A NOVEL STATISTICAL METHOD FOR MICROARRAY DATA INTEGRATION: APPLICATIONS TO CANCER RESEARCH

by
Lei Xu

A dissertation submitted to the Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March 2007

Abstract

DNA microarray technology has had a tremendous impact on cancer research and microarray gene expression data has been widely used to identify cancer gene signatures which could complement conventional histopathologic evaluation to increase the accuracy of cancer diagnosis and prognosis. However, due to the limited sample size of individual studies, there is often only a small overlap between gene signatures for specific cancers obtained in different studies. With the rapid accumulation of microarray data, it is of great interest to understand how to combine microarray data across studies of similar cancers in order to increase sample size, which could lead to the identification of more reliable gene signatures of specific cancers.

The aim of this thesis is to propose a novel statistical method, based on the top-scoring pair(s) (TSP) classifier family, for inter-study microarray data integration, and to apply it to address two of the key problems in cancer research: identification of robust cancer diagnostic and prognostic signatures.

We develop the rank-based TSP classifier family, including the TSP, *k*-TSP, and G-TSP classifiers. These classifiers only use the rank orders of gene expression values within each profile, and are therefore invariant to most data normalization methods. This property makes them extremely useful for integrating inter-study microarray data since these methods eliminate the need to perform data normalization and transformation. By incorporating the TSP method into different statistical models, we can identify robust gene signatures for cancer diagnosis and prognosis from integrated microarray data. To illustrate the efficacy of these models, we apply them to a large amount of microarray

data, and have identified several important gene signatures for cancer diagnosis and prognosis. All these signatures are properly validated on independent microarray data.

Our work has not only established new models for the identification of robust gene expression signatures from accumulated microarray data, but also demonstrated how the great wealth of microarray data can be exploited to increase the power of statistical analyses. These models will be increasingly useful as more and more microarray data is generated and becomes publicly available in near future. With the inclusion of more samples, cancer gene signatures will be continuously refined and consensus signatures will finally be reached.

Dr. Raimond L. Winslow (advisor, first reader)

Dr. Frederick Jelinek (second reader)

Dr. Trac Duy Tran (committee member)

Dr. Andreas G. Andreou (committee member)

Acknowledgements

First and foremost, I would like to thank my advisor, Professor Raimond L. Winslow, for his insightful guidance and inspiration in research, and for his consistent support and encouragement through all stages of my Ph.D. study. I always feel fortunate to work with such a wonderful mentor. The four years I have spent at the Center for Cardiovascular Bioinformatics and Modeling (CCBM) is the most rewarding and enjoyable in my student life.

In addition, I am grateful to Professor Donald Geman and Professor Daniel Q. Naiman for providing expertise and insight on machine learning and statistics to the work. I also would like to thank Professor Joseph Greenstein for his guidance and help on my first computational modeling project at CCBM, and Dr. Aik Choon Tan for many valuable discussions and assistance on technical issues.

My special thanks to Professor Frederick Jelinek, Professor Trac Duy Tran, and Professor Andreas Andreou for reading my dissertation and serving on my dissertation committee.

I am grateful to Anne Albinak and Jennifer Hopkins for their professional help in making all administrative matters run smoothly. I also would like to thank my fellow colleagues, Dr. Siamak Ardekani, Tabish Almas, Troy Anderson, Christina Yung, Robert Kazmierski, Yasmin Hashambhoy, Hongxuan Zhang, and An-Chi Wei, for making CCBM a pleasant place to work.

I would like to thank my family for their full support and endless love all along. Finally, a special thanks goes to my wife Weiwei Zhang, who always stands by me

throughout my Ph.D. study, and has given me so much love and support to complete this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 DNA Microarrays and Cancer Research

Cancer is a complex and clinical heterogeneous disease. During the past century, the clinical behavior of human cancer has been predicted on the basis of histopathologic examination using microscopy, a process that often fails to reflect the complexity of oncogenesis. The major limitation of this microscopic approach is that it can only predict general categories of cancer and cannot achieve high sensitivity and specificity of prediction in clinical practice [1]. It is known that histologically similar cancer patients may have a different clinical outcome. Consequently, there is a persistent need to find new tools which can complement conventional histopathologic evaluation to increase the sensitivity and specificity of cancer diagnosis and prognosis [2].

It is the central dogma in molecular biology that gene expression is a two-step process in which the information encoded in a gene is first transcribed into messenger RNA

(mRNA) and then translated into protein, the major structural and functional components of a cell. Because cellular processes are governed by the repertoire of expressed genes and the levels and timing of expression, it is of great value and appeal to monitor genome-wide mRNA levels in parallel [3, 4]. The recent completion of the human genome sequence, coupled with advances in biotechnology, has given birth to a new technology for cancer research – DNA microarray technology [3, 5]. DNA microarrays are used to simultaneously measure the transcript abundance (gene expression level) of mRNA for thousands of genes in a given sample. The analysis of this type of data is commonly called gene expression profiling. The introduction of DNA microarray technology has had dramatic impact on cancer research, allowing researchers to analyze expression of thousands of genes in concert and relate gene expression patterns to clinical phenotypes [6]. The DNA microarray technology offers great potential to identify molecular signatures capable of differentiating cancer from normal tissues, predicting outcome, detecting recurrence and monitoring response to cancer treatment. It also could improve our understanding of the cause and progression of cancer, for the discovery of new drug targets.

1.1.1 DNA Microarrays

DNA microarrays rely on specific hybridization of complementary nucleic acid sequences between DNA fragments (termed probes), immobilized on a solid surface in high density, and labeled RNA isolated from biological tissue of interest [7]. A typical DNA microarray consists of thousands of ordered sets of DNA fragments on a glass, filter, or silicon wafer. After hybridization, the signal intensity of each individual probe should correlate with the abundance of the labeled mRNA complementary to the probe.

DNA microarrays fall into two types based on the DNA fragments used to build an array: complementary DNA (cDNA) arrays and oligonucleotide arrays. Although a number of subtypes exist for each array type, spotted cDNA [5] arrays and Affymetrix oligonucleotide arrays [3] are the major platforms currently used by the vast majority of investigators. The choice of which microarray platform to use is based on the research needs, cost, available expertise and accessibility.

For cDNA arrays, cDNA probes, generally manufactured by polymerase chain reaction amplification of cDNA clone inserts (representing genes of interest) from cDNA libraries, are robotically spotted on glass slides or filters. The immobilized sequences of cDNA probes may range greatly in length, but are usually much longer than those of the corresponding oligonucleotide probes. The major advantage of cDNA arrays is the great flexibility in designing a custom array for specific purposes. *A priori* knowledge of cDNA sequence is not required because clones from cDNA libraries can be used and then sequenced if of interest [8, 9]. In addition, cDNA arrays usually cost only one-fourth as much as commercial ones. This flexibility and relatively lower cost makes cDNA arrays popular in academic research laboratories. However, the major disadvantage of these arrays is the amount of total input RNA needed under regular protocol [10]. Another shortcoming of cDNA arrays is that it is difficult to have complete control over the design of the probe sequences, making comprehensive coverage of all genes in a cell impossible. Furthermore, managing large clone libraries and the infrastructure of a relational database for keeping the records, sequence verification and data extraction is not an easy task for most laboratories.

With oligonucleotide arrays, probes can be synthesized directly on the surface of a silicon wafer. These probes are comprised of short segments of DNA complementary to the RNA transcripts of interest. A main advantage of these arrays is their coverage, consistency, and better quality control for the immobilized sequences. Other advantages include uniformity of probe length, the ability to discern gene splice variants, and the availability of carefully designed standard operating procedures. Another advantage particular to Affymetrix arrays is the ability to recover samples after hybridization to an array [8]. This feature makes Affymetrix arrays very attractive in situations where the amount of available tissue is limited. However, a major disadvantage is the high cost of arrays.

There are two different ways to detect the signal intensity of probes on an array after hybridization of a test sample. In the two-color fluorescence hybridization scheme of cDNA arrays, RNA from experimental and reference samples (referred to as target RNAs) are differentially labeled with two fluorescent dyes (e.g. Cye5 vs. Cye3) and hybridized onto the same array. The target RNAs will hybridize to the corresponding complementary probes that have been spotted on the array surface. When the region of the probe is fluorescently illuminated, both the experimental and reference target RNAs will fluoresce and the relative balance of green versus red fluorescence which indicates the relative expression levels of experimental and reference target RNAs can be measured. Therefore, gene expression values are reported as ratios between two fluorescent values. Alternatively, Affymetrix oligonucleotide arrays use a one-color fluorescence hybridization system where experimental RNA is labeled with a single color fluorescent dye and hybridized onto an oligonucleotide array. After hybridization, the fluorescence

intensity from each spot on the array provides a measurement of the abundance of the corresponding target RNA which is hybridized to that spot.

For the analysis and interpretation of microarray data, a number of innovative computational tools are available. These tools can be divided into unsupervised learning (i.e. clustering) and supervised learning (i.e. classification) methods. Unsupervised learning involves the aggregation of samples, genes or both into different clusters based on similarity of measured gene expression values [11]. The goal of clustering is to group together individual objects with similar properties, leading to clusters where the similarity measure is small within clusters and large between clusters. Several clustering methods from classical pattern recognition, such as hierarchical clustering, *K*-means clustering and self-organizing maps, have been applied to microarray data analysis [12, 13]. Using unsupervised learning methods, we can search for candidate genes whose expression pattern could be associated with a given biological condition. The advantage of unsupervised learning is that it is unbiased and allows for identification of significant structures in a complex dataset without any *a priori* information about the objects. However, a typical microarray study generates expression data for thousands of genes from a relatively small number of samples (usually less than 100), therefore many different relationships are possible in the dataset and the predominant cluster revealed by clustering may not necessarily provide important clues as to biological differences of interest [8].

In contrast, supervised learning methods integrate the knowledge of sample class label information into the analysis and the goal of supervised learning is to identify expression patterns (i.e. gene expression signatures) which could be used to classify unknown

samples according to their biological characteristics [14]. A training dataset, consisting of gene expression values and sample class labels, is used to select a subset of expressed genes which have the most discriminative power between the classes to be predicted and build a predictive model, also called classifier (e.g. *k*-nearest neighbors, neural network, support vector machines), which takes gene expression values of the pre-selected set of genes of an unknown sample as input and outputs the predicted class label of the sample. Supervised learning involves learning from examples (i.e. the gene expression measurements of each sample and the corresponding known class label), and therefore it depends on accurate sample class labels, which can be an issue given the limitations of histopathologic cancer diagnosis [8]. Supervised learning can be a two-class classification problem (e.g. cancer vs. normal) or multi-class classification (e.g. subtypes of a certain cancer type). With supervised learning, novel molecular diagnostic and prognostic tools based on gene expression profiling can be developed.

1.1.2 DNA Microarrays in Cancer Research

In the past several years, DNA microarray technology has been widely used in cancer research and many studies have used this technology to identify candidate gene expression signatures to predict the diagnostic category or prognostic stage of a cancer patient [12, 14-32]. We discuss the utility of DNA microarrays in cancer research and review several important recent studies on molecular diagnostic classification and clinical outcome prediction in the following sections.

Molecular Diagnostic Classification

One of the common problems in clinical cancer research is the fact that histopathological identification and classification of cancer can be quite challenging. Morphologically indistinguishable cancers may belong to clinically distinct classes even though they arise from the same origin. The ability to classify unknown samples into different categories may offer great potential for more accurate and systematic cancer diagnosis.

The use of gene expression profiling for cancer diagnosis was first demonstrated by Golub et al. [14] in a study using DNA microarrays to study the gene expression of 6,817 genes in 72 human acute leukemia tumor samples. In this study, by using unsupervised learning, leukemia tumor samples were clustered into two clusters of known subtypes of leukemia - acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) - based solely on gene expression profiling. By using supervised learning, a weighted gene voting classifier, built from a subset of informative genes chosen based on their correlation with the class distinction between AML and ALL, was successfully used to assign a group of unknown samples into the correct category. The accuracy of the classifier was assessed by both cross-validation on the initial training data and independent test on an independent set of samples. This study demonstrates the feasibility of cancer classification based solely on gene expression.

A variety of studies have subsequently used gene expression profiling for cancer classification [15, 18, 22, 25-28, 33, 34]. Bittner et al. [18] reported the discovery of a subset of cutaneous melanomas identified by mathematical analysis of gene expression in a series of 31 melanomas samples. Both the hierarchical clustering of the 31 melanoma samples and the non-hierarchical cluster affinity search technique (CAST) algorithm

identified the identical major cluster of 19 melanomas which had very similar gene expression profiles. This cluster was also a compact, readily separable grouping based on its overall similarity of expression pattern viewed by a three-dimensional multidimensional scaling (MDS) plot. On the basis of the linear correlation of global gene expression with clusters, MDS analysis was used to generate a list of genes with the most power to define the major melanoma cluster of 19 samples. Further analysis of the pattern of gene expression in phenotypically characterized cells with respect to the genes that define the major cluster of 19 cutaneous melanomas indicated that samples assigned within the major cutaneous melanoma cluster would have reduced motility and reduced invasive ability.

In another study, Hedenfalk et al. [26] demonstrated that in hereditary breast cancers, the gene expression profiles of tumors with BRCA1 mutations, tumors with BRCA1 mutations, and sporadic tumors differed significantly from each other. Statistical analyses were used to identify a set of 176 genes which could accurately distinguish tumors with BRCA1 mutations from tumors with BRCA2 mutations on the basis of gene expression profiles of these genes.

Diagnostic cancer classification studies have been performed in a wide range of cancers, including breast cancer [22], lung cancer [25, 35], prostate cancer [31, 36-38], bladder cancer [39], head and neck cancer [33], and ovarian cancer [15, 40].

For classification of multiple categories of human cancers, several recent studies explored the potential use of multiclass tumor classifiers built from microarray gene expression data to discriminate different kinds of tumor classes based on tissue origin [27, 28, 41-45]. In an effort to classify multiple common adult cancers purely by molecular

classification, Ramaswamy et al. [27] collected gene expression profiles of 16,063 genes and expressed sequence tags for 90 normal tissue samples and 218 tumor samples spanning 14 tumor types. The accuracy of a multiclass classifier based on a support vector machine algorithm was evaluated by both cross-validation and independent test data. The overall classification accuracy was approximately 78%. They also found that poorly differentiated cancers could not be accurately classified according to their tissues of origin, indicating that they have a very different gene expression patterns compared to their well differentiated counterparts.

In a similar study, Su et al. [28] reported using human tumor gene expression profiles of 175 tumor samples containing 10 cancer types and supervised machine learning algorithms to distinguish a number of carcinoma classes. They first pre-filtered to exclude unreliable gene sets, and then identified signature gene subsets typical for each cancer class from the remaining genes. For an unknown sample, the classification rule was based on the class distance (i.e. sum of absolute differences of signature genes) from the unknown sample to the members of that class in the training set. The performance of the multiclass classification method was assessed by both cross-validation and independent tumor samples. The classifier accurately predicted the anatomical site of tumor origin for 90% of the tumors studied.

Some other multiclass cancer classification studies include the method of nearest shrunken centroids by Tibshirani et al. [42], a genetic algorithm-based maximum likelihood classification method by Ooi and Tan [41], a genetic algorithm-based support vector machine method by Liu et al. [43], and a rank-based pair-wise gene comparison classification method by Tan et al. [44].

Different classification methods have been proposed for molecular classification of cancers based on gene expression profiles. These methods include classification trees [46, 47], nearest-neighbor classifiers [45, 48, 49], linear discriminant analysis (LDA) [50, 51], support vector machines (SVMs) [27, 43, 52-54], artificial neural networks (ANNs) [55, 56] and some other machine learning approaches such as boosting [57] and bagging [58, 59].

In an exploratory study, Zhang et al. [46] introduced the technique of recursive partitioning based on classification trees for classifying tumors on the basis of gene expression data and demonstrated that it could accurately discriminate colon tumor tissues from normal colon tissues. This group also expanded the technique to deterministic forests of classification trees and showed that the deterministic forests outperformed the single classification trees [47]. The main advantage of classification tree-based methods is biological interpretability of the genes forming the trees or forests.

Similarly, Giordano et al. [45] employed a nearest neighbor classifier to classify 154 cancer samples to three known cancer categories based on gene expression data. 152 of the 154 samples were correctly classified by the classifier when estimated use a cross-validation procedure. One of the advantages of nearest neighbor classifiers is its simplicity and intuitiveness. For large data sets such as microarray data sets, one drawback of this method is the computational load, both in finding the neighbors and storing the whole training data [60].

Support vector machines are among the most popular classifiers used in microarray data analysis. SVMs are designed for binary classification; however, they can be easily generalized to deal with multiclass problems. SVM classification was first applied to

cancer microarray data by Mukherjee et al. [61] and then used widely in molecular classification using microarray expression data [27, 43, 52-54].

As the first application of ANNs for diagnostic classification of cancer using microarray data, Khan et al. [56] developed a method of classifying round blue-cell tumors to four distinct diagnostic categories based on gene expression data using ANNs. The ANN models correctly classified all independent test samples into their categories. Other applications of ANNs for cancer classification were also reported [55, 62].

Several variants of linear discriminant analysis, such as eigengene-based LDA and uncorrelated LDA, were proposed to deal with the small sample sizes and large variables (genes) of typical gene expression data [50, 51]. A few studies demonstrated that recent machine learning approaches such as bagging and boosting could been used in cancer classification based on microarray data to increase the performance of classification methods [57-59].

In addition to above well-established classification in traditional statistical learning, new classification methods have been developed to deal with the so-called 'small n, large p' problem, that is, a very large number of variables (genes) relative to the number of observations (tumor samples) [63-66]. Microarray gene expression studies suggest that simple methods, such as nearest neighbor classifiers, perform as well as more complex approaches, such as neural networks and support vector machines [67].

Clinical Outcome Prediction

One of the most interesting, and most promising, applications of DNA microarrays in cancer research might be the prediction of clinical outcome based on gene expression

profiles. Although not mutually excluded from diagnostic cancer classification, clinical outcome prediction based on gene expression profiles mainly deals with the correlation between gene expression profiles and clinical outcome, and the prediction of clinical outcome using molecular prognostic signatures.

The first demonstration of expression-based correlates of clinical outcome was performed in diffuse large B-cell lymphoma (DLBCL) samples by Alizadeh et al. [17]. This study used hierarchical clustering of gene expression in B-cell malignancies and identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes correlated with germinal centre B cells (germinal centre B-like DLBCL) and the second type expressed genes normally expressed in activated peripheral blood B cells (activated B-like DLBCL). Germinal centre B-like patients had a significantly better overall survival than activated B-like DLBCL patients. The molecular classification of DLBCL on the basis of gene expression led to the discovery of previously undetected and clinically significant subtypes of DLBCL. These finding established the feasibility of clinical outcome prediction purely based on gene expression. Outcome prediction of DLBCL has been followed up with a larger patient cohort and is actively being translated into a clinical measure of the likely clinical outcome in patients with DLBCL [24, 30, 68].

Breast cancer outcome prediction has also been a major focus of microarray gene expression study. Lymph-node status at diagnosis is the most important measure for future relapse and overall outcome in breast cancer patients. Adjuvant systemic therapy (chemotherapy or hormonal therapy) reduces the risk of distant metastases by approximately one-third; however, the majority of patients receiving adjuvant therapy

would survive breast cancer long term without it [69]. A high profile study by van't Veer et al. [21] applied DNA microarrays to primary tumors of 117 patients with lymph-node-negative breast cancer to identify a 70-gene prognostic signature predictive of clinical outcome (disease-free after a period of at least five years vs. distant metastases within five years). The prognostic signature was subsequently validated on 295 patients (61 of the 295 patients included in the original cohort of 117 patients) and was confirmed to be an independent prognostic factor that outperformed currently used clinical parameters in predicting breast cancer outcome [70]. These results suggested the possibility that such prognostic signatures could be used to spare the low-risk patients the cost and toxicity of unnecessary adjuvant systemic therapy. Clinical outcome prediction of breast cancer has been expanded by other studies [12, 23, 71-75] and similar prognostic signatures have been proposed. The validation and clinical transition of these signatures has been on its way [76-78].

Outcome prediction based on microarray gene expression profiling has also been explored in a number of cancers, such as prostate cancer [19, 31], lung cancer [16, 79], renal cancer [20], and brain cancer [29]. These studies illustrated the great potential of gene expression prognostic signatures in predicting cancer clinical outcome. While exciting, further work is needed to prospectively validate these microarray-based prognostic signatures.

1.2 Motivation for Microarray Data Integration

The advent of DNA microarrays provides a powerful tool in cancer research and many studies have used this technology to identify genes that could be used as candidate

components of gene expression signatures to predict the diagnostic category or clinical outcome of a cancer patient [14, 17-21, 23, 71-74, 80-82]. However, microarray data has the characteristic of high noise. In addition, the large expense of microarray experiments (~ $1,000 per sample) and the limited number of available cancer patients make the sample size of individual microarray studies relatively small, usually less than 100, with respect to the number of variables (genes) in gene expression data, typically at the order of 10,000.

It is quite interesting, but perhaps not surprising, that the gene expression signatures from different investigations involving patients with the same cancer type are study-specific, that is, there are small overlaps between signatures across the different studies focusing on a similar cancer type [12, 17, 21, 30, 68, 72, 83-85]. These diverse results make it difficult to identify the most reliable signatures for cancer diagnosis and prediction of clinical outcome.

The disagreements in gene signatures may be partly due to the use of different microarray platforms, differences in patient selection, and experimental issues. However, one central possible cause may be related to the observation in a recent study by Ein-Dor et al. [84]. In this study, reanalysis of the van 't Veer data [21] has shown that the prognostic signature is not unique even from the same data set and it is strongly influenced by the subset of the patients used for signature identification. Given the large numbers of variables in a microarray data set and the relatively small numbers of samples used in the training set of individual studies, it is highly possible to combine genes in many ways to produce signatures with similar predictive power [86]. Therefore, the disparity between gene signatures may be mainly contributed to the relatively small

sample sizes in individual studies and it is clear that much larger numbers of samples are needed to develop more robust prognostic signatures. Indeed, in a recent study performed by Mukherjee et al. [87] concerning the influence of size of the microarray data set, it was concluded that increased sample size improves both the accuracy and significance of classification results.

Given the costs inherent in microarray studies and the rarity of certain tumor specimens, it is very difficult to performing large-scale studies. In addition, many human cancer studies involve valuable clinical specimens and are difficult to repeat. The rapid accumulation of cancer gene expression data suggests that combining microarray data sets generated in different laboratories addressing a similar question may be a useful way to increase sample size. An intriguing advantage of inter-study microarray data integration is that it increases the statistical power to capture consistent features which might be masked by the small sample size and experimental artifacts in an individual data set. This could lead to the discovery of more robust and reliable cancer signatures which may offer more accurate cancer classification and improve the statistical significance of their association to clinical outcome. On the other hand, as more and more cancer microarray data becomes publicly available, there is an urgent need for methods that enable integration of microarray data generated by different research groups.

