Gene Probe expression

Probe rank approaches for gene selection in oligonucleotide arrays with a small number of replicates

Dung-Tsa Chen1,*, James J. Chen2 and Seng-jaw Soong1

1Biostatistics and Bioinformatics Unit, Comprehensive Cancer Center, University of Alabama at Birmingham, 153 Wallace Tumor Institute, 1824 6th Avenue South, Birmingham, AL 35294, USA and 2Division of Biometry and Risk Assessment, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079, USA

Received on August 24, 2004; revised on January 6, 2005; revised on March 27, 2005
Advance Access publication April 6, 2005

ABSTRACT

Motivation: One major area of interest in analyzing oligonucleotide gene array data is identifying differentially expressed genes. A challenge to biostatisticians is to develop an approach to summarizing probe-level information that adequately reflects the true expression level while accounting for probe variation, chip variation and interaction effects. Various statistical tools, such as MAS and RMA, have been developed to address this issue. In these approaches, the probe level expression data are summarized into gene level data, which are then used for downstream statistical analysis. Since probe variation is often larger than chip variation and there is also a potential interaction effect between probe affinity and treatment effect, strategies such as a gene level analysis, may not be optimal. In this study, we propose a procedure to analyze probe level data for selecting differentially expressed genes under two treatment conditions (groups) with a small number of replicates. The probe level discrepancy between two groups can be measured by a difference of the percentiles of probe perfect-match (PM) ranks or of probe PM weighted ranks. The difference is then compared with a pre-specified threshold to determine differentially expressed genes. The probe level approach takes into account non-homogenous treatment effects and reduces possible cross-hybridization effects across a set of probes.

Results: The proposed approach is compared with MAS and RMA using two benchmark gene array datasets. Positive predictivity and sensitivity are used for evaluation. Results show the proposed approach has higher positive predictivity and higher sensitivity.

Availability: Available on request from the authors.
Contact: dtchen@uab.edu

1 INTRODUCTION

Microarray gene chip technology has advanced genomic research through analyzing mRNA expression (Baer et al., 2004; Gauthier et al., 2004; Grant et al., 2004; Yeatman, 2003). On account of dramatic reduction in labor, time and costs, this technique becomes a popular tool for studying thousands of genes simultaneously. Gene expression profiling through this technology has great promise in the biomedical area. For example, the microarray technology can be used in identification of biomarkers, evaluation of prognoses, classification of disease status and prediction of clinical outcomes (Benimetskaya et al., 2004; Nevins et al., 2003; Schaefer et al., 2004; Sotiriou et al., 2003).

While this technology has facilitated genomic research, identification of differentially expressed genes poses a unique challenge (Hsieh et al., 2003; Irizarry et al., 2003; Strand et al., 2002; Mutch et al., 2002). The oligonucleotide array uses a set of probes to interrogate a gene expression and each array consists of thousands of genes. A microarray experiment routinely collects a huge volume of information; thus, the data structure can be quite complicated. Analyzing such complex data pose a challenge to biostatisticians to develop an approach to summarizing probe-level information that truly reflects the level of a gene expression adequately while accounting for probe variation, chip variation and interaction effects. In addition, owing to resource limitation and/or sample availability, many microarray experiments, such as in vitro study, have only a small number of replicates. In such cases, statistical inferences such as the P-value significance testing or confidence interval analysis, which work well in a large sample size, often break down and become impractical.

Various statistical tools have been developed to address this issue. For example, MAS 5 (Affymetrix, 2002) employs the Tukey’s Biweight approach to summarizing a gene-expression intensity from the perfect-match (PM) and mismatch (MM) signals. Dchip (Li and Wong, 2001a) analyzes probe level intensities using a multiplicative model to decompose each probe signal into a product of gene-expression index and probe-sensitivity index. RMA (Irizarry et al., 2003) uses a log scale linear additive model in the background corrected and normalized PM intensities to estimate gene expression. These approaches can be referred to as the gene-level-based approach in which the probe level expression data are summarized into gene level measures, which then are used for a downstream statistical analysis. One advantage of this strategy is that the data dimensions are reduced into a manageable scale. Standard statistical methods, such as the t-test, can be used to select differentially expressed genes. However, this strategy has some limitations. For example, some studies have found that probe variation is often larger than chip variation (Li and Wong, 2001b; Irizarry et al., 2003). Figure 1A of a gene-expression profile displays heterogeneity of probe effects in three arrays. All three arrays show similar gene profiles, which suggest small array variation. In contrast, the range of probe intensities is quite large: some probes have low intensity, whereas others have

