bing Han

Submitted to the graduate degree program in Electrical Engineering and Computer Science and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson Xue-wen Chen, PhD

_____

Arvin Agah, PhD

_____

Jerzy Grzymala-Busse, PhD

_____

Luke Huan, PhD

_____

Tyrone Duncan, PhD

_____

Zohreh Talebizadeh, PhD

Date Defended: _____

The Dissertation Committee for Bing Han certifies that this is the approved version of the following dissertation:

DETECTING CANCER-RELATED GENES AND GENE-GENE INTERACTIONS BY MACHINE LEARNING METHODS

_____

Chairperson Xue-wen Chen

_____

_____

_____

_____

_____

Date approved: _____

Acknowledgements

I am heartily thankful to my advisor, Dr. Xue-wen Chen, whose professional guidance, leadership, and vision from the preliminary to the concluding level enabled me to develop an understanding of my research subject. I am also grateful to Dr. Talebizadeh for her expert advices in biology and proofreading my dissertation and Dr. Agah, Dr. Grzymala-Busse, Dr. Duncan, and Dr. Huan for serving on my committee.

My parents, Dehong Han and Yiqin Wang, have provided much moral and material support during the long years of my education. My mother died from esophageal cancer in 2008, before she would witness my graduation. She has been an invisible presence during the composition of these pages. I would like to mention also my parents-in-law, Mengkui Liu and Shuzhen Hu, who unconditionally helped caring for my newborn daughter during the past two years.

I most want to thank my wife, Xia Liu, for her love, sacrifice, and kind indulgence. I also credit my daughter, Jocelyn, for inspiring and amazing me every day. None of this would have been possible without the love and support of my family.

Abstract

To understand the underlying molecular mechanisms of cancer and therefore to improve pathogenesis, prevention, diagnosis and treatment of cancer, it is necessary to explore the activities of cancer-related genes and the interactions among these genes. In this dissertation, I use machine learning and computational methods to identify differential gene relations and detect gene-gene interactions. To identify gene pairs that have different relationships in normal versus cancer tissues, I develop an integrative method based on the bootstrapping K-S test to evaluate a large number of microarray datasets. The experimental results demonstrate that my method can find meaningful alterations in gene relations. For gene-gene interaction detection, I propose to use two Bayesian Network based methods: DASSO-MB (Detection of ASSOciations using Markov Blanket) and EpiBN (Epistatic interaction detection using Bayesian Network model) to address the two critical challenges: searching and scoring. DASSO-MB is based on the concept of Markov Blanket in Bayesian Networks. In EpiBN, I develop a new scoring function, which can reflect higher-order gene-gene interactions and detect the true number of disease markers, and apply a fast Branch-and-Bound (B&B) algorithm to learn the structure of Bayesian Network. Both DASSO-MB and EpiBN outperform some other commonly-used methods and are scalable to genome-wide data.

Keywords: Cancer, Bioinformatics, System Biology, Machine Learning, Differential Gene Relations, Gene-Gene Interactions

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Significance of Detecting Cancer-related Genes and Gene-Gene Interactions

Cancer is one type of fatal disease which causes about 13% of all deaths. It is generally estimated that roughly 7.2 to 7.5 million people worldwide die from cancer each year. In the development of cancer, abnormal cells divide without control and invade nearby parts of the body. Comparing to abnormal cells, healthy cells can control their own growth and death [1].

Almost all types of cancers are caused by dynamic changes in the genome. These genetic aberrations will typically affect two types of genes: oncogenes and tumor suppressor genes [2]. In the development of cancer, the dominant gain of function for oncogenes and the recessive loss of function for tumor suppressor genes deeply change the molecular mechanism regulating growth and death of cells, thus driving the progressive transformation of normal cells into tumor cells.

In the past several decades, most cancer researchers have been trying to detect cancer-related genes including both oncogenes and tumor suppressor genes. These cancer-related genes can be used as diagnostic and prognostic signatures or as potential targets for future therapy. However, cancer is a system biology disease that involves a number of fundamental cell processes such as death, proliferation, differentiation, and migration. Genes play an important role in some key signaling pathways that control these cell processes [1]. Therefore exploring the activities of

cancer-related genes and the interactions among these genes in the complex cell processes can contribute to the understanding of the underlying molecular mechanisms of cancer.

Microarray technology is a powerful tool to identify cancer-related genes and explore their activities in a biological system. Microarrays can monitor the abundance of messenger RNA (mRNA) of tens of thousands of genes simultaneously. In cancer research, microarray experiments are used to identify a list of genes which show differential activities between normal cells and tumor cells. However, a simple list of individual differentially expressed genes can only tell us which genes are altered by biological differences between different cell types and/or states. Besides detecting differential genes, it is also crucial to explore the reason of the significant alteration of gene expression level and its effect on other genes' activities. Thus an alternative method to differential genes identification is to detect differential gene relations in different cell states.

Compared to Mendelian disorders that are rare in population, some common complex diseases like various types of cancers are conjectured to be caused by two types of interactions related with multiple genetic factors: gene-gene interactions and gene-environment interactions [3]. Interactions between genes or single nucleotide polymorphisms (SNPs) in chromosomal regions are called *epistasis* [4]. Detecting epistasis associated with complex and common diseases became an important issue in human genetics [5]. While the recent development of genome-wide association studies (GWAS) [6] and the International Hapmap project [7-8] has made it possible

to identify common genetic variation or heritable risk factors in diseases from population-based data, the size of the genotyped data is typically very large and the number of combinations of all the genetic factors to be checked for the interactions is enormous, which cannot be exhaustively detected by experimental methods. Therefore, it is essential to detect causal interacting genes or SNPs by heuristic computational methods.

In this dissertation, I use machine learning and computational methods to address two problems in cancer research: (1) identifying differential gene relations and (2) detecting gene-gene interactions (epistasis). The detected differential gene relations and gene-gene interactions can help to build a new pavement towards the improvement of prevention, diagnosis and treatment of cancer.

1.2 My Contribution

One important area in microarray-based cancer research is to identify genes that are differentially expressed between cancerous and normal cells. However, a simple list of differential genes can not reflect activities and roles of cancer-related genes in a biological system. It is well known that genes interact with each other to form various biological pathways in order to carry out a multitude of biological processes. Hence, detecting differential gene relations in different cell states is a complementary approach to identifying cancer-related genes.

Several statistical methods have been proposed for the analysis of differential gene relations. Most of these methods often perform the analyses on a single

microarray dataset and typically generate unreliable results; the results from different microarray datasets and various statistical methods could hardly overlap using these methods. Therefore, the confidence level for discoveries based on these methods is low. Furthermore, these methods fail to grasp the common molecular changes in cells transitioning from a normal state to the cancerous state.

In this dissertation, a novel integrative method to detect the differentially changed gene relations in cancer versus normal tissues is presented. Comparing to those previous methods for detecting differential gene relations, I have made several contributions. First, I use bootstrapping K-S test to integrate multiple microarray datasets across different types of cancers. Integrating multiple microarray datasets can increase sample size, eliminate study-specific biases, and lead to more valid and more reliable results. Moreover, the integrative method can detect the most common altered gene relations across different types of cancer. Second, the searching process for differential gene relations is guided by the information of some key signaling pathways related with cancer. This helps to better understand the roles of cancer-related genes and their key interactions in a complex biological system. Third, a non-parametric statistical test method, K-S test, is used to compare two distributions. K-S test requires fewer assumptions for the data and may be preferred, especially, when the number of samples is small.

To detect gene-gene interactions and explore how these gene-gene interactions contribute to increase the disease risk, a number of statistical methods and machine learning methods have been proposed. Despite the success of statistical methods to

some degrees, they can only be applied to small-scale analysis due to their computational complexity. On the other hand, the common limitation of machine learning-based methods is that they typically identify a SNP set that produces the highest classification accuracy, but not necessarily has the strongest association with the diseases. Moreover, machine learning-based approaches tend to introduce many false positives, since the including of more SNPs increases classification accuracies.

To address the two critical challenges (searching and scoring) in gene-gene interaction detection, two methods are presented in this dissertation: DASSO-MB (Detection of ASSOciations using Markov Blanket) and EpiBN (Epistatic interaction detection using Bayesian Network model). Both methods are based on Bayesian Networks. Bayesian Networks provide a succinct representation of the joint probability distribution and conditional independence among a set of variables. Therefore we can use Bayesian Networks to represent the relationship between genetic variants and a phenotype (disease status).

DASSO-MB is a new Markov Blanket based method to detect gene-gene interactions in case-control studies. The Markov Blanket is a minimal set of variables, which can completely shield the target variable from all other variables based on the Markov condition property. Thus we can guarantee that the SNP set detected by DASSO-MB has a strong association with diseases and contains fewest false positives. Furthermore, DASSO-MB performs a heuristic search by calculating the association between variables to avoid the time-consuming training process as in some machine learning methods such as SVMs and Random Forests.

EpiBN is a Bayesian Network structure learning method, which employs a Branch-and-Bound technique and a new scoring function. In general, a structure learning method for Bayesian Networks first defines a scoring function reflecting the fitness between each possible structure and the observed data, and then searches for a structure with the maximum score. Comparing to Markov Blanket based methods, the merits of applying Bayesian Network structure learning method to gene-gene interaction detection include: (1) the new scoring function for Bayesian Network structure learning can reflect higher-order interactions and detect the true number of disease SNPs, and are not sample-consuming; and (2) heuristic Bayesian Network structure learning method can solve the classical XOR problem, which may hinder the applications of Markov Blanket based approaches.

The rest of my dissertation is organized as follows. Chapter 2 introduces the background about identifying differential gene relations and detecting gene-gene interactions (epistasis). It also introduces some methods on detecting differential genes and differential gene networks. Three machine learning and computational methods are respectively discussed in Chapter 3, 4, and 5.

(1) An integrative pathway analysis from multiple microarray datasets based on bootstrapping K-S test to identify differential gene relations.

(2) A Markov Blanket based approach for gene-gene interaction detection.

(3) A novel Bayesian Network structure learning method to detect gene-gene interactions.

Chapter 6 ends up with a summary of research in this dissertation and a discussion of future work.

Chapter 2 Background and Related Work

2.1 Differential Gene Detection

Cancer is one type of fatal disease caused by genetic aberrations, and identifying cancer-related genes is very important for prevention, diagnosis, and treatment of cancer. Most cancer-related genes show differential activities between normal cells and cancer cells, and thus we call these genes as differential genes. Inside the body of a healthy person, normal cells grow and die under the mechanisms that regulate cell growth and differentiation. However, if these molecular mechanisms regulating growth and death of cells are out of control, normal cells will develop into cancer cells. Typically, the transformation of normal cells into tumor cells is associated with two types of genes: oncogenes and tumor suppressor genes. Oncogenes and tumor suppressor genes act in opposite roles during the development of cancer. Oncogenes make normal cells gain self-sufficiency in growth signals and evade antigrowth signals and programmed cell death (apoptosis). On the contrary, tumor suppressor genes suppress the function of oncogenes. In cancer research, molecular biologists use microarray technology to measure gene expression, which reflects the activities of genes. Therefore we need a reliable method to detect differential genes based on microarray datasets.

2.1.1 Statistical Methods for Detecting Differential Genes

Fold Change Rule

The simplest method to identify differential genes is the fold change (FC) rule. The general fold change is as follows:

$$FC = \overline{X} / \overline{Y} \tag{2-1}$$

where $\overline{X}$ is the mean of expression value of one certain gene from normal samples, and $\overline{Y}$ is the mean of expression value of the same gene from tumor samples. Then the fold change rule will identify the gene as differential gene if $FC > m$ or $FC < 1/m$ based on the *m*-fold change rule. In this case, m is a fold increase/decrease cutoff to identify differentially expressed genes [9-10]. As a non-statistical and parametric method, how to select an optimal fold increase/decrease cutoff *m* is a big problem for fold change. We don not know which m value is better, 10, 5 or only 2. If we select a low m value, perhaps we will introduce a lot of false positives. On the contrary, if we select a high m value, we will take the risk of missing some true differential genes.

T-statistic

T-statistic is one method to detect differential genes from microarray datasets containing both normal samples and tumor samples. We define T-statistic as follows:

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{1/n + 1/m}} \tag{2-2}$$

where X represents normal microarray dataset with n samples, and Y represents tumor microarray dataset with m samples. S is the pooled sample standard deviation. There are two disadvantages for traditional T-statistic to detect differential genes. First, it assumes both X and Y have a normal distribution. But even microarray expression data after normalization may not satisfy this assumption. Second, if the expression levels of some certain genes are very low, the pooled sample standard deviation S will be extremely small for lack of sufficient information. Thus, the T-statistic will be very high, which may produce a significant bias. A variety of methods have been proposed to overcome the above two disadvantages of traditional T-statistic.

Significant Analysis of Microarrays (SAM)

One method to avoid the small variance problem of the T-statistic is to add a constant to its denominator. For instance, Tusher *et al*. proposed Significance Analysis of Microarrays (SAM) to detect differential genes [11]. The SAM statistic is

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{1/n + 1/m} + S_0} \tag{2-3}$$

where the value of S_0 is chosen to minimize the coefficient of variation. SAM also calculates the false discovery rate by the permutation of repeated measurements to estimate the percentage of differential genes identified by chance. Comparing to the standard T-statistic, SAM adds a small positive constant to the denominator of the T-statistic. By this modification, SAM will not select genes with low expression levels or small fold changes as significant differential genes. This eliminates the

small variances problem of the standard T-statistic and makes SAM more stable than standard T-statistic.

Bayes T-test

SAM is one variant of the standard T-statistic, and another variant of the standard T-statistic is Bayes T-test. Baldi and Long developed a Bayesian probabilistic framework, Bayes T-test, to solve the small variance problem in low expression level [12]. Bayes T-test estimates parameters such as population mean and variance by Bayesian method instead of sample mean and sample variance of the standard T-statistic. When deriving the variance of each gene, Bayes T-test combines the empirical variance with a local background variance associated with neighboring genes. By this way, Bayes T-test borrows some information from neighboring genes to solve the small variance problem caused by lack of samples. Baldi and Long showed that Bayes T-test outperforms the standard T-statistic on simulation data. Although Bayes T-test can analyze microarray dataset with small size for differential genes effectively, it still heavily depends on the parametric assumption.

B-statistic

Some methods based on B-statistic can also solve the small variance problem of the standard T-statistic for detecting differential genes. Lönnstedt and Speed proposed the B-statistic [13]. B-statistic is the logarithm of a ratio of probabilities, and the ratio of probabilities is equal to the probability that a gene is differentially expressed divided by the probability that the gene is not differentially expressed.

Lönnstedt and Speed estimated both probabilities from the entire microarray data by the empirical Bayes approach. The empirical Bayes approach shrinks the estimated gene-wise sample variances towards a pooled estimate (a common value) and combines information across genes. This can help B-statistic generate a more stable list of differential genes than the standard T-statistic when the number of samples is small.

2.1.2 Integrative Methods

Almost all applications of microarray technology in cancer research encounter an issue that the number of genes far exceeds the number of samples, which will lead to serious biases sometimes. Integrating multiple microarray datasets can solve this issue. There are two types of integration method: transformation methods and meta-analysis. Transformation methods transform gene expression data from different studies into a common scale and then combine these transformed data into one larger dataset. An alternative approach for integrating independent and heterogeneous microarray datasets into one large dataset is meta-analysis. Meta-analysis combines the summary statistics from each dataset. Commonly used summary statistics are significant levels (p-values), ranks of genes, and effect sizes. Both transformation methods and meta-analysis can increase sample size, eliminate study-specific biases, and lead to more reliable results.

Transformation Methods

Transformation methods are one type of integration methods, which translate gene expression measurements from different studies into a common scale and then unify these transformed data into one larger dataset [14-16]. For instance, Jiang *et al*. proposed a transformation method to integrate two cancer microarray datasets based on joint analysis [14]. First, they performed chip normalization which means the expression of each gene in each microarray was divided by the median of the microarray. Second, they filtered out genes that show significantly different expression patterns between the two datasets based on T-test. Third, they proposed a distribution transformation (disTran) method to let the two datasets have a similar distribution. Finally, they normalized each gene in the two datasets. Although Jiang *et al*. only selected a minimum number of marker genes from one dataset and applied these marker genes to the survival prediction based on the other dataset by the data normalization and transformation method [14]; however, we know that combining two transformed datasets into one larger dataset can increase sample size. This will yield more reliable results than those from one single dataset because the estimated parameters from the enlarged dataset are more confident.

Meta-analysis

An alternative approach for integrating microarray datasets is meta-analysis, and one type of summary statistic to integrate is p-value. For example, to generate a cohort of consistent differential genes from four prostate cancer microarray datasets, Rhodes *et al*. proposed a meta-analysis method to combine p-values (extreme value

probabilities) for each gene from the four microarray datasets [17]. This meta-analysis method is based on Fisher's Inverse χ^2 test, which combines p-values p_i obtained from the analysis of the *i*th dataset by

$$S = -2\sum_i \log(p_i) \tag{2-4}$$

where S follows a χ^2 distribution with $2I$ degrees of freedom under the joint null hypothesis [18]. Traditionally, we calculate the overall p-value of S based on the χ^2 distribution with $2I$ degrees of freedom. Rhodes *et al*. used a random permutation method to generate the overall p-value of *S*. They first generated 100,000 *S*s randomly and then compared the *S* with the 100,000 random *S*s. The overall p-value of *S* equaled the fraction of random *S*s that were greater than or equal to *S*. Adding weights for each dataset in Eq. (2-4) is an advanced method. We can assign weights to each dataset based on data quality or on other factors considered important and now

$$S = -2\sum_i w_i \log(p_i) \tag{2-5}$$

P-value-based meta-analysis method can increase statistical power by detecting consistent differential genes that might be false negatives in the individual microarray dataset. It is expected that true differential genes will have high p-values in most datasets. Rhodes *et al*. implemented their model on four prostate cancer microarray datasets coming from different platforms; two datasets are based on spotted cDNA technology and the remainders are based on oligonucleotide-based technology. The

resulting differential genes were more reliable and helped them to reconstruct the transcriptional events of two metabolic pathways important in prostate cancer [17].

The T-statistic and various modified T-statistics are the most widely used statistics for identifying differential genes, and integrating the T-statistic or various modified T-statistics is another meta-analysis method. Choi *et al*. proposed an integration method based on effect size model using a T-like statistic (defined as effect size) as the summary statistic for each gene from each individual dataset [19]. They applied a hierarchical model to estimating both within- and between-study variances, which they used as weights when combining the summary statistic across multiple datasets. Then they obtained an overall estimate of the average effect size through parameter estimation and model fitting. Like the method in [17], they determined the statistical significance of the average effect size by a permutation test. A better method proposed by Hu *et al*. used a quality measure to weight the importance of each gene in each experiment [20]. They incorporated this quality measure into the effect size method to model inter-study variation of gene expression profiles. There are several merits for integration methods based on effect size model. For example, effect size model uses a standard index (T-like statistic) and is a well-established statistical framework for combining different microarray datasets. Another merit of effect size model is that it has the ability to handle the variability between different microarray datasets.

The integration method based on rank is an alternative method to the above two meta-analysis methods. Meta-analysis methods based on p-values or T-statistic are

parametric methods, and their performance is heavily dependent on the estimation of parameters. To avoid this shortcoming, Breitling *et al*. proposed a non-parametric rank product (RP) method to integrate multiple microarray datasets for differential gene detection [21]. The rank product method first ranks genes by the FC criterion introduced in section 2.1.1. Assume one gene have ranks $r_1, r_2, ..., r_n$ from n microarray datasets, the rank product of this gene is as follows:

$$RP = (\prod_i r_i)^{1/n} \tag{2-6}$$

Breitling *et al*. permuted the expression value within each array in a microarray dataset to determine the statistical significance of *RP* in Eq. (2-6). Basically, the rank product method computes FC for each gene, transforms FC into rank among all genes in each microarray dataset, then searches for genes that are consistently top ranked across multiple microarray datasets. Converting FC into ranks overcome the heterogeneity among multiple microarray datasets because some researchers demonstrated that, although the differential gene lists from fold-change method had poor consistency across multiple microarray datasets, the rank orders of genes were comparable [22]. Therefore, the rank product method can detect genes that are consistently differential genes in a number of microarray datasets.

2.2 Differential Coexpression Analysis

An alternative method to differential gene identification is to detect differential coexpression patterns in different cell states. A simple list of differentially expressed

genes, however, only tells us which genes are altered by biological differences between different cell types and/or states. In a biological system, genes are well known for forming a variety of complex networks to perform different molecular functions and regulate various biological processes. Besides detecting differential genes, it is also crucial to explore the reason of significant alteration of gene expression level and its effect on other genes' activities. Hence, detecting differentially changed gene relations in different cell states is a complementary approach to identifying cancer-related genes. Researchers proposed several statistical or machine learning methods for the analysis of differential gene relations based on different score schemes of coexpression and different searching methods.