1.3 Issues with Microarray Data Integration

Ideally, gene expression data obtained in any laboratory, at any time, using any microarray technology, should be comparable. However, this is not true in reality. The lack of uniform standard poses a big obstacle to microarray data integration. Several

issues arise when attempting to integrate microarray data generated by disparate groups using different array technologies.

Several studies have shown that expression measurements from different microarray technologies, such as spotted cDNA and oligonucleotide arrays, may show poor correlation and may not be directly comparable [88-92]. The observed disagreement may be due to the differences in probe set content, deposition technologies, labeling and hybridizing protocols, as well as data extraction procedures (e.g. background correction, normalization, and calculation of expression values). For example, cDNA microarray data is usually defined as ratios between experimental and control expression values and cannot be directly compared with oligonucleotide microarray data that is defined as expression values of experimental samples.

Besides the observation of disagreement between different microarray technologies, multiple-laboratory comparison of microarrays has demonstrated that there were relatively larger differences in data obtained in labs using the same microarray technology than that obtained in the same lab using different microarray technologies [93, 94]. This suggested that microarray data obtained from different labs could not be directly compared even the data were generated using the same microarray technology.

Commercial microarrays, such as Affymetrix arrays, have produced several generations of microarrays to keep up with advances in genome sequencing. The number of known genes and the representative composition of gene sequences are frequently updated with new developments of biotechnology. As a result, probe sets representing newly discovered genes are incorporated into new generations of commercial microarrays and some existing probe sets are modified to better detect the target gene sequences. A

recent study has shown that expression measurements within one generation of Affymetrix arrays are highly reproducible, but that reproducibility across generations depends on the degree of similarity of the probe sets and the levels of expression measurements [95]. Therefore, even when using the same microarray technology, different generations of microarrays have different probe sets and duplicate spots, making direct integration difficult.

In addition, variation among data sets, resulting from technical variability, including differences in sample composition and preparation, experimental protocols and parameters, RNA quality, and array quality, pose further challenges to the integration of microarray data from independent studies.

1.4 Previous Work on Microarray Data Integration

The rapid accumulation of microarray data has created a need for methods to effectively integrate data generated with different array platforms from disparate labs. In general, integrating multiple data sets promises to yield more reliable and more valid results since analyses may be performed using a larger number of samples and the effects of individual study-specific biases are reduced. Recently, several methods have been proposed to combine inter-study microarray data at different levels in microarray study [32, 96-108]. These methods fall into two major categories based on the level at which data integrated is performed: meta-analysis, which combines results (e.g. *t*-statistics) from individual data sets to avoid the direct comparison of gene expression values, and direct integration of expression values after specific data transformation and normalization on individual data sets.

Instead of integrating microarray gene expression values, meta-analysis methods combine results of individual studies to increase the power of identifying significantly expressed genes across studies [96-98, 101, 103-107]. Rhodes et al. [97] proposed a statistical model for performing meta-analysis of four independent microarray data sets from two different microarray technologies: spotted cDNA arrays and Affymetrix arrays. Each gene in each study was treated as an independent hypothesis and significance (denoted by one *p* value and one *q* value) was assigned to each gene in each study based on random permutations. Then the similarity of significance across studies was assessed with meta-analysis methods combined with multiple inference statistical test for each possible combination of studies. A cohort of genes was identified to be consistently and significantly dysregulated in prostate cancer. Choi et al. [98, 105] introduced a new meta-analysis method, which combines the results from individual data sets in the form of effect size and has the ability to model the inter-study variation. The effect size was defined to be a standardized mean difference between cancer and normal samples in a microarray data set. The effect sizes from multiple microarray data sets were combined to obtain the estimate of the overall mean and the statistical significance was determined by permutation test extended to multiple data sets. It was demonstrated that data integration using this method promoted the discovery of small but consistent expression changes and increase the sensitivity and reliability of analysis. An extended effect size model was then proposed by Hu et al. [106] for meta-analysis of microarray data. Other meta-analysis methods include a statistical approach within a Bayesian framework to combine multiple microarray data sets [104].

Even with some reported success for meta-analysis of microarray data, one main weakness of above meta-analysis studies is that the small sample sizes typical of individual microarray studies, coupled with variation due to differences in study protocols, will affect the results from individual studies and inevitably degrade the final results of meta-analysis. In addition, a recent study has demonstrated that there is only moderate overlap between gene detection on different array platforms [89].

In contrast to meta-analysis approach where results of individual studies are combined at an interpretative level, direct integration methods integrate microarray data from different studies at expression value level after transforming the expression values to numerically comparable measures [32, 99, 100, 102, 108]. In general, the procedure can be divided into following steps. First, a list of genes common to multiple distinct microarray platforms is extracted based on cross-referencing the annotation of each probe set represented on the microarrays. Cross-referencing of expression data is usually achieved using UniGene database [109]. Next, for each individual data set, numerically comparable quantities are derived from the expression values of genes in the common list by application of specific data transformation and normalization methods. Finally, the newly derived quantities from individual data sets are combined to increase sample size and statistical methods are applied to the combined data to build diagnostic and prognostic signatures.

We summarize some methods of the direct integration category reported in microarray study literature. Ramaswamy et al. [32] re-scaled expression values of a common set of genes for each of the five microarray data sets generated by independent labs on several microarray platforms, and then combined them to form a larger data set. A gene

expression signature that distinguished primary from metastatic cancers was identified from the combined data set with increased sample size. Bloom et al. [102] used a scaling approach based on measurements for one common reference sample to integrate microarray data from different platforms. Shen et al. [99] proposed a Bayesian mixture model to transform each raw expression value into a probability of differential expression (i.e. *poe*) for each gene presenting on each independent array data set. Integrating multiple studies on the common probability scale of *poe*, they developed a 90-gene meta-signature that predicted relapse-free survival in breast cancer patients with improved statistical power and reliability. In addition to common data transformation and normalization procedures, Jiang et al. [100] proposed a distribution transformation method to transform multiple data sets to a similar distribution before data integration. Data processed with distribution transformation showed a greatly improved consistency in gene expression patterns between multiple data sets [100]. More recently, Warnat et al. [108] used two data integration methods, namely median rank scores and quantile discretization, to derive numerically comparable measures of gene expression from independent data sets generated from different microarray platforms. These transformed data were then integrated and used to build support vector machine classifiers for cancer classification. Results showed that cancer classification based on microarray data could be greatly improved by integrating multiple data sets with similar focus. The classifiers built from integrated data showed high predictive power and improved generalization performance [108].

One major limitation of these direct integration methods is that filtering of genes to a subset common to multiple distinct microarray platforms often excludes many thousands

of genes which may include potential significant genes. Furthermore, there is still no consensus on how best to perform data transformation and normalization.

1.5 Thesis Overview

The aim of this thesis is to present a novel statistical method for microarray data integration and to apply it to cancer diagnostic and prognostic signature identification. The rest of the thesis is organized as follows.

Chapter 2. A Rank-Based TSP Classifier Family. This chapter lays the methodological foundation for microarray data integration in this thesis. We first refine the top-scoring pair(s) (TSP) classifier originally proposed by Geman et al. [64] as a new approach for molecular classification. Next, two generalizations of the TSP classifier, the *k* disjoint TSP (*k*-TSP) classifier and the group-based TSP (G-TSP) classifier, are developed to deal with the problem that the TSP classifier is somewhat sensitive to small perturbation of training data. Then, the performance of the TSP classifier family is compared with that of several well-known classification methods, such as support vector machines, *k*-nearest neighbor classifiers, on individual microarray data sets. Finally, we indicate that an important property of the TSP method makes it useful for microarray data integration.

Chapter 3. Cancer Diagnosis: A New Prostate Cancer Signature Emerges from Microarray Data Integration. Here we propose a novel, simple method of integrating different microarray data sets to identify cancer signatures for individual cancer types and apply the method to prostate cancer microarray data sets. By applying the TSP method, we have identified a robust gene signature by integrating microarray data sets from three different prostate cancer studies. Cross-platform validation shows that the TSP classifier

built from the signature genes, which simply compares relative expression values, achieves high accuracy, sensitivity, and specificity on independent data sets generated using various array platforms. Our findings suggest a new model for the discovery of cancer signatures from accumulated microarray data and demonstrate how the great wealth of microarray data can be exploited to increase the power of statistical analysis.

Chapter 4. Essential Transcriptional Features of Cancer: A Common Cancer Signature from Large-Scale Integration of Microarray Data. One of the central themes in cancer research is the identification of molecular alterations common to all cancers. This chapter addresses this issue by integrating large-scale cancer microarray data across a broad range of cancer types. Specifically, we apply the G-TSP classifier and a repeated random sampling strategy to integrated training data sets and identify a common cancer signature consisting of 46 genes. These 46 genes are naturally divided into two distinct groups; those in one group are typically expressed less than those in the other group for cancer tissues. Given a new expression profile, the classifier discriminates cancer from normal tissues by ranking the expression values of the 46 genes in the cancer signature and comparing the average ranks of the two groups. This signature is then validated by applying this decision rule to independent test data. Upon further validation, this signature may be useful as a robust and objective diagnostic test for cancer.

Chapter 5. Cancer Prognosis: A Robust Breast Cancer Prognostic Signature Identified from Inter-Study Microarray Data. By using the TSP method as a novel feature selection and transformation procedure, we have integrated three independent microarray gene expression data sets of extreme samples and identified a robust 61-gene breast cancer prognostic signature from the integrated training data set of 358 samples. The signature

has shown 88.9% sensitivity and 63.0% specificity in an independent test set of 154 samples. The gene signature is highly informative in assessing the risk of developing distant metastases within five years (hazard ratio 11.9 with 95% CI 3.7-38.0). Our findings suggest that, upon further confirmation on large-scale independent data, the prognostic signature could potentially provide a powerful tool to guide adjuvant systemic treatment that could greatly reduce the cost of breast cancer treatment, both in terms of toxic side effects and health care expenditure.

Chapter 6. Conclusions and Future Work. This chapter concludes the thesis by summarizing the major results and contributions of the thesis and suggesting directions for future work.

Chapter 2

A Rank-Based TSP Classifier Family

2.1 Introduction

The TSP classifier was first introduced by Geman et al. [64] as a new classification technique for microarray data based entirely on relative gene expression values, specifically pairwise comparisons between two gene expression values. This method was motivated by two current practical and technical limitations in using microarray data for class prediction: small sample and lack of interpretability. Accurate statistical inference is difficult due to the small number of samples, typically tens, relative to the large number of variables (genes), typically thousands. The decision boundaries generated by most standard classifiers, such as *k*-nearest-neighbor classifiers, support vector machines, neural networks, are usually highly complex and thus difficult to interpret in biologically meaningful terms.

The TSP classifier is a parameter-free, data-driven machine learning method, which avoids over-fitting by eliminating the need to perform specific parameter tuning, as in other learning techniques (e.g. support vector machines, neural networks). This classifier discriminates between two classes by finding pairs of genes that achieve the largest score defined by a simple measure of discrimination (described in details in Section 2.2). This approach only uses the ranks (orderings) of gene expression values within each profile, therefore is invariant to monotonic pre-processing designed to overcome array-to-array variation, such as most data normalization methods. Whereas information is lost using a rank-based method, the results obtained by the TSP classifier on several different microarray data sets have shown that rank information within each microarray is sufficient to perform molecular classification reliably. In fact, despite its simplicity with respect to other methods, the TSP classifier achieves classification rates comparable to or exceeding the best results reported in the literature [64]. In addition, the TSP classifier provides decision rules which: i) involve very few genes; ii) are both accurate and transparent; and iii) provide specific hypotheses for follow-up studies [64].

The TSP classifier defined in Geman et al. [64] depends on the pairs of genes that achieve the largest score. It is possible for multiple gene pairs to achieve the same top score. In order to eliminate ties and select a unique pair from the top-scoring pairs, we introduce a secondary score, called the rank-score, based on the rank differences in each sample in each class. In the remainder of this thesis, the TSP classifier means the modified TSP classifier with only one unique gene pair. In some instances, the TSP may change when the training data is slightly perturbed by adding or deleting a few samples [64]. Therefore, we develop two generalizations of the TSP classifier, the *k*-TSP classifier

and the G-TSP classifier, to deal with this problem, as well as to increase the accuracy of the TSP classifier. Class prediction of the *k*-TSP classifier is made by unweighted majority voting of the *k* disjoint top-scoring gene pairs selected from the training data. The classification rule of G-TSP classifier is based on average relative ranks of two groups of genes. There are different ways to construct the two groups of genes from the training data. The *k*-TSP and G-TSP classifiers, together with the TSP classifier, form the TSP classifier family.

In the following sections, we describe the TSP classifier family in details, demonstrate the efficacy of our classifiers by comparing them to other well-known classification methods, and indicate why this classifier family can be useful for integrating multiple microarray data sets.

2.2 The TSP Classifier

Formulation of the TSP classifier is similar to what has been described in [64]. Consider P genes whose expression values $\boldsymbol{X} = \{X_1, X_2, \ldots, X_P\}$ are measured using a DNA microarray and regarded as random variables. The class label Y for each profile $\boldsymbol{X}$ is a discrete random variable that can take on values in $\{1, 2, \ldots, C\}$. We only discuss two-class classification problems (e.g. cancer vs. normal) in this thesis, therefore, we assume $C = 2$. We focus on detecting a signature gene pair (i, j) for which there is a significant different in the probability of the event $\{X_i < X_j\}$ from class 1 to class 2. The quantities of interest are $p_{ij}(c) = P(X_i < X_j \mid Y = c)$, $c \in \{1, 2\}$, that is, the probabilities of observing the event $\{X_i < X_j\}$ in each class. Let $\Delta_{ij} = \left| p_{ij}(1) - p_{ij}(2) \right|$ denote the score of the gene pair (*i,*

j). Our goal is to find a gene pair (i, j), called a signature gene pair, which achieves the highest score. Class prediction is then based on the signature gene pair.

The TSP classifiers are rank-based, meaning that the decision rules only depend on the relative ordering of the gene expression values within each profile. This should not be confused with rank-based methods for determining differentially regulated genes in which the expression values for each fixed gene are ordered across profiles.

2.2.1 Learning the TSP Classifier

Consider a gene expression profile consisting of P genes $\{1, 2, \ldots, P\}$ and suppose there are N profiles or samples, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, available for training. Here the column vector $\boldsymbol{x}_n = \{x_{1n}, x_{2n}, \ldots, x_{Pn}\}$, $n \in \{1, 2, \ldots, N\}$, denotes the P expression values for the n-th sample. These data can then be represented as a matrix of dimension $P \times N$, $[x_{pn}]_{P \times N}$, $p \in \{1, 2, \ldots, P\}$ and $n \in \{1, 2, \ldots, N\}$, where the expression value of the p-th gene, $p \in \{1, \ldots, P\}$, from the n-th sample, $n \in \{1, 2, \ldots, N\}$, is denoted by x_{pn}. Each column represents a gene expression profile of P genes and each row represents observations of a particular gene over N samples.

Let $(y_1, y_2, \ldots, y_N)$ be the vector of class labels for the N samples, where $y_n \in \{1, 2\}$, $n \in \{1, 2, \ldots, N\}$, the set of possible class labels for two-class problems. The labeled training set is then $\boldsymbol{S} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, where $\boldsymbol{x}_n$, $n \in \{1, 2, \ldots, N\}$, is the n-th column vector of the matrix of gene expression profiles. Recall that the expression profile, $\boldsymbol{X}$, and its class label, Y, are random variables. As usual, we assume the elements of $\boldsymbol{S}$

represent independent and identically distributed samples from the underlying probability distribution of ($\boldsymbol{X}$, Y).

The quantities of interest are the probabilities $p_{ij}(c) = P(X_i < X_j \mid Y = c)$, $c \in \{1, 2\}$ for each gene pair (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$. These probabilities are estimated by the relative frequencies of observations of the event $\{X_i < X_j\}$ within profiles and over samples. We compute the score, $\Delta_{ij} = \left| p_{ij}(1) - p_{ij}(2) \right|$ for every pair of genes (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$. Obviously, pairs of genes with high scores are viewed as most informative for classification.

We assume that there are N_1 samples of class 1 and N_2 samples of class 2, $N_1 + N_2 = N$, in the training data $\boldsymbol{S}$. For each pair of genes (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$, we estimate the probabilities $p_{ij}(c)$, $c \in \{1, 2\}$ based on the training set $\boldsymbol{S}$ by

$$p_{ij}(c) \approx \frac{N_{ij}^{(c)}}{N_c}, \quad c \in \{1,2\}, \tag{2.1}$$

where

$$N_{ij}^{(c)} = \left| \{n : 1 \le n \le N, x_{in} < x_{jn}, y_n = c\} \right|, c \in \{1,2\}. \tag{2.2}$$

Here, |A| denotes the number of samples in a set A. Therefore, the corresponding score is estimated by

$$\Delta_{ij} = \left| p_{ij}(1) - p_{ij}(2) \right| \approx \left| \frac{N_{ij}^{(1)}}{N_1} - \frac{N_{ij}^{(2)}}{N_2} \right|. \tag{2.3}$$

Consequently, it is sufficient to know the ranks of gene expression values within each profile to obtain all the scores Δ_{ij}, $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$. The next step is to select a unique signature gene pair achieving the largest score. If there is only one pair achieving the top score, we take this pair to be the signature gene pair. Otherwise (i.e. if multiple pairs achieve the top score), we use a more sensitive score, called the rank-score, in order to break the tie and find a unique pair from the TSPs [44, 110]. The rank-score takes into account the extent to which a gene pair inverts from one class to the other. For each gene pair (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$, the rank-score, denoted by δ_{ij}, is defined as

$$\delta_{ij} = \left| \frac{1}{N_1} \sum_{n \in \{m:1 \leq m \leq N, y_m = 1\}} (R_{in} - R_{jn}) - \frac{1}{N_2} \sum_{n \in \{m:1 \leq m \leq N, y_m = 2\}} (R_{in} - R_{jn}) \right|. \qquad (2.4)$$

Here R_{in} is the rank of the expression value of gene i within the n-th profile. Finally, we select the TSP with the highest rank-score. The algorithm of learning the TSP classifier is illustrated in Figure 2.1.

TSP Algorithm
Input: Training sample $\boldsymbol{S}$ of P genes and N arrays.
Output: TSP classifier, i.e., a signature gene pair.

1. Compute a score Δ_{ij} based on the training set for every pair of genes (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$.
2. Select the maximum score Δ_{max} from the list of scores.
3. Proceed down the list of pairs, if $\Delta_{ij} = \Delta_{max}$, recruit pair (i, j) into a candidate list.
4. Calculate the rank score δ_{ij} for each pair (i, j) in the candidate list.
5. Sort the rank score δ_{ij} from largest to smallest.
6. Select the gene pair achieving the largest δ_{ij} from the candidate list to be the TSP classifier.
7. Return the TSP classifier

Figure 2.1: Description of the TSP algorithm.

2.2.2 Prediction with the TSP Classifier

If the gene pair (*i, j*) is the signature gene pair learned from the training data, the classification decision is made by comparing the expression values of the two genes in the TSP (*i, j*) on a test sample. For a new test sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_P\}$, the class label, y, of the test sample is classified by the TSP classifier, $f_{TSP}(\boldsymbol{x})$, as

$$y = f_{TSP}(\mathbf{x}) = \underset{c \in \{1,2\}}{argmax} \begin{cases} N_{ij}^{(c)} / N_c, & if \quad x_i < x_j \\ 1 - N_{ij}^{(c)} / N_c, & otherwise \end{cases}. \tag{2.5}$$

In other words, suppose $p_{ij}(1) \geq p_{ij}(2)$. In this case, if we observe that $x_i < x_j$, then the TSP classifier votes for class 1; otherwise, i.e. if $x_i \geq x_j$, it votes for class 2. On the other hand, suppose $p_{ij}(1) < p_{ij}(2)$. Then, if we observe that $x_i < x_j$, the TSP classifier votes for class 2; otherwise, it votes for class 1. In other words, the TSP classifier chooses the class for which the observed ordering between the expression values of gene *i* and *j* is the most likely.

It is also noteworthy that the sum of misclassification probabilities over the two classes can be expressed as $1 - \Delta_{ij}$, which provides a natural justification for score maximization.

2.3 The *K*-TSP Classifier

As mentioned before, the top scoring pair may change when the training data is perturbed by adding or deleting a few samples [64]. Hence, we introduce the *k*-TSP classifier, which extends the TSP classifier, and is designed to deal with this problem, as well as increase the accuracy of the TSP classifier, by generating a more stable classifier. This is accomplished by basing classification on the *k* disjoint top scoring pars of genes

which achieve the best combined score [44]. We can view the *k*-TSP classifier as an ensemble learning approach where the intention is to combine the discriminating power of many 'weaker' rules to make more reliable predictions. In this case, there are *k* 'weaker' rules, one for using each of the *k* top-scoring pairs to classify according to Equation (2.5).

2.3.1 Learning the *K*-TSP Classifier

The learning algorithm of the *k*-TSP classifier is similar to that of TSP. It consists of first forming a list of gene pairs, sorted from the largest to the smallest according to their scores Δ_{ij}, and then breaking ties by sorting within those that achieve the same score using the rank-score δ_{ij} [44]. The *k*-TSP classifier uses the *k* top scoring disjoint gene pairs from this list. The procedure is straightforward: take the first pair (i_1, j_1), then go down the list until arriving at the first pair (i_2, j_2) which does not involve either i_1 or j_1, and continue in this manner until reaching the *k*-th disjoint pair (i_k, j_k). The parameter *k* is determined by cross-validation; with the restriction that *k* does not exceed a certain number, K_{max}, and is an odd number in order to break the tie during unweighted majority voting procedure. Figure 2.2 illustrates the *k*-TSP learning algorithm.

2.3.2 Prediction with the *K*-TSP Classifier

Suppose that the gene pairs (i_q, j_q), $q \in \{1, 2, \ldots, k\}$, are the *k* disjoint top-scoring pairs learned from the training data. Given a new test sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_P\}$, each gene pair (i_q, j_q), $q \in \{1, 2, \ldots, k\}$, determines an individual classifier $f_{TSPq}(\mathbf{x})$ according to the decision rule in Equation (2.5). This yields *k* predictions of the class label, *y*, of the new sample. The *k*-TSP classifier employs an unweighted majority voting procedure to obtain

the final prediction of the class label of the test sample $\boldsymbol{x}$; in other words, the k-TSP classifier simply chooses the class receiving the most votes:

$$y = f_{k-TSP}(\boldsymbol{x}) = \underset{c\in\{1,2\}}{argmax} \sum_{q=1}^{k} I(f_{TSPq}(\boldsymbol{x}) = c), \tag{2.6}$$

where

$$I(f_{TSPq}(\boldsymbol{x}) = c) = \begin{cases} 1 & if \quad f_{TSPq}(\boldsymbol{x}) = c \\ 0 & otherwise \end{cases}, c \in \{1, 2\}, q \in \{1, 2, \ldots, k\}. \tag{2.7}$$

***k*-TSP Algorithm**

Input: Training sample $\boldsymbol{S}$ of P genes and N arrays.

