DOI: 10.1111/j.1541-0420.2009.01292.x

Controlling False Discoveries in Multidimensional Directional Decisions, with Applications to Gene Expression Data on Ordered Categories

Wenge Guo,[1] Sanat K. Sarkar,[2] and Shyamal D. Peddada[3]

[1]Biostatistics Branch, National Institute of Environmental Health Sciences,
Research Triangle Park, North Carolina 27709, U.S.A.
email: wenge.guo@gmail.com
[2]Department of Statistics, Temple University, Philadelphia, Pennsylvania 19122, U.S.A.
email: sanat@temple.edu
[3]Biostatistics Branch, National Institute of Environmental Health Sciences,
Research Triangle Park, North Carolina 27709, U.S.A.
email: peddada@niehs.nih.gov

Summary. Microarray gene expression studies over ordered categories are routinely conducted to gain insights into biological functions of genes and the underlying biological processes. Some common experiments are time-course/dose-response experiments where a tissue or cell line is exposed to different doses and/or durations of time to a chemical. A goal of such studies is to identify gene expression patterns/profiles over the ordered categories. This problem can be formulated as a multiple testing problem where for each gene the null hypothesis of no difference between the successive mean gene expressions is tested and further directional decisions are made if it is rejected. Much of the existing multiple testing procedures are devised for controlling the usual false discovery rate (FDR) rather than the mixed directional FDR (mdFDR), the expected proportion of Type I and directional errors among all rejections. Benjamini and Yekutieli (2005, *Journal of the American Statistical Association* **100,** 71–93) proved that an augmentation of the usual Benjamini–Hochberg (BH) procedure can control the mdFDR while testing simple null hypotheses against two-sided alternatives in terms of one-dimensional parameters. In this article, we consider the problem of controlling the mdFDR involving multidimensional parameters. To deal with this problem, we develop a procedure extending that of Benjamini and Yekutieli based on the Bonferroni test for each gene. A proof is given for its mdFDR control when the underlying test statistics are independent across the genes. The results of a simulation study evaluating its performance under independence as well as under dependence of the underlying test statistics across the genes relative to other relevant procedures are reported. Finally, the proposed methodology is applied to a time-course microarray data obtained by Lobenhofer et al. (2002, *Molecular Endocrinology* **16,** 1215–1229). We identified several important cell-cycle genes, such as DNA replication/repair gene MCM4 and replication factor subunit C2, which were not identified by the previous analyses of the same data by Lobenhofer et al. (2002) and Peddada et al. (2003, *Bioinformatics* **19,** 834–841). Although some of our findings overlap with previous findings, we identify several other genes that complement the results of Lobenhofer et al. (2002).

Key words: Benjamini–Hochberg procedure; Directional FDR; Dose-response; Microarray; Multiple testing; Ordered categories; Time course.

1. Introduction

In many applications, researchers are interested in identifying trends in mean response over ordered categories in large-scale experiments. With the advent of microarray technology such experiments are common in the literature where investigators are routinely conducting experiments to investigate changes in mean gene expressions over time or dose of a chemical or cancer stage, etc. For example, Lobenhofer et al. (2002) studied the effect of $17-\beta$ estradiol on the gene expression of MCF-7 breast cancer cells as the cells progressed through various phases of cell division cycle. In another experiment, Tamoto et al. (2004) investigated the changes in gene expression with tumor progression in esophageal cancer and identified genes implicated in the early stages of esophageal squamous cell carcinoma. Recently, Bochkina and Richardson (2007) discussed the analysis of a time-course gene expression data where cells from the H2Kb muscle cell line of mouse were treated by insulin (0, 2, or 12 hours of exposure).

In studies such as those described above, identification of statistically significant genes that have similar mean expression profiles over ordered categories is often an important goal to researchers. By identifying such genes, the researchers may potentially discover co-regulated genes belonging to similar pathways and gain insights into biological functions and processes of groups of genes with similar patterns of expressions.

Peddada et al. (2003) introduced an order-restricted inference-based method for identifying significant genes and grouping them according to various patterns of inequalities. Implicitly in their methodology, two decisions are being made for each gene. First, it is decided whether or not a gene is significant using a method exercising a control over gene-specific Type I error rate. Then, a suitable inequality pattern is assigned for each selected significant gene based on the values of the underlying test statistics. The directional error that can potentially occur in addition to the usual Type I error due to assigning wrong inequality pattern to a selected significant gene has not been addressed in that paper. In this article, we take care of both the Type I and directional errors. We do that by taking a multiple testing approach to the main problem where for each gene the mean expressions are successively compared across the ordered categories and a null hypothesis signifying no particular directional pattern is formed to test against the union of all possible directional patterns. We develop a method of simultaneously testing these null hypotheses and determining a directional pattern upon rejection of each of them controlling both the Type I and directional errors in an overall sense.

Although directional error has not been discussed extensively in the literature, it is perhaps a common error that occurs in applications. While testing a null hypothesis $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$, for some single parameter θ of interest, researchers commonly conclude either $\theta > 0$ or $\theta < 0$ upon rejection of H_0 depending on the sign of the underlying test statistic, keeping the directional error controlled in addition to the Type I error. However, when multiple hypotheses are tested and the number of parameters describing directional patterns is (even moderately) larger than one, as is the case with time-course or dose-response microarray data, controlling the directional errors is a problem.