The biclustering method is one method to detect the differential coexpression of genes. Kostka and Spang observed that some genes show no differential expression between normal and tumor samples; however, in normal samples these genes display a coexpression pattern, which disappears in tumor samples [23]. Thus, they proposed a method to investigate differentially coexpressed groups of genes that displays a striking difference in the coexpression pattern between two different types of samples. Kostka and Spang chose the mean squared residual of an additive model used in biclustering method [24] for scoring coexpressed groups of genes and then searched groups of genes showing differential coexpression patterns by a greedy stochastic downhill search algorithm, which is an heuristic algorithm for finding groups of genes with low scores. One problem for Kostka and Spang's method is that they only focus on detection of groups of genes that show significant difference between

normal samples and tumor samples but omit the biological meaning of these groups of genes. Moreover, the authors can not provide the biological explanation of detected groups of genes.

Another approach to detect the differential coexpression of groups of genes is based on the hierarchical cluster method. Watson developed CoXpress to identify groups of genes that are differentially coexpressed. There are two phases in CoXpress [25]. In the first phase, CoXpress selects several groups of coexpressed genes by the hierarchical cluster method. In the second phase, CoXpress determines whether these groups of gene are differentially coexpressed between normal samples and tumor samples by a resampling approach. Generally, cluster analysis generates groups of genes that are correlated with each other. However, this is only a static analysis, which can not indicate the change of the coexpression pattern of genes. CoXpress paves a pathway for the dynamic analysis of the coexpression pattern of genes.

Some methods for the analysis of differential gene relations only focus on the detection of differential coexpresssion pattern of gene pairs. This will reduce the searching space to n^2, assuming the number of genes is n. One method to detect differential gene pairs is Liquid Association (LA). Li observed differences of gene coexpression patterns in different cellular states and attributed these changes in gene coexpression patterns to some third influential genes [26]. Therefore, Li proposed a LA method that conducts a genome-wide search and identifies the most critical influential genes that may affect the coexpression pattern for any two genes. He used the term Liquid Association to define the internal evolution of coexpression pattern

for a pair of genes. Li *et al*. also proposed a strategy to generalize LA method for multiple genes [27]. Lai *et al*. proposed a similar method to identify differential gene-gene coexpression patterns in cells from normal state to cancerous state based on expected conditional F-statistic [28].

The LA method is suitable to detect gene coexpression patterns. First, different cellular states can alter the biological roles of genes and break up the joint activities of a pair of interacting genes. Thus, if some third influential genes change the cellular state, LA method can detect association changes of gene pairs by conditioning on these third influential genes. On the other hand, the genome-wide search in LA method can find these unknown influential genes.

Another method of detecting differential coexpresssion pattern of gene pairs is based on the comparison of Pearson correlation coefficients from different types of samples. Dettling *et al*. proposed a novel approach, CorScor, to find gene pairs with joint differential expression as a complement to the widely used one-gene-at-a-time testing methods [29]. CorScor first defines the score function

$$S(\rho, \rho_0, \rho_1) = | \rho_0 + \rho_1 - \alpha\rho | \tag{2-7}$$

to help to find gene pairs with differential coexpresssion pattern, where ρ_0, ρ_1, and ρ are Pearson correlation coefficients from normal samples, tumor samples, and the whole samples. Then, CorScor performs an exhaustive search to find gene pairs with differential coexpresssion pattern. CorScor is a straightforward method, and it can be performed very quickly, which is the biggest advantage.

Detecting differential coexpresssion pattern by integrative methods can generate more confident results. The above methods for detecting differential coexpresssion pattern performed analysis on one single microarray dataset and will face the same problem as identifying differentially expressed genes where the results from different microarray datasets and various statistical methods can hardly overlap [30-31]. Choi *et al*. introduced a model to find differential coexpression patterns related to cancer by combining independent datasets for different cancers [32]. They calculated correlation values from several microarray datasets for normal samples and tumor samples. Then they used effect size model to construct two distinct gene networks for normal samples and tumor samples and compared the difference between the normal gene network and tumor gene network. Integrative methods based on multiple microarray datasets can increase the confidence of the results and grasp the common molecular changes in cells from normal state to cancerous state.

2.3 Detection of Differential Gene Networks

Detecting differential genetic signaling pathways (gene networks), which respond to different cell states, is a superior approach for cancer research [33]. The shortcoming of the above two methods to detect differential genes or differential gene relations is that they only investigate a single and isolated element in a biological system, but neglect the integrity of the entire system [34]. System biology should explore the behavior of all elements in a biological system and the relationships among them in order to model and ultimately control the mechanism of the biological

system [35]. Therefore, detecting differential genetic signaling pathways is a better tool for cancer research according to the perspective of system biology.

One key issue for detecting differential genetic signaling pathways is how to measure pathway expression. Levine *et al*. used five different measures of pathway expression to analyze gene-set activation: Z-score, Hypergeometric, Principal component analysis, Wilcoxon Z-score, and Kolmogorov-Smirnov [36]. These five measures are based on gene expression. They found most results from these five measures are the same. However, some incoherent pathways can only be identified as differential by a subset of the measures. On the other hand, Rahnenfuhrer *et al*. proposed a measure of pathway activity based on Pearson correlation coefficient of gene pairs [37]. They first calculated the mean of Pearson correlation coefficients of all possible gene pairs in a pathway and then determined the statistical significance of changes of pathway activity by a non-parametric permutation test. The above two pathway expression measures only consider one aspect of the activity of a pathway. A better measure scheme should consider changes of both genes and gene relations. The change of expression level reflects the altered activity of a gene and the change of gene relation reflects the alteration of gene functions. These two aspects of the inside mechanism in living cells are equally important.

Some differential network detection methods first identify differential genes and then construct a differential network by these differential genes. Traditionally, identifying differential genes and detecting differential subnetworks are two separate tasks. Sanguinetti *et al*. integrated identifying differential genes into the task of

differential gene network detection [38]. They first introduced a Mixture Model on Graphs (MMG) to detect differential genes. Then they identified coherent differential submodules by a simple percolation algorithm. Starting from a given node (differential gene), the percolation algorithm extends the submodule by iteratively adding all the neighboring nodes which are differential. Sanguinetti *et al*. demonstrated that certain gene networks are consistently differentially expressed and have a clear biological meaning in terms of cellular metabolic functions, which were validated by high-throughput proteomic experiments. However, Sanguinetti *et al*. only considered the difference of gene expression and omitted the importance of the alteration of gene relations.

Another approach for differential gene network detection constructs a whole gene network from a variety of biological databases first and then determines which sub-networks are differential. The assumption of this approach is that coherent sub-networks would show differential activities. Cabusora *et al*. constructed a large biological network from protein interaction, metabolic reactions and gene coexpression databases [39]. They selected seed genes from this biological network by machine learning methods such as genetic algorithms or singular value decomposition. Next they filtered sub-networks with the shortest paths between each pair of seed genes and the highest mean of Pearson correlation coefficients of all gene pairs in the sub-network. Finally, they considered these sub-networks as differential networks. The problem of Cabusora *et al*'s method is that not all coherent networks are differential networks. Short paths and high Pearson correlation coefficients for

most gene pairs in a coherent network do not mean that these genes will respond to different cell states simultaneously.

The alteration of gene network connectivity is also important in the detection of differential gene network. Fuller *et al*. argued that differential network analysis should be concerned with identifying both differentially connected and differentially expressed genes [40]. So they considered the change of the connectivity of genes and defined a measure of differential connectivity. For the *i*th gene in a gene network, they represented its connectivity in networks 1 and 2 by $k_1(i)$ and $k_2(i)$, respectively. For convenient comparison between the connectivity measures of each network, they normalized the connectivity of each gene by the maximum network connectivity as follows:

$$K_1(i) = \frac{k_1(i)}{\max(k_1)} \tag{2-8}$$

and

$$K_2(i) = \frac{k_2(i)}{\max(k_2)} \tag{2-9}$$

Next they defined the differential connectivity of the *i*th gene as:

$$DiffK(i) = K_1(i) - K_2(i) \tag{2-10}$$

To select some interesting gene modules (differential gene networks), they produced a scatter plot of differential connectivity vs. T-statistic for each gene. The scatter plot demonstrates how differential connectivity relates to the traditional T-statistic

describing differential gene expression between two networks. Fuller *et al*. found some significant sectors in the scatter plot and showed that the genes in these sectors are related with some specific molecular functions. The differential gene network detection method based on differential connectivity tries to find differential network consisting of genes that are both differentially expressed and differentially connected. Searching for differentially connected genes focuses on the preservation of modules between two cell states because genes in these modules are highly connected in coexpression networks. Although differentially connected genes may or may not be differentially expressed, changes in connectivity of genes may reveal their responses to environmental alterations.

2.4 Detection of Gene-Gene Interactions

Detecting gene-gene interactions is critical for pathogenesis, prevention, diagnosis, and treatment of complex human diseases, and designing an efficient computational method to detect gene-gene interactions presents a challenge to the bioinformatics society. Epistasis refers to the joint and interactive effect of two or more genetic variants on complex human diseases. Interactions among multiple genetic factors can result in some common complex diseases such as various types of cancers, cardiovascular disease, and diabetes. Genome-wide association study (GWAS) is an examination to check the genetic variants from individual to individual and the number of the SNPs (single-nucleotide polymorphism) to be checked in a typical GWAS is up to 10 million. Moreover, the number of possible

combinations of SNPs is enormous. Therefore, we must resort to some heuristic computational methods to detect gene-gene interactions. There are two types of methods to detect gene-gene interactions: statistical methods and machine learning methods.

2.4.1 Statistical Methods for Detection of Gene-Gene Interactions

Statistical methods are one type of computational methods for gene-gene interaction detection, and the most commonly-used parametric statistical method is logistic regression. Marchini *et al*. tried to fit the logistic regression method to three genetic interaction disease models [41]. Logistic regression predicts the probability of disease based on the combination of independent SNPs and finds an optimal logical SNP set which can generate the highest probability of disease. When used for modeling high-order interactions with small number of samples, the estimation of a large number of parameters is not confident because of the poor number of samples per parameter. This will often results in an overfitting problem.

There are some methods to overcome problems in logistic regression method, and MDR (multifactor dimensionality reduction) is one of them. Ritchie *et al*. proposed a multifactor dimensionality reduction method to detect statistical patterns of epistatis [42]. MDR first constructs a risk table for every SNP combination. If the cases/controls ratio in a cell of this risk table is larger than 1, MDR will label it as "high risk", otherwise, MDR will label it as "low risk". By the risk table, MDR can predict disease risk and will select the SNP combination with the highest prediction

accuracy. MDR is a novel method for gene-gene interaction detection in its construction of a risk table for prediction. Moreover, unlike logistic regression, MDR is non-parametric and model-free. However, the process of labeling each cell as "high risk" or "low risk" is a process of estimating parameters. This will lead to a huge number of parameters to be estimated when the size of SNP combination is large. Furthermore, MDR has two fundamental limitations: (1) MDR selects the SNP combination purely by the prediction performance. This type of method can not find true causal factors because the high prediction accuracy of a SNP set does not mean that this SNP set has a strong association with disease and might cause disease. (2) MDR employs an exhaustive searching strategy to avoid local optima. Thus MDR is impractical for large-scale datasets.

Some variants of logistic regression can also overcome problems in standard logistic regression method for detecting gene-gene interactions. Park and Hastie proposed a penalized logistic regression (stepPLR) using a forward stepwise method to detect gene-gene interactions [43]. StepPLR makes some simple modifications for standard logistic regression. For example, stepPLR combines the LR (Logistic Regression) criterion with a penalization of the L2-norm of the coefficients. This modification makes stepPLR more robust to high-order gene-gene interactions. However, stepPLR is time-consuming when estimating parameters, which is one essential limitation of regression methods. Moreover, like standard LR and MDR, stepPLR is also based on prediction, and this is the common limitation of most gene-gene interaction detection methods.

Some researchers apply the Bayesian method to detecting gene-gene interactions. Zhang and Liu proposed a Bayesian epistasis association mapping (BEAM) method. BEAM partitions SNPs into three groups: group 0 is for normal SNPs, group 1 contains disease SNPs affecting disease risk independently, and group 2 contains disease SNPs that jointly contribute to the disease risk (interactions) [44]. Given a fixed partition, BEAM can get the posterior probability of this partition from SNP data based on Bayes theory. Thus, BEAM is a Bayesian marker partition model to identify both single disease SNP and SNP combination with maximum posterior probability. Zhang and Liu used a Markov Chain Monte Carlo method to reach the optimal SNP partition in BEAM. Zhang and Liu also proposed a new B statistic to check each SNP or set of SNPs for significant associations with the disease. The experiment results on the synthetic data from six disease models demonstrated that BEAM is more powerful than other approaches such as MDR and stepwise logistic regression. However, the performance of BEAM is worse than that of some recently proposed methods for gene-gene interaction detection such as SNPHarvester [45]. One possible reason is that BEAM is over-complex. Zhang and Liu tried to detect single disease SNPs and gene-gene interactions simultaneously in BEAM, which impairs the performance of BEAM.

2.4.2 Machine Learning Methods for Detection of Gene-Gene Interactions

The alternative approaches for statistical methods to detect gene-gene interactions are machine learning methods. Machine learning methods for gene-gene

interaction detection are based on binary classification (prediction) and treat cases as positive samples and controls as negative samples in SNP data from GWAS. Chen *et al*. proposed to use the Support Vector Machine (SVM) method to detect gene-gene interactions [46]. SVM is a state-of-the-art classification/prediction method and they used SVM to select a combination of SNPs with the highest prediction accuracy and transform detecting gene-gene interactions into a process of feature selection. Chen *et al*. tried three feature selection methods: RFE (recursive feature elimination), RFA (recursive feature addition), and GA (genetic algorithm) and found that the performance of GA is the best. Jiang *et al*. adopted random forests, which is an ensemble learning technique and can also be used as a classifier/predictor, to detecting gene-gene interactions in GWAS [47]. They first ranked SNPs based on the importance of each SNP and then performed a greedy search for a small subset of SNPs with the capacity of minimizing the classification error. Both SVM and random forests show greater powers than MDR on the synthetic data. However, prediction-based methods can not detect true causal factors like MDR. Moreover, the feature selection process for prediction-based methods is time-consuming, which means that we can not apply them to genome-wide datasets directly.

2.5 Genetics of Gene Expression

In the past few years, researchers combine genetic and gene expression approaches under the name of 'genetical genomics' and study the genetic basis of variation in gene expression. Even though the significances of the association signals

in some GWAS are extraordinary, the detection of gene-gene interactions based on SNP data from GWAS can not provide us a full understanding of how genetic variants contribute to disease susceptibility. Polymorphisms in regulatory regions of gene sequence can regulate gene expression directly, and variation in gene expression is probably a major mechanism affecting the risk of complex diseases. Therefore, some researchers conduct studies of genetics of gene expression, whcih are referred to as GOGE (genetics of gene expression) and also known as expression quantitative trait loci (eQTL) studies or genetical genomics [48-49]. It has been known that gene expression levels are controlled by a combination of cis- and trans-acting regulators. However, the goal of GOGE studies is not to identify all the cis- and trans-acting regulators but to find polymorphic variants that contribute to gene expression variation. In fact, identifying the precise causal sequence variants is a challenging task. GOGE combines whole-genome genetic association studies and the microarray data of global gene expression to identify genetic factors that affect gene expression. Results from GOGE are then functionally investigated to obtain a clear map from SNPs to diseases.

GOGE studies try to identify the DNA variants (polymorphisms) that influence expression levels of genes — that is, the gene expression phenotype. There are three merits for GOGE studies. First, GOGE studies construct a bridge between variations at the DNA sequence level and variations at the RNA level. There are over 3 million SNPs, and most of them are presumably neutral, while some are functional. However, determining which SNPs are functional is challenging. GOGE studies can find

regions and ultimately variants that regulate gene expression. Furthermore, we can compare the susceptibility SNPs for human diseases from GWAS with results from GOGE studies to perform consistency test. Second, GOGE studies identify variants that influence gene expression by scanning the genome for regulators without prior knowledge of the regulatory mechanisms. Therefore, GOGE studies can identify unknown regulators of gene expression. Third, in addition to identifying regulators of individual genes, GOGE studies can be applied to genetic regulatory network analysis. GOGE studies treat gene expression as a phenotype to identify regulators that influence the expression levels of individual genes. Many GOGE studies results show that most identified regulatory variants are close to the target (regulated) gene. GOGE studies have the ability to survey the genome for regulatory variants and can uncover novel regulatory mechanisms and assign new roles to known genes by identifying regulatory variants. Thus, we can construct a more confident genetic regulatory network by combining results from GOGE studies with correlation analysis.

Chapter 3 Detecting Differential Gene Relations by Bootstrapping K-S Test

To better understand the roles of differentially expressed genes in a complex biological system, a comprehensive pathway analysis is needed to find the most common pathways, which reveal the relationship between these genes. Biological pathways are significantly influenced by those differentially expressed genes from different datasets or different statistical methods. Moreover, it is crucial to explore the reason of the significant alteration of gene expression level and its effect on other genes' activities. It is well known that in a biological system genes form a variety of complex networks to perform different molecular functions and regulate various biological processes. Hence, it is also important for us to detect gene relation alterations and to explore how these changes of gene relations affect some key pathways related to cancer. To detect the differentially changed gene relations between cancer and normal tissues [50], a novel integrative method based on multiple datasets across different microarray platforms and from various types of cancer is developed in this dissertation.

3.1 Methodology

3.1.1 Microarray Datasets and Genetic Signaling Pathways

We collect 36 microarray datasets from NCBI GEO (Gene Expression Omnibus) [51]. These microarray datasets contain both normal samples and tumor samples across 21

different types of cancer and their platforms come from one of the three platforms: GPL570 (Affymetrix GeneChip Human Genome U133 Plus 2.0 Array), GPL96 (Affymetrix GeneChip Human Genome U133 Array Set HG-U133A) and GPL91 (Affymetrix GeneChip Human Genome U95 Version Set HG-U95A). We divide every dataset into two expression data matrix: one matrix includes all normal samples and the other includes all tumor samples. To integrate multiple microarray datasets across different platforms, we map each probe in different platforms to a unique Entrez Gene ID or a unique UniGene symbol. For genes with more than one probe in one platform, we choose the probe with the highest mean expression value.

We apply our method to analyze three cancer-associated pathways. These pathways are related to three common traits in most and perhaps all types of human cancer: self-sufficiency in growth signals, insensitivity to antigrowth signals, and evading programmed cell death (apoptosis) [1]. In fact, Hanahan and Weinberg have already shown some signaling pathways to demonstrate some capabilities cancer cells acquire during tumor development in [1]. We extend these signaling pathways to three relatively complete and larger cancer-associated pathways (antigrowth signaling pathway, apoptosis pathway, and growth signaling pathway) from the cell cycle pathway, the apoptosis pathway and the MAPK pathway in KEGG [52]. We use these three pathways (i.e., cell cycle, apoptosis and MAPK pathways) as our seeds and the genes in these pathways as our seed genes. Next we construct three gene networks corresponding to the three cancer-associated pathways from HPRD (Human Proteins Reference Database, http://www.hprd.org/) and TRANSFAC [53] based on

seed genes and their interacting partners. We download the protein-protein interaction (PPI) data released by HPRD on Sep 1, 2007. This PPI dataset contains 37107 human binary protein-protein interactions whose supporting experiments are indicated as *in vivo*, *in vitro* or yeast two-hybrid. We also collect 1042 transcription factor-target gene relations on human species from TRANSFAC. So our gene networks include seed genes, protein interaction partners and transcription factors (TFs) of seed genes or target genes for which seed genes serve as their TFs.

3.1.2 Kolmogorov–Smirnov (K-S) Test

Kolmogorov–Smirnov test (K-S test) can determine whether the distributions of values in two data sets differ significantly. The two-sample K-S test is most useful for comparing two samples because it is non-parametric and distribution free [54]. The null hypothesis for this test is that two data sets are drawn from the same distribution. The alternative hypothesis is that they are drawn from different distributions.

For n iid samples $X_1,\ldots,X_n$ with some unknown distribution, we can define an empirical distribution function by

$$S_n(x)=\begin{cases}0 & \text{if} \quad x<X_1\\ k/n & \text{if} \quad X_k\le x<X_{k+1} \quad \text{for } k=1,2,\ldots,n-1\\ 1 & \text{if} \quad x\ge X_n\end{cases} \tag{3-1}$$

where $X_1,\ldots,X_n$ are ordered from smallest to largest value.