Output: k-TSP classifier, i.e., the k disjoint top-scoring gene pairs.

1. Set an upper bound on the number of top scoring pairs to be included in the final k-TSP classifier, K_{max}.
2. (Cross-validation) Repeat m times:
    a. Leave out n arrays from the training set $\boldsymbol{S}$.
    b. Compute the score Δ_{ij} and the rank score δ_{ij} on the current, reduced training set for every pair of genes (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$.
    c. Make an ordered list O of all of the gene pairs (i, j) from largest to smallest using the ordering defined by setting $(i, j) > (i', j')$ whenever either $\Delta_{ij} > \Delta_{i'j'}$ or $\Delta_{ij} = \Delta_{i'j'}$ and $\delta_{ij} > \delta_{i'j'}$.
    d. Initialize Θ at the empty list and perform the following steps for $k = 1, 2, \ldots, K_{max}$:
        i. Add the top pair (i, j) in the list O to Θ.
        ii. Remove every pair from O that involves either i or j.
        iii. If k is odd, compute the error rate for the classifier based on the k pairs in Θ.
3. Select the (odd) value of k whose average classification rate over the m loops in Step 2 is optimal, called it k_{opt}, and compute the k-TSP classifier based on the top k_{opt} scoring pairs as follows:
4. Make an ordered list O of gene pairs as in Steps 2b and 2c using the entire training set.
    a. Initialize Θ at the empty list.
    b. Repeat k_{opt} times:
        i. Add the top pair (i, j) in O to Θ.
        ii. Remove every pair from O that involves either i or j.
5. Return the k-TSP classifier with k_{opt} top gene pairs from Θ in Step 4.

Figure 2.2: Description of the k-TSP algorithm.

2.4 The G-TSP Classifier

There might be more powerful ways than the *k*-TSP classifier to use more than two genes to increase accuracy and stability. Instead of making decision based on relative expression values of two genes, as the TSP classifier does, we could construct two disjoint groups of genes, G_1 and G_2, and make prediction based on the average relative ranks of the genes in the two groups. We name this new classifier as G-TSP classifier. In Chapter 4, we will show that the G-TSP method has unique advantage in building a stable classifier when a random sampling strategy is used to identify more reliable gene signatures.

2.4.1 Learning the G-TSP Classifier

The learning algorithm of the G-TSP classifier is similar to that of *k*-TSP. The algorithm constructs two disjoint groups (sets) of genes, G_1 and G_2, with $|G_1|+|G_2| << P$, where $|G_l|$ denotes the number of genes in group G_l, $l = 1, 2$. The classifier makes a prediction based on the average relative ranks of the genes in the two groups. The initial step of the G-TSP algorithm consists of calculating the score Δ_{ij} for each pair of genes (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$, and forming a list of gene pairs, sorted from the largest to the smallest according to their scores Δ_{ij}, with ties (if there are any) broken by using the rank-score δ_{ij}. Next, the g disjoint top-scoring pairs are selected from the list. The parameter g is predefined or determined by cross-validation as in the *k*-TSP algorithm. The third step is to assign these selected genes to either G_1 or G_2. For each pair *(i, j)* among the g selected pairs, if $N_{ij}^{(2)}/N_2 < N_{ij}^{(1)}/N_1$, then gene i is assigned to G_1 and gene j is assigned to G_2; otherwise, gene j is assigned to G_1 and gene i is assigned to G_2. The learning

algorithm is illustrated in Figure 2.3. Note that there are some other ways to construct the groups G_1 and G_2 and the numbers of genes in G_1 and G_2 don't have to be the same. The algorithm illustrated here is the one to make use of the g disjoint TSPs from the training data.

G-TSP Algorithm
Input: Training sample $\boldsymbol{S}$ of P genes and N arrays.
Output: G-TSP classifier, i.e., two disjoint groups of genes, G_1 and G_2.

1. Set the number of genes to be included in each group, g.
2. Compute the score Δ_{ij} and the rank score δ_{ij} based on the training set for every pair of genes (i, j), $i, j \in \{1, 2, \ldots, P\}$, $i \neq j$.
3. Make an ordered list O of all of the gene pairs (i, j) from largest to smallest using the ordering defined by setting $(i, j) > (i', j')$ whenever either $\Delta_{ij} > \Delta_{i'j'}$ or $\Delta_{ij} = \Delta_{i'j'}$ and $\delta_{ij} > \delta_{i'j'}$.
4. Initialize G_1, and G_2 at the empty lists and repeat the following steps g times:
    a. Select the top pair (i, j) in O:
        i. If $N_{ij}^{(2)}/N_2 < N_{ij}^{(1)}/N_1$, then gene i is added to G_1 and gene j is added to G_2;
        ii. Else, gene j is added to G_1 and gene i is added to G_2.
    b. Remove every pair from O that involves either i or j.
    c.
5. Return the G-TSP classifier with G_1 and G_2.

Figure 2.3: Description of the G-TSP algorithm

2.4.2 Prediction with the G-TSP Classifier

The decision rule of the G-TSP classifier is based on the relative ranks of all the genes in G_1 and G_2. Given a new test sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_P\}$, let $w_i, i \in G = G_1 \cup G_2$, be the relative rank of the gene x_i, $i \in G$, in ascending order within G. The class label, y, of the test sample is predicted by the G-TSP classifier, $f_{G\text{-}TSP}(\boldsymbol{x})$, as

$$y = f_{G\text{-}TSP}(\boldsymbol{x}) = \begin{cases} 1 & if \ \frac{1}{|\mathrm{G}_1|}\sum_{i\in \mathrm{G}_1} w_i < \frac{1}{|\mathrm{G}_2|}\sum_{j\in \mathrm{G}_2} w_j \\ 2 & otherwise \end{cases} . \quad (2.8)$$

Notice the decision rule is based only on the relative ordering among the expression values for the genes in G.

2.5 Classification Results on Individual Data Sets

In this section, we investigate the performance of the TSP classifier family on individual microarray data sets. For this purpose, we compare the performance of TSP-family classifiers with five well-known machine learning methods: C4.5 decision trees (DT), Naive Bayes (NB), *k*-Nearest Neighbor (*k*-NN), Support Vector Machines (SVMs) and Prediction Analysis of Microarrays (PAM). We use the DT, NB, *k*-NN and SVM implemented in the WEKA machine learning package [111] and the PAM Windows version 1.22 program [42] for the experiments.

For DT and NB, we use the default parameters for these techniques. For *k*-NN, the number of neighbors, *k*, is determined using cross-validation on the training set. The SVMs are trained using sequential minimal optimization with a linear kernel.

PAM is a variation of diagonal LDA and one of the most popular statistical methods for analyzing gene expression data. PAM is a statistical technique based on the nearest shrunken centroids approach [42]. We perform cross-validation on the training set to determine the optimal amount of shrinkage (tuning parameter of PAM) for each data set. Other than that, we apply the default parameters of the PAM program.

2.5.1 Microarray Data Sets

We have collected nine publicly available microarray data sets, with sample sizes ranging from 33 to 280 and numbers of genes ranging from 2,000 to 16,063. All of the data sets, which are summarized in Tables 2.1, are related to studies of human cancer, including: colorectal, leukemia, lung, prostate, breast, central nervous system, lymphoma, bladder, melanoma, renal, uterus, pancreas, ovary, and mesothelioma. Further information about the data sets can be obtained from the related publications.

Table 2.1 Microarray data sets

Data set	Platform	No. of genes (P)	No. of samples (N)	
			Class 1	Class 2
Colon [112]	cDNA	2,000	40 (T)	22 (N)
Leukemia [14]	Affymetrix	7,129	25 (AML)	47 (ALL)
CNS [29]	Affymetrix	7,129	25 (C)	9 (D)
DLBCL [30]	Affymetrix	7,129	58 (D)	19 (F)
Lung [113]	Affymetrix	12,533	150 (A)	31 (M)
Prostate1 [31]	Affymetrix	12,600	52 (T)	50 (N)
Prostate2 [37]	Affymetrix	12,625	38 (T)	50 (N)
Prostate3 [36]	Affymetrix	12,626	24 (T)	9 (N)
GCM [27]	Affymetrix	16,063	190 (C)	90 (N)

2.5.2 Estimation of Classification Rate

In order to estimate the classification accuracy for the classifiers, we use standard leave-one-out cross-validation (LOOCV). Hence, for each sample $\boldsymbol{x}_n$ in the training set $\boldsymbol{S}$, we train a classifier based on the remaining N-1 samples in $\boldsymbol{S}$ and use that classifier to predict the class label of $\boldsymbol{x}_n$. The LOOCV estimate of the classification rate is the fraction of the N samples which are correctly classified.

Table 2.2 Classification accuracies of the classifiers estimated by LOOCV

Data set	TSP	*k*-TSP	G-TSP	DT	NB	*k*-NN	SVM	PAM
Leukemia	93.06	95.80	91.70	73.61	100.0	84.72	98.61	97.22
CNS	76.47	88.20	88.20	67.65	82.35	76.47	82.35	82.35
DLBCL	97.40	96.10	94.80	80.52	80.52	84.42	97.40	85.71
Colon	91.94	90.30	91.90	80.65	58.06	74.19	82.26	85.48
Prostate 1	95.10	91.18	94.10	87.25	62.75	76.47	91.18	91.18
Prostate 2	69.32	72.70	72.70	64.77	73.86	69.32	76.14	79.55
Prostate 3	96.97	90.90	93.90	84.85	90.91	87.88	100.0	100.0
Lung	98.34	98.90	98.30	96.13	97.79	98.34	99.45	99.45
GCM	75.36	85.40	85.40	77.86	84.29	82.86	93.21	79.29
Average	88.22	89.94	90.11	79.25	81.17	81.63	91.18	88.91

2.5.3 Results

Table 2.2 summarizes the classification accuracies, estimated by LOOCV, of the eight different classifiers on the nine two-class classification problems. In this case, the estimated classification accuracy is (TP+TN)/*N*, where TP denotes the number of correctly classified class 1 samples, TN denotes the number of correctly classified class 2 samples, and *N* is the total sample size.

Our results show that the eight classifiers can be roughly divided into two groups. Averaged over the nine problems, the top tier classifiers (SVM, G-TSP, *k*-TSP, PAM, and TSP) achieve accuracies in the vicinity of 90% and the second tier classifiers (*k*-NN, NB and DT) in the vicinity of 80%. The best classifier based on the average accuracy for the nine classification problems used in this experiment is SVM (91.18%), followed by G-TSP (90.11%), k-TSP (89.94%), PAM (88.91%) and TSP (88.22%). We do not regard these differences in accuracy as noteworthy and conclude that all five methods perform similarly. However, in terms of efficiency and simplicity, one can argue that the TSP method is superior since it uses a single pair of genes and an elementary decision rule

based solely on expression inversion. These results confirm the findings in [64] that TSP classifier family can generate accurate and interpretable decision rules for classifying microarray data.

2.6 The TSP Method and Microarray Data Integration

As mentioned early in this chapter, an intriguing feature of the rank-based TSP classifier family is that they are robust to quantization effects and are invariant to microarray data pre-processing designed to overcome array-to-array variation, such as most data normalization methods, under the very mild assumption that the normalization method is monotonic in the gene expression values within a profile, and therefore preserves the ordering of gene expression values [64]. This property makes these methods very useful for combining inter-study microarray data without the need to perform data normalization and transformation.

Besides the major issue that expression measurements from different microarray platforms have shown poor correlation and could not be compared directly, a general problem in inter-study microarray data integration is variation among data sets, resulting from biological differences among the samples of different studies, differences in the technical procedures to generate the gene expression values, and random variation. The use of an abstraction of data like ranks reduces this variation at the price of loss of some information [108]. However, for cancer classification problems, we have already demonstrated on large-scale individual microarray data sets in previous section that the TSP classifier family, based only on the ranks of gene expression values, can generate accurate and interpretable decision rules for classifying microarray data.

In the following chapters, we will demonstrate how the TSP classifier family can be used to integrate multiple microarray data sets in order to identify robust gene expression signatures for cancer diagnosis and prognosis.

Chapter 3

Cancer Diagnosis: A New Prostate Cancer Signature Emerges from Microarray Data Integration

3.1 Introduction

Prostate cancer is the most common form of cancer and the second leading cause of cancer death among men in the United States, with an estimated 234,460 new cases and 27,350 deaths in 2006 [114]. In clinical practice, serum prostate specific antigen (PSA) is widely used as a prostate cancer marker. Although PSA screening has led to earlier detection of prostate cancer, it has limited specificity as a cancer marker since increased PSA levels may be present in benign conditions such as prostate enlargement and inflammation [115, 116]

The advent of DNA microarrays provides a powerful tool in cancer research and several studies have used this technology to identify diagnostic gene expression

signatures for prostate cancer that could be used to discriminate the prostate cancer condition from a normal condition [19, 28, 36, 117-119]. The most striking finding when comparing the prostate cancer gene expression signatures from these studies is that the signatures are study-specific, that is, there are few common signature genes, also called markers, among these signatures from different studies [83]. These diverse results make it difficult to identify the most important signature genes, and the corresponding decision rules, for prostate cancer diagnosis. As pointed out early in Chapter 1, differences in results may potentially result from the relatively small sample sizes used in each study.

As more and more microarray data are publicly available, it is becoming increasingly recognized that combining microarray data obtained from different studies may be a useful way to increase sample size. This could in turn lead to the discovery of more robust diagnostic gene expression signatures for prostate cancer. However, several issues arise when attempting to integrate microarray data generated by disparate groups using different array technologies. It has been demonstrated that data from different microarray platforms are variable to the extent that direct integration of expression values from different platforms may be complicated and unreliable [120]. Even when using the same microarray technology, different generations of microarrays have different probe sets and duplicate spots, making direct integration difficult. In addition, attempt to integrate microarray data from different studies must deal with substantial technical and biological sources of variability, which further complicate the problem.

Recently, several studies have proposed to use meta-analysis to combine results from different microarray studies to increase the power of identifying gene expression signatures for prostate cancer [96, 97]. One limitation of these methods is that the small

sample sizes typical of individual studies, coupled with variation due to differences in study protocols, will inevitably degrade the results of meta-analysis. In addition, a recent study [89] has demonstrated that there is only moderate overlap between gene detection on different array platforms.

In this chapter, we propose a novel, simple method to integrate inter-study gene expression values in order to identify diagnostic gene expression signatures for prostate cancer. In this method, there is no need to perform data normalization and transformation before data integration. The method is generally applicable to many other types of microarray data for gene expression signature identification. In this work, we apply the TSP method to microarray data sets integrated from three independent prostate cancer studies to identify a robust, inter-study diagnostic gene expression signature, which is simply a pair of genes. This is feasible since this classification method is invariant to standard procedures for data normalization and transformation. Cross-platform validation shows that the classifier built from the inter-study signature (HPN and STAT6), which declares prostate cancer if the expression value of gene HPN is greater than that of gene STAT6, achieves high accuracy, sensitivity, and specificity on two independent data sets generated from different microarray platforms. In addition, the performance of the inter-study signature is compared with that of the study-specific signatures from the three individual studies on the same cross-platform test data sets. The performance of the inter-study signature is better than that of the study-specific signatures from individual data sets. Furthermore, by reviewing the prostate cancer literature, we note that HPN has been identified as a biomarker for prostate cancer in recent studies [19, 83, 119, 121]. STAT6 is also found to be closely related to prostate cancer [122]. These findings suggest that we

have identified a robust diagnostic gene expression signature for prostate cancer by directly integrating inter-study microarray data. Upon further validation on additional independent data sets, the signature could be used to develop a genomic-based, more accurate diagnostic test for prostate cancer

3.2 Methods

3.2.1 Gene Expression Data

Five prostate microarray data sets are included in this chapter. Each data set has been downloaded from publicly available gene expression repositories or supporting web sites [36-38, 117, 118, 123, 124]. The three data sets used as training samples are generated from the same Affymetrix HG-U95A platform by different labs and the remaining two data sets used as test samples are from cross-platform independent studies (Table 3.1). In this study, we focus on identifying a diagnostic gene expression signature which can distinguish primary prostate cancer from normal samples. Therefore, metastatic prostate cancer samples are not included in the study. The summaries of the training and test data sets are provided in Table 3.1. Here, the names of the first authors of individual studies are used as the names of the data sets. Details about each data set have been described in the corresponding literature.

Table 3.1 Training and test data sets

Data Set		Platform	No. of genes	No. of samples	
				Normal	Cancer
Training Set	Singh [117]	Affy. (HG-U95A)	12600	50	52
	Stuart [37]	Affy. (HG-U95A)	12625	50	38
	Welsh [36]	Affy. (HG-U95A)	12626	9	24
Test Set	LaTulippe[124]	Affy. (HG-U95A)	12626	3	23
	Lapointe[a][38]	Spotted cDNA	44160/43008	41	62

a. 22 samples (9 N / 13 C) have 44160 probes and 81 samples (32 N / 49 C) have 43008 probes.

3.2.2 TSP Classifier

The detailed description of the TSP classifier can be found in Section 2.2. An important feature of the TSP method is that it is invariant to monotonic transformations of the expression data, such as most data normalization methods. This property makes it very useful for integrating inter-study microarray data without the need to perform data normalization and transformation. The TSP classifier is applied to integrated microarray data sets to identify diagnostic gene expression signatures for prostate cancer.

3.2.3 Data Integration

By applying the TSP method, no data transformation and normalization is required before integration. Among the three individual training data sets used in this study, there are 12600 common probe sets (genes). We directly merge individual data sets, using the 12600 common probe sets, to form integrated data sets of increased sample sizes.

3.2.4 Error Estimation

In estimating the error rate of a classifier based on cross-validation, gene pair selection and corresponding classification are performed within the cross-validation loop. With n samples and leave-one-out cross-validation, this means choosing n separate top-scoring

pairs, one for each sample left out during training, then classifying the left-out sample. In particular, both the actual top score, as well as the gene pair which achieves it, may vary with the left-out sample. The estimated prediction rate is then 1- *e/n* where *e* is the number of errors observed in the cross-validation. This is the way the prediction rates reported in Table 3.2, as well as the cross-validation results given in Table 3.5, were calculated. However, in order to associate a single TSP with each training set (as in Table 3.2), and a corresponding decision rule for evaluation on an independent test set (as in Tables 3.3 and 3.4), the 'final TSP' is computed using the entire training set. As a result, all error estimates are unbiased.

3.2.5 Stability Analysis

We have designed an experiment to analyze the stability of a TSP in response to slight perturbations of the training set resulting from changing its size. This is accomplished by randomly removing a small percentage (*K*%) of samples from the original training set and generating a TSP from the reduced training set. After repeating the experiment a large number of times with different values of *K* (e.g. 1, 2, …), we calculate the appearance frequency of the TSP among all TSPs generated at each sample size, for instance, 99% appearance frequency at sample size 220. If this frequency remains very high (e.g. above 95%) when the sample sizes are slightly (e.g. 5%) different from the original training set size, we conclude that the TSP is stable for the original training set.

3.3 Results

3.3.1 Inter-Study Gene Expression Signature Identification

To investigate whether a robust diagnostic gene expression signature that distinguishes primary prostate cancer from normal samples can be identified, three microarray data sets from different studies have been collected and the TSP method has been applied to analyze the individual and integrated data sets. To avoid loss of potential signature genes, we choose to analyze inter-study microarray data obtained using the same platform (the HG-U95Av2 array; see Table 3.1). Starting from individual data sets, we gradually increase the sample size by sequentially merging two and then three data sets (see 'Data Integration' in Section 3.2.3). Applying the TSP method to the training sets, individual and integrated, generates a TSP for each of the training sets. The TSPs are listed in Table 3.2 along with scores and prediction rates. The score refers to the absolute value of the difference between two probabilities estimated from the training data - the probability that the expression of the first gene in the pair exceeds the expression of the second gene in the pair for the cancer class and the same probability for the normal class (see 'TSP Classifier' in Section 2.2). Prediction rate is estimated by leave-one-out cross-validation on each training set. In Table 3.2, underscores are used to join names of individual data sets to denote an integrated data set. For example, 'Welsh_Stuart' is the name of the integrated data resulting from the merging of the Welsh and Stuart data sets.

Results show that when sample sizes are small, different data sets generate distinct TSPs. As the sample size reaches a certain level (between 135 and 190) and continues to increase, the pair (HPN, STAT6) is consistently selected as the TSP. This pair of genes is

the inter-study diagnostic gene expression signature for prostate cancer which we have identified from data integration.

Table 3.2 TSPs from training data sets with increased sample sizes

Training data set	**Sample size**	**Probe set ID**	**Gene symbol**	**Score**	**Prediction rate (%)[b]**
Welsh	33	39608_at 32526_at	SIM2 JAM3	1.00	97.0
Stuart	88	41732_at 456_at	CTNNB1 SMARCD3	0.74	69.3
Singh	102	40282_s_at 2035_s_at	DF ENO1	0.90	95.1
Welsh_Stuart[a]	121	31971_at 34213_at	TP73L KIBRA	0.79	77.7
Welsh_Singh	135	37639_at 32198_at	HPN COMMD4	0.88	83.7
Stuart_Singh	190	37639_at 41222_at	**HPN** **STAT6**	0.75	86.8
Welsh_Stuart_Singh	223	37639_at 41222_at	**HPN** **STAT6**	0.78	88.8

a. Welsh_Stuart is the name of the integrated data set of Welsh and Stuart data sets. Other integrated data sets use similar names.
b. Prediction rates were estimated by leave-one-out cross-validation on the training sets.

3.3.2 Stability Analysis of the Inter-Study Signature

We subsequently performed an analysis of the stability of the inter-study gene expression signature (HPN, STAT6) (see 'Stability Analysis' in Section 3.2.5), where 'stability' refers to the sensitivity of the selection procedure to perturbations of the training set. To do this, small numbers of samples are randomly removed from the integrated data set Welsh_Stuart_Singh of size 223. At each sample size, we repeat the experiment 100 times and calculate the appearance frequency of the inter-study signature. The results of the analysis are shown in Figure 3.1. When 1 – 3 % of the samples are removed, the appearance frequency of the inter-study signature is 100%. The inter-study signature appears with very high frequency when less than 10% of the samples are

randomly removed from the original training set. From this analysis, we have shown that the inter-study signature is stable for the integrated training set. We carry out the same analysis on the study-specific signature selected from the Stuart data set (size 88 samples). Results are shown in Figure 3.1. When one sample (~ 1%) is removed from the training set, the appearance frequency of the Stuart signature declines by ~ 30%. With two or more samples removed from the training set, the appearance frequency declines further. Therefore, we can conclude that the Stuart signature is not stable for the Stuart training set.