A traditional approach to dealing with directional as well as Type I errors from a multiple testing point of view is to apply a method that controls the so-called mixed directional familywise error rate (mdFWER), which is the probability of one or more Type I or directional errors, a variant of the classical familywise error rate (FWER; Shaffer, 1980; Finner, 1994, 1999; Liu, 1996; Sarkar, Sen, and Finner, 2004). However, when the number of null hypotheses is large, as in the context of microarray experiments, the notion of mdFWER, just like the FWER, is too stringent, allowing little chance to make true directional as well as nondirectional discoveries. The False Discovery Rate (FDR), due to Benjamini and Hochberg (1995), is a more powerful concept of overall Type I error rate than the FWER in the context of multiple testing and is now most commonly used in large-scale scientific investigations, especially in microarray gene expression studies. A variant of it while controlling both Type I and directional errors would be more powerful than the mdFWER. Two such variants have been introduced in the literature (Benjamini, Hochberg, and Kling, 1993), the pure directional FDR that is the expected proportion of directional errors among rejected hypotheses and the mixed directional FDR (mdFDR) that is the expected proportion of Type I and directional errors among rejected hypotheses. In this article, we focus on procedures controlling the mdFDR.

Benjamini and Yekutieli (2005) gave a method with independent tests that controls the mdFDR when testing multiple simple hypotheses against two-sided alternatives. They proved that the original Benjamini and Hochberg (1995) procedure controlling the FDR at α can be augmented to make directional decision upon rejecting a null hypothesis according to the value of the corresponding test statistic without causing the mdFDR to exceed α, a result conjectured by several authors (Benjamini and Hochberg, 2000; Shaffer, 2002; Williams, Jones, and Tukey, 1999). Throughout the article we denote Benjamini and Hochberg procedure as BH procedure. Clearly, this method, referred to as the directional BH procedure, can be applied to analyze dose-response microarray data if there are only two ordered categories, but often this is not the case, as such data typically involve more than two ordered categories and the method needs to be suitably extended to accommodate such multiple categories.

We extend the BH directional FDR procedure in this article to develop our proposed multiple testing method that allows us to make a decision on the directional pattern involving multiple parameters once a null hypothesis of no pattern is rejected and maintains a control over the mdFDR. The proposed methodology is then evaluated through a simulation study and applied to the time-course microarray data in Lobenhofer et al. (2002). Our analysis of Lobenhofer's data resulted in the discovery of several cell-cycle genes that were not previously identified by Lobenhofer et al. (2002) and Peddada et al. (2003). Some of our findings complement the previous findings as detailed in Section 5. An important and unique feature of our methodology is that it permits us to specify the time interval of up (or down) regulation of a gene during the 48-hour period of the cell cycle. One of the usual objectives for conducting cell-cycle time-course experiments is to determine the phase of peak expression for a cell-cycle gene and our methodology allows us to make such determinations.

2. Notations, Definitions, and Problem Formulation

In this section, we present the multiple testing formulation of the problem of identifying expression patterns/trends over ordered categories simultaneously for all the genes, having introduced some notations and definitions related to multiple testing.

Let μ_{ij} denote the mean response of the jth variable (e.g., gene), $j = 1, \ldots, m$, in the ith ordered category, $i = 1, \ldots, p$. A problem of biological interest is to group genes by the inequalities among the mean responses, known as directional patterns or order restrictions. Some common inequalities of interest are $\mu_{1j} \leqslant \mu_{2j} \leqslant \cdots \leqslant \mu_{pj}$ (monotone pattern), $\mu_{1j} \leqslant \mu_{2j} \leqslant \cdots \leqslant \mu_{ij} \geqslant \mu_{(i+1)j} \geqslant \cdots \geqslant \mu_{pj}$, $i = 2, \ldots, p-1$ (umbrella order with peak μ_{ij}). Let $\delta_{ij} = \mu_{i+1j} - \mu_{ij}$, $i = 1, \ldots, p-1$, $j = 1, \ldots, m$. Then, the above inequalities of interest or any other inequalities can be restated in terms of the signs of the δ_{ij}'s. Let $\boldsymbol{\delta}_j = (\delta_{1j}, \ldots, \delta_{qj})'$, where $q = p - 1$. Suppose we test

$$H_{0j} : \boldsymbol{\delta}_j = \mathbf{0} \text{ against } H_{1j} : \boldsymbol{\delta}_j \neq \mathbf{0}, \qquad (1)$$

and suppose H_{0j} is rejected, then we first decide which δ_{ij}'s are nonzero before deciding their signs. The signs of the nonzero

δ_{ij}'s are determined by the sign of the corresponding test statistic. The declared signs of the δ_{ij}'s then determine a possible inequality or directional pattern. For instance, in the case of $q = 4$, suppose for a given gene j, $\boldsymbol{\delta}_j = (\delta_{1j}, \ldots, \delta_{4j})$ is found to be significantly different from a null vector, with δ_{1j} and δ_{2j} declared to be positive and negative, respectively, and δ_{3j} and δ_{4j} are zero. Then, the corresponding directional pattern is $\mu_{1j} < \mu_{2j} < \mu_{3j} = \mu_{4j} = \mu_{5j}$. We can test H_{0j} against H_{1j} for all the genes applying a suitable multiple testing method. Thus, given p ordered categories for each gene, the task of identifying directional patterns of the mean expressions over these categories for all the genes is being formulated as a multiple testing problem where H_{0j} is tested against H_{1j} simultaneously for all the genes and the signs of δ_{ij}'s are determined subsequent to the rejection of the corresponding H_{0j}.

For multiple testing of H_{0j} against H_{1j}, $j = 1, \ldots, m$, we need p-values that will provide a valid test for each of these individual testing problems and will allow us to make decisions on the individual δ_{ij}'s once a H_{0j} is rejected. For that, we consider for each j the p-value available for testing each component null hypothesis $H_{0j}^i : \delta_{ij} = 0$ against the corresponding component alternative hypothesis $H_{1j}^i : \delta_{ij} \neq 0$, for $i = 1, \ldots, q$, and apply a suitable combination method pooling these $q\,p$-values by treating H_{0j} as an intersection of the subfamily of these q component null hypotheses. Thus, $H_{0j} = \bigcap_{i=1}^q H_{0j}^i$, and $H_{1j} = \bigcup_{i=1}^q H_{1j}^i$. Before we discuss appropriate combination methods to be used, we explain how to obtain these component p-values and state the underlying assumptions.

