University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 40

Semiparametric methods for identification of tumor progression genes from microarray data

Debashis Ghosh* Arul Chinnaiyan†

*University of Michigan, ghoshd@psu.edu
†University of Michigan Pathology and Urology, arul@umich.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/umichbiostat/paper40

Copyright ©2004 by the authors.

Semiparametric methods for identification of tumor progression genes from microarray data

Debashis Ghosh and Arul Chinnaiyan

Abstract

The use of microarray data has become quite commonplace in medical and scientific experiments. We focus here on microarray data generated from cancer studies. It is potentially important for the discovery of biomarkers to identify genes whose expression levels correlate with tumor progression. In this article, we develop statistical procedures for the identification of such genes, which we term tumor progression genes. Two methods are considered in this paper. The first is use of a proportional odds procedure, combined with false discovery rate estimation techniques to adjust for the multiple testing problem. The second method is based on order-restricted estimation procedures. The proposed methods are applied to data from a prostate cancer study. In addition, their finite-sample properties are compared using simulated data.

Semiparametric methods for identification of tumor progression genes from microarray data

Debashis Ghosh[1] and Arul Chinnaiyan[2]

Departments of [1]Biostatistics and [2]Pathology and Urology

University of Michigan

Ann Arbor, MI, 48109-2029, USA

Summary

The use of microarray data has become quite commonplace in medical and scientific experiments. We focus here on microarray data generated from cancer studies. It is potentially important for the discovery of biomarkers to identify genes whose expression levels correlate with tumor progression. In this article, we develop statistical procedures for the identification of such genes, which we term tumor progression genes. Two methods are considered in this paper. The first is use of a proportional odds procedure, combined with false discovery rate estimation techniques to adjust for the multiple testing problem. The second method is based on order-restricted estimation procedures. The proposed methods are applied to data from a prostate cancer study. In addition, their finite-sample properties are compared using simulated data.

Keywords: Gene Expression; Metastasis; Mixture Models; Multiple Comparisons; