Published in final edited form as:

J Biopharm Stat. 2010 March ; 20(2): 223–239. doi:10.1080/10543400903572720.

Optimized Ranking and Selection Methods for Feature Selection with Application in Microarray Experiments

Xinping Cui^{1,2,*}, Haibing Zhao³, and Jason Wilson⁴

¹Department of Statistics, University of California, Riverside, CA 92521

²Center for Plant Cell Biology and Institute for Integrative Genome Biology, University of California, Riverside, CA 92521

³Department of Statistics, Shanghai University of Finance and Economics, Shanghai 200433, China

⁴Department of Mathematics, Biola University, La Mirada, CA 90639

Abstract

In microarray experiments, the goal is often to examine many genes, and select some of them for additional investigation. Traditionally, such a selection problem has been formulated as a multiple testing problem. When the genes of interest are genes with unequal distribution of gene expression under different conditions, multiple testing methods provide an appropriate framework for addressing the selection problems. However, when the genes of interest are a set of genes with the largest difference in gene expression under different conditions, multiple testing methods do not directly address the selection goal and sometimes lead to biased conclusions. For such cases, we propose two methods based on the statistical ranking and selection framework to directly address the selection goal. The proposed methods have an inherent optimization nature in that the selection is optimized according to either a pre-specified minimum correct selection ratio (r^* -selection) or probability of making a correct selection (P^* -selection). These methods are compared with the multiple testing method which controls the tail probability of the proportion of false positives. Both simulation studies and real data applications provide insight into the fundamental difference between the multiple testing methods and the proposed methods in the way of addressing different selection goals. It has been shown that the proposed methods provide a clear advantage over the multiple testing methods when the goal is to select the most significant genes (not all the significant genes). When the goal is to select all the significant genes, the proposed methods perform equally well as the current multiple testing methods. Another advantage provided by the proposed methods is their ability to detect noisy data and therefore suggest no sensible selection can be made.

Keywords

c-best; Microarray; Multiple testing; Probability of correct selection (PCS); Ranking and selection; Tail probability of the proportion of false positive (TPFP)

1 Introduction

Technology has reached a point that massive datasets, such as microarray data, have established themselves as a permanent part of the statistical landscape. In microarray

*corresponding xinping.cui@ucr.edu.

experiments, the goal is often to examine many genes and select some of them for additional investigation. This problem has traditionally been viewed through the prism of multiple testing, in which the selection of a set of genes is formulated as a series of hypothesis tests on the equal distribution of gene expression for all genes under examination. Since tens of thousands of genes are tested simultaneously, the multiplicity problem arises, i.e., the overall Type I error rate of multiple tests will be drastically inflated. Early proposals for dealing with the multiplicity problem focused on the control of family-wise error rate (FWER) or false discovery rate (FDR) (Benjamini and Hochberg, 1995). For more details, see the review by Dudoit et al. 2003. The FWER was quickly recognized as being too conservative for the analysis of microarray data and was extended to a less stringent error rate: generalized family-wise error rate (gFWER) (Dudoit et al., 2004). gFWER is the probability that k ($k \geq 1$) or more false positives occur when multiple tests are conducted. Recently some procedures have been proposed to control gFWER in microarray data analysis (Korn et al., 2004, Van der Laan et al., 2004ab, Lehmann and Romano, 2005, Xu and Hsu, 2007, Romano and Wolf, 2007). Van der Laan et al., (2004b) also suggested the control of the tail probability of the proportion of false positives (TPPFP) among significant findings. On the other hand, the FDR is also recognized as a useful measure of the false positive error rate in microarray studies (Allison et al., 2006). The FDR can be interpreted as the expected proportion of false positives among significant findings. A few multiple testing procedures have been proposed for FDR control under independent or dependent structure of test statistics (Benjamini and Hochberg, 1995, Yekutieli and Benjamini, 1999, Benjamini and Hochberg, 2001). Storey (2002) suggested that the positive false discovery rate (pFDR) is a more reasonable error rate to use in the analysis of microarray data than FDR, which can be interpreted as the Bayesian posterior probability that a significant finding is a false positive. He also devised a multiple testing procedure to control pFDR which has greater statistical power. Conditional false discovery rate (cFDR) (Tsai et al., 2003, Pounds and Cheng, 2004), the total error criterion (Genovese and Wasserman, 2002) and profile information criterion (Cheng et al., 2004) have also been proposed as alternatives to FDR. See Pounds (2006) and Farcomeni (2008) for the most recent reviews.

When the goal of microarray analysis is to discover differentially expressed genes for intense further study, for which case genes of interest can be considered as genes with unequal distribution of gene expression between the two groups, it is appropriate to use p-value based multiple testing procedures with the appropriate error rate controlled. However, for another common purpose of training a prognostic algorithm for predicting risk, such as gene signature as predictor of survival in cancer (Golub et al., 1999, van de Vijver et al., 2002, van't Veer et al., 2002, Pepe et al., 2003), the genes of interest should be those with large (standardized) mean distance in gene expression between the two groups. In other words, testing the null state of equal expression seems less compelling than does ranking the extent of differential expression amongst the genes since very often many genes will be differentially expressed in these studies (Pepe et al., 2003, Bickel, 2004). Sometimes limited resources require investigators to pre-specify a number of genes for further study, regardless of the number of differentially expressed genes (Smyth et al., 2002, Matsui et al., 2008). Therefore, selecting genes based on ranking the extent of differential expression amongst genes is a more directly relevant approach. Moreover, the degree of confidence in the ranking of genes is of more concern than the power of detecting "differentially expressed" genes. Pepe et al. (2003) suggested ranking genes by using receiver operating characteristic (ROC) and partial area under curve (pAUC) as statistical measures of discrimination between distributions of gene expression. ROC and pAUC focuses on the separation of distributions rather than the difference of distributions. They also proposed selection probability $P_g(k) = P[\text{Rank}(g) \leq k]$ to quantify the degree of confidence in choosing the g th gene among the top k . Lu et al. (2009) extended the selection probability for each selected gene to the ReDiscovery Rate (RDR) for the k genes selected, which is defined as the

