2007

Comparative Microarray Data Mining

Shihong Mao
Wright State University

Comparative Microarray Data Mining

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

By

SHIHONG MAO

M.S., Wright State University, 2001

2007
Wright State University

Wright State University
SCHOOL OF GRADUATE STUDIES

Nov 30th, 2007

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Shihong Mao ENTITLED Comparative Microarray Data Mining BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Guozhu Dong, Ph.D.
Dissertation Director

Thomas A. Sudkamp, Ph.D.
Director, Ph.D. Program of CS&E

Joseph F. Thomas, Jr., Ph.D.
Dean, School of Graduate Studies

Committee on Final Examination:

Guozhu Dong, Ph.D.

Michael L. Raymer, Ph.D.

Mateen M. Rizki, Ph.D.

F. Javier Alvarez-Leefmans, Ph.D., M.D.

Dale E. Courte, Ph.D.

ABSTRACT

Mao, Shihong. Ph.D., Department of Computer Science and Engineering, Wright State University, 2007. Comparative Microarray Data Mining.

As a revolutionary technology, microarrays have great potential to provide genome-wide patterns of gene expression, to make accurate medical diagnosis, and to explore genetic causes underlying diseases. It is commonly believed that suitable analysis of microarray datasets can lead to achieve the above goals. While much has been done in microarray data mining, few previous studies, if any, focused on multiple datasets at the comparative level. This dissertation aims to fill this gap by developing tools and methods for set-based comparative microarray data mining. Specifically, we mine highly differentiative gene groups (HDGGs) from given datasets/classes, evaluate the concordance of datasets generated from different platforms/laboratories, investigate the impact of variability in microarray dataset on data mining results, provide tools and algorithms for the above tasks, and identify reliable invariant HDGG patterns for better understanding diseases.

It is a big challenge to discover high-quality discriminating (emerging) patterns from high dimensional microarray datasets. We develop a novel feature-group selection method to help discover HDGGs, especially signature HDGGs that completely characterize some disease classes. In addition to giving insights on the diseases, better classification results are also obtained using HDGG-based classifiers compared with other existing classifiers.

As microarray datasets are often generated from different platforms/laboratories, it is necessary to evaluate their concordance/consistence before they can be studied together. We provide measures and techniques to quantitatively test such concordance at the

comparative level.

In addition to applying measures to evaluate the degree of variability in microarray datasets, we also develop a novel algorithm called C-loocv to effectively minimize the variability. As an indicator of the utility of C-loocv, classifiers trained from C-loocv-refined datasets become more robust and predict test samples at significantly higher accuracy over classifiers trained from original datasets.

Based on the variability minimization algorithm, we provide a novel strategy to mine invariant patterns from multiple datasets concerning a common disease. As a demonstration, invariant patterns are identified from two datasets concerning lung cancer; these patterns may shed light on the mechanism underlying the pathogenesis of lung cancer. Our methods are generic and can be applied to microarrays concerning any human diseases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to thank my advisor Guozhu Dong, for his academic guidance and support during all these years of my Ph. D. program. I am deeply grateful for his mentoring, advice and research support. His stimulating suggestions and encouragement have helped me during this research.

I would like also to express my warm and sincere thanks to the other members of my dissertation committee: Dr Michael Raymer, Dr. Matt Rizki, Dr. Francisco Alvarez-Leefmans, and Dr. Courte Dale for their precious time in reviewing the manuscript and their valuable suggestions. I am also grateful for their encouragement and support on my research and studies.

A special thank to Dr. Alvarez-Leefmans for providing me an opportunity working in his lab. During the past years there, I accumulated strong background in neuroscience research, and have successfully finished several projects with other members. I believe this valuable experience will help me a lot in my future career.

The comments and suggestions offered by all other members of Dr. Dong's laboratory with whom I have worked are also greatly appreciated; in particular, I would like to thank my fellow students Lijun Chen, Chunyu Jiang and Yin Sun.

Finally, I would like to thank all my family members and friends who have given me encouragement and support which make this thesis possible.

DEDICATION

To my wife Zhihui who encourages and supports me to overcome challenges, brings me love and happiness, and shares every moment of joy and sorrow with me…

To our first baby, Angel, who brought us lots of happiness and is blessing us now in heaven; and to our second baby, who is coming on his way and brings us bright hope…

To my parents & parents-in-law, who give me incredible inspiration and unconditional love and support for all these years…

To my sisters Yachun and Yahong, who have been tremendous supportive during my difficult times and helped me go through them…

Chapter 1: INTRODUCTION

1.1. MOTIVATION – THE PROSPECT OF MICROARRAY GENE EXPRESSION DATA

Microarray technology allows the measuring of the expression level of thousands of genes simultaneously by using gene chips. This technology provides the possibility of creating datasets that capture the information concerning all the relevant genes and proteins for many systems of biological and clinical interest. Such datasets may help scientists explore gene expression patterns and discover gene interaction networks, and perhaps even pathways underlying various diseases and biological processes. Such discoveries can in turn lead to better understanding of the physiological functions in healthy and diseased cells, and to better ways to diagnose and treat diseases. Large-scale transcription analyses using microarrays can reveal the molecular mechanisms of physiology and pathogenesis, and therefore can help scientists to develop new diagnostic and therapeutic strategies.

Recently, microarray (DNA chip) technology is becoming a very important and powerful tool in almost every field of biomedical research. This technology has been used in reproductive medical research including study of oocyte fertilization, early embryo development, implantation and some infertility-related diseases such as endometriosis and myoma (Chen et al, 2006). Microarray also brings new insights into evolutionary

biology by providing genome-wide patterns of gene expression within and between species (Ranz et al, 2006).

For cancer study, a number of cancer-related datasets such as colon cancer (Alon 1999), acute lymphoblastic leukemia (Golub 1999), breast cancer (van't Veer et al 2002) and so on have been successfully generated using microarray technology in the past decade. These datasets have also been widely studied by researchers in various fields, and the analysis results have provided valuable information for biomedical, pharmaceutical, and clinical research. In the present dissertation several public cancer-related microarray datasets were employed for data mining. The long-term goal of this project is to provide useful tools and information aiming at enhancing our understanding of diseases by developing comparative data mining methods.

Specifically, we develop several novel comparative data mining approaches on microarray gene expression data, aiming at extracting reliable patterns from microarray datasets. We propose that comparative data mining of microarray data has the potential to discover key groups of genes in cancer, which will help us to better understand its patho-physiology. Moreover, this approach will help uncovering new therapeutic targets for diseases, predicting how patients respond to specific treatments, and revealing possible regulatory relationships among genes in normal and disease situations.

1.2. OVERVIEW OF RESULTS

In this dissertation, we provide measures, tools and methodologies for the above mentioned data mining approaches. Below is a summary of the research projects we have been working on:

1. Identification of highly differentiative gene groups (HDGGs) from microarray data. The aim is to introduce methods that could do a better job given the high dimensionality challenge. We combine a new approach (gene club formation) with previous data mining algorithms for feature selection and discovering emerging (discriminating) patterns. The HDGGs mined from each dataset are considered as discriminative characteristic patterns and are important features for each specific dataset.

2. Multi-source microarray platform concordance detection. The aim is to provide measures and techniques to compare microarray gene expression data generated from different platforms and laboratories. Microarray datasets are generated from different platforms / laboratories. It is necessary to evaluate the concordance and consistence of the multi-source microarray datasets before they can be applied to clinical, pharmaceutical research and other purposes. Since no comparative methods have been applied to test the concordance of multi-source microarray datasets, we generated several comparative methods to test such data and measure their concordance with each other.

3. Minimization of microarray dataset variability. The aim is to evaluate the effect of variability on data mining results and to provide novel methods to minimize it. The inevitable variability in microarray datasets leads to less reliability and accuracy of mined patterns and models. We develop a novel method to minimize variability by eliminating highly noisy samples from datasets.

4. Identification of invariant patterns. The aim is to mine reliable patterns thought to play key roles in the pathogenesis of diseases. We provide a novel method to mine invariant patterns from multiple datasets related to a disease in particular. The invariant patterns (shared gene interactions) are considered as more reliable patterns and expected to

provide useful information for potential gene pathways for the diseases under consideration.

1.3. ORGANIZATION OF THE DISSERTATION

This dissertation deals with comparative microarray data mining, and is divided in eight chapters. Chapter 1 is an introduction chapter. Chapter 2 gives an overview of preliminary information on gene expression data, and a discussion of some important concepts which will be applied in other chapters, such as high dimension, high variability microarray data, emerging patterns, entropy-based discretization method and information gain. Chapter 3 provides a review of the literature on topics related to the present research. This chapter discusses what have been done by previous researches, the gaps that exist between current research and our research goals, and what we want to do in order to fill these gaps.

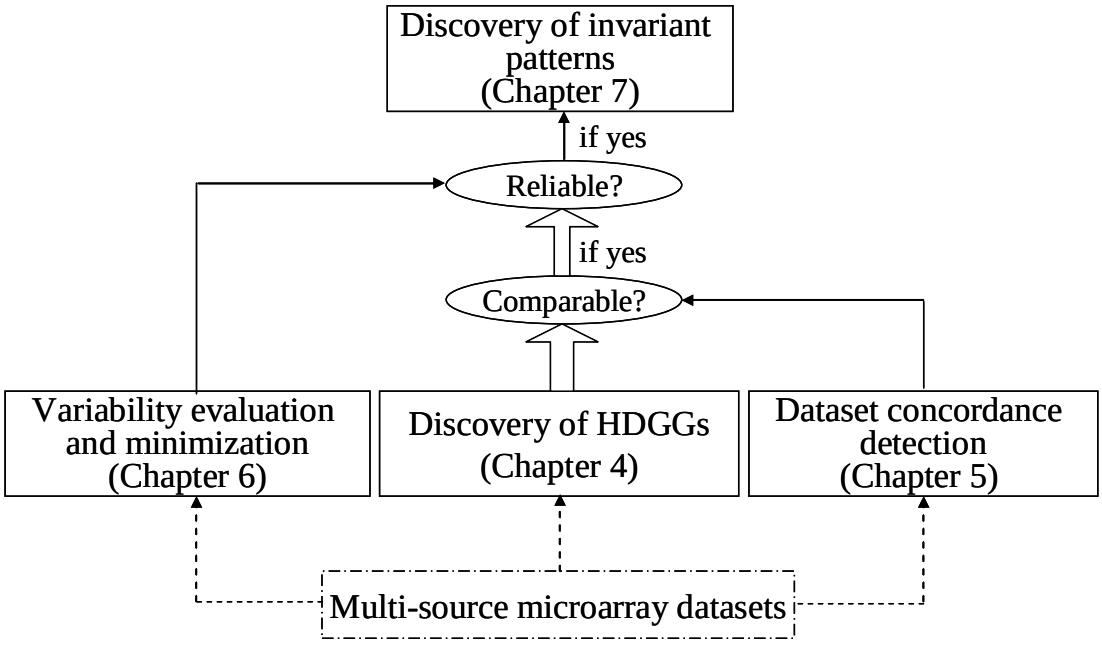

Figure 1.1: Outline of topics studied in this dissertation.

Chapters 4 to 7 comprise the results from the scientific research accomplished during the past 5 years. Chapter 4 introduces the discovery and application of HDGGs. Chapter 5 deals with the detection of concordance of microarray datasets generated from different platforms / laboratories. Chapter 6 presents our investigation on and the minimization of biological variation in microarray datasets. Chapter 7 discusses the discovery of invariant patterns from multi-microarray datasets. Chapter 8 is a summary of our work and future directions.

The relationship between the topics considered in this dissertation is illustrated in Figure 1.1.

Chapter 2: PRELIMINARIES

This chapter deals with background information regarding the techniques and terminology that will be used throughout this dissertation.

2.1. MICROARRAY GENE EXPRESSION DATA

With the development of microarray technology many kinds of microarray data have been generated such as: DNA microarrays, protein microarrays, tissue microarrays, cell microarrays, chemical compound microarrays and antibody microarrays. The most commonly used ones are DNA microarrays. In this dissertation, “microarrays” refers to “DNA microarrays”.

A DNA microarray (also commonly known as a gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots, each representing a gene probe, arrayed on a solid surface by covalent attachment to chemically suitable matrices. DNA arrays differ from other types of microarrays only in that they either measure DNA or use DNA as part of its detection system. Qualitative or quantitative measurements with DNA microarrays utilize the selective nature of DNA-DNA or DNA-RNA hybridization under high-stringency conditions and use fluorophore-based detection. DNA microarrays are commonly used for expression profiling, i.e., monitoring gene expression levels of thousands of genes simultaneously to determine whether those genes are active, hyperactive or silent in given tissues.

Figure 2.1 shows the step by step procedure to convert each gene's expression level into a real numeric value data when using microarrays. 1) A microarray chip is labeled with gene probes (left column). 2) Total mRNA is extracted from a given sample and labeled with the corresponding fluorophore, e.g. cy3 or cy5 (middle column). 3) The microarray is hybridized with labeled mRNA (middle). 4) The microarray is scanned, filtered and raw data is generated (right column). 5) After data preprocessing and normalization, the data generated can be used for further analysis and mining. Table 2.1 is an example of microarray gene expression data with three genes and six tissues (three from normal control, and three from diseased patients).

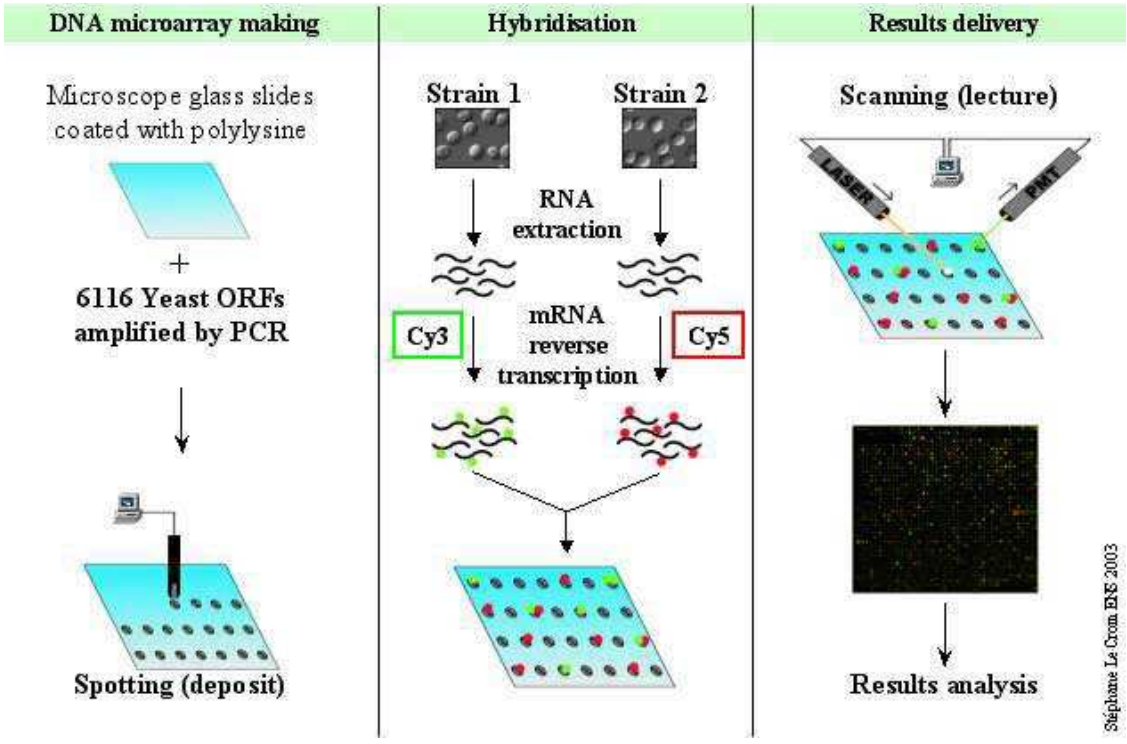

Figure 2.1: Microarray technology pipeline

ORF: open reading frame, PCR: polymerase chain reaction

Table 2.1: A sample microarray gene expression dataset

	d_1	d_2	d_3	N_1	n_2	n_3
g_1	32.0	52.3	89.3	51.1	29.7	4.5
g_2	5.9	1.7	2.6	10.0	3.2	9.1
g_3	94.4	132.7	180	73.4	55.8	120.6

d_1, d_2, d_3: diseased tissues; n_1, n_2, n_3: normal tissues; g_1, g_2, g_3: genes.

Currently microarray data are generated using various different platforms such as commercial platforms (Affymetrix, Agilent, Applied Biosystems, etc) or custom-made ones. The platforms mainly differ on what and how the gene probes are labeled on the microarray chips. For example, the probes on the Agilent platform are cDNA, which is reverse-transcribed from known mRNA. The probes can perfectly hybridize with their corresponding mRNA, although the quality of the probes may greatly affect the hybridizing results. Affymetrix is probably the most popular commercial platform so far. The probes are designed based on gene sequence analysis. Each probe on the chip is designed as one pair of oligonucleotide about 35 mer in length. One oligonucleotide is the Perfect-Match (PM), and the other one is a one-base pair Mis-Match (MM), with the corresponding gene. This probe design can effectively reduce noise in microarray data.

2.2. CHARACTERISTICS OF MICROARRAY DATA

Compared with commercial datasets, microarray gene expression datasets have quite different characteristics. First, their dimensionality, i.e. the number of features explored in gene chip, is high. The raw microarray images are transformed into gene expression matrices where the rows usually denote genes or features and the columns denote various

samples, conditions, tissues or instances. The number of features (dimensions) can be very high. Usually, there are usually thousands of gene probes in one gene chip. If a gene chip is designed to detect all genes in a human tissue sample, the number of probes may exceed 100,000.

Second, the number of samples may be small, compared with typical commercial applications. For many biomedical and pathology studies, the number is usually less than 200.

Third, microarray datasets may be very noisy containing unreliable or contaminated values. In sum, there is high variability in microarray datasets. The variability includes that inherent to measurement procedures (measurement variability) and biological variation. Measurement variability results from differences in experimental conditions, procedures and differences in microarray technologies. Biological variation is due to intrinsic characteristics of the samples.

Table 2.2: An ideal microarray gene expression dataset

	Class A			Class B		
	s_1	…	s_x	s_1	…	s_y
g_1	a_1	…	a_1	b_1	…	b_1
g_2	a_2	…	a_2	b_2	…	b_2
…	…	…	…	…	…	…
g_n	a_n	…	a_n	b_n	…	b_n

Ideally, if there is no variability, each gene's expression value across the samples within one class should be identical. Table 2.2 shows a sample microarray gene expression dataset without variability. There are two classes A and B in this dataset. There are x samples (s_1 to s_x) in A and y samples (s_1 to s_y) in B. Total n gene probes (g_1 to g_n) are used to construct the microarray chip. a_i and b_i are gene g_i's standard values in class A

and B respectively. Under the ideal situation, g_i's real value in any samples in one class should be equal to its standard value. In this case, data mining results from this dataset will be 100% accurate and reliable. Unfortunately this kind of microarray dataset never exists in reality.

Table 2.3 shows microarray dataset with measurement variability. Each gene's detected value a_{ij} and b_{ik} ($1 \leq i \leq n$, $1 \leq j \leq x$, $1 \leq k \leq y$) is shifted from its standard value by a random value α_{ij} or β_{ik}. If α_{ij} or β_{ik} is large, then the data have high variability.

Table 2.3: Microarray gene expression dataset with only measurement variability

	Class A			Class B		
	s_1	…	s_x	s_1	…	s_y
g_1	$a_1 \pm \alpha_{11}$	…	$a_1 \pm \alpha_{1x}$	$b_1 \pm \beta_{11}$	…	$b_1 \pm \beta_{1y}$
g_2	$a_2 \pm \alpha_{21}$	…	$a_2 \pm \alpha_{2x}$	$b_2 \pm \beta_{21}$	…	$b_2 \pm \beta_{2y}$
…	…	…	…	…	…	…
g_n	$a_n \pm \alpha_{n1}$	…	$a_n \pm \alpha_{nx}$	$b_n \pm \beta_{n1}$	…	$b_n \pm \beta_{ny}$

As mentioned before, besides measurement variability, there is also intrinsic biological variation in microarray datasets. Microarray datasets are typically generated using tissue samples from different patients. These patients have different characteristics such as height, weight, age, race, and sex. These differences inevitably lead to biological variations in microarray datasets. Because of biological variation, gene g_i's standard values (a_i or b_i) can not be measured. In reality, in different samples, the "real" values of g_i are different. For example, in class A, g_i's real value at sample j is a_{ij} and g_i's detected value at sample j will be $a_{ij} \pm \alpha_{ij}$. Theoretically, g_i's real value should be close to its standard value within one class; however, in real microarray dataset, some samples have

big biological variation because the genes' real values in these samples are far away from their standard value.

Table 2.4: Actual microarray gene expression dataset

	Class A			Class B		
	s_1	…	s_x	s_1	…	s_y
g_1	$a_{11} \pm \alpha_{11}$	…	$a_{1x} \pm \alpha_{1x}$	$b_{11} \pm \beta_{11}$	…	$b_{1y} \pm \beta_{1y}$
g_2	$a_{21} \pm \alpha_{21}$	…	$a_{2x} \pm \alpha_{2x}$	$b_{21} \pm \beta_{21}$	…	$b_{2y} \pm \beta_{2y}$
…	…	…	…	…	…	…
g_n	$a_{n1} \pm \alpha_{n1}$	…	$a_{nx} \pm \alpha_{nx}$	$b_{n1} \pm \beta_{n1}$	…	$b_{ny} \pm \beta_{ny}$

Table 2.4 shows a realistic microarray dataset. Both measurement variability and biological variation are included. The high variability may affect the reliability of microarray datasets. During microarray data analysis and data mining, the high variability should be considered.

2.3. EMERGING PATTERNS AND BORDER DIFFERENTIAL ALGORITHM

Emerging patterns (EPs) (Dong et al., 1999; Dong et al., 2005) are defined as patterns whose supports increase significantly from one class to another. EPs with growth rate (defined as the ratio of frequency between the classes) of infinity are called jumping EP (JEP), i.e. these patterns appear in one class but never exist in other classes. EPs have been proved to be very useful as a means of discovering distinctions inherently present between different classes of data. For example, by using emerging patterns, Li and

colleagues (Li et al., 2002, Li et al., 2003a, Li et al., 2003b) identified good diagnostic genes or gene groups from gene expression data of the Acute Lymphoblastic Leukemia vs Acute Myeloid Leukemia (ALL/AML) dataset (Golub et al., 1999) and other datasets such as colon cancer (Alon et al., 1999) and ALL (Yeoh et al., 2002).

To efficiently discover the jumping emerging patterns (JEPs) with respect to a positive dataset and a negative dataset, border manipulation algorithm was proposed (Dong et al., 1999). The border differential algorithm is the core subroutine for JEP mining and it aims to derive the difference between a pair of border of a special form. The border based algorithm achieves high efficiency by manipulating only the itemsets in the borders and avoiding the tedious process of enumerating all the individual JEPs.

2.4. ENTROPY-BASED DISCRETIZATION METHOD AND INFORMATION GAIN

Microarray gene expression data are always continuous and contain a large number of genes. Such data should be pre-processed, through operations including binning, duplicate gene removal, and gene ranking. Binning transforms continuous features into discrete features. The entropy based method (Dougherty et al., 1995) is often applied to convert the values for each gene into two intervals (bins) and to rank the genes. (More bins can be allowed, but in this dissertation we only consider two bins.) One bin will be called “high” and the other “low”. Let S be the set of all tuples and T(S, Cj) be the proportion of tuples in S that have class Cj. The entropy for S is:

$$Entropy(S) = -\sum_{j=1}^{2} T(S, C_j) * \log(T(S, C_j))$$

Let g be a gene. Each value v can divide g's values into two intervals, namely g ≤ v and g > v; let S_1 (resp., S_2) be the set of tuples in S where g's values are ≤ v (resp., > v). We define Entropy(Si) similarly as above. The class information for gene g at partition point v is

$$I(g,v) = \frac{|S_1|}{|S|} * Entropy(S_1) + \frac{|S_2|}{|S|} * Entropy(S_2)$$

where |S| denotes the cardinality of S. The information gain for g at partition point v is:

$$InfoGain(g,v) = Entropy(S) - I(g,v)$$

The value v for which InfoGain(g,v) is maximal amongst all the candidate split points is selected as the split point for g. Let InfoGain(g) denote that maximal InfoGain(g, v).

Information gain for a gene captures how strong the gene is related to the class; the larger the information gain, the stronger the relationship. We will usually rank the genes in decreasing information gain order.

Chapter 3: LITERATURE REVIEW

The main goal of the present dissertation is to find new ways to analyze microarray gene expression datasets so that we can mine more information out of them and hopefully provide valuable clues for biological and medical research. One of the strength of this research is to compare multiple datasets. In addition, microarray analysis and data mining have potential value as a diagnostic and predictive tool in various researches.

This chapter comprises a survey of the work related to our study, identify the gaps between previous studies and our research goals, and briefly introduce what we are going to do to fill those gaps. The main topics include comparative studies and data mining, feature selection, microarray data concordance detection, biological variation, instance selection and classification.

3.1. COMPARATIVE STUDIES AND DATA MINING

Comparative studies, aimed at comparing the similarity/difference between groups using comparative methods, have been applied to many fields, such as genomics, gene function comparison, text files and microarray datasets. In the field of genomics, comparative study, commonly named comparative genomics, is used to study the similarities and differences in the structure and function of hereditary information across species. This approach is used to compare genomes in genomics comparative data analysis. Recently, the availability of sequences from numerous biological species has allowed multiple

species-comparisons for identifying the relationships between species (Hood et al 1995, Dubchak et al 2000, Pennacchio et al 2001, Gottgens et al 2002). The utility of comparative sequence analysis is based on the hypothesis that important biological sequences are conserved between species due to functional constraints.

A number of recent comparative genomics studies, such as the evolutionary distance comparison between human-mice, human-birds, human-fish and human-primate, have yielded the identification of functional sequences solely through the use of genomic comparison. Enormous advances, such as the inference of function of new sequences through similarity to known sequences, have been made (Boffelli et al, 2003, Kappen et al, 2003, Harris et al, 2003, Nobrega et al, 2003, Postlethwait et al, 2000, OBrien et al, 1999).

Similarly, comparative studies have been applied on gene function comparison. Lin and colleagues (Lin et al., 2002) proposed a new concept called “functional genomic units”, which is a group of genes carrying out some common biological functions. They described an interesting attempt to use the Rosetta dataset (cDNA platform) to corroborate a Rac1 transfection obtained from Affymetrix platform.

