

MAOSA: A new procedure for detection of differential gene expression

Greg Dyson^{a,*}, C.F. Jeff Wu^b

^a *Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA*

^b *School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA*

Received 5 March 2005; received in revised form 24 August 2005; accepted 25 August 2005

Abstract

Gene expression data analysis provides scientists with a wealth of information about gene relationships, particularly the identification of significantly differentially expressed genes. However, there is no consensus on the analysis technique that will solve the inherent multiplicity problem (thousands of genes to be tested) and yield a reasonable and statistically justifiable number of differentially expressed genes. We propose the Multiplicity-Adjusted Order Statistics Analysis (MAOSA) to identify differentially expressed genes while adjusting for the multiple testing. The multiplicity problem will be eased by performing a Bonferroni correction on a small number of effects, since the majority of genes are not differentially expressed.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Microarray; Multiplicity; Order statistics

1. Introduction

High-throughput gene expression data has enabled scientists to simultaneously study a large number of gene effects via spotted glass arrays and oligonucleotide gene chips. However, there is no consensus on the choice of technique to analyze these data that will solve the multiplicity problem and yield a reasonable and statistically justifiable number of differentially expressed genes. A *multiplicity* problem arises whenever more than one hypothesis test is conducted. When conducting 100 hypothesis tests at the 0.05 level, we should expect 5 tests to be rejected even if all the null hypotheses are true. The Multiplicity-Adjusted Order Statistics Analysis (MAOSA) algorithm proposed here approaches the problem by first transforming *t*-like statistics to the

* Corresponding author.

E-mail addresses: gdyson@umich.edu (G. Dyson), jeff.wu@isye.gatech.edu (C.F. Jeff Wu).

uniform scale. Then a unique multiple testing correction is applied to the uniform order statistics to determine significance.

Microarrays can reveal a wealth of biological information to scientists, including differential expression of genes, phenotype identification and biomarker identification. However, estimation of the variability within genes and across arrays is severely limited by the typically small number of replicates. Transformations and normalization of the raw data are typically done to account for some of the variability from hybridization, scanning, arrays, etc. In general, there is a three-step process to convert raw expression data into a usable format. The raw data is initially corrected for background noise to eliminate systematic chip (or slide) biases. Log or square-root transformations are often done at this stage to bring in extremely large values. Then normalize the chips (or slides) to put them on the same scale. Often these first two steps are combined into one step, i.e., the normalization includes background correction. Finally, the data is summarized across different replicates into a test statistic. Robust methods are employed to derive a test statistic due to outliers and the inherent noisiness of gene expression data. Recent research has employed robust measurement in different stages of gene expression data analysis including: image analysis [12], gene filtering [8] and clustering methods [5,7].

The data set to illustrate MAOSA is derived from a large-scale analysis from [10]. They conducted a mixing experiment using oligonucleotide gene chips involving three groups of human fibroblast cells, with six replicates in each group. The three groups of cells are serum starved, serum stimulated and a 50/50 mixture of starved/stimulated. For the analysis in this paper, we compare the serum starved cells versus the mixture of starved/stimulated cells.

Other standard microarray analysis techniques are described here briefly. See the original papers for details. Dudoit et al. [4] proposed applying the Westfall–Young (WY) step-down technique [17] to replicated microarray data to adjust for the thousands of comparisons. The WY technique controls family-wise error rate (i.e., probability of at least one error in the family) in the strong sense (control for all possible combinations of true and false hypotheses). Tusher et al. [16] developed the Significance Analysis of Microarrays (SAM) procedure using a t -like statistic and permutation techniques. The authors use the False Discovery Rate (FDR) to calibrate the final number of significant effects.

2. Methods

2.1. Normalization for oligonucleotide gene chips

Slide effects, image effects, and hybridization effects, etc., are not of direct interest to scientists, but play a vital role in microarray analysis. The raw data from a spotted glass array or a gene chip should not be directly analyzed due to these biases. Instead, analysis ought to take place on transformed (as known as normalized) data that eliminates systematic effects from processing, hybridization, scanning, etc.

For gene chips, the Li and Wong [11] normalization is often used if the probe level data are available. In general, this normalization works well to eliminate probe level biases. However, if one examines a spatial plot of gene expression, then one would see that sometimes slide effects are not being accounted for. The Lemon et al. 50/50 mix of the starved/stimulated PM data is used in Fig. 1 to illustrate a problem that can occur if there is a bad slide. Each of the 225 squares represents an area of about 625 probes on each slide (25 probes by 25 probes). The colors are determined by the median expression of all probes within each square, relative to the empirical quantiles of the overall distribution of those median expression levels across all slides