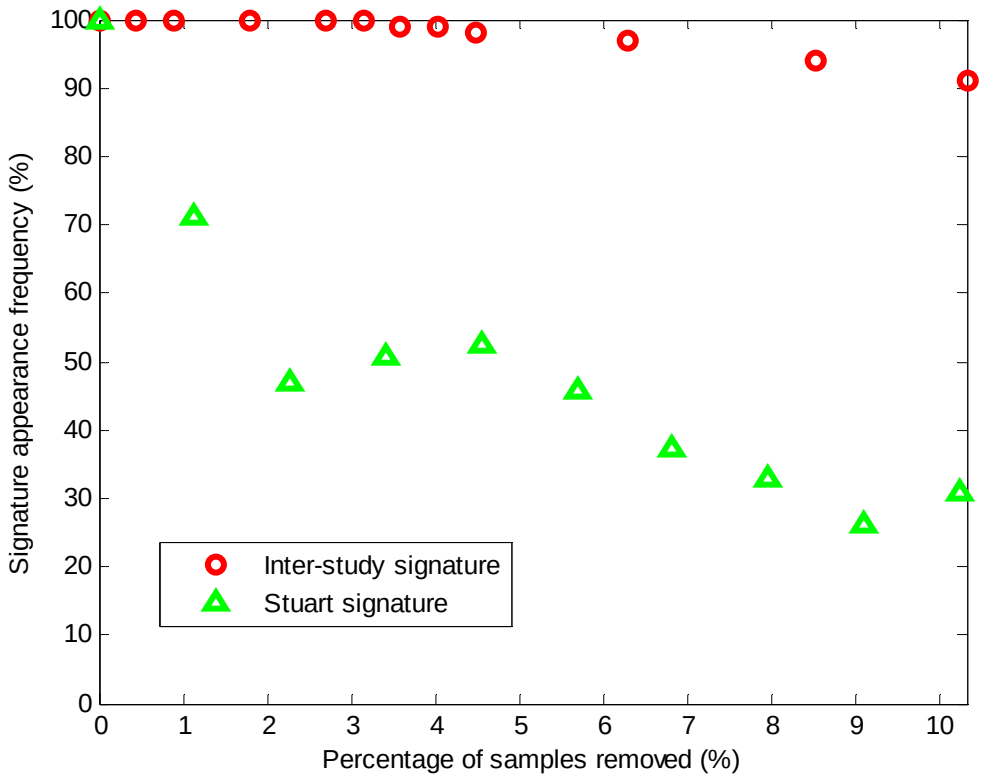


Figure 3.1: Results of the stability analysis of the inter-study signature and the Stuart signature.

3.3.3 Validation on Cross-Platform Independent Data Sets

In order to further validate the reliability and robustness of the inter-study signature, the TSP classifier, called inter-study TSP classifier, built from the inter-study signature is tested on independent cross-platform microarray data (Table 3.3). The decision rule for the inter-study TSP classifier is that if the expression value of HPN is greater than that of STAT6, a test sample is classified as prostate cancer; otherwise it is classified as normal. The performance of the classifier is measured by examining how well the classifier predicts the normal and cancer samples in the test sets. Accuracy is defined as the ratio of the number of correctly predicted samples to the total number of samples. Sensitivity (respectively, specificity) is the ratio of the number of correctly predicted cancer (respectively, normal) samples to the total number of cancer (respectively, normal) samples. The inter-study TSP classifier consistently achieves high accuracy, sensitivity and specificity on the test sets across different platforms (Table 3.3). The overall accuracy, sensitivity, and specificity of the inter-study TSP classifier are 93.8%, 91.7%, and 97.7%, respectively.

As a comparison, the TSP classifiers built from the study-specific gene expression signatures identified from individual microarray data sets are tested on the same test sets. In Table 3.4 and 3.5, we use the name of each individual training data set to represent the TSP classifier built from the study-specific signature identified from that individual data set and ‘inter-study’ to represent the inter-study TSP classifier built from the inter-study signature identified from the integrated data set Welsh_Stuart_Singh. For example, ‘Welsh’ represents the TSP classifier built from the Welsh signature identified from the Welsh data set. In each case, the decision rule is based on comparing the expression

values in the pair of the corresponding signature genes. The results of the comparison are summarized in Table 3.4. The inter-study TSP classifier outperforms all of the study-specific TSP classifiers on all test sets. The results suggest that due to the limited sample size and/or artifacts of the individual data sets, the gene expression signatures from individual data sets provide less reliable predictors of prostate cancer than the inter-study signature obtained from the integrated data. Although the inter-study signature is generated from integrated data obtained using a single microarray platform (Affymetrix HG-U95A), it can also be used for accurate classification of a novel data set obtained using spotted cDNA microarrays.

Table 3.3 Classification performance of the inter-study TSP classifier

Test data set	**No. of sample**		**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
	Normal	Cancer			
LaTulippe	3	23	96.2	95.7	100
Lapointe[a,b]	41	61	93.1	90.2	97.6
Overall	44	84	93.8	91.7	97.7

a. The corresponding clone IDs for the gene pair (HPN, STAT6) is (IMAGE:208413, IMAGE:85541).
b. One of the cancer samples has missing value for HPN and is removed from the test set.

Table 3.4 Comparison of the inter-study signature with study-specific signatures

Test data set	**TSP**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
LaTulippe	Welsh	69.2	69.6	66.7
	Stuart	84.5	82.6	100
	Singh	88.5	87.0	100
	Inter-study	96.2	95.7	100
Lapointe	Welsh	70.9	95.2	34.1
	Stuart	43.6	6.7	97.6
	Singh	43.7	6.4	100
	Inter-study	93.1	90.2	97.6

3.3.4 Consistency of Cross-Validation and Independent Test Results

In Table 3.2, we note that the prediction rates of the Welsh and Singh classifiers (again, estimated by cross-validation) is higher than that of the inter-study TSP classifier. However, when tested on independent test sets, the performance of the inter-study TSP classifier is considerably better than that of the study-specific TSP classifiers from individual studies (Table 3.4). Table 3.5 summarizes the results of cross-validation and independent test of the inter-study TSP classifier and the study-specific TSP classifiers from the three individual studies. The independent test results reported here are the overall results on the two cross-platform independent test sets with 128 total samples. For the three study-specific TSP classifiers from individual studies, cross-validation results are inconsistent with independent test results. This implies that the cross-validated estimates of error of individual studies have high variation and that the corresponding TSP classifiers are somewhat study-specific and not as reliable as the inter-study TSP classifier in classifying prostate samples generated from other independent studies. On the other hand, the inter-study TSP classifier generates consistent results between cross-validation and independent test. By integrating inter-study microarray data, the study-specific effect is reduced and more stable features of the cancer are captured by the inter-study gene expression signature.

Table 3.5 Comparison between cross-validation and independent test results

TSP	**Cross-Validation Results (%)**			**Independent Test Results (%)**		
	Accuracy	SN[a]	SP[b]	Accuracy	SN	SP
Welsh	97.0	95.8	100	70.5	88.2	36.4
Stuart	69.3	81.6	60.0	52.0	27.7	97.7
Singh	95.1	96.2	94.0	52.7	28.2	100
Inter-study	88.8	90.4	87.2	93.8	91.7	97.7

a. 'SN' denotes sensitivity. b. 'SP' denotes specificity.

3.4 Discussion

The increasing availability of gene expression microarray data has been calling for methods to effectively integrate multiple, independently generated data sets targeting the same biological question. We present a novel, simple method of integrating different microarray data sets to identify robust inter-study gene expression signatures and illustrate the method using prostate cancer data sets. By applying the TSP method, we have successfully identified a robust diagnostic gene expression signature for prostate cancer from direct integration of inter-study microarray data. The diagnostic signature is simply a top-scoring gene pairs, HPN and STAT6. The TSP classifier built from the signature, which simply compares relative expression values of HPN and STAT6, achieves high accuracy (93.8%), sensitivity (91.7%), and specificity (97.7%) on independent cross-platform microarray data sets.

Integration of microarray data across platforms can be achieved by using the subset of genes that are common to all platforms by gene mapping based on UniGene database. However, the large number of genes which are not in the common set may include potential signature genes. Therefore, in this study, we use inter-study data from the same platform (Affymetrix HG-U95A) to identify a robust prostate cancer gene expression signature. We choose Affymetrix oligonucleotide microarrays because it has been shown that Affymetrix oligonucleotide arrays are much more reproducible than spotted cDNA arrays [125]. The reason that the Affymetrix HG-U95A arrays are chosen is that there is a large amount of prostate microarray data generated on this platform among published microarray data sets. This makes it possible to increase the sample size of integrated data to a level necessary to identify a robust gene signature.

A unique pair of genes, HPN and STAT6, is consistently selected as the inter-study gene expression signature with the increase of sample sizes. However, when study-specific signatures are computed for individual microarray data sets, each of the data set generates a different signature and none of the study-specific signatures are the same as the inter-study signature. This observation implies that the dependence of the inter-study signature on the individual data sets can be significantly diminished, and the information provided about prostate cancer can be significantly increased by data integration. An advantage of inter-study microarray data integration is then that it increases the statistical power to capture consistent features which might be masked by the small sample size and experimental artifacts in an individual data set. In this sense, the inter-study signature is more reliable and more robust to variations in individual data sets.

To provide a comparison with a common approach to classifying gene expression profiles, we applied the software for a popular variation of diagonal linear discriminant analysis called Prediction Analysis of Microarrays (PAM) [42] which automatically selects an optimal number of genes ranked by a modification of the *t*-statistic. (The normalization of the standard deviation involved in 'shrunken centroids' is first performed on the three data sets separately.) We ran PAM on the integrated Welsh_Stuart_Singh data set. This resulted in a classifier based on 135 genes (with HPN at top of the list) and a classification accuracy of 86.1% as estimated by cross-validation (similar to the 88.8% accuracy of the inter-study TSP classifier; see Table 3.2). Since one of the independent test sets, LaTulippe, is generated from the same microarray platform, i.e. Affymetrix HG-U95A, we could also test the classifier learned from PAM on that data set. The accuracy of PAM and the inter-study TSP classifier are the same on

LaTulippe. However, it was not possible to evaluate PAM on the Lapointe test set, which is generated from a spotted cDNA microarray, because some of the 135 genes are not present in that data set. Moreover, even were all these genes present in the cDNA data set, the issue would remain of how to normalize the cDNA data to make them comparable to the Affymetrix HG-U95A data. This comparison demonstrates the advantage of our method in integration of inter-study microarray data.

We did not include all publicly available prostate data sets in the test sets. For some older platforms, such as the Affymetrix Hu35kA array, there is no probe set corresponding to either of the signature genes HPN or STAT6. In some cDNA microarray data sets, one of the signature genes is missing. Even with these limitations, we still obtain a reasonable number of independent test sets (total sample size 128) on differing platforms.

An interesting and important finding of the study is that although the inter-study gene expression signature is generated from integrating data obtained using a single type of microarray (the Affymetrix HG-U95A oligonucleotide microarray), it can be used to classify microarray data obtained using a different microarray technology (cDNA microarrays) with high accuracy**.** This has confirmed an observation that the rank orders of differentially expressed genes were comparable across array platforms while the fold changes showed poor correlation in a recent study [126]. It also suggests that as long as the sample size reaches a certain 'statistically significant' level, valid conclusions can be drawn from single-platform inter-study microarray data. Upon further confirmation from other studies, this finding might provide an alternative approach for microarray data analysis. Because cross-platform data integration is much more complex than single-

platform data integration, this finding, as reported in our study, will greatly facilitate microarray data integration.

One of the inter-study signature genes, HPN, has been identified as a marker of prostate cancer in recent studies [19, 83, 119, 121]. HPN encodes hepsin, a cell surface transmembrane serine protease which plays an essential role in cell growth and maintenance of cell morphology. Using both cDNA and oligonucleotide microarray technologies, hepsin was shown to be significantly over-expressed in prostate cancer samples versus normal samples, and it has been identified as a potential biomarker for screening prostate cancer [19, 118, 119, 127]. mRNA over-expression has also been validated using RT-PCR [119] and protein over-expression has been verified using tissue microarrays [19]. Magee et al. [118] also confirmed the over-expression of hepsin in prostate tumor by using *in situ* hybridization technique on an independent panel of prostate specimens. Furthermore, the expression of hepsin has been shown to have positive correlation with prostate cancer staging [127] and to promote prostate cancer progression and metastasis [121]. Thus, hepsin may be used as a diagnostic as well as prognostic marker for prostate cancer.

STAT6 encodes the signal transducer and activation of transcription 6 (Stat6), a member of STAT transcription factors located in the cytoplasm that is involved in the Jak-Stat signaling pathway. The Jak-Stat pathway is an important signaling pathway in cellular development/survival [128, 129]. It is activated by a small number of cytokines (e.g. interleukin-4) and plays a distinct role in the development of T-cells (e.g. T helper cell type 2) and in IFNγ signaling. The expression of STAT6 has been shown to be down-regulated in gastric cancer [130]. From our study, we observe that STAT6 is slightly

down-regulated in prostate cancer compared to normal samples. This down-regulation of STAT6 is necessary for the cancer cell to escape from the tumor immunosurveillance mechanism, where the tumor protects itself from being killed by the natural killer T cells [131]. It has been shown that T helper cell type 2 cytokines down-regulate anti-tumor immunity [132]. At the protein expression level, Ni et al. [122] showed that Stat6 was selectively activated in prostate cancer using Western blot analysis.

One of the goals of cancer gene expression signature identification is to translate inter-study microarray data analysis into clinically useful cancer markers. Prostate specific antigen (PSA), as a prostate tumor marker currently used in clinical practice, has, as its major limitation, low specificity. When normal serum PSA levels are defined as 4.0ng/mL or less, PSA test has a sensitivity of about 67.5% to 80% and a specificity of about 60% to 70% [133, 134]. Our study has discovered a robust gene expression signature that is sufficient to distinguish prostate cancer from normal in both training and test sets. The high sensitivity (91.7%) and specificity (97.7%) of the inter-study gene expression signature achieved on a large number (128) of independent test samples are encouraging, suggesting its potential clinical applicability. A possible application would be to make a simple diagnostic chip using the signature genes. A test sample will be predicted as cancer or normal simply by comparing the expression values of the two signature genes. Clearly, validation on a larger set of independent data will be required before the idea can be translated into clinical practice. Nevertheless, this study provides evidence for promising potential prostate cancer markers to improve the diagnostic accuracy of prostate cancer.

In conclusion, this work has not only established a new model for the discovery of robust gene expression signatures from accumulated microarray data, but also demonstrated how the great wealth of microarray data can be exploited to increase the power of statistical analyses.

Chapter 4

Essential Transcriptional Features of Cancer: A Common Cancer Signature from Large-Scale Integration of Microarray Data

4.1 Introduction

During the past century, the presence of cancer in tissues has been diagnosed on the basis of histopathology [1]. The major limitation of this approach is that it cannot achieve high accuracy of prediction in clinical practice. Therefore, there has been a persistent need to identify robust cancer signatures which could complement conventional histopathologic evaluation to increase the accuracy of cancer detection [2]. More recently, DNA microarrays have been developed as a means to simultaneously measure the transcript abundance (gene expression level) of mRNA for thousands of genes. This technology provides a potentially powerful tool for identifying molecular signatures capable of accurately detecting the presence of cancer.

Many studies have used DNA microarrays to identify cancer type-specific gene expression signatures which can discriminate certain types of cancer from normal tissues [14, 18, 19, 28, 44, 64, 81, 110, 119, 135-138]. The diversity of these signatures makes it difficult to distinguish the genes that play a crucial role in oncogenic processes from those that are spuriously differentially expressed and therefore irrelevant to the oncogenic processes. Since all cancer cells share two common characteristics, uncontrollable growth and local tissue invasion or metastasis, it is of high importance to identify a universal cancer type-independent signature to better understand cancer pathogenesis and ultimately to improve therapeutics. After such a signature is identified, it could be used as a component of genomic-based clinical diagnostic tools for cancer patients to determine the presence of cancer cells in tissues.

Recently, several studies used meta-analysis methods to identify genes differentially expressed across multiple cancer types [101, 139, 140]. In the study of Rhode et al. [101], 21 published cancer microarray data sets, spanning 12 distinct human cancer types, were collected and analyzed in an effort to identify a cancer type-independent transcriptional signature of neoplastic transformation. A statistical meta-analysis method, termed comparative meta-profiling, was proposed to compare and assess the intersection of many cancer type-dependent signatures, the goal being to identify a common cancer meta-signature. A set of 67 genes that are universally activated in most cancer types, relative to corresponding normal tissues, was characterized as a meta-signature of neoplastic transformation.

One limitation of meta-analysis of microarrays is that the small sample sizes typical of individual studies, coupled with variation due to differences in study protocols, inevitably

degrade the results of meta-analysis. An additional and major limitation of the comparative meta-profiling method is that those genes which are common to the various array platforms used in these studies are highly overrepresented in the identified meta-signature. The comparative meta-profiling method works as follows. First, an overexpression direction (e.g. cancer > normal) and significant threshold are set to define differential gene expression signatures from a set of cancer vs. normal studies (one signature per study). Then genes are sorted by the number of signatures in which they are present. Finally, a meta-signature is defined as those genes appearing in a given number of signatures [101], where the cutoff is determined by a random simulation. This way of defining a meta-signature by gene enrichment in signatures implies that many potentially informative genes which are not common to the various array platforms used in these studies may be overlooked due to the intrinsic properties of this method. As a specific example, the relationship between the numbers of total genes on two major Affymetrix microarray platforms used in the study of Rhodes et al. and the corresponding numbers of genes included in the reported common cancer meta-signature is shown in the Venn diagram of Figure 4.1. Among the 67 meta-signature genes, 59 genes are on one or both of these two microarray platforms and the other eight genes come from other microarray platforms. Almost all of the 59 meta-signature genes come from the set of 5127 genes which are common across the two microarray platforms employed in this study.

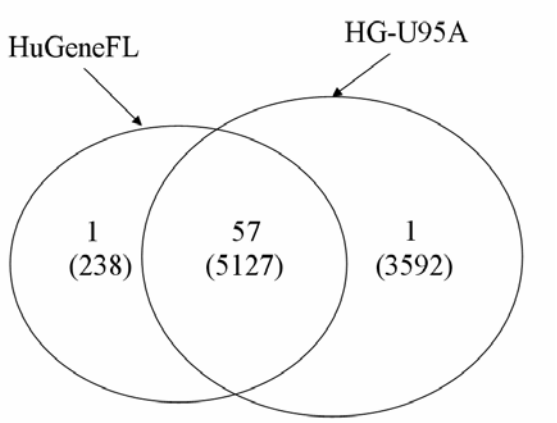


Figure 4.1: Relationship between the numbers of genes on two microarray platforms and the corresponding numbers of genes in the meta-signature of neoplastic transformation [101]. There are 5127 genes common to the two platforms, 238 only on HuGeneFL and 3592 only on HG-U95A. The numbers without parentheses are the corresponding numbers of genes in the meta-signature.

In the previous chapter, we have successfully demonstrated how the TSP method can be used to integrate inter-study microarray data to identify robust gene expression signatures for cancer diagnosis. In this chapter, we combine the G-TSP method, another member of the TSP classifier family, and a repeated random sampling strategy to identify a cancer type-independent signature by integrating large-scale microarray data from different cancer studies across almost all major human cancer types. This approach overcomes the limitations of previous meta-analyses by integrating large-scale microarray data generated using the same microarray platform, resulting in one signature per platform (rather than per study) and effectively increasing sample size. Integrating microarray data from the same platform can guarantee that all the genes on the platform will be included in the analysis, therefore avoiding losing any potential signature genes. By combining the G-TSP method and a repeated random sampling strategy, a common cancer signature, which consists of 46 distinct genes, is identified from the integrated microarray data. The G-TSP classifier, which discriminates cancer from normal samples,

is built from the signature and validated on both the training data and independent cancer microarray data. Upon further validation on large-scale independent data, the signature may be used to develop a novel cancer diagnostic tool and provide new insights regarding the mechanism of cancer.

4.2 Methods

4.2.1 Data Collection

Microarray data sets were obtained from public gene expression data repositories, including Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo), Oncomine (http://www.oncomine.org), and supporting web sites identified from the published literature. In particular, we focused on analysis of human cancer gene expression data obtained using the Affymetrix HuGeneFL, HG-U95A and HG-U133A microarray platforms. Of these collected data, 26 individual data sets generated from HuGeneFL arrays (Table 4.1) and HG-U95A arrays (Table 4.2) were used as training data and six individual data sets from HG-U133A arrays (Table 4.3) were used as independent test data. In total, 1593 microarray experiments from 32 independent studies across almost all human cancer types were used in this analysis. Studies are named using the same convention as in [101]: FirstAuthor_Tissue (e.g. Beer_Lung).