When data mining approaches are applied to compare different datasets / classes and to find similarity patterns and unique patterns between these datasets, then the comparative study is called comparative data mining. Comparative data mining has been applied on text files, which is named as comparative text mining. Zhai and colleagues (Zhai et al, 2004) proposed a generative probabilistic mixture model for comparative text mining. By using this model, they discovered certain latent common themes across all collections and summarized the similarities and differences of these collections along each common

theme. This approach could be adopted to study microarrays to find common /unique patterns from mutli-datasets.

If comparative data mining is applied to study microarray datasets, we call it "comparative microarray gene expression data mining". This term is first mentioned in this dissertation and will be its main focus. Comparative data mining is defined as the mining of similarities and differences/contrasts among multiple data classes or multiple datasets (each with or without classes).

According to this definition, few papers, if any, can be considered to fall within the categories of comparative microarray data mining. One paper mentioned comparative data mining on microarrays (Page et al 2002). The authors used comparative data mining experiments to compare various classification methods and to identify the advantages of the leading supervised learning algorithms for microarray data. It should be pointed out that comparing multiple methods on one dataset will not be considered as comparative data mining, in the sense defined above.

Many data mining methods focus on feature selection and classification methods by using microarray data (Li et al 2004). Strictly speaking, building classifiers can not be considered as comparative data mining, although it is considering two or more classes and hence it can be loosely considered as comparative mining. Other studies use comparative methods to analyze microarray data (Xing et al 2001). In their studies, they use statistical methods other than data mining for microarray data study.

So far, many methods have been proposed for the analysis of microarray data. In general, some of these methods were borrowed from data mining in other areas and ignored the intrinsic biological features of microarray data. Most importantly, few of these methods

considered the high variability in microarray datasets or the concordance between multiple datasets, which may greatly affect the reliability of data mining results. The present dissertation addresses these fundamental issues by using comparative microarray data mining.

3.2. FEATURE SELECTION

Feature selection aims at selecting a subset of relevant features for building robust learning models. It has been extensively studied in machine learning. Feature selection helps to improve the performance of learning models by enhancing generalization capability; speeding up learning processes; and improving model interpretability. In microarray studies, feature selection is called discriminative gene selection. It selects the influential genes based on their ability to distinguish between various classes of samples, such as between different types of diseases or between diseased and healthy states. Thus, feature selection can help to better understand microarray data, and tell which genes are important and how they are related with each other.

One characteristic of microarray data is their considerable number of features (genes). Among these features, not all of them carry relevant information for a particular application. It is necessary to use feature selection to select the most important components. From a biological perspective, the most common situation is a group of genes work together rather than a single gene in the genesis of a disease (Cunliffe et al., 2003, Califano et al., 2000, Segal et al., 2003). Thus, feature selection is a useful tool to detect groups of genes from microarray datasets, by considering genes' interaction with other genes.

In gene expression analysis studies, many gene (feature) selection methods have been developed. These methods include information gain, towing rule, sum minority, max minority, Gini index, sum of variances, one-dimensional SVM, t–statistics, the ratio of between-groups to within-groups sum of squares (BSS/ WSS), principal component analysis among others (Su et al., 2003). Information gain and Gini index are widely used in machine learning. Towing rule, max minority, sum minority and sum of variances are broadly applied in statistical learning theory. In these methods, the full range of expression of a given gene is split into two regions: high or low. Then the strength of this given gene with respect to the class is evaluated. One-dimensional SVM (Brown et al., 2000; Ramaswamy et al., 2001) measures the effectiveness of a future by calculating the accuracy of single-feature SVM classifiers. t -statistics was first used by Golub and colleagues to measure the class predictability of genes for two-class problems (Golub et al., 1999).

In many of the methods mentioned above, genes are typically grouped by similarity of their expression profiles. We suggest a different approach in which genes are grouped together when correlation of their expression profiles in one state is destroyed in another state. The correlation considered here is general (e.g., one gene is high whenever another is low). A highly differentiative gene group (HDGG) is a gene group in which genes are correlated with each other. A HDGG captures the following information: in the normal state, some genes are correlated (perhaps because they participate in some common pathway under normal situation) and are fully “in sync”, but in a disease state these genes are no longer “in sync” (perhaps because the pathway is disrupted).

Since microarray data have thousands of dimensions, discovery of gene groups, such as

HDGGs, is a big challenge. Exhaustive search is impossible, since the required search time grows exponentially with the number of dimensions. Li and colleagues (Li et al., 2001) use the top-k approach to get around the dimensionality hurdle by selecting the globally top-k genes in decreasing information gain (Dougherty et al.,1995) order and then applying the border differential algorithm (Dong et al., 1999, Dong et al., 2005) on these genes. The border differential algorithm can effectively handle up to 75 genes for current generation PCs, but can not finish in reasonable amount of time when higher dimensions are present. As will be shown in Chapter 4, the best HDGGs may contain genes which are very low in information gain rank, and may be missed by the top-k method. We aim to introduce methods that are more effective given the dimensionality challenge.

In order to overcome the high dimension and find HDGGs, we establish gene clubs for each given gene by using several methods that will be discussed later in this dissertation. Although the determination of gene clubs shares similarities with traditional feature selection, it has some new characteristics and is based on the interaction among genes.

3.3. MICROARRAY DATA CONCORDANCE DETECTION

Microarray technology allows simultaneous measurements of mRNA expression of thousands of genes. This technology provides the possibility of creating datasets that capture the information concerning all the relevant genes and proteins for many systems of biological or clinical interest. Such datasets are useful because they may help scientists to discover gene interaction networks, and perhaps even pathways underlying various diseases and biological processes. Such discoveries can in turn lead to better

understanding of the biological processes and diseases, and to better ways to diagnose and treat diseases.

Recent advances in microarray technology have generated large amounts of gene expression data, collected using a variety of commercial platforms from different laboratories. The concordance/consistency of the datasets from different sources should be evaluated before this technology can be successfully and reliably applied in biological/clinical practice and regulatory decision-making (Shi et al., 2004, Hackett et al., 2003, Petricoin et al., 2002b). This need was also recognized in recent publications addressing some possible factors affecting the consistency of DNA microarrays (Guo et al, 2006, Shi et al, 2006).

As mentioned in Section 2.2, two major factors, experimental noise and biological variation may cause inconsistent results in repeated experiments during data generation in microarray experiments. Experimental noise can be caused by differences in probe labeling efficiency, RNA concentration or hybridization efficiency, image analysis and so on. All types of such noise might make the experiments unrepeatable. As a result, the expression levels reported by a microarray experiment might not exactly reflect the true gene expression levels. Biological variation will be discussed in next section.

There has been wide interest in the intra- and inter-platform comparisons of gene expression values (Kuo et al., 2002, Tan et al., 2003b, Hardiman, 2004, Shi et al., 2005, Guo et al, 2006, Shi et al 2006). These studies reached different conclusions: On one hand, some cross platform comparisons reported a failure to demonstrate an acceptable level of correlation between different microarray technologies (Tan et al., 2003b, Jarvinen et al., 2004, Woo et al., 2004, Yauk et al., 2004, Mah et al., 2004, Cicatiello et

al., 2003, Marshall 2004); the authors conjectured that the difficulties in correlating data may be attributed to fundamental differences between cDNA and oligonucleotide based microarray technologies. Other studies concluded that low inter-platform consistency is due to other reasons instead of inherent technical differences among different platforms (Shi et al. 2005). Recent studies related to microarray quality control (MAQC) project showed that the microarray data from different platforms are fairly concordant (Guo et al 2006, Shi et al 2006).

It is not easy to tell how concordant two datasets are. Different criteria (testing methods) may give different answers. Most studies have focused on the expression values of individual genes. They are not applicable for comparative studies, where one compares one class of data against another class. Comparative microarray analysis can better distinguish phenotypes from related phenotypes; identify valid differentially expressed genes by combining many studies; test new hypothesis; and discover fundamental patterns of gene regulation. In order to get reliable results when using comparative methods, it is desirable to test the datasets' concordance using the same comparative methods.

The present work introduces novel comparative methods for evaluating concordance of microarray data collected from different laboratories and/or different platforms. These methods evaluate concordance by measuring quality preservation of discriminating genes and classifiers. Considering that microarray datasets are generated from different platforms, if the microarray datasets are concordant with each other with respect to discriminating genes, then the knowledge on discriminating genes gained from one platform/lab can be transferred to another platform/lab (Mao & Dong et al, in

preparation).

3.4. BIOLOGICAL VARIATION

As mentioned in Section 2.2, two major factors: experimental noise and biological variation, may cause inconsistent results in repeated experiments. Experimental noise can arise at any step of microarray experiments. It might render the experiment non-reproducible. As a result, the expression levels reported by a microarray experiment might not exactly reflect the true mRNA levels. Biological variation refers to the natural variation we would expect to encounter even under ideal experimental conditions. In other words, even if we could sidestep experimental issues, magically looking inside the cells and counting the mRNA molecules of interest, we would still expect some variation in counts between cells in the same category (Piatetsky_Shapiro et al, 2003).

Different from experimental noise, biological variation inevitably exists in microarray dataset because of the variety among tissue samples (e.g. patients). Biological variation may also affect the accuracy of data analysis and may lead to unreliable results. So far, there are few, if any, studies to investigate how biological variability affects to data mining results and how to reduce it in microarray dataset. Liu and colleagues (Liu et al, 2003) selected samples according to the patient's surviving time. This selection can be applied to some specific datasets but not to all microarray datasets.

In order to mine reliable patterns, the biological variability in microarray datasets needs to be considered. In this study, first we investigate the influence of variability on our comparative study and then we provide methods to minimize it.

3.5. INSTANCE SELECTION

Instance selection aims to search for a representative data subset that replaces the original dataset, still solving a data mining task as if the whole dataset were used. Finding a small set of representative instances for large datasets can bring various benefits to data mining practitioners (Zhu et al 2006): 1) build a learner superior to the one constructed from the whole massive data; 2) avoid working on the whole original dataset all the time; and 3) remove irrelevant instances as well as noise and/or redundant data. For most data mining tasks, such as classification and clustering tasks, the selected dataset should preferably exclude noisy instances. Many instance selection algorithms have been developed so far.

Sampling: Sampling, a basic instance selection, is a well established statistical technique that selects a part from a whole to make inferences about the whole, which is applied to overcome problems caused by high attribute dimensionality as well as large data volumes in data mining. It can profitably used to estimate characteristics of a population of interest with less cost, hihger speed, greater scope and probably greater accuracy compared to a complete enumeration. It has been applied in different domains of real world application. Many sampling-based algorithms have been proposed. According to their relations and characteristics, these methods can be classified into different categories.

Genetic algorithm (GA) based instance selection: Genetic algorithms (Holland, 1975) have been successfully applied to various problems (Goldberg, 1989). Genetic algorithm can be viewed as a general-purpose optimization technique in discrete search spaces. They are suitable for complex problems with multi-model objective functions. Their application to instance selection was proposed by Kuncheva (Kuncheva, 1995) for designing nearest neighbor classifiers. In her study, the classification performance of

selected instances was maximized. A penalty term with respect to the number of selected instances was added to the fitness function of her subsequent genetic algorithms (Kuncheva, 1999). Later on, a generic algorithm-based approach was used for simultaneously selecting instances and features (Liu, et al, 2001). Through computer simulations, the authors demonstrated that a small number of instances can be successfully selected together with only significant features by their genetic algorithm. They also demonstrated that the generalization ability of nearest neighbor classifiers was improved by the instance and feature selection in some datasets.

There are many other instance selection methods which are similar to those mentioned above. These methods can reduce the number of instances in datasets, but they didn't do anything for the noise instance removal. The following methods consider how to detect and eliminate the noisy instances.

Iterative case filtering algorithm: Iterative case filtering algorithm (ICF) was introduced by Liu and colleagues (Liu et al, 2001). They improved the repeated Wilson algorithm investigated by Tomek (Tmoek 1976) by applying a rule which identifies cases that should be deleted. In the ICF algorithm, the authors built a K-Nearest-Neighbor classifier, then found and removed the noisy instance which was wrongly predicted by the k-NN classifier. This process is repeated iteratively until no more instances need to be removed.

In all previous instance selection studies, few of them, if any, focused on microarray dataset. In this study, we introduce a new instance selection method (C-loocv) by improving ICF algorithm. C-loocv is applied to remove the noisy instance from microarray datasets and more reliable data mining results are expected to be mined from

the C-loocv refined datasets.

3.6. CLASSIFICATION

Classification aims to learn how to classify objects into one of a pre-specified set of categories or classes. A robust classification method is very important to classify the new samples into the correct category efficiently and accurately. There have been lots of studies looking for reliable classification methods, as discussed below.

Naive Bayesian: Naive Bayesian is a simple probabilistic classifier based on Bayesian' theorem with the (naive) independence assumption. Based on the rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.

K-nearest neighbor (*K*-NN): *K*-NN is a method for classifying objects based on closest training examples in the feature space, which is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. The *K*-NN classifier is a simple supervised concept learning scheme which classifies unseen instances by finding the closest previously observed instances, taking note of their classes, and predicting the class for the unseen instance (Cover et al, 1967). *K*-NN is a non-parametric classifier which has been applied to various information retrieval problems. K-NN uses an integer parameter K. Given an input x, the algorithms finds the K closest training data points to x, and predicts the label of x based on the vote of labels of the K points.

Decision Tree: In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an object to conclusions about its

target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classes and branches represent conjunctions of features that lead to those classes. Thus, each node corresponds to a sequence of predicates and their values appearing on the downward path from the root to it. Each leaf is labeled by a class. To predict the class label of an input, a path to a leaf from the root is found depending on the value of the predicate at each node that is visited.

Many classification methods have been derived from decision tree method. ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree (Quinlan 1993). It prefers smaller decision trees (simpler theories) over larger ones. However, it does not always produce the smallest tree, and is therefore a heuristic. The ID3 algorithm can be summarized as follows: (1) Take all unused attributes and count their entropy concerning test samples; (2) Choose attribute for which entropy is minimum; (3) Make node containing that attribute.

C4.5 algorithm is an extended version of ID3, which is used to generate a decision developed by Ross Quinlan (Quinlan 1993). C4.5 is often referred to as a statistical classifier. C4.5 uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. C4.5 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision.

Committee decision techniques such as AdaBoost (Freund et al., 1996) and Bagging (Breiman et al., 1996) have also been proposed to increase the prediction accuracy by

voting the member decisions of the committee. Both AdaBoost and Bagging apply a base classifier multiple times to generate a committee of classifiers using bootstrapped training data. By the bagging idea, a bootstrapped training set is generated from the original data. Boosting uses a different method to construct the committee of classifiers. It builds the individual classifier sequentially so that every new classifier is influenced by the performance of those built previously. Therefore, the samples incorrectly classified by previous models can be emphasized in the new model.

Classification and Regression Tree (CART): CART (Breiman et al. 1984) are nonparametric procedures for explaining and/or predicting a response, either categorical (then this is discriminant analysis or classification), or continuous (then this is a nonparametric regression).

Support vector machines (SVM): Support vector machines are a relatively new type of learning algorithm. SVMs were originally introduced by Vapnik and co-workers (Boser et al., 1992; Vapnik et al., 1998) and successively extended by a number of other researchers (Malossini et al., 2000, Cristianini et al., 2000). Recently, SVMs have been shown to perform well in multiple areas of biological analysis including evaluating microarray expression data (Statnikov et al., 2005, Brown et al., 2000).

Support vector machines (Vapnik, 1998) have exhibited superb performance in binary classification tasks. Intuitively, SVM aims at searching for a hyperplane that separates the two classes of data with largest margin (the margin is the distance between the hyperplane and the point closest to it). SVMs have demonstrated the ability to not only correctly separate entities into appropriate classes, but also to identify instances whose established classification is not supported by data. Although SVMs are relatively

insensitive to the distribution of training examples in each class, they may still get stuck when the class distribution is too skewed. This is obviously not the desired classification result.

Emerging patterns (EPs): EPs (Dong, Li 1999) can also serve as a classification model. By aggregating the differentiating power of EPs/JEPs, the constructed classification systems (Li et al 2002, Li et al 2001, Dong, Zhang et al 1999) used to be more accurate than other previously existing classifiers. In recent years, several new classifiers have been generated on the basis of EPs, which will be discussed next.

Prediction by collective likelihood (PCL): PCL (Li, et al., 2002) is based on the concept of emerging patterns. With the discovery of emerging patterns, PCL proceeds to calculate a classification score for every class when a test sample is presented; and the class with the highest score is predicted. The classification scores are calculated by aggregating the frequencies of multiple top-ranked EPs: the committee of patterns and their collective discriminating power show strong strength. This method makes higher accuracy of prediction for many published microarray gene expression data (Li. et al., 2003b).

Highly differentiative gene groups (HDGGs): Based on the previous studies, we propose a new concept called HDGGs. HDGGs are the specific emerging patterns whose frequency is the highest among the EPs mined from one gene club (Mao & Dong, 2005). The classifiers built using HDGGs are called HDGG-based classifiers. Our experiments indicate that such classifiers predict the samples in many microarray datasets with very high accuracy. In this dissertation, we make considerable contribution for classification problem using HDGG-based classifiers.

Chapter 4: DISCOVERY AND APPLICATION OF HIGHLY DIFFERENTIATIVE GENE GROUPS (HDGGS)

4.1. MOTIVATION

It is commonly believed that suitable analysis of microarray gene expression profile data can lead to better understanding of diseases, and better ways to diagnose and treat diseases. To achieve those goals, it is of interest to discover the gene interaction networks, and perhaps even pathways, underlying given diseases from microarray data.

Most physiological functions in human body are regulated by multiple genes instead of individual genes. With specific diseases, the alteration of some physiological functions may be controlled by a group of related genes which are interactive with each other. Such groups of genes will be referred to as highly differentiative gene groups (HDGGs). Our aim of this chapter is to give methods to find such groups of genes, from datasets collected for studying such diseases.

We note that the discovery of HDGGs is a challenging problem, due to the high dimensionality of microarray datasets.

4.2. OUR APPROACH

A highly differentiative gene group (HDGG) is defined as a set of genes which co-express in a certain manner consistently and frequently in the diseased class but never co-express in that manner in the normal class, or vice versa. For example, in Table 4.1, {g_4, g_5} is a HDGG since g_4 is low (shown as '0') and g_5 is high (shown as '1') in all diseased samples, but there are no normal samples where g_4 is low and g_5 is high. We only want to discover minimal HDGGs (in the set-containment sense) to ensure that the set of HDGGs is concise. We use frequency of HDGGs in a class to measure their strength, where high frequency indicates high strength. When a HDGG has 100% frequency in a class, we call the HDGG a signature HDGG for that class.

Table 4.1: A simple gene expression dataset

	d_1	d_2	d_3	d_4	d_5	n_1	n_2	n_3	n_4	n_5
g_1	1	0	1	1	1	0	0	0	0	0
g_2	0	0	0	1	0	1	1	1	1	0
g_3	1	1	1	1	1	1	0	1	0	1
g_4	0	0	0	0	0	1	1	0	1	0
g_5	1	1	1	1	1	0	1	0	1	0

d1…d5: diseased tissues; n1 … n5: normal tissues. g1 … g5: genes.
'0': the gene expresses low; '1': the gene expresses high.

Since microarray data have thousands of dimensions, discovering HDGGs is a big challenge. Border differential algorithm (Dong et al, 1999; Dong et al., 2005) is applied to mine HDGGs. For HDGGs mining, the required search time is exponential in the number of dimensions. Thus, exhaustive search of the whole dataset is impossible. The border differential algorithm can effectively handle up to 75 genes for current generation PCs, but can not finish in reasonable amount of time for much higher dimensions of dataset. In this chapter, we introduce methods that can do much better job given the

dimensionality challenge. Our methods are based on the novel concept of gene club.

A gene club is a set of genes in which the genes have high potential to be interactive with each other. The total number of genes in a gene club should be big enough to contain high frequency HDGGs; at the same time, it shouldn't be too large for border differential algorithm to effectively handle up under current generation PCs. Within in a gene club we can (i) efficiently discover signature HDGGs which completely characterize the diseased and the normal tissues respectively, (ii) find strongest or near strongest HDGGs containing any given genes, and (iii) find much stronger HDGGs than using previous methods.

The main idea of our approach is to select, for each given gene g, a set of genes which are highly likely to be interactive with g. We call the set of potentially interactive genes a gene club of gene g. We will consider several methods for finding good gene clubs. Although the determination of gene clubs share similarities with feature selection, it has some new characteristics and it is based on the interaction among genes.

4.3. GENE CLUB FORMATION STRATEGY

In this section, we discuss four methods for gene club formation: the independent method (IN), the iterative method (IT), the divisive and independent method (DIN), and the divisive and iterative method (DIT). Then, we show the results obtained from several public microarray datasets and discuss the importance of mined HDGGs from the gene clubs obtained using these methods.

Our gene club based methods work as follows: First, for each gene g we find a gene club for some desired cardinality k. Second, we apply the border differential algorithm (Dong

et al., 1999; Dong et al., 2005) to discover the best EPs containing g from the gene club. Finally, we remove the conditions to get the HDGGs. All these methods utilize information gain for gene groups.

4.3.1. Independent gene club formation

The independent gene club formation method (IN) forms a gene club for a gene g by selecting the genes which are independently the most interactive with g. This method is based on the notion of combined information gain, defined (for each gene g') as

$$InfoGain(g'|\, g) = InfoGain(g, g') - InfoGain(g)$$

The combined information gain captures how g' interacts with g with respect to the disease under consideration, or how much "help" g' offers to g w.r.t. the disease. The IN method works by first ranking all genes g' in decreasing InfoGain(g'| g) order, and then selecting the k-1 genes g_1, g_2, …, g_{k-1} with the highest combined information gain as the gene club for g. Figure 4.1 illustrates how the method works; the combined information gain is shown as the label for the edge from g to g', apparently, $I_1 \geq I_2 \geq \ldots \geq I_{k-1}$.

From Table 2.1, the gene club of size 3 for gene g_1 formed by IN consists of g_2 and g_4, since InfoGain($g_2|g_1$) = 0.554 > InfoGain($g_4|g_1$) = 0.502 > InfoGain(gi|g1) for i = 3 or 5.

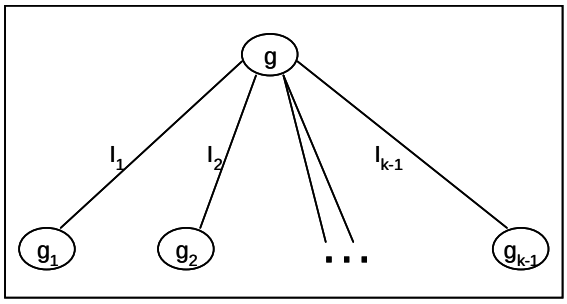

Figure 4.1: the IN method

4.3.2. Iterative gene club formation

The iterative gene club formation method (IT) forms a gene club for a gene g by selecting the genes which are iteratively the most interactive with g for a current partial gene club. This is different from the IN method, which does not consider the interaction of a new gene with other selected genes. IT is based on the notion of generalized combined information gain, defined (for each gene g' and selected genes $g_1,\ldots, g_m$) as

$$InfoGain(g' \mid g_1, g_2 \ldots g_m, g) = InfoGain(g_1, g_2, \ldots g_m, g, g') - InfoGain(g_1, g_2, \ldots g_m, g)$$

The generalized combined information gain captures how g' interacts with the current partial gene club with respect to the disease under consideration, or how much help g' offers to the partial gene club w.r.t. the disease. IT finds a gene club of size k for g as follows: First it sets the partial gene club to be {g}; it then selects the next gene g_1 having the highest InfoGain(g' | g) among all remaining genes g' and adds g_1 to the partial gene club; it then selects the next gene g_2 having the highest InfoGain(g'| g_1, g) among all remaining genes g' and adds g_2 to the partial gene club; it then selects the next gene g_3 having the highest InfoGain(g'| g_1, g_2, g) among all remaining genes g' and add g_3 to the partial gene club; this process is repeated until the $(k-1)^{th}$ gene is selected and the partial gene club becomes the final gene club of g. Figure 4.2 illustrates the iterative process, where the edge labels represent the generalized combined information gain.

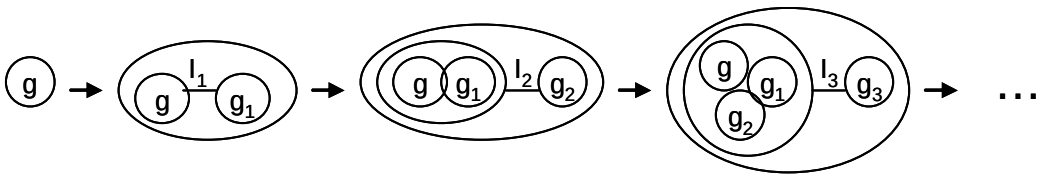

Figure 4.2: the IT method

According to the data in Table 2.1, the gene club of g_1 formed by IT consists of g_2 and g_5,

which is different from that formed by IN. One can verify that InfoGain(g_5| g_2, g_1) > InfoGain(g_4| g_2, g_1).

4.3.3. Divisive gene club formation

We now turn to two more methods which first divide the data using a split value for g, and then use one of the two previous methods to work with each partition. This is illustrated in Figure 4.3. The selected genes from the two partitions are then combined to form the overall gene club for g. If IN is used for each partition, this method is called the divisive independent method (DIN); if iterative method is applied, this method is called divisive iterative method (DIT).

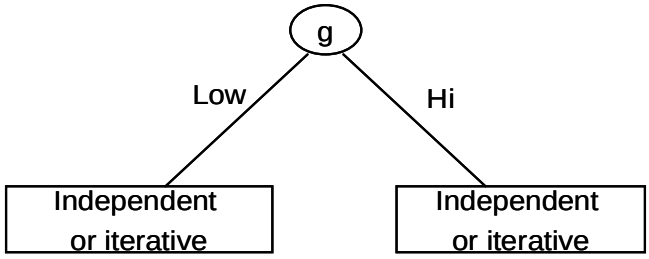

Figure 4.3: the divisive method

We need to combine the genes selected from the two partitions. For both methods, we first select 1/3 of the gene club members from each partition among genes with non-zero information gain. The final 1/3 (or more if we did not get 1/3 in the last step) is chosen from the remaining selected genes from the two partitions. In DIN, the last 1/3 is chosen according to the information gain values for the genes from the two partitions. In DIT, this is done by normalizing information gain using the partial gene club size.

