3-21-2006

Ranking-Based Methods for Gene Selection in Microarray Data

Li Chen
University of South Florida

Ranking-Based Methods for Gene Selection in Microarray Data

by

Li Chen

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Co-Major Professor: Dmitry B. Goldgof, Ph.D.
Co-Major Professor: Lihua Li, Ph.D.
Lawrence O. Hall, Ph.D.

Date of Approval:
March 21, 2006

Keywords: Parametric Methods, Nonparametric Methods, Classification and Prediction
Simulated Experiment, Biological Application

Acknowledgements

I am grateful to my faculty supervisors, Dr. Dmitry Goldgof and Dr. Lihua Li, for their help and guidance throughout the thesis study. Most of the research work presented in this thesis is taken in H. Lee Moffitt Cancer Center & Research Institute under the grant support from Dr. Lihua Li. I thank Dr. Lawrence Hall for his valuable comments and suggestions. I also appreciate the help of my fellow graduate student George Florence.

I am deeply indebted to my parents, Yongming Chen and Qizhen Zhuang, for their support, encouragement throughout my studies.

Table of Contents

List of Tables

List of Figures

Ranking-Based Methods for Gene Selection in Microarray Data

Li Chen

ABSTRACT

DNA microarrays have been used for the purpose of monitoring expression levels of thousands of genes simultaneously and identifying those genes that are differentially expressed. One of the major goals of microarray data analysis is the detection of differentially expressed genes across two kinds of tissue samples or samples obtained under two experimental conditions. A large number of gene detection methods have been developed and most of them are based on statistical analysis. However the statistical analysis methods have the limitations due to the small sample size and unknown distribution and error structure of microarray data. In this thesis, a study of ranking-based gene selection methods which have weak assumption about the data was done. Three approaches are proposed to integrate the individual ranks to select differentially expressed genes in microarray data. The experiments are implemented on the simulated and biological microarray data, and the results show that ranking-based methods outperform the t-test and SAM in selecting differentially expressed genes, especially when the sample size is small.

Chapter 1

Introduction

DNA microarrays have been used for the purpose of monitoring expression levels of thousands of genes simultaneously and identifying those genes that are differentially expressed. While the main goal of this study is to explore the ranking-based methods to improve the performance of selecting differentially expressed genes, it is essential to have a fundamental understanding of biology and microarray technology to understand gene expression profiles well.

1.1 Background

Proteins are the structural components of cells and tissues and perform many key functions of biological systems. The production of proteins is controlled by genes, which are coded in deoxyribonucleic acid (DNA), common to all cells in one being, and mostly static over one's lifetime (Parmigiani *et al.*, 2003). A gene consists of a specific DNA fragment, and can be interpreted as a construction for a protein. Protein production from genes is explained by the central dogma of molecular biology (Figure 1.1) which includes two principal stages, transcription and translation. First the gene is transcribed into messenger ribonucleic acid, abbreviated as mRNA. Second, the mRNA is translated into a protein. There is huge variation in abundance and efficiency of transcription and translation among different cell type. The distribution is responsible for the appearance

and state of a cell. Ultimately, a cell's role is determined by the proteins it produces, which in turn depend on its expressed genes.

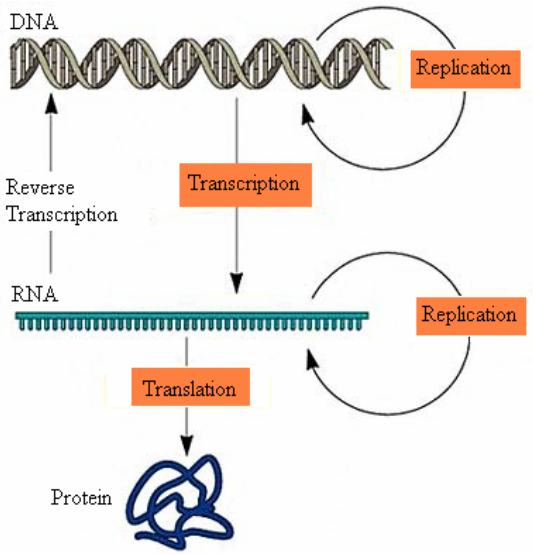


Figure 1.1 The central dogma of molecular biology from http://cats.med.uvm.edu/

Measuring changes in the mRNA levels is one of possible methods to detect differences between cells. Scientists study the kinds and amounts of mRNA produced by a cell to learn which genes are expressed, which in turn provides insights into how the cell responds to its changing needs. There are various methods for detecting and quantifying the amount of mRNA. Traditional methods in molecular biology generally have the limits that the throughput is very limited and the "whole picture" of gene function is hard to obtain. A new technology, called DNA microarray, has been of interest among biologists in the past several years. This technology can monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of gene.

1.2 Microarray technology

DNA microarrays, or DNA chips are fabricated by high-speed robotics, generally on glass, for which probes with known identity are used to determine complementary binding, thus it allows massively parallel gene expression and gene discovery studies (Parmigiani *et al.*, 2003). An experiment with a single DNA chip can generate thousands of gene expression levels simultaneously.

The two most prevalent microarray technologies are cDNA microarrays and high-density oligonucleotide arrays. The differences between them are in the manner of placement of the DNA sequences on the array and in the length of these sequences during hybridization. Accordingly, the experimental approach and the data preprocessing differ as well.

1.2.1 cDNA microarrays

In cDNA arrays, mRNA from two different biological samples (i.e. a sample of interest and a control sample) is reverse transcribed into cDNA, labeled with red and green fluorescent dyes, and distributed on the microarray. Then the cDNA competitively hybridizes to the corresponding DNA clones. Finally the remaining material is washed off and the amount of chemically bound cDNA is quantified by the intensity of the fluorescence in each spot measured by a laser scanner. This procedure was described in Schena *et al.* (1995) and DeRisi *et al.* (1997). Higher fluorescence indicates higher amounts of hybridized cDNA, which in turn indicates higher gene expression in the sample. A spot consists of a number of pixels and needs to be segmented and summarized by image analysis algorithms.

For each location on the array, a typical output consists of at least four quantities, one of each color for both the spot and the background. Sometimes these are summarized by measures of quality of the spot or the pixel intensity variability. The use of two samples in cDNA allows for measurement of relative gene expression across two sources of cDNA. Therefore it is less sensitive to the variable amount of spotted DNA, as well as other experimental variation in this way. Although this ratio is critical, there is relevant information in all four of the quantities above.

1.2.2 High-density oligonucleotide arrays

The oligonucleotide arrays, most widely used by the Affymetrix GenChip™ , are a new approach in microarray technology, based on hybridization to small, high-density arrays containing tens of thousands of synthetic oligonucleotides (Lockhart *et al.*, 1996). Compared to cDNA, its main characteristics are 1) only one biological interest sample is fluorescently labeled and hybridized to the microarray. There is no competitive hybridization, and 2) the expression of each gene is measured by comparing hybridization of the sample mRNA to a set of probes, which is composed of 11-20 pairs of oligonucleotides and each of length 25 base pairs. The first type of probe in each pair is the perfect match (PM) which exactly corresponds to the gene sequence, whereas the second is the mismatch (MM), created by changing the middle (the 13th) base of the original sequence. The idea of this construction is to provide a control mechanism for random variation and cross-hybridization.

An RNA sample is prepared, labeled with a fluorescent dye, and hybridized to an array. Arrays are then scanned, and images are produced and analyzed to obtain a fluorescence intensity value for each probe, measuring hybridization for the

corresponding oligonucleotide. For each gene, or probe set, the typical output consists of two vectors of intensities, one for PMs and one for MMs. Specifically, the PM and MM probe intensities for each probe set must be combined together to produce a summary value. Two common methods are used to produce the gene expression value. One is MAS 5.0 (Affymetrix, 2001) proposed by Affymetrix to remedy the drawbacks of being noisy for low intensity and giving negative values in the traditional average difference (AvDiff) algorithm (Affymetrix, 1999). Another method is the Robust Multi-chip Average (RMA) (Irizarry *et al.*, 2003), which consists of three steps: a background adjustment, quantile normalization (Bolstad *et al.*, 2003) and finally summarization.

1.3 Motivation and goals

DNA microarray technology makes it possible to understand the processes within the cell and to learn the functional units in the genome by analyzing the gene expression microarray data. Usually DNA microarray technology generates thousands of genes simultaneously for each sample. One of the major goals of microarray data analysis is the detection of differentially expressed genes across two kinds of tissue samples or samples obtained under two experimental conditions in this high-dimensional gene space. A large number of gene detection methods have been developed and most of them are based on statistical analysis (Parmigiani *et al.*, 2003). However, Due to the time-consuming experimental protocol, the cost and the often limited access to biological tissues of interest, a large number of microarray experiments are performed on a small number of samples only. Even in some clinical research where a large number of samples can be obtained, the number of samples in the interesting subclass which has similar clinical

information is still small due to the biological variance. These lead to a big challenge in significant gene selection because traditional statistical methods have the limitation due to the basic assumption on sample size. Another challenge is the lack of understanding of the distribution and error structure of microarray data, and therefore a statistically significant difference in the expression level may not imply the occurrence of any difference of biological or clinical significance, which increases the difficulty of detecting differentially expressed genes.

In this thesis, a new method based on ranking information is studied in order to improve performance on selecting differentially expressed genes. This method has weak assumptions about the data because of its non-parametric nature. The thesis is organized as follows:

Chapter 2 introduces the general gene selection methods. A brief overview of parametric and non-parametric statistical methods is given, as well as ranking-based methods. The chapter concludes with a discussion of advantages and disadvantages of different gene selection methods.

Chapter 3 introduces the theory of a novel ranking-based method. A detailed description of ranking-based methods is presented including three approaches of integrating individual ranks.

Chapter 4 describes the experiments conducted on the simulated microarray data to evaluate how well the proposed ranking-based gene selection methods are. Various experiments under different conditions are implemented. The discussion and summary of the results are presented in the chapter.

Chapter 5 is the application of ranking-based methods on the real biological microarray data. The applications on differentially expressed gene selection and classification/prediction are conducted on several benchmark datasets using proposed ranking-based methods and other traditional gene selection methods.

Finally, a discussion on the proposed approaches is presented in Chapter 6, as well as the further work, followed by the conclusions of the study.

Chapter 2

Overview of Gene Selection Methods

The earliest gene selection approach used a simple fold-change criterion to detect the differentially expressed genes, but it is known to be unreliable because statistical variability was not taken into account (Chen *et al.*, 1997). Since then, a large number of sophisticated gene detection methods have been proposed and most of them are based on statistical analysis. In general, there are two types of statistical tests: parametric and non-parametric tests. A parametric test assumes the data to be known or follow a certain distribution, whereas a non-parametric test does not make such an assumption. Ranking-based methods appear to be an alternative approach which are concerned with the rank information among the genes rather than the actual gene expression levels.

2.1 Parametric methods

The two-sample t-test (Devore and Peck, 1997) is a traditional parametric hypothesis testing method for the selection of differentially expressed genes. Under the normality assumption of the expression levels, the t-statistic follows a t-distribution in a standard t-test. The probability value (p-value) of the t-statistic for each gene expression is the chance of getting the t-statistic as or more extreme than the observed one, under the hypothesis of no differential expression. A small p-value indicates that the hypothesis of no differential expression is not true and the gene is differentially expressed.

We can calculate the *T*-value for each gene using following equation.

$$T_i = \frac{|m_{i1} - m_{i2}|}{\sqrt{\left(\frac{(n_1 - 1)s_{i1}^2 + (n_2 - 1)s_{i2}^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (2.1)$$

Where n_1 and n_2 are the number of cases in two classes and i is the index of the gene; m_{i1} and m_{i2} are the mean values and s_{i1} and s_{i2} are the standard deviations respectively for ith gene. The p-value can be identified by *T*-value and a gene is considered as differentially expressed with significance when its p-value is less than some threshold, usually chosen as 0.01 or 0.05.

The performance of the t-test depends on the sample size and how well the assumption of normality of expression intensities was met. Since the level of replication within treatments is often low for microarray experiments and the expression intensities may not be normally or even symmetrically distributed, the performance of the t-test is usually poor. For example, in a tumor-versus-normal comparison every tumor sample may contain some amount of normal tissue which leads to non-symmetrical distribution of microarray data.

Instead of under a single normality assumption on microarray data, a set of mixture model approaches are proposed based on the assumption that the observed intensity values are distributed as a mixture of two distributions. Newton *et al.* (2001) modeled the expression levels in the two channels of a cDNA microarray with a Gamma-Gamma hierarchical model and used Bayesian and empirical Bayesian techniques to identify differentially expressed genes in two tissue types.

McLachlan *et al.* (2002) proposed a mixture of t-distributions model. This model estimates the distribution of log likelihood ratio statistic for testing one versus two

components hypothesis in the mixture model for each gene considered individually, using a parametric boostrap approach.

Kendziorshi *et al.* (2003) developed the Parametric Empirical Bayes (EBarrays) approach to compute the posterior probability under one of the two proposed hierarchical model assumption of the expression levels, one based on the assumption of Gamma distributed measurements and the other based on log-normally distributed measurements. In EBarrays the expression data are assumed to come from the specific parametric model and a constant coefficient of variation of expression levels is assumed.