expected number of top k genes being within the top k again, divided by k . The nonparametric bootstrap method was used to estimate the selection probability and RDR. Zhang et al. (2006) proposed to calculate the probability that an uninformative (not differentially expressed) gene ranks at or above the observed ranking position by a given ranking method. Saeys et al. (2007) reviewed selection techniques. Matsui et al. (2008) was the first group to advocate the application of statistical ranking and selection framework for gene selection and studied the sample size calculation for achieving minimum probability of correct selection. Cui and Wilson (2008a, 2009) showed how to calculate the probability that a given selection is correct for the d -best and g -best selections which are two extensions of the traditional definition of correct selection.

In this article, we propose two new methods for selecting the *best* t of k populations when the goal is to select either the most significant ones or all the significant ones. The key element of both methods lies in the utilization of the probability of c -best correct selection in the framework of ranking and selection. The c -best correct selection corresponds to the "goal I" selection criterion in the fixed subset-size approach by Mahamunulu (1967). Besides the proposed methods directly address a specific selection goal, they also have a unique optimization feature so that the optimal number of most significant genes or all significant genes are selected. Therefore, our methods ensure that the truly top-ranking genes are selected with high probability, which increases the likelihood of biologically relevant results. In what follows, the background of c -best correct selection and the corresponding probability as well as the two new selection methods are given in Section 2. In Section 3, two simulation studies are conducted to compare the proposed selection methods to the TPPFP based multiple testing procedure (MTP) under different selection goals and different levels of noise in data. We selected TPPFP for two reasons: (i) it is representative of a family of increasingly popular multiple testing procedures (Farcomeni, 2008), and (ii) it has the highest amount of mathematical and conceptual similarity and is therefore most suitable for comparison. It will be shown in simulation studies that one of the proposed methods has an advantage over TPPFP when the goal is to select the most significant genes (not all the significant genes), and it performs equally well as TPPFP when the goal is to select all significant genes. Application to two well-studied microarray datasets is given in Section 4 and the paper closes with a discussion in Section 5.

2 c-Best Selection

Consider i.i.d. observations X_{ij} , $j = 1, \dots, n$, arising from population π_i with location parameter family $G(x - \theta_i)$, $i = 1, \dots, k$, denoted by $X_{ij} \stackrel{iid}{\sim} G(x - \theta_i)$. Let $\theta = (\theta_1, \dots, \theta_k)$ and $\theta_{(1)} \leq \dots \leq \theta_{(k)}$ denote the ordered parameters in θ . Assuming θ and the ordering of its elements are unknown, the goal is to select the *best* t populations. Under the traditional ranking and selection framework (Bechhofer, 1954), *best* t is defined as those populations with the largest parameters, $\theta_{(k-t+1)}, \dots, \theta_{(k)}$. Alternatively, *best* t can be defined as those populations with the smallest parameters, $\theta_{(1)}, \dots, \theta_{(t)}$. Suppose the sample statistic $Y_i = Y(X_{i1}, X_{i2}, \dots, X_{in})$ is an estimate of θ_i with continuous cumulative distribution function $F(y - \theta_i)$, $i = 1, \dots, k$. Let $Y_{[1]} \leq \dots \leq Y_{[k]}$ denote the ordered statistics of $Y_1, \dots, Y_k$ and $Y_{(i)}$ denote the sample statistic Y that corresponds to the i th ordered population with parameter $\theta_{(i)}$. Accordingly, the selection procedure, or "rule" R , is

$$R: \text{select } Y_{[k-t+1]}, \dots, Y_{[k]} \text{ as the best } t \text{ populations.}$$

Rule R results in a *correct selection* when the event CS_t occurs, where

$$CS_t = \{\{Y_{[k-t+1]}, Y_{[k-t+2]}, \dots, Y_{[k]}\} = \{Y_{(k-t+1)}, Y_{(k-t+2)}, \dots, Y_{(k)}\}\}.$$

The probability of correctly selecting the best t populations is defined as follows

$$P(CS_t) = P\left(\max_{1 \leq i \leq k-t} Y_{(i)} \leq \min_{k-t+1 \leq j \leq k} Y_{(j)}\right) \quad (1)$$

$$= \int_{-\infty}^{\infty} \prod_{j=k-t+1}^k \bar{F}(y - \theta_{(j)}) d \left\{ \prod_{i=1}^{k-t} F(y - \theta_{(i)}) \right\}, \quad (2)$$

where $\bar{F} = 1 - F$ (see Bechhofer (1954) Equation 8). Alternatively, correct selection, CS_t , can be defined in terms of index sets (Cui and Wilson, 2008a).