For every $i = 1, \ldots, q$ and $j = 1, \ldots, m$, suppose we use the absolute value of a test statistic T_{ij} for testing H_{0j}^i against H_{1j}^i. Let $T_{ij} \sim F_{ij}(t, \delta_{ij})$ for some continuous cdf F, which is symmetric about 0 under H_{0j}^i and gets stochastically larger or smaller as δ_{ij} either increases or decreases from 0. In other words, $F_{ij}(t, \delta_{ij}) \leqslant$ or $\geqslant F_{ij}(t, 0)$ according to $\delta_{ij} >$ or < 0, with $F_{ij}(0, 0) = \frac{1}{2}$. Under this setting, a right-tailed test based on the absolute value of T_{ij} will be considered for testing H_{0j}^i against H_{1j}^i, with the corresponding two-sided p-value being defined as $\widetilde{P}_{ij} = 2\min\{F_{ij}(T_{ij}, 0), 1 - F_{ij}(T_{ij}, 0)\}$. By the assumed distributional property of T_{ij}, it is easy to verify that under H_{0j}^i, the two-sided p-value $\widetilde{P}_{ij}$ satisfies

$$Pr\{\widetilde{P}_{ij} \leqslant p\} \leqslant p, \text{ for any } p \in (0, 1). \tag{2}$$

Given p-values for testing H_{0j}^i against H_{1j}^i, for $i = 1, \ldots, q$, a number of combination methods (or methods of pooling the p-values) are available in the literature for testing the intersection null hypothesis $H_{0j} = \bigcap_{i=1}^q H_{0j}^i$ against the alternative $H_{1j} = \bigcup_{i=1}^q H_{1j}^i$. Among these, however, the Bonferroni and Simes methods are often used in multiple testing and allow one to make decisions on the individual δ_{ij}'s. For a review of these methods, one may see Bernhard, Klein, and Hommel (2004). Let $\widetilde{P}_{(1)j} \leqslant \cdots \leqslant \widetilde{P}_{(q)j}$ be the ordered versions of $\widetilde{P}_{ij}, i = 1, \ldots, q$, for a fixed $j = 1, \ldots, m$. Then, in the Bonferroni test, the pooled (or adjusted) p-value is given by $P_j = q\widetilde{P}_{(1)j}$; whereas, in the Simes test, it is given by $P_j = \min_{1 \leqslant i \leqslant q}\{q\widetilde{P}_{(i)j}/i\}$. While the Bonferroni test does not require any dependence structure in the underlying p-values, the Simes test requires a certain type of positive dependence condition that is often satisfied in multiple testing applications (Sarkar and Chang, 1997). Upon rejection of H_{0j} using the Bonferroni pooled p-value at a level α, the ith component of null hypothesis H_{0j}^i can be rejected if $\widetilde{P}_{ij} \leqslant \alpha/q$. For the test based on the Simes pooled p-value, H_{0j}^i corresponding to every $\widetilde{P}_{ij} \leqslant \widetilde{P}_{(R_j)j}$ is rejected, where $R_j = \max\{i : \widetilde{P}_{(i)j} \leqslant \frac{i}{q}\alpha\}$, if the maximum exists; otherwise, $R_j = 0$.

Now, suppose the pooled p-value P_j, based on either Bonferroni or Simes test, is available to us for every $j = 1, \ldots, m$, to carry out a multiple testing procedure to test H_{0j} against H_{1j} simultaneously for all $j = 1, \ldots, m$. We use the multiple testing method of Benjamini and Hochberg (1995) (the BH method) that is designed to control the false discovery rate (FDR). The FDR, for any given multiple testing procedure, is the expected proportion of false rejections (Type I errors) among all rejections, an overall measure of Type I error rate commonly used in microarray studies. More formally, with V the number of falsely rejected true null hypotheses among $H_1, \ldots, H_m$ and R the total number of rejected hypotheses among $H_1, \ldots, H_m$, it is defined as

$$FDR = E\left\{\frac{V}{R \vee 1}\right\}, \tag{3}$$

where $R \vee 1 = \max(R, 1)$. This method with a control of the FDR at a given level α is a step-up test as follows: Given ordered p-values $P_{(1)} \leqslant \cdots \leqslant P_{(m)}$ with the corresponding null hypotheses $H_{(1)}, \ldots, H_{(m)}$, find $k = \max\{1 \leqslant j \leqslant m : P_{(j)} \leqslant j\alpha/m\}$ and reject those $H_{(j)}$ for which $P_{(j)} \leqslant P_{(k)}$, provided this maximum exists, otherwise, accept all the null hypotheses.

When a $H_{0j} : \boldsymbol{\delta}_j = \mathbf{0}$ is rejected using the BH method and further decisions are being made on the signs of the component δ_{ij}'s in the corresponding $\boldsymbol{\delta}_j$, a directional error might occur due to wrong assignments of the signs. For instance, if there is a component δ_{ij} in $\boldsymbol{\delta}_j = (\delta_{1j}, \ldots, \delta_{qj})$ that is truly positive (or negative) but declared to be negative (or positive) while deciding on the signs of the δ_{ij}'s upon rejection of $H_{0j} : \boldsymbol{\delta}_j = 0$, a directional error occurs. So, we need to control such directional errors as well. A convenient and practical way of doing that would be to use an error rate combining both Type I and directional errors in the FDR framework and make sure that it is controlled. One such error rate is the mdFDR, the sum of the FDR and the pure directional FDR (dFDR). The dFDR is defined as

$$dFDR = E\left\{\frac{S}{R \vee 1}\right\}, \tag{4}$$

where S denotes the total number of false null hypotheses among $H_1, \ldots, H_m$ that are correctly rejected but at least one directional error has been made while deciding upon the signs of the components. In other words, S is the number of rejected hypotheses H_j's such that $\boldsymbol{\delta}_j \neq \mathbf{0}$ and for some $i = 1, \ldots, q$, δ_{ij} is declared to be positive when $\delta_{ij} \leqslant 0$, or δ_{ij} is declared to be negative when $\delta_{ij} \geqslant 0$. Thus, more formally, the mdFDR is defined as

$$mdFDR = FDR + dFDR = E\left\{\frac{V + S}{R \vee 1}\right\}, \tag{5}$$

the expected proportion of Type I and directional errors among all rejections.