Normalization is used since it is not very meaningful to compare information gain over gene groups with large size differences.

4.3.4. Converged gene ranking

Since there are thousands of genes in one dataset, it is impractical to establish gene clubs for every single gene. A better strategy is to use top k ranked genes as seeds to form gene clubs because the top ranked genes have stronger relationship with the specific class. In each club, the HDGGs are mined with border differential algorithm.

As mentioned before, some genes may not be strongly related with the specific disease individually, but they are very important for the disease under consideration when they combine with other genes. In another word, they are important for the disease under gene group based condition instead of individual condition. Therefore, we need to rank the genes with new criterion which is based on genes' participation in high-quality gene groups.

Converged gene ranking is such a method that gives the genes which are involved in HDGGs more frequently a higher rank. After the genes are ranked according to their information gain, the converging method is applied to re-rank these genes. The following is the algorithm for converged gene ranking:

(1). Ranking the genes in decreasing information gain order.

(2). Using all four gene club methods to obtain the high frequency HDGGs (using top ranked genes as seeds). In each gene club, the HDGGs with the highest frequency are chosen. H is the set of chosen HDGGs.

(3).Calculating the weight for every gene in each HDGG in H:

$$Weight(g) = \sum_{g \in h, h \in \mathrm{H}} Frequency(g) \qquad h : \mathrm{HDGG},\ H : set\ of\ HDGGs$$

(4). Re-ranking the genes according to their weight.

(5). Repeating steps (1) to (4) until the gene rank order reaches steady state.

As will be seen later, the set of top ranked genes using information gain gene ranking method are different from that using converged gene ranking method. Therefore, the seeds for gene club formation are also different. With converged gene ranking, we discover many high frequency HDGGs which are missed by information gain gene ranking method. Therefore, this strategy can be considered as an auxiliary method which can help to find higher frequency HDGGs from microarray datasets.

4.4. BUILDING HIGH ACCURACY HDGG-BASED CLASSIFIERS

HDGGs can also be used to build classifiers to diagnose diseases with very high accuracy. The following pseudo-code describes how to build HDGG-based classifiers. (The parameter k can be determined by the user based on the available computation power. In this research, we choose k = 20 as its default value unless stated otherwise). The details on building HDGG-based classifier can be found in Chapter 6.

1. Find the top k genes g_1, …, g_k ranked by information gain;

2. For each gene g_i, find its gene club using the IT method, and then find the strongest HDGGs containing g_i using border differential algorithm;

3. Among the mined HDGGs, a total of 2 * k HDGGs are chosen as classifier according to our criteria (see below). Among all these chosen HDGGs, k HDGGs are from one class whereas the other k HDGGs from the other class;

4. For a given test sample T, T's score in class x(C_x) is calculated according to the frequency of each chosen HDGGs in C_x:

$$S(T)\_C_x = \sum_{i=1}^{k} \frac{frequency(HDGG_i \ in \ C_x)}{n_x}$$

where n_x is the total number of samples in C_x.

By comparing the scores of T in all classes, T is categorized into the class in which T has the highest score.

Criteria for choosing HDGGs: We apply border differential algorithm to obtain plenty of HDGGs from gene clubs. Among them, we need to choose the typical HDGGs as classifier according to following criteria:

(1) High HDGG's frequency, the higher the better. This is the most important one.

(2) Low gene overlapping in the chosen HDGGs, the lower the better. This strategy is called gene diversity.

(3) Tie breaking. Two or more HDGGs may have the same frequency and same gene diversity. We need to make a rule to break the tie. The following are our rules:

(i). Comparing the length of each HDGG, shorter HDGGs ranks higher;

(ii). Comparing the gene's rank between two or more equivalent HDGGs, the HDGGs which contain the lowest ranked gene should be ranked lower. This is also called "smallest first rule". For example, two HDGGs {1, 3, 5} and {2, 3, 4}, both have the same length of three genes, gene '4' ranks higher than gene '5', so HDGG {2, 3, 4} ranks higher than {1, 3, 5}.

4.5. RESULTS AND EVALUATION

We evaluate the effectiveness of our methods in terms of (i) the high-strength HDGGs discovered, (ii) the ability to find the strongest HDGGs, (iii) the improvement of strength of discovered HDGGs over the top-k gene method, and (iv) the meaningful biological functions of our HDGGs compared with previous gene group methods. The experiments were conducted on the following datasets: (1) colon cancer data (Alon et al., 1996), which has 2000 genes and 62 tissue samples (22 normal ones and 40 cancer ones); (2) prostate cancer data (Singh et al., 2002), which has 12600 genes and 102 tissue samples (50 normal ones and 52 cancer ones); (3) breast cancer data (van't Veer et al 2002), which has 24481 genes and 78 tissue samples (44 non-relapse ones and 34 relapse ones); (4) ovarian cancer data (Petricoin et al 2002a), which has 15154 genes and 253 tissue samples (91 normal ones and 162 tumor ones); (5) leukemia data (Golub, 1999), which has 7129 genes and 72 tissue samples (47 ALL and 25 AML).

4.5.1. High strength HDGGs

Table 4.2 lists the top 10 HDGGs and EPs in diseased tissues and normal tissues for the colon cancer data. The HDGGs can be obtained by removing the high/low signs from the EPs. Each number represents a gene, with '1' for the highest ranked gene according to the information gain order. The signs of '+' and '–' represent "high" and "low" respectively, e.g. '1+' is for "gene 1 is high", and '4-' is for "gene 4 is low". We give the accession number and description for the genes which occur in the HDGGs, and their split values in Table 4.8.

We observe that some of the signature HDGGs involves genes ranked at 1089. This

implies that such genes are very weak for characterizing the cancer by themselves, but they can completely characterize the cancer when combined with several other genes.

Table 4.2: The top 10 HDGGs in diseased (left) and normal tissues (right) of colon data

Emerging patterns	Count	Frequency (%)	Emerging patterns	Count	Frequency (%)
{1+ 4- 112+ 113+}	40	100	{12- 21- 35+ 40+ 137+ 254+}	22	100
{1+ 4- 113+ 116+}	40	100	{12- 35+ 40+ 71- 137+ 254+}	22	100
{1+ 4- 113+ 221+}	40	100	{20- 21- 35+ 137+ 254+}	22	100
{1+ 4- 113+ 696+}	40	100	{20- 35+ 71- 137+ 254+}	22	100
{1+ 108- 112+ 113+}	40	100	{5- 35+ 137+ 177+}	21	95.5
{1+ 108- 113+ 116+}	40	100	{5- 35+ 137+ 254+}	21	95.5
{4- 108- 112+ 113+}	40	100	{5- 35+ 137+ 419-}	21	95.5
{4- 109+ 113+ 700+}	40	100	{5- 137+ 177+ 309+}	21	95.5
{4- 110+ 112+ 113+}	40	100	{5- 137+ 254+ 309+}	21	95.5
{4- 112+ 113+ 700+}	40	100	{7- 21- 33+ 35+ 69+}	21	95.5

4.5.2. Ability to find top strength HDGGs and improvement of strength over top k method

Experiments showed that our methods can often find the strongest HDGGs, and they can find EPs whose frequency is very close to the strongest for cases when our methods cannot find the strongest EPs. The experiments are set up as follows: From the colon cancer data, we picked the top 75 genes under the information gain rank. (The number 75 is chosen for reasons discussed in section 4.2) We then use the border differential algorithm to exhaustively mine all the EPs in these 75 genes. For each g of these 75 genes, let SupBEP(g) be the highest frequency (support) of the discovered EPs containing g. We then apply the four methods introduced in this chapter, together with the top-k method, on these 75 genes, using a gene club size of 20. For each gene g, we check whether a given method can find an EP containing g with frequency of SupBEP(g) from

the 75 genes. We considered how often each of the methods is able to find the strongest EP.

We first let S consist of the top 75 genes ranked by information gain. The IT method is the best, which can find the strongest EPs for about 82.5% of the genes. We also found that the average frequency (over the 75 genes) of the strongest EPs found by IT is more than 98% of the average support of the strongest EPs. In contrast, the top-k method can only find the strongest EPs for 32.5% of the genes, and the average frequency of the strongest EPs found by that method is about 77% of the average frequency of the strongest EPs. The other three of the new methods are slightly worse than IT. Figure 4.4 shows the performance of all methods in terms of the average frequency of the strongest EPs for the top 20 genes. In the figure, EX denotes the exhaustive method, and GT20 denotes the top-k method with k = 20.

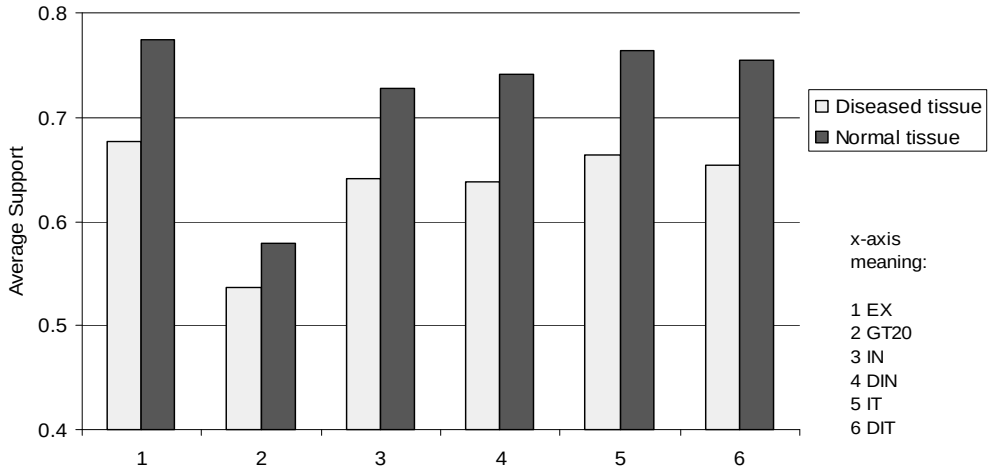

Figure 4.4: Average frequency of strongest EPs found by six methods over top 75 genes in colon cancer data

The experiments over randomly selected sets of 75 genes also showed similar

performance. Experiments reported in Figure 4.4 show that the new methods improve over the top-k method by a large margin. Indeed, the IT method improved the average frequency by 47.1% and 29.8% respectively, in diseased tissues and normal tissues, over the top-k method.

We also conducted experiments for gene club size of 35. Every method except EX improved. The relative performance of the methods is similar to the case when the gene club size is 20.

4.5.3. From HDGGs to gene functions and disease understanding

By using our HDGGs based method to group genes, we found that many genes in the HDGGs are related to the disease under consideration (i.e. colon cancer in our example). Indeed, from the signature HDGGs in Table 4.2, we found several genes have known biological functions. For example, in the HDGG {1, 4, 112, 113}, gene 1 (Chang et al., 2002) has been studied intensively. It is one of the major mediators of the inflammatory response and a potent angiogenic factor. There is a close relation between the level of this gene and the state of illness: the higher the expression level it is, the more serious the patients' condition. Gene 4 encodes a kind of cysteine and glycine-rich protein, which may be involved in regulatory processes important for cellular development and differentiation (Wang et al., 1992). Gene 112 regulates cell growth; it is believed to have some tissue-specific functions, although its specific function is still under study (Nomura et al., 1994). Gene 113's function includes the following: it accelerates differentiation of select human hematopoietic cells; it encodes a protein which is a receptor in erythropoiesis; it may play a role in angiogenesis (Wang et al., 2002). The fact that {1+,

4-, 112+, 113+} is a pattern characterizing the colon diseased tissue is consistent with the function of the four involved genes.

Some genes in HDGGs, especially the low-ranked genes, have not received enough investigation. The membership of these genes in HDGGs indicates that these genes are important for the disease under consideration. For example, gene 113, gene 216, gene 1089 and so on are low-ranked genes. We were not able to find definitive published results indicating the function of such genes. We believe that these genes should be studied further in the biology and medicine fields.

The HDGGs may be used to suggest research directions to find gene functions and to discover new pathways. Many of the cancers considered in this paper are still the subject of extensive research (Sugiyama et al 2005), and the majority of the pathways are still to be found in large NIH supported projects (Newcomb 2003). These may be reasons why we only found the APC pathway related to colon cancer in the literature and the web, and the APC gene (accession number M74088) for this pathway was not included in the colon cancer data.

We also observed that, for some diseases such as leukemia and lung cancer, our gene club methods produced smaller improvement over the top-k approach than for other diseases. This happened because the top-k approach has achieved very high average support (89% in leukemia and 94% in lung cancer data) already – there is little room for further improvement. Interestingly, this implies that the important gene groups for these diseases only involve top ranked genes under the entropy measure. We suggest that this might be used as an indication that these diseases have relatively low disease complexity. On the other hand, colon and prostate cancers may have high disease complexity, since

there are important gene groups for these diseases that involve genes that are ranked quite low under the entropy measure.

4.5.4. Other datasets

We also conducted experiments on several other datasets. The results are mostly similar to that for the colon cancer data. For the prostate data, the improvement of the IT method over the top-k method is 16.97% in diseased tissues and 35.99% in normal tissues, respectively (see Figure 4.5). We list some of the top HDGGs in Table 4.7. For the breast cancer data, the improvement of the IT method over the top-k method is 6.03% and 14.55% respectively (see Figure 4.6). For the ovarian data, the improvement is 4.41% and 0.92% respectively (see Figure 4.7).

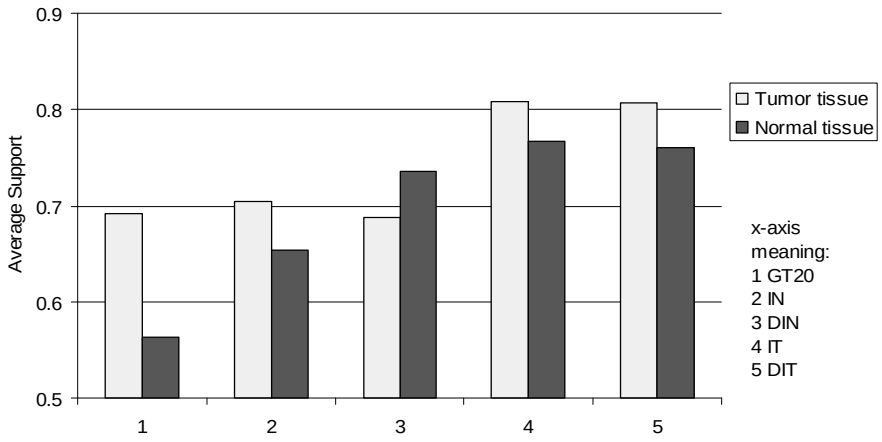

Figure 4.5: Average frequency of strongest EPs by different methods for prostate data

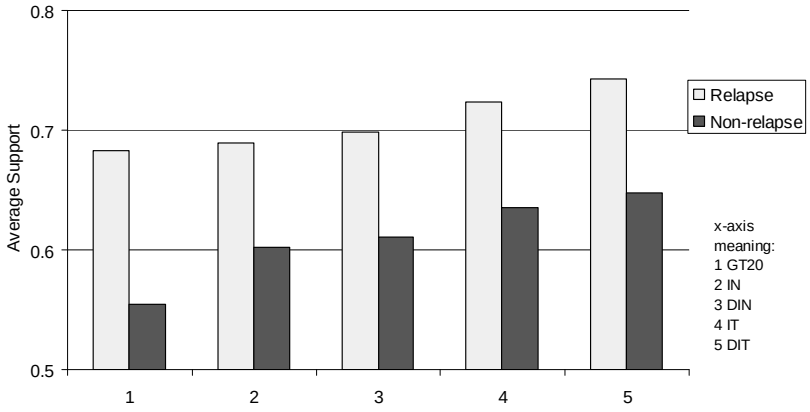

Figure 4.6: Average frequency of strongest EPs by different methods for breast cancer data

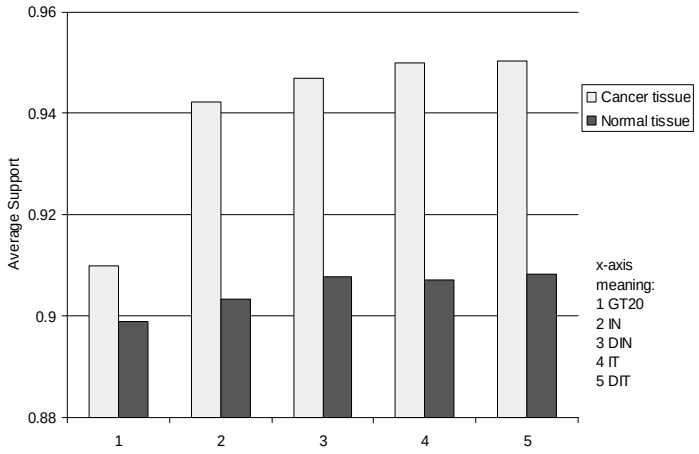

Figure 4.7: Average frequency of strongest EPs by different methods for ovarian data

4.5.5. Gene ranking by converged method

We re-ranked the gene by using HDGG-converged method. On average, most datasets can reach converged stage within 10 - 20 cycles. Table 4.3 lists the sets of top 20genes ranked by information gain and by converged gene ranking method using colon cancer data.

Table 4.3: Top 20 genes using information gain ranking and converged ranking

Ranking order	Information gain ranking	Converged gene ranking
1	**1670**	**1422**
2	**248**	681
3	**492**	575
4	**764**	**1670**
5	1771	**1041**
6	**624**	1923
7	**1041**	**1581**
8	**1422**	**624**
9	512	1632
10	1770	174
11	244	1634
12	**257**	**764**
13	779	1559
14	**398**	**257**
15	896	1885
16	**1581**	**492**
17	1292	**248**
18	651	580
19	1226	1327
20	42	**398**

In the table, the first column is the gene rank order; the second column is the information gain based ranking result and the third column is the result by gene converged ranking method. The number in the table is the gene's index. We notice that 50% of top 20 genes of the two methods are different. We use bold-number to mark the shared genes (in both top 20 genes of the two methods). The converged ranking is gene group-based, whereas information gain ranking is individual gene-based. The experimental results show that some important genes for one class are not the top genes in the individual gene based rank.

In colon cancer dataset, the average frequency of top 20 HDGGs obtained by information

gain ranking is 71.4% whereas that from converged gene ranking is 89.4%. The average frequency increased 18%. Using other microarray gene expression data to compare these two ranking methods, we also observed that the average frequency increases in some degree (data not shown).

4.5.6. Comparison of HDGGs based classification method with other methods

In this section, we compare our HDGGs-based classifier with other classifiers for predicting several published microarray cancer datasets. For a given dataset, if one classifier can predict the samples with lower error rate, we say that classifier is more robust. The error rate of a classifier is defined as the number of samples in a dataset that are incorrectly predicted by the classifier. It is also called test error rate, which is widely used in biomedical fields.

Table 4.4: The error rates of six classification algorithms

Datasets	HDGGs	PCL	SVM	C4.5		
				Boosting	Bagging	Single
Ovarian	3.4	4	5	5	8	10
Lung Cancer	2	3	1	27	18	27
ALL subtypes	4	4	7	14	10	21

Using three public datasets (Ovarian, Lung cancer (Hong et al, 1991) and ALL subtypes) which have been partitioned into training and testing sets already, we directly compare the classification results produced by our methods and some existing classification algorithms. Our method (HDGGs-based) has lower error rate than other algorithms (Table 4.4). The numbers shown in the table are the error rate. There is no testing data in

ovarian cancer data, so we used 10-fold cross validation method to estimate the error rate. We repeated the 10-fold cross validation method for 10 times, and then got the average of error rates for ovarian dataset. Our results indicate that HDGGs can be used to build very good classifiers and HDGGs are very important feature for a dataset.

4.6. DISCUSSION OF HDGGS AND FUTURE WORK

Some other methods do not work for HDGG mining: Since frequent itemsets are anti-monotone (i.e. all subsets of a frequent itemset are frequent), one may be tempted to think that some frequent-item or frequent-itemset based methods can be used to efficiently find the high-frequency EPs and HDGGs. For example, one may remove all the non-frequent items and then mine the reduced data set. For example, suppose that we want to find EPs whose frequency is at least 70% in the diseased tissues. We can first find the frequent items whose frequency is $\geq$ 70% in the diseased tissues. Then, for each tissue (sample), we remove the non-frequent items. The reduced samples will then be used to do border differential against the normal tissues.

Table 4.5: Minimum, maximum and average tuple length in reduced dataset of colon data

Threshold %	Diseased class			Normal class		
	Min	Max	Avg	Min	Max	Avg
100	179	179	179	541	541	541
90	423	534	510	730	835	820
80	596	943	864	863	1051	1010
70	705	1196	1075	912	1220	1147
60	795	1364	1216	982	1362	1267
50	868	1491	1322	1107	1494	1395

This method will not work for microarray gene expression data, because the reduced data

sets will still have very high dimensionality. Table 4.5 lists the maximum, minimum, and average length of the reduced tuples for thresholds ranging from 50% to 100%. Observe that, even for the threshold of 100%, the average tuple lengths are still 179 and 541 in the diseased and normal classes, respectively. Moreover, the method is not desirable since it requires the user to give a threshold.

Duplicated genes: There are thousands of gene probes in one microarray chip. During hybridization reaction, many different gene probes may hybridize with one gene sequence. These gene probes are called duplicate gene probes though their sequences are different. In our data analysis, we only keep one copy and eliminate the rests before gene club formation among all duplicated probes.

The following is our criteria for identifying duplicate (or equivalent) genes: We consider two genes g_1 and g_2 as duplicate genes if (a) InfoGain(g_1 | g_2) = InfoGain(g_2 | g_1) = 0, and (b) NormMutualInfo(g_1, g_2) = 100%, where NormMutualInfo = (Entropy(g_1) + Entropy(g_2) – Entropy(g_1, g_2)) / max (Entropy(g_1), Entropy(g_2)).

In the colon cancer dataset, there are 106 genes which have duplicated copies, and the total number of duplicated copies is 449 (so 343 of these are removed). Table 4.6 lists the duplicated genes among the top 500 genes in colon cancer data (Alon et al., 1996). Each row contains a set of genes which are equivalent to each other.

Table 4.6: Duplicated genes in top 500 gene group of colon data

11	12							
112	117							
168	169	170	171					
369	373	375	377	380	385	394	395	402
371	401							
376	387	388	391					
378	400							
382	383							

Disease complexity: We also observe that, for some diseases such as leukemia and lung cancer, our gene club methods produce smaller improvement over the top-k approach than for other diseases. This happens because the top-k approach has achieved very high average frequency (89% in leukemia and 94% in lung cancer data) already – there is little room for further improvement. Interestingly, this implies that the important gene groups for these diseases only involve top ranked genes under the entropy measure. We suggest that this might be used as an indication that these diseases have relatively low disease complexity. On the other hand, colon and prostate cancers may have high disease complexity, since there are important gene groups for these diseases that involve genes that are ranked quite low under the entropy measure.

Limitation: It is commonly known that microarray datasets are highly variable. The high variability in microarray data may affect the reliability of the discovered HDGGs. Therefore, it is necessary to test the effect of variability on our data mining results. In Chapters 5-7, we are going to investigate and minimize the effect of variability, and mine reliable HDGGs from highly variable microarray datasets.

4.7. APPENDIX

Table 4.7 lists the top HDGGs in normal and diseased classes in prostate cancer dataset (Singh et al., 2002). There are 52 disease samples and 50 normal samples in this dataset.

Table 4.7: The top 10 HDGGs in diseased (left) and normal tissues (right) of prostate cancer data

Emerging patterns	Count	Support (%)	Emerging patterns	Count	Support (%)
{07- 331- 557+ 5011-}	51	98.1	{11- 19- 20+ 41+}	43	86
{07- 331- 564+ 5011-}	51	98.1	{11- 20+ 41+ 3890+}	43	86
{07- 331- 708+ 5011-}	51	98.1	{11- 20+ 41+ 122-}	43	86
{07- 331- 719- 5011-}	51	98.1	{11- 41+ 78-}	43	86
{07- 557- 657- 5011-}	51	98.1	{19- 41+ 78- 122-}	43	86
{07- 564+ 657- 713+ 5011-}	51	98.1	{01+ 06- 2002+}	42	84
{07- 657- 708+ 5011- }	51	98.1	{04- 11- 19- 41+}	42	84
{07- 657- 719- 5011-}	51	98.1	{04- 11- 41+ 122-}	42	84
{01- 947- 1271-}	50	96.1	{04- 11- 41+ 3890+}	42	84
{01- 1271- 2083-}	50	96.1	{04- 18+ 507+ 1937+}	42	84

Table 4.8 is the description of the genes that are involved in HDGGs in colon cancer data.

We named the genes that are involved in HDGGs discriminating genes (DGs).