Cox and Wong (2004) proposed a mixture model for large numbers of tests, in which the distribution of the test statistic was modeled as a mixture of two normal distributions, one corresponding to the null hypothesis being true, and the other to its being false.

Dean and Raftery (2005) used a normal-uniform mixture model. Non-differentially expressed genes are modeled with a Gaussian density since they have a true log ratio of zero. Differentially expressed genes are modeled as a uniform distribution since these genes can be viewed as outliers from the main distribution of non-differentially expressed genes. The whole data are modeled by a weighted mixture of these densities, where the weights correspond to the prior probabilities of being in each of the two groups.

Currently, most statistical analysis of microarray data depends on the assumption that the data is normally distributed with variances not dependent on the mean of the data (Pan, 2002). More and more studies suggest that microarray data violate these assumptions dramatically. Several alternative models have been proposed for the

measurement error in microarray data by Ideker *et al.* (2001) and Durbin *et al.* (2002). These models all reflect the observation that the variance of expression data of a gene increases with its mean. Durbin *et al.* (2002) proposed a two-component model. Under this model, the measured expression intensity is a linear combination of a normal random variable and a lognormal random variable. The normal component dominates at low expression levels, while the lognormal component dominates at high expression levels. Based on this model, Wang *et al.* (2004) derived a generalized likelihood ratio (GLR) test to identify differentially expressed genes from microarray data.

2.2 Nonparametric methods

The assumptions relating to sample size and distribution in parametric statistical methods are often not met by microarray data. In this case, non-parametric analyses may be more appropriate to employ.

The Wilcoxon test is a nonparametric method which is commonly used to check if there is a difference between Treatment and Control subjects. The Wilcoxon rank sum test is based on the sum W_s of the ranks of the observations in one of the groups (Lehmann, 1975). In the usual sense, small or large values of W_s correspond respectively to under expression or over expression. In other word, genes with small or large ranks contribute most to the difference between two mutation types. The p-value for each gene was derived using the test statistics of the Wilcoxon distribution. However the Wilcoxon test does not use all the information available for all the genes from microarray data thus may have low power to detect differential gene expression (Thomas *et al.*, 2001; Pan, 2002).

Efron *et al.* (2001) developed an Empirical Bayes approach that calculates a posteriori probabilities of effect for the individual genes. It avoids parametric assumptions about gene expression by using a simple nonparametric mixture prior to model the population of affected and unaffected genes. After a long series of preprocessing steps, each gene yields a one-dimensional test statistic whose marginal distribution turns out to be known and whose null distribution (i.e., on equivalent expression) can be nonparametrically estimated.

Tusher *et al.* (2001) proposed a nonparametric statistical method, Significance Analysis of Microarrays (SAM), which identifies genes with statistically significant changes in expression by assimilating information from a set of gene-specific t-tests. Compared to the t-test, it first performs permutation and computes test statistics for each permutation, then adds a s_0 term to deal with cases when the variance gets too close to zero. The process is as follows:

$$d_i = \frac{m_{i1} - m_{i2}}{s_i + s_0}, \qquad s_i = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1 - 1)s_{i2}^2 + (n_2 - 1)s_{i2}^2}{n_1 + n_2 - 2}} \tag{2.2}$$

Here, d_i is the Score, s_i is Standard Deviation. Others symbols are same as in Equation (2.1). Specifically, s_0 is chosen as the percentile of the s_i values that makes the coefficient of variation of d_i approximately constant as a function of s_i. This has the added effect of dampening large values of d_i that arise from genes whose expression is near zero.

Dudoit *et al.* (2002) proposed a nonparametric t-test with family-wise error rate (FWER), the probability that at least one of the true null hypotheses is rejected, for multiple comparisons by using a permutation analysis on Welsh's t-statistics.

Pan *et al.* (2001) proposed a nonparametric mixture model method (MMM) that uses a mixture of normal distribution as to estimate each of the two distributions of the test statistics and the null statistics. The mixture model is fitted by maximum likelihood using the expectation-maximization (EM) algorithm after running several times. A comparison of these two distributions by means of a likelihood ratio test, or simply using the tail distribution of the null statistic, can identify genes with significantly changed expression.

One of the limitations of MMM is starting the EM algorithm with random values as the parameters of the normal basis functions to estimate distributions makes the results depend highly on the exact initialization, and always causes variations in the results. In addition, the results of the MMM may not be repeatable when dealing with a small number of replicates. Zhao and Pan (2003) proposed a modified method to construct the test and null statistics. Najarian *et al.* (2004) proposed a novel mixture model method which used K-means clustering method in estimating the distributions to improve the repeatability, and robustness of the mixture model.

Grant *et al.* (2002) applied a nonparametric method of controlling the false discovery rate (FDR), the expected proportion of rejected null hypotheses that are true (Benjamini and Hochberg 1995), and used the permutation method of Dudoit *et al.* (2000) to control the family-wise error rate.

2.3 Ranking-based methods

A set of ranking-based methods have been proposed as alternative gene selection methods. A Ranking-based method evaluates the rank information among genes rather

than the actual gene expression levels. It appears to be a robust choice for microarray data, which are often nonnormal and contain outliers. Zimmerman and Zumbo (1993) demonstrated that the Wilcoxon rank sum test is more powerful than the t-test when outliers (unusual extreme data values) are present. The Wilcoxon rank sum test first ranks gene expression values for each gene across all experiments, and then tests for the equality of means of the two ranked samples.

Park *et al*. (2001) scored genes based on the number of permutations of expression values required to make that gene into a perfectly discriminating marker, where all high expression values belong to one group of experiments and all low expression values belong to the other group. Significance of scores was assessed based on column permutations of the data set and comparison of the distribution of scores from permuted data to that of the original data.

Neuhauser and Senske (2004) introduced the Baumgartner-Weiβ-Schindler test based on ranks for the detection of differentially expressed genes in replicated microarray experiments. The combined score B for each gene is calculated by the nonparametric statistic introduced by Baumgarner et al. (1998). The Baumgartner-Weiβ-Schindler test is less conservative than the Wilcoxon test and more powerful, because the exact permutation distribution of B is less discrete than that of the rank sum.

Martin *et al*. (2004) proposed a new analysis method, Rank Difference Analysis of Microarrays (RDAM), which replaced raw signal by its rank, expressed on a 0-100 scale as a normalizing procedure. Rank difference between individual experiment points was calculated as the variation. Finally RDAM estimated the total number of truly varying genes by assigning a p-value to each gene variation.

Breitling *et al*. (2004) implemented a ranking-based test statistic, RankProducts (RP), as a non-parametric method for detecting differentially expressed genes in microarray experiments. Individual ranks for each comparison under two different conditions are calculated and rank product is used to integrate the individual ranks for each gene.

2.4 Discussion

For all above parametric and non-parametric gene detection methods, it is important to understand the underlying models of microarray data and then apply the proper methods to find the significant differentially expressed genes. Parametric statistical methods are mostly based on some known assumptions to estimate the parameters, such as normal distribution, gamma distribution or mixture model with different distributions. In nonparametric statistical methods, the basic idea rests on constructing a null statistic such that its distribution is the same as the null distribution of the test statistic, and thus the null distribution of the test statistic can be estimated using the constructed null statistics.

Statistical gene selection methods including parametric and nonparametric methods are widely used in microarray gene selection applications. They perform well in most cases because they treat the genes as arising from some population which takes full advantage of the level of information sharing among genes. (Parmigiani *et al*., 2003)

One of the problems in these statistical methods is the sample sizes which are usually small in microarray experiments. In this case, the statistical assumption on the sample size is hard to meet. For example, the asymptotic (i.e. large sample) justification

for the t-tests is not applicable and the normality assumption may not hold in small sample size. In most nonparametric methods permutation is used to construct null statistic in the case of small sample size, however it is known to overfit the data if the sample size is too small (Pan, 2002). Figure 2.1 shows performance of the simulation study on differentially expression gene selection using the t-test and SAM respectively at different number of sample sizes. The experiment will be described in detail in Chapter 4. From the figure we can see that as the sample size decreases, the performance of the t-test and SAM drop down consistently, especially when the sample size is less than 10.

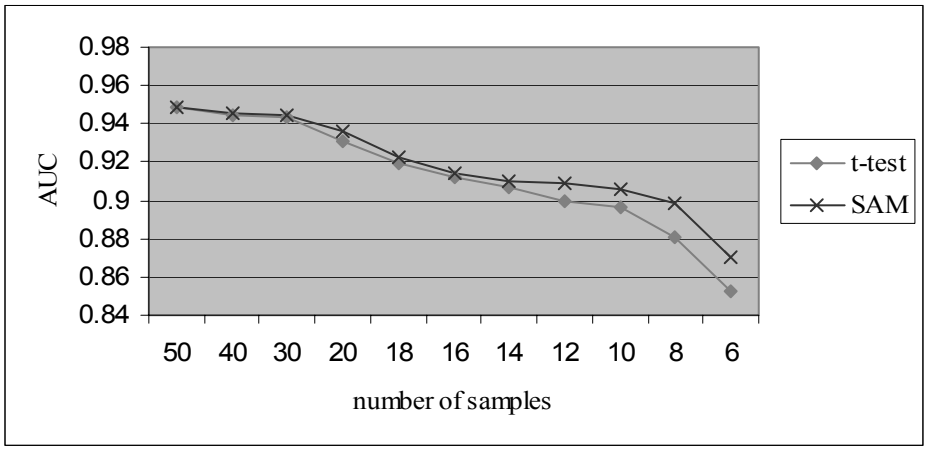


Figure 2.1 Performance of gene selection methods of t-test and SAM at different number of sample sizes

Another potential problem is that it is hard to know the real distribution of microarray data and therefore the accurate estimation of parameters of interest, even if the sample size is large enough. One of the reasons is that there are a large number of potential sources of random and systematic measurement error in microarray studies. The performance of parametric statistical gene selection methods strongly depends on how well the expression intensities of microarray data fit the assumed model. Nonparametric

methods have much weaker modeling assumptions, but they still have the assumption that random errors have symmetric distributions, and after proper standardization, the random errors from all the genes have a common distribution.

In many situations, statistical methods will detect some significant differentially expressed genes with small fold-changes. However, it is usually unlikely that genes with very small fold-changes have any significant biological difference even if they are statistically significant, which results in high false positive in differentially expressed genes selection for a large number of methods.

Ranking-based methods are more intuitive than statistical methods because they use rank information rather than actual gene expression values. There is no specific distribution to be assumed for ranking-based methods. They are considered to be robust to outliers, normalization schemes, and systematic errors such as chip-to-chip variation. (Park *et al*., 2001). However, there may be a slight loss of information and they can be conservative and computer-intensive (Neuhauser and Senske 2004).

Most of the current gene selection methods in use today evaluate each gene in isolation and ignore the gene to gene correlations (Piatetsky-Shapiro and Tamayo, 2003). Clearly, this was done to keep the formulation of the methodology simple and differentially expressed genes are selected one by one. From a biological perspective, however, genes with similar biological functions are often co-regulated. The exact correlation structure of all the gene expression levels could be extremely complicated and could be largely unknown in most problems. Perhaps a compromise could be reached by some multivariate statistical techniques such as the multivariate t-test, clustering, Principal Component Analysis (PCA) etc (Darghici, 2003). However, the difficulties of

these methods are how to decide the total number of differentially expressed genes and the pattern among the differentially expressed genes.

Chapter 3

Ranking-Based Gene Selection Methods

3.1 Introduction

It has been discussed that there are several limitations of statistical methods in analyzing microarray data due to the small sample size and unknown distribution and error structure of microarray data. A statistically significant difference in the expression level may not imply the occurrence of any difference of biological or clinical significance. Hence we explore the ranking-based gene selection methods. The idea of a ranking-based gene selection method is based on the observation that genes with very small fold-changes have little significant biological or clinical difference even if they are significant statistically. Ranking-based methods do not rely on estimating the measurement variance for each single gene and thus are particularly useful when this estimate becomes unreliable due to a small number of samples.

Among the ranking-based methods we discussed in chapter 2, Breitling *et al.* (2004) implemented a ranking-based test statistic, RankProducts (RP), as a non-parametric method for detecting differentially expressed genes in microarray experiments. The most prominent difference between Breitling's ranking-based method and other ranking-based methods is that rank information in RP indicates the correlation among genes, while other ranking-based methods, such as the Wilcoxon rank sum test, consider each gene independently and the rank information of each gene only indicates the

relationship across all experiments for this gene. However, there are many problems which need to be solved. First, RP is only one of the approaches to integrate the individual ranks. Other possible approaches using different integrating criteria are not considered, which may have advantages under some conditions. Second, little systematic understanding of the ranking-based methods is achieved so far. Third, the performance of ranking-based methods on biological experiments is hard to evaluate because no truth about differentially expressed genes is known in most biological experiments. It is necessary to design the experiments to evaluate the performance of ranking-based gene selection methods under a variety of conditions based on simulated and biological microarray data. Therefore, the theory of ranking-based methods for gene selection is analyzed systematically, and several integrating approaches are proposed in this study.

3.2 Gene selection with ranking information

Compared to previous statistical gene selection methods, the assumptions made for ranking-based methods are relatively weak. They are that (1) relevant expression changes affect only a minority of genes, (2) measurements are independent between replicate arrays, (3) most changes in expression are independent of each other, and (4) measurement variance is about equal for all genes. Ranking-based methods consist of rankings of genes and integration of individual ranks.