Table 4.1 Microarray data from Affymetrix HuGeneFL arrays

Study	Normal samples	Size	Cancer samples	Size
Beer_Lung [79]	Normal Lung	10	Lung Adenoarcinoma	86
Dyrskjot_Bladder [39]	Normal Bladder	4	Bladder Cancer	40
Hippo_Gastric [141]	Normal Gastric Tissues	8	Gastric Cancer	22
Hsiao_Normal [142]	Normal Tisues	59		
Lancaster_Ovarian [143]	Normal Ovary	3	Ovarian Adenocarcinoma	31
Logsdon_Pancreas [144]	Normal Pancreas	5	Pancreatic Adenocarcinoma	10
Pomeroy_Brain [145]	Normal Cerebellum	4	Atypical Teratoid/Rhabdoid Tumors	10
			Primitive Neuroectodermal Tumors	8
			Malignant Gliomas	10
			Medulloblastoma	10
Quade_Myometrium [146]	Normal Myometrium	4	Leiomyosarcoma	14
Ramaswamy_Multi [27]	Normal Prostate	9	Prostate Cancer	10
	Normal Uterus	6	Uterine Cancer	10
	Normal Whole Brain	8	Glioblastoma	20
	Normal Breast	5	Breast Adenocarcinoma	11
	Normal Lung	7	Lung Adenoarcinoma	11
	Nromal Colon	11	Colorectal	11
	Normal Germinal Center	6	Lymphoma	22
	Normal Bladder	7	Bladder Cancer	11
			Melanoma	10
	Peripheral Blood	5	Leukemia	30
	Normal Kidney	12	Renal Cell Carcinoma	11
	Normal Pancreas	10	Pancreatic Adenocarcinoma	11
	Normal Ovary	4	Ovarian Carcinoma	11
			Mesothelioma	11
Rickman_Brain [147]	Normal Temporal Lobe	6	Glioma	45
Welsh_Ovarian [15]	Normal Ovary	4	Ovarian Carcinoma	22
Zhan_Myeloma [148]	Normal Plasma Cell-Bone Marrow	30	Multiple Myeloma	74
Total	Normal Tissues	227	Cancer Tissues	572

Table 4.2 Microarray data from Affymetrix HG-U95A arrays

Study	Normal samples	Size	Cancer samples	Size
Bhattacharjee_Lung [25]	Normal Lung	17	Small Cell Lung Carcinoma	6
			Lung Carcinoid	20
			Squamous Cell Lung Carcinoma	21
Cromer_Head-Neck [149]	Normal Uvula	4	Head-Neck Squamous Cell Carcinoma	34
Dehan_Lung [150]	Normal Lung	9	Lung Adenocarcinoma	7
			Lung Squamous Cell Carcinoma	17
			Lung Adenosquamous	1
Frierson_Salivary [151]	Normal Salivary Gland	6	Salivary Carcinoma	16
Giordano_Adrenal [152]	Normal Adrenal Cortex	3	Adrenocortical Carcinoma	11
Gutmann_Brain [153]	Normal White Matter	3	Pilocytic Astrocytoma	8
Huang_Thyroid [154]	Normal Thyroid	8	Thyroid Carcinoma	8
Shai_Brain [155]	White Matter	7	Glioblastoma Multiforme	35
Stearman_Lung [156]	Normal Lung	19	Lung Tumor	20
Su_Multi [157]	Normal Tissues	63	Tumor Tissues	18
Su_Tumors [28]			Prostate Cancer	24
			Bladder/Ureter	8
			Breast	21
			Colorectal	21
			Gastroesophagus	11
			Kidney	10
			Liver	6
			Ovary	23
			Pancreas	6
			Lung Adenocarcinoma	12
			Lung Squamous Cell Carcinoma	12
Welle_Normal [158]	Normal Muscle	12		
Yanai_Normal [159]	Normal Tissues	24		
Yu_Prostate [160]	Normal Prostae	16	Primary Prosate Carcinoma	35
Total	Normal Tissues	191	Cancer Tissues	411

Table 4.3 Microarray data from Affymetrix HG-U133A arrays

Study	Normal samples	Size	Cancer samples	Size
Gordon_Lung [161]	Normal Lung	4	Malignant Pleural Mesothelioma	40
	Normal Pleura	5		
Hoffman_Myometrium [162]	Normal Myometrium	5	Uterine Leiomyomas	5
Lenburg_Kidney [163]	Normal Kidney Tissue	5	Renal Cell Carcinoma	12
Talantov_Skin [164]	Normal Skin	7	Melanoma	45
Wachi_Lung [165]	Normal Lung	5	Squamous Lung Cancer	5
Yoon_Soft_Tissue [166]	Normal Soft Tissue	15	Soft Tissue Sarcoma	39
Total	Normal Tissues	46	Cancer Tissues	146

4.2.2 G-TSP Classifier

The detailed description of the G-TSP classifier can be found in Section 2.4. The development of the G-TSP method is motivated by following considerations. It is known that many genes are involved in the oncogenic processes; therefore, in order to better understand cancer pathogenesis, we need to identify a common cancer signature which consists of more than just a few genes. The TSP method previously used only identifies a pair of signature genes, thus is inappropriate in this study. Another motivation for developing the G-TSP method is related to the observation that, in some cases, one gene may pair with different genes to form a TSP when the training data is perturbed by adding or deleting a few samples. This may imply that the gene consistently appears in the TSP may be closely correlated to cancer while the other genes occasionally paired with it might be irrelevant to cancer. We want to keep those genes which may play a crucial role in oncogenic processes in the common cancer signature and eliminate those genes which may be irrelevant to cancer. When combined with repeated random

sampling, the G-TSP method provides the flexibility to keep one gene of a TSP in a signature while excluding the other one from it. The *k*-TSP method doesn't provide the flexibility to keep one gene of a TSP because it only recruits pairs of genes to a signature.

4.2.3 Data Integration

Since rank is invariant to monotonic data transformations, by applying the rank-based G-TSP method, which performs on-chip comparisons within each microarray, no data normalization and transformations are required before integration. For microarray data generated from the same platform, we directly merge individual data sets using the common genes across all the data sets to form an integrated data set of increased sample size. To merge data from different generations of the same array technology, we first select the common probe sets across the platforms and then merge data using the common probe sets. In that case, a large number of genes, which are not in the common set and may include potential signature genes, will be excluded from analysis. In this study, we focus on integrating microarray data generated using same platform to avoid losing any potential signature genes. Specifically, we integrate microarray data generated from both Affymetrix HuGeneFL and HG-U95A microarray platforms to form two corresponding integrated data sets with large (> 500) sample sizes. These two integrated data sets are used as training data sets to identify a common cancer signature.

4.2.4 Repeated Random Sampling and Signature Gene Selection

A recent study by Michiels et al. has shown that molecular signatures strongly depend on the selection of patient samples in the training sets and they advocate the use of repeated random sampling for validation [167]. Motivated by their work, we combine the

G-TSP algorithm with a repeated random sampling strategy in order to obtain a more robust and reliable common cancer signature. We randomly select *T%* of the total samples from an integrated training set, with *T* chosen close to 100, specifically $T = 90$, in order to reasonably represent the original training samples. We then apply the G-TSP algorithm to the selected subset of the original training set to construct two groups, G_1 and G_2, of genes, with a predefined value of *g*, the number of genes in each group. After repeating this experiment a large number of times, we calculate the appearance frequency of each gene in G_l, $l = 1, 2$. A frequency threshold *F* (default, $F = 80\%$) is set to select those genes whose appearance frequency exceeds the threshold in G_l, $l = 1, 2$. The final common cancer signature consists of all the genes picked from G_1 and G_2 for each of the two integrated training data sets, HuGeneFL and HG-U95A.

4.2.5 Class Prediction

To assess the classification accuracy of the common cancer signature, we build a G-TSP classifier based on all the signature genes. The signature genes picked up from G_1's (respectively, G_2's) of the two integrated training sets by repeated random sampling form the group G_1 (respectively, G_2) of the final G-TSP classifier. Note that the way to construct the two groups, G_1 and G_2, is slight different from that in Section 2.4.1. The classifier makes prediction according to the decision rule in Equation (2.8), that is, it votes for class 1 (i.e. normal) if the average relative rank of the genes in G_1 is less than that of the genes in G_2; otherwise, it votes for class 2 (i.e. cancer). We not only use the classifier to predict samples from data sets involved in identifying the signature, but also validate it on independent data sets. The Fisher's exact test is used to assess the significance of the classification accuracy. The classification accuracy, as well as its

statistical significance, is reported for all the individual data sets, training and test, in this study.

4.3 Results

4.3.1 Common Cancer Signature

We directly merge 12 (respectively, 14) cancer/normal microarray data sets generated from Affymetrix HuGeneFL (respectively, HG-U95A) (Table 4.1 & 4.2), using the common 7069 (respectively, 12532) probe sets among all these data sets to form an integrated training data set with 799 (respectively, 602) samples. These data sets span 21 tissue types, including lung, breast, bladder, ovarian, pancreas, brain, prostate, uterus, colon, blood, kidney, uvula, salivary gland, thyroid gland, liver, skin, gastric tissue, myometrium, bone marrow, adrenal cortex and gastroesophagus. For each of the two integrated data sets, we randomly select 90% of the total samples from the integrated data set and then apply the G-TSP algorithm (Figure 2.3) to the selected data to construct two groups, G_1 and G_2, of genes. The parameter g for the G-TSP algorithm is set to be $g = 20$. After the experiment is repeated 1000 times, the appearance frequency for each gene which is present in any of the 1000 G_1's (respectively, G_2's) is calculated. For the default frequency threshold $F = 80\%$, the appearance frequency of 24 genes (13 in G_1 and 11 in G_2) from the HuGeneFL integrated data set and 25 genes (12 in G_1 and 13 in G_2) from the HG-U95A integrated data set exceeds F (Table 4.4). There are three genes (CLEC3B, COX7A1 and KIAA0101) which are selected from both integrated data sets. Therefore, a common cancer signature, which consists of the 46 genes (24 in G_1 and 22 in G_2)

obtained from the two integrated data sets, is identified from the integrated microarray data.

Table 4.4 Common cancer signature genes

Microarray platform	G1		G2	
	Gene symbol	Probe set ID	Gene symbol	Probe set ID
HuGeneFL	BOP1	D50914_at	**COX7A1**	M83186_at
	PON2	L48513_at	CXCL12	U19495_s_at
	NME1[a]	X17620_at	ALDH1A1	M31994_at
	CKS2[a]	X54942_at	SELP	M25322_at
	CCT3	X74801_at	CD36	Z32765_at
	KIAA0101[a]	D14657_at	CSRP1	M76378_at
	FOXM1	U74612_at	C9orf61	L27479_at
	MAP3K11	L32976_at	MYH11	AF001548_rna1_at
	RAB13	X75593_at	LTC4S	U50136_rna1_at
	ARPC1B	AF006084_at	DEFA4	X65977_at
	HMGA1	L17131_rna1_at	**CLEC3B**	X64559_at
	TYMS	D00596_at		
	DNMT1	X63692_at		
HG-U95A	SOX4[a]	33131_at	TEK	1596_g_at
	C7orf24	41696_at	FXYD1	32109_at
	POSTN	1451_s_at	ABCA8	35717_at
	BAZ1B	32261_at	**CLEC3B**	36569_at
	KIAA0101[a]	38116_at	CBX7	36894_at
	RECQL	34684_at	TNXA /// TNXB	38508_s_at
	FAT	40454_at	SH3BP5	38968_at
	SIPA1L3	37831_at	CA4	40739_at
	MARCKSL1	36174_at	FBXO9	38990_at
	CKAP4	32529_at	**COX7A1**	39031_at
	KIF14	34563_at	GABARAPL1	35785_at
	SUB1	36171_at	ADH1B	35730_at
			PTGDS	216_at

a. These genes are also identified as common cancer signature genes in Rhode et al. [101].

For F = 90% and 70%, the common cancer signatures consist of slightly different genes. (For F = 90%, 39 out of the above 46 genes appear and for F = 70%, 10 more

genes are added to the 46 genes.) For the rest of this chapter, we will focus on the 46-gene common cancer signature corresponding to $F = 80\%$.

4.3.2 Validation of the Signature on the Training Data

To validate the reliability and robustness of the common cancer signature, a G-TSP classifier, which predicts cancer vs. normal status, is built based on all the signature genes, with 24 genes in G_1 and 22 genes in G_2 as indicated above. Recall that the classification rule for the G-TSP classifier is that if the average relative rank of the genes in G_1 is less than that of the genes in G_2, a test sample is classified as normal; otherwise it is classified as cancer. The expression values of the 46 signature genes are illustrated in Figure 4.2 using the Stearman_Lung data set. Distinct patterns of expression values of the genes in G_1 and G_2 can be observed for normal and cancer samples. The classifier is then used to assess the prediction accuracy of the signature on the training data sets spanning a wide range of cancer types. For the 26 individual data sets which have been integrated and used to identify the signature, the classification accuracy and the *p*-values of the Fisher's exact test are shown in Table 4.5 and 4.6. The classifier achieves high accuracy (> 85%) on 19 of 26 data sets and the overall accuracy is about 86%. From the *p*-values of the Fisher's exact test, we learn that the classification is significant (*p*-value < 0.03) on 18 of 22 data sets. There is no *p*-value available for four data sets which only have samples from one class. The classifier is both significant (*p*-value < 0.03) and accurate (> 85%) on 14 of 22 data sets. The results suggest that we have identified a common cancer signature for most, if not all, human cancer types.

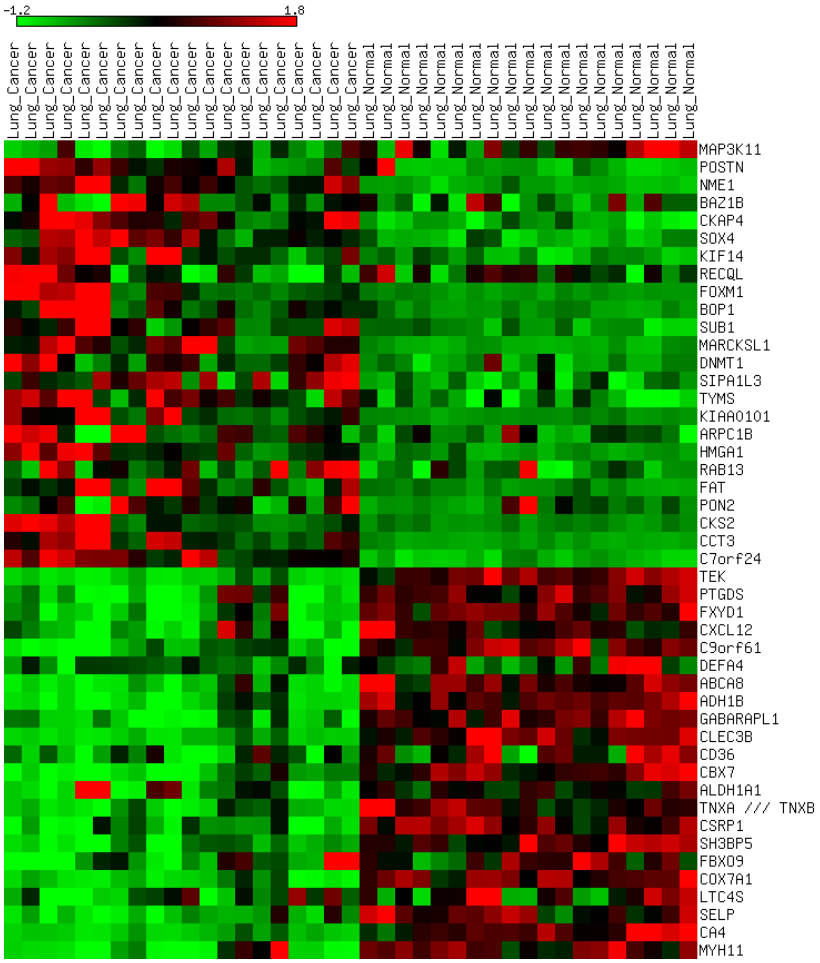


Figure 4.2: Common cancer signature which can discriminate cancer from normal samples. The Stearman_Lung data is used to illustrate the gene expression values of the signature genes in the figure. The heatmap is generated by the matrix2png software [168]. The expression value for each gene is normalized across the samples to zero mean and one standard deviation (SD) for visualization purposes. Genes with expression levels greater than the mean are colored in red and those below the mean are colored in green. The scale indicates the number of SDs above or below the mean.

Table 4.5 Class prediction of the signature on training data from HuGeneFL arrays

Study	No. of normal	N of cancer	Accuracy (%)	*P*-value
Beer_Lung	10	86	95.8	8.87E-11
Dyrskjot_Bladder	4	40	95.5	1.18E-03
Hippo_Gastric	8	22	76.7	2.67E-01
Hsiao_Normal	59	0	91.5	N/A[a]
Lancaster_Ovarian	3	31	91.2	1.00
Logsdon_Pancreas	5	10	100	3.33E-04
Pomeroy_Brain	4	38	97.6	3.48E-04
Quade_Myometrium	4	14	77.8	2.29E-02
Ramaswamy_Multi	90	190	77.9	1.28E-20
Rickman_Brain	6	45	94.1	2.87E-04
Welsh_Ovarian	4	22	100	6.69E-05
Zhan_Myeloma	30	74	72.1	2.88E-01
Overall	227	572	85.0	6.70E-68

a. For the data sets with samples from only one class, no *p*-value is available.

Table 4.6 Class prediction of the signature on training data from HG-U95A arrays

Study	No. of normal	No. of cancer	Accuracy (%)	*P*-value
Bhattacharjee_Lung	17	47	90.6	6.34E-10
Cromer_Head-Neck	4	34	97.4	4.74E-04
Dehan_Lung	9	25	85.3	3.82E-05
Frierson_Salivary	6	16	95.5	9.38E-05
Giordano_Adrenal	3	11	92.9	1.10E-02
Gutmann_Brain	3	8	100	6.06E-03
Huang_Thyroid	8	8	75	5.59E-02
Shai_Brain	7	35	85.7	6.36E-05
Stearman_Lung	19	20	89.7	1.28E-07
Su_Multi	63	18	81.5	1.15E-05
Su_Tumors	0	154	93.5	N/A[a]
Welle_Normal	12	0	100	N/A
Yanai_Normal	24	0	91.7	N/A
Yu_Prostate	16	35	54.9	1.83E-02
Overall	191	411	87.0	3.65E-77

a. For the data sets with samples from only one class, no *p*-value is available.

4.3.3 Independent Data Validation of the Signature

To further validate the generality and robustness of the common cancer signature, the G-TSP classifier built from the signature is tested on six independent test data sets generated from a different generation of Affymetrix microarray platforms, HG-U133A (Table 4.3). The six independent data sets represent six different human cancer types, one of which is not represented in the training data sets (Yoon_Soft_Tissue). The prediction accuracy and statistical significance (i.e. *p*-value of the Fisher's exact test) of the common cancer signature are listed in Table 4.7. The signature significantly (*p*-value < 0.005) discriminates cancer from normal samples with very high accuracy (> 95%) on four of the six data sets, including the new cancer type data set. On the other two data sets, the signature achieves much higher accuracy (> 75%) than that of coin-flipping but only marginally significant (*p*-value = 0.083 and 0.107). The independent test results have validated that the signature is common to a wide range of cancer types and may be used to detect the presence of cancer cells in tissues.

Table 4.7 Validation of the signature on independent HG-U133A data

Study	No. of normal	No. of cancer	Accuracy (%)	*P*-value
Gordon_Lung	9	40	95.9	1.75E-07
Hoffman_Myometrium	5	5	80.0	8.33E-02
Lenburg_Kidney	5	12	76.5	1.07E-01
Talantov_Skin	7	45	98.1	3.44E-07
Wachi_Lung	5	5	100	3.97E-03
Yoon_Soft_Tissue	15	39	96.3	6.76E-11
Overall	46	146	94.3	9.74E-30

4.4 Discussion

The advent of DNA microarray has had a tremendous impact on cancer research. This technology provides a novel molecular tool, complementary to histopathologic examination, to assess the presence of cancer cells in patient tissues. The rapid accumulation of cancer microarray data makes it possible to integrate a large amount of microarray gene expression data across a wide range of cancer types to identify a universal cancer signature to detect cancer cells, regardless of the tissue from which the cancer is derived. In this study, by integrating microarray data and applying the G-TSP method combined with repeated random sampling, we have identified a robust cancer gene expression signature common to almost all major human cancer types. The discriminative power of the signature has been validated on both data sets involved in identifying the signature and independent test data sets. The G-TSP classifier built from the signature, which simply compares the average relative ranks of two groups of genes, achieves high accuracy on most of the training and test data sets with statistical significance. Although the signature has the potential to be developed as a robust and objective clinical diagnostic test for cancer, larger number of samples will be required to further refine and validate it.

An intriguing advantage of inter-study cancer microarray data integration is that it increases the statistical power to capture essential, cancer type-independent gene expression features, which might be masked by specific features of individual cancer types and small sample sizes of individual data sets. In this sense, the signature is reliable and robust to variations in individual cancer data sets. The universal cancer signature described here may play a crucial role in oncogenic processes and be used to improve our

understanding of cancer pathogenesis and ultimately design improved anticancer treatments. It also suggests the possible existence of therapeutic targets common to different cancer types.

It is not surprising that many of the signature genes (BOP1, KIAA0101, CCT3, ARPC1B, CKAP4, ALDH1A1, CD36, CLEC3B, TEK, CBX7) have been reported to be associated with specific types of cancer in the literature and some other genes (NME1, TYMS, POSTN, FOXM1, HMGA1, DNMT1, KIF14, CXCL12, SELP) have been previously found to be associated with a variety of distinct human cancer types. As defined by Gene Ontology Consortium (www.geneontology.org), the common signature genes are involved in cell cycle (MAP3K11, NME1, CKS2, MYH11), regulation of transcription (BAZ1B, SOX4, FOXM1, SUB1, CBX7, DNMT1, HMGA1, RAB13), DNA metabolism (RECQL, CBX7, DNMT1, TYMS, HMGA1), cellular biosynthesis (NME1, PTGDS, KIF14, TYMS, LTC4S), cellular protein metabolism (MAP3K11, TEK, KIF14, FBXO9, HMGA1, CCT3) and other important biological processes, such as cell organization and cell adhesion. These findings are consistent with the fact that all cancer types share the common features of uncontrollable cell growth and local tissue invasion, and therefore the genes that are essential to these cellular processes are possible signature genes among almost all cancer types.

We compare the common cancer signature to the meta-signature of neoplastic transformation reported in [101]. It is not surprising that there is a reasonable overlap between the two signatures (Table 4.4). On one hand, this implies that both studies have identified some molecular features common to all cancer types. On the other hand, the difference between the two signatures may result from the two major differences between

the methods: (a) the comparative meta-profiling method overlooks a large number of potential signature genes due to its intrinsic properties while our method includes all possible genes in the analysis; (b) their analysis has focused on genes overexpressed in one direction (cancer > normal) while our signature includes genes overexpressed in both directions as illustrated in Figure 4.2. We then compare the classification results of the two methods. For the data sets which have been used in both studies (e.g. Beer_Lung, Rickman_Brain, Welsh_Ovarian), the G-TSP classifier consistently achieves more accurate and more significant classification results than corresponding results reported in [101].

One limitation of our proposed method for microarray data integration is that it can only directly integrate microarray data generated from the same standard microarray platforms. Even with this limitation, we still obtain a large number of samples (> 500) on each of the two microarray platforms used in this study. With the rapid increase of available microarray data and the standardization of microarray technologies, the mass of microarray data generated from the same platforms will continue to grow, which will make our method become increasingly useful.

It is quite interesting that a similar study on common cancer biomarkers was published very recently [169]. The uniqueness of the study is that the researchers have generated microarray data across various cancer types using the same spotted cDNA microarray, and therefore no data integration is needed. By applying a gene pairing method to a training set with 201 samples of various normal and cancer tissues, a subset of 14 genes have been identified as common cancer biomarkers with high predictive power (87%) in segregating cancer from normal samples. The major limitation of the study is that the

cancer samples are dominated by only a few cancer types (colon, melanoma, ovarian and esophageal cancers). Therefore, the biomarkers identified in the study may not really be common to a broad range of cancer types. In our study, motivated by the work of Rhodes et al., we collected a broader range of microarray gene expression data for about 20 cancer types and each of them is reasonably represented in the training data sets. The signature identified in our study has been validated on independent data sets of various cancer types, including one cancer type which is not represented in the training data sets.