Then the Kolmogorov-Smirnov statistic for a given function $S(x)$ is

$$D_n = \max_x |S_n(x) - S(x)| \tag{3-2}$$

D_n will converge to 0 if the sample comes from distribution $S(x)$ [54]. Moreover, the cumulative distribution function of Kolmogorov distribution is

$$K(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty}e^{-(2i-1)^2\pi^2/(8x^2)} \tag{3-3}$$

It is easy to prove that $\sqrt{n}D_n = \sqrt{n}\max_x |S_n(x) - S(x)|$ will converge to Kolmogorov distribution [54]. Therefore if $\sqrt{n}D_n > K_\alpha = \Pr(K \le K_\alpha) = 1 - \alpha$, the null hypothesis for the Kolmogorov-Smirnov test will be rejected at level α.

For the case of determining whether the distributions of two data vectors differ significantly, the Kolmogorov-Smirnov statistic is

$$D_{n,m} = \max_x |S_n(x) - S_m(x)| \tag{3-4}$$

and the null hypothesis will be rejected at level α if

$$\sqrt{\frac{nm}{n+m}}D_{n,m} > K_\alpha \tag{3-5}$$

The p-value from K-S test can measure the confidence of the comparison result against the null hypothesis. It is obvious that the smaller the p-value, the more confident we are to reject the null hypothesis.

3.1.3 Bootstrapping K-S Test

We use the Kolmogorov-Smirnov test (K-S test) to determine whether the distributions of values in two datasets differed significantly. Assume that we have n microarray datasets and a list of *m* genes, we denote the expression data matrix for normal samples as:

$$N^k = \begin{pmatrix} X_{11}^k & X_{12}^k & \dots & X_{1p(k)}^k \\ X_{21}^k & X_{22}^k & \dots & X_{2p(k)}^k \\ . & . & . & . \\ X_{m1}^k & X_{m2}^k & \dots & X_{mp(k)}^k \end{pmatrix} \quad k=1,\dots,n \tag{3-6}$$

and the expression data matrix for tumor samples as:

$$T^l = \begin{pmatrix} Y_{11}^l & Y_{12}^l & \dots & Y_{1q(l)}^l \\ Y_{21}^l & Y_{22}^l & \dots & Y_{2q(l)}^l \\ . & . & . & . \\ Y_{m1}^l & Y_{m2}^l & \dots & Y_{mq(l)}^l \end{pmatrix} \quad l=1,\dots,n \tag{3-7}$$

where $p(k)$ is the number of normal samples in the kth dataset and $q(l)$ is the number of tumor samples in the lth dataset.

For these two types of expression data matrix, each row represents one gene and each column represents one sample. The correlation coefficient for gene *i* and gene *j* from the kth normal sample can be calculated by

$$NPC_{ij}^k = \frac{\sum_{a=1}^{p}(X_{ia}^k - \overline{X}_i^k)(X_{ja}^k - \overline{X}_j^k)}{\sqrt{\sum_{a=1}^{p}(X_{ia}^k - \overline{X}_i^k)^2}\sqrt{\sum_{a=1}^{p}(X_{ja}^k - \overline{X}_j^k)^2}} \tag{3-8}$$

where $\overline{X}_i^k$ and $\overline{X}_j^k$ are the average of expression levels for gene *i* and gene *j*.

The correlation coefficient for every gene pair from tumor samples can be calculated similarly.

We use bootstrapping K-S test to detect some gene relations with different PC (Pearson coefficient) distribution.The bootstrapping method can give us an empirical distribution of p-value θ, with which, we can estimate the probability that the distribution of two PC vectors are different. In our computational experiment, for a gene pair, if its value of $\Pr(\theta < 0.05)$ is larger than 0.8, we consider it as a pair of genes with the correlation relation significantly different between normal and cancer cells.

Our method can be described as follows:

Step1. Compute *n* correlation coefficient Matrices $NPC^1 — NPC^n$ from the normal samples in *n* datasets for every gene pairs. For example, NPC^1 is an $m \times m$ Matrix from normal samples in the 1st dataset and NPC_{ij}^1 represent the correlation coefficient between gene *i* and gene *j*.

Step2. Compute *n* correlation coefficient Matrixes $TPC^1 — TPC^n$ from the tumor samples in the *n* datasets for every gene pair.

Step3. For every gene pair (gene *i* and gene *j*), let

$NPC_{ij} = [\, NPC_{ij}^1 \quad NPC_{ij}^2 \quad NPC_{ij}^3 \ \ldots NPC_{ij}^n \,]$

$TPC_{ij} = [\, TPC_{ij}^1 \quad TPC_{ij}^2 \quad TPC_{ij}^3 \ \ldots TPC_{ij}^n \,]$

Step4. Perform the following:

For k=1 to *N*

Do generate bootstrapping samples *NPC* and *TPC* from NPC_{ij} and TPC_{ij} respectively

θ_k = p-value of K-S test on *NPC* and *TPC*.

End –For

Output $\Pr(\theta < 0.05) = \#(\theta < 0.05) / N$.

During step 4, we generate *N* bootstrapping samples NPC and TPC by repeatedly sampling with replacement from the original NPC_{ij} and TPC_{ij} respectively. When using bootstrapping, we randomly extract a new element from the original sample every time and then put it back before extracting the next element until the sample size of the bootstrapping sample NPC(TPC) is the same as that of the original NPC_{ij}(TPC_{ij}). Therefore each element in the original sample can be selected many times.

3.2 Experimental Results

We apply the bootstrapping K-S test method for analyzing three cancer related pathways: antigrowth signaling, apoptosis, and growth signaling pathways. The experimental results of our method on these three genetic signaling pathways are demonstrated in this section.

Antigrowth signaling pathway

Antigrowth signals can control proliferation in normal samples. Cancer cells have the ability to evade these antiproliferation signals. In the antigrowth signaling pathway**,** transforming growth factor beta (TGFβ) initiates this pathway by binding to two TGFβ receptors: Tgfbr1 and Tgfbr2 [1]. These two activated Tgfβ receptors can phosphorylate Smad2, Smad3, and Smad4 [55]. The SMAD family proteins then transduce antigrowth signals to cell cycle inhibitors: p21, p16, p27, and p15, which can inhibit the action of cyclin-CDK complex. The cyclin-CDK complex can phosphorylate RB and make RB dissociate from the E2F/RB complex to liberate E2F

to activate the cell cycle procession from G1 to S phase (Figure 3.1A).

There are 19 genes in the antigrowth signaling pathway (Figure 3.1A). We can find 689 unique genes related to these 19 genes from TRANSFAC and HPRD. Among these 708 genes, there are 4215 paired gene interactions, among which the correlation relations of 47 gene pairs are identified as significantly changed between normal and cancer cells. Among these 47 relations, we detect a cluster around SMAD family proteins which contains 15 relations with different distribution between normal samples and tumor samples (Figure 3.1B). Most of them come from large-scale protein-protein interaction experiments without the associated molecular function. For example, (Smad1—Arl4d), (RHOD—Smad2) and (WEE1--Smad3) in [56], (PAPOLA—Smad2), (SNRP70—Smad5), (GPNMB—Smad4), (PSMD11-Smad3) and (Smad9—MBD1) in [57] and (EWSR1—Smad4) in [58], all of them are detected from large-scale protein-protein interaction experiments without annotation of molecular function. Our results indicate that although their associated functions and internal mechanisms are still unclear, these gene pairs are related to the Tgfβ-SMAD signaling pathway and the relation between the two genes in each pair is significantly different in cancer and normal cells. Additionally, we identify some differentially changed relations with known molecular functions as listed below:

(1) MAGI2 (a.k.a. ARIP1)—Smad3. MAGI2 (ARIP1) can interact with Smad3 and overexpression of ARIP1 can significantly suppress Smad3-induced transcriptional activity [59]. We can validate this from the boxplot for MAGI2 (ARIP1)--Smad3 (Figure 3.2A). In normal samples, MAGI2 (ARIP1)

and Smad3 show a high positive correlation, while they have a high negative correlation in tumor samples.

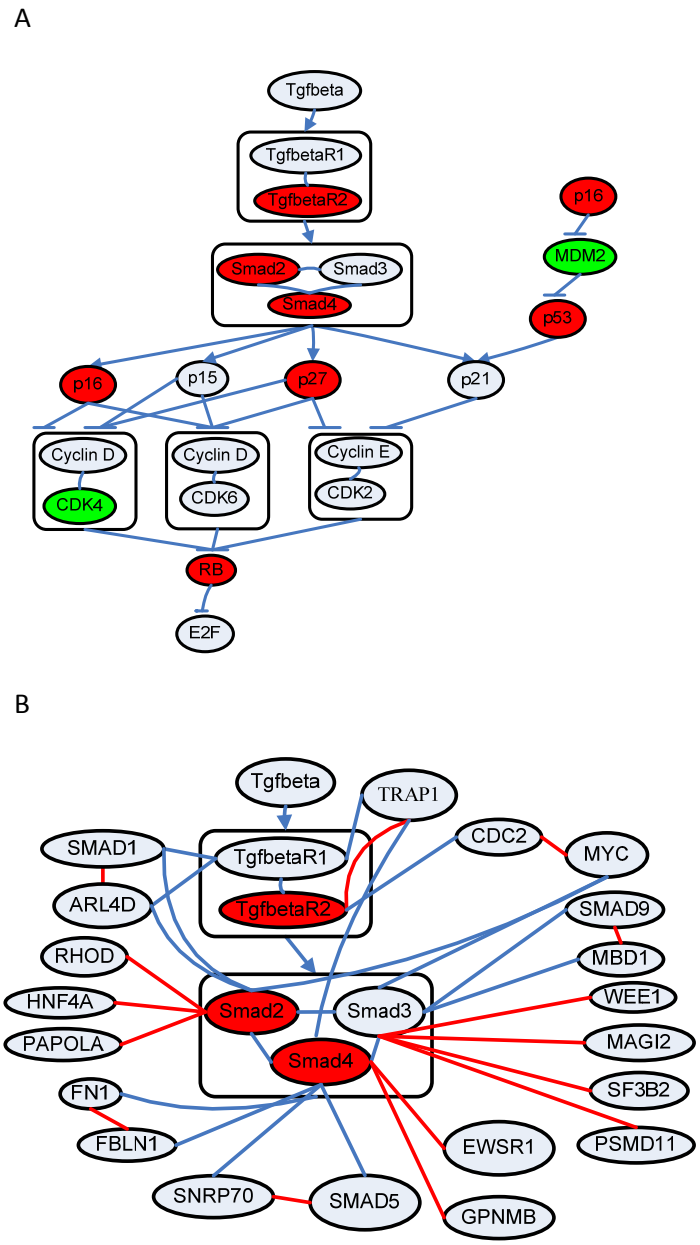

Figure 3.1 Antigrowth signaling pathway. (**A**) Antigrowth signaling pathway. Nodes and edges represent human proteins and protein-protein interactions respectively. Edges with direction represent a regulatory relation. → means an activating relation and --| means an inhibitory relation. (**B**) Cluster around smads. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes and green nodes represent oncogenes.

(2) EWSR1—Smad4. Although the experiment type of the interaction between EWSR1 and Smad4 is yeast two-hybrid [58], mutations in EWSR1 are known to cause Ewing sarcoma and other members of the Ewing family of tumors [60]. From the boxplot for EWSR1--Smad4, we find that the third quartile is the densest part of the whole distribution for both normal and tumor samples. But the third quartile for normal samples shows a positive correlation, while it shows a negative correlation for tumor samples (Figure 3.2B). So we suspect that EWSR1 can suppress the activity of Smad4 in tumor samples.

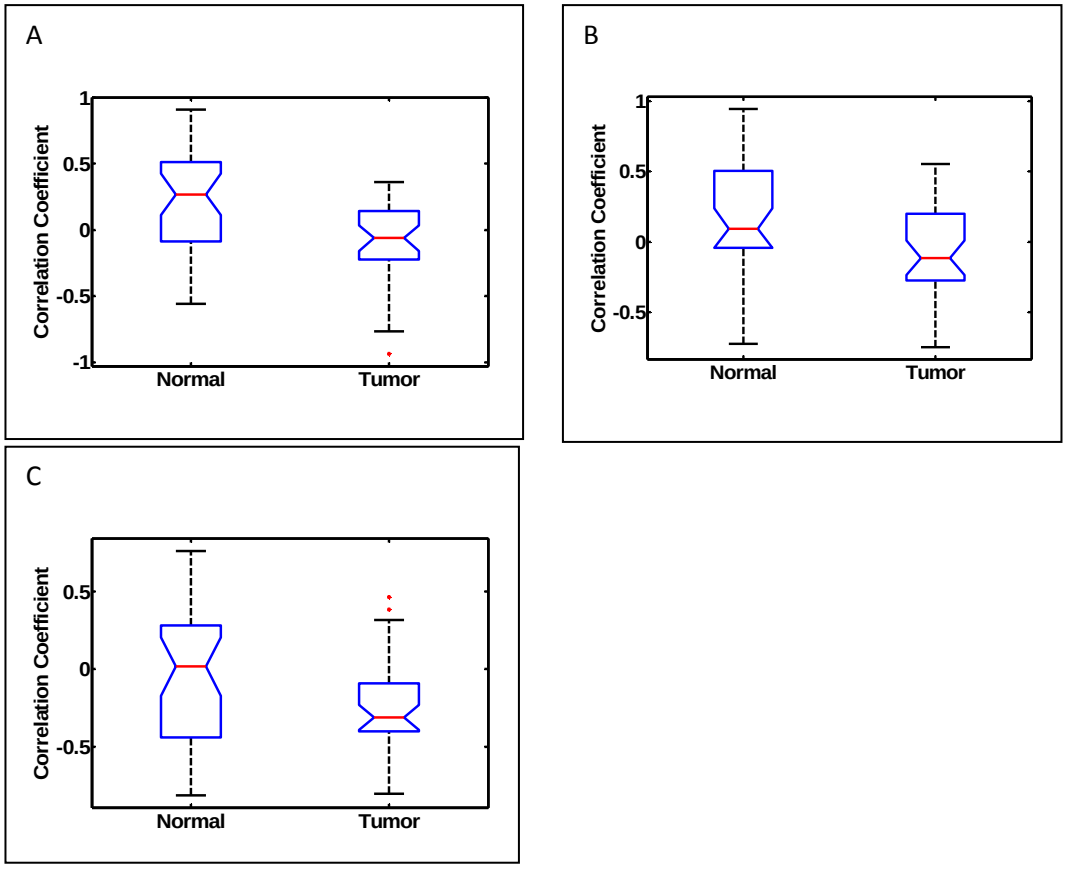

Figure 3.2 Boxplots for differential gene relations in antigrowth signaling pathway. (**A**) Boxplot for MAGI2 (ARIP1)—Smad3. $\Pr(\theta < 0.05) = 0.986$. (**B**) Boxplot for EWSR1—Smad4. $\Pr(\theta < 0.05) = 0.954$. (**C**) Boxplot for TRAP1—TgfbetaR2. $\Pr(\theta < 0.05) = 0.944$.

(3) TRAP1—Tgfbr2. TRAP1 has been shown to bind to TGFβ receptors and play a role in TGFβ signaling pathway. TRAP1 can interact with Smad4 and affect the SMAD-mediated signal transduction pathway. Mutant TRAP1 can prevent the formation of the Smad2-Smad4 complex to inhibit the TGFβ signaling pathway [61]. Thus in the boxplot for TRAP1—Tgfbr2 (Figure 3.2C), the densest quartile for tumor samples shows a high negative correlation.

Apoptosis pathway

Cancer cells have the ability to evade programmed cell death or apoptosis. TNFα, FASL, TRAIL and other genes can initiate apoptosis by binding to their receptors such as TNFR1, FAS, and TRAIL-R. A lot of apoptosis signals go through mitochondria. Mitochondria can help transduce the apoptosis signals by releasing cytochrome C (Cytc) which is a potent catalyst of apoptosis. There are two different Bcl-2 family members: proapoptotic members (Bid, BAD) and antiapoptotic members (Bcl-2, Bcl-xl), which activate and inhibit, respectively, the release of Cytc. Finally, two key caspases (Casp8 and Casp9) activate other downstream caspases that perform the cascading events of cell death (Figure 3.3A) [1].

In our result, we detect 33 relations with different distributions in the apoptosis pathway and some are supported by existing evidences. Examples include (Figure 3.3B):

(1) PUMA—Bcl-XL (BCL2L1). PUMA can interact with Bcl-XL and meanwhile PUMA can also neutralize and antagonize all the Bcl-2-like proteins [62]. From

the boxplot for PUMA—Bcl-XL, we can find that Bcl-XL and PUMA show a higher negative correlation in normal samples than in tumor samples (Figure 3.4A).

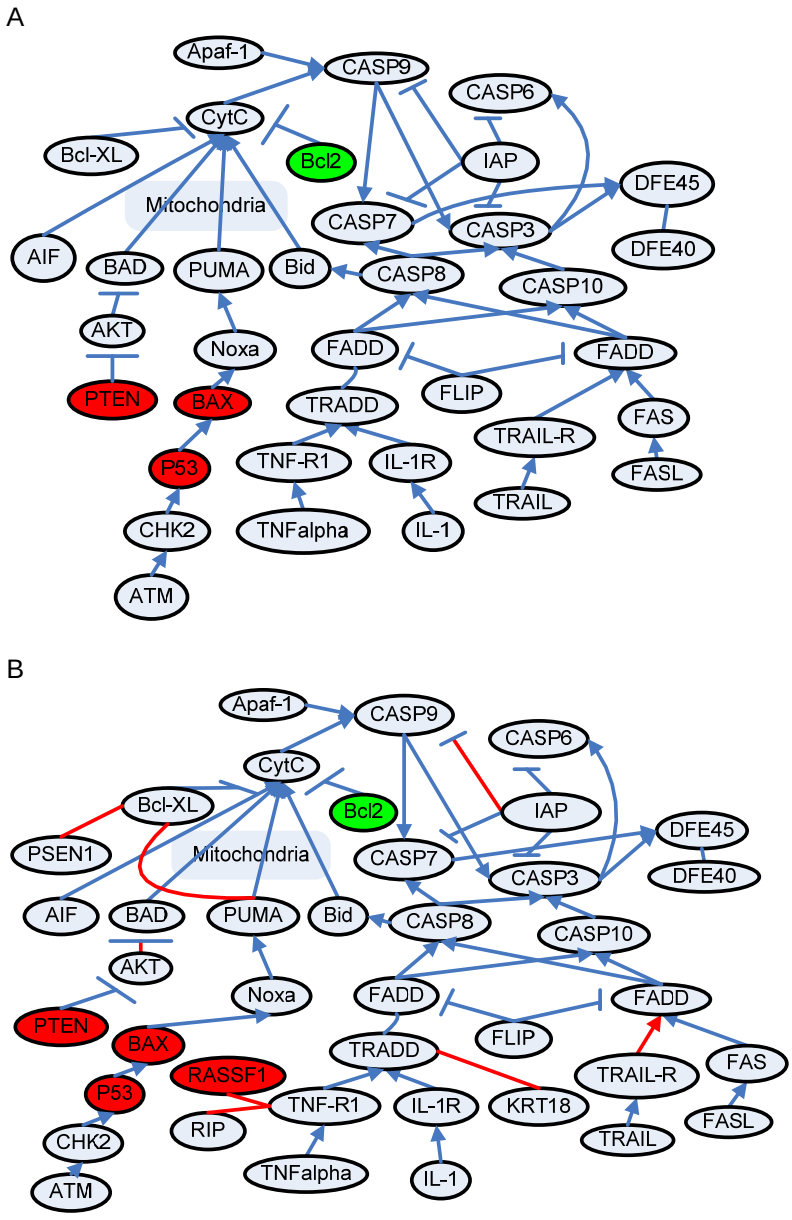

Figure 3.3 Apoptosis pathway. (**A**) Apoptosis pathway. (**B**) Differentially changed gene relations in apoptosis pathway. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes and green nodes represent oncogenes.

(2) AKT1—BAD. Active forms of Akt can phosphorylate BAD *in vivo* and *in vitro* to prevent it from promoting cell death [63]. In the boxplot for AKT1--BAD, the first quartile which is the densest for normal samples shows a higher positive correlation than the second quartile (densest) of tumor samples (Figure 3.4B). So we speculate that Akt can suppress BAD's activity in tumor samples.