3.2.1 Ranking of genes

The ranking of genes is derived from the simple fold-change (FC) criterion. For a specific gene *g*, we define its up-regulated rank r_g^{up} in comparing two samples as the position of the gene g in the list of genes sorted by decreasing FC. Similarly, down-

regulated rank r_g^{down} as the position of the gene *g* in the list of genes sorted by increasing FC. r_g^{up}=1 means gene *g* is the most strongly up-regulated gene and r_g^{down}=1 means gene *g* is the most strongly down-regulated gene. For single-channel arrays, e.g., Affymetrix GeneChip arrays, the rank is taken over all possible pairwise comparisons. A set of rank values are obtained for each gene on each comparison.

Consider one microarray dataset which has two conditions represented as A and B, such as treated vs. untreated samples, diseased vs. normal tissues, there are *M* samples in condition A, and *N* samples in condition B. We will generate two ranking tables, respectively, one for up-regulated and the other for down-regulated. Each table has *M*×*N* comparisons. $r_g^{up}(i, j)$ represents the up-regulated ranking of gene *g* in the comparison of i^{th} sample in condition A and j^{th} sample in condition B. Similarly, $r_g^{down}(i, j)$ represents the down-regulated ranking of gene *g* in the comparison of i^{th} sample in condition A and j^{th} sample in condition B.

3.2.2 Integration of individual ranks

Individual ranks give out the information of expression level change on each comparison pair. For a typical experiment, a set of sample comparisons will be taken. Some criteria are used to integrate the individual rank information for the measurement of the significance of the genes.

3.2.2.1 Rank products (RP)

The rank products combining approach has been discussed in detail (Breitling *et al.*, 2004). Here we present a modified rank product approach which combines the up- and down-regulated genes using a ranking measure score.

For a microarray dataset we defined above, the up- and down-regulated RP values for a specific gene *g* are calculated by equation (3.1) and (3.2) respectively.

$$RP^{up}(g) = \prod_{i \in M, j \in N} r_g^{up}(i, j) \tag{3.1}$$

$$RP^{down}(g) = \prod_{i \in M, j \in N} r_g^{down}(i, j) \tag{3.2}$$

The smaller the RP value, the more likely the gene is significant differentially expressed. We combine the RP values of each gene under the up- and down-regulation RP values. Each gene has two RP values: one is its up-regulated value and another is its down-regulated value. The smaller one is assigned as the ranking measured score of the gene. Genes are ranked according to their ranking RP values. Using this method, the most up-regulated and down-regulated genes are ranked at the top with small RP values, which means that they are more likely significant differentially expressed. Finally we can get the sorted genes with their RP values.

After getting the ranked genes, we need to determine the significant levels to know how many genes are truly significant differentially expressed. The normalized rank product values which are divided by the maximum value in the gene measured score could be interpreted as p-values, as the measurement describes the probability of observing gene *g* at a certain rank without significant change (Breitling *et al.*, 2004). Note that this interpretation is valid when all ranks are equally likely, which is the case when the replicates are independent, genes have equal variance and none of them are differentially expressed. These are exactly the assumptions described.

3.2.2.2 Rank average/summation (RS)

Since RP can be considered as an average log rank of the differential gene expression, we can use rank average/summation (RS) to evaluate the significance of differentially expressed genes instead of rank products. Similar to RP, the up- and down-regulated RS values for a specific gene *g* are defined in equations (3.3) and (3.4) respectively.

$$RS^{up}(g) = \sum_{i \in M, j \in N} r_g^{up}(i, j) \tag{3.3}$$

$$RS^{down}(g) = \sum_{i \in M, j \in N} r_g^{down}(i, j) \tag{3.4}$$

The smaller the RS value, the more likely the gene is significant differentially expressed. Also, we need to combine the RS values of each gene under up and down regulation RS values. The combining approach is similar to RP except that we use rank summation as the ranking measured score.

In the same way as RP, the significance levels of differentially expressed genes in RS method can also be determined by the ranking measured scores of the genes. It is obvious that genes with the smallest RS values are the most interesting candidates and the biologist can then select some of them for further study.

It is hard to determine the significance levels for most gene selection methods. For example, in many cases a large number of genes will be detected as significantly differentially expressed by the t-test when the significance level is set to be $p<0.05$. However, in many applications, what we need is the information on which ones are the most significant genes in terms of the ranking, instead of a cutoff point of the significance level. For instance, in the application of classification and predication problems using

microarray data, the number of genes is usually specified by some algorithms. In this case, it is enough to have the rank information of genes in selecting the differentially expressed genes that we need.

3.2.2.3 Rank-based committee decision method (RC)

Due to the variation of the biological samples and microarray array analyses, the gene expression value is usually "noisy". Some of them may even be outlying from true values, which may give wrong information if we combine all individual ranks in RP and RS. Here we introduce the committee decision method based on ranks (RC) which can have a certain tolerance to noise and outliers.

The committee decision method (Black 1963) is based on a simple premise that the significant differentially expressed genes should be at the top in most of individual gene ranks among all paired comparisons.

For a comparative study with M samples in condition A and N samples in Condition B, there are a total of $M \times N$ ranking values for each gene. A gene will be selected as differentially expressed by the RC method if it is ranked among the top K_0 ($K_0 < K$, K is the number of genes) in more than C_0 ($C_0 < M \times N$) ranking lists of comparison study. A committee decision method will select a set of differentially expressed genes decided by the two parameters K_0 and C_0. Changing the parameters of K_0 and C_0, we will obtain different set of significant genes. Mathematically, this can be described as follows:

The committee score of gene g in the up-regulated ranking is defined as:

$$RC^{up}(g) = \sum_{i \in M, j \in N} \cup\, r_g^{up}(i, j) \tag{3.5}$$

where

$$\cup r_g^{up}(i,j) = \begin{cases} 1 & r_g^{up}(i,j) \le K_0 \\ 0 & r_g^{up}(i,j) > K_0 \end{cases} \tag{3.6}$$

If $RC^{up}(g)$ is greater than C_0, gene g is considered to be significantly differentially expressed gene under the parameters of K_0 and C_0. The number of selected genes is decided by the parameters of K_0 and C_0. We can get different sets of significant differentially expressed genes under the different parameters of K_0 and C_0.

Similarly, committee decision methods can be applied to choose the differentially expressed genes from down-regulated genes at parameters of K_0 and C_0.

The overall differentially expression genes at parameters of K_0 and C_0 can be obtained by taking the union set of up- and down- regulated genes at the parameters of K_0 and C_0.

The difference between the committee decision method and RP, RS for gene selection is that a ranking committee score in the RC method depends on parameter K_0. Genes fall into two sets: significant and non significant differentially expressed. A gene is decided to be differentially expressed or non-differentially expressed based on C_0. So it is hard to explicitly select a fixed number of genes.

3.3 Discussion

Three ranking-based methods are described in this chapter to integrate the individual ranks. Compared to traditional statistical methods, a relatively weak assumption was made on microarray data. It has the biological intuition that a significant gene has large relevant changes while small changes may have statistical but rarely

biological significance. Ranking-based gene selection methods do not rely on estimating the measurement variance for each single gene and thus are particularly useful when this estimate becomes unreliable due to a low number of samples.

Chapter 4

Evaluation on Simulated Microarray Datasets

In a practical situation, determining differentially expressed genes is one of the most challenging tasks in microarray data analyses partly because it is hard to decide which gene is truly significant differentially expressed. In this case, a set of experiments on simulated microarray data in which differentially expressed genes can be specified was conducted in order to evaluate the performance of proposed ranking-based methods.

4.1 Criteria for evaluation

Receiver Operating Characteristic (ROC) curve (Witten, 2000) is widely used in numerous gene selection methods to evaluate their performance. ROC curves can show how well the method discriminates between true positives and true negatives if truly differentially expressed (DE) genes are known. Specifically, genes are ranked in terms of scores obtained by different gene selection methods, and an ROC curve is a plot of the true positive (TP) rate against the false positive (FP) rate at different cutoff points in a specific ranked gene set. An ROC curve demonstrates several things in a specific DE gene selection method. First, it shows the tradeoff between sensitivity and specificity (i.e. any increase in sensitivity will be accompanied by a decrease in specificity). Second, the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test will be. Similarly, the closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test will be. Finally the area under the

curve (AUC) is a measure of test accuracy, which is a value between 0.5 and 1. The bigger the AUC value, the more accurate the gene selection method is. ROC curves will be used to evaluate the performance of the proposed ranking-based gene selection methods and several typical statistical methods on the simulated microarray datasets in this chapter and one benchmark microarray dataset which have known DE genes in Chapter 5.

4.2 Simulation study

4.2.1 Simulation of microarray gene expression data

Normal distribution has been widely used for gene expression data simulation (Lonnstedt and Speed, 2002; Gottardo 2002; Bickel 2004). Non-DE genes are drawn from normal distribution independently and DE genes are generated from different distribution or the same distribution with different parameters. In most previous simulation studies, the mean and variance in the normal distribution are fixed, however, there is much variance on the biological microarray data coming from the real situation. Lonnstedt and Speed (2002) used an inverse gamma prior distribution for the variance of all genes. Non-DE genes have a fixed zero mean and DE genes are produced from a normal prior distribution in this model.

In our study, a model named as normal-normal-gamma model is used to generate the simulated microarray data, in which the expression values of each gene i were independently drawn from the normal distribution $N(\mu_i, \sigma_i^2), i = 1,2,3.....N$. The means of each gene μ_i were randomly generated from a normal prior distribution $N(\mu, \sigma^2)$ and the standard deviations σ_i were generated independently from a gamma prior distribution

with shape= σ_1 and scale = γ_1. For DE gene, the means and standard deviations in two conditions are selected randomly and separately from the same distributions but independently for each condition.

To make the simulation data biologically meaningful, we used a real ovarian cancer dataset as the reference for modeling. The parameters in normal-normal-gamma model are selected in such a way that the simulated data are similar in statistical distribution to the real data. The ovarian cancer dataset contains 55 samples and 22,283 genes. After fitting a normal distribution to the means and a gamma distribution to the standard deviations of 22,283 genes in the dataset, the estimated parameters for normal-normal-gamma model are $\mu = 6.7, \sigma = 1.685, \sigma_1 = 3.875, \gamma_1 = 9.386$. The marginal densities of the ovarian cancer data and simulated data are shown in Figure 4.1, which illustrate how similar the simulated data is to the ovarian cancer data. Without loss of generality, to reduce the computational load, 10,000 genes were simulated in this study using this model. The percentage of DE genes is also set to be adjustable in the experiments.

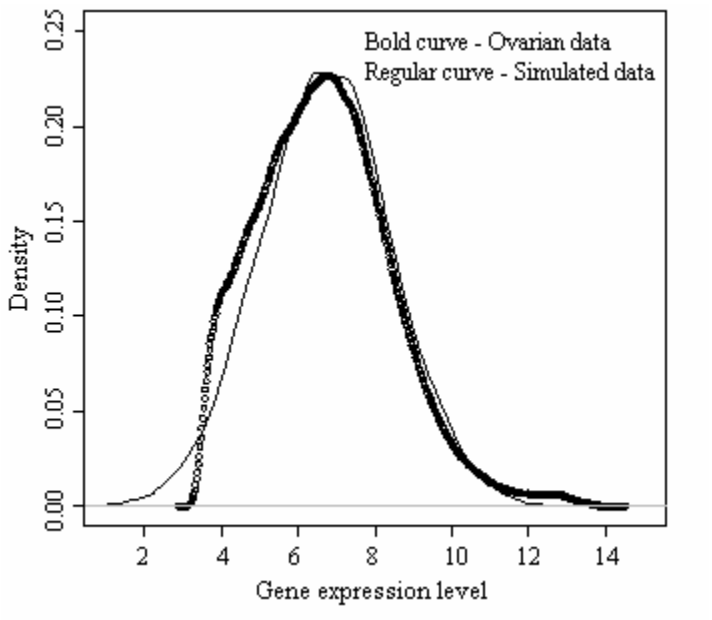


Figure 4.1 The marginal densities of the ovarian cancer data and simulated data

It is well known that due to the time-consuming experimental protocol, the cost and the often limited access to biological tissue of interest, a large number of microarray experiments are performed on a small number of samples only. Even in some clinical research where a large number of samples can be obtained, the number of samples in the interested subclass which has similar clinical information is still small due to the biological variance. For example, the clinical information can include the age of the patients, predicted survival time (short/ long), stage epithelial cancer (III/IV), surgical debulking (optimal/suboptimal), etc. Traditional statistical methods used for gene selection have to take this problem into account and most of them have a limitation due to the small sample size. In order to simulate the biological experiment scenario, we will mainly focus on cases with small sample size in the following study to compare the performance of proposed ranking-based gene selection methods with the traditional statistic methods.

4.2.2 Simulation experiments

The simulation experiments are performed under various conditions including different sample sizes, different percentages of DE genes, and different noise levels to simulate various scenarios. The classical Student's t-test and SAM, which are parametric and non-parametric statistical methods respectively, were chosen as the references to compare the performance of ranking-based methods (RP, RS, and RC) in gene selection. Since the samples are randomly generated, all experiments under each condition are repeated 10 times in order to get more reliable results.

4.2.2.1 Different sample sizes

To explore the effect of sample size on selection performance of different methods, we test the selection using datasets with different sample sizes. Figure 4.2 shows the performance of different gene selection methods, which are the t-test, RS, RP and SAM, with different sample size. The number of samples was chosen to vary from 50 to 6, which are equally divided into condition A and condition B. The percentage of DE genes is kept the same at 1% in each dataset.

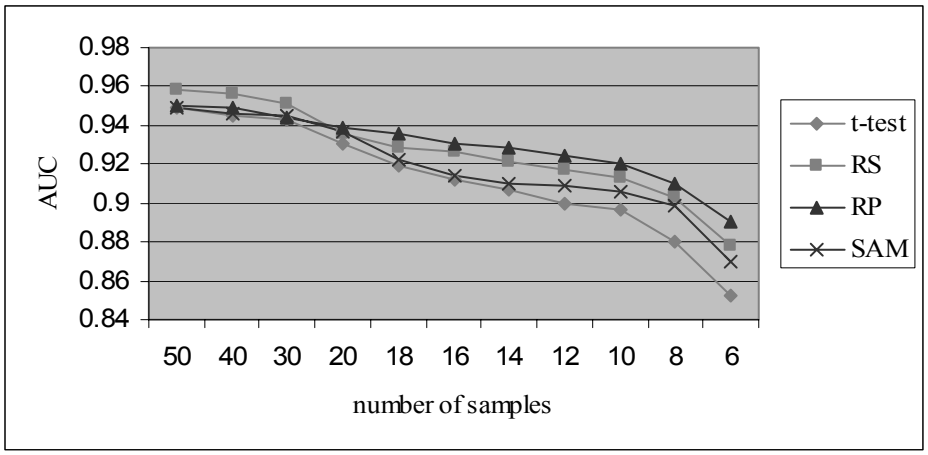


Figure 4.2 Performance of the differentially expressed gene selection with different number of samples for t-test, RS, RP, SAM

It is observed that there is not much difference between ranking-based methods (RP, RS) and statistical methods (t-test and SAM) in selecting DE genes when the sample size is greater than 30. As the sample size decreases, the performance for all methods dropped but ranking-based methods outperform the t-test and SAM consistently. The t-test drops dramatically when sample size is smaller than 10. In order to have a further comparison on the small sample size, we will focus on the results of the experiments when sample size is varying from 12 to 6.

Figure 4.3, 4.4, 4.5 and 4.6 give the ROC curves for the DE gene selection with different number of samples, which are 12, 10, 8 and 6 respectively.

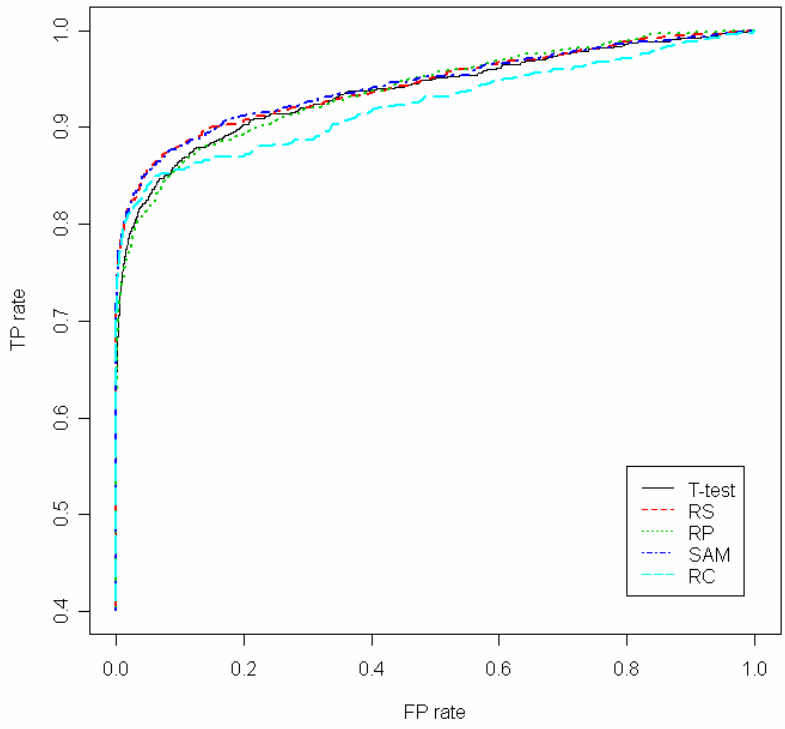


Figure 4.3 ROC curves for the DE gene selection with 12 simulated samples

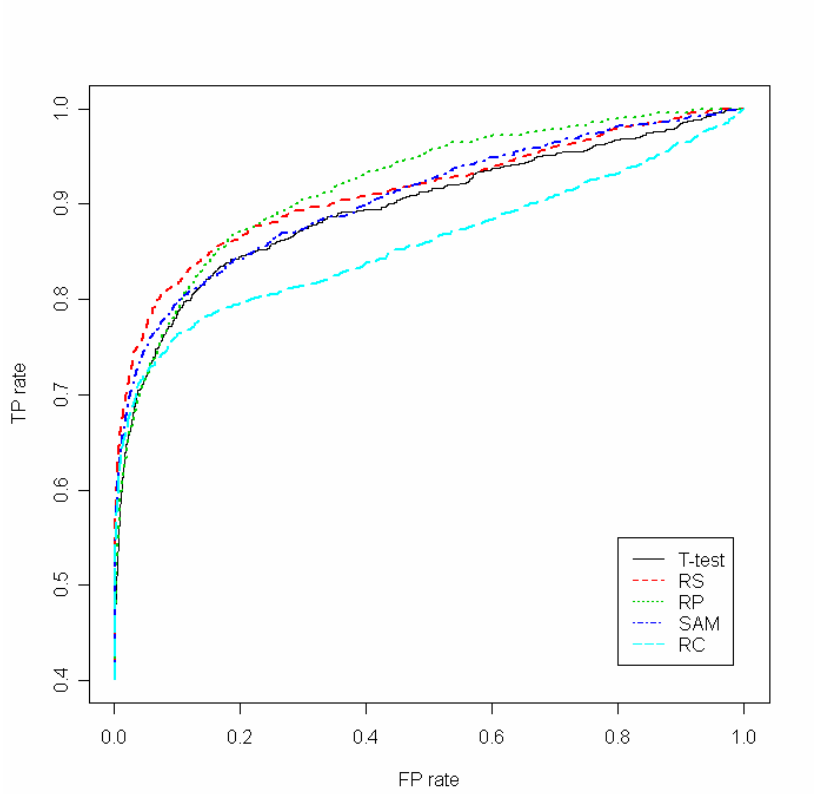


Figure 4.4 ROC curves for the DE gene selection with 10 simulated samples

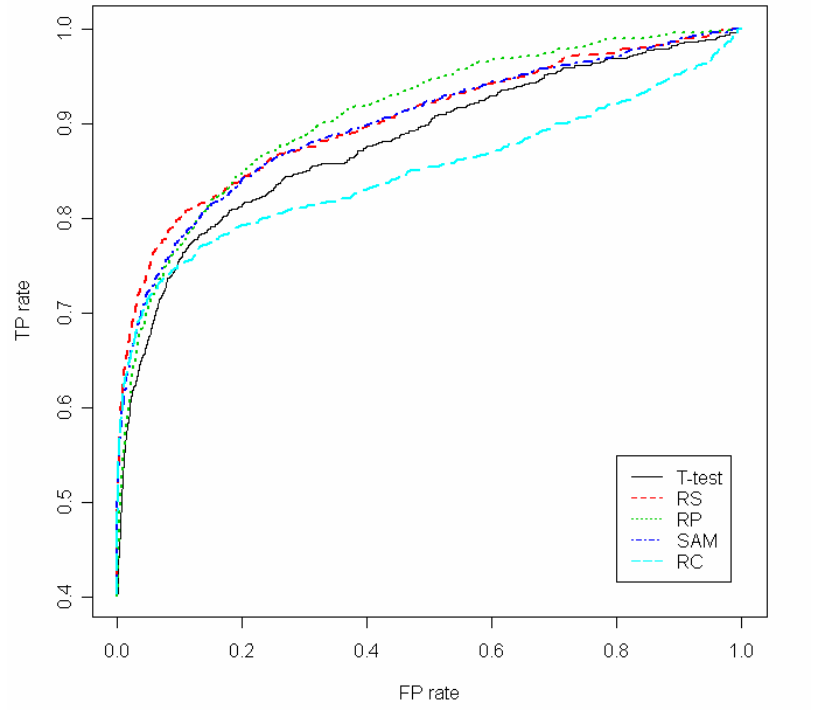


Figure 4.5 ROC curves for the DE gene selection with 8 simulated samples

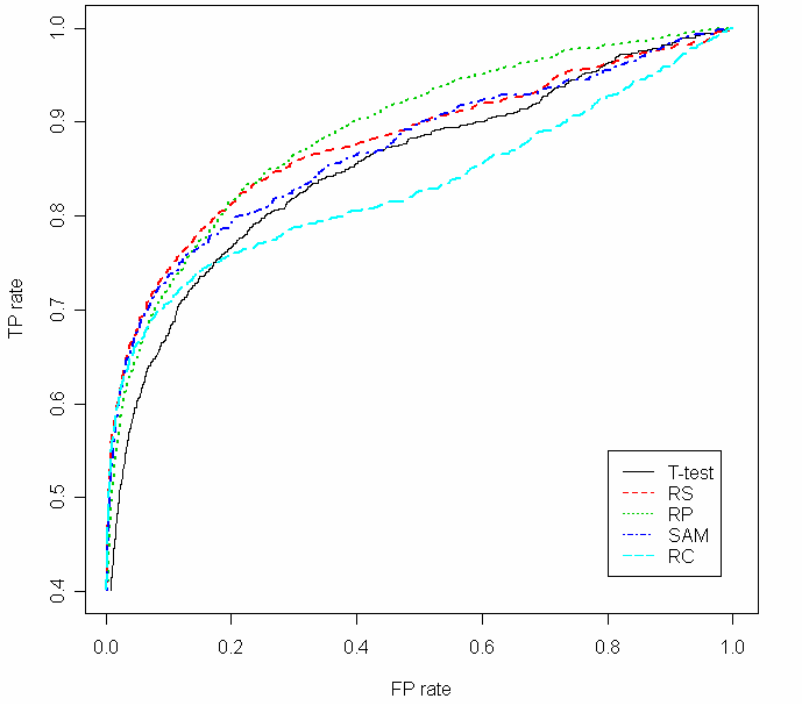


Figure 4.6 ROC curves for the DE gene selection with 6 simulated samples

Note that the TP rate axis in the figures above is adjusted to begin from 0.4. The only reason is we observed that there is not much difference in ROC curves among the methods when the true positive is less than 0.4 and it will make it easier to see the difference among different selection methods at high TP rate.

In the RC gene selection method, parameter C_0 is chosen to be 30, 20, 12 and 7 for experiment with 12, 10, 8 and 6 samples respectively. Parameter K_0 is chosen from the set {20, 40, 60, 80 … 10000} one by one to generate the ROC curve. However, the choice of the parameters in RC was somewhat arbitrary which were decided based on experimental conditions and may not be the optimal ones.

In order to get a quantitative measurement of the efficiencies of the different methods under different conditions, AUC is calculated and compared. Considering the

fact that in many cases the selection is meaningful only when the false positive rate is low, Partial AUC at a range of the FP rate is used to compare the performance of different methods. Here the cutoff point of the FP rate is chosen as 0.2 to calculate the partial AUC at low FP rate. Table 4.1 gives the partial AUC when the FP rate is between 0 and 0.2 and Table 4.2 gives the overall AUC for each method under each condition.

Table 4.1 Partial AUC for different sample sizes when the FP rate is between 0 and 0.2

Methods	12 samples	10 samples	8 samples	6 samples
T-test	0.1536	0.1519	0.1441	0.1299
RS	0.1618	0.1598	0.1543	0.1441
RP	0.1579	0.1535	0.1487	0.1399
SAM	0.1584	0.1550	0.1508	0.1418
RC	0.1472	0.1459	0.1447	0.1378

Table 4.2 Overall AUC for different sample sizes

Methods	12 samples	10 samples	8 samples	6 samples
T-test	0.8997	0.8962	0.8805	0.8525
RS	0.9176	0.9127	0.9027	0.8782
RP	0.9249	0.9199	0.9095	0.8900
SAM	0.9087	0.9058	0.8990	0.8702
RC	0.8475	0.8590	0.8494	0.8308

From the figures and tables above, we can see that when the sample size is decreasing, the performance of all methods has the decreases. RP is the best method in overall and RS has best performance when the FP rate is low. RS and RP are statistically significantly better than the t-test and SAM in each condition ($p<0.01$ for RS vs. t-test, RP vs. t-test, RP vs. SAM and RS vs. SAM in terms of TP rate). RC has lowest overall AUC value, but it is better than the t-test at the low FP rate when the sample size is small.

It is observed that the gap between the t-test and RS and RP becomes large with decreasing sample size. As it is shown in Figure 4.2, ranking-based methods have less

change compared to the t-test and SAM when the sample size changes, which shows that they are less sensitive to the change of sample size, especially when the sample size is small.

4.2.2.2 Different percentages of DE genes

In order to test the effectiveness of the proposed methods at different numbers of DE genes, the percentage of DE genes is set to vary in the range of 1%-5%. The percentage of DE genes is small because of the assumptions that only a minority of genes are in fact changed in an experiment. The sample size is fixed at 10 for each experiment.

Figure 4.7, 4.8, 4.9 and 4.10 show the ROC curves for the DE gene selection with different percentages of DE genes using the t-test, RP, RS, SAM and RC respectively. The sample size is 10 and percentage of the DE genes varies from 2% to 5%. ROC cures for the DE gene selection with 1% DE genes were shown in Figure 4.4 which is considered as a baseline experiment. The parameters in the RC gene selection method are same as the ones in the baseline experiment.

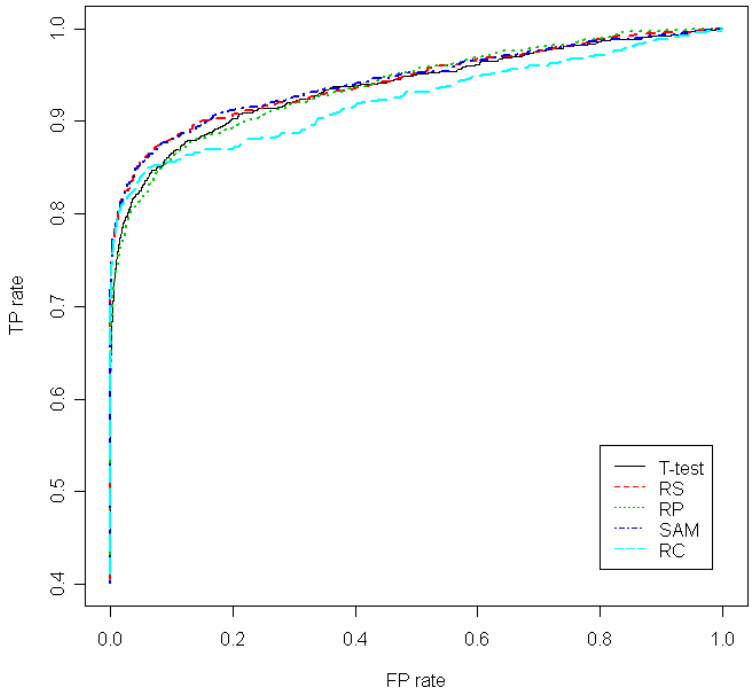


Figure 4.7 ROC curves for the DE gene selection with 2% DE genes

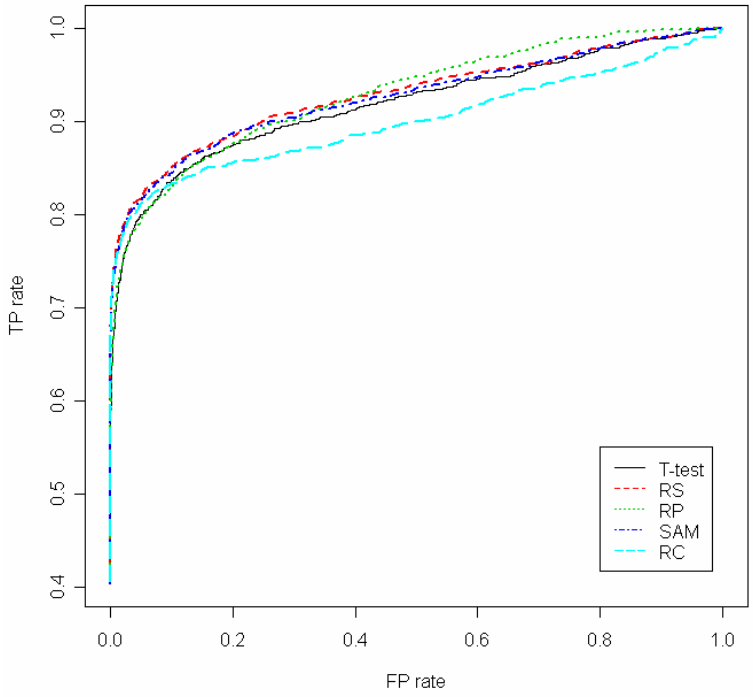


Figure 4.8 ROC curves for the DE gene selection with 3% DE genes

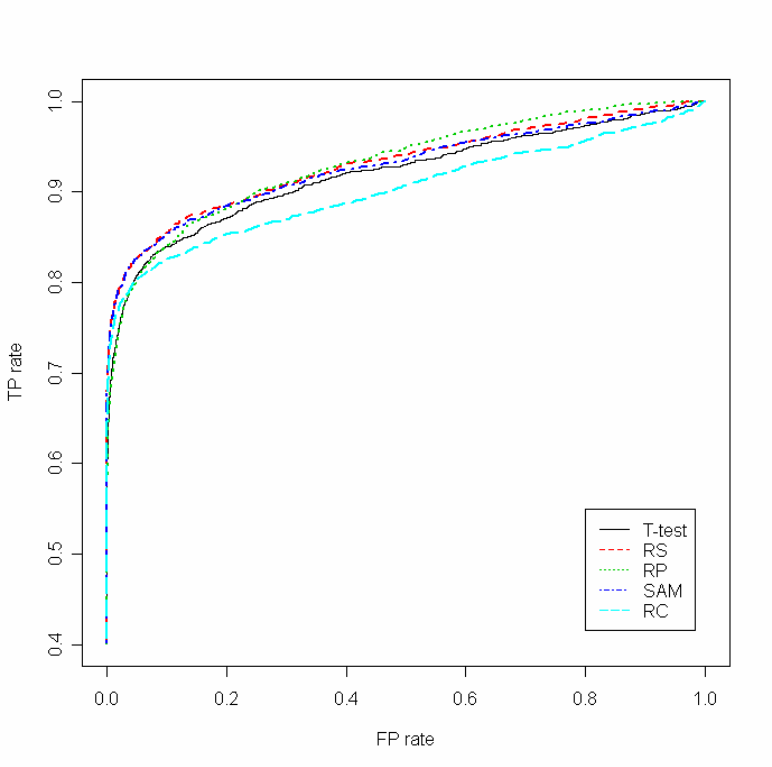


Figure 4.9 ROC curves for the DE gene selection with 4% DE genes

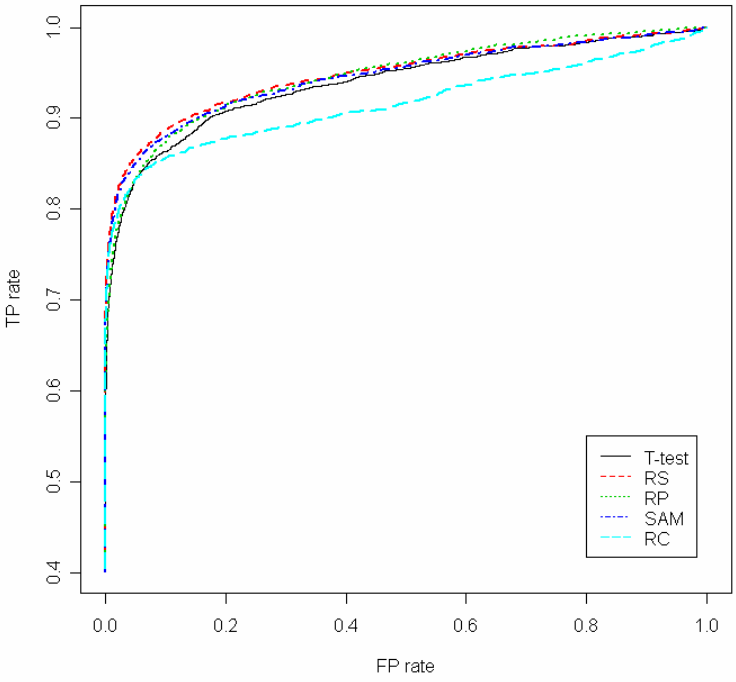


Figure 4.10 ROC curves for the DE gene selection with 5% DE genes

Table 4.3 gives the partial AUC for different methods with different percentages of DE genes when the FP rate is between 0 and 0.2. Table 4.4 is the overall AUC for different methods with different percentages of DE genes.

Table 4.3 Partial AUC for different percentages of DE genes when the FP rate is between 0 and 0.2

Methods	1%	2%	3%	4%	5%
T-test	0.1519	0.1695	0.1636	0.1641	0.1694
RS	0.1598	0.1737	0.1676	0.1686	0.1745
RP	0.1535	0.1682	0.1613	0.1645	0.1708
SAM	0.1550	0.1736	0.1699	0.1681	0.1732
RC	0.1459	0.1680	0.1633	0.1622	0.1668

Table 4.4 Overall AUC for different percentages of DE genes

Methods	1%	2%	3%	4%	5%
T-test	0.8962	0.9372	0.9181	0.9197	0.9391
RS	0.9127	0.9427	0.9278	0.9304	0.9477
RP	0.9199	0.9379	0.9286	0.9314	0.9458
SAM	0.9058	0.9433	0.9252	0.9273	0.9450
RC	0.8590	0.9238	0.9001	0.9023	0.9162

It is observed that when the percentage of DE genes is increasing, compared to the baseline, the performance for each method is improved. Overall RS and RP are generally better than other gene selection methods (Table 4.4), and RS performs best when the FP rate is low (Table 4.3). The figures show that there is a statistically significant improvement between RS and t-test, RP and t-test ($p<0.01$ for RS vs. t-test and RP vs. t-test). RP and RS are comparable to SAM on the dataset with 2% percentage DE genes ($p=0.458$ for RP vs. SAM and $p=0.058$ for RS vs. SAM), but they have significant improvement to SAM when the percentage of DE genes increases ($p<0.01$ for RP vs. SAM and RS vs. SAM). RC is comparable to the t-test although it has lowest overall AUC values.

4.2.2.3 Different noise levels

It is common that gene expression data may be noisy due to the variation of the biological samples and microarray experiments. In this case, experiments using different noise level datasets are conducted in order to evaluate the effect of noise on the selection performance of different methods. We assume that the noise follows the normal distribution and the noise level is determined by the standard deviation. However, the distribution of the noise is not necessary to be the normal distribution since ranking-based methods do not have the assumption on the distribution of microarray data. Noise is added into each gene expression value in the original dataset, which is randomly generated from the normal distributions.

Figure 4.11, 4.12 and 4.13 give the comparison of ROC curves among these gene selection methods when we add different noise levels to the original dataset. The noise follows the normal distribution with the mean 0 and standard deviation (sd) 0.2, 0.5 and 1 respectively. The sample size is 10 and there are 1% DE genes in the dataset. Parameters in the RC gene selection method are same as the ones in the baseline experiment.

Table 4.5 gives the partial AUC of different gene selection methods at different noise levels when the FP rate is between 0 and 0.2. Table 4.6 shows the overall AUC of different gene selection methods at different noise levels.

Table 4.5 Partial AUC for different noise levels when the FP rate is between 0 and 0.2

Methods	No noise	Sd=0.2	Sd=0.5	Sd=1.0
T-test	0.1519	0.1427	0.1281	0.1013
RS	0.1598	0.1503	0.1338	0.1071
RP	0.1535	0.1450	0.1319	0.1068
SAM	0.1550	0.1457	0.1340	0.1073
RC	0.1459	0.1490	0.1254	0.1011

Table 4.6 Overall AUC for different noise levels

Methods	No noise	Sd=0.2	Sd=0.5	Sd=1.0
T-test	0.8962	0.8765	0.8405	0.7672
RS	0.9127	0.8916	0.8497	0.7781
RP	0.9199	0.8998	0.8549	0.7779
SAM	0.9058	0.8884	0.8522	0.7785
RC	0.8590	0.8499	0.8162	0.7526