Definition 1 Let the indices corresponding to the top t statistics be denoted by set s (for "sample"). Let the indices corresponding to the top t parameters, $\{\theta_{(k-t+1)}, \theta_{(k-t+2)}, \dots, \theta_{(k)}\}$ be denoted by the set A_t . Then the correct selection of t out of k parameters is defined by

$$CS_t = \{s = A_t\}, \text{ for } t = 1, 2, \dots, k.$$

Notice that when the elements of θ are strictly unequal there is only one s satisfying CS_t . In the case of ties, more than one s may satisfy CS_t , thereby increasing $P(CS_t)$. However, $P(CS_t)$ often becomes too small to be useful when k is extremely large and the parameters (and statistics) follow a continuous bounded distribution and are therefore relatively dense (Cui and Wilson, 2008a). Under such a parameter configuration, which is very common in microarray gene expression data, it is very difficult to obtain the unique best t populations. Therefore, we extend the definition of correct selection and the corresponding PCS as follows.

Definition 2 Let the indices corresponding to the top t statistics be denoted by the set s . Let the indices corresponding to the top t parameters be denoted by the set A_t . Then

$$CS_{c,t} = \{s: |s \cap A_t| \geq c\},$$

where $|\cdot|$ denotes cardinality. A set s that satisfies $CS_{c,t}$ is defined to be **c-best**. The probability that a c-best selection occurs $P(CS_{c,t})$ is given as follows:

$$P(CS_{c,t}) = \sum_{g=1}^{|S|} \int_{-\infty}^{\infty} \prod_{j=k-t+1}^k \bar{F}(y - \theta_{s_{g,j}}) d \left\{ \prod_{i=1}^{k-t} F(y - \theta_{s_{g,i}}) \right\}, \quad (3)$$

where S is the set of all c-best sets; $(\theta_{s_{g,1}}, \dots, \theta_{s_{g,(k-t)}}, \theta_{s_{g,(k-t+1)}}, \dots, \theta_{s_{g,k}})$ is the g^{th} combination of parameters $\theta_{(i)}$, $i = 1, 2, \dots, k$ which satisfies $CS_{c,t}$ (Cui and Wilson, 2009).

Remark 1 A c-best set is a set of size t that contains any c or more of the top t populations. c-best corresponds to "goal I" of Mahamunulu (1967).

The probability of making a c -best selection, $P(CS_{c,t})$, may be estimated by using a parametric bootstrap method. Suppose $X_{ij} \stackrel{iid}{\sim} G(x - \theta_i)$, $j=1, \dots, n$, and θ_i is estimated by $Y_i = Y(X_{i1}, X_{i2}, \dots, X_{in}) \sim F(y - \theta_i)$, $i = 1, \dots, k$. For example, $X_{ij} \sim N(\theta_i, \sigma^2)$ and $\hat{\theta}_i = Y_i = \text{avg}(X_{i1}, X_{i2}, \dots, X_{in}) \sim N(\theta_i, \sigma^2/n)$, where σ^2 is known. The estimate of $P(CS_{c,t})$ can be calculated as follows.

For the b th bootstrap, $b = 1, \dots, B$,

1. Obtain a bootstrap sample Y_i^* of statistic Y_i by sampling from $F(y - \hat{\theta}_i)$, $i = 1, \dots, k$.
2. Let $Y_{[1]}^* \leq \dots \leq Y_{[k]}^*$ denote the ordered statistics of $Y_1^*, \dots, Y_k^*$ and $\hat{\theta}_{[1]} \leq \dots \leq \hat{\theta}_{[k]}$ denote the ordered statistic of $\hat{\theta}_1, \dots, \hat{\theta}_k$. Denote the set of indices of populations with $Y_{[k-t+1]}^*, \dots, Y_{[k]}^*$ by s and indices of populations with $\hat{\theta}_{[k-t+1]}, \dots, \hat{\theta}_{[k]}$ by $\hat{A}_t$. Calculate $m_b = |s \cap \hat{A}_t|$, where $|\cdot|$ denotes cardinality.

The estimated $P(CS_{c,t})$ is

$$\widehat{P(CS_{c,t})} = \frac{\sum_{b=1}^B I(m_b \geq c)}{B}$$

The $\widehat{P(CS_{c,t})}$ can also be estimated by using a non-parametric bootstrap method if sampling from $F(y - \hat{\theta}_i)$ is replaced by sampling from data $\{X_{ij}\}_{j=1}^n$. For further details, such as convergence speed and accuracy, see Cui and Wilson (2008b, 2009). The algorithm is implemented in the PCS package on R (www.r-project.org). Note: the parameter "c" in this paper is "L" in the package).

In microarray studies, the goal is often to examine many genes, and select some of them for additional investigation. Although how many genes should be selected is usually unknown, the minimum and maximum number of genes to be selected, as denoted by $T_{\min}$ and $T_{\max}$, can possibly be determined based on the practitioners' research goal, prior knowledge, and cost constraints. Denote t as the number of populations to be selected within a predefined range, $T_{\min} \leq t \leq T_{\max}$. Let c be the minimum number of correctly selected populations, $1 \leq c \leq t$, so that $r = c/t$ (correct selection ratio, CSR). Below we introduce two algorithms that can be used to determine the optimal value of t as well as the t most significant genes. The first algorithm is based on the optimization of $P(CS_{c,t})$ and the second algorithm is based on the optimization of the correct selection ratio r .

Algorithm 1 r^* -selection

Fix CSR r^* , $0 < r^* < 1$. Calculate $P_t = \widehat{P(CS_{\lceil r^* t \rceil, t})}$ for $T_{\min} \leq t \leq T_{\max}$, where $\lceil r^* t \rceil$ denotes the next integer not smaller than $r^* t$ (ceiling function). Select the t with the largest P_t .