(3) KRT18—TRADD. TRADD is a KRT18-interacting protein. KRT18 may inactivate TRADD to prevent interactions between TRADD and the activated TNFR1 and then affect TNFα-induced apoptosis [64]. So in the boxplot for KRT18—TRADD, normal samples show a higher positive correlation (Figure 3.4C).

(4) TNFR1—RIPK1 (RIP). The interaction between the death domain of TNFα receptor-1 (TNFR1) and TRADD can trigger distinct signaling pathways leading to apoptosis. TRADD also interacts strongly with another death domain protein, RIP and RIP plays an important role in the TNF signaling cascades leading to apoptosis [65]. In the boxplot for TNFR1—RIPK1, TNFR1 and RIPK1 show a preference for high positive correlation in normal samples (Figure 3.4D).

(5) TNFR1—RASSF1. RASSF1A is a tumor suppressor gene and the apoptosis initiation by TNFα or TRAIL recruit RASSF1A and MAP-1 to form complexes. RASSF1A and MAP-1 are the key links between death receptors and the apoptotic machinery [66]. We can verify this by the Boxplot for TNFR1—RASSF1. In most normal samples, they show a high positive correlation. In most tumor samples, they show a zero or negative correlation

(Figure 3.4E).

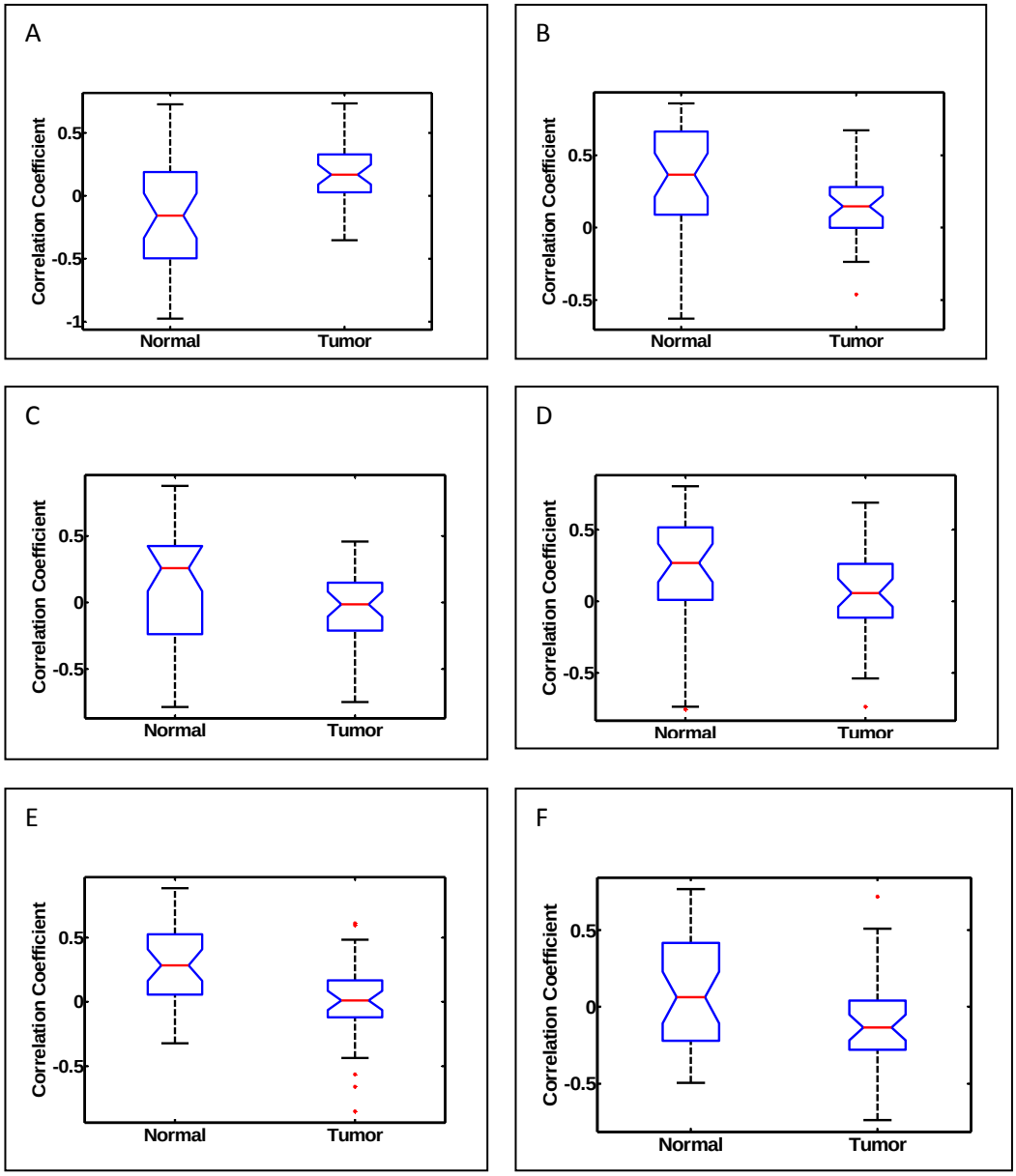

Figure 3.4 Boxplots for differential gene relations in apoptosis pathway. (**A**) Boxplot for PUMA—Bcl-XL(BCL2L1). $\Pr(\theta < 0.05) = 0.998$. (**B**) Boxplot for AKT1—BAD. $\Pr(\theta < 0.05) = 0.859$. (**C**) Boxplot for KRT18—TRADD. $\Pr(\theta < 0.05) = 0.991$.(**D**) Boxplot for TNFR1—RIPK1(RIP). $\Pr(\theta < 0.05) = 0.831$. (**E**) Boxplot for TNFR1—RASSF1. $\Pr(\theta < 0.05) = 0.946$. (**F**) Boxplot for IAP—CASP9. $\Pr(\theta < 0.05) = 0.826$.

(6) IAP—CASP9. Inhibitor of apoptosis (IAP) suppresses the activities of caspases and inhibits different apoptotic pathways [67]. Therefore IAP and CASP9 show a high negative correlation in tumor samples (Figure 3.4F).

Among the eight differential gene relations in Figure 3.3B, three of them are in the seed pathway: TRAIL-R→FADD, IAP→CASP9 and AKT→BAD, which demonstrates the effectiveness of the proposed method.

Growth signaling pathway

Cancer cells have the ability to produce their own growth promoting signals. EGF, TGFα and PDGF are activated and then bind to their receptors to transduce the growth signals. The activated growth factor receptors can then activate the SOS-Ras_Raf_Mapk cascade [1]. In the growth signaling pathway (Figure 3.5), Ras, JUN and Fos are oncogenes.

We find 68 relations with different distributions in growth signaling pathway and we discuss three relations here:

(1) RASSF2—KRAS. Although different forms of Ras are frequently thought as oncogenes, they also have the ability to incite antigrowth effects such as cell cycle arrest, differentiation, and apoptosis. RASSF2 can bind directly to K-Ras. Moreover, RASSF2 can inhibit the growth of tumor cells and the activated K-Ras can enhance this ability [68]. This is why RASSF2 and RAS show a preference for a high positive correlation in normal samples in the boxplot (Figure 3.6A).

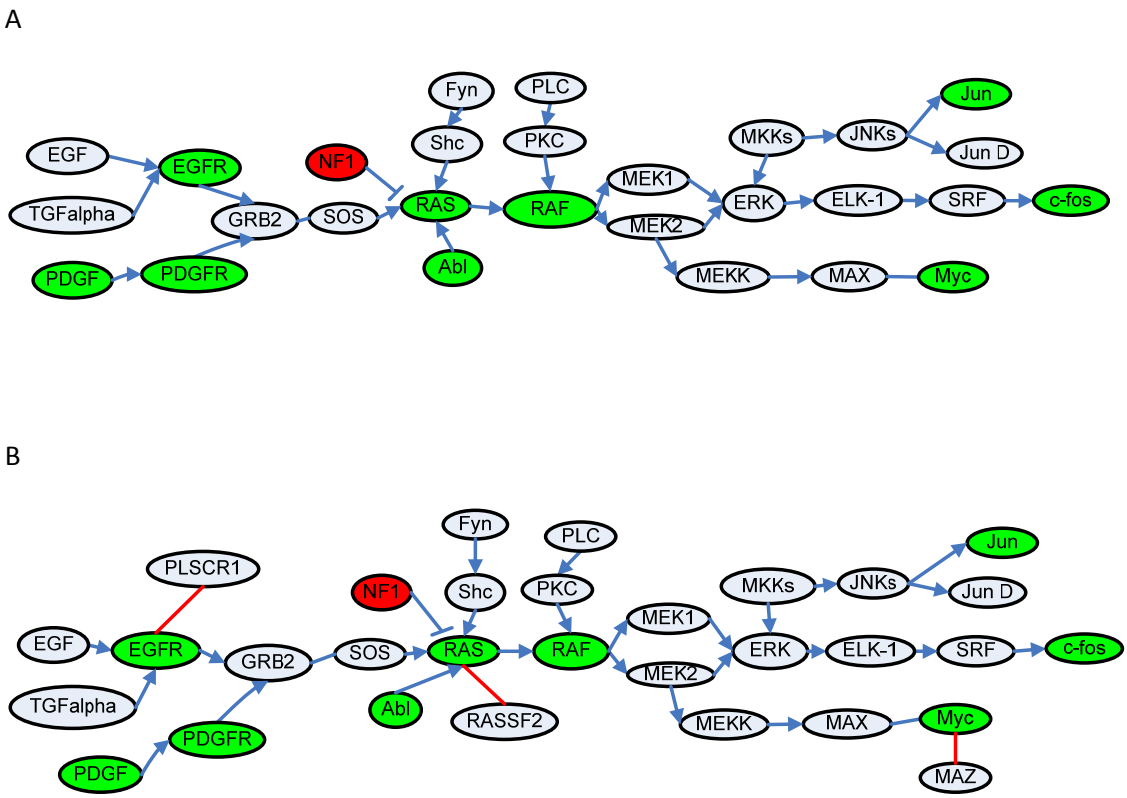

Figure 3.5 Growth signaling pathway. (**A**) Growth signaling pathway. (**B**) Differentially changed relations in growth signaling pathway. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes and green nodes represent oncogenes.

(2) MAZ—MYC. MAZ family can increase the oncogene MYC's transcriptional activity [69]. As expected, MAZ and MYC demonstrate a higher positive correlation in tumor samples (Figure 3.6B).

(3) PLSCR1—EGFR. Activated epidermal growth factor receptors (EGFR) can both physically and functionally interact with PLSCR1. PLSCR1 can interact with Shc and then accelerate the activation of Src kinase through the EGF receptor while Src can initiate some activating pathway for the oncogene JUN [70]. Thus in the

boxplot for PLSCR1—EGFR, the densest quartile for normal samples shows a low negative correlation while the densest quartile for tumor samples shows a low positive correlation (Figure 3.6C).

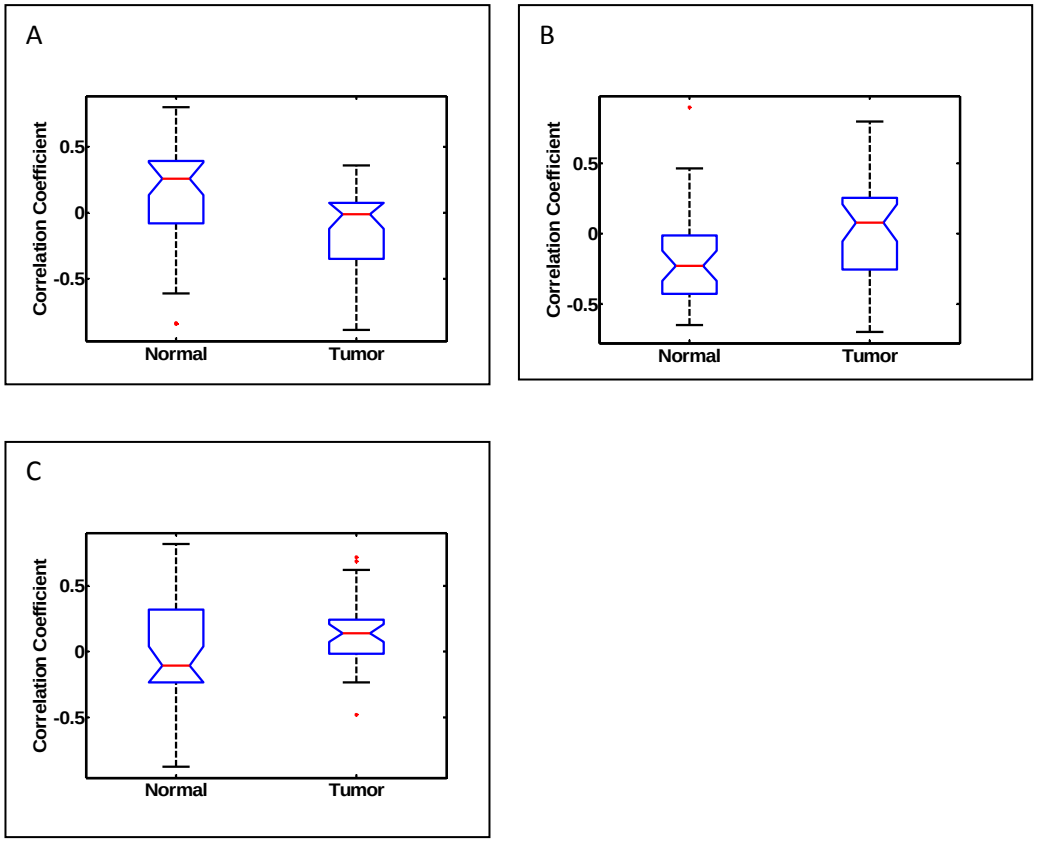

Figure 3.6 Boxplots for differential gene relations in growth signaling pathway**.** (**A**) Boxplot for RASSF2—KRAS. $\Pr(\theta < 0.05) = 0.983$. (**B**) Boxplot for MAZ—MYC. $\Pr(\theta < 0.05) = 0.833$. (**C**) Boxplot for PLSCR1—EGFR. $\Pr(\theta < 0.05) = 0.963$

Chapter 4 Markov Blanket Based Method for Detecting Gene-Gene Interactions

Some common complex diseases such as various types of cancers, cardiovascular disease, and diabetes are influenced by multiple genetic variants [5]. Therefore, detecting high-order epistasis (gene-gene interaction), which refers to the interactive effect of two or more genetic variants on complex human diseases, can help to unravel how genetic risk factors confer susceptibility to complex diseases [4]. However, the very large number of SNPs checked in a typical GWAS (more than 10 million) and the enormous number of possible SNP combinations make detecting high-order gene-gene interactions from GWAS data computationally challenging [71-72]. Moreover, how to measure the association between a set of SNPs and the phenotype presents another grand statistical challenge.

Some statistical and machine learning methods for gene-gene interaction detection are introduced in Section 2.4. These statistical and machine learning methods can also be grouped into two categories: prediction/classification-based methods and association-based methods. Prediction/classification-based methods try to find the best set of SNPs which can generate the highest prediction/classification accuracy including, for example, multifactor dimensionality reduction (MDR) [42, 73-75], penalized logistic regression (stepPLR [43], lassoPLR [76]), Support Vector Machines (SVMs) [46], and random forest [47]. Some prediction/classification-based methods can only be applied to small-scale analysis (i.e., a small set of SNPs) due to their computational complexity. Moreover, almost all prediction/classification-based

methods tend to introduce many false positives, which may result in a huge cost for further biological validation experiments. BEAM is a scalable and association-based method [44]. One drawback of BEAM is that identifying both single disease SNP and SNP combinations simultaneously makes BEAM over-complex and weakens its power.

To address the challenges in gene-gene interaction detection and overcome the drawbacks of existing methods, I propose a novel Markov Blanket based method, DASSO-MB (Detection of ASSOciations using Markov Blanket), to detect gene-gene interactions in case-control studies [77]. The Markov Blanket is a minimal set of variables, which can completely shield the target variable from all other variables based on Markov condition property. Thus, DASSO-MB can detect the SNP set that shows a strong association with diseases with the fewest false positives. Furthermore, the heuristic search strategy in DASSO-MB can avoid the time-consuming training process as in SVMs and Random Forests.

4.1 Markov Blanket

Bayesian Networks are probabilistic graphical models representing a joint probability distribution *J* over a set of random variables $\{X_1, X_2, ..., X_n\}$ by a directed acyclic graph (DAG) *G* and encode the Markov condition property: each node is conditionally independent of its non-descendents given its parents [78]. In this case, the joint probability distribution *J* can be represented as

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \quad (4\text{-}1)$$

where $Pa(X_i)$ denotes the set of parents of X_i in *G*.

For three random variables *X*, *Y* and *Z*, if the probability distribution of *X* conditioned on both *Y* and *Z* is equal to the probability distribution of *X* conditioned only on *Y*, i.e., $P(X \mid Y,Z) = P(X \mid Y)$, *X* is conditionally independent of *Z* given *Y*. This conditional independence is represented as $(X \perp Z \mid Y)$. Similarly, $(X \not\perp Z \mid Y)$ represents conditional dependence.

Definition 1 (Faithfulness) *A Bayesian Network N and a joint probability distribution J are faithful to each other if and only if every conditional independence entailed by the DAG of N and the Markov Condition is also present in J* [79].

Theorem 1. *If a Bayesian Network N is faithful to a joint probability distribution J, then: (1) nodes X and Y are adjacent in N if and only if X and Y are conditionally dependent given any other set of nodes, (2) for the triplet of nodes X, Y , and Z in N, X and Z are adjacent to Y , but Z is not adjacent to X, X→Y←Z is a subgraph of N if and only if X and Z are dependent conditioned on every other set of nodes that contains Y .*

We can define the Markov Blanket of a target variable of *T*, *MB(T)*, as a minimal set for which $(X \perp T \mid MB(T))$, for all $X \in V - \{T\} - MB(T)$ where *V* is the variable set in Bayesian Network *N*. The Markov Blanket of a variable *T* is a minimal set of variables which can completely shield variable *T* from all other variables. All other

variables are probabilistically independent of the variable *T* conditioned on the Markov Blanket of variable *T*.

Theorem 2. *If Bayesian Network N is faithful to its corresponding joint probability distribution J, then for every variable T, MB*(*T*) *is unique and is the set of parents, children, and spouses of T.*

We show an example of the Markov Blanket in Figure 4.1. The *MB*(*T*) of the variable *T* is the set of gray-filled nodes $\{B, L, M, D, X\}$ and variable *S* and *U* are independent of *T* conditioned on $\{B, L, M, D, X\}$

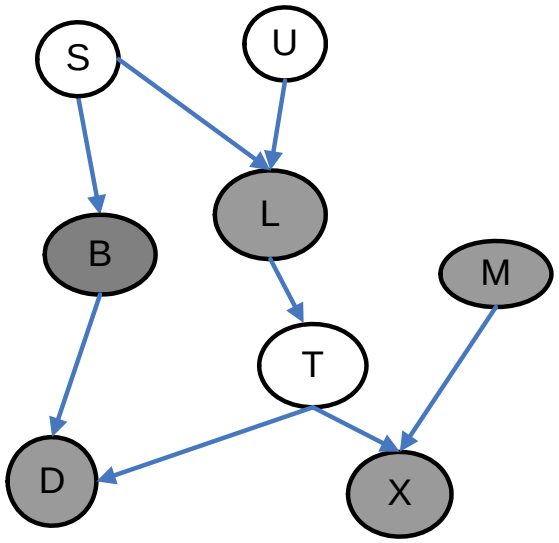

Figure 4.1 Markov Blanket in a Bayesian Network. The gray-filled nodes are the Markov Blanket of node T.

Given the definition of Markov Blanket, the probability distribution of *T* is completely determined by the values of variables in *MB*(*T*). Therefore, the detection of Markov Blanket has been applied for optimal variable selection problem [80]. In addition, the Markov Blanket can be used for causal discovery because *MB*(*T*) is the union of direct cause variables (parents), direct effect variables (children), and direct

cause variables (spouse) of direct effect variables of *T*. Thus the Markov Blanket learning method is suitable for detection of gene-gene interactions in genome-wide case-control studies, e.g., to identify a minimal set of SNPs which may cause the disease for further experiments.

4.2 G^2 Test

The G^2 test is commonly used to test independence and conditional independence between two variables for discrete data as an alternative to the χ^2 test because G^2-values are additive and can be applied to more complicated statistical designs [79, 81-82]. The null hypothesis for G^2 test is that the two variables are independent.