*To whom correspondence should be addressed.
patterns suggest that MM also partially measures RNA concentration. The intensity of MM is supposed to be greater than the intensity of PM by hybridization, the intensity of PM is supposed to be greater than the intensity of MM. Consequently, PM is used to detect mRNA concentration of a target gene. Since PM is also contaminated with background and cross-hybridization, the proposed procedure is based on the two-stage rank approach using the probe PM rank, instead of the PM intensity for data analysis (Chen et al., 2002). Thus, inclusion of MM with PM probably underestimates mRNA concentration.

The proposed procedure focuses on a two-group comparison with a small number of replicates. We compare the proposed approaches with the MAS and RMA methods using two benchmark datasets, an Affymetrix Latin square dataset (U95A) and a Gene Logic Spike-in dataset. Both datasets have been used to develop and validate MAS and RMA. Since the concentrations of spike-in genes (treated as known differentially-expressed genes) are known in these datasets, we are able to estimate positive predictivity (i.e. the predictive value of a positive test) and sensitivity of these approaches. Thus, we use the two measures to evaluate performance.

2. METHODS

The proposed procedure is based on the two-stage rank approach using the probe PM rank, instead of the PM intensity for data analysis (Chen et al., 2004). The PM rank is defined as the rank of PM intensity over all probes in the gene chip. The rationale is that the rank analysis provides a better treatment for alleviating the effects of extreme values that can be serious in oligonucleotide array data. We present a (un-weighted) rank and a weighted rank approach described below.

2.1 Probe rank approach

Let \( Y_{i,j} \) be a PM rank for the \( j \)-th probe in the \( i \)-th gene on array \( k \). Assume group \( A \) has arrays \( a_1, \ldots, a_{k_A} \) and group \( B \) has arrays \( b_1, \ldots, b_{k_B} \) (e.g. treatment versus control groups). Consider a difference of the two percentiles from the two groups.

\[
D_{A,B}^{[\alpha,\beta]} = \frac{P_{\beta} - P_{\alpha}}{n} \left[ \frac{1}{n_A} \sum_{i=1}^{n_A} Y_{i,j,a_i} - \frac{1}{n_B} \sum_{i=1}^{n_B} Y_{i,j,b_i} \right]
\]

where \( P_{\alpha} \) could be 0 (i.e. the minimum), \( P_{\beta} \) could be 1 (i.e. the maximum), and \( n \) is the size of each group. The proposed procedure focuses on a two-group comparison with a small number of replicates. We compare the proposed approaches with the MAS and RMA methods using two benchmark datasets, an Affymetrix Latin square dataset (U95A) and a Gene Logic Spike-in dataset. Both datasets have been used to develop and validate MAS and RMA. Since the concentrations of spike-in genes (treated as known differentially-expressed genes) are known in these datasets, we are able to estimate positive predictivity (i.e. the predictive value of a positive test) and sensitivity of these approaches. Thus, we use the two measures to evaluate performance.

2. METHODS

The proposed procedure is based on the two-stage rank approach using the probe PM rank, instead of the PM intensity for data analysis (Chen et al., 2004). The PM rank is defined as the rank of PM intensity over all probes in the gene chip. The rationale is that the rank analysis provides a better treatment for alleviating the effects of extreme values that can be serious in oligonucleotide array data. We present a (un-weighted) rank and a weighted rank approach described below.