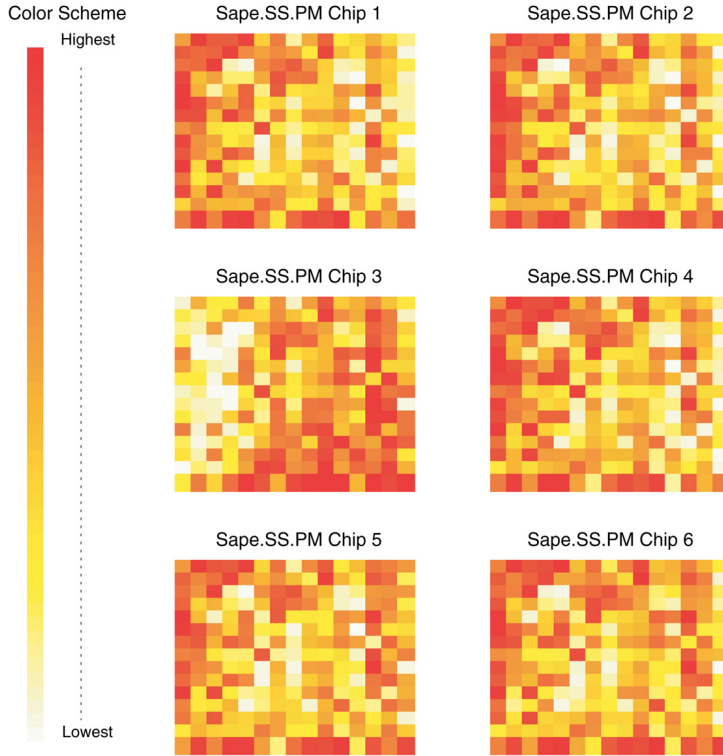


Fig. 1. Lemon et al. 50/50 mix of starved/stimulated PM data, normalized by the Li and Wong method.

in an experiment (replicates). For example, the 95th percentile of all 6 (number of replicates) \times 225 (15 by 15) squares is 700. A square with a median above 700 would receive a dark red color. The darker the color, the higher the intensity. Clearly the third slide (especially in the lower right-hand and upper left-hand corners) is different from the other five slides. This could be due to any number of reasons, including chip irregularities or under-hybridization. The Li–Wong normalization does not correct this difficulty; leading us to believe that their method should be adapted to adjust for these location effects. Other methods, including a quantile based approach [1] also do not correct for these chip location effects.

A simple method that does correct for these spatial differences is employed for the analysis here. Assume that the probe-level data has the following form

$$Y_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk} + \epsilon_{ijk}, \quad (1)$$

where Y_{ijk} represents the observed expression, μ_k the overall mean for gene chip k , α_{ik} the effect of row i in chip k , β_{jk} the effect of column j in chip k , and ϵ_{ijk} the residual. In this formulation, the PM and MM probe values are not separated. The parameters in (1) are estimated on a chip-by-chip basis using median polish. The resultant residual value ϵ_{ijk} is used as the normalized data matrix. The spatial plot from Fig. 1 is reproduced using the median polished data in Fig. 2. When the median polish normalization is used, the difficulty found in chip 3 is no longer present. With replicated chips, it is reasonable to use a spatial plot to verify a normalization technique. Across different experiments, there is no expectation that the expression is the same in all regions.

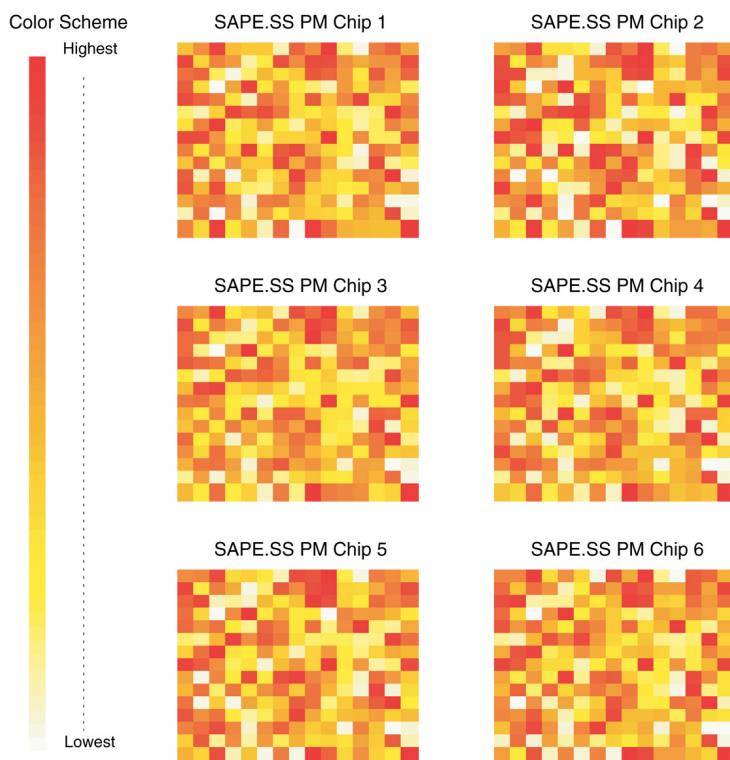


Fig. 2. Lemon et al. 50/50 mix of starved/stimulated, normalized with median polish.