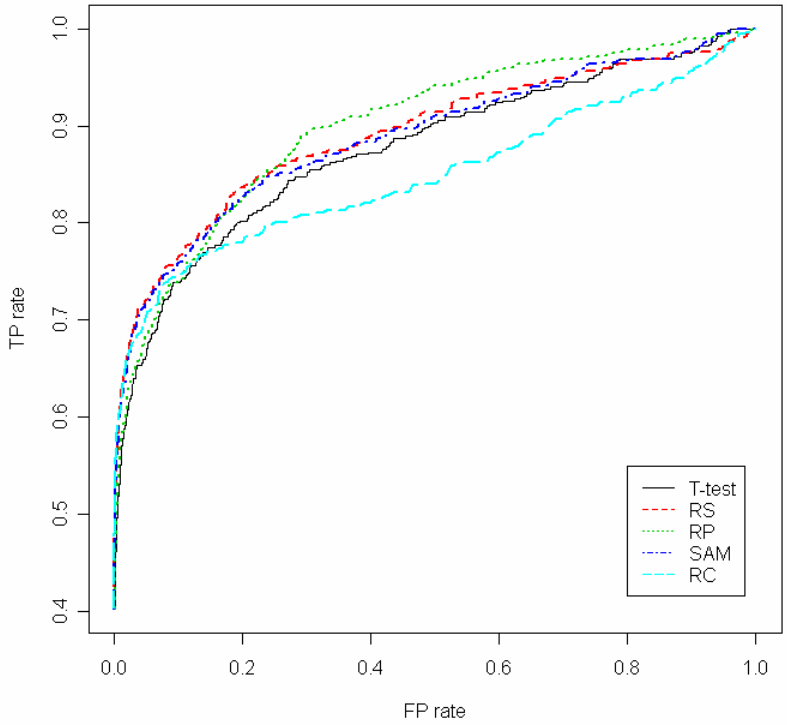


Figure 4.11 ROC curves for the DE gene selection with 0.2 sd noise level

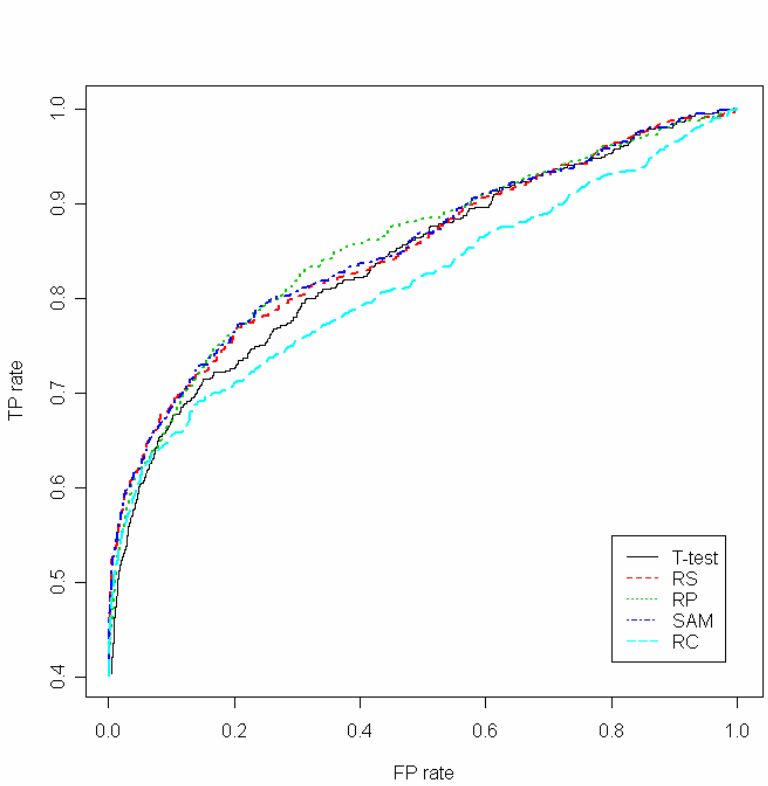


Figure 4.12 ROC curves for the DE gene selection with 0.5 sd noise level

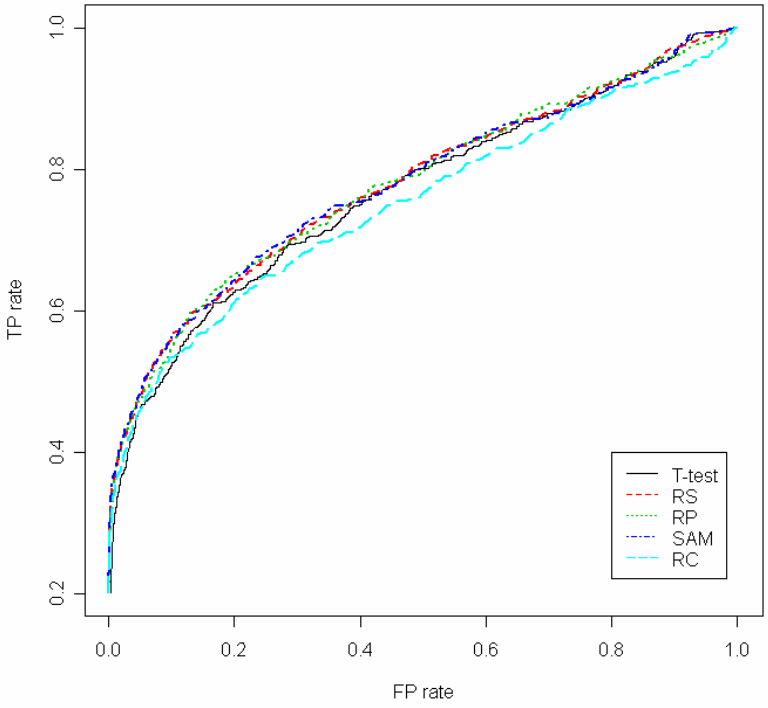


Figure 4.13 ROC curves for the DE gene selection with 1.0 sd noise level

From the figures we can see that the performance for each method drops when the noise level increases. However, RP and RS are consistently better than the t-test method, especially when the false positive rate is low (Table 4.5 and Table 4.6). SAM turns better when the noise level is high. Figure 4.14 shows the relative decrease of overall AUC in ratio compared to the baseline experiment for each method when the noise level increases to 1.0 standard deviation. RC decreases least compared to other methods which shows that RC is less sensitive to the noise.

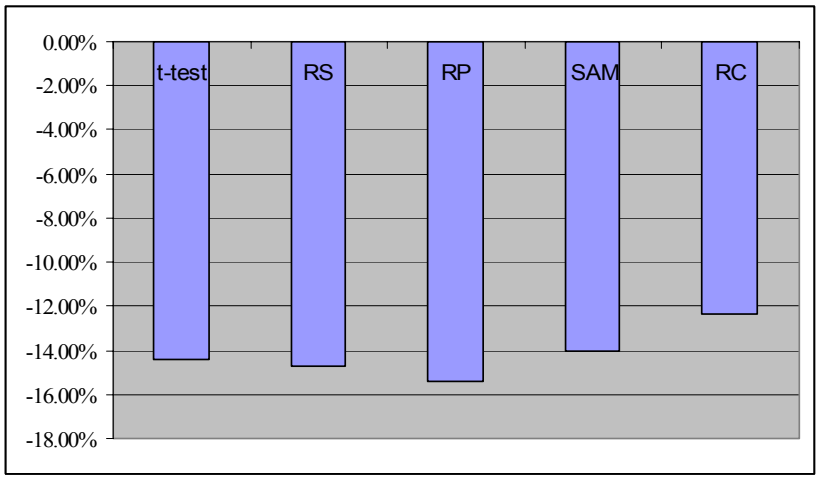


Figure 4.14 Relative decrease of overall AUC in ratio compared to baseline experiment for each method when the noise level increases to 1.0 standard deviation

4.3 Summary

RP and RS perform very well on the simulated microarray datasets compared to the t-test and SAM under different conditions, especially in the small sample size. Specifically, RP has the best overall efficiency in the DE gene selection and RS achieves better performance when the FP rate is low than others. RC has better performance than the t-test when the FP rate is low although it is not as good as the t-test overall. RC is less sensitive to the noise compared to other methods when noise level is increasing.

Chapter 5

Application on Biological Microarray Datasets

5.1 "Truth" of biological data

In real biological microarray datasets, it is usually hard to know which genes are truly significant differentially expressed under different biological conditions. The difference in gene expression value may simply result from the biological varieties and the experimental errors. Therefore, a validation of gene selection methods using real biological microarray data is a challenging task. One possible solution is to specify the DE genes in experiments. For example, the benchmark dataset provided by the Affymetrix GeneChip™ makes it possible by providing us the known spiked-in DE genes. However, specified DE genes may not reflect the true biological information. The second approach to circumvent the problem of unknown "truth" is to evaluate the sample classification/prediction performance by using the "DE" genes identified by the selection method. In contrast to the first approach, this is an indirect evaluation approach. It is based on the assumption that DE genes have better predictive abilities than non DE genes, which means that the better gene selection method should have better performance in classification and prediction.

In the classification and prediction applications, K-fold cross validation is widely used (Witten, 2000) as an evaluation method. The dataset is divided into k subsets, and

the classification and prediction are repeated k times in K-fold cross validation. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. The final result is the average output across all k trials. Every sample gets to be in a test set exactly once, and gets to be in a training set k-1 times. The disadvantage is that the result may have a big variation in different experiments because the dataset is randomly divided. In order to reduce this kind of variation, cross validation will be repeated several times by using multiple different splits of the dataset into k folds. Although k can take any value, k=10 has been experimentally shown in literature to achieve a reasonable estimate of error (Witten, 2000).

5.2 Gene selection on Affy spike-in experimental data

The spike-in experiment represents a portion of the data used by Affymetrix GeneChip^{TM} to develop their MAS 5.0 preprocessing algorithm. The data feature 14 human genes spiked-in at a series of 14 known concentrations (0, 2^{-2}, 2^{-1} … 2^{10} pm) according to a Latin square design including 12 612 null genes. Each 'row' of the Latin square (given spike-in gene at a given concentration) was replicated (typically three times, two rows 12 times, 59 arrays in total). More details about this data are available at http://www.affymetrix.com/analysis/download_center2.affx. We utilize RMA (Irizarry *et al.*, 2003) to summarize probe level gene expression data. A portion of this dataset that presents a two-group comparison problem with 12 replicates in each group was used in this study. The subsets of samples were selected from them randomly with different sample sizes.

Figure 5.1, 5.2, 5.3 and 5.4 provide the ROC curves for the DE gene selection using t-test, RP, RS, SAM and RC when the sample size is 4, 6, 8, and 10 respectively. As in the simulation study, we repeat 10 times under each condition to get more reliable results. Parameter C_0 in RC gene selection method is set to 2, 7, 12, and 20 corresponding to sample size 4, 6, 8, and 10, respectively. We choose 500 equidistant cutoff points for K_0 to get the TP rate and FP rate pairs to generate ROC curves. However, the choices of the parameters in RC may not be the optimal ones. Note that the axis of TP rate is adjusted to show the difference better among methods in each ROC curve.

Table 5.1 lists the partial AUC for different methods at different sample sizes when the FP rate is between 0 and 0.2. Table 5.2 gives the overall AUC for each condition.