Table 4.8: Description of genes involved in HDGGs in Table 4.2

Gene number	Splitting point	Accession number	Description
1	59.83	M26383	monocyte-derived neutrophil-activating protein mRNA
2	1696	M63391	Human desmin gene
3	379.4	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
4	842.3	M76378	Human cysteine-rich protein (CRP) gene, exons 5 an exon 6
5	84.88	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
6	230	X12671	Heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
7	275	R36977	TRANSCRIPTION FACTOR IIIA
8	735.8	J02854	MYOSIN REGULATORY LIGHT CHAIN 2
9	447	M22382	MITOCHONDRIAL MATRIX PROTEIN P1
10	88.9	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA
11	1048	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
12	390.4	M16937	Human homeo box c1 protein, mRNA
13	400	H40095	MACROPHAGE MIGRATION INHIBITORY
14	289	U30825	Human splicing factor SRp30c mRNA
15	334	H43887	COMPLEMENT FACTOR D PRECURSOR
16	84.2	H51015	H.sapiens mRNA for p cadherin
17	417.3	X57206	GTP-BINDING NUCLEAR PROTEIN RAN

18	494.2	R10066	PROHIBITIN (Homo sapiens)
19	75.43	T96873	HYPOTHETICAL PROTEIN IN TRPE 3'REGION
20	2598	T57619	40S RIBOSOMAL PROTEIN S6 (Nicotiana tabacum)
21	735.6	R84411	SMALL NUCLEAR RIBONUCLEOPROTEIN
35	58.51	M36634	Human vasoactive intestinal peptide (VIP) mRNA
40	356.4	R28373	HEMOGLOBIN BETA CHAIN (HUMAN)
69	97.4	R39209	IMMUNODEFICIENCY VIRUS TYPE I ENHANCER-BINDING PROTEIN 2
71	454.7	H17434	NUCLEOLIN (HUMAN)
108	3239	Z24727	H.sapiens tropomyosin isoform mRNA, complete CDS
109	282.3	J03040	SPARC PRECURSOR (HUMAN);contains MSR1 repetitive element
110	123.6	K03460	Human alpha-tubulin isotype H2-alpha gene, last exon
112	99.38	D14812	Human mRNA for ORF
113	155.5	T51849	TYROSINE-PROTEIN KINASE RECEPTOR ELK PRECURSOR
116	24.29	R49459	TRANSFERRIN RECEPTOR PROTEIN (Homo sapiens)
117	36.63	H49515	SIGNAL RECOGNITION PARTICLE 68 KD PROTEIN
136	189.2	X61118	Human TTG-2 mRNA for a cysteine rich protein with LIM motif
137	26.81	R06601	METALLOTHIONEIN-II (Homo sapiens)
177	24.6	T40578	40S RIBOSOMAL PROTEIN S6 (Homo sapiens)
188	147.5	M31303	Human oncoprotein 18 (Op18) gene
216	31	H66786	ESTROGEN SULFOTRANSFERASE (Bos Taurus)
254	29.03	H64807	PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
263	373.3	M69135	Human monoamine oxidase B (MAOB) gene, exon 15
309	1393	H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM
696	126.4	M59807	NATURAL KILLER CELLS PROTEIN 4 PRECURSOR
700	116.5	H87465	PRE-MRNA SPLICING FACTOR SRP75 (Homo sapiens)
1089	373.5	R80855	MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3
1261	81.16	M23254	Human Ca2-activated neutral protease large subunit (CANP) mRNA

Chapter 5: MICROARRAY GENE EXPRESSION DATA CONCORDANCE DETECTION

5.1. MOTIVATION

The microarray technology has great potential for improving our understanding of biological processes, medical conditions and diseases. Often, microarray datasets are collected using different microarray platforms (provided by different companies) under different conditions in different laboratories. The cross-platform and cross-lab concordance of the microarray technology needs to be evaluated before it can be successfully and reliably applied in biological/clinical practice.

It has been realized that different testing methods may cause different concordance results. In previous studies, the cross platform and cross lab concordance of microarray data has been tested using statistical methods, but not with comparative study yet. In this chapter, we will detect the cross platform and cross lab concordance under the comparative perspective.

5.2. OUR APPROACH

In this chapter, we introduce novel comparative methods for evaluating concordance of microarray data collected from different platforms and different laboratories. Our methods evaluate this concordance by measuring quality preservation of discriminating genes and classifiers. The discriminating genes are the genes that participate in HDGGs,

and the classifiers are the HDGG-based classifiers which have been discussed in Chapter 4. They are used to test the platform/lab concordance under the comparative perspective.

Our rationale for classifier/discriminating gene transferability is: Considering that microarray datasets are generated from different platforms, if the microarray datasets are concordant with each other with respect to discriminating genes or HDGG based classifiers, then the knowledge on discriminating genes/HDGG based classifiers gained from one platform/lab can be transferred to another platform/lab. This is called classifier/discriminating gene transferability.

We apply classifier transferability to test the degree of classifier-based concordance between different platforms/laboratories; we use consistency rate to detect concordance (consistency) between the datasets before and after most of the gene-level noise is removed; and we also use *P*-value to quantitatively measure the concordance between platforms/laboratories. Conclusions on concordance based on our methods mostly agree with previous conclusions obtained using other methods, except the cases involving one particular platform. It should be noted that our methods are very general, and can be applied when two or more tissue types/classes are available.

5.3. MATERIALS AND METHODS

5.3.1 Gene expression data

In this chapter, we evaluate our methodology by using some microarray gene expression data provided by the Microarray Quality Control (MAQC) project in terms of inter-lab and cross-platform concordance. We describe the MAQC data briefly below; more details can be found in MAQC main paper (Shi et al., 2006). The datasets were generated using

more than 10 different platforms in more than 30 different labs. We use the data generated using four major platforms (Applied Biosystem(ABI), Affymetrix(AFX), Agilent on color array (AG1), and GE Healthcare(GEH)), leading to a total of 12 datasets. Among them, AG1 is a fairly new technology compared with the traditional two color Agilent platform.

For each platform, there are three repeated datasets (explained below), each from one of three laboratories; this design makes both inter/intra platform comparison possible. More specifically, four standard mRNAs are used in each lab/platform combination, which implies that biological variation has been eliminated. The four mRNAs are named as A, B, C (75%A + 25%B) and D (25%A + 75%B). mRNA A is the universal human reference RNA (SUHRR) provided by the Stratagene; and mRNA B is the ambion human brain reference RNA (AHBRR) from the Ambion. While mRNA A and B are primitive, mRNA C and D are mixtures of A and B with the proportions given above. In each dataset, there are five repeats for each mRNA, giving rise to a total of 20 chip data. Since C contains more A than B whereas D contains more B than A, mRNA A & C are grouped into one class, whereas mRNA B & D into the other class; Table 5.1 gives a schematic explanation. This division is used in all datasets.

Table 5.1: MAQC style dataset structure

	Class 1										Class 2									
	mRNA A					mRNA C					mRNA B					mRNA D				
	t_1	.	.	.	t_5	t_6	.	.	.	t_{10}	t_{11}	.	.	.	t_{15}	t_{16}	.	.	.	t_{20}
g_1																				
…																				
g_n																				

One MAQC dataset: $g_1..g_n$ are genes; $t_1 \ldots t_{20}$ are tissues;

The concordance among laboratories/platforms is tested by comparing the datasets generated from these lab-platform combinations. Observe that, all of the datasets from a common platform have the same set of genes and RNA samples, and they use the same gene IDs and RNA sample IDs to refer to the genes and samples. For datasets generated from different platforms, the same samples are used but they may use different sets of genes.

5.3.2. Discovery of discriminating genes

One main idea in this chapter is to use the transferability of discriminating genes from one dataset to another to evaluate the concordance of two given datasets. Discriminating genes are genes which are highly correlated with a class, or genes which participate in highly differentiative gene groups (HDGGs). The discriminating genes are transferable from one lab/platform combination to another, if the discriminating genes discovered from the dataset generated from the first lab/platform combination are highly likely also discriminating genes for the dataset generated by the second combination, and vice versa. Intuitively speaking, high discriminating-gene transferability implies that one can use discriminative knowledge gained in one platform/lab combination in another platform/lab combination. In this chapter, we use discriminating patterns that occur in one class but never occur in the other class. Such discriminating patterns are referred to as HDGGs (Mao & Dong 2005) in data mining studies. HDGGs have been proved to be very useful for discovering the inherent distinctions between different classes of data, and they have been very useful for building highly accurate classifiers (see Chapter 4).

The MAQC data we use have only 20 samples. A very high proportion of HDGGs mined from such data contain only one gene. Thus, the HDGGs from MAQC data can be considered to be equivalent to highly frequent jumping EPs (JEP), which are defined as emerging patterns that appear in one class but never exist in other classes.

As mentioned in Chapter 4, the genes which are involved in HDGGs (or highly frequent JEPs) are considered as discriminating genes. If a JEP involves just one gene, then the JEP has a condition of one of the following two forms: "$g \leq v$" or "$g > v$", where g is a gene and v is a value; the condition asks whether gene g's expression value (in a tissue under consideration) is $\leq$ or $>$ than v. A multi-gene JEP is a set (or conjunction) of several such conditions; we refer to a JEP with k conditions as a k-gene JEP. Multi-gene JEPs capture interactions among genes which only happen in one class but never in the other classes. Since many biological functions are regulated by multiple genes, we collected the DGs from 2-gene and 3-gene JEPs, besides those from one-gene JEPs. After using the entropy-based method (Dougherty et al., 1995) to find the split value for each gene, the so-called "iterative gene club formation algorithm" (Mao & Dong, 2005) was employed to discover 2-gene and 3-gene JEPs from the two classes of each MAQC microarray dataset. We selected the JEPs having 100% frequency in the class they occur, so the discriminating genes in such JEPs are frequently involved in discriminative interactions among genes.

5.3.3. Classifier transferability

If a classifier built from the two classes in one dataset is applied to predict the sample's class type in another dataset and the prediction accuracy is high, and vice versa, the classifiers can be transferred between the two datasets. High classifier transferability represents high similarity between the two datasets. Since MAQC datasets were derived from different labs/platforms, high classifier transferability between pairs of such datasets indicates high concordance among these labs/platforms. Thus, classifier transferability is a good criterion to test the platforms' concordance. Intuitively speaking, high classifier transferability implies that one can use diagnosis knowledge gained in one platform/lab condition to predict what may be happening in another platform/lab condition.

A classifier is a function (or computer program) for classifying objects without class label to one of a pre-specified set of categories or classes. It is trained from data having class labels. In terms of microarray data, the goal is to build highly accurate classifiers that may be used to predict class membership for new microarray samples. Some selected sets of emerging patterns have been used as classifiers to predict new samples in microarray data with considerable predicting accuracy (Li, J et al., 2002, Li, Dong et al., 2001, Dong et al., 1999). Therefore, JEPs can also act as very good classifiers to test classifier transferability.

Here, discriminating genes (DGs) gathered from one-gene jumping emerging patterns (JEP) were exploited as classifier to check the classifier transferability between any given dataset pair. We used the DGs to build a voting classifier as follows: For each discriminating gene g_i, suppose u_i is g_i's split value such that "$g_i \ \theta \ u_i$" is a one-gene JEP (where θ is either $\leq$ or $>$), and suppose C_1 and C_2 are respectively the majority classes of

g_i's low/high intervals. Then g_i can be used as a low-level classifier as follows: for an arbitrary tissue T, if $T(g_i) < u_i$, then T is predicted to be member of C_1; otherwise, T is predicted as member of C_2. We now have a low-level classifier for each of the DGs. The voting classifier's final decision of T's membership is reached by the voting result of all of the DGs.

In assessing classifier transferability, for each dataset D (or lab/platform combination) we get the discriminating genes as a classifier. Then we test the classifier on any other dataset D'. We use the accuracy of the classifier on D' as the numerical measure for classifier transferability.

5.3.4. Discretized-bin consistency rate between dataset pair

We also propose another concept called discretized bin consistency rate, or consistency rate (CR) to measure the concordance between (datasets generated from different) laboratories/platforms. The CR between a given pair of datasets is defined as the percentage of binary bits (interval values, where each interval is a bin in entropy-based discretization (Dougherty et al., 1995) for the two classes in a given dataset) whose values are consistent between the two datasets. For example, in Table 5.2, there are 4 tissues and 3 genes in each of the datasets D and D'. g_2's expression value in t_3 is "0" in D but it is "1" in D'. This is the only inconsistency between D and D'; all other corresponding bits in matching tissues for matching genes are consistent. The total number of bits in each dataset is 4 * 3 = 12, and the number of inconsistent bits is 1. So the consistency rate between this dataset pair is 11/12 = 91.7%.

Table 5.2: Discretized microarray dataset pair (D & D') sample

	D				D'			
	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
g_1	0	0	1	1	0	0	1	1
g_2	1	1	0	0	1	1	1	0
g_3	0	0	1	1	0	0	1	1

$t_1 \ldots t_4$: tissues; $g_1 \ldots g_3$: genes; '0': the gene expresses low; '1': the gene expresses high.

As mentioned earlier, discriminating genes have been proven to be of great importance, because they contain the most valuable information for classification in each specific dataset. Our hypothesis is that working on the discriminating genes may give us more accurate as well as useful results for analyzing the relationship between different datasets. On the other hand, the expression value of non-discriminating genes may be inherently random for biological reasons, and is unimportant or less important for most medical/biological studies. When used in concordance analysis, the non-discriminating genes may contribute a large proportion of inconsistency. In this study, two methods of calculating the consistency rate between a dataset pair are performed and compared, (1) one using all genes while (2) the other one using the discriminating genes only.

5.3.5. Calculation of *P*-value

Another criterion to check cross lab/platform concordance is *P*-value; *P*-value is used to measure how much evidence we have against the null hypotheses (which states that there is no difference between two given datasets). Small *P*-values suggest that the null

hypothesis is unlikely to be true. Traditionally, researchers will reject a hypothesis if the *P*-value is less than 0.05. Here, the *P*-value is supposed to quantitatively show how similar two given datasets are. If *P*-value is less than 0.05, the two datasets are considered as coming from different populations, i.e. they are not concordant with each other. We calculate the *P*-value by using the permutation test, which involves the random exchange of data between two datasets in order to determine the relationship between them.

For concordance analysis, we focus on the set of common discriminating genes between two given datasets D_1 and D_2. Using methods discussed earlier, we can find a set of DGs from the two classes in D_1, and similarly a set of DGs from D_2. The genes in both DG sets were defined as the common DGs. The rationale for our permutation-based approach is: if two given datasets are similar, then a permutation of samples (for identical tissues) between them will cause very small change to the set of common DGs. In other words, the set of common DGs for the two datasets after the permutation should be very similar to the set for the original two datasets. Thus, by comparing the set of common DGs from the dataset pair before and after the permutation, we can detect whether the original dataset pair is concordant or not. We perform a sequence of permutation tests and get a series of sets of common DGs, and use that series to derive the *P*-value (Note: While a large number of common discriminating genes is an indication that the two datasets are similar, that number alone cannot give us the confidence interval, but the *P*-value would be able to do so.)

More specifically, we did the following: (i) we identified the set of common discriminating genes (CG_o) from the original dataset pair; (ii) we did the same for the dataset pair in the result of random permutations; (iii) we repeated step (ii) m times and

got m sets ($CG_1 \ldots CG_m$) of common DGs; (iv) let $F = |CG_o|$, $F_i = |CG_o \cap CG_i|$ these numbers ($F_1 \ldots F_m$) had a mean value (*mu*) and standard deviation (sigma); (v) finally, we calculated the *P*-value from '*F*', '*mu*' and 'sigma' using Chebyshev inequality and tested whether (*F*) and ($F_1 \ldots F_m$) came from the same population. Let *f* denote the random variable of the number of common DGs that occur in CG_o. (Observe that $F = |CG_o \cap CG_o|$.) We used the Chebyshev inequality to estimate (an upper bound of) the *P*-value, which is a frequently used method (Saw J., 1984), as follows:

$$P(|\ f - mu\ | \geq |\ F - mu\ |) \leq \frac{1}{((F - mu)\ /\ sigma\ )^2}$$

This estimates the area of the curve in Figure 5.1 where |*f-mu*| is larger than or equal to |*F-mu*| (i.e. situations where *f* is more extreme than *F*), which gives us a confidence interval on a given dataset pair's concordance.

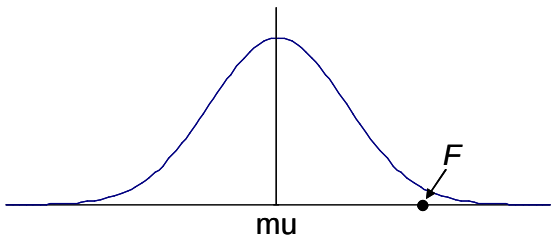

Figure 5.1: Possible distribution of f after many permutations between a dataset pair

5.3.6. Cross platform comparison

The concordance of four different platforms (ABI, AFX, AG1 and GEH) is evaluated. Essentially the same methods as described above are used, except that, in order to make the comparison of data from two platforms meaningful, the common genes used in the two platforms should be determined and the gene expression values should also be

normalized (see below).

Gene probe matching: Different platforms may use different gene probes. Thus, we need to find the matching genes between different platforms in order to test their concordance. To find matching gene probes, we use Unigene IDs. UniGene has been widely used to match genes on different microarrays, and UniGene IDs are considered to be the common gene identifier between platforms (Wang et al., 2005). In this study, utilizing UniGene IDs, 16140 common genes are identified as being present on all four of the analyzed platforms. Gene expression values are averaged in cases where multiple probes for a given UniGene ID are present on the chip.

Per-gene baseline adjustment: The gene expression values generated using different platforms can not be directly compared because of different labeling methods and different probe sequences used which may give rise to variable signals for the same target (gene). A per-gene baseline adjustment is performed to normalize these datasets.

Suppose datasets D_1 and D_2 share m genes ($g_0 \ldots g_{m-1}$) and n tissues ($t_0 \ldots t_{n-1}$). Let $V_1(g_k, t_j)$ denote gene g_k's expression value at tissue t_j in D_1 and $V_2(g_k, t_j)$ denote the same in D_2 where $0 \leq k < m$ and $0 \leq j < n$. Define:

$$MaxD_{i,k} = Max(V_i(g_k, t_j) \mid 0 \leq j < n)$$

$$MinD_{i,k} = Min(V_i(g_k, t_j) \mid 0 \leq j < n)$$

We use the following formula to generate dataset D_1'.

$$V_1'(g_k, t_j) = \frac{V_1(g_k, t_j) - MinD_{1,k}}{MaxD_{2,k} - MinD_{2,k}} + MinD_{1,k}$$

A similar formula is applied to $V_2(g_k, t_j)$ to generate dataset D_2', where we exchange the subscript 1 with the subscript 2. The concordance between D_1 and D_2 can be investigated

by checking D_1' and D_2 (or equivalently D_1 and D_2') using classifier transferability, consistency rate and *P*-value, as discussed in previous sections.

5.3.7. Absolute distance (AD) and artificial data

For concordance analysis, it is desirable to have a quantitative measure of the amount of difference between any two given datasets. Here we define one such measure called absolute distance (AD):

Suppose D and D' are two datasets which share m genes (g_0 …g_{m-1}) and n tissue (t_0…t_{n-1}). Let $V(g_i, t_j)$ denote a gene g_i's expression value at tissue t_j in dataset D and $V'(g_i, t_j)$ denote the same in D'; let R_i denote g_i's expression value range in D, that is, $R_i = \max(V(g_i, t_j) \mid 0 \le j < n) - \min((V(g_i, t_j) \mid 0 \le j < n)$. The absolute distance from D to D' (denoted $AD_{D \to D'}$) is:

$$AD_{D \to D'} = ((\sum_{i=0}^{m-1} (\sum_{j=0}^{n-1} \frac{|V(g_i, t_j) - V'(g_i, t_j)|}{R_i}) / n) / m) * 100\%$$

The absolute distance (AD) between D and D' is defined as:

$$AD = (AD_{D \to D'} + AD_{D' \to D}) / 2$$

R_i is an important feature for g_i in dataset D. For every tissue, $V(g_i, t_j)$ is in $[\min((V(g_i, t_j) \mid 0 \le j < n), \min((V(g_i, t_j) \mid 0 \le j < n) + R_i)]$. In D, $V(g_i, t_j)$ varies less than R_i from tissue to tissue. If the other dataset D' is highly concordant with D, then, in every tissue, $V'(g_i, t_j)$ should be equal or very close to $V(g_i, t_j)$. On the other hand, if D' is not concordant with D, the difference between $V(g_i, t_j)$ and $V'(g_i, t_j)$ may be larger than R_i. Thus, the relative difference between $V(g_i, t_j)$ and $V'(g_i, t_j)$ (normalized by R_i) is a good criterion to define absolute distance.

Apparently, if for every gene g_i, if $|V(g_i, t_j) - V'(g_i, t_j)| = 0$, then AD between D and D' is

0, which means that D and D' are highly concordant; if $|V(g_i, t_j) - V'(g_i, t_j)| = R_i$, then AD between D and D' is 100%. In this situation, according to our results, the *P*-value between D and D' is less than 0.01, which means D and D' are not concordant. Therefore, this absolute distance measure can be used to indicate how concordant a dataset pair is, from highly concordant to non-concordant.

In order to evaluate the absolute distance between any dataset pair, we create a series of artificial datasets with known absolute distance from a given original dataset. These artificial data may serve as benchmarks to identify how far apart the tested dataset pair is. The strategy for creating artificial data is as follows: for a given original dataset D_{ori} and one known absolute distance AD_k, we create the artificial dataset D_{modk} which differs from D_{ori} by AD_k by modifying every gene's expression value $V(g_i, t_j)$ in D_{ori} with value $\Delta_{i,j}$. $\Delta_{i,j}$ is randomly chosen and satisfy the following two constraints:

(i). $-R_i \leq \Delta_{i,j} \leq R_i$ for gene g_i,

(ii). $AD_k = (\sum_{i=0}^{m-1}(\sum_{j=0}^{n-1}\frac{\Delta_{i,j}}{R_i})/n)m * 100\%$, where R_i is gene g_i's expression range in D_{ori}.

In this study, 10 artificial datasets ($D_1 \ldots D_{10}$) are generated from each D_{ori}. The AD between each D_{ori} and the 10 associated D_{modk} are 10%, 20%, ….100%, respectively. A total of six D_{ori} datasets are used for creating D_{modk}. Each of the first four of these D_{ori} datasets is constructed from the three repeats of one of the four platforms by averaging as follows: For each platform, let $V_m(g_i, t_j)$ denote the gene expression value of t_j on g_i in the m^{th} repeat of the given platform, and let $V(g_i, t_j)$ denote the same in the D_{ori} dataset to be constructed from the three repeats; then $V(g_i, t_j) = (V_1(g_i, t_j) + V_2(g_i, t_j) + V_3(g_i, t_j))/3$. The fifth D_{ori} is obtained by (1) normalizing (as discussed earlier) and (2) averaging the expression values of those genes shared by all four platforms of the previous four D_{ori}

datasets. The sixth D_{ori} is created randomly, where the "expression value" of each gene at each tissue is randomly generated.

5.4. RESULTS

5.4.1 Concordance test by classifier transferability

We used classifier transferability to evaluate both cross-lab concordance and cross-platform concordance.

For cross-lab concordance, let D and D' be two datasets respectively generated by two laboratories using a common platform. We mined discriminating genes from each of the two datasets, say D, and then used them to build a classifier to predict the class of samples in the other dataset D'. The classifier's prediction accuracy is always 100%. This means that microarray datasets from different laboratories using any of the four given platforms as a common platform are highly concordant.

Next, the same method was applied to test the cross-platform concordance. Table 5.3 shows the results. The values in the table are the average accuracy achieved by two classifiers, each of which was built from data generated in one of the two platforms to predict tissue samples generated in the other of the two platforms. Our results indicate that the three platforms (AFX, ABI and GEH) are highly concordant with each other, but the AG1 platform is less concordant with the other three platforms.

Table 5.3: Classifier transferability between platforms

	AFX	AG1	GEH
ABI	100%	62.5%	100%
AFX		62.5%	100%
AG1			60%

5.4.2 Consistency rate (CR) analysis

In order to figure out the influence of noise to microarray dataset's concordance, and explain the discordant results in previous papers, the consistency rate (CR) is estimated before and after the noise genes are removed. As discussed earlier, CR is calculated in two ways involving two different sets of genes, where one way uses all genes, and the other uses the discriminating genes only. The results are listed in Tables 5.4 and 5.5.

Table 5.4 shows the consistency rate in intra-platform comparison. If all genes are included, the consistency rate is around 75% for any two labs within one platform (left); however, if only the discriminating genes are considered, the consistency rate is between 92% and 98.5% for any dataset pairs (right).

Table 5.4: Consistency rate between laboratories

	All genes			**Discriminating genes**		
	L_1 vs L_2	L_1 vs L_3	L_2 vs L_3	L_1 vs L_2	L_1 vs L_3	L_2 vs L_3
ABI	73.5%	74.6%	73.4%	98.3%	98.2%	98.5%
AFX	73.5%	74.0%	75.7%	97.9%	97.3%	98.1%
AG1	79.3%	76.1%	77.0%	98.0%	98.2%	97.0%
GEH	64.6%	72.6%	66.0%	92.7%	97.3%	92.5%

L_1, L_2 and L_3 are datasets generated from three different laboratories; each entry in the table is the consistency rate (CR) between the corresponding laboratory pair

Table 5.5 shows the consistency rate in cross-platform comparison by using either all genes or discriminating genes only. The discriminating gene-based CR between any dataset pair is considerably higher than all-gene-based CR except for cases where AG1 is involved. This result is consistent with classifier transferability result in Section 5.4.1. It

should be noted that the 50% CR is what would be achieved for a pair of random datasets after discretization, which has been confirmed by our experiments.

Table 5.5: Consistency rate across platforms

	All genes			**Discriminating genes**		
	AFX	AG1	GEH	AFX	AG1	GEH
ABI	73.5%	50.0%	69.1%	88.7%	50.4%	83.7%
AFX		50.0%	69.9%		50.2%	84.1%
AG1			49.6%			50.4%

Each entry in the table is the consistency rate between a platform pair

5.4.3 Permutation and *P*-value

To quantitatively measure the concordance between two platforms/laboratories, *P*-value is calculated directly between the pairs of datasets generated by the two platform/laboratories. The results are shown in Tables 5.6 and 5.7.

Table 5.6: *P*-value between intra-platform dataset pair

	L_1 vs L_2	L_1 vs L_3	L_2 vs L_3
ABI	0.3097	0.3068	0.3950
AFX	0.5222	0.6449	0.5109
AG1	0.4099	0.3435	0.3604
GEH	0.2017	0.3475	0.2102

According to the calculated *P*-values, there is no statistical significance ($P > 0.05$) between different laboratories using a common platform (see Table 5.6), which implies that the laboratories are concordant with each other if they use the same platform. The *P*-values for the dataset pairs from different platforms are shown in Table 5.7. There is no statistical significance between the dataset pairs from platform ABI, AFX and GEH ($P > 0.05$), which again implies concordance; however, the dataset from AG1 is significantly

different from the other three platforms ($P < 0.05$), which implies non-concordance.

Table 5.7: *P*-value between cross-platform dataset pair

	AFX	AG1	GEH
ABI	0.198	0.011	0.201
AFX		0.012	0.286
AG1			0.010

5.4.4 Absolute distance (AD) as bridge

It takes much time to calculate reliable *P*-value directly by permutation due to two reasons: (i) In order to get reliable *P*-value, a fairly large number (at least hundreds) of permutations are required. (ii) For each resulting dataset pair from these permutations, we need to compute the discriminating genes by examining up to n^3 sets of genes, where n is the total number of genes. Therefore, to get *P*-value between a dataset pair directly by permutation is not very effective. It will be desirable if we can find an alternative way to obtain *P*-value such as from known consistency rate. In our approach, we use some artificial datasets with known degree of absolute distance (AD) to serve as a bridge to relate *P*-value and consistency rate (CR). Ideally we want to have a one-to-one correspondence between AD and CR, and between AD and *P*-value.

A. Repeatability of artificial datasets

Since the artificial datasets were created randomly (see Section 5.3.7), for a given AD one may create different datasets at different times. This may cast doubt on whether the measures and the correlations generated from such artificial datasets are completely repeatable. Therefore, the repeatability of the correlation between AD and CR/*P*-value derived from such artificial datasets needs to be checked.

Six series of artificial datasets from six original datasets are used to check the repeatability of the correlation between AD and CR, and between AD and *P*-value. Figures 5.2 and Figure 5.3 show the results. Overall, when discriminating genes were used to calculate CR and *P*-value, the relationships between AD and CR, and between AD and *P*-value were very reproducible and independent of the dataset used, and the correlations were also quite close to a one-to-one correspondence. This is the case despite the fact that the six original datasets are very different from each other; in fact, different datasets (from different platforms) may use sets of genes, the gene expression value ranges vary considerably between datasets, and one of the six original datasets was generated randomly.

The CR between each artificial dataset and its original dataset was calculated. (Recall that each artificial dataset was generated from its original dataset for a given AD) Six series of artificial datasets produce six CR/AD plots (Figure 5.2). The repeatability of CR/AD plots is confirmed by the almost exact overlapping of these plots. The maximum CR difference between the six curves is less than 3% at each absolute distance (X-axis). Thus, this correlation is considered as very repeatable.

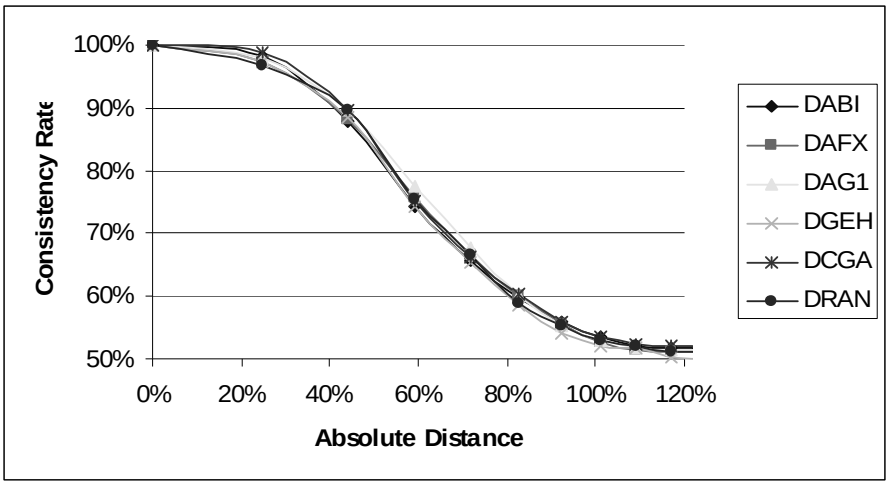

Figure 5.2: Correlation between absolute distance and consistency rate measure, and repeatability

Each curve was generated from one of the six series of artificial datasets. The name of the curve corresponds to the Dori of the series of artificial datasets; CGA in DCGA stands for Common Gene Averaging, and RAN in DRAN stands for RANdom.

P-values were calculated by permutation test on the six series of artificial datasets described above. Six series of *P*-value/AD curves are plotted (Figure 5.3). When the AD is greater than 60%, the six curves are highly similar to each other. Though there were minor differences between *P*-values in these curves when AD is less than 60%, we are still able to draw consistent conclusion regarding the dataset pair concordance ($P>0.05$). Since all *P*-values were created by random permutation, the more repeats of permutation we do, the more reliable the *P*-value will be. In this study *P*-values were calculated by 100 times of random permutations.

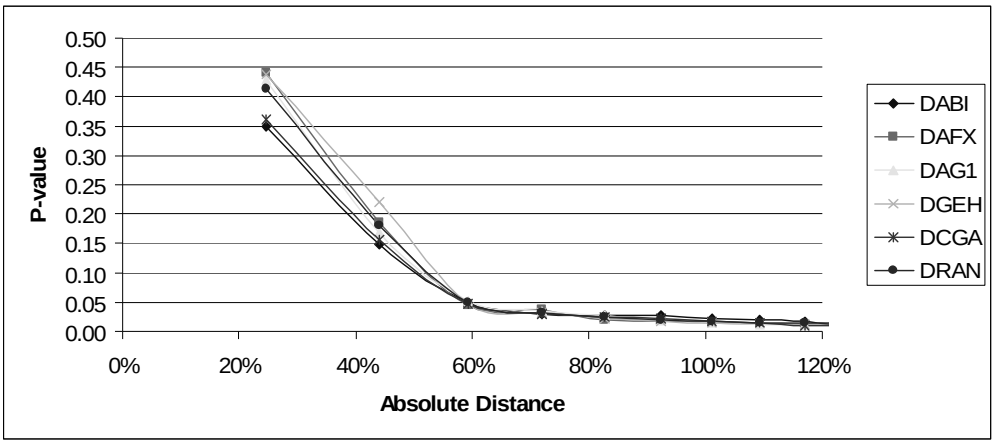

Figure 5.3: Correlation between absolute distance and *P*-value, and repeatability

B. *P*-value vs consistency rate curve

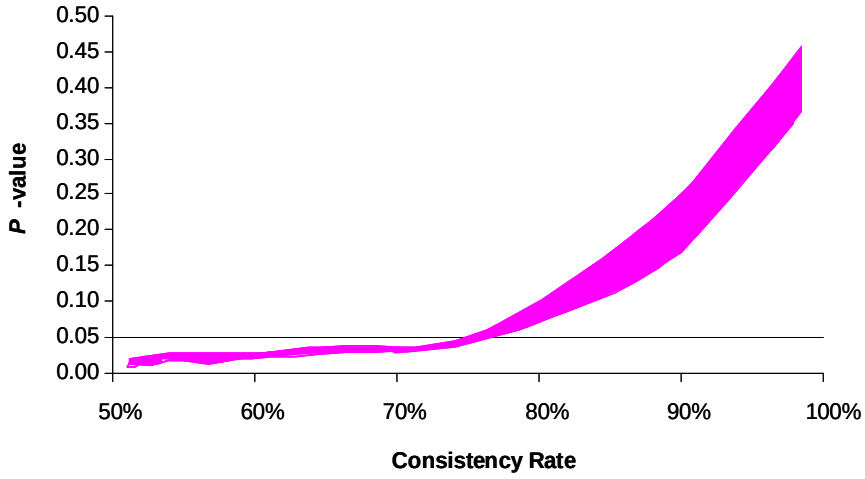

Figure 5.4: Consistency rate vs *P*-value

AD has a near one-to-one correspondence with CR (Figure 5.2), and also good correlation with *P*-value (Figure 5.3). Although the correlation between *P*-value and CR, established using AD, is not exactly a one-to-one function, AD is still be a valuable way to tell us how to link consistency rate and *P*-value, by means of controlled data

modification. More specifically, from any given CR we can identify a unique AD (Figure 5.2), and from each AD value we can identify a small range of *P*-values. The correlation between CR and *P*-value is plotted in Figure 5.4; Table 5.8 lists the correlation in table form for ease of reference.

This curve can become useful tool for estimating *P*-value if consistency rate between dataset pair is known. First, if a CR is given, the range of the corresponding *P*-value can be estimated using Figure 5.4 (the range is specified by the pink area). For example, given a CR of 90% between a dataset pair, the corresponding *P*-value should be between 0.15-0.25 (and so this dataset pair is concordant). Second, Figure 5.4 shows that if the CR between a dataset pair is lower than 73%, the *P*-value between that dataset pair is lower than 0.05, which implies significant difference, or the dataset pair is not concordant.

Table 5.8: CR vs *P*-value

CR	*P*-value	
	Lower bound	Higher bound
97.9%	0.3484	0.4396
88.4%	0.1476	0.2206
75.4%	0.0447	0.0515
66.3%	0.0307	0.0378
59.7%	0.0200	0.0276
55.4%	0.0177	0.0267
52.9%	0.0150	0.0225
51.9%	0.0129	0.0198
51.3%	0.0103	0.0184

C. Accuracy testing of *P*-value vs CR curve

To check the accuracy of the CR/*P*-value plotted in Figure 5.4, the *P*-values were estimated using the CR listed in Tables 5.4 and 5.5. Next, the estimated *P*-values were compared with the actual *P*-value shown in Tables 5.6 and 5.7. As expected, the results

were consistent with each other. This suggested the plot we got is accurate.

In order to further test the accuracy of the *P*-value/CR plot, five pairs of datasets are created randomly. The CR and *P*-value between each dataset pair are investigated. Moreover, the *P*-value is estimated from CR using the *P*-value/CR plot in Figure 5.4. Table 5.9 shows the result. One can see that the actual *P*-value and estimated *P*-value are also matched well, which suggests that the plot in Figure 5.4 is accurate.

Table 5.9: Comparison of *P*-value from randomly dataset pairs

Dataset pairs	CR	*P*-value from permutation	Derived *P*-value from CR
1	58.7%	0.01	0.008 - 0.01
2	55.5%	0.0093	0.008 - 0.01
3	92.5%	0.2465	0.019 – 0.270
4	84.4%	0.1026	0.08 – 0.11
5	73.7%	0.0483	0.046 – 0.049

5.5 DISCUSSION AND CONCLUSION

In this chapter, three comparative analysis methods were introduced to test the microarray datasets' concordance. These methods were designed for comparative studies and focused on discriminating genes in the datasets: classifier transferability tests the degree of concordance between platforms/laboratories; consistency rate detects concordance (consistency) between the datasets before and after most of the gene-level noise is removed (each gene value is discretized to have only 0 and 1); and *P*-value quantitatively measures the concordance between platforms/laboratories. In the *P*-value based method, numerous and random permutations between datasets were conducted before the concordance decision is made. Analyzing the datasets provided by MAQC

project, our results showed that microarray data from different labs within any one of the four platforms are fairly concordant at discriminating gene level; moreover, cross-platform microarray data derived from ABI, AFX and GEH are also concordant. Our methods are fairly general, requiring just two tissue types and can be applied to cases with two or more tissue types or classes.

Our methods have been successfully applied to the datasets with two tissue types. If there are more than two tissue types in any given datasets, our methods can still be applied. We have two strategies to handle such datasets. (1) We can randomly divide the tissues into two classes, and then apply our methods. (2) We can also apply our methods for each pair of tissue types in the dataset pair, and combine the results for all tissue type pairs together and find out the concordance between the dataset pair. (If there are n tissue types in the given dataset, then our methods should be applied n * (n - 1) / 2 times.) Compared with the first strategy, this strategy is more accurate, although more time consuming.

Noise exists in microarray data. However, having minor differences doesn't mean that two datasets must be inconsistent. Some studies (Kuo et al., 2002, Kothapalli et al., 2002) concluded that the datasets from different platforms are not concordant. They reached this conclusion by comparing all genes in the microarray data, without filtering out the noise. In our study, we calculated consistency rate between dataset pairs from different laboratories and different platforms. If all genes were included, we got conclusion that the microarray data from different platforms have poor reproducibility. However, when only considering discriminating genes for analysis, we found that, besides very high concordant rate within platforms, 3 out of 4 platforms are also concordant with each other except the AG1 platform.

Recent publications derived from the MAQC project used some other methods to conclude that all of the 4 platforms are fairly concordant with each other (Shi et al, 2006, Guo et al, 2006). Fold-change ranking with a nonstringent *P*-value cutoff method was used to rank the genes in these studies. Only the top ranked genes were selected for platform-concordance analysis for noise removal. After noise removal from the microarray datasets, it was found that the microarray datasets from different laboratories/platforms are fairly concordant.

Guo and colleagues also noticed the importance of discriminating genes in concordance evaluation, although they used the fold change method to discover DGs. It should be pointed out that the fold change method uses ratio of average of values in the two classes, whereas the entropy-based method aims to find split values to separate the classes as cleanly as possible. So their criteria for discriminating gene selection are very different from ours. We believe that the entropy-based discriminating gene selection method used in this paper is more biological meaningful. Instead of arbitrarily selecting highly differentially expressed genes, we selected the genes that participate in discriminating gene groups; such groups are highly related to one of the classes in the dataset. Thus, these selected genes may provide insights on gene interaction networks, or even pathways for the specific disease. Moreover, the entropy based method (Dougherty et al., 1995) has been widely used in the data mining and machine learning communities (Han et al, 2006).

It was realized that different gene selection criteria led to different concordance results. Guo et al used six different gene selection methods to choose and rank genes. They got different concordance results from their different gene selection methods. In this study,

we used comparative methods to select discriminating genes for platform concordance testing. Compared with their results, we reached the same conclusion that the datasets within a given platform are highly concordant, and that the non-AG1 platforms are fairly concordant with each other. However, we assessed that the AG1 platform has low concordance with other platforms according to the datasets provided by MAQC project, whereas Guo et al concluded that the AG1 platform is also concordant with other platforms. While the agreement results reinforce previous results and indicate that most platforms are concordant with each other, the disagreement was the results of using different gene selecting and testing methods.

Our conclusion can be helpful for the future comparative studies. For example, since AG1 and other platforms are not concordant under comparative discriminating gene selection, the datasets from AG1 can't be analyzed together with the datasets from other platforms in comparative studies, such as discriminating gene data mining. On the contrary, these datasets can be analyzed together under fold-change ranking with a nonstringent *P*-value cutoff method.

It should be noted that, the concordance results reported in this paper were based on datasets from MAQC, which were derived from well controlled experiments and identical reference RNAs. In real world situations one also needs to consider lab-practice variability and biological variability, which will be discussed in Chapter 6.

5.6 APPENDIX

Since discriminating genes play a key role in microarray dataset, it is tempting to use the percentage of number of common discriminating genes (which we refer to this approach as the basic discriminating gene transferability) to directly measure the concordance

between a dataset pair, or to use the binning cut-point difference of common discriminating genes to test the concordance. If one or both of these two methods can serve as such a measure, the concordance between a dataset pair can be easily determined. Unfortunately, none of these methods work for microarray gene expression data, because for concordance determination, both methods are sufficient condition, but not necessary condition. In other words, two datasets with a small number of common discriminating genes or with a large cut-point difference may still be highly concordant.

We created four datasets pairs with controlled degree of concordance to test the above two methods. D_k and D_k' is one dataset pair, where $0<k\leq4$; N_k is the number of discriminating genes in dataset D_k; N_k' is the number of discriminating genes in dataset D_k'; C_k is the number of common discriminating gene in both D_k and D_k'. The CDG Ratio (common discriminating gene ratio which is mined using basic discriminating gene transferability) between one dataset pair is calculated by the following formula:

$$\text{CDG ratio}_k = (C_k * 2) / (N_k + N_k') * 100\%.$$

The CP Diff (cut point difference) is calculated as the following:

$$CP\ Diff_k = (\sum_{i=0}^{n-1}(|V_k(c_i) - V_k'(c_i)|/R_{ki}))/n * 100\%$$

where $V_k(C_i)$ and $V_k'(C_i)$ denote gene g_i's cut point at dataset pair k, and R_i is g_i's expression value range. The testing results are shown in Table 5.10.

Table 5.10: Comparison of dataset concordance using different methods

Dataset pairs	CP diff	CDG Ratio	Classifier transferability	Consistency rate	*P*-value
1	17.0%	81.6%	100%	99.5%	0.6563
2	17.8%	71.3%	82.5%	89.4%	0.1326
3	17.3%	72.2%	74.0%	78.6%	0.0527
4	18.2%	52.8%	50.5%	52.5%	0.0082

Four pairs of datasets were created in a controlled manner: dataset pair 1 is highly concordant between each other; dataset pair 2 is concordant; dataset pair 3 is on the boundary of concordance and dataset pair 4 is not concordant.

In Table 5.10, besides cut-point difference and CDG Ratio, the classifier transferability, CR and *P*-value result are also listed for comparison. According to the table, classifier transferability, CR and *P*-value can act as very good criteria to scale the concordance between dataset pair, but cut point difference and CDG ratio can't.

The following are weakness of these two methods. Suppose there are m tissues in dataset D. So each gene g_i has m expression values in D. (1) the basic discriminating gene transferability uses the number of common discriminating genes as criterion. For any discriminating gene g_i, if its expression value at one tissue is slightly modified so that it crosses the old binning cut point, then this gene might not be a discriminating gene anymore. Thus, one expression modification in one tissue may affect the whole gene's property, which makes this method very vulnerable. (2) g_i's cut point is highly dictated by the two expression values which are closet to the cut point and which envelope the cut point So the cut point difference between dataset pair is mostly dictated by those two expression values of g_i and is insensitive to the others. Thus, it can't represent the behavior of g_i, and it can't serve as an indicator for concordance test either.

Our analysis of the data on these two methods also indicated the following interesting observations: (1) most consistent genes between a dataset pair are discriminating genes. (2) Consistency rate for non-discriminating genes is positively correlated to the consistency rate for discriminating genes, although the consistency rate for the former is much lower than the latter.

Chapter 6: MINIMIZING VARIABILITY OF MICROARRAY DATASETS

6.1. MOTIVATION

During comparative microarray data mining, it is very important to get reliable patterns and models in order to gain high-quality understanding the mechanism of specific diseases. It has been well known that there is variability in microarray datasets and the variability may affect the data mining results. So far, few reports, if any, used comparative methods to study this topic. In this chapter, we investigate the effect of variability, develop novel methods to reduce the variability, and improve the reliability of the mined patterns and models.

6.2. OVERVIEW

Ideally, if there were no variability in microarray datasets, the mined patterns and models should represent the intrinsic features of the classes they belong to. They should be independent of the chosen patient samples, laboratories and platforms which were used to generate the microarray datasets. Roughly speaking:

- Data mining results from multiple microarray datasets concerning a common disease should be identical or highly similar;

- Data mining results from different subsets of one microarray dataset should be identical or highly similar.

Unfortunately, variability exists in microarray datasets. Such variability includes measurement variability, biological variation and so on (Chapter 2, Section 2.2). Microarray datasets are generated from different laboratories using different platforms. Experimental noise exists during the data generation of microarray experiments, which might cause inconsistent results and produce measurement variability. We have tested this effect of experimental noise in Chapter 5. Our results showed that, at comparative level, the datasets from a common platform, different labs are highly concordant.

At the same time, microarray datasets are typically generated using tissue samples from different patients. These patients have different characteristics such as height, weight, age, race, sex etc. These differences inevitably lead to biological variations in the microarray datasets and may affect data mining results and conclusions. Compared with measurement variability, which randomly distributes through the datasets, biological variation is sample specific. Some samples may bring high degree of biological variation to the dataset and thus influence the data mining results.

In this chapter, we consider how to test and minimize the variability by identifying and eliminating noisy tissue samples from datasets. The noisy samples may have high degree of biological variation, or they may have high degree of measurement variability.

More specifically, first we use four measurements to evaluate the degree of variability. Then we develop a novel method which is called Converged Leave-One-Out-Cross Validation (C-loocv) to identify and remove the samples which have high degree of the variability in microarray datasets. This process is called biological variability

minimization (BVM). Next we compare the data mining results from the original datasets with those from the C-loocv refined datasets. And finally we show the advantage of C-loocv application in classification.

6.3. MATERIALS AND METHODS

6.3.1. Microarray datasets for variability evaluation

In this chapter, two microarray datasets which study lung cancer will be used to evaluate the effect of variability on our data mining results. These two datasets were generated in different laboratories. One dataset was created by Bhattachariee and colleagues (Bhattacharjee et al, 2001) in Harvard medical school, which is called Harvard lung cancer dataset or Harvard in this study. The other one was created by Beer and colleagues (Beer et al, 2002) in University of Michigan, which is named Michigan lung cancer dataset, or Michigan in this study. Both datasets study lung cancer using samples from different patients. These two datasets made it possible to check the influence of variability on the data mining results.

There are totally 12,600 DNA probes (genes) and 203 samples in Harvard microarray dataset. 17 samples come from normal lung (Normal), and the other 186 samples are from lung cancer patients (Tumor). These tumor samples are further divided into five subtypes: (1) lung adenocarcinomas (n = 127); (2) squamous cell lung carcinomas (n = 21); (3) pulmonary carcinoids (n = 20); (4) small-cell lung carcinomas (n = 6) and (5) Other adenocarcinomas (n = 12). In our study, all tissue samples in five tumor subtypes are considered as lung cancer samples and categorized in one group. Thus, there are two classes in this dataset: Normal class (n = 17) and Tumor class (n = 186).

Compared with Harvard dataset, Michigan dataset is relatively smaller in size and simpler in structure. There are 7129 gene probes and 96 samples in Michigan dataset. Among 96 samples, 10 of them come from normal controls; the other 86 samples are from lung cancer patients. These cancer samples are separated into two subgroups: 67 are stage I and 19 are stage III lung cancer. Same as Harvard dataset, in our study, the cancer samples from two subgroups are categorized into one class. So there are also two classes in Michigan dataset: Normal class (n = 10) and Tumor class (n = 86).

Even though both lung cancer datasets were created using a common platform (Affymetrix), the total number and the order of gene probes on the two chips are different. Thus, the common genes between these two datasets should be identified. Same as Chapter 5, we will use UniGene IDs to identify the common genes between these two datasets. Gene expression values are averaged in cases where multiple probes for a given UniGene ID are present on the chip.

Because variability in microarray datasets includes measurement variability and biological variation, it is desirable to test the effect of measurement variability and biological variation separately. MAQC datasets make it possible. The details about MAQC data structure have been introduced in Chapter 5. Since all MAQC datasets use same mRNA as tissue samples, biological variation is supposed to have been eliminated. Thus, MAQC datasets should be very appropriate to test the effect of variability caused only by measurement variability.

6.3.2. Measurements of the degree of variability

It is difficult to measure the degree of variability in microarray datasets directly. So far, no study has been reported to define the degree of variability in microarray datasets; as a

result, no attempt has been made to measure it. Here, we define the degree of variability between datasets (see note below) as the similarity of data mining results mined from such microarray datasets. High degree of variability produces low similarity of data mining results, and vice versa.

In this chapter, we measure the similarity of data mining results mined from microarray datasets to evaluate the degree of variability. Roughly speaking, agreements on each of the following four types of data mining results obtained from datasets will be used to evaluate the degree of variability, (1) the sets of top 20 ranked genes from each dataset; (2) classifier transferability (CT) between datasets; (3) the sets of discriminating genes (DGs); and (4) the sets of top frequency HDGGs.

More specifically, first, we rank the genes in each dataset in decreasing information gain order and obtain the top 20 gene sets from each dataset. Meanwhile, we discover highly differentiative gene groups (HDGGs) from each dataset by using "iterative gene club formation algorithm" (Mao & Dong, 2005) and border differential algorithm (Dong et al, 1999). Recall that genes involved in HDGGs are discriminating genes.

Next, the similarity of the top gene sets between datasets is evaluated using Jaccard similarity coefficient (JSC). Identical criteria are also used to measure the similarity of DG sets and HDGGs sets. The Jaccard similarity coefficient (JSC) is calculated to check the similarity of two sets, which is defined as the size of the intersection divided by the size of union of the sample sets. Thus, the range of JSC is [0, 1].

Finally, the discriminating genes from each dataset are used as classifier to test classifier transferability between datasets. (Details about classifier transferability have been described in Chapter 5). The range of classifier transferability (CT) is [50%, 100%].

Note: the term “datasets” used in this chapter means the datasets which study a common disease. Such datasets are suitable for analyzing the degree of variability. If only one microarray dataset is under consideration, we make several subsets by randomly choosing samples from each class in this dataset. These subsets are used as “datasets” for variability analysis.

6.3.3. Converged Leave-One-Out Cross-Validation algorithm (C-loocv)

In microarray datasets, the tissue samples come from a variety of patients. Some samples may be highly different from the others in the same class. These noisy samples cause high degree of variability and should be eliminated from microarray datasets. Liu and colleagues (Liu et al, 2001) developed a so-called ICF algorithm to eliminate noisy samples from datasets. In this study, we develop a new algorithm (C-loocv) by improving Liu’s algorithm and apply it on microarray datasets. C-loocv is created by combining leave-one-out cross validation algorithm with our HDGG-based classifiers. In our microarray datasets, C-loocv will be applied to identify and eliminate noisy samples.

A. Leave-One-Out-Cross-Validation (LOOCV)

Leave-one-out cross-validation (LOOCV) has received much attention since it has been shown to give an almost unbiased estimator of the generalization properties of statistical models, and therefore provides a sensible criterion for model selection and comparison.

The original purpose of LOOCV is to compare the robustness of classifiers. In a given dataset, one test sample (an unobserved output value ‘y’) is left out and the other samples (an observed vector ‘x’) are used to train the classifiers. The trained classifiers are

applied to predict 'y'. The classifiers with better predicting accuracy are considered as more robust. Whenever a new classifier is developed, LOOCV is a common method to test this classifier's robustness by comparing with other classifiers.

During classifier's robustness comparison, it has been noticed that in some datasets, several specific samples could not be predicted correctly no matter which classifiers were applied. This fact led us to consider the reliability of the dataset itself other than that of the classifiers.

As mentioned before, many factors, such as biological variation, high risk normal controls, misdiagnosis patients and so on, may cause the test samples to be categorized into the incorrect class during microarray dataset generation. According to this fact, we will develop algorithms to identify and remove those wrong-predicted samples from datasets in order to improve reliability.

B. Building high accuracy HDGG-based classifiers

HDGG-based classifiers have shown improvement in predicting accuracy over the other classifiers. The algorithm for building HDGG-based classifier was mentioned in Chapter 4. Here, we discuss in detail how to build this classifier.

Suppose a cancer related microarray dataset D has n tissue samples ($t_0 \ldots t_{n-1}$). These samples can be divided into two classes according to its class type: "Normal" class C_N which has n_1 samples, and "Tumor" class C_T which has n_2 samples; the HDGGs are mined from C_N and C_T and will be used to build a HDGG-based classifier. Let's denote the ranked tumor HDGGs in C_T as $HDGG^T_1$, $HDGG^T_2$ … $HDGG^T_x$ in descending order of their frequency in C_T. Similarly, denote the ranked normal HDGGs in C_N as $HDGG^N_1$,

$HDGG^N_2$ … $HDGG^N_y$ also in descending order of their frequency. Next, some normal HDGGs and tumor HDGGs are chosen to build a classifier with gene diversity strategy as mentioned in Chapter 4, i.e. each discriminating gene occurs no more than once in the chosen HDGGs. For example, if one discriminating gene 'g' appears in $HDGG^T_1$, then all other $HDGG^T$s which contain 'g' are not allowed to serve in the classifier.

Suppose the following are the HDGGs selected for use in the classifier, using the gene diversity strategy:

Tumor class (C_T): $HDGG^T_{x1}$, $HDGG^T_{x2}$…$HDGG^T_{xi}$, where $x_1<x_2<x_i\leq x$

Normal class (C_N): $HDGG^N_{y1}$, $HDGG^N_{y2}$…$HDGG^N_{yj}$, where $y_1<y_2<y_j\leq y$

In order to predict the class type of a test sample T, the score of the class label of T needs to be calculated. Suppose we use k (k < i and k < j) top-ranked HDGGs from C_T and C_N in the scoring process. Then we define the score of T in the C_T class as

$$S(T)\_C_T = (\sum_{z=1}^{k} \frac{frequency(HDGG^T_{xz})}{n_2}) / k$$

and the score of T in the C_N class as

$$S(T)\_C_N = (\sum_{z=1}^{k} \frac{frequency(HDGG^N_{yz})}{n_1}) / k$$

Finally, the test sample T's score is: score(T) = S(T)_C_T – S(T)_C_N.

According to the formula, one can see that for any sample T, $-1 \leq score(T) \leq 1$. If score(T) > 0, then sample T is predicted as tumor; if score(T) < 0, then T is predicted as normal; score(T) = 0 means we have a tie and T's class is difficult to decide. If score(T) is close to 1 or -1, T typically belongs to the corresponding class (Tumor or Normal).

The parameter k can be determined by the user based on the available computation power. In this research, in order to obtain the maximum information from the training datasets and make the classifiers as diversity as possible, we keep the number of HDGGs in classifiers as large as possible. Thus, in this study, k = i (if i < j), or k = j (if j < i); in other words, k is the minimum of the numbers of top-ranked HDGGs for the two classes.

C. The Converged loocv (C-loocv) Algorithm:

The C-loocv algorithm is designed to identify noisy samples by comparing each testing sample's score using HDGG-based classifiers. This design requires that the scores from every sample (T) be comparable. We modify the definition of score(T) in order to fit the C-Loocv algorithm:

For any sample t_i, if t_i's original class is "Tumor", the score of t_i is defined as:

$$score(t_i) = S(t_i)\_C_T - S(t_i)\_C_N$$

Otherwise, if t's original class type is "Normal", then:

$$score(t_i) = S(t_i)\_C_N - S(t_i)\_C_T$$

This modification ensures that tumor samples and normal samples can be compared together.

The following is C-loocv algorithm.

Suppose the given cancer related microarray dataset D has n tissue samples ($t_0 \ldots t_{n-1}$).

Step 1: Every sample t_i in D is separated as test sample (so the rest are used as training samples); this yields a score(t_i);

Step 2: Rank the samples in descending order of their scores;

Step 3: Remove the samples whose scores are less than threshold τ;

Step 4: In the remaining samples, repeat steps 1 to 3 until every sample's score is larger than τ. Now the "core dataset" is obtained;

Step 5: Use the core dataset as training data to build classifiers and to predict the discarded samples in step 3 and rank them in descending order;

Step 6: Restore the samples whose scores are larger than τ, and permanently discard the samples whose scores are less than τ. Suppose the number of discarded samples is n_1, then the number of remaining samples in D is $n - n_1$;

Step 7: Repeat steps 1 to 6 on D until the dataset reaches steady state, i.e. no more samples are added or removed; we have now reached the converged state.

In this study, the threshold τ is set to 0. We can adjust the threshold in order to find the best optimization for any given datasets. According to score(T)'s formula, if τ is set larger than 0, then some test samples which weakly belong to their class are also discarded; if τ is set less than 0, the test samples which are slightly predicted wrong are kept in the dataset.

6.4. RESULTS OF EVALUATING OUR METHOD

As mentioned before, four measurements are used here to test the degree of variability in datasets. We focus on two measurements: one is the sets of top 20 genes, the other one is classifier transferability (CT). The other two measurements (sets of discriminating genes (DGs) and sets of HDGGs) are used as reference.

6.4.1. Degree of measurement variability

Since variability includes measurement variability and biological variation, it is desirable to test measurement variability alone. According to MAQC style dataset structure, there is no biological variation in MQAC datasets. Thus, we can test the degree of variability caused only by measurement variability using these datasets.

Two datasets (ABI_L1 and ABI_L2) from ABI platform are chosen. Using entropy-based gene ranking method, several thousands of genes have equally information gain and can be ranked as top genes simultaneously. We further re-rank these top ranked genes according to their gene index number in their respective datasets (small index number ranks higher). Table 6.1 shows the top 20 genes from each dataset.

Table 6.1: Top 20 genes in two datasets generated with ABI platform

Rank order	ABI_L1	ABI_L2
1	**1**	**1**
2	**4**	**4**
3	**5**	**5**
4	**6**	**6**
5	**7**	**7**
6	**10**	**10**
7	15	13
8	**16**	**16**
9	**17**	**17**
10	**18**	**18**
11	**24**	**24**
12	**26**	**26**
13	**27**	**27**
14	**31**	**31**
15	**33**	**33**
16	**34**	**34**
17	**37**	**37**
18	**44**	39
19	47	42
20	55	**44**

According to Table 6.1, we notice that the sets of top 20 genes from two datasets are quite similar. 85% of genes in the top 20 gene sets are identical (bold font). The Jaccard similarity coefficient (JSC) between these two sets is 0.739.

Classifier transferability (CT) between these two datasets is also calculated: CT = 100%.

We also calculate the JSC value from the sets of DGs and the sets of HDGGs. JSC = 0.676.

According to these results, we conclude that measurement variability exists in the datasets and moderately affect the top 20 gene ranking. But it has no effect on classifier transferability. This is consistent with the results in Chapter 5.

6.4.2. Artificial dataset with maximal degree of variability

In Section 6.4.1, we tested our measurements using MAQC datasets which have minimal degree of variability (no biological variation). We also need to test our measurements with the datasets which have maximal degree of variability. Clearly, for any two datasets, if JSC = 0 and CT = 50%, then the degree of variability in those datasets is maximal.

In order to obtain such datasets, we randomly swap several samples between two classes in one dataset. In ABI_L1 dataset, we exchange several samples in t_1 … t_{10} with the corresponding samples in t_{11} … t_{20} (see Table 5.1). After sample swap, the modified dataset is named as ABI_L1'. Top 20 genes are ranked from two datasets (ABI_L1 and ABI_L1'). As expected, there is no common gene shared by the two sets of top 20 genes (JSC = 0 and CT = 50%). In fact, after we exchange only four samples (t_1 vs $t_{11}$and t_2 vs t_{12}), we have gotten maximal degree of variability. The top 20 genes in both sets have already become completely different (Table 6.2).

Table 6.2: Top 20 genes before vs after sample swap

Rank order	ABI_L1	ABI_L1'
1	1	11535
2	4	441
3	5	500
4	6	580
5	7	666
6	10	694
7	15	792
8	16	1086
9	17	1334
10	18	1381
11	24	1550
12	26	1565
13	27	1669
14	31	1711
15	33	1802
16	34	1828
17	37	1942
18	44	2277
19	47	2809
20	55	2882

In Table 6.2, the second column is the top 20 genes from original ABI_L1 dataset. The third column is the top 20 genes from dataset ABI_L1', in which two pair of samples are swapped.

We use our C-loocv algorithm to identify noisy samples from ABI_L1'. These swapped samples can easily be identified and eliminated. After noisy sample removal, the data mining results from the refined datasets become highly consistent.

6.4.3. Variability in lung cancer datasets

As mentioned before, both Harvard and Michigan datasets study lung cancer and both of them were generated using the Affymetrix platform. We have proved that datasets

without biological variation, which come from a common platform and different laboratories, are highly concordant (see Chapter 5). Thus, if no biological variation exists in the two lung cancer datasets, the top 20 gene sets from both datasets should be identical or very similar, and classifier transferability (CT) between these two datasets should be 100% (see Section 6.4.1).

Top 20 gene sets: Table 6.3 shows the sets of top 20 genes mined from Harvard and Michigan datasets. The number in the table is the gene's index number, i.e, each number represents one specific gene.

Table 6.3: Top 20 genes in Harvard and Michigan lung cancer dataset

Rank order	Michigan	Harvard
1	**845**	**2910**
2	**1344**	640
3	**1814**	**2688**
4	2040	**2607**
5	**2910**	4351
6	**3895**	5180
7	711	3284
8	2275	**845**
9	**2607**	**1344**
10	**2688**	2555
11	3581	1981
12	**3945**	**3895**
13	4020	3911
14	4145	4598
15	3136	2956
16	223	123
17	997	**1814**
18	*2287*	**3945**
19	2508	4657
20	2736	4165

The top 20 genes from the two datasets are not very similar. Only 40% of the genes are shared in both gene sets. We use bold-number to mark the common genes in both top 20

genes from two datasets. The JSC_{top20} is equal to 0.25. The rankings of some specific genes differ greatly between two datasets. For example, gene 2040 ranks top 4 in Michigan, but ranks very low (top 91) in Harvard. We also mined the HDGGs and discriminating genes (DGs) from both datasets. The JSC_{DG} is equal to 0.184, and JSC_{HDGG} is equal to 0.056 between Michigan and Harvard datasets. The details about the sets of DGs and the sets of HDGGs are listed in Section 6.6, Table 6.6 and Table 6.7.

Classifier transferability: CT between two datasets is not very high. When Harvard dataset is used to build HDGG-based classifier to test the tissue samples in Michigan dataset, the predicting accuracy is 55.2%. When Michigan dataset is used to build HDGG-based classifier and predict Harvard dataset, the accuracy is 83.6%. Therefore, the classifier transferability (CT) between Harvard and Michigan datasets is CT = (55.2% + 83.6%) / 2 = 69.4%.

These inconsistent data mining results indicate that there is high degree of variability in Michigan and/or Harvard datasets. Therefore, we will minimize the variability in next section.

6.4.4. Biological variability minimization (BVM)

Both Michigan and Harvard datasets are treated with C-loocv in order to find and remove the noisy samples. It takes less than 10 cycles before the datasets reach steady state. After C-loocv processing, the new datasets are named as refined (Michigan/Harvard) datasets. In contrast, before C-loocv, the datasets are called original datasets.

Several samples are removed by C-loocv. In Michigan, 5 samples are eliminated. These

samples include 3 tumor samples (stage I lung cancer, patient number 16, 19 and 54) and 2 normal samples (patient number 88 and 90). In Harvard, 13 samples are eliminated. They are 10 tumor samples (patient 15, 17, 21, 76, 83, 117, 121, 132, 135 and 137) and 3 normal samples (patient number 193, 199 and 203). After BVM, there are 91 samples in refined Michigan dataset and 190 samples in refined Harvard dataset.

JSC from refined datasets: Top 20 genes are mined from both refined datasets. The results show in Table 6.4.

Table 6.4: Top 20 genes in refined Harvard and Michigan lung cancer datasets

Rank order	Michigan	Harvard
1	711	640
2	**845**	**845**
3	**1344**	**1344**
4	**1814**	**2607**
5	2040	**2688**
6	2275	**2910**
7	2287	4351
8	**2607**	**2555**
9	**2910**	**3895**
10	3136	**2956**
11	3581	3197
12	**3895**	3284
13	4020	**123**
14	5269	**1814**
15	**123**	**1981**
16	**1981**	1241
17	**2555**	3911
18	**2688**	4704
19	**3945**	2916
20	**2956**	**3945**

The similarity of data mining results from the two refined datasets is improved. 60% of the genes appears in both top 20 genes sets (bold font) and $JSC_{(top20,\ BVM)} = 0.43$. Recall

that before BVM, JSC_{top20} = 0.25. We also compare sets of top 40, top 60 and top 80 genes from both datasets before and after BVM. All of them show that after BVM, the Jaccard similarity coefficient value increases. We test the discriminating gene sets and HDGG sets too. After BVM, the number of common genes in both discriminating gene sets and HDGG sets also increased. $JSC_{(DG,\ BVM)}$ = 0.40 and $JSC_{(JEP,\ BVM)}$ = 0.357 in contrast with JSC_{DG} = 0.184 and JSC_{JEP} = 0.056. Detailed gene lists can be found in Table 6.6 and Table 6.7 in Section 6.6.

Classifier transferability (CT) between refined datasets: The CT between refined datasets improved. Using HDGG-based classifier built from Harvard tests Michigan, the accuracy is 83.7%; meanwhile, using classifier built from Michigan test Harvard, the accuracy is 93.8%. Thus, the CT reaches (83.7% + 93.8%)/2 = 88.8%. Recall that before BVM, the transferability is only 69.4%. CT significantly improved due to our BVM.

The above results indicate that using C-loocv, we can effectively minimize the biological variability and improve the reliability of datasets.

6.5. CLASSIFICATION IMPROVEMENT

BVM with C-loocv has many advantages and significances in data mining. Besides minimizing the degree of variability in microarray datasets, BVM also increases the predicting accuracy during classification process. This was mentioned in Section 6.4.4: classifier transferability improved significantly when using classifiers built from refined datasets. Therefore, C-loocv can be a very good algorithm to improve the robustness of trained classifiers.

Many classifiers such as SVM, PCL, HDGG-based classifiers and so on, can predict unknown samples with very high accuracy. But in some datasets, some specific samples can not be predicted correctly no matter which classification methods/classifiers were applied.

High degree of variability in these datasets is one of the major reasons. In addition, the difference of gene expression range between training and testing datasets is another factor. By using C-loocv to identify and remove noisy samples from training datasets, the situation can be significantly improved. Further, when the baseline is adjusted between training and testing datasets, we obtain even better results. Next, we will show that our C-loocv algorithm gives very good results on several public microarray datasets.

A. Prostate cancer microarray dataset:

Prostate cancer is another very common cancer worldwide. One microarray dataset focusing on prostate cancer was generated several years ago (Singh et al, 2002). This dataset was created using Affymetrix platform. It includes 12,600 gene probes and 136 tissue samples. The samples were further divided into two subsets: 102 samples were used as training dataset and 34 samples as testing dataset. Both datasets contain tumor samples and normal samples. This is a very well designed dataset for testing the robustness of classification methods, i.e. building different classifiers from training dataset and using each of them to predict the samples in testing dataset. The robustness of each classifier can be evaluated by their predicting accuracy.

Until now, no classification methods can predict the samples in Singh's testing dataset with high accuracy. The highest predicting accuracy is around 50% so far (Tan et al

2003b). We also tested our HDGG-based classifier on this dataset. The predicting accuracy is nearly 50%. Many reasons may be responsible for the poor predicting accuracy. Among them, biological variability is one of the major factors.

In order to minimize biological variability, we use C-loocv algorithm to refine the training dataset and eliminate noisy samples. After BVM, totally 20 samples are eliminated. These discarded samples include 15 tumor samples and 5 normal samples. When a classifier is built from the refined training dataset and applied to test the original testing dataset, the predicting accuracy increases to 73.5%. Though this accuracy rate is not very high, it improved greatly compared with the highest previous accuracy rate.

As discussed in Section 5.3.6, each individual gene's expression value in two datasets should be normalized to identical or similar range before these two datasets can be studied together. In Singh's dataset, nearly one-third of the gene's expression values in training and testing dataset were not in the same range. For example, gene 8's expression range is [-11, 94] across 102 samples in training dataset, but it is [414, 2017] across 34 samples in testing dataset.

Per gene range based normalization (Chapter 5) was applied to normalize training and testing dataset, the predicting accuracy reaches 85.3%. If C-loocv algorithm is applied on the top of gene range based normalization, the predicting accuracy increases to 94.1% (only 2 of 34 testing sample were wrongly predicted). (Table 6.5)

B. Breast cancer microarray dataset:

Breast cancer is the most prevalent cancer in women in the US. One microarray dataset to study breast cancer was created by van't Veer and colleagues (van't Veer et al 2002). This

dataset is also divided into training dataset (78 samples) and testing dataset (19 samples). The predicting accuracy was low using current classification methods. When using HDGG-based classifier built from training dataset, 6 samples were wrongly predicted. After gene range based normalization between training and testing datasets, 5 samples were still incorrectly predicted.

Then, C-loocv algorithm is used to refine the training dataset and eliminate noisy samples. Totally 31 samples were eliminated from training dataset. When HDGG-based classifiers are built from the refined training dataset on top of gene range based normalization, the predicting result improved significantly. Only 2 samples in original testing dataset were mis-predicted (Table 6.5).

C. Leukemia dataset:

In some public cancer related microarray datasets, such as Leukemia dataset (Golub 1999), the testing samples have been predicted with very high accuracy using current classifiers. However, we can get higher predicting accuracy by using baseline adjustment and C-loocv algorithm.

Leukemia dataset includes both training dataset (38 samples) and testing dataset (34 samples). With our HDGG-based classifier built from the original training dataset, 4 samples were predicted incorrectly. After baseline adjustment between the training and testing dataset and C-loocv, our HDGG-based classifier built from the refined training dataset achieved a predicting accuracy of 94.1% (only 2 wrongly predicted samples). (Table 6.5)

Table 6.5: predicting accuracy improvement with C-loocv and baseline adjustment

Datasets	Original	Baseline adjustment	C-loocv	C-loocv with Baseline adjustment
Prostate cancer	50%	85.3%	73.5%	94.1%
Breast cancer	68.4%	73.7%	68.4%	89.5%
Leukemia	88.2%	91.2%	91.2%	94.1%

Table 6.5 shows the improvement of prediction accuracy using baseline adjustment or/and C-loocv. The values in the table are the predicting accuracy rate in percentage. HDGG-based classifier was used for predicting testing samples. The second column shows the results with the classifier trained from the original training dataset. The training dataset was treated with baseline adjustment (3^{rd} column) or C-loocv (4^{th} column) respectively before the classifier was built. The fifth column demonstrates the results where both treatments were performed.

6.6. CONCLUSION AND DISCUSSION

High degree of variability in microarray datasets affects data mining results, makes them unreliable, and affects the robustness of classifiers. Because of the existence of noisy samples, the training process may be misled and the established classifiers may not fully embody the intrinsic difference between the two classes within one dataset. In other words, they are not as robust as they should be since the most significant differences between two classes may have been quenched by the noisy samples. Thus, it is hard to get high predicting accuracy using such classifiers.

In microarray data mining, we are the first one to evaluate the effect of variability on data mining results using comparative methods and to provide methods to minimize

variability. Our attempt turned out to be promising and useful.

We developed the C-loocv algorithm to minimize variability in microarray datasets by improving Liu's algorithm. Compared with Liu's algorithm, our method is more general. HDGG-based classifiers have been proven to be very robust (Chapter 4). LOOCV examines the instances one by one and treats every instance equally. Every instance in a given training dataset participates to decide the testing instance's class type. This is better than the k-NN method, in which only k instances are used to predict the testing instance. The choice of value 'k' is important because different k value may cause different predicting result, and it is a difficult issue.

The C-loocv algorithm effectively eliminates noisy samples and reduces variability. After datasets are refined by C-loovc, the predicting accuracy increases significantly, and the consistency of data mining results is also improved.

However, though the consistency of data mining results is improved by BVM, it did not reach our expected level. For example, there is still big divergence of data mining results between Harvard and Michigan datasets as shown in Table 6.3, Table 6.6 and Table 6.7. We also tested another dataset (Singh et al, 2002) and obtained similar results. We mined top 20 genes from original training and testing datasets in this prostate dataset, and only 2 genes were shared in the two top 20 gene sets. After BVM with C-loocv, we got 6 common genes in both sets (recall that the predicting accuracy has been dramatically improved after C-loocv). This large gene difference (14 genes were difference between two sets) indicated that the training dataset and testing dataset are not very consistent with each other, even though biological variability has been minimized.

The fact that the ultimate results did not reach 100% consistency might have been caused

by many reasons. One of them is the tradeoff of biological variability versus bias. Most microarray datasets have large number of genes and relatively small number of samples. These samples may not fully reflect the intrinsic class features, and they may lead to data mining results which are not fully consistent.

During biological variability minimization, if a small value of threshold τ is chosen, fewer samples are eliminated and the data mining results may not improve significantly. On the other hand, if larger value of τ is chosen, more samples are removed and few samples remain, the bias may rise and mislead the results.

Based on our results in this chapter, we realize that the variability and some other reasons make big impact on the HDGGs we mined. In order to get more reliable HDGGs from current highly variable microarray datasets, we are going to mine so called "invariant patterns" from several microarray datasets which focus on a common disease. The methods and results will be shown in Chapter 7.

6.7. APPENDIX

In this section we will show the sets of discriminating genes and the sets of HDGGs mined from Michigan and Harvard lung cancer datasets.

6.7.1 The sets of discriminating genes (DGs)

Discriminating genes have been proved to be of great importance, because they contain valuable information in each specific dataset. The similarity between the sets of discriminating genes from two datasets is another good criterion to test biological variability. If there is high degree of biological variability in datasets, the discriminating genes and the number of discriminating genes from these datasets will become quite

different. Table 6.6 shows the DG sets from two lung cancer datasets before and after biological variability minimization (BVM).

Table 6.6: Discriminating genes in Harvard and Michigan lung cancer datasets

Original datasets		Refined datasets	
Michigan	Harvard	Michigan	Harvard
640	223	**123**	**123**
845	483	**640**	**640**
1210	**640**	**845**	796
1344	**845**	**1344**	**845**
2302	997	1814	961
2555	1086	1981	1035
2607	**1344**	**2040**	1100
2688	1526	2275	**1344**
2910	1690	**2287**	1600
2956	1814	**2555**	1797
3104	2040	**2607**	1814
3207	2275	**2688**	1981
3550	2287	**2910**	**2040**
4351	2508	3136	**2287**
4657	**2607**	**3581**	**2555**
	2688	**3895**	**2607**
	2736	3945	**2688**
	2910	**4020**	**2910**
	2956	4145	2941
	3136	**4165**	2956
	3581		3012
	3784		3038
	3895		3284
	3945		**3581**
	4020		**3895**
	4145		**4020**
	4180		**4165**
	4366		4351
	4598		4657

The contents in Table 6.6 are the discriminating gene index numbers. We use the bold font to show discriminating genes shared by both Harvard and Michigan datasets. Before BVM, there are 16 and 29 discriminating genes in Michigan and Harvard datasets

respectively. Among them only 7 common DGs are in both sets (JSC_{DG} = 18.4%). After refining the datasets by BVM, the numbers of discriminating genes in two datasets become 20 and 29 respectively, and 14 of them are shared by two datasets ($JSC_{(DG, BVM)}$ = 40%). The improvement of common DGs indicates that C-loocv algorithm can effectively reduce biological variability.

6.7.2. The sets of HDGGs (JEPs)

HDGGs have been proved to be very useful for discovering the inherent distinctions between different classes within one dataset (Chapter 4). It is also one important characteristic of a given dataset. The similarity of the sets of HDGGs from two datasets is also a good criterion to test biological variability. Table 6.7 lists the top 10 signature JEPs in diseased class in Michigan and Harvard datasets. The details about the sets of HDGGs in these two datasets will be described in Chapter 7.

Note: In order to show more information, we leave '+' and '-' sign in the table, so the patterns in the table are actually JEPs. When the signs are removed, each JEP will become one HDGG. There are only 9 JEPs in original/refined Harvard dataset.

Table 6.7: HDGGs (JEPs) in Harvard and Michigan lung cancer datasets

Original datasets		Refined datasets	
Michigan	Harvard	Michigan	Harvard
{2910- }	**{2910- }**	{711-}	{640-}
{845- }	{3550- 4351- }	**{845-}**	**{845-}**
{1344- }	{3550- 3207- }	**{1344-}**	**{1344-}**
{1814- }	{2607- 3550- }	{1814-}	**{2607-}**
{2040- }	{2607- 2302- }	{2040-}	**{2287-}**
{3895- }	{2607- 1210+ }	{2275-}	{2688-}
{3136+ 483- }	{3550- 4351- }	**{2287-}**	**{2910-}**
{3136+ 483- }	{4351- 2302- }	**{2607-}**	{2956-}
{3136+ 5126- }	{4351- 1210+ }	**{2910-}**	{4351-}
{3136+ 65- }		{3136+}	

Comparing the HDGGs mined from two lung cancer datasets, we find that only one HDGG is shared by two original datasets (first & second columns) whereas five HDGGs are shared by two C-loocv refined datasets (third & fourth columns).

Chapter 7: DISCOVERY OF INVARIANT PATTERNS

7.1. MOTIVATION AND OUR APPROACH

During our microarray dataset study, we observed that the current microarray datasets we focused on have very high variability. The existence of variability affects the quality and transferability of mined patterns. In our study, among the mined HDGGs from one dataset, it is difficult to tell which HDGGs are intrinsic patterns and which ones are artifacts caused by variability. Most importantly, some intrinsic patterns may have been covered up by the variability and can not be mined from such highly variable datasets. Even though biological variability has been minimized by our C-loocv algorithm, other factors also exist in microarray datasets that influence the mined patterns.

In order to better understand the mechanism underlying diseases, we will mine "invariant patterns" from different datasets concerning a common disease. We define invariant pattern (IVP) as following: if a pattern is a signature HDGG in one dataset and it has very high support in all other datasets (even though it might not be a jumping emerging pattern), we call this pattern an *invariant pattern*. In contrast, other HDGGs are defined as variant patterns (VP).

In this chapter, (1) we mine both invariant patterns and variant patterns from multiple microarray datasets concerning a common disease; (2) we prove that invariant patterns are more related with the disease of interest than variant patterns; (3) we demonstrate again that C-loocv is an effective algorithm for minimizing variability in datasets, and for

helping us to mine more invariant patterns.

We hope that invariant patterns can help shed light on the mechanism of the given disease underlying the given datasets based on the following rationale. If one pattern is involved in the disease-specific gene interaction networks and pathways, this pattern is very likely to be a HDGG, or to have very high "verification" frequency in any microarray datasets concerning this disease. On the other hand, if a pattern has very high frequency in one dataset, but has very low frequency in other datasets concerning the same disease, it is more reasonable to deduce that this pattern might be an artifact caused by noise or technical differences.

7.2. MATERIALS AND METHODS

We use the same lung cancer microarray datasets (namely Michigan and Harvard) as used in Chapter 6. The details about them were described in Chapter 6. After treating the datasets by C-loocv algorithm, we obtain two more datasets: Refined Michigan (RM) and Refined Harvard (RH). For our convenience, the two datasets not treated with C-loocv are called Original Michigan (OM) and Original Harvard (OH) respectively. Thus, totally four datasets are studied in this chapter (OM, RM, OH and RH).