In conclusion, by combining large-scale microarray data, a robust common cancer signature has been identified. Upon more large-scale validation, it could be developed as a component of genomic-based clinical diagnostic tools for cancer patients. Further studies of the signature might also improve our understanding of cancer and identify new drug targets.

Chapter 5

Cancer Prognosis: A Robust Breast Cancer Prognostic Signature Identified from Inter-Study Microarray Data

5.1 Introduction

Breast cancer is the most common form of cancer and the second leading cause of cancer death among women in the United States, with an estimated 212,920 new cases and 40,970 deaths in 2006 [114]. The main cause of breast cancer death comes from its metastases to distant sites. Early diagnosis and adjuvant systemic therapy (hormone therapy and chemotherapy) substantially reduce the risk of distant metastases. However, adjuvant therapy has serious short-term and long-term side effects and involves high medical costs [170]. Therefore, highly accurate prognostic tests are essential for clinicians to decide at diagnosis which patients are at high risk of developing metastases and should receive adjuvant therapy. Currently, the most widely used treatment

guidelines, St. Gallen [171] and the US National Institutes of Health (NIH) [170] consensus criteria, assess a patient's risk of distant metastases based on clinical prognostic factors such as tumor size, lymph node status, and histologic grade. These guidelines cannot accurately identify at-risk patients and about 70-80% of patients defined as being at risk by these criteria and receiving adjuvant therapy would have survived without it [172]. Many patients who would be cured by local or regional treatment alone are 'over-treated' and suffer toxic side effects of adjuvant therapy unnecessarily. Therefore, there is an urgent need for new prognostic tests to precisely define a patient's risk of developing metastases to ensure that the patient receives appropriate therapy.

Over the past few years, a number of studies have identified prognostic gene expression signatures whose prediction of breast cancer outcome is superior to conventional prognostic tests [21, 23, 70-73, 75]. Among them, the two largest studies have attempted to identify gene expression signatures strongly predictive of distant metastases. van't Veer et al. [21] applied a supervised method to identify a 70-gene signature, capable of predicting a short interval to distant metastases, in a cohort of 78 young breast cancer patients (< 55 years of age) with lymph-node-negative tumors. The signature was validated in a cohort of 295 patients with either lymph-node-negative or lymph-node-positive breast tumors [70]. Using a different microarray platform, Wang et al. [72] derived a 76-gene prognostic signature from 115 lymph-node-negative patients who had not received adjuvant systemic treatment. The signature could be used to predict distant metastasis within 5 years in breast cancer patients of all age groups with lymph-node-negative tumors and was subsequently validated by a set of 171 lymph-node-

negative patients. These studies have shown that gene expression signatures would result in a substantial reduction of the number of patients who would receive unnecessary adjuvant systemic treatment, thereby preventing over-treatment in a considerable number of breast cancer patients.

The most striking observation when comparing the signatures from different studies is the lack of overlap of signature genes. For instance, in the studies of van't Veer et al. and Wang et al., despite the similar clinical and statistical designs, there is an overlap of only three genes in the two gene signature lists. These diverse results make it difficult to identify the most predictive genes for breast cancer prognosis. The disagreements in gene signatures may be partly due to the use of different microarray platforms, differences in patient selection, and experimental issues. However, in a recent study [84], reanalysis of the van't Veer data has shown that the prognostic signature is not unique even from the same data set and it is strongly influenced by the subset of the patients used for signature selection. This observation indicates that given the small number of samples in the training sets, many genes might show correlations with clinical outcome and differences between these correlations are small. Therefore, it is possible to combine genes in many ways to generate different signatures with the similar predictive power when validated on internal test sets [86]. In general, these signatures cannot be validated on independent novel data sets [71]. Another independent reanalysis on other microarray data sets has shown very similar findings [173]. Given the large numbers of features (genes) (~ 10,000 to 40,000) of microarray data and the relatively small numbers of samples (patients) (~ 100) used in the training set of each study, it is highly possible to accidentally find a set of genes with good predictive power on internal test sets. In light of the observations of

above reanalysis studies, the disagreements in gene signatures obtained from different data sets are not surprising. We believe that much larger numbers of samples (patients) are needed to develop more robust prognostic signatures.

In this chapter, by using the TSP method as a novel feature selection and transformation procedure, we integrate three independent microarray gene expression data sets in order to increase the sample size to identify a robust prognostic gene expression signature for breast cancer. The feature selection and transformation procedure generates a new rank-ordered feature set with much smaller number of features (usually less than 1,000) than the number of the original features (~ 22,000), from the integrated training data set. All the samples included in this study are from lymph-node-negative patients who have not received adjuvant systemic treatment. Each new feature is a pair of genes and each feature value is either 1 or -1. By applying this novel feature selection procedure, no data normalization is needed before data integration. The number of genes in a prognostic gene expression signature is optimized by using an *m*-fold cross-validation scheme on the integrated training data of 358 samples. The optimal number of features (gene pairs) is estimated to be 40, corresponding to 61 genes. (Some genes appear in more than one feature.) A correlation-based classifier is built from the prognostic gene expression signature by using a cut-off correlation coefficient and it classifies patients as poor-outcome, meaning they are likely to metastasize, or good-outcome, meaning that they are unlikely to develop distant metastases. The prognostic gene expression signature is validated by an independent microarray data set of 159 patients. Upon further validation on large-scale independent data, the prognostic gene

expression signature may be used to develop a novel breast cancer prognostic test and help to avoid over-treatment of newly diagnosed patients.

5.2 Methods

5.2.1 Patient Samples Selection

Four breast cancer microarray data sets are included in this chapter. Each data set has been downloaded from publicly available gene expression repositories (e.g. Gene Expression Omnibus) or supporting web sites [72, 75, 82, 174]. All the four data sets are generated from the same Affymetrix HG-U133A microarray platform. Here, the names of the first authors of individual studies are used as the names of the data sets. Three data sets, Miller with 251 patients, Sotiriou with 189 patients and Wang with 286 patients, are used as training data and the other one, Pawitan with 159 patients, is used as independent test data. The reason to make such choice is that detailed clinical information has been provided for the three data sets and has been used to select specific patients for training. Little clinical information is provided for the Pawitan study. For the Miller, Sotiriou and Pawitan studies, because the gene expression data sets provided by them have undergone cross-sample normalization, we have downloaded the raw .CEL files and calculated expression values using the Affymetrix GeneChip Operating Software (version 1.4). There is an 85-patient overlap between Miller and Sotiriou data sets, so we have excluded the replicate samples from our study. Detailed patient information in each study has been described in the corresponding literature.

In this chapter, we focus on developing a prognostic gene expression signature which can identify patients who are likely to develop distant metastases within five years.

Motivated by a recent study [175], we employ the idea of using extreme patient samples, which are more informative in identifying a prognostic signature, to form training data. Extreme patients are either short-term survivors who got a poor-outcome within a short period or long-term survivors who were maintaining a good-outcome after a long follow-up time. Specifically, we select patients who developed distant metastases (relapse) within five years as poor-outcome samples and patients who were free of distant metastases (relapse) during the follow-up for a period of at least eight years as good-outcome samples. The sharp contrast between short-term and long-term survivors should identify more informative and reliable genes for a prognostic signature. Only early stage lymph-node-negative patients who had not received adjuvant systemic treatment are included in the training data because adjuvant treatment is likely to modify patient outcome. The selection is irrespective of age, tumor size and other clinical parameters. After applying the above selection criteria, a total number of 358 patients from the three training data sets are used to identify a prognostic signature. The numbers of selected patients from each training data set are listed in Table 5.1.

Table 5.1 Training data sets – lymph-node-negative with no adjuvant treatment

Data set	**No. of patients**	**No. of good-outcome**	**No. of poor-outcome**
Miller [82]	106	92	14
Sotiriou [75]	43	30	13
Wang [72]	209	114	95
Total	358	236	122

5.2.2 Data Integration

Since ranks of gene expression values within a profile are invariant to monotonic data transformations within each microarray, by applying the rank-based TSP method as a

feature selection and transformation procedure (see 'Feature Selection and Transformation' in Section 5.2.3), no data normalization and transformations are required before integration. We directly merge gene expression data of the patients from three training data sets in Table 5.1, using the 22283 probe sets on Affymetrix HG-U133A microarray, to form an integrated training data set of 358 samples.

5.2.3 Feature Selection and Transformation

Here, we use the idea of the TSP method to develop a novel feature selection and transformation procedure to select the most discriminating candidate genes for breast cancer prognosis from an integrated training data set. After applying this procedure, we will obtain a ranked list of new features with much smaller number of features. Each new feature is a pair of genes with a feature value being 1 or -1. The procedure is described as follows.

The feature selection procedure consists of forming a list of gene pairs, sorted from the largest to the smallest according to their scores Δ_{ij} as defined in Equation (2.3), and selecting the top K pairs. The K top-ranked gene pairs are the most discriminating candidate gene pairs for breast cancer prognosis. During the process, we have transformed the original feature (gene) set of 22,283 features to a new rank-ordered feature (gene pair) set of only K features. For each gene pair (i, j) (i.e. a new feature) among the K selected pairs, if $X_{in} < X_{jn}$, $n = 1, 2, \ldots, N$, then the value of the new feature for the n-th sample is set to be -1; otherwise, the value of the new feature for the n-th sample is set to be 1. After this procedure, the original $P \times N$ training data matrix is reduced to a $K \times N$ data matrix ($K \ll P$). The parameter K approximately sets an upper

bound for the number of genes in a prognostic signature. In our practice, there are always more than *K* distinct genes among the top *K* gene pairs. Given that the numbers of genes in published breast cancer prognostic signatures are mostly less then 100, the parameter *K* is set to be 500 in this study to make sure we can identify a prognostic signature with a reasonable number of genes.

5.2.4 Prognostic Gene Expression Signature Identification

The number of gene pairs in a prognostic signature is optimized by using the *m*-fold cross-validation scheme. The procedure is as follows: (1) divide the integrated training data set into *m* disjoint subsets of approximately equal sample size; (2) leave out one subset and put together the other *m*-1 subsets to form a training set; (3) generate a feature list of *K* rank-ordered gene pairs from the training set by using the TSP-based feature selection procedure; (4) generate a feature value data set (with the *K* features in the same order as in the feature list) for both the training set and the left-out subset as described in Section 5.2.3; (5) sequentially add subsets of five top-ranked features from the feature list to form a series of classifiers with different numbers of features, until all the *K* features are used; (6) classify the left-out samples using each classifier and count the number of misclassified samples for each classifier (For each left-out sample, classification is made on the basis of the correlations of its feature values with the mean feature values of the training samples from the good-outcome and the poor-outcome patients, respectively. Note that the numbers of features are various for different classifiers.); (7) repeat steps (2)-(6) exhaustively for all *m* subsets in step (1); (8) group classifiers based on the number of features in each classifier and count the total number of misclassified samples for each group. The number of features in the group of classifiers with the minimum total

number of misclassified samples is the optimal number, say K_{opt}, of features for the final prognostic signature. The final prognostic signature includes K_{opt} top-ranked features (gene pairs) of the feature list generated from the original integrated training set using the feature selection and transformation procedure.

5.2.5 Class Prediction

In clinical practice, when selecting breast cancer patients for adjuvant systemic therapy, a lower number of poor-outcome patients assigned to the good-outcome category should be attained. The conventional guidelines (e.g. St. Gallen and NIH) for breast cancer treatment usually predict breast cancer outcome with about 90% sensitivity but only 20 - 30% specificity. Sensitivity (respectively, specificity) is defined as the rate of the number of correctly predicted poor-outcome (respectively, good-outcome) patients to the total number of poor-outcome (respectively, good-outcome) patients. Our prognostic classifier should maintain the high standard of sensitivity (above 90%) while at the same time achieve the highest specificity. Therefore, when we try to build a classifier from the identified prognostic signature, we set a threshold of 90% for its sensitivity. We only consider prognostic classifiers whose sensitivities are higher than the threshold as candidate classifiers. For this purpose, we build a classifier from the prognostic signature as follows: (1) generate a feature (gene pair) value data set for the original integrated training data using the features in the prognostic signature; (2) calculate the correlation coefficient of feature values of each sample with the mean feature values of the training samples from the good-outcome patients; (3) order the samples according to their correlation coefficients calculated in the previous step; (4) get a cutoff correlation coefficient for each possible sensitivity value between 90% and 100% by correlation-

based classification in which we classify samples with a correlation coefficient above the cutoff value as good-outcome and below the cutoff value as poor-outcome; (Given the finite number of poor-outcome samples in the training data, only finite number of possible sensitivities are achievable. At each sensitivity level, the cutoff value is the one achieving the highest specificity.) (5) calculate the odds ratio for each possible sensitivity between 90% and 100%. (The odds ratio is the ratio of the odds that a patient is poor-outcome in the group of patients who are classified as poor-outcome to the odds that a patient is poor-outcome in the group of patients who are classified as good-outcome. A prognostic classifier with a higher odds ratio is generally preferred to those with lower odds ratio.) The cutoff value achieving the maximum odds ratio is the optimal cutoff value, say c_{opt}, for our classifier. Given a test sample, we first generate its feature values using the features in the prognostic signature; then calculate the correlation coefficient of these feature values with the mean feature values of the good-outcome patients from the integrated training samples. The classification rule is that if the correlation coefficient is greater than c_{opt}, the test sample is classified as good-outcome; otherwise, it is classified as poor-outcome.

5.2.6 Statistical Analysis

We compute the odds ratio of our prognostic signature for developing distant metastases within five years between the patients in the poor-outcome group and good-outcome group as determined by our correlation based classifier. *P*-values associated with odds ratios are calculated by the Fisher's exact text. We also plot the Kaplan-Meier curve of the signature on the independent test data with *p*-values calculated by log-rank

test. The Mantel-Cox estimation of hazard ratio of distant metastases within five years for the signature is also reported. All the statistical analyses are performed using MATLAB.

5.3 Results

5.3.1 A Prognostic Signature from Integrated Microarray Data

We directly merge the three microarray data sets in Table 5.1, using the 22283 probe sets on Affymetrix HG-U133A microarray, to form an integrated training data set. The integrated data set consists of 122 extreme poor-outcome samples (distant metastases within five years after surgery) and 236 extreme good-outcome samples (free of distant metastases during the follow-up for a period of at least eight years after surgery). The number of gene pairs in our prognostic signature is optimized by using a 40-fold cross-validation method on the integrated training data as described in Section 5.2.4. The relationship between the number of features (gene pairs) in a prognostic classifier and the total number of misclassified samples evaluated by the 40-fold cross-validation is plotted in Figure 5.1. The optimal number of features for our prognostic signature is $K_{opt} = 40$ and the prognostic classifier built from an optimal prognostic signature with 40 gene pairs achieves an unbiased classification accuracy of 66.8% estimated from the 40-fold cross-validation. Our final prognostic signature consists of the 40 top-ranked features (gene pairs) from the feature list generated from the original integrated training data using the feature selection and transformation procedure. Because some genes appear in more than one gene pair, the 40 top-ranked gene pairs in our prognostic signature include 61 distinct genes (Table 5.2).

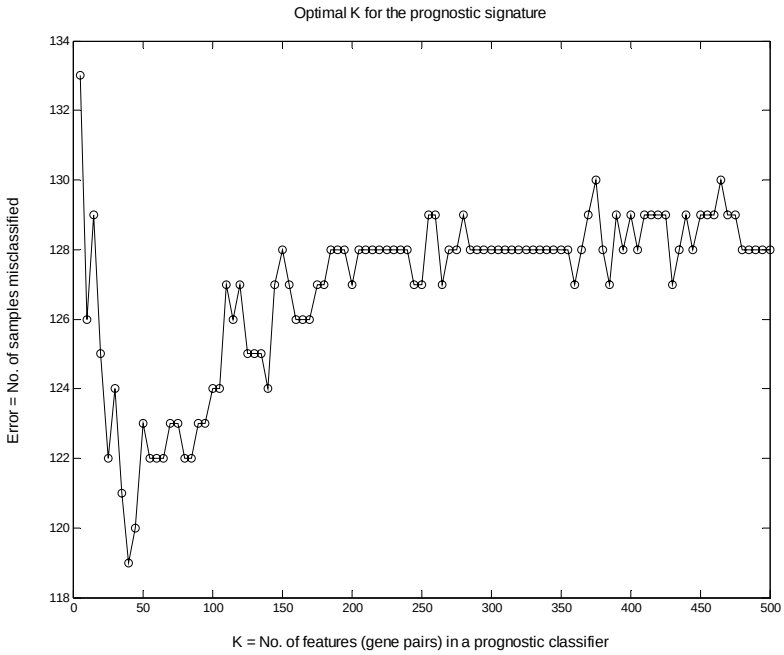


Figure 5.1: Relationship between the number of the features in a prognostic classifier and the total number of misclassified samples evaluated by 40-fold cross-validation. The optimal number of features is 40 in the above plot.

Table 5.2 Genes in the identified prognostic signature

Probe set	Gene symbol	Gene title
204641_at	NEK2	NIMA (never in mitosis gene a)-related kinase 2
212188_at	KCTD12	potassium channel tetramerisation domain containing 12
212022_s_at	MKI67	antigen identified by monoclonal antibody Ki-67
205264_at	CD3EAP	CD3e molecule, epsilon associated protein
206687_s_at	PTPN6	protein tyrosine phosphatase, non-receptor type 6
209574_s_at	C18orf1	chromosome 18 open reading frame 1
210199_at	CRYAA	crystallin, alpha A
204177_s_at	KLHL20	kelch-like 20 (Drosophila)
203010_at	STAT5A	signal transducer and activator of transcription 5A
205034_at	CCNE2	cyclin E2
222077_s_at	RACGAP1	Rac GTPase activating protein 1
203332_s_at	INPP5D	inositol polyphosphate-5-phosphatase, 145kDa
211584_s_at	NPAT	nuclear protein, ataxia-telangiectasia locus
209671_x_at	TRAC	T cell receptor alpha constant
208952_s_at	LARP5	La ribonucleoprotein domain family, member 5
211581_x_at	LST1	leukocyte specific transcript 1
205395_s_at	MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)
202602_s_at	HTATSF1	HIV-1 Tat specific factor 1
204817_at	ESPL1	extra spindle poles like 1 (S. cerevisiae)
215783_s_at	ALPL	alkaline phosphatase, liver/bone/kidney
204825_at	MELK	maternal embryonic leucine zipper kinase
206545_at	CD28	CD28 molecule
206364_at	KIF14	kinesin family member 14
208079_s_at	AURKA	aurora kinase A
210966_x_at	LARP1	La ribonucleoprotein domain family, member 1
204498_s_at	ADCY9	adenylate cyclase 9
206211_at	SELE	selectin E (endothelial adhesion molecule 1)
201890_at	RRM2	ribonucleotide reductase M2 polypeptide
204847_at	ZBTB11	zinc finger and BTB domain containing 11
203214_x_at	CDC2	cell division cycle 2, G1 to S and G2 to M
204605_at	CGRRF1	cell growth regulator with ring finger domain 1
210042_s_at	CTSZ	cathepsin Z
203595_s_at	IFIT5	interferon-induced protein with tetratricopeptide repeats 5
202114_at	SNX2	sorting nexin 2
91816_f_at	RKHD1	ring finger and KH domain containing 1
213139_at	SNAI2	snail homolog 2 (Drosophila)
219716_at	APOL6	apolipoprotein L, 6
218009_s_at	PRC1	protein regulator of cytokinesis 1
219579_at	RAB3IL1	RAB3A interacting protein (rabin3)-like 1

Table 5.2 (continue) Genes in the identified prognostic signature

Probe set	Gene symbol	Gene title
221824_s_at	MARCH8	membrane-associated ring finger (C3HC4) 8
219493_at	SHCBP1	SHC SH2-domain binding protein 1
212747_at	ANKS1A	ankyrin repeat and sterile alpha motif domain containing 1A
217427_s_at	HIRA	HIR histone cell cycle regulation defective homolog A (S. cerevisiae)
36545_s_at	SFI1	Sfi1 homolog, spindle assembly associated (yeast)
218883_s_at	MLF1IP	MLF1 interacting protein
219512_at	C20orf172	chromosome 20 open reading frame 172
221193_s_at	ZCCHC10	zinc finger, CCHC domain containing 10
221521_s_at	GINS2	GINS complex subunit 2 (Psf2 homolog)
218726_at	DKFZp762E1312	hypothetical protein DKFZp762E1312
221273_s_at	LOC653323	similar to tripartite motif protein 32 (predicted)
214973_x_at	IGHD	immunoglobulin heavy constant delta
211881_x_at	IGLJ3	immunoglobulin lambda joining 3
218143_s_at	SCAMP2	secretory carrier membrane protein 2
212911_at	DNAJC16	DnaJ (Hsp40) homolog, subfamily C, member 16
213689_x_at	RPL5	Ribosomal protein L5
214955_at	TMPRSS6	transmembrane protease, serine 6
218830_at	RPL26L1	ribosomal protein L26-like 1
219298_at	ECHDC3	enoyl Coenzyme A hydratase domain containing 3
211251_x_at	NFYC	nuclear transcription factor Y, gamma
213007_at	KIAA1794	KIAA1794
221529_s_at	PLVAP	plasmalemma vesicle associated protein

Now, we need to build a prognostic classifier which could achieve a high sensitivity (~ 90%) and a much higher specificity than those (20 - 30%) of the conventional criteria, from the prognostic signature. Following the steps in Section 5.2.5, we have calculated the odds ratio for each possible sensitivity value between 90% and 100%. The relationship between the sensitivity and the odds ratio is plotted in Figure 5.2. The best odds ratio (= 32.3) is achieved when the sensitivity is 97.5% (119 out of the 122 poor-outcome samples). At the 97.5% sensitivity level, the highest specificity is 44.9% (106 out of the 236 good-outcome samples) and the cutoff correlation coefficient achieving

that specificity is c_{opt} = 0.446. The odds ratio at the 100% sensitivity level has been excluded from consideration because it is an infinite number and when the sensitivity is 100%, the highest possible specificity is 7.2%, which is lower than those of the conventional guidelines.