2.2. Summary value for each gene within one slide

After normalization, the next step is to summarize the PM and MM probe values into one number for each gene. For each of the 7129 genes (or probe sets) on the gene chip, there are approximately 20 PM and MM probe pairs. Therefore for each gene, the expression within each chip is defined as

$$\tau_i = \text{Median}_j\{\text{PM}_{ij} - \text{MM}_{ij}\}, \quad (2)$$

for $i = 1, \dots, 7129$, $j = 1, \dots, N_i$, where N_i is the number of probes in probe set (or gene) i . There exists one number for each gene in each experiment.

Alternatively, an analysis on the PM-only data may be employed. This would entail using $\tau_i = \text{Median}_j\{\text{PM}_{ij}\}$ as the test statistic instead of (2). It is still debatable whether the MM probes any utility. Cope et al. [2] were unable to conclude that the MM values provide any utility. Irizarry et al. [9] suggested that “the MMs are a mixture of probes for which (i) the intensities are largely due to non-specific binding and background noise and (ii) the intensities include transcript signal just like the PMs”. Since the MM probes are in some instances measuring actual signal or binding, discarding them will cause a loss of information. There is no “best” way to utilize the MM probes, but a simple difference in medians should alleviate some concern.

2.3. Calculate a test statistic for each gene across replicates

For a replicated comparable gene chip experiment, construct the two-sample t -statistic as the test statistic for the l th gene ($l = 1, \dots, n$), using robust measures instead of the mean and

standard deviation:

$$\text{rts}_i = \frac{\text{Med}\{\tau_{i,1}\} - \text{Med}\{\tau_{i,2}\}}{\sqrt{\text{MAD}\{\tau_{i,1}\}^2/n_{i,1} + \text{MAD}\{\tau_{i,2}\}^2/n_{i,2} + c}}. \quad (3)$$

In (3), $\tau_{i,k}$ refers to the vector of summary values for gene i in condition k ($k = 1, 2$) from the replicated experiments, $n_{i,k}$ is the number of replicates for condition k ($k = 1, 2$) and c is a constant included to insure that genes that merely have low variability across samples are not mistakenly called significant. The normalization should have eliminated the systematic biases that exist in the experiment, so the median and median absolute deviation (MAD) are reasonable and consistent estimates of the location and scale for each gene effect. Noting that there are typically a small number of replicates in a microarray experiment, we choose these robust measures since outliers in the data affect them less than the mean and standard deviation. Other papers have suggested criteria for choosing such a c , including the minimization of the coefficient of variation [16] and using the 90th percentile of the rest of the denominator [6]. These correction methods lack a strong statistical justification. In the next section, a different method for determining c is described that is based on the normality assumption about the statistics in (3).

2.4. Determine c using the first four moments of the normal distribution

This analysis will not make the strong assumption that the rts_i (3) follow a normal distribution. Instead, we will assume that its *middle portion* follows a normal distribution. Therefore it is necessary to determine the expressions for the first four moments of the normal distribution when data from the upper and lower tails are truncated. Let $X \sim N(\mu, \sigma^2)$. Define $Y = \{X : a < X < b\}$. Define $\Phi(x, y, z)$ as the cumulative distribution function of a normal with mean y and standard deviation z evaluated at x and $\phi(x, y, z)$ as the density of a normal distribution with mean y and standard deviation z evaluated at x . It can be shown that

$$\begin{aligned} E[Y] &= \frac{\mu[\Phi(b, \mu, \sigma) - \Phi(a, \mu, \sigma)] + \sigma^2[\phi(a, \mu, \sigma) - \phi(b, \mu, \sigma)]}{P(a < X < b)}, \\ E[Y^2] &= \frac{(\mu^2 + \sigma^2)[\Phi(b, \mu, \sigma) - \Phi(a, \mu, \sigma)]}{P(a < X < b)} \\ &\quad + \frac{\mu\sigma^2[\phi(a, \mu, \sigma) - \phi(b, \mu, \sigma)] + \sigma^2[a\phi(a, \mu, \sigma) - b\phi(b, \mu, \sigma)]}{P(a < X < b)}, \\ E[Y^3] &= \frac{(\mu^3 + 3\mu\sigma^2)[\Phi(b, \mu, \sigma) - \Phi(a, \mu, \sigma)] + \sigma^2[a^2\phi(a, \mu, \sigma) - b^2\phi(b, \mu, \sigma)]}{P(a < X < b)} \\ &\quad + \frac{(\mu^2\sigma^2 + 2\sigma^4)[\phi(a, \mu, \sigma) - \phi(b, \mu, \sigma)] + \mu\sigma^2[a\phi(a, \mu, \sigma) - b\phi(b, \mu, \sigma)]}{P(a < X < b)}, \\ E[Y^4] &= \frac{(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4)[\Phi(b, \mu, \sigma) - \Phi(a, \mu, \sigma)]}{P(a < X < b)} \\ &\quad + \frac{(\mu^3\sigma^2 + 5\mu\sigma^4)[\phi(a, \mu, \sigma) - \phi(b, \mu, \sigma)]}{P(a < X < b)} \\ &\quad + \frac{(\mu^2\sigma^2 + 3\sigma^4)[a\phi(a, \mu, \sigma) - b\phi(b, \mu, \sigma)]}{P(a < X < b)} \\ &\quad + \frac{(\mu\sigma^2)[a^2\phi(a, \mu, \sigma) - b^2\phi(b, \mu, \sigma)] + \sigma^2[a^3\phi(a, \mu, \sigma) - b^3\phi(b, \mu, \sigma)]}{P(a < X < b)}. \end{aligned}$$