2.1 Probe rank approach

Let \( Y_{i,j} \) be a PM rank for the \( j \)-th probe in the \( i \)-th gene on array \( k \). Assume group \( A \) has arrays \( a_1, \ldots, a_{k_A} \) and group \( B \) has arrays \( b_1, \ldots, b_{k_B} \) (e.g. treatment versus control groups). Consider a difference of the two percentiles from the two groups.

\[
D_{A,B}^{[\alpha,\beta]} = \frac{P_{\beta} - P_{\alpha}}{n} \left[ \frac{1}{n_A} \sum_{i=1}^{n_A} Y_{i,j,a_i} - \frac{1}{n_B} \sum_{i=1}^{n_B} Y_{i,j,b_i} \right]
\]

where \( P_{\alpha} \) could be 0 (i.e. the minimum), \( P_{\beta} \) could be 1 (i.e. the maximum), and \( n \) is the size of each group. The proposed procedure focuses on a two-group comparison with a small number of replicates. We compare the proposed approaches with the MAS and RMA methods using two benchmark datasets, an Affymetrix Latin square dataset (U95A) and a Gene Logic Spike-in dataset. Both datasets have been used to develop and validate MAS and RMA. Since the concentrations of spike-in genes (treated as known differentially-expressed genes) are known in these datasets, we are able to estimate positive predictivity (i.e. the predictive value of a positive test) and sensitivity of these approaches. Thus, we use the two measures to evaluate performance.

Fig. 1. Heterogeneity of probe effects and interaction between probe and treatment effects. (A) displays heterogeneity of probe effects for a gene profile in three replicate arrays (different line for each array). The probe set for this gene has 16 probes. All three arrays show a similar gene profile across the 16 probes. Such similar patterns suggest small array variation. On the other hand, variation among probe intensities is quite large. For example, probes 2, 3 and 9 have high intensity, but probes 1, 6, 7 and 8 have low intensity (at least 16-fold change between the lowest and highest intensity). The pattern indicates heterogeneity of probe effects. (B) shows interaction between probe and treatment effects for a gene profile in two control arrays (denoted by two solid lines) and the two treated arrays (denoted by two dashed lines). The probe set for this gene has 16 probes. The first half of the probes shows higher expression in the treated arrays than in the control arrays. However, such expression differences disappear for the other probes. The pattern of partial differential expressions indicates interaction between probe and treatment effects (i.e. changes in expression due to treatment depend on the probe effects).

High intensity. Such large variation among probes in a gene set indicates simple summary statistics, such as mean or median, may not be sufficient to reflect the full information of expressions. Another problem occurs when there is an interaction effect between probe affinity and treatment effect (Chen et al., 2004). For example, alternative splicing, one form of interaction, occurs in 40% of human genes (Modrek and Lee, 2002) and yields various functional forms for the genes. Since alternative splicing causes some sequences to be truncated or modified, this feature is likely to result in some probes being unable to bind to the target gene and thus produces low intensity for these probes in oligonucleotide gene chips. Figure 1B illustrates one example of an interaction between probe and treatment effects. Large expression differences between treatment and control arrays occur in the first half of the probes, but the remaining probes show similar expressions. Therefore, the use of gene level analysis is likely to miss the target gene that may have potential biological implications, such as alternative splicing.

In this paper, we propose a procedure to analyze PM probe level data for selection of differentially expressed genes as an alternative to the gene level analysis. Here we consider only PM probe level data because of potential biases introduced by the use of MM. MM is designed to identify background intensity and cross-hybridization, whereas PM is used to detect mRNA concentration of a target gene. Since PM is also contaminated with background and cross-hybridization, the intensity of PM is supposed to be greater than the intensity of MM. However, it is common to see a substantial portion of probe pairs with MM > PM (30–50%) and a high correlation between PM and MM (Chen et al., 2002; Irizarry et al., 2003). Such patterns suggest that MM also partially measures RNA concentration.

Thus, inclusion of MM with PM probably underestimates mRNA concentration.