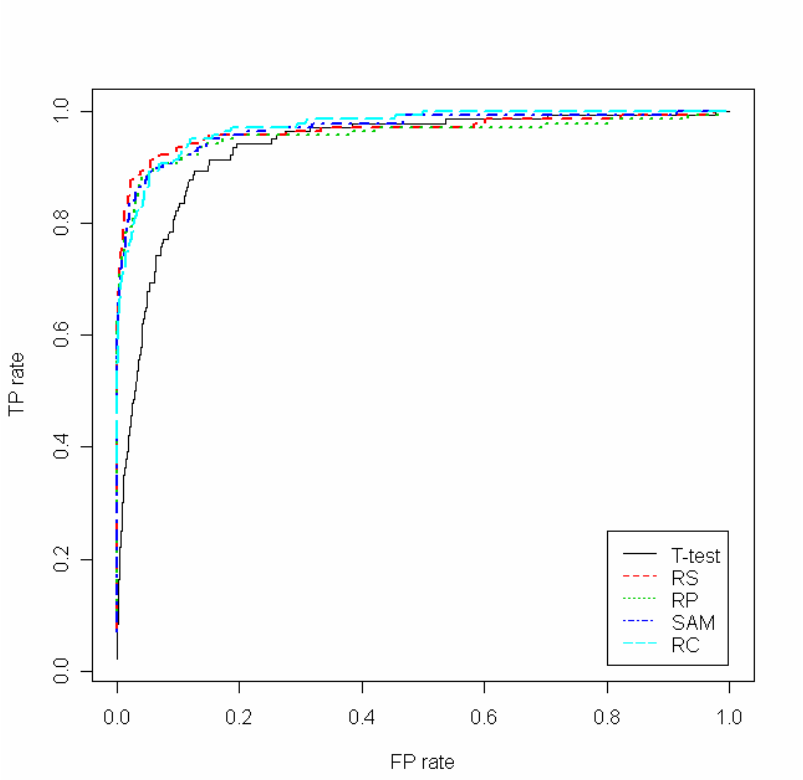


Figure 5.1 ROC curves for the DE gene selection with 4 spike-in samples

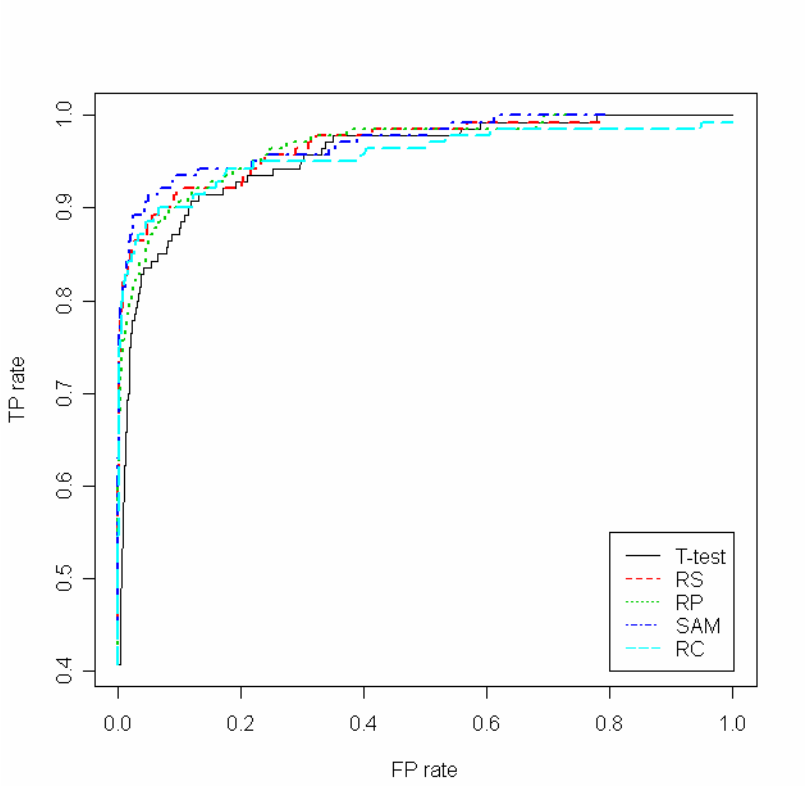


Figure 5.2 ROC curves for the DE gene selection with 6 spike-in samples

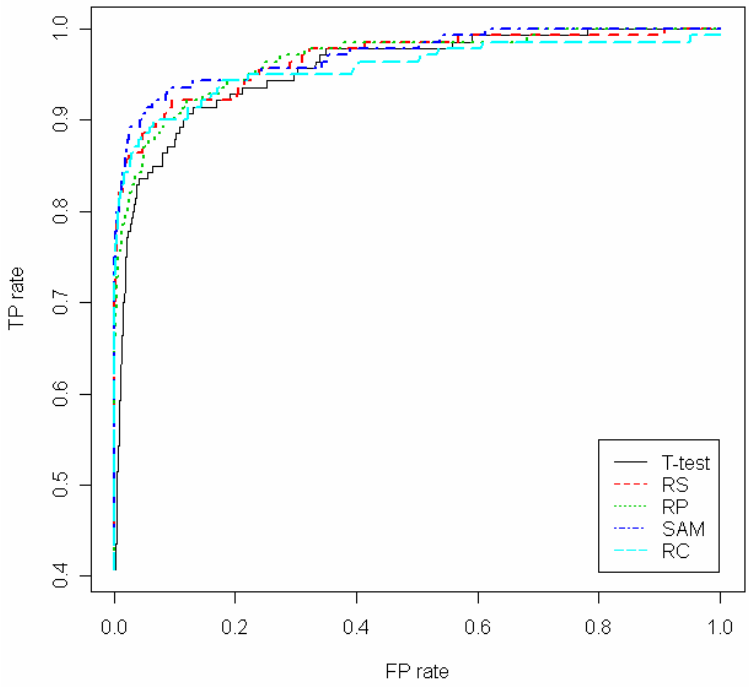


Figure 5.3 ROC curves for the DE gene selection with 8 spike-in samples

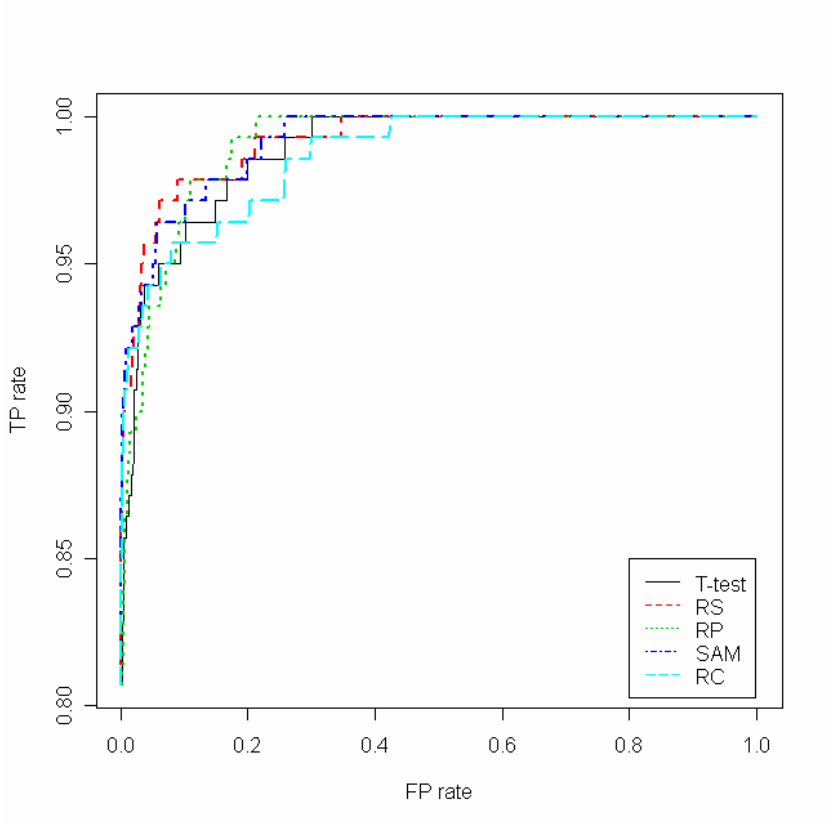


Figure 5.4 ROC curves for the DE gene selection with 10 spike-in samples

Table 5.1 Partial AUC for different sample sizes when the FP rate is between 0 and 0.2 in the spike-in experiment

Methods	4 samples	6 samples	8 samples	10 samples
T-test	0.1497	0.1696	0.1847	0.1891
RS	0.1826	0.1789	0.1898	0.1925
RP	0.1788	0.1767	0.1874	0.1898
SAM	0.1792	0.1830	0.1917	0.1918
RC	0.1783	0.1769	0.1858	0.1887

Table 5.2 Overall AUC for different sample sizes in the spike-in experiment

Methods	4 samples	6 samples	8 samples	10 samples
T-test	0.9347	0.9552	0.9795	0.9880
RS	0.9655	0.9667	0.9845	0.9914
RP	0.9573	0.9665	0.9827	0.9897
SAM	0.9691	0.9716	0.9889	0.9913
RC	0.9736	0.9575	0.9837	0.9864

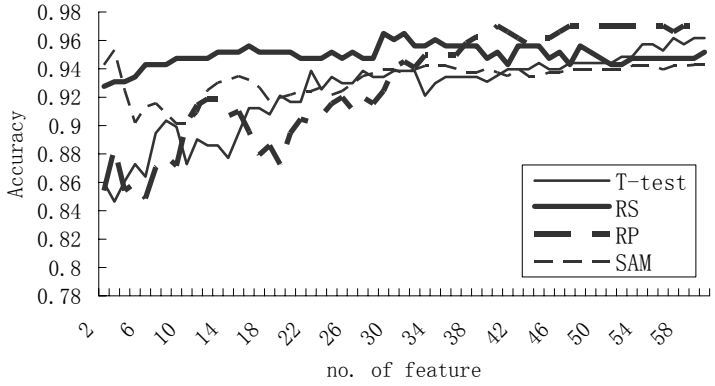
Figure 5.1 shows that all the ranking-based methods have statistically significant improvements compared to the t-test when the sample size is 4 ($p<0.01$ for RP vs. t-test, RS vs. t-test and RC vs. t-test). RC gets best overall performance in this case. When the sample size increases, all methods get better. Among these, RP and RS consistently outperform the t-test ($p<0.01$ for RP vs. t-test and RS vs. t-test), but are comparable to SAM ($p>0.01$ for RP vs. SAM and RS vs. SAM). RC has a little drop compared to other methods when the sample size is 10.

5.3 Leukemia prediction

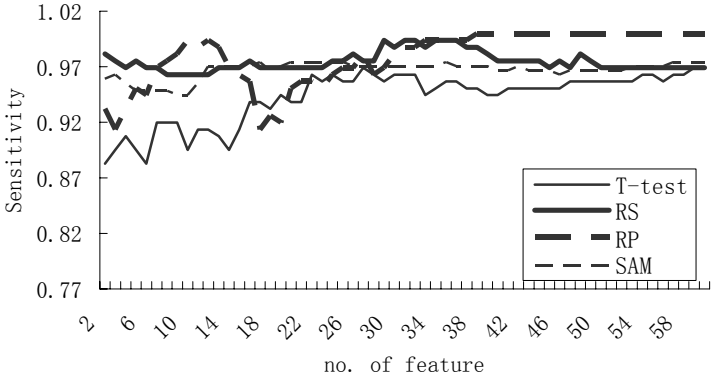
The leukemia dataset used for this analysis is available to the public (Golub *et al.* 1999). It includes 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples and each sample has 7129 genes. Our goal is to evaluate the performance of different gene selection methods by comparing the accuracy of prediction using the genes selected by each gene selection method.

Support Vector Machines (SVMs) (Vapnik, 1998) were used to build the classifiers and the features were selected by the top genes ranked in different gene selection methods. The number of features varied from 2 to 60. 10-fold cross validation was repeated 10 times for each case to get the average results. A comparison was made on the performance of RP, RS, SAM and t-test. RC is not included because the number of selected DE genes is determined by the parameters C_0 and K_0 and it is hard to compare it at a gene number specified by other methods.

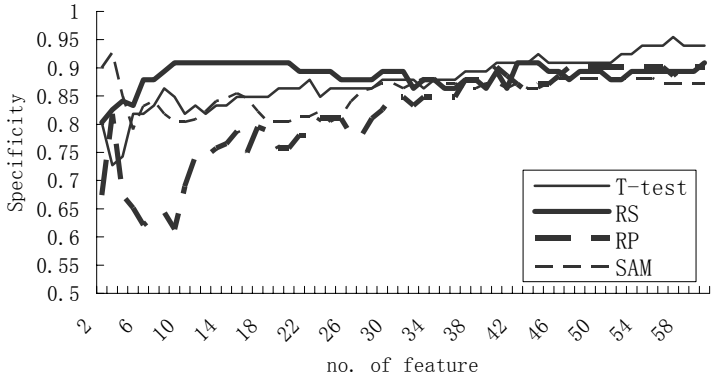
Figure 5.5 shows the comparison of performance of 10-fold cross validation with different number of features selected by the t-test, SAM, RP and RS. The performance is compared in terms of accuracy, sensitivity and specificity.



(a)



(b)



(c)

Figure 5.5 Performance comparison of 10-fold cross validation at different number of features selected by t-test, SAM, RS and RP on the leukemia dataset
(a) accuracy (b) sensitivity (c) specificity

The average accuracies across all feature numbers of the t-test, RS, RP, and SAM were 92.37%, 94.94%, 92.83% and 93.22% respectively. The best accuracy is 97.04% which is achieved by RP when the number of features was greater than 47. RS performs well compared to other methods when the number of features was between 6 and 30, especially it had good specificities in this range. RP performs well when number of feature is increasing and its sensitivities are 100% when the number of feature is greater than 37.

For a comparison of these different gene selection methods, a set of statistical tests were performed to measure the significance of improvements in accuracy, sensitivity and specificity respectively. As the normality test fails, the Wilcoxon test is used to test the significance of the differences. Listed in Table 5.3 are the p-values of the Wilcoxon test for the t-test compared to RP and RS, Table 5.4 shows the p-values of Wilcoxon test for SAM compared to RP and RS. The tables show that RS has statistically significant improvement compared to SAM and it is outperform the t-test on accuracy and sensitivity. RP is comparable to the t-test and SAM. However, the statistical test may have high type I error because of 10 times 10-fold cross validation (Dietterich, T.G., 1998).

Table 5.3 Statistical Wilcoxon test of performance improvement for t-test, RP and RS on the leukemia dataset

	Method 1	Method 2	Wilcoxon test p-value
Accuracy	T-test	RP	0.308
	T-test	RS	7.033e-09
Sensitivity	T-test	RP	7.971e-13
	T-test	RS	1.339e-12
Specificity	T-test	RP	6.378e-06
	T-test	RS	0.08516

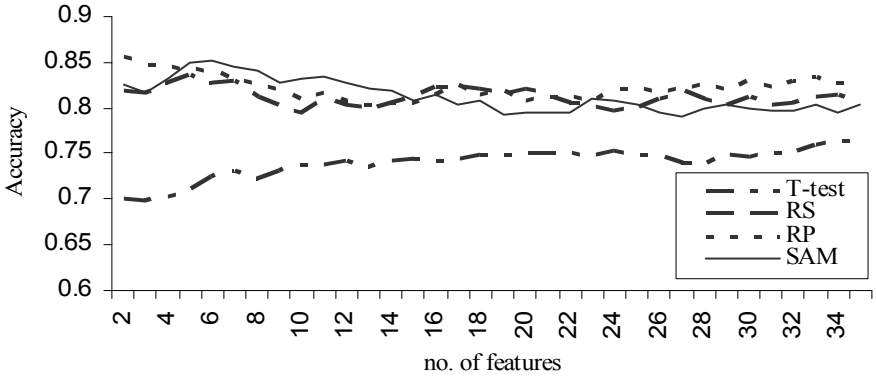
Table 5.4 Statistical Wilcoxon test of performance improvement for SAM, RP and RS on the leukemia dataset

	Method 1	Method 2	Wilcoxon test p-value
Accuracy	SAM	RP	0.6845
	SAM	RS	6.766e-15
Sensitivity	SAM	RP	9.215e-4
	SAM	RS	2.814e-06
Specificity	SAM	RP	0.001
	SAM	RS	1.062e-09

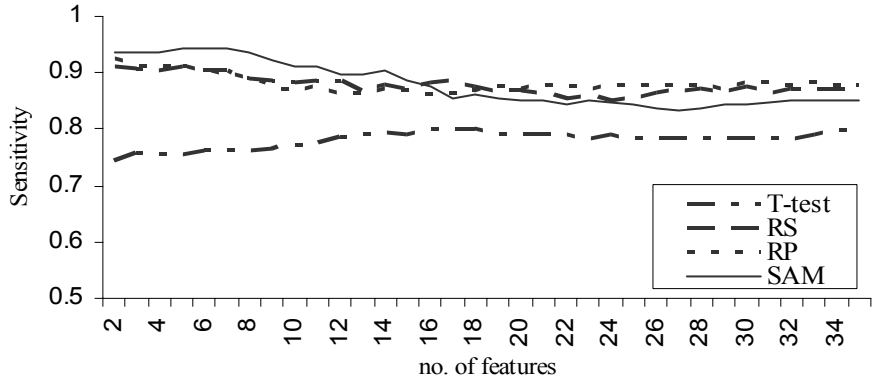
5.4 Colon cancer detection

The colon cancer dataset is a well-known benchmark microarray dataset which can be obtained from the website http://microarray.priceton.edu/oncology/affydata/. The colon cancer data contain 62 tissue samples including 22 normal and 40 colon cancer tissues. Each sample has 2000 gene expression values. Performance is compared between ranking-based methods (RP and RS) and the t-test and SAM in terms of accuracy, sensitivity and specificity by using 10-fold cross validation. Support Vector Machines (SVMs) were used as the classifiers and the features were selected by the top genes as ranked by the different gene selection methods. The number of features varied from 2 to 35 and we also repeat 10 times for each case to get more reliable results.