Assume that we have a contingency table to record and analyze the joint distribution of two variables. The count in a particular cell in a contingency table, x_{ij}, is the value of a random variable from *N* samples with a multinomial distribution. Let $x_{i\bullet}$ represent the sum of elements in all cells along the *i*th row, and $x_{\bullet j}$ denote the sum of the counts in all cells along the *j*th column. If these two variables are independent based on the null hypothesis, the expected value of the random variable x_{ij} is:

$$E(x_{ij}) = \frac{x_{i\bullet} x_{\bullet j}}{N} \tag{4-2}$$

We can compute the conditional independence from appropriate marginal distributions in a similar way. For instance, to determine whether the first variable is

independent of the second conditioned on the third, we can calculate the expected value of a cell x_{ijk} as

$$E(x_{ijk}) = \frac{x_{i\bullet k} x_{\bullet jk}}{x_{\bullet\bullet k}} \tag{4-3}$$

For *n* cells in a contingency table, assume that the observed numbers are denoted by $O_1, O_2, \ldots, O_n$ and the corresponding expected numbers by $E_1, E_2, \ldots, E_n$, then, the G^2 is given by

$$G^2 = 2\sum_i^n O_i \ln(\frac{O_i}{E_i}) \tag{4-4}$$

which has an asymptotical distribution as chi-square (χ^2) with appropriate degrees of freedom. The degrees of freedom (*df*) for the G^2 test between two variables A and B can be calculated as:

$$df = (Cat(A)-1)\times(Cat(B)-1) \tag{4-5}$$

and the degrees of freedom (*df*) for the G^2 test between A and B conditional on the third variable C can be calculated as:

$$df = (Cat(A)-1)\times(Cat(B)-1)\times\prod_{i=1}^{n} Cat(C_i) \tag{4-6}$$

where *Cat(X)* is the number of categories of the variable *X* and *n* is the number of variables in C. Here in Eq. (4-5) and Eq. (4-6) we assume that there are no empty cells in the contingency table. If there are some empty cells in the contingency table, we

should reduce the degrees of freedom from Eq. (4-5) or Eq. (4-6) by the number of empty cells.

As described in section 4.4, the proposed DASSO-MB uses G^2 to test the association and independence between SNPs and disease status.

4.3 Markov Blankets Learning Methods

There are several Markov Blanket learning methods such as: Koller-Sahami (KS) algorithm [83], Grow-Shrink (GS) algorithm [84], Incremental association Markov Blanket (IAMB) algorithm [85], Max-Min Markov Blanket (MMMB) algorithm [86], HITON_MB [80], and PCMB [87].

Koller-Sahami (KS) algorithm is the first algorithm to employ Markov Blanket for feature selection. However, there is no theoretical guarantee for Koller-Sahami (KS) algorithm to find optimal MB set [83]. The GS algorithm [84] and IAMB methods [85] are two similar algorithms with two search procedures: forward phase and backward phase. In the forward phase, the nodes of MB(T) are admitted into MB, while in the backward phase false positives are removed from MB. Under the assumptions of faithfulness and correct independence test, both the GS algorithm and IAMB are proved correct [85]. Comparing to GS algorithm, IAMB might achieve a better performance with fewer false positives admitted during the forward phase. A common limitation for GS algorithm and IAMB is that both methods require a very large number of samples to perform well. IAMB can be revised in two ways: (1) after each admission step in forward phase, perform a backward conditioning phase to

remove false positives to keep the size of MB(T) as small as possible, and (2) substitute the backward conditioning phase with the PC algorithm instead [79]. In other words, the backward phase will perform the independence test conditioned on all subsets of the current Markov Blanket. Tsamardinos *et al*. proposed three IAMB variants: interIAMB, IAMBnPC, and interIAMBnPC [85]. They also proved the correctness of interIAMBnPC. The time complexity of IAMB is O(|MB|×N) where |MB| is the size of MB and N is number of variables.

To overcome the data inefficient problem of IAMB and its variants, Max-Min Markov Blanket (MMMB) algorithm [86], HITON_MB [80], and PCMB [87] are proposed. All these three algorithms take a divide-and-conquer method that breaks down the problem of identifying Markov Blanket of variable T into two subproblems: First, identifying parents and children of T (PC(T)) and, second, identifying the spouses of T. Meanwhile, they have the same two assumptions as IAMB (i.e. faithfulness and correct independence test) and take into account the graph topology to improve data efficiency. However, results from MMPC/MB and HITON-PC/MB are not always correct since some descendants of T other than its children will enter PC(T) during the first step of identifying parents and children of T [87]. PCMB can be proved correct in [87]. In every loop, PCMB first remove unrelated variables, then PCMB use IAMBnPC method to admit one feature and remove false positives. The problem of PCMB is that the PC algorithm performs an exhaustive conditional independence test, which is very time consuming. The reason that PC algorithm was used in PCMB and interIAMBnPC is that PC algorithm is a more sample-efficient

method and is sound under the assumption of faithfulness [85]. In fact if the size of Markov Blanket is large, PC algorithm still needs a lot of samples to guarantee its performance. There is no theoretical proof and guarantee that the PC algorithm admits less false positives than other methods.

4.4 DASSO-MB

Detecting gene-gene interactions is a special application of Markov Blanket learning method because we only need to detect the parents of the target variable T and don't need to design a complex algorithm to detect spouses of T. Here target variable T is the disease status labels and the parents of T are those disease SNPs. The Barkov Blanket of T, MB(T), only contains the parents of T.

All Markov Blanket learning methods are based on the following two Theorems.

Theorem 3. *If a variable belongs to MB(T) which only contains the parents of T, then it will be dependent on T given any subset of the variable set* $\mathrm{V}-\{\mathrm{T}\}$.

Proof: This is a direct consequence of **Theorem 1** because now MB(T) only contains the parents of T. □

Theorem 4. *If a variable is not a member of MB(T), then conditioned on MB(T), or any superset of MB(T), it will be independent of T.*

Proof: Let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ represent four mutually disjoint variable sets. Any probability distribution p satisfies the weak union property: $X \perp (Y \cup W) \mid Z \Rightarrow X \perp Y(Z \cup W)$ [88]. Based on the definition of Markov Blanket,

we get that $X \notin MB(T) \Rightarrow (X \perp T \mid MB(T))$. Thus, by the weak union property, we have $(X \perp T \mid (MB(T) \bigcup S)$ for any subset $S \subseteq V - \{T\} - \{X\} - MB(T)$. □

We use a Markov Blanket based algorithm, DASSO-MB, to detect gene-gene interactions (Algorithm 4.1). Let T denote the disease status and V the set of all variables containing T and all SNPs. There are two types of phases in DASSO-MB: forward phase and backward phase. In each loop of the forward phase, if one variable shows a maximal G^2 score conditioned on $MB(T)$ and is dependent on target variable T, it will be admitted into $MB(T)$. This admission operation is followed by a backward phase to remove false positives by conducting conditional independence tests. If no more variable will be added into $MB(T)$ in the forward phase, we will enter the final backward phase to remove variables that do not belong to $MB(T)$. Comparing to IAMB, DASSO-MB adds a backward phase after each step of selecting a variable in the forward phase to remove false positives, make the size of $MB(T)$ as small as possible and therefore improve the sample-efficiency. In addition, it uses subset S of $MB(T)$ rather than the remaining set $MB(T) - \{Y\}$ while conducting the conditional independence tests in the backward phase. Here we let the size of subset S of $MB(T)$ be larger than zero and exclude the empty set because of the joint effect of set of SNPs on the disease status. These two changes can make the detected results more reliable.

Algorithm 4.1 DASSO-MB

/*Initialization*/
V : set of all variables; T: Target variables;
MB(T)= ϕ;
/*DASSO-MB algorithm*/
Begin procedure
  Forward-MB;
  Backward-MB;
End procedure

/* Forward phase */
Begin Forward-MB
Repeat
  For all $x_i \in V - MB(T) - \{T\}$;
    $g(x_i) = G^2(x_i : T \mid MB(T))$;
    $X = \arg\max(g(x_i))$;
      If $(X \perp\!\!\!\perp T \mid MB(T))$
        $MB(T) = MB(T) \cup \{X\}$;
      End If
  End For
  Backward-MB;
  Until MB(T) has not changed;
End

/*Backward phase*/
Begin *Backward-MB*
  For all $Y \in MB(T)$
    If $\exists S \subseteq (MB(T) - \{Y\})$
    s.t. $(Y \perp T \mid S)$ and $size(S) > 0$
      $MB(T) = MB(T) - \{Y\}$;
    End If
  End For
End

4.5 Experimental Results

4.5.1 Epistatic Models

We evaluate the proposed DASSO-MB on simulated data sets, which are generated from three commonly-used disease models developed elsewhere [41, 44]. We show the three disease models in Table 4.1.

Table 4.1 Three two-locus epistatic models

Model 1	AA	Aa	aa
BB	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
Bb	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$
bb	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	$\alpha(1+\theta)^4$
Model 2	AA	Aa	aa
BB	α	α	α
Bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
bb	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$
Model 3	AA	Aa	aa
BB	α	α	α
Bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

Table 4.1 lists the disease odds for these three epistatic models, where α is the baseline effect and θ is the genotypic effect. Assume an individual has genotype

g_A at locus A and genotype g_B at locus B in a two-locus epistatic model, then the disease odds are defined as

$$p(D \mid g_A, g_B) / p(\overline{D} \mid g_A, g_B) \tag{4-7}$$

where $p(D \mid g_A, g_B)$ is the probability that an individual has the disease given genotype (g_A, g_B) and $p(\overline{D} \mid g_A, g_B)$ is the probability that an individual does not have the disease given genotype (g_A, g_B).

In Model1 the odds of disease increase in a multiplicative mode both within and between two loci. For example, an individual with *Aa* at locus A has larger odds which are $1+\theta$ times relative to those of an individual who is homozygous *AA*; the *aa* homozygote has further increased disease odds by $(1+\theta)^2$. We can also find similar effects on locus B. Finally the odds of disease for each combination of genotypes at loci A and B can be obtained by the product of the two within-locus effects. Model2 demonstrates two-locus interaction multiplicative effects because at least one disease-associated allele must be present at each locus to increase the odds beyond the baseline level. Moreover the increment of the disease-associated allele at loci A or B can further increase the disease odds by the multiplicative factor $1+\theta$. Model3 specifies two-locus interaction threshold effects. Like Model2, Model3 also requires at least one copy of the disease-associated alleles at both loci A and B. However the increment of the disease-associated allele does not increase the risk further. We call this as disease threshold effect. It means a single copy of the disease-associated allele at each locus is required to increase odds of disease and this

is the disease threshold. But after the disease threshold has already been met, having both copies of the disease-associated allele at either locus has no additional influence on disease odds.

To generate data, we need to determine three parameters associated with each model: the marginal effect of each disease locus (λ), the minor allele frequencies (MAF) of both disease loci, and the strength of linkage disequilibrium (LD) between the unobserved disease locus and a genotyped locus. LD is a non-random association of alleles at different loci and is quantified by the squared correlation coefficient r^2 calculated from allele frequencies [89]. The prevalence of a disease is the proportion the total number of cases of the disease in the population and in this paper we assume that the disease prevalence is 0.1 for all these three disease models [41]. The marginal effect of each disease locus (λ) can be determined by the baseline effect α and the genotypic effect θ in Table 4.1 and the minor allele frequencies (MAF) of both disease loci. So first we fix λ, the disease prevalence and MAF of both disease loci. Then we numerically derive the model parameters θ and α. Based on θ and α, we calculate the conditional probability of each genotype combination given disease status which is necessary for generating data [90]. We set parameters for each model as follows:

- Model1: λ=0.3; r^2=0.7, 1.0; MAF=0.05, 0.1, 0.2, 0.5.
- Model2: λ=0.3; r^2=0.7, 1.0; MAF=0.05, 0.1, 0.2, 0.5.
- Model3: λ=0.6; r^2=0.7, 1.0; MAF=0.05, 0.1, 0.2, 0.5.

For each non-disease marker, we randomly chose its MAF from a uniform distribution in [0.0. 0.5]. We generate 50 datasets and each dataset contains 100 markers genotyped for 1,000 cases and 1,000 controls based on each parameter setting for each model.

4.5.2 Simulation Analysis

We compare the DASSO-MB algorithm with four commonly used methods: BEAM, SVM, MDR, and stepPLR on the three simulated disease models. We use power as our evaluation criterion, which is defined as the proportion of simulated datasets in which all diseases associated markers are identified without any false positives, to measure the performance of each method.

BEAM uses a Bayesian marker partition model to partition SNPs into three groups: group 0 contains markers unlinked to the disease, group 1 contains markers contributing independently to the disease, and group 2 contains markers that jointly influence the disease. After the partition step by MCMC, candidate SNPs or groups of SNPs are further filtered by the B statistic [44]. The BEAM software is downloaded from http://www.fas.harv-ard.edu/~junliu/BEAM. We set the p-value threshold of the B statistic as 0.1.

For SVM, we use LIBSVM with a RBF kernel to detect gene-gene interactions [91]. A grid search is used for selecting optimal parameters. Instead of using the exhaustive greedy search strategy for SNPs as in [46], which is very time-consuming and infeasible to large-scale datasets, we turn to a search strategy used in [47]. First

we rank SNPs based on the mutual information between SNPs and disease status label that is 0 for the control and 1 for the case. Then, we use a sliding window sequential forward feature selection (SWSFS) algorithm in [47] based on SNPs rank. The window size in SWSFS algorithm determines how robust the algorithm could be and we set it to 20.

Since MDR algorithm can not be applied to a large dataset directly, we first select top 10 SNPs by ReliefF [92], a commonly-used feature selection algorithm, and then MDR performs an exhaustive search for a model consisting of no more than four SNPs that can maximize cross-validation consistency and prediction accuracy. When one model has the maximal cross-validation consistency and another model has the maximal prediction accuracy, MDR follows statistical parsimony (selects the model with fewer SNPs).

For stepPLR, we download the R package from CRAN (ftp://200.17.202.1/CRAN/ web/packages/stepPlr). StepPLR provides both stepwise forward and backward methods for feature selection procedure. We use both methods and set the regularization parameter λ to default value (10^{-4}) for the L2 norm of the coefficients.

The results on the simulated data are shown in Figure 4.2. As can be seen, among the five methods, the DASSO-MB algorithm performs the best. BEAM is the second best. Interestingly, BEAM prefers to assign the two disease-associated markers to group 1, which means that BEAM considers that the two disease SNPs affect the disease independently. In most cases, the powers of both MDR and SVM are much

smaller than those of the DASSO-MB and BEAM algorithms. For the MDR algorithm, the poor performance may be due to the use of ReliefF to reduce SNPs from a very large dimensionality.

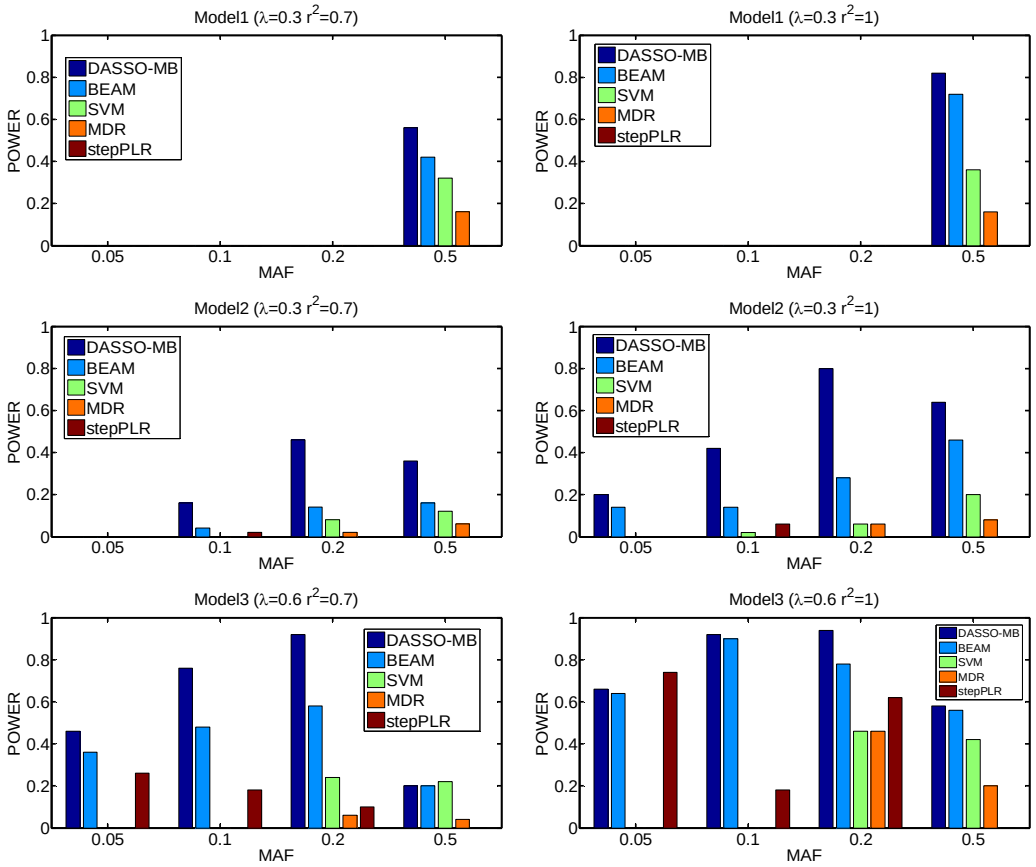

Figure 4.2 Performance comparison of DASSO-MB, BEAM, SVM, MDR, and stepPLR. The power is defined as the proportion of simulated datasets whose result only contains all disease associated markers without any false positives.

In some other studies, the definition of power is not in a strict sense. For example, in [44, 47], the power is defined as the proportion of 50 data sets in which all associated markers are identified at a significance threshold of 0.1 after Bonferroni correction. In other words, false positives are allowed in the final SNP sets. Accordingly, we also evaluate the methods in terms of the power defined as the

proportion of simulated datasets in which two diseases associated markers are identified with no more than two false positives. The results of those three models are shown in Table 4.2. In parentheses we list the average number of false positives. From Table 4.2, we can see that the DASSO-MB again outperforms other algorithms. Furthermore, the DASSO-MB algorithm finds SNP sets with fewer false positives.

Table 4.2 Comparison of performance of DASSO-MB, BEAM and SVM. We show the number of datasets in which two disease-associated markers can be identified with no more than two false positives. The average number of false positives is in the parentheses.

Model 1 (r^2 =0.7)	**MAF**			
	0.05	0.1	0.2	0.5
DASSO-MB	0(0)	0(0)	0(0)	32(0.16)
BEAM	0(0)	0(0)	0(0)	22(0.05)
SVM	1(3)	1(3)	0(0)	33(0.7)
Model 1 (r^2 =1)	**MAF**			
	0.05	0.1	0.2	0.5
DASSO-MB	0(0)	0(0)	0(0)	46(0.11)
BEAM	0(0)	0(0)	0(0)	36(0.07)
SVM	0(0)	0(0)	1(2)	43(0.76)
Model 2 (r^2 =0.7)	**MAF**			
	0.05	0.1	0.2	0.5
DASSO-MB	0(0)	8(0)	26(0.12)	18(0)
BEAM	0(0)	2(0)	10(0.3)	9(0.11)
SVM	0(0)	2(1.5)	14(0.93)	21(0.8)
Model 2 (r^2 =1)	**MAF**			
	0.05	0.1	0.2	0.5
DASSO-MB	10(0)	22(0.05)	42(0.05)	33(0.03)
BEAM	8(0.13)	7(0)	17(0.24)	27(0.11)
SVM	1(2)	3(0.67)	22(1.18)	33(0.94)
Model 3 (r^2 =0.7)	**MAF**			
	0.05	0.1	0.2	0.5
DASSO-MB	24(0.04)	44(0.14)	47(0.02)	11(0.09)
BEAM	21(0.14)	24(0)	32(0.09)	11(0.09)
SVM	1(1)	6(1.83)	29(0.83)	29(0.83)
Model 3 (r^2 =1)	**MAF**			
	0.05	0.1	0.2	0.5
DASSO-MB	34(0.03)	50(0.08)	49(0.04)	31(0.06)
BEAM	33(0.03)	47(0.04)	43(0.09)	31(0.1)
SVM	5(1.6)	23(1.52)	42(0.64)	38(0.55)

Compared to the strict definition of power, a difference we can see is that for MAF > 10%, SVM actually detects the two disease associated markers in more datasets than BEAM, however, at the cost of introducing more false positives.