It is important to point out that the goal of this article is to identify expression patterns of m genes over p ordered categories. For each gene it is biologically relevant to consider its expression pattern as a whole across p ordered categories rather than viewing this to be a problem of testing qm separate hypotheses which ignores the intrinsic pattern over ordered categories. Thus, rather than viewing it as a problem of performing qm tests, we treat it as a problem of performing a set of m tests each involving q-dimensional hypotheses. In addition, we want to emphasize that while making directional decisions for the components of $\boldsymbol{\delta}_j$, no directional errors are being made when $\boldsymbol{\delta}_j = \mathbf{0}$. In contrast, when making directional decisions regarding a nonnull $\boldsymbol{\delta}_j$, a directional error is made if a component δ_{ij} for which $\delta_{ij} \neq 0$ is declared to be positive or negative.

In the next section, we develop methods to control the mdFDR. This extends the following directional BH procedure of Benjamini and Yekutieli (2005) from dimension one (i.e., $q = 1$) to a general dimension.

Definition 1 (The level-α directional BH procedure)

(1) *Apply the BH method at level α to test $H_{0j} : \delta_{1j} = 0$ against $H_{1j} : \delta_{1j} \neq 0$ simultaneously for $j = 1, \ldots, m$, based on the two-sided p-values $\widetilde{P}_{1j}, j = 1, \ldots, m$.*
(2) *Let R denote the total number of null hypotheses rejected.*
(3) *For every $j = 1, \ldots, m$, with $\widetilde{P}_{1j} \leqslant \frac{R}{m}\alpha$, declare $\delta_{1j} >$ or < 0 according to $T_{1j} > 0$* or < 0.

It controls the mdFDR at α under independence of the underlying test statistics.

3. Multidimensional Directional FDR Controlling Procedures

We introduce in this section our proposed method of controlling the mdFDR while testing $H_{0j} : \boldsymbol{\delta}_j = \mathbf{0}$ against $H_{1j} : \boldsymbol{\delta}_j \neq \mathbf{0}$, simultaneously for all $j = 1, \ldots, m$, and making further decisions on the signs of the δ_{ij}'s upon rejection of the corresponding H_{0j}. It is based on the Bonferroni pooled p-values.

Procedure 1

(1) *Apply the BH method at level α to test H_{0j} against H_{1j} simultaneously for $j = 1, \ldots, m$, based on the Bonferroni pooled p-values P_j, $j = 1, \ldots, m$.*
(2) *Let R denote the total number of null hypotheses rejected.*
(3) *For every $i = 1, \ldots, q$ and $j = 1, \ldots, m$ with $\widetilde{P}_{ij} \leqslant \frac{R}{qm}\alpha$, if $T_{ij} > 0$, declare $\delta_{ij} >$ or < 0 according to $T_{ij} > 0$ or < 0.*

Theorem 1: *With independent q-dimensional test statistics $\mathbf{T}_j = (T_{1j}, \ldots, T_{qj})$, $j = 1, \ldots, m$, the mdFDR of Procedure 1 is less than or equal to α.*

Remark 1. Proof of Theorem 1 is provided in the Web Appendix. Benjamini and Yekutieli (2005) gave an indirect proof of this theorem in the special case when $q = 1$ using an approach that relates to the FDR-adjusted confidence intervals for selected parameters they developed in the same paper. However, it is not apparent how one could adapt their proof to the present case involving multiple parameters. So, we provide a direct proof in a more general setting.

Remark 2. In Theorem 1, we assume that q-dimensional test statistics $\mathbf{T}_j$'s are independent. However, within each $\mathbf{T}_j$, we do not impose any restriction on T_{ij}'s.

It would be tempting to develop an alternative method based on the Simes pooled p-values as follows:

Procedure 2.

(1) *Apply the BH method at level α to test H_{0j} against H_{1j} simultaneously for $j = 1, \ldots, m$, based on the Simes pooled p-values P_j, $j = 1, \ldots, m$.*
(2) *Let R denote the total number of null hypotheses rejected.*
(3) *For every $j = 1, \ldots, m$, let $\widetilde{P}_{(1)j} \leqslant \cdots \leqslant \widetilde{P}_{(q)j}$ be the ordered values of $\widetilde{P}_{ij}, i = 1, \ldots, q$. Let $R_j = \max\{i : \widetilde{P}_{(i)j} \leqslant \frac{i}{q} \cdot \frac{R}{m}\alpha\}$, if the maximum exists; otherwise $R_j = 0$. For every i and j with $\widetilde{P}_{ij} \leqslant \frac{R_j}{q} \cdot \frac{R}{m}\alpha$, declare $\delta_{ij} > 0$ or < 0 according to $T_{ij} > 0$ or < 0.*

Remark 3. As the Simes test is known to be more powerful than the Bonferroni test (Simes, 1986), Procedure 2 would be more powerful than Procedure 1. Unfortunately, however, it would not control the mdFDR, as the associate editor pointed out. Consider, for instance, $m = 1$. The augmented test in this procedure in this case reduces to the step-up test with Simes critical values for the q hypotheses. Assume that $q = 10$ and that for half of the hypotheses $\delta_{i1} = 0$ and for the remaining δ_{i1} is very large. Then the familywise error rate (FWER) of the step-up test with Simes critical values (for the test of the q hypotheses) is not controlled; see also Hommel (1988). However, in this scenario, mdFDR $\geqslant$ FWER. Therefore, Procedure 2 loses the control of the mdFDR in this situation. So, we do not formally propose it in this article as a multidimensional directional FDR controlling procedure, though we will consider it along with Procedure 1 in our simulation studies in the next section.