Schneider [14] discusses properties of the truncated normal distribution, including an expression to calculate all moments. Here a different approach is taken using the quantiles of the normal distribution to determine the cutoff points a and b . The resulting expressions are much cleaner. Let $X \sim N(\mu, \sigma^2)$, $a = \Phi^{-1}(\delta/2, \mu, \sigma)$, and $b = \Phi^{-1}(1 - \delta/2, \mu, \sigma)$, where $\Phi^{-1}(x, y, z)$ denotes the x th quantile of a normal distribution with mean y and standard deviation z . Then $Y = \{X : \Phi^{-1}(\delta/2, \mu, \sigma) < X < \Phi^{-1}(1 - \delta/2, \mu, \sigma)\}$. Therefore,

$$\begin{aligned}
 E[Y] &= \mu, \\
 E[Y^2] &= \mu^2 + \sigma^2 - \frac{2\sigma^2 \Phi^{-1}(1 - \delta/2) \phi(\Phi^{-1}(1 - \delta/2))}{1 - \delta}, \\
 E[Y^3] &= \mu^3 + 3\mu\sigma^2 - \frac{6\mu\sigma^2 \Phi^{-1}(1 - \delta/2) \phi(\Phi^{-1}(1 - \delta/2))}{1 - \delta}, \\
 E[Y^4] &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \\
 &\quad - \frac{(12\mu^2\sigma^2 + 6\sigma^4 + 2\sigma^4 \Phi^{-1}(1 - \delta/2)^2) \Phi^{-1}(1 - \delta/2) \phi(\Phi^{-1}(1 - \delta/2))}{1 - \delta}.
 \end{aligned} \tag{4}$$

In (4), $\Phi(x)$ and $\phi(x)$ denote the standard normal cumulative distribution and density functions, respectively, evaluated at x . Using these simplified expressions, we observe how the moments of a sample from a normal distribution vary as δ changes. The closeness of the middle $1 - \delta$ of the distribution to what is expected is determined, in terms of the first four moments. In reality, this moment-hunting will choose the δ for which the first four moments resemble the values we expect, calculated from (4). It is a four-dimensional optimization problem that can be solved by using an objective function to relate the four moments.

The application to gene expression data is as follows. Many genes have low variability across different experiments. Using a standard t -statistic will incorrectly yield many of these low variability genes as significant. Therefore, it is necessary to add a term to the denominator of the t -statistic to alleviate this concern. This constant c , added to the denominator of (3), is chosen to minimize the objective function in (5), where all summations are over the middle $1 - \delta$ of the distribution for the statistics in (3):

$$\begin{aligned}
 f(x, y, c, \delta) &= c + \log \left(\left[\left(\frac{1}{n} \sum_i \left(\frac{x_i}{y_i + c} \right)^4 \right)^{1/4} - E[Y^4]^{1/4} \right]^2 \right. \\
 &\quad + \left[\left(\frac{1}{n} \sum_i \left(\frac{x_i}{y_i + c} \right)^3 \right)^{1/3} - E[Y^3]^{1/3} \right]^2 \\
 &\quad + \left[\left(\frac{1}{n} \sum_i \left(\frac{x_i}{y_i + c} \right)^2 \right)^{1/2} - E[Y^2]^{1/2} \right]^2 \\
 &\quad \left. + \left[\left(\frac{1}{n} \sum_i \left(\frac{x_i}{y_i + c} \right) \right) - E[Y] \right]^2 \right).
 \end{aligned} \tag{5}$$

In (5), x denotes the numerator and y denotes the denominator (excluding c) of the middle $1 - \delta$ of the distribution of the statistics in (3). In addition, n is the number of effects in the middle $1 - \delta$ of this distribution. The values $E[Y]$, $E[Y^2]$, $E[Y^3]$ and $E[Y^4]$ are calculated from (4).