The method we use to mine invariant patterns from any two given microarray datasets is the following: a) in each given microarray dataset, every gene's expression value is discretized as 0 or 1 using the entropy based method; b) the iterative gene club formation algorithm and border differential algorithm are applied to mine HDGGs from each of the dataset; c) the HDGGs mined from the diseased class in each dataset are collected; d) the frequency of each HDGG is determined in both datasets; e) the invariant patterns are

determined based on the HDGGs' frequency in each dataset.

In this chapter, the above method will be applied to mine HDGGs and IVPs from dataset pairs: one dataset from Michigan (original or refined) and the other one from Harvard (original or refined). Specifically, we will mine HDGGs and IVPs from four pairs of datasets (OM vs OH; RM vs OH; OM vs RH; RM vs RH) respectively. The data mining results from dataset pairs (OM vs OH and RM vs RH) are of our great interest and will be studied thoroughly. The results from dataset pairs RM vs OH and OM vs RH will be used as reference.

After finding the invariant patterns and variant patterns, the known biological functions of the discriminating genes within those patterns are investigated and evaluated. The number and percentage of invariant patterns within mined HDGGs from different dataset pairs are also compared and discussed.

7.3. RESULTS ON INVARIANT HDGG PATTERNS

Table 7.1 lists the HDGGs mined from OM and OH datasets. The columns from left to right are: HDGG's index, emerging patterns which include the involved discriminating genes in the HDGG, frequency of the HDGG in the OM dataset, frequency of the HDGG in the OH, and whether the HDGG is considered as an invariant pattern, respectively. We use the frequency of 95% as threshold to determine whether a HDGG is an IVP, i.e. if one HDGG's frequency in both datasets is higher than 95%, we consider it as an IVP. Otherwise, it is a VP.

Table 7.1: IVPs and VPs from OM and OH datasets

Index	Emerging patterns	Frequency in OM (%)	Frequency in OH (%)	IVP
1	{845 -}	100.0%	98.4%	Y
2	{1344 -}	100.0%	97.8%	Y
3	{1814 -}	100.0%	96.8%	Y
4	{2040 -}	100.0%	88.7%	
5	{2910 -}	100.0%	99.5%	Y
6	{3895 -}	100.0%	97.3%	Y
7	{3136 +, 483 -}	100.0%	13.4%	
8	{3136 +, 4126 -}	100.0%	5.4%	
9	{3136 +, 65 -}	100.0%	9.7%	
10	{3136 +, 115 -}	100.0%	1.1%	
11	{3136 +, 725 -}	100.0%	13.4%	
12	{3136 +, 4228 +}	100.0%	13.4%	
13	{3136 +, 3829 +}	100.0%	15.1%	
14	{3136 +, 3925 +}	100.0%	12.4%	
15	{3136 +, 1214 +}	100.0%	8.1%	
16	{3136 +, 1722 +}	100.0%	8.1%	
17	{3136 +, 2126 +}	100.0%	14.5%	
18	{3136 +, 4757 +}	100.0%	12.9%	
19	{483 -, 4228 +, 5126 -}	100.0%	27.4%	
20	{483 -, 4228 +, 115 -}	100.0%	12.4%	
21	{640 -}	97.7%	98.4%	Y
22	{2688 -}	98.8%	98.4%	Y
23	{2607 -, 3550 -}	97.7%	99.5%	Y
24	{2607 -, 2302 -}	74.4%	99.5%	
25	{2607 -, 1210 +}	8.1%	99.5%	
26	{4351 - , 3550 -}	90.7%	99.5%	
27	{4351 - , 2302 -}	67.4%	99.5%	
28	{4351 - , 1210 +}	8.1%	99.5%	
29	{2607 -, 2302 -}	74.4%	99.5%	
30	{2607 -, 1210 +}	8.1%	99.5%	
31	{2302 -, 4351 –}	67.4%	99.5%	

Among the 31 HDGGs mined from the two original datasets, 8 of them are IVPs because their frequencies are higher than 95% in both datasets, and the other 23 HDGGs are VPs which exhibit high frequency in only one dataset.

Table 7.2 lists the IVPs and VPs mined from RM and RH datasets. A total of 26 HDGGs are mined from these two datasets and 14 of them are IVPs.

Table 7.2: IVPs and VPs from RM and RH datasets

Index	Emerging patterns	Frequency in RM (%)	Frequency in RH (%)	IVP
1	{640 -}	100.0%	100.0%	Y
2	{845 -}	100.0%	100.0%	Y
3	{1344 -}	100.0%	100.0%	Y
4	{1814 -}	100.0%	99.4%	Y
5	{2040 -}	100.0%	99.4%	Y
6	{2275 -}	100.0%	40.3%	
7	{2287 -}	100.0%	100.0%	Y
8	{2607 -}	100.0%	100.0%	Y
9	{2910 -}	100.0%	100.0%	Y
10	{3136 +}	100.0%	2.3%	
11	{3581 -}	100.0%	98.9%	Y
12	{3895 -}	100.0%	98.9%	Y
13	{4020 -}	100.0%	99.4%	Y
14	{2688 -}	98.8%	100.0%	Y
15	{2956 -}	97.6%	100.0%	Y
16	{4351 -}	100.0%	100.0%	Y
17	{3284 -, 4657 -}	91.9%	100.0%	
18	{3284 -, 796 +}	90.7%	100.0%	
19	{3284 -, 2941 +}	69.9%	100.0%	
20	{3284 -, 1600 +}	20.5%	100.0%	
21	{3284 -, 3038 +}	45.8%	100.0%	
22	{3284 -, 961 +}	30.1%	100.0%	
23	{3284 -, 1035 +}	66.3%	100.0%	
24	{3284 -, 1100 +}	13.3%	100.0%	
25	{3284 -, 1797 +}	21.7%	100.0%	
26	{3284 -, 3012 +}	69.9%	100.0%	

From these two Tables, we notice that biological variability has big impact on the mined HDGGs. Some HDGGs have very high frequency in one dataset, but have quite low frequency in the other dataset. For example, in Table 7.1, patterns {3136+, 115-}, {3136+, 65- } have 100% frequencies in OM dataset; but their frequencies are less than 10% in

OH dataset. Patterns {4351- , 1210+}, {2607-, 1210+} have nearly 100% frequencies in OH dataset, but have lower than 10% frequencies in OM. These results demonstrate that biological variability greatly affects data mining result and reduce the quality of mined HDGGs. Therefore, further study is needed to sort out real valuable HDGGs, which are IVPs as our concern, for the understanding of diseases.

In the following sections, first we prove that IVPs are more related with cancer diseases than VPs; then we demonstrate again that the C-loocv algorithm can effectively improve the invariance of datasets.

7.3.1. Biological function comparison of the discriminating genes within IVPs versus VPs

As shown in Table 7.2, 14 IVPs and 12 VPs are identified from two refined lung cancer datasets (RM vs RH). In order to evaluate the quality of these patterns, one good and direct way is to investigate the biological functions of each discriminating gene (DG) in these patterns since the set of DGs (total 14 DGs) in IVPs are completely different from that in VPs (total 13 DGs). Each discriminating gene's function will be described in Section 7.4.2.

Note: If there were overlapping of DGs in IVPs with those in VPs, we should have further investigated the relationship between these DGs' involved in one given HDGG.

According to the gene's functions by biological studies, we find that among the 14 DGs within IVPs, 7 DGs have been shown to be related with tumors. The ratio is 7/14 = 50.0%. In contrast, among the 13 DGs within VPs, only 3 genes have been proven to be related with tumors, and the ratio is 3/13 = 23.1%.

In our cancer study, theoretically, if no variability exists in microarray datasets, all DGs

in both IVPs and VPs should be related with tumors. Therefore, the above results indicate that, using datasets with high variability, IVPs are more robust than VPs to understand diseases. They may provide useful clues for finding some unknown gene's functions and for discovering the potential gene interactions in tumor occurrence.

We also investigated the functions of the discriminating genes listed in Table 7.1, which were mined from the original datasets (OM vs OH). We randomly chose several discriminating genes for function analysis and got similar results.

7.3.2. C-loocv effectively improves the quality of mined HDGGs

We compare the IVPs and VPs mined from different dataset pair combinations. These HDGGs are listed in Table 7.1, 7.2, 7.4 and 7.5.

(1) The numbers of IVPs mined from the refined dataset pairs are larger than that from the original ones. Meanwhile, the percentage of IVPs in HDGGs also increases (Table 7.3). Table 7.3 lists the number and the percentage of IVPs in HDGGs mined from 4 different dataset pairs.

Table 7.3: The number and percentage of IVPs mined from four dataset pairs

Index	Datasets	No. of IVPs	No. of HDGGs	% of IVPs
1	OM vs OH	8	31	25.8%
2	OM vs RH	11	36	30.6%
3	RM vs OH	11	22	50.0%
4	RM vs RH	14	26	53.8%

(2) Comparing the set of IVPs from different dataset pair, we notice that one IVP {2607, 3550} is mined from the original datasets; in contrast, a different IVP {2607} is mined from the refined datasets, which is the subset of the former pattern. This means that one discriminating gene 2607 may be enough to determine the disease. The occurrence of

gene 3550 in pattern {2607, 3550} may be an artifact.

(3) Some IVPs fail to be mined from the original datasets, but they can be mined from the refined datasets. By comparing the sets of IVPs mined from different dataset pairs in Table 7.3, we find the relationship among the IVP sets: A $\subset$ B $\subset$ D, and A $\subset$ C $\subset$ D (Figure 7.1). Here, A is the set of IVPs mined from dataset pair OM vs OH; B is IVP set from OM vs RH; C is the IVP set from RM vs OH, and D is the IVP set from RM vs RH dataset pair. Note: We consider the IVP {2607, 3550} and IVP {2607} as the same pattern.

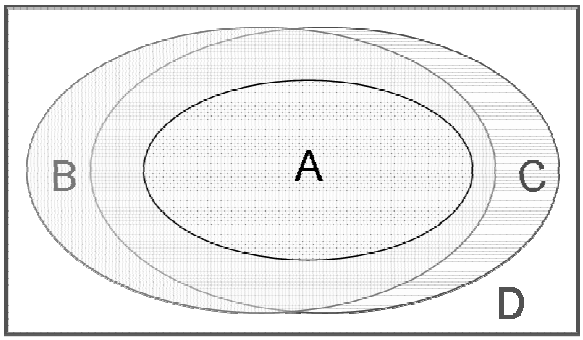

Figure 7.1: Relationship of the sets of IVPs from different dataset pairs

Several IVPs such as {2040}, {2956}, {4351}, which failed to be mined from the original datasets, have very high frequency in both refined datasets. The discriminating genes in these IVPs are proven to be related to tumors.

From the above results, we conclude that: (A) C-loocv effectively minimizes the variability of microarray datasets; (B) C-loocv improves the overall quality of the HDGGs with respect to IVPs mined from the refined datasets.

7.3.3. Discussion

In order to better understand diseases, we identified invariant patterns from multiple datasets concerning a common disease. In this chapter, we obtained both invariant

patterns and variant patterns from the datasets concerning lung cancer. Compared with variant patterns, there is a higher proportion of discriminating genes in invariant patterns which are known to be related to tumors. This suggests that invariant patterns are more valuable for revealing the mechanism of specific diseases.

Our C-loocv algorithm, which was developed to minimize the biological variability of datasets, can effectively help us to mine high-quality IVPs from microarray datasets. Indeed, the quality of HDGGs mined from the C-loocv refined datasets is higher than that from the original datasets.

95% of frequency was used as threshold to determine invariant patterns in this study. We also tried 90% and 85% of frequency as thresholds and got similar results: (A) the invariant patterns were more related with tumors than variant patterns; (B) the number and the proportion of invariant patterns in mined HDGGs were higher from the refined datasets than that from original datasets.

With the development of modern molecular biology, more and more genes have been identified and sequenced. However, the specific functions of many genes are still unknown or still under investigation. Functional genomics is a relatively new field of molecular biology that attempts to use numerous known gene-sequence data to determine unknown gene functions and interactions. It usually takes a long time to understand one specific gene's function. We hope that our data mining results can provide valuable clues for gene function study.

According to our results, the genes in invariant patterns tend to be potentially important for the occurrence of lung cancer. So far, more than 40% of the discriminating genes within these IVPs have not been found to be functionally related with tumors. We

recommend further study by biomedical scientists to determine the exact functions of these genes.

7.4. APPENDIX

7.4.1. IVPs and VPs in OM vs RH and in RM vs OH datasets

Table 7.4 lists the IVPs and VPs mined from dataset OM vs RH. A total of 36 HDGGs are mined from these two datasets and 11 of them are IVPs.

Table 7.4: IVPs and VPs from OM and RH dataset pair

Index	Emerging patterns	Frequency in OM (%)	Frequency in RH (%)	IVP
1	{845 -}	100.0%	100.0%	Y
2	{1344 -}	100.0%	100.0%	Y
3	{1814 -}	100.0%	99.4%	Y
4	{2040 -}	100.0%	99.4%	Y
5	{2910 -}	100.0%	100.0%	Y
6	{3895 -}	100.0%	98.9%	Y
7	{3136 +, 483 -}	100.0%	2.3%	
8	{3136 +, 5126 -}	100.0%	0.0%	
9	{3136 +, 65 -}	100.0%	1.1%	
10	{3136 +, 115 -}	100.0%	2.3%	
11	{3136 +, 725 -}	100.0%	2.3%	
12	{3136 +, 4228 +}	100.0%	2.3%	
13	{3136 +, 3829 +}	100.0%	2.3%	
14	{3136 +, 3935 +}	100.0%	2.3%	
15	{3136 +, 1214 +}	100.0%	0.0%	
16	{3136 +, 1722 +}	100.0%	1.1%	
17	{3136 +, 2126 +}	100.0%	2.3%	
18	{3136 +, 4757 +}	100.0%	2.3%	
19	{483 -, 4228 +, 5126 -}	100.0%	9.7%	
20	{483 -, 4228 +, 115 -}	100.0%	73.9%	
21	{640 -}	97.7%	100.0%	Y
22	{2287 -}	96.5%	100.0%	Y
23	{2607 -}	98.8%	100.0%	Y
24	{2688 -}	98.8%	100.0%	Y
25	{2956 -}	96.5%	100.0%	Y
26	{4351 -}	91.9%	100.0%	
27	{3284 -, 4657 -}	91.9%	100.0%	

28	{3284 -, 796 +}	90.7%	100.0%	
29	{3284 -, 2941 +}	70.9%	100.0%	
30	{3284 -, 1600+}	20.9%	100.0%	
31	{3284 -, 3038 +}	48.8%	100.0%	
32	{3284 -, 961 +}	66.3%	100.0%	
33	{3284 -, 1035 +}	23.3%	100.0%	
34	{3284 -, 1100 +}	34.9%	100.0%	
35	{3284 -, 1797 +}	22.1%	100.0%	
36	{3284 -, 3012 +}	72.1%	100.0%	

Table 7.5 lists the IVPs and VPs mined from OH vs RM datasets. A total of 22 HDGGs are mined from these two datasets and 11 of them are IVPs.

Table 7.5: IVPs and VPs from RM and OH datasets

Index	Emerging patterns	Frequency in RM (%)	Frequency in OH (%)	IVP
1	{845 -}	100.0%	98.4%	Y
2	{1344 -}	100.0%	97.8%	Y
3	{1814 -}	100.0%	96.8%	Y
4	{2040 -}	100.0%	88.7%	
5	{2275 -}	100.0%	58.1%	
6	{2287 -}	100.0%	96.2%	Y
7	{3136 +}	100.0%	15.6%	
8	{3581 -}	100.0%	80.1%	
9	{3895 - }	100.0%	97.3%	Y
10	{4020 -}	100.0%	95.2%	Y
11	{2910 -}	100.0%	99.5%	Y
12	{640 -}	100.0%	98.4%	Y
13	{2688 -}	98.8%	98.4%	Y
14	{2607 -, 3550 -}	100.0%	99.5%	Y
15	{2607 -, 2302 -}	68.7%	99.5%	
16	{2607 -, 1210 +}	8.4%	99.5%	
17	{4351 -, 3550 -}	100.0%	99.5%	Y
18	{4351 -, 2302 -}	68.7%	99.5%	
19	{4351 -, 1210 +}	8.4%	99.5%	
20	{2607 -, 2302 -}	68.7%	99.5%	
21	{2607 -, 1210 +}	8.4%	99.5%	
22	{2302 -, 4351 -}	68.7%	99.5%	

7.4.2. Biological functions of the related genes

Table 7.6 lists the genes that are involved in the IVPs and HDGGs mined from both the original datasets and the refined datasets. In this section, we introduce the genes listed in Table 7.2 and the known functions of these genes. Meanwhile, we also randomly choose several genes from Table7.6 and discuss their functions.

Table 7.6: Description of DGs involved in HDGGs in Tables 7.1 and 7.2

Index number	Gene number	Uni gene name	Description
1	65	GTF2B	general transcription factor IIB
2	115	RPS6KA1	ribosomal protein S6 kinase, 90kDa, polypeptide 1
3	483	POLR2C	polymerase (RNA) II (DNA directed) polypeptide C, 33kDa
4	640	PTPRH	protein tyrosine phosphatase, receptor type, H
5	725	KDR	kinase insert domain receptor (a type III receptor tyrosine kinase)
6	796	CTNNA2	catenin (cadherin-associated protein), alpha 2
7	845	FRAP1	FK506 binding protein 12-rapamycin associated protein 1
8	961	RPL18	ribosomal protein L18
9	1035	ANXA3	annexin A3
10	1100	DRP2	dystrophin related protein 2
11	1210	VAMP2	vesicle-associated membrane protein 2 (synaptobrevin 2)
12	1214	ZNF345	zinc finger protein 345
13	1344	PPP3CC	protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform (calcineurin A gamma)
14	1600	FXN	Frataxin
15	1722	MAGEA2	melanoma antigen family A, 2
16	1797	DDT	D-dopachrome tautomerase
17	1814	PTPRU	protein tyrosine phosphatase, receptor type, U
18	2040	SLC2A5	solute carrier family 2 (facilitated glucose/fructose transporter), member 5
19	2126	MAGEA5	melanoma antigen family A, 5
20	2275	3.8-1	MHC class I mRNA fragment 3.8-1
21	2287	SGCD	sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)
22	2302	OPRK1	opioid receptor, kappa 1
23	2607	ZNF268	zinc finger protein 268

24	2688	NCOA1	nuclear receptor coactivator 1
25	2910	RNF113A	ring finger protein 113A
26	2941	MDH1	malate dehydrogenase 1, NAD (soluble)
27	2956	HSD17B4	hydroxysteroid (17-beta) dehydrogenase 4
28	3012	CAMP	cathelicidin antimicrobial peptide
29	3038	NEUROD1	neurogenic differentiation 1
30	3136	ERP29	endoplasmic reticulum protein 29
31	3284	CYP11A1	cytochrome P450, family 11, subfamily A, polypeptide 1
32	3550	BLMH	bleomycin hydrolase
33	3581	GLB1	galactosidase, beta 1
34	3829	PENK	Proenkephalin
35	3895	FASN	fatty acid synthase
36	3925	TARS	threonyl-tRNA synthetase
37	4020	ATP6V0D1	ATPase, H+ transporting, lysosomal 38kDa, V0 subunit d1
38	4126	AGC1	aggrecan 1 (chondroitin sulfate proteoglycan 1, large aggregating proteoglycan, antigen identified by monoclonal antibody A0122)
39	4228	CD3G	CD3g molecule, gamma (CD3-TCR complex)
40	4351	PRM1	protamine 1
41	4657	CCT5	chaperonin containing TCP1, subunit 5 (epsilon)
42	4757	KCNB1	potassium voltage-gated channel, Shab-related subfamily, member 1
43	5126	HTATIP	HIV-1 Tat interacting protein, 60kDa

Gene 65 (general transcription factor IIB): general transcription factor IIB is a ubiquitous factor required for transcription initiation by RNA polymerase II. It was suggested that TFIIB serves as a bridge between the "TATA"-binding factor (TFIID) and RNA polymerase II during pre-initiation complex assembly. Recently, it was also found that GTFIIB can be a target of acidic activators.

Gene 483 (polymerase (RNA) II (DNA directed) polypeptide C, 33kDa): This gene encodes the third largest subunit of RNA polymerase II, the polymerase responsible for synthesizing messenger RNA in eukaryotes. The product of this gene contains a cysteine rich region and exists as a heterodimer with another polymerase subunit, POLR2J. These two subunits form a core subassembly unit of the polymerase. The expression of this

gene is regulated during muscle differentiation (Corbi et al 2005).

Gene 640 (PTPRH) and **gene 1814** (PTPRU): The proteins encoded by these two genes are a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. The genes were shown to be expressed primarily in brain and liver, and at a lower level in heart and stomach. It was also found that these two genes expressed in several cancer cell lines, but not in the corresponding normal tissues (Trapasso et al 2004).

Gene 796 (catenin (cadherin-associated protein), alpha 2): The protein encodes by this gene is a subunit of alpha N-catenin, a linker between cadherin adhesion receptors and the actin cytoskeleton. It is essential for stabilizing dendritic spines in rodent hippocampal neurons in culture. It has been proven that alpha N-catenin is a key regulator for the stability of synaptic contacts (Abe et al 2004).

Gene 845 (FK506 binding protein 12-rapamycin associated protein 1 or Frap1): The protein encoded by this gene belongs to a family of phosphatidylinositol kinase-related kinases. The known function for this protein is kinase activity and binding. It has been reported that Frap is a candidate gene for the plasmacytoma resistance locus Pctr2 and can act as a tumor suppressor gene (Bliskovsky et al 2003).

Gene 961 (ribosomal protein L18): This gene encodes the large subunit of ribosomal protein. This is one of the proteins that binds and probably mediates the attachment of the 5S RNA into the large ribosomal subunit, where it forms part of the central protuberance.

Gene 1035 (annexin A3): ANXA3 is present in healthy epithelial cells, and is relatively less abundant in individual tumor cells of increasing Gleason pattern (GP), despite

exhibiting higher overall tissue abundance in tumors. ANXA3 staining was predominantly cytoplasmic. Strongly staining single cells, possibly phagocytes, were interspersed in highly dedifferentiated GP5 tumor areas among tumor cells without measurable ANXA3. (Wozny et al 2007).

Gene 1100 (dystrophin related protein 2): DRP protein is the members of the dystrophin family, which performs a critical role in the maintenance of membrane-associated complexes at points of intercellular contact in vertebrate cells. Dystrophin related protein 2 is predicted to resemble certain short C-terminal isoforms of dystrophin and dystrophin-related protein 1 (DRP1 or utrophin). DRP2 is expressed principally in the brain and spinal cord.

Gene 1210(vesicle-associated membrane protein 2 (synaptobrevin 2)): Vesicle-associated membrane protein (VAMP) (or synaptobrevin), a type II membrane protein of small synaptic vesicles, is essential for neuroexocytosis because its proteolysis by tetanus and botulinum neurotoxins types B, D, F and G blocks neurotransmitter release. It implies the existence of a synaptophysin-VAMP-2 complex is helpful for the processes of vesicle docking and fusion with the presynaptic membrane (Washbourne et al 1995).

Gene 1344 (protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform (calcineurin A gamma)): The putative function of this gene includes: Calcium-dependent, calmodulin-stimulated protein phosphatase. This subunit may have a role in the calmodulin activation of calcineurin.

Gene 1600 (frataxin): Frataxin is a small protein, localized to the mitochondrion. The function of frataxin is not entirely clear, but it seems to be involved in assembly of irosulfur clusters. Deficiency of frataxin is the cause of Friedrich’s ataxia, a hereditary

trinucleotide repeat disorder.

Gene 1797 (D-dopachrome tautomerase): This gene's expression is tightly related with Macrophage migration inhibitory factor (MIF) activity. When UVB light was used to induce an experimental inflammation in normal human skin, the D-dopachrome tautomerase's expression increases significantly accomplishing with skin's inflammation (Sonesson et al 2003).

Gene 2040 (solute carrier family 2 (facilitated glucose/fructose transporter), member 5): Another name of this gene is GLUT5, which is expressed on the brush border membrane of human small intestinal enterocytes (Davidson et al 1992). GLUT5 is a fructose transporter and may be largely responsible for the uptake of fructose from the lumen of the small intestine (Burant et al 1992). Godoy and colleagues (Godoy et al 2006) used situ RT-PCR and ultrastructural immunohistochemistry confirmed GLUT5 over-expression in breast cancer. The extensive expression of GLUT2 and 5 (glucose/fructose and fructose transporters, respectively) in malignant human tissues indicates that fructose may be a good energy substrate in tumor cells.

Gene 2275 (MHC class I mRNA fragment 3.8-1): specific function is under investigation.

Gene 2287 (sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)): The protein encoded by this gene is one of the four known components of the sarcoglycan complex, which is a subcomplex of the dystrophin-glycoprotein complex (DGC). DGC forms a link between the F-actin cytoskeleton and the extracellular matrix. This protein is expressed most abundantly in skeletal and cardiac muscle. The mutations in this gene have been associated with autosomal recessive limb-girdle muscular dystrophy and dilated cardiomyopathy. Alternatively spliced transcript variants encoding distinct

isoforms have been observed.

Gene 2607 (zinc finger protein 268): ZNF268 plays a role in the development of human fetal liver and the differentiation of blood cells. There are many splicing isoforms of ZNF 268 genes. ZNF268c mRNA was detected only in tumor cells. ZNF268a, ZNF268b1 and ZNF268b2, were also detected in tumor cell lines (Shao et al 2006).

Gene 2688 (nuclear receptor coactivator 1): The nuclear receptor coactivator 1(NCOA1) is a transcriptional co-reluatory protein which is recruited to DNA promotion sites by ligand activated nuclear receptors. NCOA1 accumulates histone which makes downstream DNA more accessible to transcription. NCOA1 is also frequently called steroid receptor coactivator-1(SRC-1). It has been reported that enhanced androgen receptor activity through elevated expression of SRC-1 in the development of more aggressive disease in men with prostate cancer (Agoulnik et al 2005).

Gene 2910 (RNF 113A): RNF 113A is also called RNF113 or ZNF183 which encodes a ring finger protein 113A. It is a novel gene whose function cannot directly be inferred from its sequence analysis. RNF113A is a ubiquitously expressed protein that contains a RING type zinc finger and a C3H1 type zinc finger.