Algorithm 2 P^* -selection

Fix PCS P^* , $0 < P^* < 1$. Calculate $r_t = \widehat{P(CS_{c,t})}/t$ for $T_{\min} \leq t \leq T_{\max}$, where $c = \lceil P^* t \rceil$. Select the t with the largest r_t .

Remark 2 Note that in P^* -selection, c is a unique number determined by the value of P^* . In r^* -selection, if for some r^* there are ties of P_t , we choose the largest value of t in order to select more populations while maintaining the fixed CSR r^* .

3 Simulation Study

In microarray studies, the problem of selecting "interesting" genes for additional investigation by examining tens of thousands of genes has traditionally been formulated as a hypothesis testing problem. In this article, we propose two algorithms based on maximizing the probability of correct selection (PCS) or the correct selection ratio for the gene selection problem. The fundamental difference between the proposed framework and the MTP lay in the fact that the former directly addresses the selection problem while the later indirectly solves the selection problem when "interesting" is equivalent to "significant". However, for some microarray studies, more than 50% genes are "significantly" differentially expressed, but not all of them are "interesting" (Pepe et al., 2003) and the application of MTP might lead to biased selection. The following two simulations provide some insight into this difference.

The type I error rates controlled by the MTP include family-wise error rate (FWER) and generalized family-wise error rate (gFWER) which measure the tail probability of the number of false positive and false discovery rate (FDR) which measure the expectation of the proportion of false positives among the rejected hypotheses. The tail probability of the proportion of false positives among the rejected hypotheses (TPPFP) has recently been introduced as another class of type I error rate and defined as follows (Van der Laan et al., 2004b):

$$TPPFP(q) = \Pr\{V/t > q\}$$

where q is a user-supplied constant $q \in (0, 1)$, t is the number of rejected hypotheses, V is the number of false positive among the rejected hypotheses. By definition, if the 'best' is defined as 'significant', it is easy to see that the CSR $r = 1 - V/t$ and $P(CS_{c,t}) = \Pr\{(t - V) > c\}$. Note that both $TPPFP(q)$ and $P(CS_{c,t})$ involve the joint distribution of random variables V and t . Also conditional on t , $TPPFP(1 - c/t) = 1 - P(CS_{c,t})$. Therefore, we consider the $TPPFP$ based MTP to be most comparable to the proposed methods.

In the first simulation, we consider the following configuration of parameters: $\theta = \{\theta_i\}_{i=1}^{5000}$, where $\theta_i = 3$, $i = 1, \dots, 4950$, $\theta_i = 7$, $i = 4951, \dots, 4975$ and $\theta_i = 10$, $i = 4976, \dots, 5000$. Genes with $\theta = 7$ and $\theta = 10$ are considered as differentially expressed genes. For illustration purposes, we also assume that the estimator Y_i of θ_i follow a normal distribution with mean θ_i and common known standard deviation $\sigma = 0.5$ for $i = 1, \dots, 5000$. Therefore, the results are based on a parametric bootstrapping method. However, when there is no distribution assumption for statistics Y_i , a resampling scheme with replacement can easily be implemented to generate bootstrap samples from the empirical distribution of data and similar results can be obtained if the sample size is large (Cui and Wilson, 2009).

To calculate the TPPFP adjusted p-values, the follow steps are taken:

1. Calculate z -scores using $z_i = (Y_i - 3)/0.5$ and the associated raw p-values using the standard normal distribution.
2. Each raw p-value is then adjusted by Sidak step-down procedure for strong control of FWER.
3. Each Sidak adjusted p-value is then adjusted again by "fwer2tpfp" procedure from the multtest package with $q = 1 - r$ for each specified r^* in r^* -selection and calculated r in P^* -selection.

Two significance cutoff values of α for TPPFP adjusted p-value were chosen that correspond to the largest first gap and largest second gap in the p-values. In practice, the value of α is usually specified in advance and it is obvious to see the advantage of r^* -selection and P^* -selection in that the selection is optimized based on data. In order to calibrate the r^* -selection (or P^* -selection) and the α in TPPFP so the methods are on equal footing, α is set according to the gap between adjusted p-values so that the selection using TPPFP adjusted p-value is also optimized.

The simulation was repeated 1,000 times and 100 bootstraps were performed each time for r^* -selection and P^* -selection. Figure 1a is the histogram of sample statistics Y where the populations corresponding to $\theta = 7$ and $\theta = 10$ are clearly separated from the populations corresponding to $\theta = 3$. Figure 1b is the scatter plot of $-\log_{10}(\text{p-value})$ versus the rank of p-values for one simulation with $q = 0.05$, i.e., $r = r^* = 0.95$ (all p-values in Figure 1b are TPPFP adjusted). As expected, the plot shows two gaps in the p-values. The first 30 p-values were near zero, the negative logarithm of which is very large and hence do not appear on the plot. This is the first gap and we therefore set $\alpha = p_{(30)}$ (any value between $p_{(30)}$ and $p_{(31)}$ will lead to the same selection) and declaring the most significant genes to be those corresponding to $p_{(i)}$, $1 \leq i \leq 30$, which is larger than the target of 25. For 1,000 simulations, recorded in Table 1, the average first gap was 35.7 with standard deviation 2.6. When r^* was set to 0.60 or 0.75, TPPFP was considered to be unable to detect the first gap since it occurred at 57.2 for $r^* = 0.60$, and 45.7 for $r^* = 0.75$. The second gap is illustrated in Figure 1b around $t = 53$ using $r^* = 0.95$. Table 1 reports that the largest gap between p-values occurred at $t = 52.6$ with standard deviation 1.0 over 1,000 simulations. For r^* -selection, we calculated $\hat{P}_t$ with $1 \leq t \leq 100$. The values of $\hat{P}_t$ were plotted against t in Figure 1c. When $r^* = 0.90, 0.95$, two peaks appear at $t = 25, 52$. For 1,000 simulations, the average first peaks are 26.9 with standard deviation 2.1 and 25.1 with standard deviation 0.5 respectively and the average second peaks are 53.9 with standard deviation 0.5 and 50.1 with standard deviation 0.6 respectively. For P^* -selection, we calculated $\hat{r}_t$ with $1 \leq t \leq 100$. The values of $\hat{r}_t$ are plotted against t in Figure 1d. Two peaks appear at $t = 25, 50$ for different values of P^* . For 1,000 simulations, the average first peaks and second peaks appear close to 25 and 50 respectively, for different values of P^* .