Chapter 5 Detecting Gene-Gene Interactions using Bayesian Networks with a New Scoring Function

Chapter 4 demonstrates that the Markov Blanket based method, DASSO-MB, outperforms other commonly used statistical and machine learning methods in gene-gene interaction detection. In this Chapter, I propose a Bayesian Network structure learning method, EpiBN (Epistatic interaction detection using Bayesian Network model), to detect gene-gene interactions. Comparing to Markov Blanket based methods, the merits of applying Bayesian Network structure learning method to gene-gene interaction detection include: (1) the new scoring function for Bayesian Network structure learning in EpiBN can reflect higher-order interactions and detect the true number of disease SNPs, and are not sample-consuming; and (2) heuristic Bayesian Network structure learning method can solve the classical XOR problem, which may hinder the applications of Markov Blanket based approaches.

5.1 Bayesian Networks

A Bayesian Network is a directed acyclic graph (DAG) *G* consisting of nodes corresponding to a random variable set $\{X_1, X_2, ..., X_n\}$ and edges between nodes, which determine the structure of *G* and therefore the joint probability distribution of the whole network [78, 93]. The DAG *G* encodes the Markov condition property: each variable is conditionally independent of its nondescendants, given its parents in *G*. By applying the Markov condition property, the joint probability distribution *J* can be represented as

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \quad (5\text{-}1)$$

where $Pa(X_i)$ denotes the set of parents of X_i in G. Therefore, there are two components in a Bayesian Network. The first component is the DAG G reflecting the structure of the network. The second component, θ, describes the conditional probability distribution $P(X_i \mid Pa(X_i))$ to specify the unique distribution J on G.

Bayesian Networks provide models of causal influence and allow us to explore causal relationships, perform explanatory analysis, and make predictions. Genome-wide association studies attempt to identify the gene-gene interaction among a set of SNPs, which are associated with one certain type of disease. Therefore, we can use Bayesian Networks to represent the relationship between genetic variants and a phenotype (disease status), as shown in Figure 5.1. The n SNP nodes and the disease status/label node form a two-layer Bayesian Network and we want to determine which SNP nodes are the parent nodes of the disease status node.

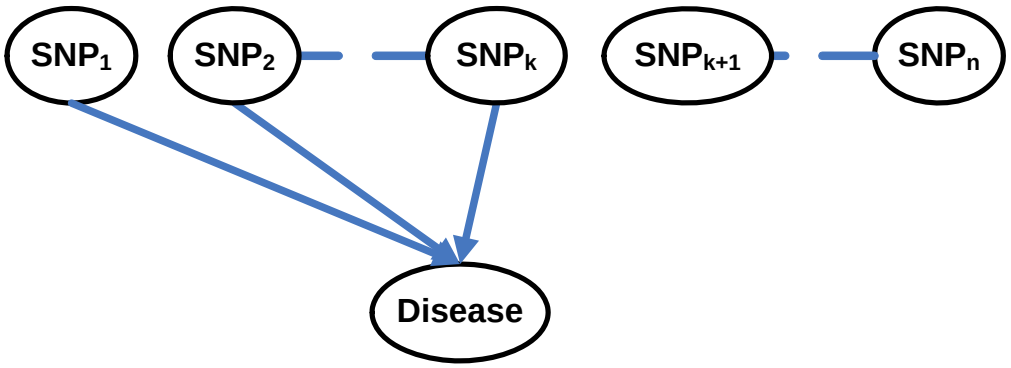

Figure 5.1 A Bayesian Network for detecting gene-gene interactions in genome-wide association studies. Genome-wide association studies attempt to identify the k-way gene-gene interaction among SNPs: SNP_1, SNP_2,…,SNP_k, which are associated with disease.

By modeling the association between SNPs and the disease status based on Bayesian Networks, we transform detecting gene-gene interactions into structure learning of Bayesian Networks from GWAS data. There are two types of structure learning methods for Bayesian Networks: constraint-based methods and score-and-search methods. The constraint-based methods first build the skeleton of the network (undirected graph) by a set of dependence and independence relationships. Next they direct links in the undirected graph to construct a directed graph with d-separation properties corresponding to the dependence and independence determined [79, 94-95]. Even though constraint-based methods are developed with a rigorous theoretical foundation, errors in conditional dependence and independence will affect the stability of constraint-based methods, and this problem is especially serious when the number of samples is small. The score-and-search methods view a Bayesian Network as a statistical model and transform the structure learning of Bayesian Networks into a model selection problem [96]. To select the best model, a scoring function is needed to indicate the fitness between a network and the data. Then the learning task is to find the network with the highest score. Thus, score-and-search methods typically consist of two components, (1) a scoring function, and (2) a search procedure. In this disseartation, I focus on structure learning approaches for Bayesian Networks based on score-and-search methods because score-and-search methods are more robust for small data sets than constraint-based methods.

5.2 A New BN Scoring Function

One of the most important issues in score-and-search methods is the selection of scoring function. A natural choice of scoring function is the likelihood function. However, the maximum likelihood score often overfits the data because it does not reflect the model complexity. Therefore, a good scoring function for Bayesian Networks' structure learning must have the capability of balancing between the fitness and the complexity of a selected structure. There are several existing scoring functions based on a variety of principles, such as the information theory and minimum description length (e.g. BIC score, AIC score, and MDL score) [97-99] and Bayesian approach (BDe score) [100].

Suppose that a dataset D includes n variables $\{X_1, X_2, ..., X_n\}$ and N samples, we can write a general information-based scoring function as:

$$\log P(D \mid S) = \log P(D \mid \hat{\theta}_S, S) - C(S) f(N) \tag{5-2}$$

$$C(S) = \sum_{i=1}^{n} q_i (r_i - 1) \tag{5-3}$$

where $\hat{\theta}_S$ is an estimate of parameters from the maximum likelihood method for the structure S, q_i is the number of configurations of the parent set $Pa(X_i)$ of X_i, r_i is the number of states of X_i, $C(S)$ represents the structure complexity, and $f(N)$ is a penalization function. The first term of this score scheme measures the fitness between the structure and data, and the second term reflects structure complexity. With the maximum likelihood method [96], we can get

$$\log(P(D \mid \hat{\theta}_S, S)) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log(N_{ijk} / N_{ij}) \quad (5\text{-}4)$$

where N_{ijk} is the number of instances where X_i takes its k-th value and the set of variables $Pa(X_i)$ takes its j-th configuration; N_{ij} is the number of instances where the set of variables $Pa(X_i)$ takes its j-th configuration. Obviously, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

If we set $f(N)=1$, we get the AIC score as

$$\log P(D \mid S) = \log P(D \mid \hat{\theta}_S, S) - C(S) \quad (5\text{-}5)$$

If we set $f(N)=1/2\log(N)$, then the BIC score is

$$\log P(D \mid S) = \log P(D \mid \hat{\theta}_S, S) - 1/2C(S)\log(N) \quad (5\text{-}6)$$

The BIC score and AIC score are derived from some approximations when the number of samples N approaches infinity [101]. If the number of samples is small, the approximation in the inference of both AIC score and BIC score can not hold any more and the structure penalty term in Eq. (5-5) and Eq. (5-6) are not suitable. When using information-based scores in the Bayesian Network model to detect gene-gene interactions, the BIC score is too strict and prefers to select simple structures, while the AIC score prefers to select complex structures [102].

We herein describe a new information-based scoring function to detect gene-gene interactions by Bayesian Network model. In the Bayesian Network for gene-gene interaction detection in Figure 5.1, we are only concerned with one target node, the disease status node, and we want to detect its parent SNP nodes. We represent the local structure around the disease status node as *LDS* (Local Disease Structure), which consists of the disease status node and edges from candidate

disease SNP nodes to the disease status node. Because of the decomposability property of information-based scoring function, the AIC score for *LDS* is:

$$\begin{aligned} &\log P(D \mid LDS) = \log P(D \mid \hat{\theta}_{LDS}, LDS) - C(LDS) \\ &= \sum_{j=1}^{q} \sum_{k=1}^{r} N_{jk} \log(N_{jk} / N_j) - q(r-1) \end{aligned} \tag{5-7}$$

where $C(LDS)$ is the complexity of the local disease structure, q is the number of configurations of parent SNP nodes, r is the number of states of the disease status node, N_{jk} is the number of instances where the disease status node takes its k-th value and the parent SNP nodes take their j-th configuration, N_j is the number of instances where the parent SNP nodes take their j-th configuration, and $N_j = \sum_{k=1}^{r} N_{jk}$.

We start our search from an empty local disease structure LDS_0, and we can obtain the AIC score for LDS_0:

$$\begin{aligned} &\log P(D \mid LDS_0) = \log P(D \mid \hat{\theta}_{LDS_0}, LDS_0) - C(LDS_0) \\ &= \sum_{k=1}^{r} N_k \log(N_k / N) - (r-1) \end{aligned} \tag{5-8}$$

where N_k is the number of instances in which the disease status node takes its k-th value, and $N = \sum_{k=1}^{r} N_k$.

For further inference, we use X for the target disease status node and use $Pa(X)$ for its parent SNP nodes. Based on the concept of mutual information and Eq. (5-7) and Eq. (5-8), the mutual information between X and $Pa(X)$ can be expressed as [103]:

$$MI(X, Pa(X)) = \frac{\log P(D \mid \hat{\theta}_{LDS}, LDS) - \log P(D \mid \hat{\theta}_{LDS_0}, LDS_0)}{N} \tag{5-9}$$

i.e. the mutual information between X and $Pa(X)$ coincides with the difference between the log-likelihood of LDS and LDS_0.

The G^2 test is commonly used to test independence and conditional independence between two variables for discrete data. From the general formula for G^2, we know that the value of G^2 can also be calculated from mutual information [81]. Thus, we can write the G^2 test value between X and $Pa(X)$ as:

$$G^2(X, Pa(X)) = 2N(MI(X, Pa(X))) \tag{5-10}$$

The number of degrees of freedom for G^2 test between X and $Pa(X)$ is:

$$\begin{aligned} &DF(G^2(X, Pa(X))) = (Cat(X) - 1)(Cat(Pa(X)) - 1) \\ &= (r-1)(q-1) \end{aligned} \tag{5-11}$$

where $Cat(V)$ is the number of categories of the variable V, and thus $Cat(X) = r$ and $Cat(Pa(X)) = q$ [79].

It is interesting to note that the difference between the complexity of LDS and LDS_0 is equal to the degree of freedom of $G^2(X,Pa(X))$ by

$$\begin{aligned} &C(LDS) - C(LDS_0) = (r-1)q - (r-1) \\ &= (r-1)(q-1) = DF(G^2(X, Pa(X))) \end{aligned} \tag{5-12}$$

By applying Eq. (5-7)-(5-12), the difference of AIC scores between LDS and LDS_0 is:

$$\begin{aligned}
&\log P(D \mid LDS) - \log P(D \mid LDS_0) \\
&= (\log P(D \mid \hat{\theta}_{LDS}, LDS) - C(LDS)) - (\log P(D \mid \hat{\theta}_{LDS_0}, LDS_0) - C(LDS_0)) \\
&= N(MI(X, Pa(X)) - (r-1)(q-1) \\
&= 1/2(G^2(X, Pa(X)) - 2DF(G^2(X, Pa(X))))
\end{aligned} \tag{5-13}$$

Thus, the AIC score becomes:

$$\begin{aligned}
&\log P(D \mid LDS) \\
&= 1/2(G^2(X, Pa(X)) - 2DF(G^2(X, Pa(X)))) + \log P(D \mid LDS_0)
\end{aligned} \tag{5-14}$$

where $\log P(D \mid LDS_0)$ is a constant.

The distribution of G^2 asymptotically approximates to that of chi-square with the same number of degrees of freedom [79]. The chi-square distribution with k degrees of freedom has a variance of $2k$, and therefore $2DF(G^2(X,Pa(X)))$ is the variance of the corresponding G^2 distribution. Since $G^2(X,Pa(X))$ reflects the bias, the AIC score in Eq. (5-14) indicates a trade-off between bias and variance in terms of the G^2 statistic $G^2(X,Pa(X))$ and its variance. One problem for the AIC score in Eq. (5-5), Eq. (5-7), and Eq. (5-14) is that it assumes that the noise variance is equal to one, which is not true especially when applied to discrete data like SNP data [104-105]. We can confirm this by comparing $2DF(G^2(X,Pa(X)))$ with the true variance of $G^2(X,Pa(X))$ from data. There is a large deviation between them when $Pa(X)$ contains more than two parent nodes. The more parent nodes $Pa(X)$ contains, the larger the deviation is. One simple but practical way to consider and estimate the noise variance in AIC score is replacing $2DF(G^2(X,Pa(X)))$ in Eq. (5-14) with the true variance of $G^2(X,Pa(X))$ from data , and our new epistatic scoring function (EpiScore) becomes:

$$EpiScore(LDS:D) = \log P(D \mid LDS)$$
$$= 1/2(G^2(X, Pa(X)) - Variance_D(G^2(X, Pa(X)))) + \log P(D \mid LDS_0) \quad (5\text{-}15)$$

where $Variance_D(G^2(X,Pa(X)))$ comes from the estimation of the variance of the corresponding G^2 distribution from data. Our new scoring function estimates the penalty term from data to guarantee its reliability.

5.3 A Branch-and-Bound Algorithm for Local Structure Learning in Bayesian Networks

The computational task in score-and-search methods is to find a network structure with the highest score. The searching space consists of a super-exponential number of structures and thus exhaustively searching optimal structure from data for Bayesian Networks is NP-hard [106]. One simple heuristic search algorithm is greedy hill-climbing algorithm. In greedy hill-climbing algorithm, there are three types of operators that change one edge at each step: add an edge, remove an edge, and reverse an edge. By these three operators, we can construct the local neighborhood of the current network. Then we select the network with the highest score in the local neighborhood to get the maximal gain. This process can be repeated until it reaches a local maximum. However, greedy hill-climbing algorithm cannot guarantee a global maximum [96]. Other structure learning methods for Bayesian Networks include Branch-and-Bound (B&B) [107-109], genetic algorithms, [110] and Markov chain Monte Carlo [111].

We employ B&B algorithm in our study because the B&B algorithm can guarantee the optimal results in a significantly reduced search time compared to

exhaustive search. Our EpiBN method uses B&B algorithm to search a local disease structure that maximizes the EpiScore in Eq. (5-15). The pseudo code of EpiBN is shown in Algorithm 5.1. In EpiBN, the procedure BN_B&B starts from an empty parent node set and constructs a depth-first search tree to find the optimal parent (disease SNPs) set for the disease status node. In our B&B search, instead of using the pruning strategy as in [107-108], which sets a lower bound for the MDL score to prune the search tree, we stop the recursive calls when we observe that the score will decrease on the children state of the current state. This strategy cannot guarantee

Algorithm 5.1 EpiBN

Input: Data D, Disease status node, all n SNP nodes

Output: Disease SNP nodes, which has the maximum EpiScore on Disease status node

Procedure [S_1 P_1] = BN_B&B (V_1)

Input: SNP node set V_1.

Output: EpiScore S_1, parent SNP node set P_1.

Begin

1. Compute EpiScore $tempS_1$ for V_1, S_1=$tempS_1$, P_1=V_1
2. **IF** V_1=null **then** i=0 **else** i= V_1 (end)
3. For $i+1 \le q \le n$

   Begin

   (1) $V_2 = V_1 \cup q$ Compute EpiScore $tempS_2$ for V_2

   (2) **IF** $tempS_2 > tempS_1$ **then** [S_2 P_2] = BN_B&B (V_2)

   (3) **IF** $S_2 > S_1$ **then** S_1=S_2, P_1= P_2

   End

End

global optima theoretically. However, it will significantly speed up the search process and perform well practically.

5.4 MCMC Screening Method for Real Datasets

Even though the B&B algorithm uses a lower bound to reduce the searching space, it still has an exponential time complexity in the worst case and is not feasible to be directly applied to real GWAS data. Therefore, an efficient screening method is necessary. Traditional screening methods assign a score to every single SNP and select a subset of SNPs with high scores. However, these methods ignore the joint effect of SNPs on disease and are not suitable for detecting gene-gene interactions from real GWAS data.

In this dissertation, we use the Markov chain Monte Carlo (MCMC) method [111] to perform the screening process. In the Bayesian Network for gene-gene interaction detection, we use a Metropolis-Hastings method to build a Markov chain to get the posterior probability for each edge from the SNP nodes to the disease status node. At each step of the Markov chain, we use two types of moves: add an edge and remove an edge. The proposed move is accepted with probability

$$\alpha = \min\{1, R_\alpha\} \tag{5-16}$$

where

$$R_\alpha = \frac{\#(nbd(LDS))P(LDS'|D)}{\#(nbd(LDS'))P(LDS|D)} \tag{5-17}$$

where #(*nbd*(*LDS*)) is the cardinality of the neighborhood of the current local disease structure and *LDS'* is the candidate local disease structure in each step of the

Markov chain. Since *LDS* and *LDS'* differ in one move, the ratio $\#(nbd(LDS))/\#(nbd(LDS'))$ is one. In addition, the posterior probability of the local disease structure, $P(LDS \mid D)$, is that $P(LDS \mid D) \propto P(D \mid LDS)P(LDS)$ and we take a uniform distribution over the considered local disease structures. Therefore, the acceptance ratio in Eq. (5-17) becomes:

$$R_{\alpha} = P(D \mid LDS') / P(D \mid LDS) \tag{5-18}$$

The likelihood of local disease structure, $P(D \mid LDS)$, can be calculated by Eq. (5-15).

Based on the result from MCMC method, we select SNP nodes associated with edges whose posterior probabilities larger than 0. Since we consider the association of multiple SNPs with disease status at each step of the Markov chain in our MCMC method, the potential disease SNPs related with gene-gene interactions will be kept in the final subset of SNPs.

5.5 Experimental results

In this section, we assess the proposed EpiBN method on both simulated datasets and real biological datasets.

5.5.1 Analysis of Simulated Data

We first evaluate the proposed EpiBN method on four simulated data sets, which are generated from three commonly used two-locus epistatic models in section 4.5 and one three-locus epistatic model developed in [44]. We show the three-locus epistatic

model (model4) in Table 5.1. There are three disease loci in model4 [44]. Some certain genotype combinations can increase disease risk in model4 and there are almost no marginal effects for each disease locus.

Table 5.1 A three-locus epistatic model

Model 4	AA		
	BB	Bb	bb
CC	α	α	α
Cc	α	α	$\alpha(1+\theta)$
cc	α	$\alpha(1+\theta)$	α
	Aa		
	BB	Bb	bb
CC	α	α	$\alpha(1+\theta)$
Cc	α	$\alpha(1+\theta)$	α
cc	$\alpha(1+\theta)$	α	α
	aa		
	BB	Bb	bb
CC	α	$\alpha(1+\theta)$	α
Cc	$\alpha(1+\theta)$	α	α
cc	α	α	α

For model1, model2, and model3, we use the same parameters as in section 4.5. For model4, we arbitrarily set $\theta = 7$ because there are almost no marginal effects in model4. We first generate 50 datasets and each dataset contains 100 markers genotyped for 1,000 cases and 1,000 controls based on each parameter setting for

each model. To measure the performance of each method, we use power as our evaluation criterion, which is defined as the proportion of simulated datasets in which all diseases associated markers are identified without any false positives.

We compare the EpiBN algorithm with three methods: BEAM, SVMs, and MDR on the four simulated disease models. The results on the simulated data are shown in Figure 5.2 and Figure 5.3. As can be seen, among the four methods, the EpiBN method performs the best, and BEAM is the second best. One possible reason is that BEAM tries to detect single disease locus and epistatic interactions simultaneously. This strategy makes BEAM unnecessarily over-complex. In most cases, the powers of both MDR and SVM are much smaller than those of the EpiBN and BEAM algorithms.

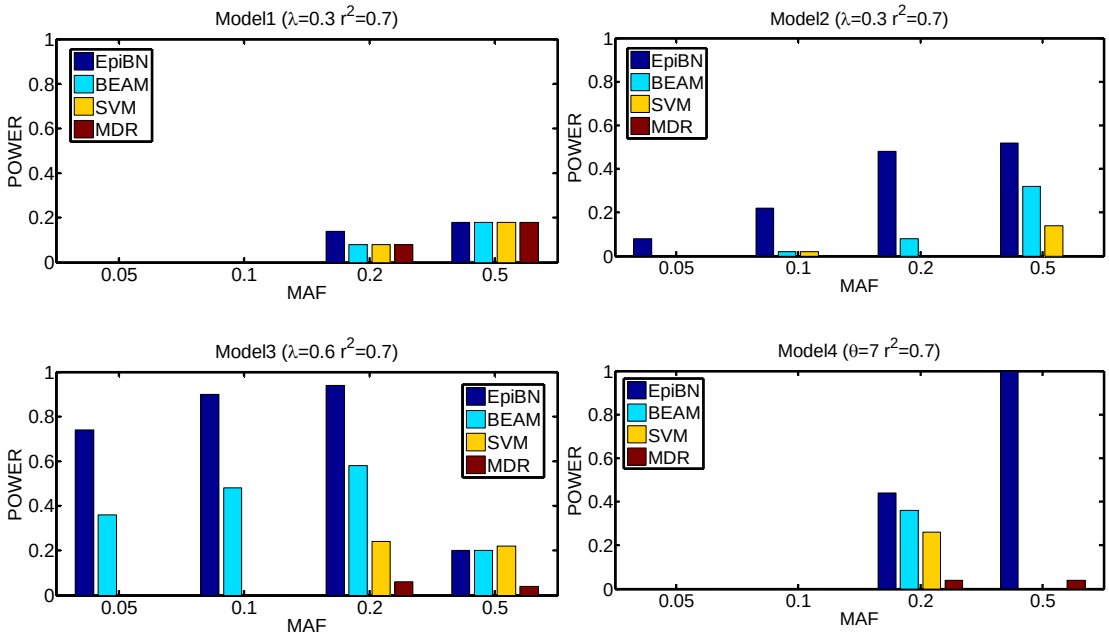

Figure 5.2 Performance comparison of EpiBN, BEAM, SVM and MDR (r^2 =0.7).