Gene 2941 (malate dehydrogenase 1, NAD (soluble)): Malate dehydrogenase catalyzes the reversible oxidation of malate to oxaloacetate, utilizing the NAD/NADH cofactor system in the citric acid cycle. The protein encoded by this gene is localized to the cytoplasm and may play pivotal roles in the malate-aspartate shuttle that operates in the metabolic coordination between cytosol and mitochondria.

Gene 2956 (hydroxysteroid (17-beta) dehydrogenase 4): The peroxisomal 17β-hydroxysteroid dehydrogenase type 4 (17β-HSD 4, gene name HSD17B4) catalyzes the

oxidation of estradiol with high preference over the reduction of estrone. The expression of 17β-HSD 4 has been detected in several human cancer cell lines (Launoit et al 1999).

Gene 3012 (cathelicidin antimicrobial peptide): The cathelicidin antimicrobial peptide (CAMP) is an important innate defense peptide. It showed the expression of CAMP in nasal mucosa supporting its role in innate defenses against inhaled pathogens (Ooi et al 2007).

Gene 3038 (neurogenic differentiation 1): This gene encodes a member of the NeuroD family of basic helix-loop-helix (bHLH) transcription factors. The protein forms heterodimers with other bHLH proteins and activates transcription of genes that contain a specific DNA sequence known as the E-box. It regulates expression of the insulin gene, and mutations in this gene result in type II diabetes mellitus.

Gene 3136(ERp29): ERp29 is a recently discovered ER resident that has been implicated in secretory protein synthesis and appears to be of similar prevalence to the established major reticuloplasmins (Hubbard et al. 2000). The novel protein sequence of ERp29 exhibits characteristic features of a reticuloplasmin (signal peptide, ER retention motif), and localization to the ER lumen was comprehensively supported at the biochemical level (Demmer et al. 1997). Hubbard found that cancer cells have more ERp29 than normal cells, and suggested that if it does help make key cellular components, perhaps it could be targeted at preventing cancer growth (Shnyder S., Hubbard M., 2002).

Gene 3284 (cytochrome P450, family 11, subfamily A, polypeptide 1): This gene encodes a member of the cytochrome P450 superfamily of enzymes. The encoded enzyme catalyzes many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids which includes the biosynthesis of sex-steroid

hormones. Recently, the relationship between common variation in CYP11A and breast cancer risk among African-Americans, Latinas, Japanese-Americans, native Hawaiians have been reported (Setiawan et al 2006).

Gene 3550 (bleomycin hydrolase): The normal physiological role of BLM hydrolase is unknown, but it catalyzes the inactivation of the antitumor drug BLM (a glycopeptide) by hydrolyzing the carboxamide bond of its B-aminoalaninamide moiety thus protecting normal and malignant cells from BLM toxicity.

Gene 3581 (galactosidase, beta 1): galactosidase, beta 1 encodes a protein called beta-galactosidase. A deficiency of (GLB1) causes G(M1)-gangliosidosis which is a lysosomal storage disorder (Caciotti et al 2005).The GLB1 gene gives rise to the GLB1 lysosomal enzyme and to the elastin binding protein (EBP), involved in elastic fiber deposition.

Gene 3895 (Fatty acid synthase): Fatty acid synthase (FAS) is a multienzyme protein required for the conversion of acetyl coenzyme A and malonyl coenzyme A to palpitate. High levels of FAS expression have been found in many human cancers, including breast, prostate and colon (Notarnicola et al 2006).

Gene 4020 (ATPase, H+ transporting, lysosomal 38kDa, V0 subunit d1): This gene encodes a protein or proteins that contain an ATPase, V0/V1 complex. It is implied that this gene is in the proton-transporting two sector ATPase complex, which is involved in ATP synthesis coupled proton transport. The function of this enzyme is a hydrogen ion transporting ATPase activity, rotational mechanism and a hydrogen ion transporting ATP synthase activity, rotational mechanism protein.

Gene 4351(protamine 1): Gene 4351 encodes a protein called Protamines. Protamines is

a kind of sperm nuclear protein, which is directly related with male infertility (Iguchi et al, 2006).

Gene 4657 (chaperonin containing TCP1, subunit 5 (epsilon)): This gene encodes a molecular chaperone that is member of the chaperonin containing TCP1 complex (CCT), also known as the TCP1 ring complex (TRiC). The complex folds various proteins, including actin and tubulin. Alternate transcriptional splice variants of this gene have been observed but have not been thoroughly characterized.

Chapter 8: CONCLUSION AND DISCUSSION

Microarrays enable high-throughput parallel gene expression analysis, and their use has grown exponentially during the past decade. We are now in a position where suitable data mining results using the public microarray datasets can be used to identify hypothesis about various biological mechanisms. Comparative microarray data mining could better distinguish phenotypes, propose new hypothesis, identify differentially expressed genes, and discover fundamental patterns of gene expression and regulation.

In this chapter, we will give a brief summary of this dissertation about comparative study of microarray data mining, and highlight some major contributions. Meanwhile, some future research directions are also described based on the results of this dissertation.

8.1. SUMMARY

In this dissertation, we used comparative data mining methods to study certain public microarray datasets. Our goal is to mine intrinsic patterns from cancer related public microarray datasets and to provide valuable clues for biologists to further study cancer diseases. In order to reach this goal, we provided novel methods for testing the concordance of microarray datasets generated from multi-platforms and multi-laboratories, investigating the effect of biological variability on our data mining results, and finally, mining invariant patterns from multi-microarray datasets. We believe that

such patterns could provide valuable information for future cancer study.

Below is a brief summary of this dissertation in a Chapter-by-Chapter manner.

In Chapter 1, we gave the motivations and set the research goal for our study. We also highlighted the outline of this dissertation and briefly introduced the results obtained in this dissertation.

In Chapter 2, we presented some preliminaries on the techniques and terminologies that were used throughout this dissertation. This chapter introduced the high dimensional microarrays, the procedure of microarray data generation, the characteristics of microarray datasets, emerging patterns, border differential algorithms, entropy based discretized method, information gain and so on.

In Chapter 3, we surveyed the existing works related to the topics studied in this dissertation. We discussed the researches reported by previous papers, and most importantly, we identified the gaps between previous studies and our current research goal. What we did in our research was to fill those gaps. We focused on the following topics: comparative microarray gene expression data mining, feature selection, microarray data concordance detection, instance selection, and classification etc.

In Chapter 4, we combined a new feature selection approach with a previous data mining algorithm to discover emerging patterns, which are named highly differentiative gene groups (HDGGs). The HDGGs mined from one dataset are considered as discriminative characteristic patterns and HDGGs are important features for each specific dataset.

Since there are more than thousands of dimensions in microarray data, it is a big challenge to mine HDGGs. In Chapter 4, we introduced novel methods that did a better job to overcome the high dimensionality challenge. To be specific, first, we provided new

ideas to create a relatively small gene group called gene club. Within one gene club, all genes are potentially interactive with each other. Next, we applied border differential algorithm to mine HDGGs from the original data projected on each given gene club. Some genes in mined HDGGs have been confirmed to be related with cancer diseases. Finally, HDGGs have also been used to build classifiers, which are named HDGG-based classifiers.

In Chapter 5, we provided novel measures and techniques to test the concordance of microarray gene expression data. Microarray data were collected using different microarray platforms (provided by different companies) under different conditions in different laboratories. It is necessary to test if these microarray data are comparable, reliable and consistent before they can be applied for clinical, pharmaceutical research and other purposes.

In previous studies, the cross platform and cross lab concordance of microarray data have been tested with many methods, such as biological experiments, statistical methods and so on. But the platform/lab concordance has not been examined with comparative method yet. It has been realized that different testing methods may lead to different results. Therefore, in Chapter 5, we developed several novel comparative methods to examine the concordance of datasets from cross platforms/labs.

In Chapter 6, we defined the degree of variability in microarray datasets, developed measurements for testing the variability, and investigated the effect of variability on our data mining results. To be specific, we studied the effect of two types of variability, measurement variability and biological variation, in microarray datasets.

We also provided novel method (C-loocv) to minimize the biological variability. After

biological variability was minimized, the data mining results from refined datasets were evaluated, and showed a good improvement of the reliability. More importantly, the HDGG-based classifiers trained from refined datasets became more robust, and predicted test samples with much higher accuracy.

In Chapter 7, we discovered certain invariant patterns from multiple microarray datasets concerning a common disease. We studied two microarray datasets derived from the samples of patients with common disease. These datasets were generated from a common platform but different laboratories. We mined HDGGs from each dataset and discovered the shared gene interactions (which are called invariant patterns) by comparing these HDGGs' generality and specificity among different datasets. Since variability affects the data mining results, the invariant patterns should be more reliable to provide useful information for understanding the potential gene pathways for diseases.

In the above 7 chapters we presented our approaches for comparative microarray data mining. Experimental evaluations have been conducted for those proposed approaches, and the results showed that our approaches are very promising and effective. However, limitations were also observed in some of the approaches, which have been suggested as potential future works.

8.2. CONTRIBUTIONS

Overall, we made the following contributions in this dissertation:

(1) We conducted extensive study to mine HDGGs from high dimensional microarray datasets. Our methods are better than previous studies in many ways including: (A) we improved the strength of discovered patterns compare with previous studies; (B) we

discovered strongest HDGGs (100% frequency); (C) our HDGGs were proven to be biological meaningful; (D) the discovered HDGGs were used to build the so-called HDGG-based classifiers, which showed higher predicting accuracy on many public microarray datasets compared with other classifiers.

(2) Using comparative methods, we quantitatively tested the concordance of microarray datasets collected from same/different platforms and different laboratories. This was the first attempt to use comparative methods for evaluating microarray dataset concordance. We tested four popular, commercial platforms: Applied Biosystem (ABI), Affymetrix (AFX), Agilent one color array (AG1) and GE Healthcare (GEH). Our results showed that the datasets from any common platform but different laboratories were highly concordant; the datasets from different platforms were also concordant with each other except the datasets generated from the Agilent one color array platform.

(3) This dissertation was the first attempt to define the degree of variability, measure the effect of variability on data mining results, and minimize variability in microarray datasets. After variability was minimized by C-loocv algorithm, the data mining results from datasets became more consistent. Furthermore, the robustness of our HDGG-based classifiers built from C-loocv refined datasets was significantly improved.

(4) We provided novel method to mine high quality patterns from highly variable microarray datasets. The so-called “invariant patterns” have been proven to be more reliable for helping understand diseases, and they tend to be potentially important for the occurrence of diseases.

We believe that those contributions not only are useful for DNA microarray dataset studies, but also provide valuable information for pharmaceutical and clinical research.

8.3. FUTURE WORKS

While the work in this dissertation has addressed many problems of current microarray data mining, it could only do so with a limited depth. We believe that this dissertation has laid the foundation for a wide variety of potential research and applications. There are several relevant research topics that remain open:

A. Classification method improvement

In microarray data study, good classifiers are very important for accurately predicting the samples (patients) and diagnosing the diseases. Many classification methods have been developed. Currently, no classifier can predict the test samples with 100% accuracy in any datasets. Therefore, newer classifiers with higher predicting ability still need to be developed. HDGG-based classifiers building from HDGGs have been proven to be very robust. Invariant patterns can be used to improve the reliability of HDGG based classifiers. IVP-based classifiers building from IVPs are expecting to be more reliable.

B. Studying more microarray datasets which focus on any common disease

In this dissertation, we provided generic methods to effectively identify valuable patterns, aiming to shed light on the intrinsic mechanism underlying diseases of interest. For our study, we used lung cancer related microarray datasets to mine invariant patterns. For future research, our methods may be employed for microarray datasets concerning other tumors or diseases. When IVPs are mined from multi-microarray datasets which study a common disease, they will be of great help for understanding the mechanism of any

given diseases.

C. Comparative study on multiple diseases

It is also desirable to study microarray datasets which study multiple diseases. By comparing IVPs' generality and specificity among the datasets which study different diseases, the shared and unique gene interactions may be discovered and the shared and unique gene networks may be established. The information for the potential gene pathways for the set of diseases may also be provided.

BIBLIOGRAPHY

Abe K., Chisaka O., et al (2004). Stability of dendritic spines and synaptic contacts is controlled by alpha N-catenin. Nature Neuroscience 7(4):357-63.

Agoulnik I., Vaid A., et al (2005). Role of SRC-1 in the promotion of prostate cancer cell growth and tumor progression. Cancer research 65:7959-7967.

Alon U., Barkai N., et al (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc National Academy of Science 12:6745-50.

Beer D., et al (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine 9 (816).

Bhattacharjee A., et al (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. PNAS. 98 (24), 13790-13795.

Bliskovsky V., Ramsay E., et al (2003). Frap, FKBP12 rapamycin-associated protein, is a candidate gene for the plasmacytoma resistance locus Pctr2 and can act as a tumor suppressor gene. Proc. Natl. Acad. Sci. USA 100:14982-7.

Boffelli D., Mcauliffe J., et al (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391–1394.

Boser B., Guyon I., et al (1992). A training algorithm for optimal margin classifiers. In proceedings of the 5th Annual ACM Workshop on Computation Learning Theory ACM Press. Pittsburgh, PA 144-152.

Breimann L., (1996). Bagging predictors. Machine Learning 24:123-140.

Breiman L., Friedman J., et al (1984). Classification and regression trees. Wadsworth International Group, Belmont CA.

Brown M., Grundy W., et al (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA, 97, 262–267.

Burant C., Takeda J., et al (1992). Fructose transporter in human spermatozoa and small intestine is GLUT5. Journal of Biological Chemistry 267:14523-14526.

Caciotti A., Donati M., et al (2005). Role of beta-galactosidase and elastin binding protein in lysosomal and nonlysosomal complexes of patients with GM1-gangliosidosis. Human mutation 25(3):285-92.

Califano A., et al (2000). Analysis of gene expression microarrays for phenotype classification. Proceedings of ISMB.

Chang Y., Hsieh S., et al (2002). Lymphotoxin beta receptor induces interleukin 8 gene expression via NF-kappaB and AP-1 activation. Exp Cell Res. 278(2):166-74.

Chen H., Tzeng C., (2006). Applications of microarray in reproductive medicine. Chang Gung Med J. 29(1):15-24.

Cicatiello L., Scafoglio C., et al (2003). Analysis of the estrogen-responsive transcriptome from breast cancer cells: a comparative analysis with three different microarray platforms.

Corbi N., Bruno T., et al (2005). RNA Polymerase II subunit 3 is retained in the cytoplasm by its interaction with HCR, the psoriasis vulgaris candidate gene product. Journal of cell science. 118:4253-4260.

Cover T., Hart P., (1967). Nearest neighbor pattern classification. Institute of electrical and electronics engineers transactions on information theory, IT-13:21-27.

Cristianini N., Shawe-Taylor J., (2000). An introduction to support vector machines. Cambridge Unversity Press, Cambridge, www.support-vector.net

Cunliffe H., et al (2003). The Gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. Cancer research, 63:7158-7166.

Davidson N., Hausman A., et al (1992). Human intestinal glucose transporter expression and localization of GLUT5. American Journal Physiology 262: C795-C800.

Demmer J., Zhou C., (1997). Molecular cloning of ERp29, a novel and widely expressed resident of the endoplasmic reticulum. FEBS Lett 402:145–150.

Dong G., Li J., (1999). Efficient mining of emerging patterns: discovering trends and differences. Proc. of ACM SIGKDD international conference on knowledge discovery and data mining. 43–52.

Dong G., Li J., (2005). Mining border descriptions of emerging patterns from dataset pairs. Knowledge and Information Systems. 8: 178 – 202.

Dong G., Zhang X., et al (1999). Classification by aggregating emerging patterns. Proc. 2nd Int'l Conf. on Discovery Science (DS'99), Tokyo, Japan, 30-42.

Dougherty J., Kohavi R., et al (1995). Supervised and unsupervised discretization of continuous features. Proc. of International Conference on Machine Learning. 94–202.

Dubchak I., Brudno M., et al (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. Genome research 10:1304–1306.

Freund Y., Shapire R., et al (1996). Experiments with a new boosting algorithm. Machine

learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Bari, Italy 148-156.

Godoy A., Ulloa V., et al (2006). Differential subcellular distribution of glucose transporters GLUT1-6 and GLUT9 in human cancer: Ultrastructural localization of GLUT1 and GLUT5 in breast tumor tissues. Journal of cellular physiology 207:614-7.

Goldberg D., (1989). Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley.

Golub T., Slonim D., et al (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

Gottgens B., Barton L., et al (2002). Transcriptional regulation of the stem cell leukemia gene (SCL) – comparative analysis of five vertebrate SCL loci. Genome research 12:749–759.

Guo L., Lobenhofer E., et al (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nature Biotechnology 24:1162-1169.

Hackett J., Lesko L., (2003). Microarray data – the US FDA, industry and academia. Nat Biotechnol 21(7):742-743.

Han J., Kamber M., (2006) Data Mining: Concepts and Techniques (2nd edition). Morgan Kaufmann.

Hardiman G., (2004). Microarray platforms --- comparisons and contrasts. Pharmaco-genomics 5:487-502.

Harris T., Lee R., et al (2003). WormBase: a cross-species database for comparative genomics. Nucleic acids research 31:133-137.

Holland J., (1975). Adaptation in natural and artificial systems. Ann Arbor, MI:

University of Michigan Press.

Hong, Z., and Yang, J., (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern recognition, Vol. 24:317-324.

Hood L., Rowen L., et al (1995). Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. Ann N Y Acad. Sci. 758:390–412.

Hubbard M., McHugh N., et al (2000). Isolation of ERp29, a novel endoplasmic reticulum protein, from rat enamel cells: evidence for a unique role in secretory-protein synthesis. Eur J Biochem 267:1945–1956.

Iguchi N., Yang S., et al (2006). An SNP in protamine 1: a possible genetic cause of male infertility? Journal of Medical Genetics 43:382-384.

Jarvinen A., Hautaniemi S., et al (2004). Are data from different gene expression microarray platforms comparable? Genomics 83(6):1164-1168.

Kappen C., Yaworsky P., (2003). Mutation of a putative nuclear receptor binding site abolishes activity of the nestin midbrain enhancer. Biochim Biophys Acta 1625:109–115.

Kothapalli R., Yoder S., et al (2002). Microarray results: how accurate are they? Bioinformatics 3: 1-10.

Kuncheva L., (1995). “Editing for the k-nearest neighbors rule by a genetic algorithm”, Pattern recognition letters, 16:809:814.

Kuncheva L., Jain L., (1999). “Nearest neighbor classifier: Simultaneous editing and feature selection”, pattern recognition letters, 20: 1149-1156.

Kuo W., Jenssen T., et al (2002). Analysis of matched mRNA measurements from two

different microarray technologies. Bioinformatics, 18:405-412.

Launoit Y., Adamski J., (1999). Unique multifunctional HSD17B4 gene product: 17beta-hydroxysteroid dehydrogenase 4 and D-3-hydroxyacylcoenzyme A dehydrogenase / hydratase involved in Zellweger syndrome. Journal of Molecular Endocrinology 22:227-240.

Li J., Liu H., et al (2003). Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics 19(Suppl. 2): i93-i102.

Li J., Dong G., et al (2001). Making use of the most expressive jumping emerging patterns for classification. Knowledge and Information Systems 3: 131-145.

Li J., Liu H., et al (2003). Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic Leukemia(ALL) patients. Bioinformatics 19:71-78.

Li J., Wong L., (2002). Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns. Bioinformatics 18:725-734.

Li T., Zhang C., et al (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20(15):2429-2437.

Lin S., Liao X., et al (2002). Using functional genomic units to corroborate user experiments with the rosetta compendium. Duke University, Durham, NC.

Liu H., Motoda H., (2001). Instance selection and construction for data mining. Kluwer academic publishers. Boston/Dordrecht/London.

Liu H., Li J. et al (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome informatics 13: 51-60.

Liu H., Li J., et al (2003). Use of extreme patient samples for outcome prediction from gene expression data. Bioinformatics 1-9.

Mah N., Thelin A., et al (2004). A comparison of oligonucleotide and cDNA-based microarray systems. Physiol Genomics 16: 361–370.

Malossini A., Blanzieri E., et al (2000). Assessment of SVM Reliability for microarrays data analysis. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics 42:46.

Mao S., Dong G., (2005). Discovery of highly differentiative gene groups from microarray gene expression data using the gene club approach. JBCB 3: 1263-1280.

Mao S., Wang C., Dong G., (2007). Evaluation of inter-laboratory and cross-platform concordance of DNA microarrays through discriminating genes and classifier transferability. Submitted to Journal of bioinformatics (under revision).

Marshall E., (2004). Getting the noise out of gene arrays. Science 306(5696):630-631.

Newcomb PA. (2003 -) Colon Cancer Pathways: Hyperplastic Polyps & Adenomas, Reaserch Grant: National Institutes of Health (NIH).

Nobrega M., Pennacchio L., (2003). Comparative genomic analysis as a tool for biological discovery. The physiology society 554:31-39.

Notarnicola M., Altomare D., et al (2006). Fatty acid synthase hyperactivation in human colorectal cancer: relationship with tumor side and sex. Oncology 71:327-332.

OBrien S., Menotti-Raymond M., et al (1999). The promise of comparative genomics in mammals. Science 286:458-481.

Ooi E., Wormald P., et al (2007). Human cathelicidin antimicrobial peptide is up-regulated in the eosinophilic mucus subgroup of chronic rhinosinusitis patients.

American Journal of Rhinology 21:395-401.

Page D., Zhan F., et al (2002). Comparative data mining for microarrays: A case study based on multiple myeloma. Technical report, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin.

Pennacchio L., Rubin E., (2001). Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109.

Petricoin E., Ardekani A., et al (2002). Use of proteomic patterns in serum to identify ovarian cancer. Mechanisms of disease 359:572-577.

Petricoin E., Hackett J., et al (2002). Medical applications of microarray technologies: a regulatory science perspective. Nat Genet 32(Suppl):474-479.

Piatetsky_Shapiro G., et al (2003). Microarray data mining: facing the challenges. ACM SIGKDD Explorations Newsletter. 5:1-5.

Postlethwait J., Ian G., et al (2000). Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome research 10:1890–1902.

Quinlian J., (1993). C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA.

Ramaswamy S., Tamayo P., et al (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl Acad. Sci. USA, 98, 15149–15154.

Ranz J., Machado C., (2006). Uncovering evolutionary patterns of gene expression using microarrays. Trends Ecol Evol. Jan;21(1):29-37.

Saw J., Yang M., et al (1984). Chebyshev inequality with estimated mean and variance. *The American Statistician*, 38:130-132.

Segal E., (2003). Decomposing Gene Expression into Cellular Processes. Proceedings of

PSB 8:89-100.

Setiawan V., Cheng I., et al (2006). A systematic assessment of common genetic variation in CYP11A and risk of breast cancer. Cancer research 66(24):12019-25.

Shao H., Zhu C., et al (2006). KRAB-containing zinc finger gene ZNF268 encodes multiple alternatively spliced isoforms that contain transcription regulatory domains. International journal of molecular medicine 18:457-463.

Singh D., Phillip G., et al (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer cell 1:203-209.

Shi L., Frueh F., et al (2005). The MAQC (Microarray Quality Control) Project: calibrated RNA samples, reference datasets, and QC metrics and thresholds. In The 11th Annual FDA Science Forum: Advancing Public Health Through Innovative Science: 27 – 28 April:Washington, DC:D-11.

Shi L., Tong W., et al (2004). QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. Expert Rev Mol Diagn 4(6):761-777.

Shi L., Tong W., et al (2005). Microarray scanner calibration curves: characteristics and implications. BMC Bioinformatics 6(Suppl 2):S11.

Shi L., Reid L., et al (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology 24:1151-1161.

Shnyder S., Hubbard M., (2002). ERp29 is a ubiquitous resident of the endoplasmic reticulum with a distinct role in secretary protein production. The Journal of Histochemistry & Cytochemistry 50: 557–566.

Sonesson B., Rosengren E., et al (2003). UVB-induced inflammation gives increased d-

dopachrome tautomerase activity in blister fluid which correlates with macrophage migration inhibitory factor. Experimental dermatology 12:278-282.

Statnikov A., Aliferis C., et al (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21:631-643.

Su Y., Murali M., et al (2003). RankGene: identification of diagnostic genes based on expression data. Bioinformatics 19:1578-1579.

Sugiyama Y., Farrow B., et al (2005). Analysis of differential gene expression patterns in colon cancer and cancer stroma using microdissected tissues. Gastroenterology 128: 480-6.

Tan A., Gilbert D., (2003). Ensemble machine learning on gene expression data for cancer classification. Applied bioinformatics 2: S75-S83.

Tan P., Downey T., et al (2003). Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Research 31(19):5676-5684.

Tomek I., (1976). An experiment with the edited nearest-neighbor rule. IEEE Transactions on systems, Man and cybernetics, SMC-6(6):448-452.

Trapasso F., Yendamuri S., et al (2004). Restoration of receptor-type protein tyrosine phosphatase h function inhibits human pancreatic carcinoma cell growth in vitro and in vivo. Carcinogenesis 25: 2107—2114.

van't Veer L., Dai H., (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530 – 536.

Vapnik V., (1998). Statistical Learning Theory, Wiley, New York

Wang Z., Miura N., et al (2002). Receptor tyrosine kinase, EphB4 (HTK), accelerates

differentiation of select human hematopoietic cells. Blood. 99(8):2740-7.

Wang H., He X., et al (2005). A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics 6:71.

Washbourne P., Schiavo P., et al (1995). Vesicle-associated membrane protein-2 (synaptobrevin-2) forms a complex with synaptophysin. Biochemistry Journal 305:721.

Woo Y., Affourtit J., et al (2004). A comparison of cDNA, oligonucleotide, and Affymetrix genechip gene expression microarray platforms. J Biomol Tech 15(4):276-284.

Wozny W., Schroer K., et al (2007). Differential radioactive quantification of protein abundance ratios between benign and malignant prostate tissues: Cancer association of annexin A3. Protenomics 7(2):313-22.

Xing P., Karp M., (2001). CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. Bioinformatics 1:1-9.

Yauk C., Berndt M., et al (2004). Comprehensive comparison of six microarray technologies. Nucleic Acids Research 32(15):e124.

Yeoh E., Ross M., et al (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell 1:133-143.

Zhai C., Velivelli A., et al (2004). A cross collection mixture model for comparative text mining. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining 743 – 748.

Zhu X., Wu X., (2006). Scalable representative instance selection and ranking. Pattern recognition. ICPR 18th International Conference.