4. A Simulation Study

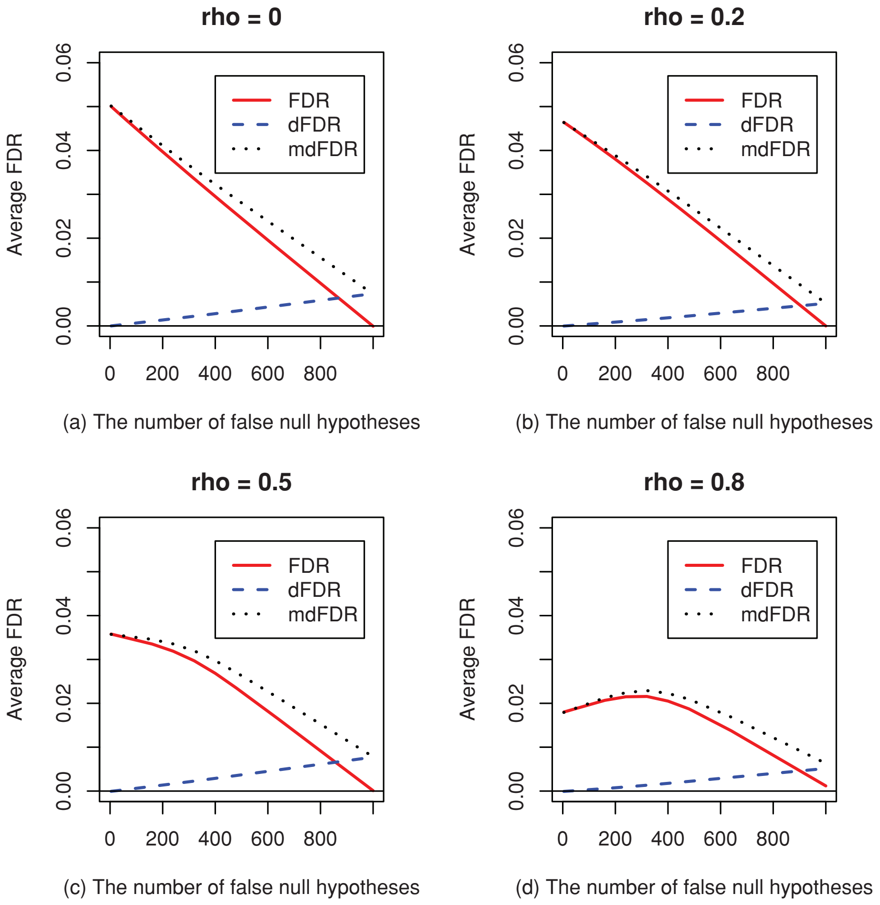
A simulation study was performed to evaluate the performance of our proposed method, Procedure 1. Specifically, investigated the following:

(i) How does it perform in terms of FDR, dFDR, mdFDR, and power under independence as well as different types of dependence?
(ii) How does it perform in terms of the above operating characteristics under independence across genes when we benchmark it against Procedure 2 and the procedure that makes no adjustment to the gene-specific p-values, that is, simply uses $\widetilde{P}_{(1)j}$ as the pooled p-value?
(iii) How does the performance of Procedure 1 under the independence across the genes change as the dimension q increases?

We generated $q+1$ independently distributed m-dimensional random normal vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_{q+1}$, where the components Z_{ij}, $j = 1, \ldots, m$, in each $\mathbf{Z}_i$ are dependent with $Z_{ij} \sim N(\mu_{ij}, 1)$ and have a common correlation ρ. Let $\delta_{ij} = (\mu_{i+1,j} - \mu_{ij})/\sqrt{2}, i = 1, \ldots, q; j = 1, \ldots, m$. Of the m parameter vectors $\boldsymbol{\delta}_j = (\delta_{1j}, \ldots, \delta_{qj})$, $j = 1, \ldots, m$, m_0 were set to a null vector and the δ_{ij}'s in 50%, 25%, and 25% of the remaining $m - m_0$ $\boldsymbol{\delta}_j$'s were selected randomly from the intervals $(-0.75, 0.75)$, $(-4.25, -2.75)$, and (2.75, 4.25), respectively. For each $i = 1, \ldots, q$, and $j = 1, \ldots, m$, the statistic $T_{ij} = (Z_{i+1,j} - Z_{ij})/\sqrt{2}$ for testing $H_{0j}^i : \delta_{ij} = 0$ vs. $H_{1j}^i : \delta_{ij} \neq 0$ and the corresponding two-sided p-value $\widetilde{P}_{ij} = 2\{1 - \Phi(|T_{ij}|)\}$ were then computed, where $\Phi(\cdot)$ is the standard normal cdf. The pooled p-values were calculated according to the Bonferroni and Simes tests, respectively. Procedures 1 and 2 were applied to their respective lists of pooled p-values for testing the m null hypotheses described in (1). We also considered the so-called no-adjustment procedure, which is same as Procedure 1 or 2, except for every hypothesis H_j we do not make any adjustment for its corresponding p-value $P_j = \widetilde{P}_{(1)j}$. For each of these procedures, the number of true null hypotheses that are rejected (Type I errors), the number of $\boldsymbol{\delta}_j$'s corresponding to the false null hypotheses the signs of whose components do not completely match with those assigned by the procedure (directional errors), and the sum of these two numbers (Type I and directional errors) were noted. Finally, the following three proportions among the total number of rejected null hypotheses were calculated—the proportion of Type I errors (the observed value of $V/R \vee 1$), the proportion of directional errors (the observed value of $S/R \vee 1$), and the proportion of Type I and directional errors (the observed value of $(V+S)/R \vee 1$). These steps were repeated 10,000 times and the simulated values of the FDR, dFDR, and mdFDR were obtained by averaging out the 10,000 values of the above three proportions.