The proposed procedure focuses on a two-group comparison with a small number of replicates. We compare the proposed approaches with the MAS and RMA methods using two benchmark datasets, an Affymetrix Latin square dataset (U95A) and a Gene Logic Spike-in dataset. Both datasets have been used to develop and validate MAS and RMA. Since the concentrations of spike-in genes (treated as known differentially-expressed genes) are known in these datasets, we are able to estimate positive predictivity (i.e. the predictive value of a positive test) and sensitivity of these approaches. Thus, we use the two measures to evaluate performance.
be 100 (i.e. the maximum) when sample size is small (e.g. 2 or 3 in an in vitro study), and $n$ represents the total number of probes in an array (e.g. there are 201,807 probes in a HG-u95A gene chip). $D_{i}^{A,B}$ is a measure of difference between groups A and B. The measure between groups B and A is defined similarly. Denote the number of probes for gene $i$ as $J_i$. A gene $i$ is considered to be an altered gene if

$$\frac{1}{J_i} \sum_{j=1}^{J_i} \left( b_{i,j}^{(B)} - b_{i,j}^{(A,B)} \right) > \left( \sum_{j=1}^{J_i} b_{i,j}^{(A)} \right)$$

or

$$\frac{1}{J_i} \sum_{j=1}^{J_i} \left( b_{i,j}^{(B)} - b_{i,j}^{(B,A)} \right) > \left( \sum_{j=1}^{J_i} b_{i,j}^{(B)} \right),$$

where $b_{i,j}^{(E)}$ represents a probe level threshold with $E$ as $(A, B)$ or $(B, A)$. $b_{i,j}^{(E)}$ represents a gene level threshold, and $I$ is an indicator function. The two thresholds, $b_{i,j}^{(E)}$ and $b_{i,j}^{(E)}$, are prespecified constants. The probe level threshold $b_{i,j}^{(E)}$ provides a cutoff for probe discrepancy between the two groups, and the gene level threshold, $b_{i,j}^{(E)}$, provides a cutoff for determining differential expressions.

### 2.2 Probe weighted rank approach

Based on our experiences of microarray data analysis, the probe rank has limitations for detecting expression differences for genes with extremely high intensity (e.g. in the 98th–100th percentile). For genes within this range, their ranks tend to be similar. As a result, it becomes difficult to identify altered genes in this range because of a very small rank difference. In practice, this situation is rare and probably occurs in a high abundance gene. In this case, the gene intensity is likely in the range of high percentile at each experimental group (i.e. treatment and control). Though the difference of intensity between the two experimental groups may be substantial (e.g. >2-fold for treatment versus control), the rank difference remains relatively small. To overcome the problem, we introduce a weighting factor to the rank. By giving the weighting factor, a gene with high intensity is likely to have its rank score amplified substantially, because the weighting factor is proportional to the probe intensity. As a result, the difference of the weighted rank scores becomes large between treatment and control groups.

Mathematically, the weighted rank approach is the same as the rank approach except $Y_{i,j,k}$ is multiplied by a weight $w_{i,j,k}$. The weight $w_{i,j,k}$ for the $j$-th probe in the $i$-th gene on array $k$, is defined as $\log_2(P_{M_{i,j,k}}/\sum_{m=1}^{M} \sum_{j=1}^{J_i} \log_2(P_{M_{m,j,k}})$ where PM is intensity for the $j$-th probe in the $i$-th gene of array $k$, $M$ is the total number of genes in an array, and $J_i$ represents the total number of probes for gene $i$. The PM weighted rank $Y_{i,j,k}^{(weighted)}$ becomes $Y_{i,j,k} \times w_{i,j,k}$. The probe weighted rank approach uses $Y_{i,j,k}^{(weighted)}$ to compute percentile difference (i.e. $D_{i,j}^{A,B}$ and $D_{i,j}^{B,A}$) and selects differentially expressed genes. By giving a larger weight on high intensity probes, the probe weighted rank approach can increase the power of detecting expression differences for high abundant genes better than the rank approach does.