The first term in (5) is included so that smaller values of c are selected. It is not our goal to have c dominate the denominator of (3), but rather to ensure that genes with low variability across the experiments are not mistakenly declared to have significant differential expression. This method is aimed at detecting *large* significant effects. The interest is not to locate *small* transcript variability that may be biologically important. Use of the roots corresponding to the degree of moments (e.g., $\frac{1}{2}$ for the second moment) in (5) puts the expressions of the four moments on the same scale. In addition, log is taken to make it easier to distinguish between different values of c and limit the effect that the first term will have on the chosen c . Eq. (5) is simply an adaptation of the classic “observed minus expected squared minimization” technique found in goodness-of-fit tests, for example. This choice of c will ensure that the middle $1 - \delta$ of the test statistic is normally distributed. Other correction methods (including no correction at all) will just assume this conclusion without testing its validity. By forcing the majority of the gene expressions to follow a normal distribution, we have an estimated distribution on which the inference is made.

A numerical minimization procedure is used to solve for c . It is necessary to input the values of μ and σ^2 in order to compute $E[Y]$, $E[Y^2]$, $E[Y^3]$, $E[Y^4]$ in (4) for use in (5). To avoid even a slight effect of outliers, the median of the input vector (i.e., set $c = 0$ in (3)) is used as the estimate for μ , while the MAD is used as the estimate for σ . This criterion for selecting c ensures that $1 - \delta$ of the distribution of the statistics in (3) will mimic a normal distribution. Implicitly this assumes that $1 - \delta$ of this distribution is not differentially expressed.

For different values of δ , the optimal c is calculated. The final selection of δ will depend on the prior knowledge of the experimenter and the resultant minimum value. One can examine a plot of δ versus c versus the minimum value to see where the optimal regions are.

For the Lemon et al. starved versus starved/stimulated data, the minimum value attained depends mostly on δ . Fig. 3 displays two plots which related δ to c and the minimum value attained. A loess smoother was added to the plots to help visualize the distribution. The constant c does not seem to be an important component in the determination of the minimum value. Both smoothing curves obtain a minimum value at around $\delta = 0.10$. At that δ , $c = 1.92$, with a minimum value of -9.73 . This is the c used in (3) to compute the test statistics used in the later stages of the analysis. Note that this δ assumes that approximately 90% of the gene effects are null. Fig. 4 demonstrates the difference between the distribution of the test statistics (3) with no adjustment ($c = 0$) and the adjustment set forth in this section ($c = 1.92$) for the Lemon et al. data. From the right-hand picture, it appears that the adjustment has achieved its purpose of forcing the middle $1 - \delta = 90\%$ of the data to mimic a normal distribution. Using no adjustment will yield a more curvilinear distribution in the middle part (between -2 and 2 on the x -axis) of the distribution of (3).

2.5. Analysis using beta statistics

An analysis can be done using a well-known fact regarding order statistics. Suppose we have a sample X_i , $i = 1, \dots, n$ from a distribution function F . It is known from the probability integral transformation (PIT) that $F(X_i)$, $i = 1, \dots, n$, are uniformly distributed over $(0, 1)$. In essence, a random sample from any distribution function can be converted to a uniform $(0, 1)$. In addition $F(X_{(i)})$, $i = 1, \dots, n$, has a beta distribution with parameters i and $n - i + 1$ [3]. Since there are nice properties when using uniform order statistics, results below will only be concerned with the standard uniform density, although the results are applicable to any distribution because of the PIT.

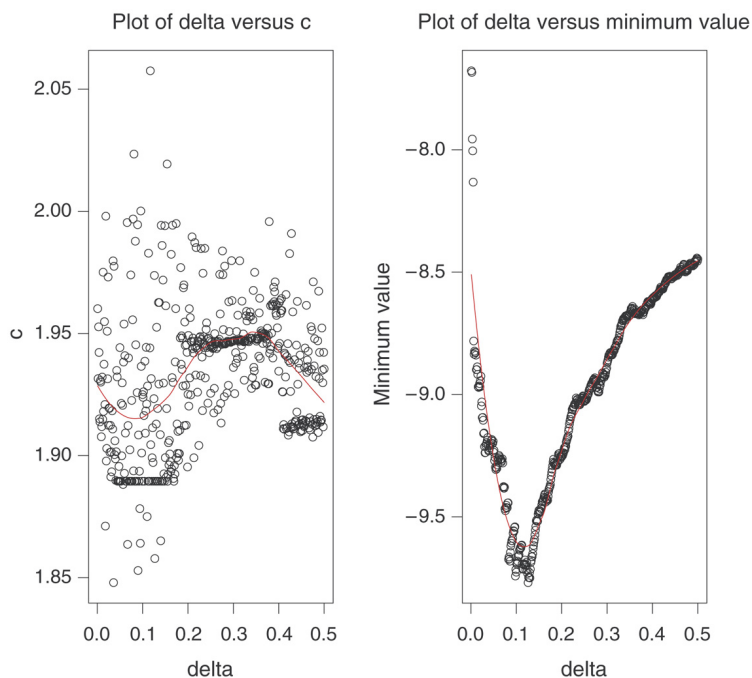


Fig. 3. Optimal c determination based on (4) for Lemon et al. data.