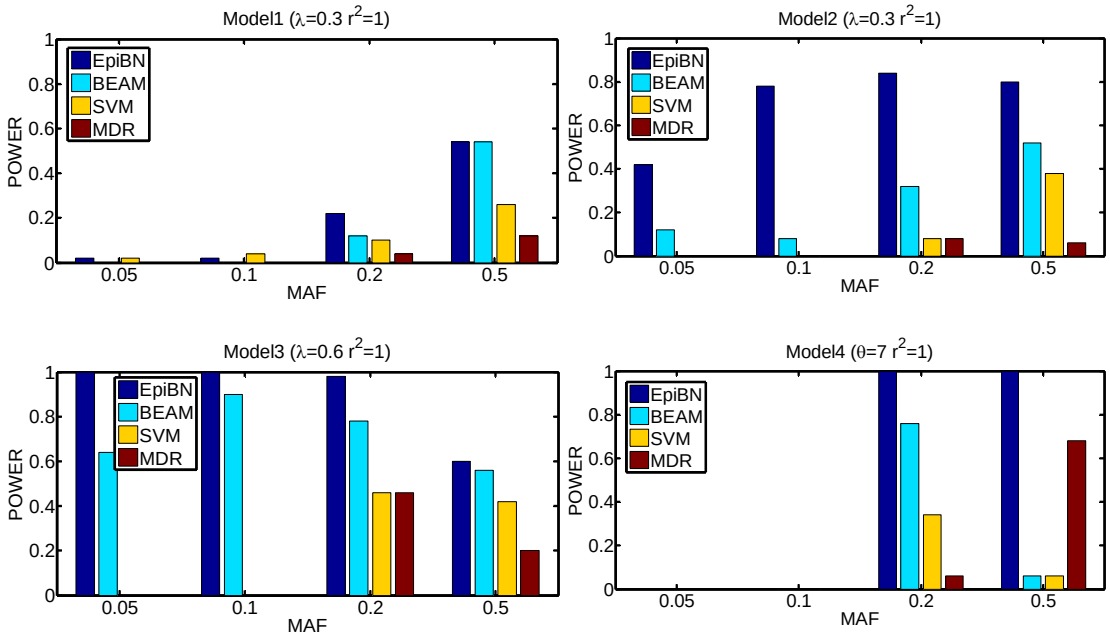

Figure 5.3 Performance comparison of EpiBN, BEAM, SVM and MDR (r^2 =1).

Typically, GWAS can not generate a large number of samples due to the high experiment cost. Thus, the performance of various computational methods for gene-gene interaction detection in case of small samples is important. We explore the effect of the number of samples on the performance of EpiBN, MDR, BEAM and SVM. The parameters used are: $\lambda = 1.1$ for model1, $\lambda = 0.9$ for model2, $\lambda = 1.8$ for model3, and $\theta = 7$ for model4. To test the scalability of EpiBN on large number of SNPs, we generate synthetic datasets containing different number of markers (40, 200, and 1000) genotyped for different number of samples (100, 200, 300, 400, 600, 1000, and 2000) with $r^2 = 1$ and MAF=0.5.

The results are shown in Figure 5.4, Figure 5.5, and Figure 5.6. We find that EpiBN is more sample-efficient than other methods in that it can achieve the highest

power when the number of samples is the same. In addition, it needs fewer samples to reach the perfect power comparing to other methods. BEAM is still the second best. The powers of both MDR and SVM are still smaller than those of the EpiBN and BEAM algorithms. However, MDR and SVM demonstrate a better performance comparing to Figure 5.3 and Figure 5.4. This is perhaps due to the fact that increasing the marginal effect size λ for model1-model3 makes the detecting task suitable for the pre-filtering based methods such as MDR and SVM. The result from model4 is particularly interesting: EpiBN exhibits overwhelming superiority over other three methods, as EpiBN yields a perfect power even the number of samples is small (around 600), which indicates that EpiBN is especially suitable for detecting epistatic interactions with weak or no marginal effects. From Figure 5.4, Figure 5.5, and

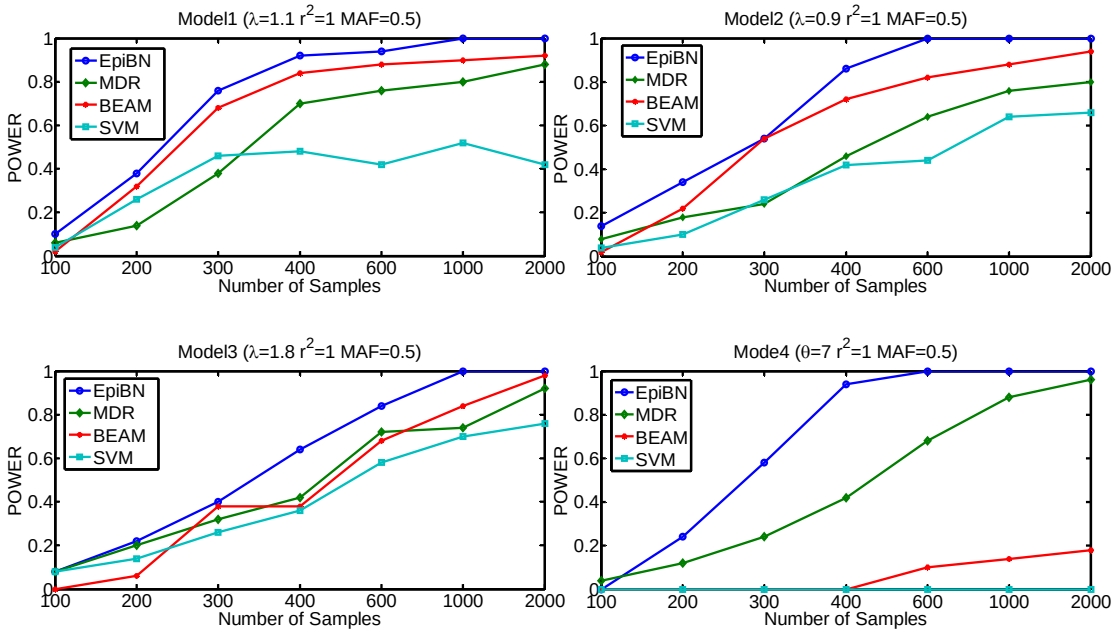

Figure 5.4 Comparison of sample efficiency on datasets with 40 SNPs.

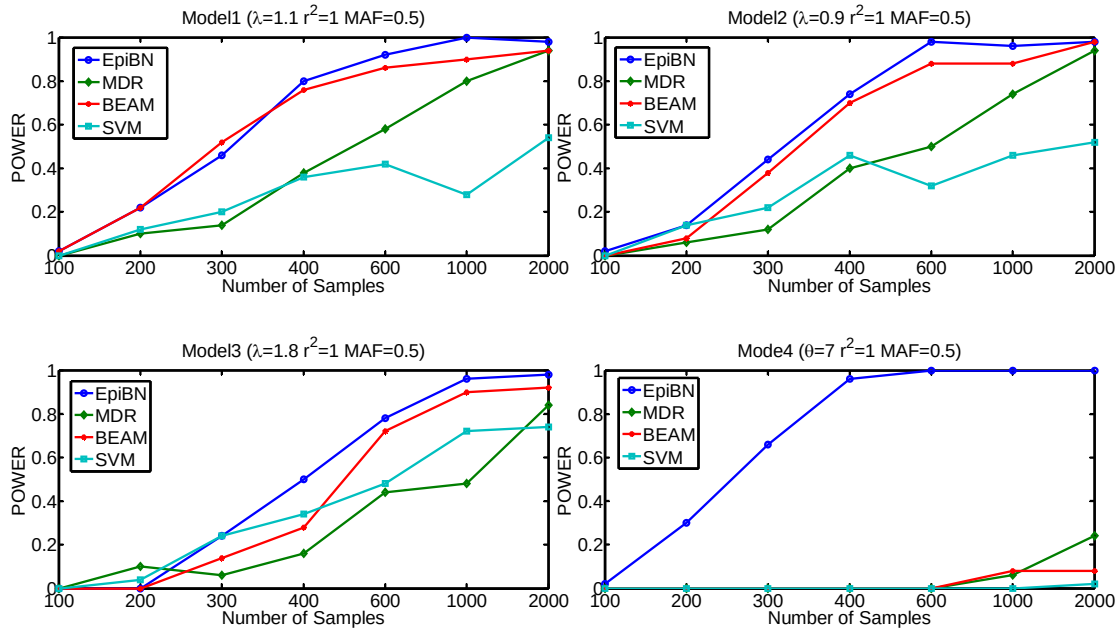

Figure 5.5 Comparison of sample efficiency on datasets with 200 SNPs.

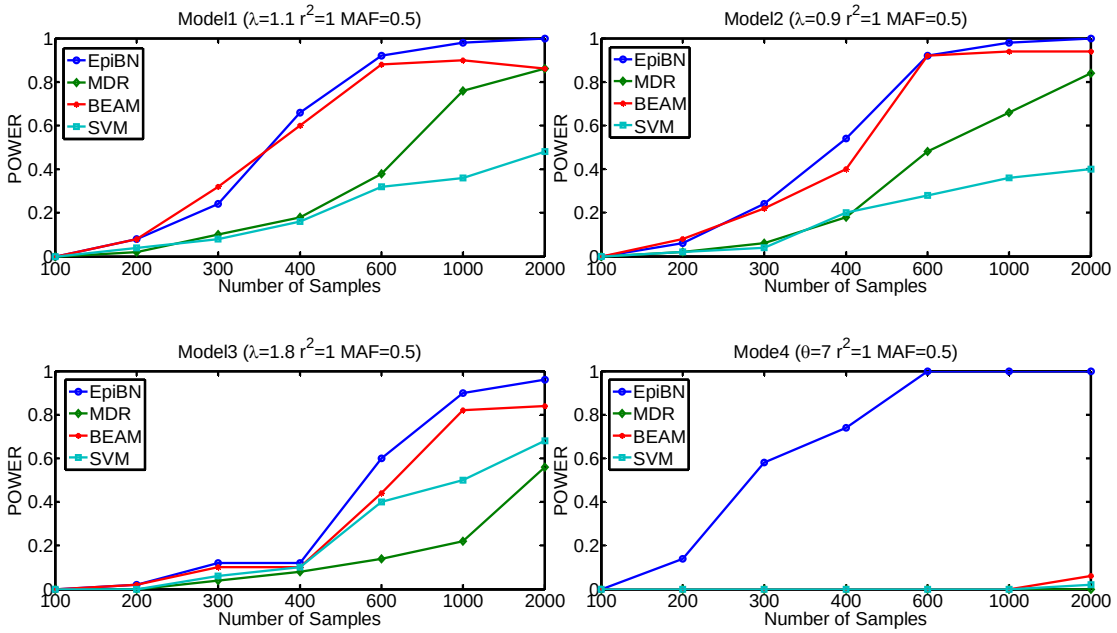

Figure 5.6 Comparison of sample efficiency on datasets with 1000 SNPs.

Figure 5.6, we can also find that increasing the number of genotyping markers, like adding some noise to the data, will impair the power of all methods, especially in case of small samples.

5.5.2 Analysis of AMD Data

From the results on simulated data, EpiBN demonstrates a better performance than three other methods. Notice that a real genome-wide case-control association study may require genotyping of 30,000–1,000,000 common SNPs. In this section, we show that EpiBN algorithm can also handle large-scale datasets in real genome-wide case-control studies. We consider an Age-related Macular Degeneration (AMD) dataset, which contains 116,204 SNPs genotyped with 96 cases and 50 controls [112]. AMD (OMIM 603075) [113] is a common genetic disease related to the progressive visual dysfunction in age over 70 in the developed country. A GWA study was successfully conducted on this disease finding two associated SNPs, rs380390 and rs1329428 ('rs': assigned reference SNP ID by dbSNP [114]) in non-coding region of the gene for complement factor H (*CFH*), which is located on chromosome 1 in a region linked to AMD [112].

In the phase of preprocessing data, we remove non-polymorphic SNPs and those that significantly deviated from Hardy-Weinberg Equilibrium (HWE). We also remove all SNPs that have more than five missing genotypes. After filtering, there are 97, 327 SNPs lying in 22 autosomal chromosomes remained.

We first perform the screening process and select 51 potential disease SNPs related with AMD by MCMC method (see detail in section 5.4). Among these 51 selected SNPs, EpiBN detects two associated SNPs: rs380390 and rs2402053, which have a G^2 test p-value of 5.36×10^{-10}. The first SNP, rs380390, is already found in [112] with a significant association with AMD. The other SNP detected by the EpiBN algorithm is SNP rs2402053, which is intergenic between TFEC and TES in chromosome 7q31 [115].

Even though no evidences show that rs2402053 is related with AMD, it is worth noting that mutations in some genes on 7q31-q32 are revealed in patients with retinal disorders [116-117]. Therefore, rs2402053 may be a new genetic factor, on chromosome 7q, contributing to the underlying mechanism of AMD. The real mechanism of interaction between rs380390 and rs2402053 should be explored further by biological experiments.

5.5.3 Analysis of LOAD Data

Late-onset Alzheimer's disease (LOAD) is the most common form of Alzheimer's disease and usually occurs in persons over 65. It causes patients' degeneration of the ability of thinking, memory, and behavior. The apolipoprotein E (APOE) gene is one genetic factor that accounts for affecting the risk of LOAD. There are three common variants of the APOE gene: $\varepsilon2$, $\varepsilon3$, and $\varepsilon4$. The appearance of the $\varepsilon4$ allele in a person's APOE genotype increases the LOAD risk. Rieman et al. conducted genome-wide association studies to detect other generic risk factors related with

LOAD [118]. They found 10 SNPs showing significant association with LOAD in the APOE $\varepsilon 4$ carriers. All these 10 SNPs are in the GRB-associated binding protein 2 (GAB2) gene.

We download the LOAD GWAS data from http://www.tgen.org/neurogenomics/data. After pre-processing, we have 287,479 SNPs and 1408 samples (857 cases and 551 controls). EpiBN keeps APOE as one parent of the disease status node and identifies two other SNPs: rs1931565 and rs4505578, which may interact with APOE and affect the LOAD risk. The rs1931565 SNP is intergenic between ABCA4 and ARHGAP29 in chromosome 1p22. ABCA4 is related with some brain-related diseases including stargardt disease 1, early-onset severe retinal dystrophy and age-related macular degeneration. On the other hand, some ABC transporter family genes such as ABCA1, ABCA2, ABCA7, and ABCA12 are associated with Alzheimer's disease [119]. Therefore, we can speculate that the interaction among rs1931565, rs4505578 and APOE may affect some brain functions and therefore increase the LOAD risk.

Our results do not contain any of the 10 SNPs in GAB2 found in [118]. One reason is that Rieman et al. only explored two-locus interactions related with LOAD. In fact, the gene-gene interactions are very complicated. If we restrict the number of genetic risk factors as two, we will miss some potential disease SNPs associated with complex diseases.

Chapter 6 Conclusion and Future Work

Cancer is a system biology disease and two types of genes: oncogenes and tumor suppressors play the central role in the process of transforming normal cells into tumor cells. These cancer-related genes may cooperate with each other or affect other genes to regulate some fundamental cell processes such as death, proliferation, differentiation, and migration. Thus, identifying differential gene relations and gene-gene interactions associated with cancer can contribute to the understanding of the underlying molecular mechanisms of cancer and therefore help to improve pathogenesis, prevention, diagnosis, and treatment of cancer.

6.1 Summary of Research

In this dissertation, I use machine learning and computational methods to address two problems in cancer research: (1) identifying differential gene relations and (2) detecting gene-gene interactions (epistasis). Over the past two decades, a lot of high-throughput techniques have been developed to generate different types of cancer research data such as gene expression, chip-on-chip, next generation sequencing, RNA-seq, and genome-wide association studies (GWAS). In this dissertation, I focus on gene expression data and GWAS data and perform the analysis at two levels: genes and genetic variants. Identifying differential gene relations can reveal the activities of cancer-related genes in a biological system. On the other hand, detecting gene-gene interactions can determine genes that influence the phenotype (disease and non-disease).

To identify gene pairs that have different relationships in normal versus cancer tissues, I develop an integrative method based on the bootstrapping K-S test to evaluate a large number of microarray datasets generated from 21 different types of cancer. The significant alteration of gene relations can greatly extend our understanding of the molecular mechanisms of human cancer. My method avoids the disadvantage of the traditional t-test, which only considers the mean and variance of samples and fails in the analysis of microarray data with small numbers of samples. Instead of the t-test, I propose the use of the bootstrapping K-S test method to detect gene pairs with different distributions of Pearson correlation coefficient values in normal and tumor samples. The experimental results demonstrate that our method can find meaningful alterations in gene relations and open a potential door for further cancer research.

For gene-gene interaction detection, I propose to use two Bayesian Network based methods, DASSO-MB and EpiBN, to address two critical challenges: searching and scoring. DASSO-MB is based on the concept of Markov Blanket in Bayesian Networks. Comparing with many computational methods used for identification of gene-gene interactions, DASSO-MB can increase power and reduce false positives. This is critical in saving the potential costs of biological experiments and being an efficient guideline for pathogenesis research. However, DASSO-MB is sample-consuming and the greedy searching strategy in DASSO-MB is not suitable for detecting some interaction models with no independent main effects for each disease locus. To address the problems of DASSO-MB, I propose EpiBN, a Bayesian

Network structure learning method. In EpiBN, I develop a new scoring function, which can reflect higher-order gene-gene interactions and detect the true number of disease markers, and apply a fast Branch-and-Bound algorithm to learn the structure of a two-layer Bayesian Network containing only one target node. To make my method scalable to GWAS data, I use a MCMC method to perform the screening process. The experimental results demonstrate that EpiBN outperforms some other commonly-used methods and is scalable to GWAS data.

6.2 Future Work

6.2.1 New Score Scheme for Differential Gene Network Detection

Most cancer research methods based on microarray technology only focus on identifying differential genes as biomarkers for cancer detection or future therapy and detecting differential gene relations is a complementary approach. An obvious drawback of these two methods is that they ignore the importance of cancer research at system level. System biology explores the interactions between subsets of genes or all genes and how these interactions regulate functions and behaviors of a biological system.

In order to understand the inside mechanism of cancer, we must examine the alteration of system structures and dynamics related with cellular functions and biological processes, rather than just a simple list of differential genes or differential gene relations. Identifying all differential genes and differential gene relations in an organism is like listing all the malignant parts in a system. While such a list provides

a catalog of the individual components for further research, it is not sufficient to understand the complex mechanism underlying different types of cancer. We need to know how these parts affect and change the inside mechanism in a system. Therefore detecting differential signaling pathways which respond to different cell states is a better choice for cancer research. It is an extension of the method for detecting differential genes or differential gene relations.

We have already introduced several different methods for detecting differential gene networks in Chapter 2. There are several shortcomings of these methods. First, it is hard to measure the alteration of a pathway or to determine whether a pathway has been significantly altered. For example, Levine *et al*. use five score schemes as pathway 'activation metrics' [36]. Rahnenfuhrer *et al*. analyze the change of activity of a pathway for different samples based on the calculation of correlation coefficients [37]. These two methods only consider a part of the activity of a pathway. Second, all the methods for differential pathway analysis are based on a single dataset thus lack the power of integrative methods. Integrating microarray datasets obtained from different laboratories can combine complementary pieces of information in various datasets, enable broader understanding of gene regulation, and achieve more reliable and more valid results. Moreover, integrative methods for multiple microarray datasets across different types of cancers help us identify deregulated signaling pathways that are common to all types of cancer or specific to some certain types of cancer. Third, several differentia pathway analysis methods extract pathways from

KEGG to determine whether these pathways are significantly altered or not. However, pathways in KEGG database give us incomplete information.