### 2.3 Criteria for evaluation

We consider the two diagnostic measures, positive predictivity and sensitivity, to evaluate the performance of the proposed methods and compare it with the RMA and MAS methods. Positive predictivity is defined as the proportion of the number of selected (truly) altered genes over the total number of selected genes. Sensitivity is defined as the proportion of the number of selected (truly) altered genes over the total number of altered genes. Note that $1$−positive predictivity is the proportion of false positive genes among the selected genes, an empirical false discovery rate. A positive predictivity value close to 1 indicates that a selected gene is likely to be truly differentially expressed, but the number of selected altered genes may be few. For example, let us suppose there are 20 altered genes and 1000 non-altered genes. If a test selects only 1 altered gene and 0 non-altered genes, then positive predictivity will be 1, but the test will miss the other 19 altered genes. On the other hand, a sensitivity value close to 1 suggests the procedure has selected most altered genes, but the number of false positive genes may be large. Using the above example, if a test selects all 20 altered genes as well as all 1000 non-altered genes (i.e., all 1020 genes are selected), then the sensitivity will be 1, but positive predictivity will be only 1.67%. Since the relation between positive predictivity and sensitivity is not straightforward, it is challenging to find a method with high positive predictivity and high sensitivity values for gene selection.

### 3 IMPLEMENTATION

#### 3.1 Data examples

Two benchmark oligonucleotide array datasets are used for evaluation.

1. **Affymetrix Latin Square dataset (59 arrays):** The human data set (Affymetrix HG-u95A gene chip) consists of a series of genes spiked-in at known concentrations and arrayed in a Latin Square format. They represent a subset of the data used to develop and validate the Affymetrix Microarray Suite (MAS) 5.0 algorithm. The Latin Square design was made by 14 spiked-in gene groups in 14 experimental conditions. The concentration of the 14 gene groups in the first experiment was 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM. Each subsequent experiment rotated the spike-in concentrations by one group. Each experimental condition had two or three replicates. However, two conditions were repeated four times with three replicates each time. Therefore, we had 20 groups used for pairwise comparison to study gene selection. After removing comparisons between replicated groups, there were 178 pairwise comparisons. For detailed information, please see the website: http://www.affymetrix.com/support/datasets.affx.

2. **Gene Logic Spike-in datasets** ([http://qcelotus02 genie.com/datasets.nsf](http://qcelotus02 genie.com/datasets.nsf)): Three datasets were generated from the experiments where 10 or 11 bacterial control cRNAs (i.e. spike-in genes) were spiked to Affymetrix HG-u95A gene chips at dilution series and Latin square series. We pool the three datasets for evaluation. Below is a brief description for each dataset.

- (a) Spike-in dataset (19 arrays): Ten spike-in genes were hybridized at the same concentration on each gene array. There were 14 concentrations used for the 10 spike-in genes in the experiments: 0, 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, 100 and 150 pM. Each experiment had one concentration for all 10 spike-in genes (e.g. experiment 1 had 0.5 pM for the 10 spike-in genes and experiment 2 had 0.75 pM for the 10 spike-in genes). There were seven experimental groups with more than one replicate (i.e. groups with a concentration of at least 5 pM for the spike-in genes). We use the seven experimental groups to perform pairwise comparisons (21 comparisons) for gene selection.

- (b) Acute myeloid leukemia (AML) dataset (32 arrays): There were 11 spike-in genes in a hybridization mix to prepare 11 samples from an AML tumor cell line. An incomplete Latin Square experiment was carried out in the 11 samples with 12 concentrations for the spike-in genes: 0.5, 1, 1.5, 2, 3, 5, 12.5, 25, 37.5, 50, 75 and
Fig. 2. Sensitivity and positive predictivity. (A) is the result of sensitivity versus positive predictivity for Affymetrix Latin dataset. (B) is for Gene Logic dataset. In each figure, a solid curve (-o-) represents the rank approach, a dashed curve (-Δ-) is for the weighted rank approach, a dotted curve (-+-) is for MAS and a dot-dash curve (-×-) is for RMA. The numbers displayed at each curve are either probe level thresholds or cutoff points of \( P \) values. The pattern of curves of sensitivity versus positive predictivity indicates positive predictivity is higher in both rank approaches than in the RMA and MAS for a fixed sensitivity.

---

100 pM (e.g. sample 1 had 0.5, 1, 1.5, 2, 3, 5, 12.5, 25, 37.5, 50 and 75 pM for spike-in genes A1–A11, respectively; and sample 2 had 1, 1.5, 2, 3, 5, 12.5, 25, 37.5, 50, 75 and 100 pM for spike-in genes A1–A11, respectively). Each sample group had two or three gene arrays. We use the 11 sample groups for pairwise comparisons (55 comparisons).