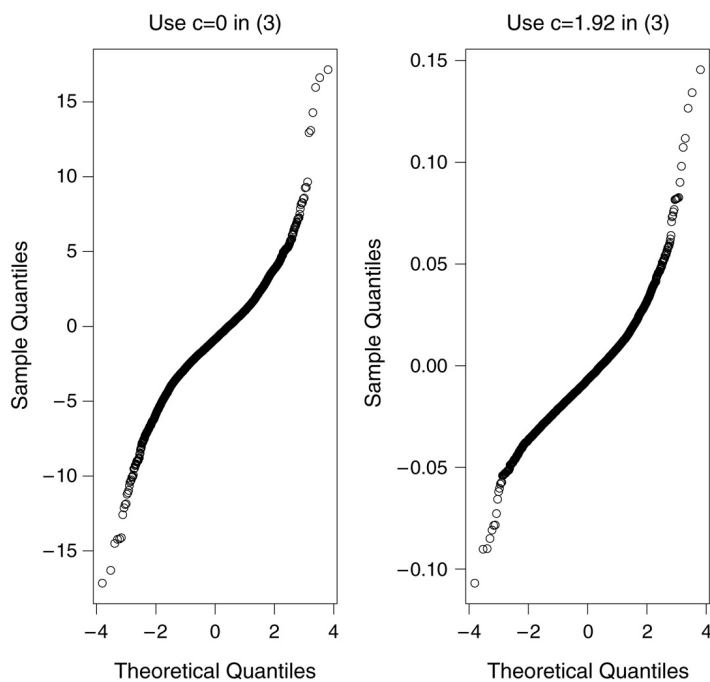


Fig. 4. Comparison of the distribution of the test statistics in (3) using $c = 0$ or $c = 1.92$.

The joint distribution of the order statistics of a random variable can be determined. In particular let X_1, \dots, X_n be independent and identically distributed random variables from an absolutely continuous distribution function F . Denote the order statistics as $X_{(1)}, \dots, X_{(n)}$. For $X \sim F$, David [3] showed that $F(X_{(s)}) - F(X_{(r)}) \sim \text{Beta}(s - r, n - (s - r) + 1)$. Therefore the distribution of the difference in uniform order statistics depends only on the length of the difference, and not on the particular order statistics being subtracted. A crucial element is the correct distribution function F . Hence, any analysis should include a method, like in [Section 2.3](#) to ensure that the data mimics a known distribution.

Based on these distributional results, a test can be developed to determine change points in order statistics. In a sense, the following procedure will identify the point at which the ordered data shifts away from the uniform distribution. In the case of gene expression data, this procedure will allow the identification of gene effects that are significantly different from the large amount of unexpressed genes.

2.6. MAOSA outline

A new procedure for detecting differential expression, called *Multiplicity-Adjusted Order Statistics Analysis* (MAOSA), is proposed. It consists of the following steps:

- (i) Determine the optimal c and calculate the summary statistics in (3) as discussed in [Section 2.4](#).
- (ii) Convert the test statistics from (i) to the uniform scale, assuming that they have a normal distribution. Use the median and MAD of the test statistics as the mean and standard deviation in the probability integral transformation.
- (iii) Compute test statistics using the uniform statistics derived in (ii) and set $r = 1$, i.e., compare each subsequent order statistic to the minimum order statistic. This analysis is two-sided in that it looks for significance in both the upper and lower tails.
- (iv) Compare the test statistics from (iii) to the theoretical beta quantiles. When making this comparison, use the Bonferroni correction on a small number effects since the majority of gene effects can be ignored.
- (v) Determine the segments of genes that have significant differential expression.

Note that an alternative normalization technique or test statistic may be employed in the early stages of this algorithm. However, after step (i) it is necessary that the gene summary statistics have a known distribution (in this case, normal) to proceed with the algorithm. There are a host of other methods to ascertain statistical significance of normally distributed statistics, from a simple p -value from a t -distribution to non-parametric methods. The proposed method is compared to the SAM and WY methods introduced in [Section 1](#).

The Bonferroni adjustment employed in step (iv) works in the following way. First, eliminate effects that are clearly not significant. The choice of c allows us to ignore the vast majority of effects. Since we assume that the test statistic has a normal null distribution, effects that fall within ± 2 s.d. should not be declared significant. Hence, the Bonferroni adjustment is employed only when we move past those null effects as we have helped define by the selection of c . Test statistics from (iii) that do not pass a non-corrected p -value (say, 0.10) cutoff are excluded. This cutoff should be no less than the δ chosen to lessen the possibility of a type II error. There are effects left in the lower and upper tails. For the effects in the upper tail, start at the smallest indexed order statistic remaining and determine the first occurrence of r consecutive significant genes. A similar procedure is done for the remaining genes in the lower tail. In this