Table 1 summarizes the mean and standard deviation (sd) of the locations of the two gaps/peaks found by the different methods. If the goal is to select all the significant genes as "interesting" genes for further study, as suggested by Table 1, then both TPPFP based MTP and r^* -selection method perform equally and the performance of P^* -selection method appear to be best and robust. If the goal was to select the most significant genes as "interesting" genes for further study, both r^* -selection method and P^* -selection method perform equally well and outperform TPPFP based MTP.

In the second simulation study, we assessed how r^* and P^* selection methods perform differently from TPPFP in a control-treatment comparison when no genes are differentially expressed and the goal is to select all the significant genes. In particular, we considered the case where a tiny difference observed between the control and treatment groups was not due to a biological reason, but to system bias such as different batches of chips and/or different experiment dates for the control and treatment groups. Although various normalization methods have been proposed to remove such bias, it cannot be completely eliminated due to the complexity of system bias. The simulation study was designed as follows.

The control group had $k = 5000$ means, $\{\theta\}_{i=1}^{500} = 3, \{\theta\}_{i=501}^{1000} = 4, \{\theta\}_{i=1001}^{1500} = 5, \dots, \{\theta\}_{i=4501}^{5000} = 12$. The standard deviations σ_i were chosen so that the coefficient of variation, $CV_i = \sigma_i/\theta_i$, was fixed at $CV_i = 1/12$ for each gene. The σ_i 's were therefore $\theta_i/12$, with values $\frac{3}{12}, \frac{4}{12}, \dots, \frac{12}{12}$.

The treatment group had the same standard deviations as the control group, but the means were shifted by 0.4 : $\eta = \theta + 0.4$. Samples of size $n = 20$ were simulated from normal distributions and Welch t -statistics, Y_i , were calculated for each gene (Figure 2a).

Using $r^* = 0.90$, TPPFP selected an average of 61.8 genes ($sd = 7.8$, determined from 1000 simulations) at the 0.01 level of significance, despite the fact that none should be selected. Using more stringent controls of $r^* = 0.95$ at $\alpha = 0.001$, TPPFP still selected an average of 9.9 genes ($sd = 3.1$). By contrast, r^* and P^* -selection always selected the bound $T_{\max}$. This selection held for all $T_{\max}$ studied, $T_{\max} = 100, 101, \dots, 300$. In other words, the selection decision would change as the prespecified maximum value of t changes. Therefore, r^* and P^* -selection did not consider any particular selection as optimal. A single simulation for r^* -selection is shown in Figure 2c where the graph also indicates no selection should be made. The reason is that for $r^* = 0.75, 0.90$ and 0.95 there is no peak and PCS quickly moves to zero. For $r^* = 0.60$, although PCS is non-zero, there is no peak. It monotonically increases, which indicates that the more genes selected, the more likely the CSR of 0.60 will be achieved. Since PCS monotonically increases with t before it reaches 1, arbitrary $T_{\max}$ would always be selected. For P^* -selection, Figure 2d shows that the data is noisy in the same way as just explained for $r^* = 0.60$. In short, r^*/P^* -selection correctly indicate the user should not make a selection due to too much noise (or no signal) in the data, whereas TPPFP is unable to discriminate the noise and incorrectly selects a relatively large number of erroneous genes. Such phenomena can occur in practice as shown in the Leukemia data example in the next section.

4 Applications

4.1 Apo AI Knock Out Data

Callow et al. (2000) studied the effect of "knocking out" the Apo AI gene in the livers of mice. There were 8 control and 8 treatment cDNA microarrays, from which Welch two-sample t -statistics, Y_i , $i = 1, 2, \dots, 6384$ were calculated (Dudoit et al., 2003). The processed data was taken from Bioconductor's *multtest* library (www.bioconductor.org). The ApoAI data set was selected because it has biologically verified results reported, with extensive statistical multiple comparison work, in Dudoit et al. (2003). The top $t = 8$ genes were verified for biological significance (Dudoit et al., 2003), which implies that this dataset has good signal near the top.