Figure 1 presents the simulated FDR, dFDR, and mdFDR; Figure 2 presents standard deviation of the simulated mdFDR; and Web Figure 1 presents the simulated average power (the proportion of false null hypotheses that are correctly rejected with correct assigned signs) of Procedure 1 plotted against the number of false null hypotheses for $m = 1000$, $q = 5$, $\alpha = 0.05$, and $\rho = 0$ (independence), 0.2, 0.5, and 0.8.

Some interesting observations can be made from Figure 1. With increasing number of truly significant genes, the FDR steadily decreases to zero as long as the dependence across the genes is low or moderately high, while the dFDR slowly increases from zero to a value slightly less than 0.01, no matter



(a) The number of false null hypotheses

(b) The number of false null hypotheses

(c) The number of false null hypotheses

(d) The number of false null hypotheses

Figure 1. Performance of Procedure 1 under dependence across genes in terms of its control of the FDR, dFDR, and mdFDR for $m = 1000$, $q = 5$, $\alpha = 0.05$, and $\rho = 0$, 0.2, 05 and 0.8. This figure appears in color in the electronic version of this article.

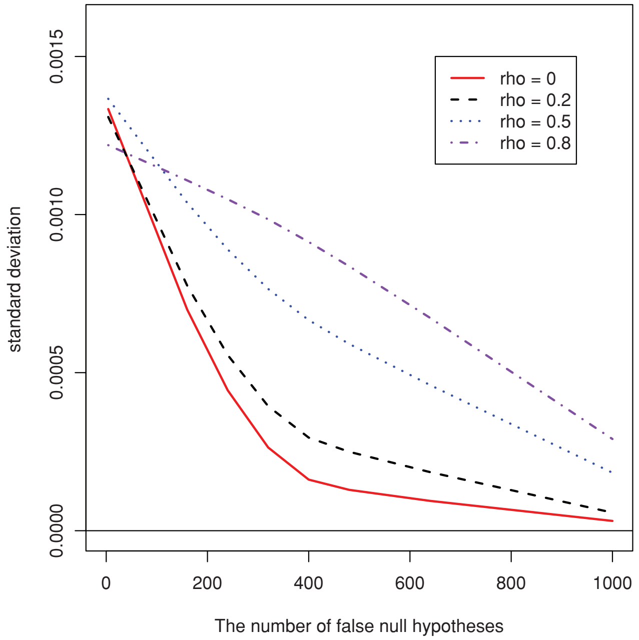


Figure 2. Standard deviation of the mdFDR of Procedure 1 under dependence across genes for $m = 1000$, $q = 5$, $\alpha = 0.05$, and $\rho = 0$, 0.2, 05, and 0.8. This figure appears in color in the electronic version of this article.

what the dependence across the genes is, as long as it is nonnegative. Consequently, when the genes are not too highly dependent, with increasing number of truly significant genes although the mdFDR decreases, implying that Procedure 1 as an mdFDR controlling procedure becomes more conservative, it does not however reach zero (see Figures 1(a)–1(c)). When the genes are highly dependent, as we see from Figure 1(d), Procedure 1 becomes less conservative as the number of truly significant genes begins to increase from zero, but eventually it becomes more conservative as this number becomes larger.

Also, as seen from Figure 2, the standard deviation of the estimated mdFDR is very small. From Web Figure 1 we see that as the dependence across genes increases, the change in power is small. Overall, the effect of dependence across genes on the performance of the proposed procedure is relatively small. As suggested by the associate editor, under the above simulation settings, we also evaluated the performance of Procedure 1 for $\boldsymbol{\delta}_j = (100, 0, \ldots, 0)$, in which one component is extremely large and the rest are zero. For such nonnull $\boldsymbol{\delta}_j$, there is a high chance that it is detected to be nonnull and its one or more zero components are declared to be positive or negative. That is, in such scenarios, it is more possible to make directional errors. Web Figure 2 presents the simulated FDR, dFDR, and mdFDR and Web Figure 3 presents the standard deviation of the simulated mdFDR. As expected, in the case of a nonnull pattern, the dFDR is increasing as the number of false nulls increases, and is much larger than that for the previous scenario. However, the mdFDR is still controlled under a prespecified level (see Web Figure 2). As seen from Web Figure 3, the standard deviation of the estimated mdFDR is still very small for such scenarios.

Web Figure 4 presents an answer to question (ii). As we see from this figure, Procedures 1 and 2 behave quite similarly, at least when the dependence across the genes is not of concern, in terms of controlling the FDR, dFDR, and mdFDR and the power, though Procedure 2 is slightly more liberal as expected. Also as expected, if no adjustment is made to gene-specific p-values, we lose the control of the FDR and mdFDR, with the maximum reaching 0.2. It seems surprising that, even without any adjustment to gene-specific p-values, the dFDR always remains low, though it becomes larger compared to that for Procedures 1 and 2 as the number of false nulls increases.

Web Figures 5 and 6 provide an answer to question (iii). It is interesting to note that the performance of Procedure 1 in terms of controlling the FDR, dFDR, and mdFDR is unaffected by the dimension q when the dependence across the genes is not present. The power, of course, increases with increasing dimension.

5. An Application to Time-Course Gene Expression Data

Lobenhofer et al. (2002) investigated the effect of estrogen on the expression of cell-cycle genes as MCF-7 breast cancer cells go through the cell division cycle. A normal cell division cycle consists of four major phases, namely, the G1 (or Gap 1), S (Synthesis), G2 (or Gap 2), and M (Mitosis) phase. Genes involved in the cell cycle (known as *cell-cycle genes*) are expected to attain peak gene expression during the phase in which they have a specific biological function in the cell cycle.