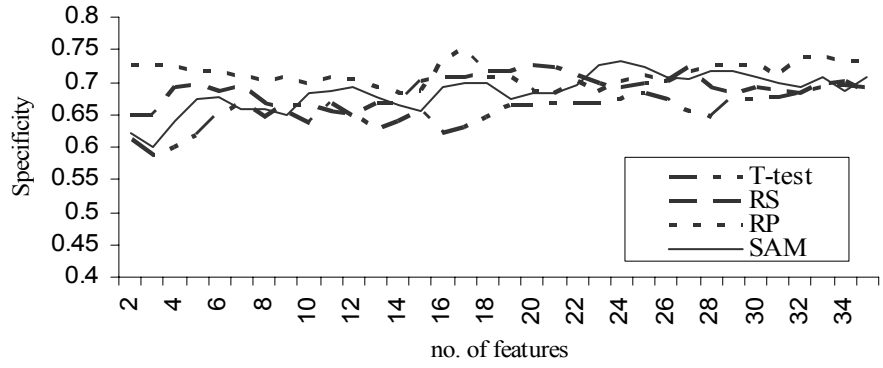
Figure 5.6 shows the comparison of performance of 10-fold cross validation with different number of features selected by the t-test, SAM, RP and RS.



(a)



(b)



(c)

Figure 5.6 Performance comparison of 10-fold cross validation at different number of features selected by t-test, SAM, RS and RP on the colon cancer dataset
(a) accuracy (b) sensitivity (c) specificity

Figure 5.6 shows that the performance of 10-fold cross validation for both RP and RS has significant improvements compared to the t-test. The average accuracies on different number of features for RP and RS are about 82.30% and 81.27% while only about 74.01% for the t-test. For a comparison of these three different gene selection methods, a set of statistical tests were performed to measure the significance of improvements in accuracy, sensitivity and specificity respectively. As the normality test fails, the Wilcoxon test is used to test the significance of the differences. Listed in Table 5.5 are the p-values of the Wilcoxon test. The p-values show that there is a significant difference between RP and t-test, and RS and t-test. However, the statistical test may have high type I error because of 10 times 10-fold cross validation (Dietterich, T.G., 1998).

Table 5.5 Statistical Wilcoxon test of performance improvement for t-test, RP and RS on the colon cancer dataset

	Method 1	Method 2	Wilcoxon test p-value
Accuracy	T-test	RP	1.406e-12
	T-test	RS	1.406e-12
Sensitivity	T-test	RP	1.326e-12
	T-test	RS	1.339e-12
Specificity	T-test	RP	9.375e-12
	T-test	RS	1.869e-06

The average of accuracies of SAM across all the feature number is 81.27%. Table 5.6 shows the statistical Wilcoxon test of performance improvement for SAM, RP and RS. It is shown that RP and RS are comparable to SAM. SAM performs well in accuracy when the number of feature is less than 16, but it has lower specificities than RP. RS and RP perform well in both accuracy and sensitivity when the number of features is greater than 16. There is no significant difference on accuracy between RP and RS, while RP has better performance on specificity compared to RS.

Table 5.6 Statistical Wilcoxon test of performance improvement for SAM, RP and RS on the colon cancer dataset

	Method 1	Method 2	Wilcoxon test p-value
Accuracy	SAM	RP	0.013
	SAM	RS	0.976
Sensitivity	SAM	RP	0.659
	SAM	RS	0.999
Specificity	SAM	RP	5.272e-05
	SAM	RS	0.594

The performance of 10-fold cross validation on the colon cancer data in this study is not good as the ones in the paper of Guyon, *et al.*, (2002) in which the data was pre-processed using extensive methods and genes are selected utilizing SVMs based on Recursive Feature Elimination (RFE). In our study, only gene selection methods are concerned and RP and RS are shown to have statistically significant improvement compared to t-test in the same experimental conditions.

5.5 Summary

The biological experiments show that ranking-based methods perform well in the DE gene selection and sample classification/prediction applications. In the Affy Spike-in experiment, RP and RS outperform the t-test in selecting DE genes, especially when the sample size is small. RC results in better performance than the t-test when the FP rate is low and better than SAM when the sample size is small, but its performance drops when the sample size increases. RP and RS also show good performance in accuracy compared to the t-test and SAM implemented by 10-fold cross validation in the leukemia dataset and they have statistically significant improvement compared to the t-test in the colon cancer dataset.

Chapter 6

Discussion and Conclusion

6.1 Discussion

Ranking-based methods are proposed for selecting differentially expressed genes in microarray data. We proposed three methods to combine the individual ranks. They are Rank Product (RP), Rank Average/Summation (RS) and Committee Decision method on ranks (RC). The simulation experiments show that RP and RS perform very well compared to the t-test and SAM under different conditions, especially with small sample size. Specifically, RS has better performance in low false positive rate and RP has the best overall performance. How to integrate RS and RP to get more accurate performance could be further studied.

RC is less sensitive to the noise compared to the other methods when the noise level increases in the simulation experiments. However, the overall performance of RC is not as good as other ranking-based methods. One of the reasons is that it is difficult to choose the optimal parameters K_0 and C_0 in RC, which will have large influence on the performance. The choice of the parameters in RC in the study was somewhat arbitrary which were decided based on experimental conditions and may not be the optimal ones. Further study can be explored to find the optimal parameters in RC to improve the performance of gene selection.

In the simulation experiment, the performance of RP, RS and t-test were similar when the sample size is large. However, the experiments on classification/prediction on

two benchmark microarray datasets show that RP and RS still significantly outperform the t-test in selecting differentially expressed genes, although there is a relative large number of samples for the t-test. Gene expression in simulation microarray data has a normal distribution and the t-test can perform well when the sample size is large. While this assumption may not be met in biological microarray data, the t-test will then have poor performance and ranking-based methods show better ability since they have a relatively weak assumption about the data. More experiments on biological microarray data should be done in further studies.

One of the potential advantages of Ranking-based methods is that they are considered to be independent of microarray platform since no distribution and variance need to be estimated. Experiments on simulated and biological microarray data could be conducted to evaluate the performance in gene selection on the integrated data from different sources and/or experiments.

Additionally, other integration approaches to combine the individual ranks in ranking-based methods could be explored besides RP, RS and RC. The determination of significant levels for differentially expressed genes can also be studied in further.

6.2 Conclusion

Ranking-based gene selection methods use rank information among genes rather than actual gene expression levels. It has the biological intuition that a significant gene has large relevant changes while small changes may have statistical but rarely biological significance. The proposed ranking-based methods in this thesis consider the correlation among the genes. They make relatively weak assumption about the data and do not rely

on estimating the measurement variance for each single gene, which is particularly useful when this estimate becomes unreliable due to a low number of samples. Results in simulation and biological experiments show that ranking-based methods perform better than the t-test and SAM in selecting differentially expressed genes, especially when the sample size is small.

References

Affymetrix. (1999) Affymetrix Mcroarray Suite Users Guide. santa Clara, CA, version 4.0 edition.

Affymetrix. (2001) Affymetrix Mcroarray Suite Users Guide. santa Clara, CA, version 5.0 edition.

Baumgartner, W., Weiβ, P. and Schindler, H. (1998) A nonparametric test for the general two-sample problem. *Biometrics*, 54, 1129-1135.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* B 57, 289-300.

Bickel, D. (2004) Degrees of differential gene expression. *Bioinformatics,* Advance Access.

Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1-3): 83-92.

Chen,Y., Dougherty, E.R. and Bittner, M.L. (1997) Ratio-based decisions and quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2, 364-367.

Cox, D. and Wong, M.Y. (2004) A simple procedure for the selection of significant effects. *J R Stat Soc Ser* B, 66:395-400.

Dean, N. and Raftery, A.E. (2005) Normal uniform mixture differential gene expression detection for cDNA microarrays, *BMC Bioinformatics*, 6:173.

DeRisi, J.L., Iyerr, V.R., Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686.

Dietterich, T.G., (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895—1924.

Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica.*, 12(1), 111–139.

Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(Suppl. 1), S105–S110.

Efron, B., Tibshirani, R. Storey, J.D. (2001) Tusher V. Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*. 96: 1151-1160.

Golub, T.R., Slonim, D.K., Tampayo,P.,m Huard, C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 285, 531-537.

Gottardo R. (2002) Statistical Analysis of Microarray Data: A Bayesian approach, *Biostatistics*, 1:1-37.

Grant, G. R., E. Manduchi, and C. J. Stoeckert Jr. (2002) Using non-parametric methods in the context of multiple testing to identify differentially expressed genes, *Methods of microarray data analysis*, eds. S.M. Lin and K.F. Johnson, Kluwer Academic Publishers: Boston; 37-55.

Guyon, I., Weston, J., and Barnhill, S. (2002) Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, 46, 389-422.

Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2001) Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, 7, 805–817.

Irizarry, R.A., Bolstad, B.M, Collin, F., and Cope, L.M. (2003) Bridget Hobbs and Terence P. Speed, Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15.

Kendziorshi, C.M., M.A. Newton, H.Lan, and M.N. Gould. (2003) On Parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22: 3899-3914.

Lehmann, E.L. (1975) Nonparametrics: Statistical Methods Based on Ranks. San Francisco, CA: Holden-Day.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnlogy*, 14, 1675 - 1680.

Lonnstedt, I. and Speed, T. (2002) Replicated Microarray data, *Statistical Sinica.*, 12:31-46.

Martin DE, Demougin P, Hall MN, and Bellis M. (2004) Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics.* 11;5:148.

McLachlan, G. J., R. W. Bean, and D. Peel. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics,* 18:413–422.

Neuhauser M, Senske R. (2004) The Baumgartner-Weiβ-Schindler test for the detection of differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 20(18):3553-64.

Newton, M.A., Kendziorski C.M., Richmond C.S., Blattner F.R. and Tsui K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational Biology*, 8:37-52.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18, 546–554.

Pan, W., Lin J., and Le, C.T. (2001) A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data. Research report 2001-011, Division of Biostatistics, University of Minnesota. /*Functional & Integrative Genomics*. 3:117–124, 2003.

Pan, W., Lin, J., and Le, C.T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach, *Genome Biol*, 3(5): research0022.1–research0022.10.

Park PJ, Pagano M, and Bonetti M. (2001) A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac Symp Biocomput.* 52-63.

Piatetsky-Shapiro, G.. and Tamayo, P. (2003) Microarray data mining: facing the challenges. *ACM SIGKDD Explorations Newsletter*, 5(2)1-5.

Schena M., Shalon D., Davis R., and Brown P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467-470.

Thomas,J.G., Olson,J.M., Tapscott,S.J. and Zhao,L. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genomic Res.,* 11, 1227–1236.

Tusher, V., Tibshirani R., and Chu C. (2001) Significance Analysis of Microarrays Applied to Transcriptional Response to Ionizing Radiation, *Proceedings of the National Academy of Sciences,* 98: 5116-5121.

Wang, S. and Ethier, S. (2004) A generalized likelihood ratio test to identify differentially expressed genes from microarray data. *Bioinformatics,* 20(1):100-4.

Zhao, Y. and Pan, W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics,* 19(9):1046-54.

Zimmerman, D.W. and Zumbo, B.D. (1993) The relative power of parametric and nonparametric statistical methods. In Keren, G. and Lewis,D. (eds), A handbook for data analysis in the behavioral sciences: methodological issues. *Lawrence Erlbaum Associates*, Hillsdale, HJ, pp. 481-517.

Bibliography

Black, D. (1963) The Theory of Committees and Elections. Cambridge University Press.

Devore, J. and Peck, R. (1997) Statistics: the Exploration and Analysis of Data, 3rdedn, Duxbury Press, Pacific Grove CA.

Draghici,S. (2003) Data analysis for DNA microarrays. Chapman & Hall/CRC.

Parmigiani, G., Elizabeth, S., Garrett, R. A. Irizarry, S. and Zeger, L. (2003) The Analysis of Gene Expression Data, Methods and Software, Springer.

Vapnik, V.N. (1998) The Statisitcal Learning Theory. Springer.

Witten, I.H. and Frank, E. (2000) Data Mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufmann Publishers, San Fransisco, CA.