Figures 3a and 3b were obtained after applying r^* -selection and P^* -selection to the data. Instead of assuming the statistics Y_i were Normal with common known variance, we used a shifted Student's- t distribution with the degrees of freedom estimated using the method of moments for each gene (see Cui and Wilson, 2008b). For r^* -selection we used $r^* = 0.90$ with $1 < t < 50$, which selected 6 genes. P^* -selection selected 6 genes at $P^* = 0.90$ and pre-specified $1 < t < 50$ (Figure 3ab). Dudoit et al. (2003) reported 8 genes statistically significant using the maxT procedure (Westfall and Young, 1993) at $\alpha = 0.05$. According to our calculations using maxT, however, only 6 had p-values below $\alpha = 0.05$, while 8 had p-values below $\alpha = 0.10$. Using the same maxT statistics in TPPFP with $\alpha = 0.10$, and $r^* = 0.90$, 8 genes were chosen. Since the top 8 genes are known to be significant all of the above choices are without error. All three of the methods, therefore, perform similarly in this example.

4.2 Leukemia Data

Golub et al. (1999) sought to create a classifier for leukemia tissue into acute lymphoblastic (ALL) or acute myeloid (AML). They obtained a training sample of 27 Affymetrix microarrays from ALL patients and 11 arrays from AML patients. This leukemia dataset was selected because it has come to be used as a benchmark dataset in microarray studies.

To apply r^* -selection and P^* -selection to it, we obtained the processed data on the $k = 3051$ genes from Bioconductor's *multtest* library (www.bioconductor.org) (see Dudoit et al., 2003, p. 90). Instead of assuming the statistics Y_i were normal with common known variance, we used a shifted student's- t distribution with the degrees of freedom estimated using the method of moments for each gene.

In the original paper, Golub et al. (1999) found their classifier to work about the same whether they selected the top 10, 50 or 200 genes. More recently, Park et al. (2007) applied a leave-one-out cross-validation approach with three different selection methods to it (Golub-correlation, Support Vector Machines, and Fisher's Discriminant Analysis). They concluded that different splittings of the training and test sets gave significantly different classification performance results, and that the different selection methods were highly inconsistent with one another. This implies that few genes stand out above the others, which means that the dataset is noisy. On the other hand, Ge et al. (2003) used Welch t -statistics with max T multiple testing adjusted p-values (Westfall and Young, 1993) and concluded 92 genes were significantly differentially expressed at $\alpha = 0.05$. TPPFP found 109, 115, or 138 significant for $r^* = 0.95, 0.90$, or 0.75 . Both r^* -selection and P^* -selection exhibit non-informative behavior. That is, unless one limits the initial selection to small t , e.g. $1 \leq t \leq 5$, then the methods indicate the largest pre-specified value of t should be selected, as discussed in the second simulation of the previous section. For example, if $T_{\max} = 300$, then r^* -selection and P^* -selection would select $t = 300$ (see Figure 3cd). This conclusion would not necessarily be incorrect, accounting for both CSR and PCS; it is just indicative of noisy data. Thus, while there is some tension in the literature as to how many genes really stand out above the others, it is safe to conclude the data is highly noisy. This is in agreement with the original conclusions of Golub et al. (1999), Park et al. (2007), and Dudoit et al. (2003). This does not imply that the $t = 92$ genes declared significant by the max T test are not statistically significant. It does, however, suggest that they may not be truly the best (Cui and Wilson, 2008a) or they may not have biological significance (Bickel, 2004).

5 Discussion

The motivation for this study was the prevalence of multiple testing methods for extremely large k applications where the true goal is not testing, but selection. Although the applications of the present paper focused on microarrays, the theory and methods are fully general. We are currently investigating its application in the identification of microorganisms that are associated with a human disease process. We have investigated two flexible selection methods, r^* -selection and P^* -selection, based upon the PCS statistic, and compared them to the TPPFP based MTP. The comparisons between r^* -selection, P^* -selection, and TPPFP based MTP revealed that, (1) P^* -selection is the robust and most accurate method if the goal is to select all significant genes, while r^* -selection and TPPFP based MTP perform similarly. If the goal is to select most significant genes, then r^* -selection and P^* -selection perform equally well while selection due to TPPFP based MTP is largely biased, (2) When the data is noisy, r^* -selection and P^* -selection are able to send a clear message that selection should not be made whereas TPPFP based MTP failed to do so since any MTP allows for a given level of false positives and it is expected to see some rejections even when the groups do not differ.

In light of the results presented, we recommend r^* -selection or P^* -selection for large k population problems where the user wants to optimize selection according to *either* r^* (for fixed CSR find maximum PCS) *or* P^* (for fixed PCS find maximum CSR), but not both. We also recommend r^* -selection or P^* -selection when the user is uncertain about the noise level of the data. We recommend TPPFP based MTP when the primary goal is to choose the largest number of populations. Relaxing r^* or P^* will generally increase the number

selected by TPPFP based MTP, but not for r^* -selection and P^* -selection because of their optimization nature. As a reviewer pointed out, it is possible that TPPFP could be calculated for different levels of P^* and trends of r^* could be checked in an optimization manner analogous to P^* -selection. However, the TPPFP statistic was designed to select as many genes as possible, within constraints, whereas the PCS statistic is designed to select the best set of genes. Hence, it is not surprising that PCS offers an advantage in an optimization context. Lastly, P^* -selection gives a more consistent selection size for different values of P^* and we recommend it over r^* -selection in general. The gain in optimization quality with r^*/P^* -selection comes at the cost of higher computation time.

In order to apply r^* -selection or P^* -selection, some guidelines for parameter selection can be given. In general, it is best to select $T_{\min} \geq 5$, when possible, because the PCS statistic can be biased upwards for very small values of t . For a discussion of the unbiasedness of the statistic for moderate t , see Wilson (2008). $T_{\max}$ is determined by practical considerations, e.g. "How many genes do we have funds / lab time to investigate when selected?" Consideration should be given regarding what is expected from theory, e.g. small numbers (Apo AI data) or large numbers (Leukemia data)? It has been shown through the examples that since the purpose of r^* -selection and P^* -selection is optimization, a range of r^*/P^* values indicate the same selection decisions, which is a positive feature of these methods.