According to Lobenhofer et al. (2002), most estradiol treated MCF-7 cells are expected to go through S, G2, and M phases in 12–36 hours after treatment and complete the cycle in 48 hours. Genes involved in cell growth and related activities are expected to have maximum expression (or minimum expression if they are anti-growth) during 1 or 4 hours and then monotonically decrease (or increase) in expression as cells go through the remaining phases. On the other hand, genes involved in DNA synthesis, repair, and mitosis would have maximum (or minimum) expression during 12 to 36 hours. Thus, such genes may have an *Umbrella* (or *Inverted umbrella*) shaped pattern with a peak or trough during 12 to 36 hours time period. However, according to Lobenhofer et al. (2002), the cells may be asynchronous as they complete the cell division cycle at 48 hours after the exposure. For this reason, the expression of some of the cell-cycle genes may not return to their baseline values at 48 hours but may attain a plateau.

Before exposing the MCF-7 breast cancer cells to estrogen, Lobenhofer et al. (2002) first synchronized all the cells to G1 phase by depriving the cells of serum for 24 hours. Synchronization of cells to the same phase at the beginning of the experiment is important for obtaining reliable gene expression data. They then harvested estradiol-treated cells after 1, 4, 12, 24, 36, or 48 hours of treatment. Gene expressions using cDNA microarray chips were obtained at each time point. Each cDNA microarray chip consisted of 1900 gene probes. With eight replicates at each time point, there were a total of 48 microarray chips across the six time points.

Motivated by the above observations, in this section we apply the proposed methodology to identify some cell-cycle genes by considering four ordered categories of time points, namely, 1 hour ($T1$), 4 hours ($T2$), mid group (i.e., the union of 12, 24, 36 hours) ($T3$), and 48 hours ($T4$) after treatment. Thus, the sample sizes in the four groups are 8, 8, 24, and 8, respectively. Since the major cell division related activity takes place during the 12 to 36 hours time interval, we combined those time periods together to contrast that period from initial cell growth period (1, 4 hours) and the end of mitosis (48 hours).

Suppose $\mu_{T1,j}$, $\mu_{T2,j}$, $\mu_{T3,j,}$, and $\mu_{T4,j}$ denote the mean gene expression of gene j, $j = 1, 2, \ldots, 1900$, during time periods $T1$, $T2$, $T3$, and $T4$, respectively. Using notations from the previous section, we let $\delta_{1j} = \mu_{T1,j} - \mu_{T2,j}$, $\delta_{2j} = \mu_{T2,j} - \mu_{T3,j,}$, and $\delta_{3j} = \mu_{T3,j} - \mu_{T4,j}$.

Let T_{ij} denote the test statistic associated with the parameter δ_{ij}, $i = 1, 2, 3$ and $j = 1, 2, \ldots, 1900$. In this application T_{ij} is the usual two-sample t-test statistic and since the underlying data are not necessarily normally distributed, nonparametric bootstrap methodology based on 10,000 bootstrap samples is used for computing the p-values $\widetilde{P}_{ij}$ associated with hypotheses on δ_{ij}. We then calculate the Bonferroni pooled p-value P_j for each gene j after computing $\widetilde{P}_{ij}, i = 1, 2, 3$.

By applying our proposed method, Procedure 1 to the list of the pooled p-values P_j's, we identified 86 differentially expressed genes at level $\alpha = 0.05$ of which 19 had an umbrella-shaped response, 3 inverted umbrella, 32 increased in expression from $T1$ to $T3$ and then plateaued (i.e., for some gene j, $\mu_{T1,j} \leqslant \mu_{T2,j} \leqslant \mu_{T3,j} = \mu_{T4,j}$, with at least one strict inequality). An opposite response was seen with 22 genes that had decreased expression from $T1$ to $T3$ and then plateaued (i.e., for some gene j, $\mu_{T1,j} \geqslant \mu_{T2,j} \geqslant \mu_{T3,j} = \mu_{T4,j}$, with at least one strict inequality). We also discovered 10 genes that had a flat expression until $T3$ and then a decrease in response from $T3$ to $T4$.

Comparing our results with those of Lobenhofer et al. (2002) and Peddada et al. (2003), we found that of the 86 genes we identified, 39 were also identified in at least one of the two previous papers. This included 8 of 13 DNA replication/repair genes identified by Lobenhofer et al. (2002). Among the five that were not identified by our procedure, we note that except for MCM7, which may be significant at $\alpha = 0.10$, all others had large p-values that were not significant even at $\alpha = 0.20$. Interestingly, in addition to MCM3 that was identified by both Lobenhofer et al. (2002) and Peddada et al. (2003), we identified a well-known cell-cycle gene MCM4 (`http://www.cyclebase.org`).

An important step in DNA synthesis during the S phase is the binding of complex proteins to DNA for recruiting other proteins necessary for DNA synthesis. One such complex protein is the replication factor C. Lobenhofer et al. (2002) identified one subunit of this protein, known as replication factor C3. Later the order-restricted inference-based methodology of Peddada et al. (2003) identified two additional subunits of this protein, namely, replication factors C4 and C5. Interestingly, the proposed methodology identified subunits C2, C3, and C5 as significant genes, thus reinforcing the earlier findings and adding one more subunit to the previous list of replication factor C. Furthermore, based on the proposed methodology it is possible to conclude that the subunits C2, C3, and C5 have peak expression during the 12, 24, or 36 hours time period where the DNA synthesis and replication takes place.

Furthermore, similar to the order-restricted inference procedure of Peddada et al. (2003), the proposed methodology identified the cyclin-dependent kinase inhibitor 1 A (p21 and Cip 1) as repressed during 12 to 36 hours. This gene was not identified by Lobenhofer et al. (2002).

A complete list of all genes identified by this procedure is provided in the Supplementary Materials.

6. Concluding Remarks

In microarray gene expression studies, researchers are often not only interested in identifying differentially expressed genes under different biological conditions, but are also interested in detecting trends in mean response over ordered categories. For instance, in the simple case of two categories (normal vs. tumor tissue), researchers are not only interested in identifying significant genes across these two categories, but they are also interested in further identifying the down- and up-regulated genes. As the number of ordered categories increases, the trends or directional patterns become complex and the number of directional patterns increases. Except for the usual Type I errors, this also potentially results in a relatively high frequency of directional errors. Hence, it is important to develop statistical methods of identifying trends in mean response over ordered categories while maintaining a control over both the Type I and directional errors.