(c) Tonsil dataset (36 arrays): The experimental design is similar to the one for the AML dataset. The same 11 spike-in genes were spiked into a hybridization mix to prepare 12 samples from a tonsil tissue. There were 12 concentrations arranged in an incomplete Latin square experiment: 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, and 100 pM. Each sample group had 3 gene arrays. The number of pair-wise comparisons for the 12 sample groups is 66.

3.2 Evaluation

We compare RMA, MAS, probe rank, and probe weighted rank approaches with respect to sensitivity and positive predictivity. Gene selection criteria are briefly described as follows. For the rank and weighted rank approaches, we set \( P_a \) as 0 and \( P_b \) as 100 to compute \( D_{i,j}^{\text{A,B}} \). The gene level threshold \( P_{\text{gene,E}} \) is fixed at 50%. Given a probe-level threshold (e.g. 0.05), we identify a set of genes with differential expression between two experimental groups. The sensitivity and positive predictivity values are then calculated based on the results of gene selection. Since there are multiple experimental groups in each dataset, we calculate sensitivity and positive predictivity values for all pairwise comparisons among groups. Then we compute the average of sensitivity values and the average of positive predictivity values, denoted by \( \text{Sensitivity}_{\text{avg}} \) and \( \text{PP}_{\text{avg}} \), respectively. By varying the probe-level threshold from 0 to 1, we can construct a curve for the \( \text{Sensitivity}_{\text{avg}} \) versus \( \text{PP}_{\text{avg}} \) plot. For RMA and MAS, gene-level intensity measurements are generated and normalized using the default values by Bioconductor (Irizarry et al., 2003). The two-sample \( t \)-test is then applied with a \( P \) value cutoff (i.e. significance level), such as 0.01, to select differentially expressed genes. The curve of the \( \text{Sensitivity}_{\text{avg}} \) versus \( \text{PP}_{\text{avg}} \) for RMA and MAS is constructed by varying different \( P \) value cutoffs.

3.3 Results

The plots of \( \text{Sensitivity}_{\text{avg}} \) versus \( \text{PP}_{\text{avg}} \) are shown in Figure 2A and B for the two datasets. Each figure consists of four curves from the four methods for various \( P \) value cutoffs or probe level thresholds. The four methods show a similar ‘A’-shape pattern. That is, both \( \text{Sensitivity}_{\text{avg}} \) and \( \text{PP}_{\text{avg}} \) initially increase (starting at 0), and after reaching the peak or plateau, \( \text{PP}_{\text{avg}} \) begins to decrease as \( \text{Sensitivity}_{\text{avg}} \) becomes large. However, the four curves appear to have two distinct patterns. RMA and MAS form an unbalanced bell shape. The curve is truncated at 0 and has a long tail on the right side (i.e. skew to the right). When \( \text{Sensitivity}_{\text{avg}} \) is ~0.2, \( \text{PP}_{\text{avg}} \) reaches maximum at different values (0.3–0.7) in the two datasets. The \( \text{PP}_{\text{avg}} \)
is close to 0 for a Sensitivityavg value >0.8 in most cases. In addition, a different P value cutoff can result in substantial changes of PPavg and Sensitivityavg. For example, the cutoff of P value at 0.0001 yields the highest PPavg, but a low Sensitivityavg. In contrast, when the cutoff P value increases to 0.05, Sensitivityavg becomes large, but PPavg is ≈0.

For the rank and weighted rank approaches, the corresponding curves show a plateau pattern. The maximum value of PPavg is >0.8 in most cases. The curves often reach the plateau at Sensitivityavg = 0.2. However, the plateau ends (i.e. PPavg begins to decrease) at different sensitivity values (Sensitivityavg = 0.8 and 0.5 for the Affymetrix Latin square dataset and the GeneLogic dataset, respectively). Moreover, a different probe level threshold also results in different PPavg and Sensitivityavg in both rank approaches. Specifically, when the probe level threshold is very small (i.e. the threshold <0.001 and <0.01 for the rank and weighted rank approaches, respectively), PPavg and Sensitivityavg do not change much. In contrast, when the probe level threshold becomes large (e.g. from 0.01 to 0.2), the degree of decrease in Sensitivityavg is larger than that of increase in PPavg. Overall, the weighted rank approach often has a higher PPavg value than the rank approach given the same Sensitivityavg value.