In this study, we attempted as much as possible to make a fair comparison between the methods. Ultimately, however, this broke down. For example, for TPPFP we used the null hypothesis $\theta_i = 3$, $i = 1, \dots, 5000$, which subtly offers an advantage due to the p-values being calculated from the truth. Although r^* -selection performed sensibly, $\hat{P}$ shrunk rapidly for large r^* under the configurations of θ attempted. In practice, improving estimation of θ could help. Both TPPFP and r^*/P^* -selection are sensitive to different configurations of θ . One approach to improving estimation of the θ employed in this study would be to cluster the data into the number of groups the populations lie in, and let $\hat{\theta}$ be the group means. Improving estimation of θ in general is an area of our current investigation.

Another issue for further research is that studies based on microarray technology have shown inconsistent results when they are executed by different labs or different platforms. This has raised concern about the reproducibility of the modern high-throughput technology (Tan et al., 2003, Marshall 2004, Shi et al., 2006). To evaluate the likelihood that a set of selected populations is reproducible across labs or studies, Lin et al. (2006) proposed the Reproducibility Probability Score method and Zhang et al. (2006) proposed the Probability of Gene's Ranking Position. The c-best selection approach of this paper could also be applied to address reproducibility. When extremely k populations are involved in the ranking and selection procedure, it is virtually impossible to have a unique set of t populations stand out Wilson (2008). However, the chance of observing any c ($< t$) or more of the top t populations can be high. For example, suppose two different labs select different sets of $t = 100$ of genes, each including 50 of the 100 truly best genes ($CS_{50,100}$). Now, while the percentage of genes overlapping between the two labs may be anywhere from 0 to 100%, the results may still be considered reproducible in the sense that the two labs obtained the same rate of 50% of the true top genes. This suggests the gene selection is reproducible as long as the c or more out of t populations selected are among the truly top t populations.

Acknowledgments

The authors would like to thank the two referees for their valuable comments that have led to a considerable improvement in the paper. Thanks also to the Institute for integrative Genome Biology for providing the bioinformatics cluster. This work was supported by NIH R01AI078885 to X. C. and by a grant from the University Of California, Riverside (AES-CE RSAP A01869) to H. Z. and J. W.

References

- Allison DB, Cui XQ, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nature* 2006;7:55–65.
- Bechhofer RE. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics* 1954;25(1):16–39.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 1995;57(1):289–300.
- Benjamini Y, Hochberg Y. The control of the false discovery rate under dependency. *Annals of Statistics* 2001;29:1165–1188.
- Bickel DR. Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics* 2004;20(5):682–688. [PubMed: 15033875]
- Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. Microarray expression profiling identifies genes with altered expression in hdl deficient mice. *Genome Research* 2000;10(12):2022–2029. [PubMed: 11116096]
- Cheng C, Pounds S, Boyett J, Pei D, Kou M-L, Roussel M. Significance threshold selection criteria for massive multiple comparisons with applications to dna microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 2004;3 Article 36.
- Cui X, Wilson J. On the probability of correct selection for large k populations with application to microarray data. *Biometrical Journal* 2008a;50(5):870–833. [PubMed: 18932145]
- Cui, X.; Wilson, J. Technical Report. Vol. 297. Riverside: University of California; 2008b. On how to calculate the probability of correct selection for large k populations.
- Cui X, Wilson J. A simulation study on the probability of correct selection for large k populations. *Communications in Statistics - Simulation and Computation* 2009;38:1244–1255.
- Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science* 2003;18(1):71–103.
- Dudoit S, Van der Laan MJ, Pollard KS. Multiple testing. part i. single-step procedures for control of general type i error rates. *Statistical Applications in Genetics and Molecular Biology* 2004;3(1) article 13.
- Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* 2008;17:347–388. [PubMed: 17698936]
- Ge, Y.; Dudoit, S.; Speed, TP. Technical report. U. C. Berkeley Statistics Department; 2003. Resampling-based multiple testing for microarray data analysis.
- Genovese C, Wasserman L. Operating characteristic and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society* 2002;64:499–517.
- Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E. Molecular classification of cancer class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–537. [PubMed: 10521349]
- Korn E, Troendle J, McShane L, Simon R. Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 2004;124:379–398.
- Lehmann E, Romano J. Generalizations of the familywise error rate. *Annals of Statistics* 2005;33(3): 1138–1154.
- Lin G, He X, Ji H, Shi L, Davis R, Zhong S. Reproducibility probability score-incorporating measurement variability across laboratories for gene selection. *Nature Biotechnology* 2006;24(12): 1476–1477.
- Lu X, Gamst A, Xu R. Rdcurve: a non-parametric method to evaluate the stability of ranking procedures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2009;99(1) doi.10.1109/TCBB.2008.138.
- Mahamunulu DM. Some fixed-sample ranking and selection problems. *Annals of Mathematical Statistics* 1967;38:1078–1091.
- Marshall E. Getting the noise out of gene arrays. *Science* 2004;306(5696):630–631. [PubMed: 15499004]