Table 1
MAOSA analysis of Lemon et al. data

α	# Significant	FDR (%)
0.01	975	2.69
0.0001	824	0.46
0.000001	681	0.13
0.00000001	613	0.00

phase, the r comparisons are adjusted via the Bonferroni correction, i.e., use level α/r . The selection of this small number, r , will depend on the data and the expected number of significant effects. For example, there are 7129 genes in the Lemon et al. data. There were 1199 genes that passed the initial screening, with 61 in the lower tail and 1138 in the upper tail. Next, start at the smallest index of the 1138 order statistics in the upper tail, in this case, order statistic number 5992. Starting from this order statistic, find the first r consecutive ordered test statistics that pass the Bonferroni corrected significance level. All subsequent test statistics (after the initial r significant) are also declared significant. Once some ordered effects are significant, the subsequent (i.e., larger) effects are also significant. The same logic is used in half-normal plots for detecting significance in factorial experiments. (For details on half-normal plots, see [18]). Table 1 displays the results for the Lemon et al. data, using $r = 10$. The first column (α) indicates the significance level employed, the second column (# Significant) the number of significant effects, and the third column the FDR. This example is further discussed in Section 3.

Other test statistics were considered in step (iii) of the above algorithm, including comparing all order statistics to the maximum order statistic. In addition, Stephens [15] suggested using

$$Z = \frac{(n - i + 1)U_{(i)}}{i(1 - U_{(i)})}, \quad (6)$$

which has an $F_{2i, 2(n-i+1)}$ distribution. However, we chose to use the proposed technique due to its simplicity and nice distributional properties.

3. Results

The analysis presented in this section compares the cells grown in the starved environment to those grown in the mixed starved/stimulated environment from the Lemon et al. data cited in Section 1. We follow the procedure outlined in Section 2 for the data pre-processing. For this example, using $\delta = 0.10$ in (4) yields $c = 1.92$ in (3). Using this δ value tacitly assumes that about 10% of the data is differentially expressed. Some methods for determining differential expression as well as methods for assessing the validity of such an analysis require permuted data sets. Therefore set the c for test statistics computed from the permuted data sets to be 1.92. This reduces the bias by keeping the permuted data on the same level as the original.

Continuing with the outline from Section 2.6, the test statistics are converted to the uniform scale using PIT. These uniform statistics are used to produce beta test statistics. Then using a two-sided hypothesis test, calculate p -values for each test statistic versus its respective theoretical beta distribution. Gene indices near 0 correspond to genes with the most negative value from (3), while gene indices near 7100 correspond to genes with the most positive value from (3). Genes with the smallest p -values lie on the extremes.

The next step is to determine which genes are significant. The obvious method to use involves theoretical quantities from the beta distribution. Since there will be multiple testing issues, a

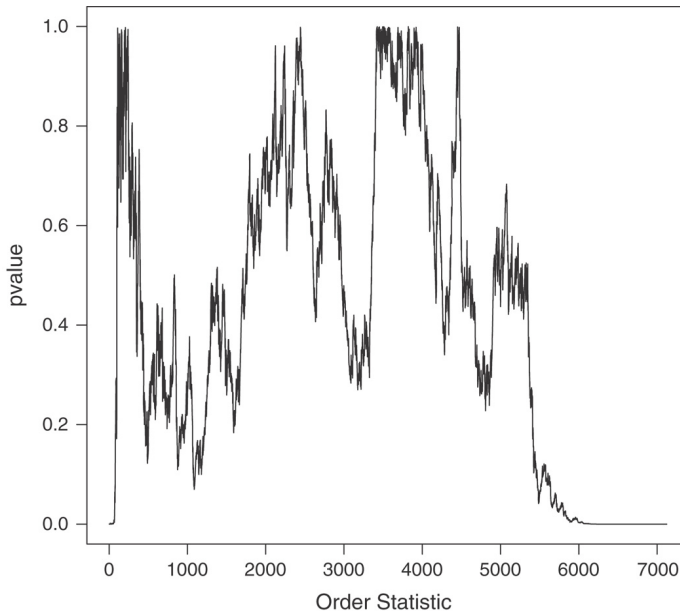


Fig. 5. p -values for Lemon et al. data, using the MAOSA procedure.

correction is necessary. We use the *Bonferroni adjustment* since it requires no independence assumption. However, enforcing the Bonferroni correction over 7129 tests would be impractical. Therefore, to deal with the multiplicity problem in a prudent manner, we only enforce the correction on a small subset of tests, in this example, 10. See [Section 2.6](#) for a detailed explanation on how the Bonferroni correction was employed in this example. Ten was chosen because we assume that only a small number of effects would be significant based on a WY analysis which found 7 significant effects when controlling type I error at 0.05. [Table 1](#) shows the analysis results using MAOSA as described in [Section 2.6](#).

The number of significant genes for this data are 975, 824, 681 and 613 for various significance levels α given in [Table 1](#). An estimate of the error rate for each of these “significance sets” can be obtained using the FDR. Permute the condition labels while maintaining balance between conditions. Compute the robust statistics (3), keeping the same $c = 1.92$ for each permuted data set. Then using the same α values from [Table 1](#) and MAOSA, compute the number of significant effects for each permuted data set. The average number of significant genes over all 400 permuted data sets for an α value divided by the number of significant genes found by the actual data set (for the same α value) is the FDR.