4 DISCUSSION

The goal of this study is to propose a new approach to improve the selection of altered genes when the replicates are small. As we know, microarray technology is often used as an exploratory tool to select potentially altered genes. Owing to some uncertainty of the technology (e.g. chip variation, probe affinity and sample preparation), the selected genes require confirmatory tests (e.g. RT–PCR) to validate the results. However, because of technical, cost and time limitations, often only a small portion of the selected genes is chosen for confirmatory tests. On the other hand, the number of genes selected by a bioinformatical method is often larger than the number that biological researchers can perform in confirmatory tests. Therefore, a procedure with most altered genes selected (i.e. high sensitivity) and with few or none of the false positive genes selected (i.e. high positive predictivity) is desirable for gene selection. For this reason, we consider the use of both sensitivity and positive predictivity to examine our approaches. The curve of sensitivity versus positive predictivity, in fact, is equivalent to the conventional ROC curve (i.e. plot of sensitivity versus 1–specificity or plot of true positive versus false positive). However, the conventional ROC curve focuses on one aspect of evaluation of a screening test without considering the other curve (e.g. a high sensitivity does not guarantee a high or low positive predictivity). In contrast, we present a unique plot of sensitivity versus positive predictivity to effectively evaluate a gene selection procedure.

Figures 2A and B show that our rank approaches have higher positive predictivity values than the other two approaches given a fixed sensitivity. Specifically, the highest positive predictivity value is 70–90% for the rank approaches versus 30–70% for the RMA and MAS in the two datasets. Moreover, the highest positive predictivity occurred at a larger sensitivity value in both rank approaches than in the RMA and MAS. Taking both together, it suggests that the rank approaches have a higher sensitivity and higher positive predictivity to ensure that altered genes are selected and the selected genes are truly differentially expressed.

In addition, the plateau pattern in both rank approaches suggests that positive predictivity is in the robust to medium range of sensitivity. That is, we can maintain a high positive predictivity while increasing sensitivity from low medium to high medium, such as from 40 to 60%. Furthermore, when the probe level threshold becomes small, such as 0.001, positive predictivity and sensitivity become stable. For the Affymetrix dataset, positive predictivity is at 60–70% and sensitivity is ~80%. For the Gene Logic dataset, positive predictivity is 50–60% and sensitivity is 70%. This observation indicates that without any prior information, the use of a small scale of probe level threshold will still yield reasonable positive predictivity and sensitivity values. Finally, since the plateau is higher in the weighted rank approach than in the rank approach, the weighted rank approach has better performance than the rank approach does in yielding high positive predictivity. One possible explanation is that the weighted rank approach gives a larger weight in high intensity so that the approach has a higher power to detect expression differences for high abundant genes.

In contrast, the RMA and MAS approaches show that a cutoff P value at 0.001–0.0001 yields the highest positive predictivity but a small sensitivity. As the cutoff P value increases, sensitivity increases, but positive predictivity decreases rapidly to 0. The result indicates that the use of a simple t-test with or without P value adjustment for multiple testing does not help secure a high positive predictivity and a high sensitivity simultaneously for the RMA and MAS approaches.

One may wonder why the four approaches have both sensitivity and positive predictivity decrease to 0 as the cutoff P value becomes small or the probe level threshold becomes large. This pattern seems counterintuitive because a very low P value as the cutoff should reduce sensitivity, but increase positive predictivity as shown in Table 1. This can be explained from Bayes’ rule and uncertainty of microarray technology. Let Y denote a binary variable for the status of altered gene or non-altered gene (i.e. Y = 1 if it is an

Table 1. Behavior of sensitivity versus positive predictivity as a function of smaller P-value in altered genes*