- Matsui S, Zeng S, Yamanaka T, Shaughnessy J. Sample size calculations based on ranking and selection in microarray experiments. *Biometrics* 2008;64:217–226. [PubMed: 17680829]
- Park CH, Jeon M, Pardalos P, Park H. Quality assessment of gene selection in microarray data. *Optimization Methods and Software* 2007;22(1):145–154.
- Pepe MS, Longton G, Anderson GL, Schummer M. Selecting differentially expressed genes from microarray experiments. *Biometrics* 2003;59:133–142. [PubMed: 12762450]
- Pounds SB. Estimation and control of multiple testing error rates for microarray studies. *Briefings in Bioinformatics* 2006;7(1):25–26. [PubMed: 16761362]
- Pounds S, Cheng C. Improving false discovery rate estimation. *Bioinformatics* 2004;20:1737–1745. [PubMed: 14988112]
- Romano JP, Wolf M. Control of generalized error rates in multiple testing. *Annals of Statistics* 2007;35(4):1378–1408.
- Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;23(19):2507–2517. [PubMed: 17720704]
- Shi L, Reid L, Jones W, Shippy R, Warrington J, Baker S, Collins P, de Longueville F, Kawasaki E, Lee K, et al. The microarray quality control MAQC shows inter- and intra-platform reproducibility of gene expression measurements. *Nature Biotechnology* 2006;24(9):1151–1161.
- Smyth, GK.; Yang, YH.; Speed, T. Statistical issues in cDNA microarray data analysis. In: Brownstein, MJ.; Khodursky, AB., editors. *Functional Genomics: Methods and Protocols*. Totowa, NJ: Humana Press; 2002.
- Storey J. A direct approach to false discovery rates. *Journal of Royal Statistical Society B* 2002;64:479–498.
- Tan P, Downey T, Spitznagel ELJ, Xu P, Fu D, Dimitrov D, Lempicki R, Raaka B, Cam M. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research* 2003;31(19):5676–5684. [PubMed: 14500831]
- Tsai C, Hsueh H, Chen J. Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* 2003;59:1071–1081. [PubMed: 14969487]
- van de Vijver M, He Y, van't Veer L, Dai H, Hart A, Voskuil D, Schreiber G, Peterse J, Roberts C, Marton M, Parrish M, Atsma D, Witteveen A, Glas ALD, van der Velde T, Bartelink H, Rodenhuis S, Rutgers E, Friend S, R B. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 2002;347:1999–2009. [PubMed: 12490681]
- Van der Laan MJ, Dudoit S, Pollard K. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* 2004b;3(1):1–25.
- Van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. part ii. step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology* 2004a;3(1) article 14.
- van't Veer L, Dai H, van de Vijver M, He Y, Hart A, Mao M, Peterse J, van der Kooy K, Marton M, Witteveen A, Schreiber G, Kerkhoven R, Roberts C, Linsley P, Bernards R, Friend S. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–535. [PubMed: 11823860]
- Westfall, PH.; Young, SS. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley and Sons; 1993.
- Wilson, J. PhD thesis. Riverside: University of California; 2008. On the Probability of Correct Selection for Large k Populations.
- Xu H, Hsu J. Applying the generalized partitioning principle to control the generalized familywise error rate. *Biometrical Journal* 2007;49(1):52–67. [PubMed: 17342949]
- Yekutieli D, Y B. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 1999;82:171–196.
- Zhang C, Lu X, Zhang X. Significance of gene ranking for classification of microarray samples. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2006;3(3):312–320. [PubMed: 17048468]

Figure 1.

Graphs of three-point distribution. All graphs are of a single representative simulation (not mean simulation results). a) Histogram of sample statistics Y . b) Plot of $-\log_{10}$ of the TPFP p-values versus t to investigate gaps. c) Plot of maximum PCS $\hat{P}$ versus t at four levels of r^* . d) Plot of maximum CSR $\hat{r}$ versus t at four levels of P^* .

Figure 2.

Graphs of the noise distribution. All graphs are of a single representative simulation (not mean simulation results). a) Histogram of sample statistics Y for the first group b) Plot of $-\log_{10}$ of the TPPFP p-values versus t to investigate gaps. c) Plot of maximum PCS $\hat{P}$ versus t at four levels of r^* . d) Plot of maximum CSR $\hat{r}$ versus t at four levels of P^* .

Figure 3.
Graphs of r^* -selection and P^* -selection results from the ApoAI and leukemia data.

Table 1

Gap locations selected by different methods. The means and standard deviations (in parentheses) of 1000 simulations are shown for given values of r^* and P^* . TPPFP was controlled at the level of r^* and the selection for the first gap was the largest p-value near zero while the selection for the second peak was the largest p-value gap. The selection range for the first peak of the three point distribution with r^*/P^* -selection was $10 \leq t \leq 40$, for the second peak it was $10 \leq t \leq 100$. A ** entry indicates the first gap went undetected and selection signified the second gap.

Gap	r^* (P^*)	TPPFP	r^* -sel	P^* -sel
first	0.6 (0.5)	**	**	25.3 (0.8)
	0.75	**	**	25.7 (0.8)
	0.90	37.0 (2.8)	26.9 (2.1)	25.9 (0.5)
	0.95	35.7 (2.6)	25.1 (0.5)	25.7 (0.6)
second	0.6 (0.5)	84.0 (1.4)	92.4 (3.2)	50.2 (0.5)
	0.75	66.9 (1.3)	66.9 (1.2)	50.5 (0.5)
	0.90	55.6 (1.0)	53.9 (0.5)	50.0 (0.4)
	0.95	52.6 (1.0)	50.1 (0.6)	50.2 (0.5)