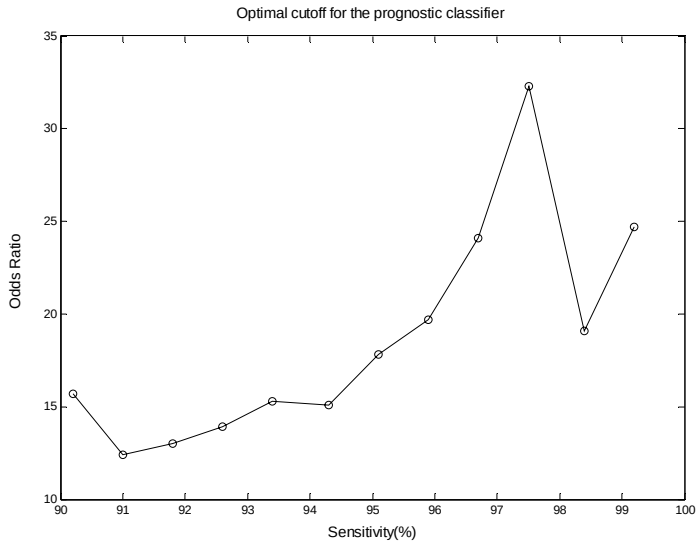


Figure 5.2: The relationship between the sensitivity and the odds ratio of a prognostic classifier built from our prognostic signature. The best odds ratio (= 32.3) is achieved when the sensitivity is 97.5%.

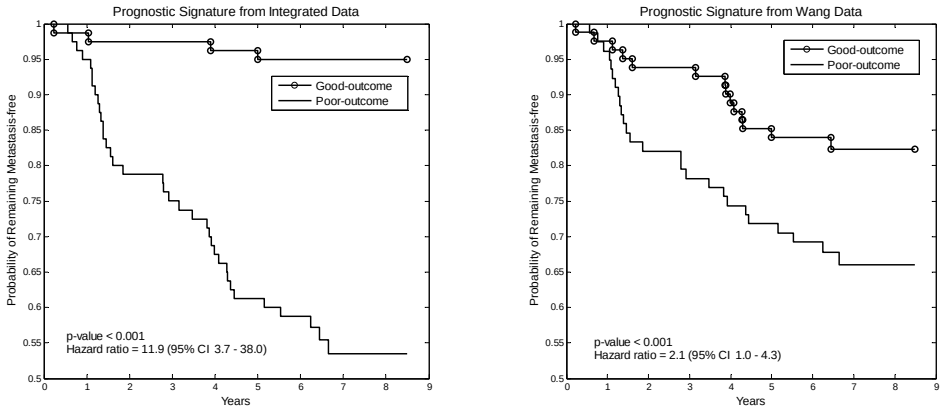
5.3.2 Validation of the Signature on Independent Data

To validate the prognostic classifier built from the prognostic signature, an independent data set, the Pawitan data set [174], of 159 primary breast cancer patients is used as independent test data. The data set includes both patients with lymph-node-negative tumors and patients with lymph-node-positive tumors. The patients had or had not received adjuvant systemic therapy. As reported in most of the literature, we use the prognostic classifier to predict the development of distant metastases within five years. Of the 159 patients, 35 patients developed distant metastases (relapse) within five years

(poor-outcome), and 119 patients were free of distant metastases (relapse) during the follow-up for a period of at least five years (good-outcome). Note that the definition of good-outcome for patients in the validating data is different from the definition in the training data because we have used extreme samples to identify the prognostic signature. The prognostic classifier achieves an accuracy of 68.8% with a sensitivity of 88.6% (31 out of the 35 poor-outcome samples) and a specificity of 63.0% (75 out of the 119 good-outcome samples) on the 154 samples included in the validating data set. The remaining five patients, who either developed distant metastases after five years or were free of distant metastases with a follow-up period less than five years, are not included in the validating data set. We compute the odds ratio of the prognostic classifier for developing metastases within five years between the patients in the poor-outcome group and in the good-outcome group as determined by the prognostic classifier. The prognostic classifier has a high odds ratio of 13.2 (95% confidence interval: 4.37 – 39.92) with a Fisher's exact test p-value < 0.0001.

To obtain a more useful estimate of the clinical outcome, we apply the prognostic classifier to all of the 159 samples in the Pawitan data set and calculate the probability of remaining free of distant metastases according to the prognostic classifier by using Kaplan-Meier analysis. The Kaplan-Meier curve of the prognostic signature shows a significant difference (p-value < 0.001) in the probability of remaining free of distant metastases between the patients in the poor-outcome group and those in the good-outcome group (Figure 5.3). The p-value is computed by the use of log-rank test. The Mantel-Cox estimation of hazard ratio for distant metastases within five years in the

poor-outcome group as compared to the good-outcome group is 11.9 (95% confidence interval: 3.7 – 38.0, p-value < 0.001).



(A) (B)

Figure 5.3: Kaplan-Meier analysis of the probability of remaining free of distant metastases among 159 patients between the good-outcome group and the poor-outcome group according to the prognostic classifier from the integrated data (A) and the one from a specific data set – the Wang data (B). CI denotes confidence interval. The p-value is calculated by the log-rank test.

5.3.3 Comparison of the Signature to a Study-Specific Signature

To evaluate the potential statistical power gained by integrating multiple data sets to increase sample size, we compare the predictive power of our prognostic signature with that of a study-specific prognostic signature identified from a single data set, the Wang data set [72], using the same method. As listed in Table 5.1, the Wang data set consists of 95 extreme poor-outcome samples (distant metastases within five years) and 114 extreme good-outcome samples (free of distant metastases during the follow-up for at least eight years). By using a 40-fold cross-validation method, the optimal number of gene pairs for the Wang prognostic signature is reached at K_{opt} = 35. The cutoff correlation coefficient

for the Wang prognostic classifier is found to be c_{opt} = 0.177. The Wang prognostic classifier is tested on the same independent test data, the Pawitan data. The overall accuracy, sensitivity and specificity of the Wang classifier are 57.8% (89 out of 154 total samples), 62.9% (22 out of the 35 poor-outcome samples) and 56.3% (67 out of the 119 poor-outcome samples), respectively. The odds ratio of the Wang signature is 2.2 (95% confidence interval: 1.0 – 4.7, *p*-value = 0.02). Although the Kaplan-Meier curve of the Wang signature plotted in Figure 5.3 shows a significant difference between the patients in the poor-outcome and good-outcome groups, the estimated hazard ratio of 2.1 (95% confidence interval: 1.0 – 4.3, *p*-value < 0.001) is much lower than that of the prognostic signature from the integrated data. The comparisons have shown that the prognostic signature from the integrated data is far superior to the Wang prognostic signature from a single data set. These results have highlighted the value of data integration and the robustness of the prognostic signature identified from the integrated data. By integrating multiple microarray data, the study-specific effect is reduced and more robust features of breast cancer prognosis are captured by our prognostic signature.

5.4 Discussion

In this chapter, by using the TSP method as a novel feature selection and transformation procedure, we integrate three independent microarray gene expression data sets of extreme samples and identify a robust 61-gene breast cancer prognostic signature from the integrated training data set of 358 samples. All of the patients in the training set had lymph-node-negative tumors and had not received adjuvant systemic treatment, so the identification of the prognostic signature is not subject to potential confounding factors related to lymph node status or systemic treatment. A correlation-

based classifier, which predicts whether a breast cancer patient will develop distant metastases within five years after initial treatment, is built from the prognostic signature and achieves a sensitivity of 88.9% and a specificity of 63.0% on an independent test data set of 154 samples. The test set includes patients, who had or had not received adjuvant systemic treatment, with either lymph-node-negative or lymph-node-positive tumors, indicating that our prognostic signature could be applied to all breast cancer patients independent of age, tumor size, tumor grade, lymph node status, and systemic treatment.

Comparison with the conventional treatment guidelines (e.g. St. Gallen and NIH) is instructive. While maintaining almost the same level of sensitivity (~ 90%), our 61-gene prognostic signature can achieve a much higher specificity of 63%, compared with 20-30% by the St. Gallen and NIH criteria. This means that our signature can spare a significant number of good-outcome patients from unnecessary adjuvant therapy, whiling ensuring roughly the same number of poor-outcome patients to receive adjuvant therapy as recommended by the conventional guidelines. Therefore, our prognostic signature, if further confirmed on large-scale independent data, could potentially provide a useful means of guiding adjuvant systemic treatment that could greatly reduce the cost of breast cancer treatment and improve the quality of patients' lives.

The strengths of our study, compared with previous studies, are the larger number of homogeneous patients (lymph-node-negative tumors without adjuvant systemic treatment) in the training set, and an external independent test set. In each of the two major breast cancer prognostic signature studies [21, 72], the training and validation data are generated by the same study group from the same population. Actually, they just randomly divide the whole data set into a training set and a validation set. In this case, the training data

and the validation data are likely to be more homogenous, therefore the study-specific prognostic classifier identified from the training data gives over-promising results when it is validated on the homogenous validation data. That might be the reason why the two major prognostic signatures, although validated internally with about 90% sensitivity and about 50% specificity, cannot be validated in an external data set [71]. Dividing the whole data set into training and validation subsets decreases the sample size and can produce other sources of bias [84]. In our study, we increase the sample size by integrating multiple microaray data with patients from different population. By selecting a homogeneous subgroup of patients and combining data from multiple studies, the derived prognostic signature is less sensitive to study-specific factors. An intriguing advantage of inter-study data integration is that it increases the statistical power to capture essential prognostic features which might be masked by study-specific features and small sample size of individual data sets. In this sense, our prognostic signature is more robust to inter-study variability and may facilitate external validation of the signature.

Comparison of our prognostic signature with the two major signatures of van't Veer et al. and Wang et al. is not straightforward because of differences in patients, microarray platforms, and algorithms. The study of van't Veer et al. uses an Agilent array platform and our study uses an Affymetrix array platform. There are only 12 out of the 40 gene pairs in our prognostic signature present on the Agilent Hu25K array and only 36 of the 70 genes in the van't Veer signature present on the Affymetrix HG-U133A array. Therefore, we can neither validate the van't Veer signature on our validation data nor validate our signature on their data set. There is a two-gene overlap between the van't

Veer signature and our signature (CCNE2 and PRC1). Since the data set in Wang et al. is included in our training set, we cannot independently validate our signature on the data set. On the other hand, in order to validate the signature of Wang et al., we need to know the oestrogen receptor (ER) status of our test samples because the classification rule of their signature is depend on ER status, which is absent from our validation data. Again, there is a two-gene overlap between the signature of Wang et al. and our signature (CCNE2 and MLF1IP). It is noteworthy that the gene CCNE2 is included in all of the three signatures and is reported to be related to breast cancer [176]. CCNE2 could be a potential target for the rational development of new cancer drugs.

To assess the benefit of data integration, we compare the predictive power of our prognostic signature with that of a study-specific prognostic signature identified from the Wang data set using the same method. When validated on the same independent test data, our signature consistently outperforms the study-specific signature in terms of sensitivity (88.9% vs. 62.9%), specificity (63.0% vs. 56.3%), odds ratio (13.2 vs. 2.2), hazard ratio (11.9 vs. 2.1), and Kaplan-Meier analysis (Figure 5.3). The findings have shown that a prognostic signature from a single data set is study-specific and could not be validated on external data, while a prognostic signature from the integrated data is more robust to study-specific factors and could possibly be validated on external data.

Recently, some studies have shown that combining gene expression data and conventional clinical data (e.g. tumor size, grade, ER status) could lead to improved breast cancer prognosis [177, 178]. Our approach does not rule off the possibility of combining gene expression signatures with traditional clinical factors to maximize the use of available information for breast cancer prognosis. In this study, due to the lack of

clinical information for some of the training data sets, we could not incorporate clinical information into the development of our prognostic signature. Whenever the clinical information is publicly available, it will be interesting to combine them with the integrated gene expression data to identify a better prognostic signature.

In conclusion, the work in this chapter has not only identified a robust breast cancer prognostic signature by integrating currently available microarray data, but also proposed a simple but novel model for integrating multiple microarray data in order to develop a more reliable and robust prognostic signature. This approach will be increasingly useful as more and more microarray data will be generated and become publicly available in near future. With the inclusion of more samples, a prognostic signature will be continuously refined and a consensus signature will finally be reached.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The main objective of this thesis is to propose a novel statistical method, based on the TSP classifier family, for inter-study microarray data integration, and to apply it to address two of the key problems in cancer research: cancer diagnosis and clinical outcome prediction.

Profiling gene expression using DNA microarrays has had a tremendous impact on cancer research. Gene expression profiling of cancers has expanded exponentially in the past decade and represents the largest category of research using DNA microarray technology. One of the big challenges in microarray data analysis is that only a small number of samples, relative to a very large number of variables, can be analyzed in a given study. This makes model development and data interpretation difficult, given that

cancer is a complex and heterogeneous disease. Therefore, it is not surprising that there are only small overlaps between gene signatures from different studies focusing on a similar cancer problem. We believe that much more samples are needed to identify robust cancer signatures, and to take full advantage of microarray technology in cancer research. Given the high cost of microaray studies and the rarity of tumor specimens, it is a nature way to integrate microarray data from different studies to increase sample size.

In Chapter 2, following the previous work in [64], we systematically develop a family of rank-based classifiers, named TSP classifier family, including the (refined) TSP classifier, the *k*-TSP classifier, and the G-TSP classifier. These classifiers only use the rank orders of gene expression values within each profile, therefore are invariant to most within-array data normalization methods. This intriguing property makes them extremely useful for integrating inter-study microarray data without the need to perform data normalization and transformation. When compared to several leading classification methods on a large number of microarray data sets, all members of the TSP classifier family achieve comparable or better classification results. These results have demonstrated that the TSP classifier family may be used to identify cancer diagnostic and prognostic gene signatures capable of classifying different cancer conditions. Its demonstrated classification power, together with its invariant to data normalization, makes the TSP classifier family very attractive in microarray data integration, the goal being to identify robust cancer signatures.

In Chapter 3, we present a novel model for integrating multiple microarray data sets to identify cancer diagnostic signatures for individual cancer types. Prostate cancer microarray data sets are used to illustrate the model. By applying the TSP method, no

data normalization and transformation are required before integration. We directly merge three microarray data sets generated from the same Affymetrix oligonucleotide arrays to form integrated data sets of increased sample sizes. A robust prostate cancer diagnostic signature, which consists of only two genes, HPN and STAT6, has been identified from the integrated data sets. The TSP classifier built from the signature, which simply compares relative expression values of the two signature genes, has been successfully validated on two independent microarray data sets generated from two different microarray platform. A comparison of the signature identified from data integration with those identified from individual data sets has further verified the statistical power gained by increased sample size. Upon further validation on a large number of independent data, a potential application of the diagnostic gene signature would be to make a simple diagnostic chip to improve specificity and sensitivity of current prostate cancer diagnostic tests.

In Chapter 4, we have addressed the question of how to identify a common cancer signature from publicly available large-scale microarray data by data integration. A statistical model, which combines the G-TSP method with repeated random sampling, is proposed to integrate large-scale microarray data and to distinguish those genes that play a crucial role in oncogenic processes from those that are irrelevant to cancer. This model overcomes the limitations of meta-analyses by integrating large-scale microarray data generated from the same microarray platform, resulting in one signature per platform (rather than per study) and effectively increasing sample size. Integrating microarray data from the same platform can guarantee that all the genes on the platform will be included in the analysis, therefore avoiding losing any potential signature genes. A common cancer

signature, which consists of 46 genes that may play a crucial role in oncogenic processes, has been identified from about 1,500 microarray gene expression profiles. The G-TSP classifier, which discriminates cancer from normal samples, is built from the common cancer signature and validated on both data sets involved in identifying the signature and independent data. The cancer type-independent signature identified in this study may be used to improve our understanding of cancer pathogenesis and ultimately to develop new cancer drugs. Upon additional independent validation, the signature can be developed as a component of genomic-based clinical diagnostic tools for cancer patients.

In Chapter 5, we introduce a new model for applying the TSP method in microarray data integration. In this model, besides data integration, the TSP method is also used as a novel feature selection and transformation procedure to combine variable selection and transformation with dimensionality reduction. After the feature selection and transformation procedure is applied to a microarray data set, different classification methods can be employed to the data set generated on the new feature space. In this chapter, we use this model to address another important problem in cancer research: cancer prognosis. Specifically, a robust prognostic gene signature for breast cancer is identified from integrated microarray data using this model. A correlation-based classifier, which predicts whether a breast cancer patient will develop distant metastases within five years after initial treatment, is built from the prognostic signature and validated on independent data. The signature is highly informative in assessing the risk of developing distant metastases within five years. The prognostic signature, if further confirmed on large-scale independent data, could potentially provide a useful means of guiding adjuvant systemic treatment that could spare a significant number of breast cancer

patients from unnecessary adjuvant treatment, and therefore improve the quality of patients' lives.

In this thesis, we have proposed simple but novel models, based on the TSP classifier family, for identifying reliable and robust cancer signatures from inter-study microarray data integration. By applying these models to a large amount of cancer microarray data, we have identified several important gene signatures for cancer diagnosis and prognosis. All these signatures are properly validated on independent microarray data sets. A major advantage of inter-study microarray data integration is that it increases the statistical power to capture consistent features which might be masked by the small sample size and experimental artifacts in an individual data set. This could lead to the discovery of more robust and reliable cancer signatures which may offer more accurate cancer classification and improve the statistical significance of their association to clinical outcome.

In conclusion, our work has not only established new models for the identification of robust cancer gene signatures from accumulated microarray data, but also demonstrated how the great wealth of microarray data can be exploited to increase the power of statistical analyses. These models will be increasingly useful as more and more microarray data is generated and becomes publicly available in near future. With the inclusion of more samples, cancer gene signatures will be continuously refined and consensus signatures will finally be reached.

6.2 Future Work

The TSP-based data integration models have produced promising results for cancer diagnostic and prognostic signature identification. In the studies illustrated in this thesis,

we focus on integrating microarray gene expression data generated from the same platforms, avoiding losing any potential signature genes. There are some other ways to take advantage of the TSP method in data integration and we will outline a few as future research directions.

In some cases, there are only limited amount of microarray data generated from the some platform from different studies addressing a similar question. In this situation, we can consider to integrate data from different generations of the same microarray technology, for example, Affymetrix arrays. Some commercial microarray companies provide information on gene mapping among different generations of their microarrays. We can use this information to get a list of genes common to all arrays used in the data integration study. The models proposed in this thesis can then be applied to the integrated data sets of the list of common genes to increase statistical power and to derive reliable gene signatures. In this way, we gain statistical power at the price of loss of potential signature genes.

By using the TSP method as a feature selection and transformation procedure, we can not only integrate multiple microarray data sets, but also incorporate clinical and pathological data into the analysis. This could lead to improved cancer signatures which are mixtures of genes and clinical parameters. Further more, other high-throughput data, such as protein expression data, and tissue microarray data, can be effectively combined into the integrated microarray data in similar manners to correlate changes in gene expression profiles with changes in proteomic or tissue profiles.

Bibliography

1. Liotta, L. and E. Petricoin, *Molecular Profiling Of Human Cancer.* Nature Reviews Genetics, 2000. **1**: p. 48-56.
2. Bast, R.C., Jr., et al., *Translational Crossroads for Biomarkers.* Clin Cancer Res, 2005. **11**(17): p. 6103-6108.
3. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nature Biotechnology, 1996. **14**: p. 1675-1680.
4. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarray.* Nature Genetics Supplement, 1999. **21**: p. 33-37.
5. Schena, M., et al., *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.* Science, 1995. **270**(5235): p. 467-470.
6. Lonning, P.E., T. Sorlie, and A.-L. Borresen-Dale, *Genomics in breast cancer - therapeutic implications.* Nature Clinical Practice Oncology, 2005. **2**(1): p. 26-33.
7. Southern, E., K. Mir, and M. Shchepinov, *Molecular interactions on microarrays.* Nature Genetics Supplement, 1999. **21**: p. 5-9.

8. Wadlow, R. and S. Ramaswamy, *DNA microarrays in clinical cancer research.* Current Molecular Medicine, 2005. **5**: p. 111-120.

9. Bucca, G., et al., *Gene Expression Profiling of Human Cancers.* Ann NY Acad Sci, 2004. **1028**(1): p. 28-37.

10. Luo, J., et al., *Looking beyond morphology:cancer gene expression profiling using DNA microarrays.* Cancer Investigation, 2003. **21**(6): p. 937-949.

11. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* PNAS, 1998. **95**(25): p. 14863-14868.

12. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* PNAS, 2001. **98**(19): p. 10869-10874.

13. Quackenbush, J., *Computational analysis of microarray data.* Nature Reviews Genetics, 2001. **2**(2): p. 418-427.

14. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science, 1999. **286**(5439): p. 531-537.

15. Welsh, J.B., et al., *Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer.* PNAS, 2001. **98**(3): p. 1176-1181.

16. Wigle, D.A., et al., *Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Survival.* Cancer Res, 2002. **62**(11): p. 3005-3008.

17. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-511.

18. Bittner, M., et al., *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, 2000. **406**(6795): p. 536-540.

19. Dhanasekaran, S.M., et al., *Delineation of prognostic biomarkers in prostate cancer.* Nature, 2001. **412**: p. 822-826.

20. Takahashi, M., et al., *Gene expression profiling of clear cell renal cell carcinoma: Gene identification and prognostic classification.* PNAS, 2001. **98**(17): p. 9754-9759.

21. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**: p. 530-536.

22. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**: p. 747-752.

23. Ma, X.-J., et al., *A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.* Cancer Cell, 2004. **5**(6): p. 607-616.

24. Rosenwald, A., et al., *The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma.* N Engl J Med, 2002. **346**(25): p. 1937-1947.

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

26. Hedenfalk, I., et al., *Gene-Expression Profiles in Hereditary Breast Cancer.* N Engl J Med, 2001. **344**(8): p. 539-548.

27. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* PNAS, 2001. **98**(26): p. 15149-15154.

28. Su, A.I., et al., *Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures.* Cancer Res, 2001. **61**(20): p. 7388-7393.

29. Pomeroy, S.L., et al., *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**(6870): p. 436-442.

30. Shipp, M.A., et al., *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nat Med, 2002. **8**(1): p. 68-74.

31. Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**(2): p. 203-209.

32. Ramaswamy, S., et al., *A molecular signature of metastasis in primary solid tumors.* Nat Genet, 2003. **33**(1): p. 49-54.

33. Belbin, T.J., et al., *Molecular Classification of Head and Neck Squamous Cell Carcinoma Using cDNA Microarrays.* Cancer Res, 2002. **62**(4): p. 1184-1190.

34. Thomas, J.G., et al., *An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles.* Genome Res., 2001. **11**(7): p. 1227-1236.

35. Virtanen, C., et al., *Integrated classification of lung tumors and cell lines by expression profiling.* PNAS, 2002. **99**(19): p. 12357-12362.

36. Welsh, J.B., et al., *Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer.* Cancer Res, 2001. **61**(16): p. 5974-5978.