For future studies, I would like to detect differential gene networks by an integrative method using multiple microarray datasets. A new score scheme should be used to measure the alteration of a gene network by considering both gene expression and gene relation. The change of expression level reflects the altered activity of a gene and the change of gene relation can reflect the alteration of gene function. So we treat gene and gene relation as equally important entities.

6.2.2 Detect Substantial SNP-Gene Pairs

Genetical genomics provide the genetic basis of variation in gene expression. However, one serious issue of genetical genomics in case-control studies is that genetical genomics can only find strongly associated SNP-gene pairs. How these SNPs influence disease susceptibility via affecting the activity of genes is still mysterious for us. Therefore a novel method is needed to detect substantial SNP-gene pairs related with the alteration of disease susceptibility.

We show the relations of GWAS, genetical genomics (GOGE, eQTL) and detection of differential genes in Figure 6.1, and Figure 6.1 demonstrates that GOGE studies connect SNP data and gene expression data as a bridge. By analyzing Figure 6.1, we can find that there are two types of phenotype data: expression data and label/disease status. Expression data are one type of intermediate phenotype data that are related to DNA sequence variants. The assumption of GOGE studies is that

genetic variants can exert effects on gene expression and these changes of gene expression will cause diseases. But the issue of current GOGE studies is that no consistent result can be obtained in Figure 6.1. In other words, causal SNPs detected by GOGE studies are not the SNPs showing a strong association with disease status. The genes affected by genetic variants are also not differential genes. This issue may arise from the association calculation method between SNPs and genes in GOGE studies. GOGE studies only calculate associations between SNPs and genes by expression data and SNP data in case/control studies, omitting the importance of sample labels. Therefore, I would like to design a novel method to detect some essential SNP-gene pairs. Among these detected SNP-gene pairs, SNPs should be associated with disease status substantially. Genes should show a sound differential activity between different types of samples. In the meanwhile these SNPs' influence on the differential activity of genes from controls to cases should be confirmed.

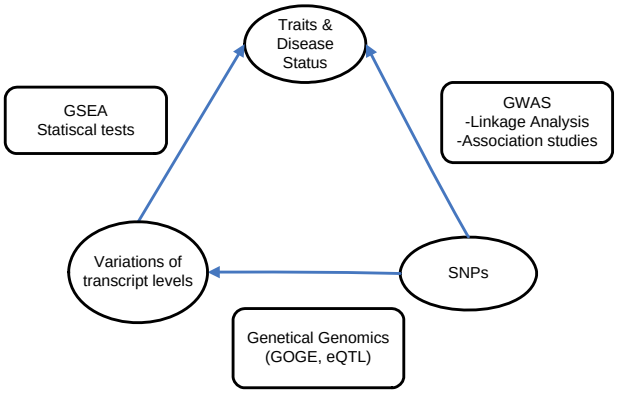

Figure 6.1 Relations of GWAS, genetical genomics (GOGE, eQTL) and detection of differential genes/pathways.

Reference

[1] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell,* vol. 100, pp. 57-70, Jan 7 2000.

[2] A. G. Knudson, "Two genetic hits (more or less) to cancer," *Nat Rev Cancer,* vol. 1, pp. 157-62, Nov 2001.

[3] J. H. Moore*, et al.*, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics,* vol. 26, pp. 445-55, Feb 15 2010.

[4] H. J. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans," *Hum Mol Genet,* vol. 11, pp. 2463-8, Oct 1 2002.

[5] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nat Rev Genet,* vol. 10, pp. 392-404, Jun 2009.

[6] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nat Rev Genet,* vol. 6, pp. 95-108, Feb 2005.

[7] "The International HapMap Project," *Nature,* vol. 426, pp. 789-96, Dec 18 2003.

[8] "A haplotype map of the human genome," *Nature,* vol. 437, pp. 1299-320, Oct 27 2005.

[9] J. DeRisi*, et al.*, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nat Genet,* vol. 14, pp. 457-60, Dec 1996.

[10] M. Schena*, et al.*, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science,* vol. 270, pp. 467-70, Oct 20 1995.

[11] V. G. Tusher*, et al.*, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.

[12] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes," *Bioinformatics,* vol. 17, pp. 509-19, Jun 2001.

[13] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistical Sinica,* vol. 12, pp. 31-46, 2002.

[14] H. Jiang*, et al.*, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," *BMC Bioinformatics,* vol. 5, p. 81, Jun 24 2004.

[15] K. Y. Kim*, et al.*, "Novel and simple transformation algorithm for combining microarray data sets," *BMC Bioinformatics,* vol. 8, p. 218, 2007.

[16] P. Warnat*, et al.*, "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes," *BMC Bioinformatics,* vol. 6, p. 265, 2005.

[17] D. R. Rhodes, *et al.*, "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Res,* vol. 62, pp. 4427-33, Aug 1 2002.
[18] R. A. Fisher, *Statistical methods for research workers*, 7th ed. Edinburgh,: Oliver and Boyd, 1938.
[19] J. K. Choi, *et al.*, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics,* vol. 19 Suppl 1, pp. i84-90, 2003.
[20] P. Hu, *et al.*, "Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models," *BMC Bioinformatics,* vol. 6, p. 128, 2005.
[21] R. Breitling, *et al.*, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Lett,* vol. 573, pp. 83-92, Aug 27 2004.
[22] T. Yuen, *et al.*, "Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays," *Nucleic Acids Res,* vol. 30, p. e48, May 15 2002.
[23] D. Kostka and R. Spang, "Finding disease specific alterations in the co-expression of genes," *Bioinformatics,* vol. 20 Suppl 1, pp. i194-9, Aug 4 2004.
[24] Y. Cheng and G. M. Church, "Biclustering of expression data," *Proc Int Conf Intell Syst Mol Biol,* vol. 8, pp. 93-103, 2000.
[25] M. Watson, "CoXpress: differential co-expression in gene expression data," *BMC Bioinformatics,* vol. 7, p. 509, 2006.
[26] K. C. Li, "Genome-wide coexpression dynamics: theory and application," *Proc Natl Acad Sci U S A,* vol. 99, pp. 16875-80, Dec 24 2002.
[27] K. C. Li, *et al.*, "A system for enhancing genome-wide coexpression dynamics study," *Proc Natl Acad Sci U S A,* vol. 101, pp. 15561-6, Nov 2 2004.
[28] Y. Lai, *et al.*, "A statistical method for identifying differential gene-gene co-expression patterns," *Bioinformatics,* vol. 20, pp. 3146-55, Nov 22 2004.
[29] M. Dettling, *et al.*, "Searching for differentially expressed gene combinations," *Genome Biol,* vol. 6, p. R88, 2005.
[30] L. Ein-Dor, *et al.*, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics,* vol. 21, pp. 171-8, Jan 15 2005.
[31] D. B. Allison, *et al.*, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet,* vol. 7, pp. 55-65, Jan 2006.
[32] J. K. Choi, *et al.*, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics,* vol. 21, pp. 4348-55, Dec 15 2005.
[33] S. Ma and M. R. Kosorok, "Identification of differential gene pathways with principal component analysis," *Bioinformatics,* vol. 25, pp. 882-9, Apr 1 2009.
[34] H. Kitano, "Computational systems biology," *Nature,* vol. 420, pp. 206-10, Nov 14 2002.

[35] H. Kitano, "Systems biology: a brief overview," *Science,* vol. 295, pp. 1662-4, Mar 1 2002.
[36] D. M. Levine*, et al.*, "Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways," *Genome Biol,* vol. 7, p. R93, 2006.
[37] J. Rahnenfuhrer*, et al.*, "Calculating the statistical significance of changes in pathway activity from gene expression data," *Stat Appl Genet Mol Biol,* vol. 3, p. Article16, 2004.
[38] G. Sanguinetti*, et al.*, "MMG: a probabilistic tool to identify submodules of metabolic pathways," *Bioinformatics,* vol. 24, pp. 1078-84, Apr 15 2008.
[39] L. Cabusora*, et al.*, "Differential network expression during drug and stress response," *Bioinformatics,* vol. 21, pp. 2898-905, Jun 15 2005.
[40] T. F. Fuller*, et al.*, "Weighted gene coexpression network analysis strategies applied to mouse weight," *Mamm Genome,* vol. 18, pp. 463-72, Jul 2007.
[41] J. Marchini*, et al.*, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nat Genet,* vol. 37, pp. 413-7, Apr 2005.
[42] M. D. Ritchie*, et al.*, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Am J Hum Genet,* vol. 69, pp. 138-47, Jul 2001.
[43] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics,* vol. 9, pp. 30-50, Jan 2008.
[44] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nat Genet,* vol. 39, pp. 1167-73, Sep 2007.
[45] C. Yang*, et al.*, "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics,* vol. 25, pp. 504-11, Feb 15 2009.
[46] S. H. Chen*, et al.*, "A support vector machine approach for detecting gene-gene interaction," *Genet Epidemiol,* vol. 32, pp. 152-67, Feb 2008.
[47] R. Jiang*, et al.*, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics,* vol. 10 Suppl 1, p. S65, 2009.
[48] W. Cookson*, et al.*, "Mapping complex disease traits with global gene expression," *Nat Rev Genet,* vol. 10, pp. 184-94, Mar 2009.
[49] V. G. Cheung and R. S. Spielman, "Genetics of human gene expression: mapping DNA variants that influence gene expression," *Nat Rev Genet,* vol. 10, pp. 595-604, Sep 2009.
[50] B. Han*, et al.*, "Integrating multiple microarray data for cancer pathway analysis using bootstrapping K-S test," *J Biomed Biotechnol,* vol. 2009, p. 707580, 2009.
[51] T. Barrett*, et al.*, "NCBI GEO: mining millions of expression profiles--database and tools," *Nucleic Acids Res,* vol. 33, pp. D562-6, Jan 1 2005.
[52] H. Ogata*, et al.*, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res,* vol. 27, pp. 29-34, Jan 1 1999.

[53] V. Matys, *et al.*, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res,* vol. 31, pp. 374-8, Jan 1 2003.
[54] W. J. Conover, *Practical nonparametric statistics*, 3rd ed. New York: Wiley, 1999.
[55] G. C. Blobe, *et al.*, "A novel mechanism for regulating transforming growth factor beta (TGF-beta) signaling. Functional modulation of type III TGF-beta receptor expression through interaction with the PDZ domain protein, GIPC," *J Biol Chem,* vol. 276, pp. 39608-17, Oct 26 2001.
[56] M. Barrios-Rodiles, *et al.*, "High-throughput mapping of a dynamic signaling network in mammalian cells," *Science,* vol. 307, pp. 1621-5, Mar 11 2005.
[57] F. Colland, *et al.*, "Functional proteomics mapping of a human signaling pathway," *Genome Res,* vol. 14, pp. 1324-32, Jul 2004.
[58] J. F. Rual, *et al.*, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature,* vol. 437, pp. 1173-8, Oct 20 2005.
[59] H. Shoji, *et al.*, "Identification and characterization of a PDZ protein that interacts with activin type II receptors," *J Biol Chem,* vol. 275, pp. 5485-92, Feb 25 2000.
[60] J. Ban, *et al.*, "EWS-FLI1 in Ewing's sarcoma: real targets and collateral damage," *Adv Exp Med Biol,* vol. 587, pp. 41-52, 2006.
[61] J. U. Wurthner, *et al.*, "Transforming growth factor-beta receptor-associated protein 1 is a Smad4 chaperone," *J Biol Chem,* vol. 276, pp. 19495-502, Jun 1 2001.
[62] L. Chen, *et al.*, "Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function," *Mol Cell,* vol. 17, pp. 393-403, Feb 4 2005.
[63] L. del Peso, *et al.*, "Interleukin-3-induced phosphorylation of BAD through the protein kinase Akt," *Science,* vol. 278, pp. 687-9, Oct 24 1997.
[64] H. Inada, *et al.*, "Keratin attenuates tumor necrosis factor-induced cytotoxicity through association with TRADD," *J Cell Biol,* vol. 155, pp. 415-26, Oct 29 2001.
[65] H. Hsu, *et al.*, "TNF-dependent recruitment of the protein kinase RIP to the TNF receptor-1 signaling complex," *Immunity,* vol. 4, pp. 387-96, Apr 1996.
[66] S. Baksh, *et al.*, "The tumor suppressor RASSF1A and MAP-1 link death receptor signaling to Bax conformational change and cell death," *Mol Cell,* vol. 18, pp. 637-50, Jun 10 2005.
[67] Q. L. Deveraux, *et al.*, "IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases," *Embo J,* vol. 17, pp. 2215-23, Apr 15 1998.
[68] M. D. Vos, *et al.*, "RASSF2 is a novel K-Ras-specific effector and potential tumor suppressor," *J Biol Chem,* vol. 278, pp. 28045-51, Jul 25 2003.
[69] H. Tsutsui, *et al.*, "Members of the MAZ family: a novel cDNA clone for MAZ from human pancreatic islet cells," *Biochem Biophys Res Commun,* vol. 226, pp. 801-9, Sep 24 1996.

[70] M. Nanjundan, *et al.*, "Plasma membrane phospholipid scramblase 1 promotes EGF-dependent activation of c-Src through the epidermal growth factor receptor," *J Biol Chem,* vol. 278, pp. 37413-8, Sep 26 2003.
[71] B. A. McKinney, *et al.*, "Machine learning for detecting gene-gene interactions: a review," *Appl Bioinformatics,* vol. 5, pp. 77-88, 2006.
[72] S. K. Musani, *et al.*, "Detection of gene x gene interactions in genome-wide association studies of human population data," *Hum Hered,* vol. 63, pp. 67-84, 2007.
[73] L. W. Hahn, *et al.*, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics,* vol. 19, pp. 376-82, Feb 12 2003.
[74] J. H. Moore, *et al.*, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *J Theor Biol,* vol. 241, pp. 252-61, Jul 21 2006.
[75] M. D. Ritchie, *et al.*, "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity," *Genet Epidemiol,* vol. 24, pp. 150-7, Feb 2003.
[76] T. T. Wu, *et al.*, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics,* vol. 25, pp. 714-21, Mar 15 2009.
[77] B. Han, *et al.*, "A Markov blanket-based method for detecting causal SNPs in GWAS," *BMC Bioinformatics,* vol. 11 Suppl 3, p. S5, 2010.
[78] X.-W. Chen, *et al.*, "Improving Bayesian Network Structure Learning with Mutual Information-Based Node Ordering in the K2 Algorithm," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 20, pp. 628-640, May 2008.
[79] P. Spirtes, *et al.*, *Causation, prediction, and search*, 2nd ed. Cambridge, Mass.: MIT Press, 2000.
[80] C. F. Aliferis, *et al.*, "HITON: a novel Markov Blanket algorithm for optimal variable selection," *AMIA Annu Symp Proc,* pp. 21-5, 2003.
[81] R. R. Sokal and F. J. Rohlf, *Biometry : the principles and practice of statistics in biological research*, 3rd ed. New York: Freeman, 1995.
[82] J. H. McDonald, *Handbook of Biological Statistics* 2nd ed. Baltimore, Maryland: Sparky House Publishing, 2009.
[83] D. Koller and M. Sahami, "Toward Optimal Feature Selection," presented at the 13th conference on machine learning, Bari, Italy, 1996.
[84] D. Margaritis and S. Thrun, "Bayesian Network Induction via Local Neighborhoods," presented at the Neural Information Processing Systems 12, Denver, Colorado, USA, 1999.
[85] I. Tsamardinos, *et al.*, "Algorithms for Large Scale Markov Blanket Discovery," presented at the The 16th International FLAIRS Conference, St. Augustine, FL, 2003.

[86] I. Tsamardinos, *et al.*, "Time and Sample Efficient Discovery of Markov Blankets And Direct Causal Relations," presented at the The ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003.
[87] J. M. Peña, *et al.*, "Towards scalable and data efficient learning of Markov boundaries," *International Journal of Approximate Reasoning,* vol. 45, pp. 211-232, 2006.
[88] J. Pearl, *Probabilistic reasoning in intelligent systems : networks of plausible inference*. San Mateo, Calif.: Morgan Kaufmann Publishers, 1988.
[89] J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: models and data," *Am J Hum Genet,* vol. 69, pp. 1-14, Jul 2001.
[90] J. Li and Y. Chen, "Generating samples for association studies based on HapMap data," *BMC Bioinformatics,* vol. 9, p. 44, 2008.
[91] C.-c. Chang and C.-J. Lin. (2001, LIBSVM: A library for support vector machines. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm
[92] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning,* vol. 53, pp. 23-69, 2003.
[93] X. W. Chen, *et al.*, "An effective structure learning method for constructing gene networks," *Bioinformatics,* vol. 22, pp. 1367-74, Jun 1 2006.
[94] J. Cheng, *et al.*, "Learning Bayesian networks from data: an information-theory based approach," *Artif. Intell.,* vol. 137, pp. 43-90, 2002.
[95] J. Pearl, *Causality : models, reasoning, and inference*, 2nd ed. Cambridge, U.K. ; New York: Cambridge University Press, 2009.
[96] D. Heckerman, *et al.*, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Mach. Learn.,* vol. 20, pp. 197-243, 1995.
[97] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control,* vol. 19, pp. 716-723, 1974.
[98] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics,* vol. 6, pp. 461-464, 1978.
[99] J. Rissanen, "Stochastic Complexity and Modeling," *The Annals of Statistics,* vol. 14, pp. 1080-1100, 1986.
[100] G. F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Mach. Learn.,* vol. 9, pp. 309-347, 1992.
[101] K. P. Burnham, *Model selection and multimodel inference : a practical information-theoretic approach*, 2nd ed. New York: Springer, 2002.
[102] B. Han and X. W. Chen, "bNEAT: a Bayesian network method for detecting epistatic interactions in genome-wide association studies," *BMC Genomics,* vol. 12 Suppl 2, p. S9, 2011.
[103] L. M. d. 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.
[104] X. Shen and J. Ye, "Adaptive Model Selection," *Journal of the American Statistical Association,* vol. 97, pp. 210-221, 2002.

[105] T. Hastie, *et al.*, *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer, 2001.
[106] D. M. Chickering, *et al.*, "Large-Sample Learning of Bayesian Networks is NP-Hard," *J. Mach. Learn. Res.,* vol. 5, pp. 1287-1330, 2004.
[107] J. Suzuki, "Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: An Efficient Algorithm Using the B & B Technique," presented at the 13th International Conference on Machine Learning Bari, Italy, 1996.
[108] J. Tian, "A Branch-and-Bound Algorithm for MDL Learning Bayesian Networks," presented at the Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, 2000.
[109] X.-w. Chen, "An improved branch and bound algorithm for feature selection," *Pattern Recogn. Lett. ,* vol. 24, pp. 1925-1933, 2003.
[110] M. L. Wong, *et al.*, "Using Evolutionary Programming and Minimum Description Length Principle for Data Mining of Bayesian Networks," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 21, pp. 174-178, 1999.
[111] P. Giudici and R. Castelo, "Improving Markov Chain Monte Carlo Model Search for Data Mining," *Machine learning,* vol. 50, pp. 127-158, 2003.
[112] R. J. Klein, *et al.*, "Complement factor H polymorphism in age-related macular degeneration," *Science,* vol. 308, pp. 385-9, Apr 15 2005.
[113] A. Hamosh, *et al.*, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res,* vol. 30, pp. 52-5, Jan 1 2002.
[114] S. T. Sherry, *et al.*, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res,* vol. 29, pp. 308-11, Jan 1 2001.
[115] E. S. Tobias, *et al.*, "The TES gene at 7q31.1 is methylated in tumours and encodes a novel growth-suppressing LIM domain protein," *Oncogene,* vol. 20, pp. 2844-53, May 17 2001.
[116] S. J. Bowne, *et al.*, "Mutations in the inosine monophosphate dehydrogenase 1 gene (IMPDH1) cause the RP10 form of autosomal dominant retinitis pigmentosa," *Hum Mol Genet,* vol. 11, pp. 559-68, Mar 1 2002.
[117] K. Nikopoulos, *et al.*, "Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy," *Am J Hum Genet,* vol. 86, pp. 240-7, Feb 12 2010.
[118] E. M. Reiman, *et al.*, "GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers," *Neuron,* vol. 54, pp. 713-20, Jun 7 2007.
[119] L. Bertram, *et al.*, "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database," *Nat Genet,* vol. 39, pp. 17-23, Jan 2007.