<table>
<thead>
<tr>
<th>Threshold (P value)</th>
<th>10^-1</th>
<th>10^-2</th>
<th>10^-3</th>
<th>10^-4</th>
<th>10^-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>3/20 = 0.15</td>
<td>3/20 = 0.15</td>
<td>3/20 = 0.15</td>
<td>2/20 = 0.10</td>
<td>1/20 = 0.05</td>
</tr>
<tr>
<td>Positive predictivity</td>
<td>3/5 = 0.60</td>
<td>3/4 = 0.75</td>
<td>3/3 = 1</td>
<td>2/2 = 1</td>
<td>1/1 = 1</td>
</tr>
</tbody>
</table>

*Suppose the total number of altered gene is 20 and the 5 smallest P values are 2 × 10^-2, 2 × 10^-3, 2 × 10^-4, 2 × 10^-5 and 2 × 10^-6. Also assume the first two P-values belong to non-altered genes and the last three are for the altered genes. Results in the Table show that as sensitivity goes down to 0, positive predictivity becomes non-decreasing for the use of very small P-values as the cutoff.
altered gene; otherwise \( Y = 0 \). And \( T \) represents another binary variable for the status of a gene being selected by a test (i.e. \( T = 1 \) if selected; otherwise \( T = 0 \)). Then positive predictivity can be expressed as \( \text{Prob}(Y = 1|T = 1) \), sensitivity as \( \text{Prob}(T = 1|Y = 1) \) and specificity as \( \text{Prob}(T = 0|Y = 0) \). By Bayes’ theorem, positive predictivity is a function of sensitivity, 1—specificity, and the probabilities \( \text{Prob}(Y = 1) \) and \( \text{Prob}(Y = 0) \):

\[
\begin{align*}
\text{prob}(Y = 1|T = 1) &= \frac{\text{prob}(T = 1|Y = 1) \times \text{prob}(Y = 1)}{\text{prob}(T = 1|Y = 1) \times \text{prob}(Y = 1) + (1 - \text{prob}(T = 0|Y = 0)) \times \text{prob}(Y = 0)}.
\end{align*}
\]

Given a very low \( P \) value as the cutoff, sensitivity is close to 0. If the measurement errors of microarray technology cause the selection of some false positive genes selected, 1—specificity is near to 0, but not equal to 0. Under the assumption of a small portion of altered genes (i.e. \( \text{prob}(Y = 1) \geq 0 \)), the proportion of non-altered genes is near to 1 (i.e. \( \text{prob}(Y = 0) \geq 1 \)). When sensitivity decreases to 0, 1—specificity will dominate the denominator. As a result, positive predictivity also decreases to 0.

Since the testing datasets have a small number of replicates (e.g. 2 or 3) in each group, the \( P \)-value significance testing approach may break down in the case of small sample sizes. This can be seen in the \( t \)-test for MAS and RMA. In contrast, our rank approaches employ the simple statistics of rank percentile difference, \( D_{A,B} \) or \( D_{B,A} \), to measure probe expression discrepancies. The large positive values imply the differential expressions for the corresponding probe. For studies using a few samples, such as in vitro studies or in the testing datasets, because the biological variation is well controlled, the two parameters \( P_A \) and \( P_B \) can be setup as 0 and 100, respectively. In other words, we compare the smallest rank percentile in one group with the largest rank percentile in the other group. For studies involving large sample sizes, \( P_A \) and \( P_B \) could be changed to 20–30 and 70–80 (e.g. lower quartile and upper quartile), respectively, to eliminate outliers.

In this evaluation, the rank and weighted rank approaches use 50% as the gene level threshold. The reason for the use of this setting (i.e. 50%) is to control for cross-hybridization effect and meanwhile, to increase the possibility of identifying genes due to alternative splicing. If this threshold is far <50% (e.g. 20%), the chance will increase in selecting false positive genes owing to cross-hybridization. In practice, we find the use of 50% for the gene level threshold is workable in the testing datasets and in our previous studies. As the microarray technology improves, especially in the control of cross-hybridization effect, the threshold value can be reduced.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Cancer Institute (5P30 CA-13148, 1P50 CA89019, P50 CA83591 and U54 CA100949). The authors gratefully acknowledge the helpful comments by the two anonymous referees. We thank Laura Gallitz for the secretarial assistance.

REFERENCES


D.-T.Chen et al