37. Stuart, R.O., et al., *In silico dissection of cell-type-associated patterns of gene expression in prostate cancer.* PNAS, 2004. **101**(2): p. 615-620.

38. Lapointe, J., et al., *Gene expression profiling identifies clinically relevant subtypes of prostate cancer.* PNAS, 2004. **101**(3): p. 811-816.

39. Dyrskjot, L., et al., *Identifying distinct classes of bladder carcinoma using microarrays.* Nature Genetics, 2003. **33**: p. 90-96.

40. Ono, K., et al., *Identification by cDNA Microarray of Genes Involved in Ovarian Carcinogenesis.* Cancer Res, 2000. **60**(18): p. 5007-5011.

41. Ooi, C.H. and P. Tan, *Genetic algorithms applied to multi-class prediction for the analysis of gene expression data.* Bioinformatics, 2003. **19**(1): p. 37-44.

42. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* PNAS, 2002. **99**(10): p. 6567-6572.

43. Liu, J.J., et al., *Multiclass cancer classification and biomarker discovery using GA-based algorithms.* Bioinformatics, 2005. **21**(11): p. 2691-2697.

44. Tan, A.C., et al., *Simple decision rules for classifying human cancers from gene expression profiles.* Bioinformatics, 2005. **21**(20): p. 3896-3904.

45. Giordano, T.J., et al., *Organ-Specific Molecular Classification of Primary Lung, Colon, and Ovarian Adenocarcinomas Using Gene Expression Profiles.* Am J Pathol, 2001. **159**(4): p. 1231-1238.

46. Zhang, H., et al., *Recursive partitioning for tumor classification with gene expression microarray data.* PNAS, 2001. **98**(12): p. 6730-6735.

47. Zhang, H., C.-Y. Yu, and B. Singer, *Cell and tumor classification using gene expression data: Construction of forests.* PNAS, 2003. **100**(7): p. 4168-4172.

48. Ben-Dor, A., et al., *Tissue Classification with Gene Expression Profiles.* Journal of Computational Biology, 2000. **7**(3-4): p. 559-583.

49. Li, L., et al., *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001. **17**(12): p. 1131-1142.

50. Ye, J., T. Li, and R. Janardan, *Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2004. **01**(4): p. 181-190.

51. Shen, R., et al., *Eigengene-based linear discriminant model for tumor classification using gene expression microarray data.* Bioinformatics, 2006. **22**(21): p. 2635-2642.

52. Peng, S., et al., *Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines.* FEBS Letters, 2003. **555**(2): p. 358-362.

53. Chu, F. and L. Wang, *Applications of support vector machines to cancer classification with microarray data.* International Journal of Neural Systems, 2005. **15**(6): p. 475-484.

54. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-914.

55. Bicciato, S., et al., *Pattern identification and classification in gene expression data using an autoassociative neural network model.* Biotechnology and Bioengineering, 2003. **81**(5): p. 594-606.

56. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nat Med, 2001. **7**(6): p. 673.

57. Dettling, M. and P. Buhlmann, *Boosting for tumor classification with gene expression data.* Bioinformatics, 2003. **19**(9): p. 1061-1069.

58. Dudoit, S. and J. Fridlyand, *Bagging to improve the accuracy of a clustering procedure.* Bioinformatics, 2003. **19**(9): p. 1090-1099.

59. Dettling, M., *BagBoosting for tumor classification with gene expression data.* Bioinformatics, 2004. **20**(18): p. 3583-3593.

60. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer series in statistics. 2001: Springer.

61. Mukherjee, S., et al., *Support Vector Machine Classification of Microarray Data*. 1998: *Technical Report CLCL Paper 182/AI Memo 1676 MIT*.

62. Liu, B., et al., *A combinational feature selection and ensemble neural network method for classification of gene expression data.* BMC Bioinformatics, 2004. **5**: p. 136.

63. Qiu, P., Z.J. Wang, and K.J.R. Liu, *Ensemble dependence model for classification and prediction of cancer and normal gene expression data.* Bioinformatics, 2005. **21**(14): p. 3114-3121.

64. Geman, D., et al., *Classifying gene expression profiles from pairwise mRNA comparison.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1): p. 19.

65. Antonov, A.V., et al., *Optimization models for cancer classification: extracting gene interaction information from microarray expression data.* Bioinformatics, 2004. **20**(5): p. 644-652.

66. Li, L. and H. Li, *Dimension reduction methods for microarrays with application to censored survival data.* Bioinformatics, 2004. **20**(18): p. 3406-3412.

67. Speed, T., *Statistical Analysis of Gene Expression Microarray Data*. Interdisciplinary Statistics. 2003: CHAPMAN & HALL/CRC.

68. Lossos, I.S., et al., *Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes.* N Engl J Med, 2004. **350**(18): p. 1828-1837.

69. Cole, B.F., et al., *Polychemotherapy for early breast cancer: an overview of the randomised clinical trials with quality-adjusted survival analysis.* The Lancet, 2001. **358**(9278): p. 277-286.

70. van de Vijver, M.J., et al., *A Gene-Expression Signature as a Predictor of Survival in Breast Cancer.* N Engl J Med, 2002. **347**(25): p. 1999-2009.

71. Naderi, A., et al., *A gene-expression signature to predict survival in breast cancer across independent data sets.* Oncogene, 2006: p. advance online publication doi: 10.1038/sj.onc.1209920.

72. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**: p. 671-679.

73. Chang, H.Y., et al., *Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.* PNAS, 2005. **102**(10): p. 3738-3743.

74. Sotiriou, C., et al., *Breast cancer classification and prognosis based on gene expression profiles from a population-based study.* PNAS, 2003. **100**(18): p. 10393-10398.

75. Sotiriou, C., et al., *Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis.* J Natl Cancer Inst, 2006. **98**(4): p. 262-272.

76. Buyse, M., et al., *Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer.* J Natl Cancer Inst, 2006. **98**(17): p. 1183-1192.

77. Foekens, J.A., et al., *Multicenter Validation of a Gene Expression-Based Prognostic Signature in Lymph Node-Negative Primary Breast Cancer.* J Clin Oncol, 2006. **24**(11): p. 1665-1671.

78. Fan, C., et al., *Concordance among Gene-Expression-Based Predictors for Breast Cancer.* N Engl J Med, 2006. **355**(6): p. 560-569.

79. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nature Medicine, 2002. **8**(8): p. 816-824.

80. Yu, Y.P., et al., *Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy.* J Clin Oncol 2004.05.158, 2004. **22**(14): p. 2790-2799.

81. Chen, X., et al., *Gene Expression Patterns in Human Liver Cancers.* Mol. Biol. Cell, 2002. **13**(6): p. 1929-1939.

82. Miller, L.D., et al., *From The Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.* PNAS, 2005. **102**(38): p. 13550-13555.

83. Nelson, P.S., *Predicting prostate cancer behavior using transcript profiles.* The Journal of Urology, 2004. **172**: p. S28-S33.

84. Ein-Dor, L., et al., *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 2005. **21**(2): p. 171-178.

85. Ring, B. and D. Ross, *Microarrays and molecular markers for tumor classification.* Genome Biology, 2002. **3**(5): p. comment2005.1 - 2005.6.

86. Jenssen, T.-K. and E. Hovig, *Gene-expression profiling in breast cancer.* The Lancet, 2005. **365**(9460): p. 634-635.

87. Mukherjee, S., et al., *Estimating Dataset Size Requirements for Classifying DNA Microarray Data.* Journal of Computational Biology, 2003. **10**(2): p. 119-142.

88. Kuo, W.P., et al., *Analysis of matched mRNA measurements from two different microarray technologies.* Bioinformatics, 2002. **18**(3): p. 405-412.

89. Mah, N., et al., *A comparison of oligonucleotide and cDNA-based microarray systems.* Physiol. Genomics, 2004. **16**(3): p. 361-370.

90. Tan, P.K., et al., *Evaluation of gene expression measurements from commercial microarray platforms.* Nucl. Acids Res., 2003. **31**(19): p. 5676-5684.

91. Woo, Y., et al., *A Comparison of cDNA, Oligonucleotide, and Affymetrix GeneChip Gene Expression Microarray Platforms.* J Biomol Tech, 2004. **15**(4): p. 276-284.

92. Zhu, B., et al., *Comparison of gene expression measurements from cDNA and 60-mer oligonucleotide microarrays.* Genomics, 2005. **85**(6): p. 657-665.

93. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nat Meth, 2005. **2**(5): p. 345-350.

94. Wang, H., et al., *A study of inter-lab and inter-platform agreement of DNA microarray data.* BMC Genomics, 2005. **6**(1): p. 71.

95. Nimgaonkar, A., et al., *Reproducibility of gene expression across generations of Affymetrix microarrays.* BMC Bioinformatics, 2003. **4**(1): p. 27.

96. Ghosh, D., et al., *Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.* Functional & Integrative Genomics, 2003. **3**: p. 180-188.

97. Rhodes, D.R., et al., *Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer.* Cancer Res, 2002. **62**(15): p. 4427-4433.

98. Choi, J.K., et al., *Combining multiple microarray studies and modeling interstudy variation.* Bioinformatics, 2003. **19**(90001): p. 84i-90.

99. Shen, R., D. Ghosh, and A. Chinnaiyan, *Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data.* BMC Genomics, 2004. **5**(1): p. 94.

100. Jiang, H., et al., *Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes.* BMC Bioinformatics, 2004. **5**(1): p. 81.

101. Rhodes, D.R., et al., *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* PNAS, 2004. **101**(25): p. 9309-9314.

102. Bloom, G., et al., *Multi-Platform, Multi-Site, Microarray-Based Human Tumor Classification.* Am J Pathol, 2004. **164**(1): p. 9-16.

103. Stevens, J. and R.W. Doerge, *Combining Affymetrix microarray results.* BMC Bioinformatics, 2005. **6**(1): p. 57.

104. Wang, J., et al., *Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies.* Bioinformatics, 2004. **20**(17): p. 3166-3178.

105. Choi, J.K., et al., *Integrative analysis of multiple gene expression profiles applied to liver cancer study.* FEBS Letters, 2004. **565**(1-3): p. 93-100.

106. Hu, P., C. Greenwood, and J. Beyene, *Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models.* BMC Bioinformatics, 2005. **6**(1): p. 128.

107. Zhou, X.J., et al., *Functional annotation and network reconstruction through cross-platform integration of microarray data.* Nature Biotechnology, 2005. **23**(2): p. 238-243.

108. Warnat, P., R. Eils, and B. Brors, *Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.* BMC Bioinformatics, 2005. **6**(1): p. 265.

109. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology.* Nucl. Acids Res., 2003. **31**(1): p. 28-33.

110. Xu, L., et al., *Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data.* Bioinformatics, 2005. **21**(20): p. 3905-3911.

111. Frank, E. and I.H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*. 1999: Morgan Kaufmann.

112. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* PNAS, 1999. **96**(12): p. 6745-6750.

113. Gordon, G.J., et al., *Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma.* Cancer Res, 2002. **62**(17): p. 4963-4967.

114. Jemal, A., et al., *Cancer Statistics, 2006.* CA Cancer J Clin, 2006. **56**(2): p. 106-130.

115. Carter, H.B. and W.B. Isaacs, *Improved Biomarkers for Prostate Cancer: A Definite Need.* J Natl Cancer Inst, 2004. **96**(11): p. 813-815.

116. Stamey, T.A., Caldwell, M., McNeal, J. E., Nolley, R., Hemenez, M., Downs, J., *The prostate specific antigen era in the United States is over for prostate cancer: what happened in the last 20 years?* The Journal of Urology, 2004. **172**: p. 1297-1301.

117. Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**(2): p. 203.

118. Magee, J.A., et al., *Expression profiling reveals hepsin overexpression in prostate cancer.* Cancer Research, 2001. **61**: p. 5692-5696.

119. Luo, J., et al., *Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling.* Cancer Research, 2001. **61**: p. 4683-4688.

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

121. Klezovitch, O., et al., *Hepsin promotes prostate cancer progression and metastasis.* Cancer Cell, 2004. **6**(2): p. 185.

122. Ni, Z., et al., *Selective activation of members of the signal transducers and activators of transcription family in prostate carcinoma.* Journal of Urology, 2002. **167**: p. 1859-1862.

123. Rhodes, D.R., Yu, J., Sharnker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., Chinnaiyan, A.M., *ONCOMINE: A Cancer Microarray Database and Integrated Data-mining Platform.* Neoplasia, 2004. **6**(1): p. 1-6.

124. LaTulippe, E., et al., *Comprehensive Gene Expression Analysis of Prostate Cancer Reveals Distinct Transcriptional Programs Associated with Metastatic Disease.* Cancer Res, 2002. **62**(15): p. 4499-4506.

125. Yauk, C.L., et al., *Comprehensive comparison of six microarray technologies.* Nucl. Acids Res., 2004. **32**(15): p. e124.

126. Yuen, T., et al., *Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays.* Nucl. Acids Res., 2002. **30**(10): p. e48.

127. Stamey, T.A., et al., *Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia.* Journal of Urology, 2001. **166**: p. 2171-2177.

128. O'Shea, J.J., M. Gadina, and R.D. Schreiber, *Cytokine signaling in 2002: new surprises in the Jak/Stat pathway.* Cell, 2002. **109**: p. S121-S131.

129. Calo, V., et al., *STAT proteins: from normal control of cellular events to tumorigenesis.* Journal of Cellular Physiology, 2003. **197**: p. 157-168.

130. Sakakura, C., et al., *Differential gene expression profiles of gastric cancer cells estabilshed from primary tumour and malignant ascites.* British Journal of Cancer, 2002. **87**: p. 1153-1161.

131. Dunn, G.P., et al., *Cancer immunoediting: from immunosurveillance to tumor escape.* Nature Immunology, 2002. **3**(11): p. 991-998.

132. Terabe, M., et al., *NKT cell-mediated repression of tumor immunosurveillance by IL-13 and the IL-4R-STAT6 pathway.* Nature Immunology, 2000. **1**(6): p. 515-520.

133. Catalona, W.J., Smith, D. S., Ornstein, D. K., *Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination: Enhancement of specificity with free PSA measurement.* JAMA, 1997. **277**(18): p. 1452-1455.

134. Brawer, M.K., *Prostate-specific antigen: Current status.* CA A Cancer Journal for Clinicians, 1999. **49**(5): p. 264-281.

135. Yagi, T., et al., *Identification of a gene expression signature associated with pediatric AML prognosis.* Blood, 2003. **102**(5): p. 1849-1856.

136. Iacobuzio-Donahue, C.A., et al., *Exploration of Global Gene Expression Patterns in Pancreatic Adenocarcinoma Using cDNA Microarrays.* Am J Pathol, 2003. **162**(4): p. 1151-1162.

137. Watson, M.A., et al., *Molecular Characterization of Human Meningiomas by Gene Expression Profiling Using High-Density Oligonucleotide Microarrays.* Am J Pathol, 2002. **161**(2): p. 665-672.

138. Higgins, J.P.T., et al., *Gene Expression Patterns in Renal Cell Carcinoma Assessed by Complementary DNA Microarray.* Am J Pathol, 2003. **162**(3): p. 925-932.

139. Yang, X., S. Bentink, and R. Spang, *Detecting Common Gene Expression Patterns in Multiple Cancer Outcome Entities*. Biomedical Microdevices, 2005. **7**(3): p. 247-251.

140. Segal, E., et al., *A module map showing conditional activity of expression modules in cancer.* Nat Genet, 2004. **36**(10): p. 1090.

141. Hippo, Y., et al., *Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays.* Cancer Res, 2002. **62**(1): p. 233-240.

142. Hsiao, L.-L., et al., *A compendium of gene expression in normal human tissues.* Physiol. Genomics, 2001. **7**(2): p. 97-104.

143. Lancaster, J.M., et al., *Gene expression patterns that characterize advanced stage serous ovarian cancers.* Journal of the Society for Gynecologic Investigation, 2004. **11**(1): p. 51-59.

144. Logsdon, C.D., et al., *Molecular Profiling of Pancreatic Adenocarcinoma and Chronic Pancreatitis Identifies Multiple Genes Differentially Regulated in Pancreatic Cancer.* Cancer Res, 2003. **63**(10): p. 2649-2657.

145. Pomeroy, S.L., et al., *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**: p. 436-442.

146. Quade, B.J., G.L. Mutter, and C.C. Morton, *Comparision of Gene expression in Uterine Smooth Muscle Tumors*. 2003, Gene Expression Omnibus: GSE764.

147. Rickman, D.S., et al., *Distinctive Molecular Profiles of High-Grade and Low-Grade Gliomas Based on Oligonucleotide Microarray Analysis.* Cancer Res, 2001. **61**(18): p. 6885-6891.

148. Zhan, F., et al., *Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells.* Blood, 2002. **99**(5): p. 1745-1757.

149. Cromer, A., et al., *Idenfification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis.* Oncogene, 2003. **23**(14): p. 2484-2498.

150. Dehan, E. and N. Kaminski, *Non Small Cell Lung Cancer*. 2004, Gene Expression Onmibus: GSE1987.

151. Frierson, H.F., Jr., et al., *Large Scale Molecular Analysis Identifies Genes with Altered Expression in Salivary Adenoid Cystic Carcinoma.* Am J Pathol, 2002. **161**(4): p. 1315-1323.

152. Giordano, T.J., et al., *Distinct Transcriptional Profiles of Adrenocortical Tumors Uncovered by DNA Microarray Analysis.* Am J Pathol, 2003. **162**(2): p. 521-531.

153. Gutmann, D.H., et al., *Comparative Gene Expression Profile Analysis of Neurofibromatosis 1-associated and Sporadic Pilocytic Astrocytomas.* Cancer Res, 2002. **62**(7): p. 2085-2091.

154. Huang, Y., et al., *Gene expression in papillary thyroid carcinoma reveals highly consistent profiles.* PNAS, 2001. **98**(26): p. 15044-15049.

155. Shai, R., et al., *Gene expression profiling identifies molecular subtypes of gliomas.* Oncogene, 2003. **22**(31): p. 4918-4923.

156. Stearman, R.S., et al., *Analysis of Orthologous Gene Expression between Human Pulmonary Adenocarcinoma and a Carcinogen-Induced Murine Model.* Am J Pathol, 2005. **167**(6): p. 1763-1775.

157. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes.* PNAS, 2002. **99**(7): p. 4465-4470.

158. Welle, S., A. Brooks, and C. Thornton, *Computational method for reducing variance with Affymetrix microarrays.* BMC Bioinformatics, 2002. **3**(1): p. 23.

159. Yanai, I., et al., *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.* Bioinformatics, 2005. **21**(5): p. 650-659.

160. Yu, Y.P., et al., *Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy.* J Clin Oncol, 2004. **22**(14): p. 2790-2799.

161. Gordon, G.J., *Malignant pleural mesothelioma*. 2005, Gene Expression Omnibus: GSE2549.

162. Hoffman, P.J., et al., *Uterine Fibroid and Normal Myometrial Expression Profiles*. 2003, Gene Expression Omnibus: GSE593.

163. Lenburg, M., et al., *Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data.* BMC Cancer, 2003. **3**(1): p. 31.

164. Talantov, D., et al., *Novel Genes Associated with Malignant Melanoma but not Benign Melanocytic Lesions.* Clin Cancer Res %R 10.1158/1078-0432.CCR-05-0683, 2005. **11**(20): p. 7234-7242.

165. Wachi, S., K. Yoneda, and R. Wu, *Squamous Lung Cancer*. 2005, Gene Expression Omnibus: GSE3268.

166. Yoon, S.S., et al., *Gene expression of human soft tissue sarcoma*. 2005, Gene Expression Onmibus: GSE2719.

167. Michiels, S., S. Koscielny, and C. Hill, *Prediction of canceroutcome with microarrays: a multiple random validation strategy.* Lancet, 2005. **365**(9458): p. 488-492.

168. Pavlidis, P. and W.S. Noble, *Matrix2png: a utility for visualizing matrix data.* Bioinformatics, 2003. **19**(2): p. 295-296.

169. Basil, C.F., et al., *Common Cancer Biomarkers.* Cancer Res, 2006. **66**(6): p. 2953-2961.

170. Eifel, P., et al., *National Institutes of Health Consensus Development Conference Statement: Adjuvant Therapy for Breast Cancer, November 1-3, 2000.* J Natl Cancer Inst, 2001. **93**(13): p. 979-989.

171. Goldhirsch, A., et al., *Meeting Highlights: International Expert Consensus on the Primary Therapy of Early Breast Cancer 2005.* Ann Oncol, 2005. **16**(10): p. 1569-1583.

172. Early Breast Cancer Trialists' Collaborative, G., *Polychemotherapy for early breast cancer: an overview of the randomised trials.* The Lancet, 1998. **352**(9132): p. 930-942.

173. Michiels, S., S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* The Lancet, 2005. **365**(9458): p. 488-492.

174. Pawitan, Y., et al., *Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.* Breast Cancer Research, 2005. **7**(6): p. R953 - R964.

175. Liu, H., J. Li, and L. Wong, *Use of extreme patient samples for outcome prediction from gene expression data.* Bioinformatics, 2005. **21**(16): p. 3377-3384.

176. Payton, M., et al., *Deregulation of cyclin E2 expression and associated kinase activity in primary breast tumors.* Oncogene, 2002. **21**(55): p. 8529-8534.

177. Pittman, J., et al., *Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes.* PNAS, 2004. **101**(22): p. 8431-8436.

178. Sun, Y., et al., *Improved breast cancer prognosis through the combination of clinical and genetic markers.* Bioinformatics, 2006: p. btl543.

Vita

EDUCATION

Johns Hopkins University, Baltimore, MD, U.S.A.

Ph.D. in Electrical and Computer Engineering	March 2007
M.S.E. in Applied Mathematics and Statistics	Jan 2005

Zhejiang University, Hangzhou, Zhejiang, P.R.China

M.S. in Biomedical Engineering	Jun 1999
B.S. in Biomedical Engineering	Jun 1996

PUBLICATIONS

Xu, L., Tan, A. C., Naiman, D. Q., Geman, D. and Winslow, R. L. (2005). "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data", Bioinformatics, 21, 3905-3911.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L. and Geman, D. (2005). "Simple decision rules for classifying human cancers from gene expression profiles", Bioinformatics, 21, 3896-3904.

Hinch, R., Greenstein, J. L., Tanskanen, A., Xu, L. and Winslow, R. L. (2004). "A simplified local control model of calcium induced calcium release in cardiac ventricular myocytes", Biophys. J., 87: 3723-3736.

Xu, L., Tan, A. C., Winslow, R. L. (2007). "Robust breast cancer prognostic signature from integration of multiple microarray data", submitted.

Xu, L., Geman, D. and Winslow, R. L. (2006). "Large-scale integration of cancer microarray data identifies a robust common cancer signature", submitted.