The approach proposed in this article provides such a methodology. Differently from existing statistical methods (Peddada et al., 2003; Lin et al., 2007), we have formulated the problem of identifying trends in mean response over ordered categories as a multiple testing problem involving successive comparisons and further directional decisions on the multidimensional parameter of each gene. To deal with this problem, we have first suggested a general multidimensional BH-type directional procedure using the Bonferroni test for controlling the mdFDR, an overall measure of both Type I and directional errors within the framework of the FDR, and theoretically proved that the proposed procedure controls the mdFDR at a prespecified level when the underlying test statistics are independent across the genes. We evaluated the performance of the introduced procedure in the case of dependence through a simulation study. Finally, the whole proposed methodology has been applied to analyze a time-course microarray data and some interesting results have been obtained.

Although our focus was on identifying individual gene expression profile or trend over the ordered categories in some common microarray experiments such as time-course or dose-response experiments, the proposed methodology can also be applied in National Toxicology Program (NTP) studies, where researchers are interested in determining whether for a given tumor type there is a significant dose effect and then identifying its dose-response profile.

The methodology proposed in this article provides an interesting starting point towards addressing the complex yet important problem of controlling both Type I and directional errors in multiple testing involving multidimensional parameters. The mdFDR controlling property of the proposed directional BH procedure has been established under the assumption that the underlying test statistics are independent across the genes. When gene expressions are obtained by drawing samples from same subjects over time, such an assumption need not be valid. In such cases, not only do we have dependence among gene expressions at a given time point but there may be temporal dependence among gene expressions at different time points. It will be interesting to theoretically investigate the performance of the proposed directional BH procedure under such complex dependence structures. In addition, it will also be interesting to develop more powerful adaptive BH directional FDR procedure by exploiting knowledge of the proportion of true null hypotheses.

7. Supplementary Materials

Web Appendices and Figures referenced in Sections 3 and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tib.org`. A complete list of all genes identified in Section 5 are also available at the Biometrics website.

Acknowledgements

The research of the first and third authors is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04) and the research of the second author is supported by a grant from U.S. National Science Foundation (DMS-0603868). We thank Grace Kissling and David Dunson for several useful comments that led to improved presentation of the manuscript. We also thank the editor, the associate editor (AE), and the referees for their comments that have led to substantial improvement in the presentation of the manuscript. We are particularly thankful to the AE for spotting an error in the proof of Theorem 1, for providing an interesting example showing that Procedure 2 cannot control the mdFDR (described in Remark 3), and for making several other important comments that led to improved simulation studies and better exposition of the article.

References

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

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25,** 60–83.

Benjamini, Y., Hochberg, Y., and Kling, Y. (1993). False discovery rate control in pairwise comparisons. Working paper 93-2, Department of Statistics and Operation Research, Tel Aviv University.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* **100,** 71–93.

Bernhard, G., Klein, M., and Hommel, G. (2004) Global and multiple test procedures using ordered p-values—A review. *Statistical Papers* **45,** 1–14.

Bochkina, N. and Richardson, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics* **63,** 1117–1125.

Finner, H. (1994). Testing multiple hypotheses: General theory, specific problems, and relationships to other multiple decision procedures. Habilitationsschrift, Fachbereich IV Mathematik, Univ. Trier.

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *Annals of Statistics* **27,** 274–289.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on modified Bonferroni test. *Biometrika* **75,** 383–386.

Lin, D., Shkedy, Z., Yekutieli, D., Burzykowski, T., Göhlmann, H., Bondt, A., Perera, T., Geerts, T., and Bijnens, L. (2007). Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resampling-based inference. *Statistical Applications in Genetics and Molecular Biology* **6**(1), 26.

Liu, W. (1996). Control of directional errors with step-up multiple tests. *Statistics and Probability Letters* **31,** 239–242.

Lobenhofer, E., Bennett, L., Cable, P., Li, L., Bushel, P., and Afshari, C. (2002). Regulation of DNA replication fork genes by 17 beta-estradiol. *Molecular Endocrinology* **16,** 1215–1229.

Peddada, S., Lobenhofer, E., Li, L., Afshari, C., Weinberg, C., and Umbach, D. (2003). Gene selection and clustering for time-course and dose response microarray experiments using order-restricted inference. *Bioinformatics* **19,** 834–841.

Sarkar, S. K. and Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* **92,** 1601–1608.

Sarkar, S. K., Sen, P. K., and Finner, H. (2004). On two results in multiple testing. In *Recent Developments in Multiple Comparisons.* IMS Lectures Notes-Monograph Series, 47, Y. Benjamini, F. Bretz, and S. Sarkar (eds), 89–99. Beachwood, Ohio: Institute of Mathematical Statistics.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Annals of Statistics* **8,** 1342–1347.

Shaffer, J. P. (2002). Multiplicity, directional (type III) errors, and the null hypothesis. *Psychological Methods* **7,** 356–369.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.

Tamoto, E., Tada, M., Murakawa, K., Takada, M., Shindo, G., Teramoto, K., Matsunaga, A., Komuro, K., Kanai, M., Kawakami, A., Fujiwara, Y., Kobayashi, N., Shirata, K., Nishimura, N., Okushiba, S., Kondo, S., Hamada, J., Yoshiki, T., Moriuchi, T., and Katoh, H. (2004). Gene-expression profile changes correlated with tumor progression and lymph node metastasis in esophageal cancer. *Clinical Cancer Research* **10,** 3629–3638.

Williams, V. S., Jones, V., and Tukey, J. W. (1999). Control errors in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics* **24,** 42–69.

Received July 2008. *Revised April* 2009.
Accepted April 2009.