[Fig. 5](#) displays the p -values for the 7128 genes other than the one corresponding to the minimum order statistic. Most of the significant effects are the larger order statistics. For this data, it means that the significant genes are more expressed in the mixed condition rather than the starved.

As described above, a permutation method was used to help determine the error rates of the sets of significant effects. The FDRs for each of the sets of significant effects are listed in [Table 1](#). With such low FDRs, 975 would be our choice for the number of significant effects. An analysis using SAM was also done on the same data. [Table 2](#) summarizes these results. Since MAOSA has a lower FDR for all subsets of significant effects, we believe that it is at least comparable to SAM.

Table 2
SAM analysis of Lemon et al. data

Δ	# Significant	FDR (%)
1.476	979	9.03
1.544	824	7.90
1.630	681	6.69
1.696	612	5.97
3.000	55	1.20

It is necessary to explain why the error control rates for both the MAOSA and SAM are so low when calling hundreds of genes significant. Pan et al. [13] postulated that the SAM method (and similarly the MAOSA method) assumes that under the null hypothesis no genes are differentially expressed. For testing a small subset of genes, this is reasonable. However, with hundreds or thousands of tests, it may be necessary to adjust the reference distribution by adding significant effects. Further development is needed here.

4. Discussion

A new method for identifying differentially expressed genes for oligonucleotide gene chips was proposed in this paper. The MAOSA technique first transforms the test statistics by assuming that the middle part of the distribution of the test statistics follows a known distribution. Then these test statistics are converted to the uniform scale, using the probability integral transformation. Analysis then proceeds with the uniform order statistics. To prudently correct for multiple testing, a Bonferroni adjustment is done on a small number of effects. This method can also incorporate a priori biological knowledge in the analysis with the selection of c . This technique can also be applied to other types of large data including spotted glass microarrays.

Acknowledgements

This research was supported in part by NSF grant DMS 0305996. The authors would like to thank the Associate Editor and the two anonymous reviewers for their suggestions which improved this paper.

References

- [1] B.M. Bolstad, R.A. Irizzary, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* 19 (2003) 185–193.
- [2] L.M. Cope, R.A. Irizarry, H.A. Jaffee, Z.J. Wu, T.P. Speed, A benchmark for affymetrix GeneChip expression measures, *Bioinformatics* 20 (2004) 323–331.
- [3] H.A. David, *Order Statistics*, John Wiley and Sons, New York, 1981.
- [4] S. Dudoit, Y.H. Yang, M.J. Callow, T. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* 12 (2002) 111–140.
- [5] G. Dyson, C.F.J. Wu, ICI: A new approach to explore between-cluster relationships with applications to gene expression data, Georgia Tech School of Industrial and Systems Engineering — Statistics Group Technical Report, 11/2005, 2005.
- [6] B. Efron, R. Tibshirani, J.D. Storey, Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* 96 (2001) 1151–1160.
- [7] M.A. Hibbs, N.C. Dirksen, K. Li, O.G. Troyanskaya, Visualization methods for statistical analysis of microarray clusters, *BMC Bioinformatics* 6 (115) (2005).

- [8] S. Imoto, T. Higuchi, S.Y. Kim, E. Jeong, S. Miyano, Residual bootstrapping and median filtering for robust estimation of gene networks from microarray data, *Computational Methods in Systems Biology Lecture Notes in Computer Science* 3082 (2005) 149–160.
- [9] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics* 4 (2003) 249–264.
- [10] W.J. Lemon, J.J.T. Palatini, R. Krahe, F.A. Wright, Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays, *Bioinformatics* 18 (2002) 1470–1476.
- [11] C. Li, W.H. Wong, Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proceedings of National Academy of Science USA* 98 (2001) 31–36.
- [12] Q.H. Li, C. Fraley, R.E. Bumgarner, K.Y. Yeung, A.E. Raftery, Donuts, scratches and blanks: robust model-based segmentation of microarray images, *Bioinformatics* 21 (2005) 2875–2882.
- [13] W. Pan, J. Lin, C. Le, A mixture model approach to detecting differentially expressed genes with microarray data, *Functional and Integrative Genomics* 3 (2001) 117–124.
- [14] H. Schneider, *Truncated and Censored Samples from Normal Populations*, Marcel Dekker, Inc., New York, 1986.
- [15] M.A. Stephens, Tests for the uniform distribution, in: R.B. D’Agostino, M.A. Stephens (Eds.), *Goodness-of-fit Techniques*, Marcel Dekker, Inc., New York, 1986.
- [16] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of National Academy of Science USA* 98 (2001) 5116–5121.
- [17] P.H. Westfall, S.S. Young, *Resampling-based Multiple Testing: Examples and Methods for p -value Adjustment*, John Wiley and Sons, New York, 1993.
- [18] C.F.J. Wu, M. Hamada, *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley and Sons, New York, 2000.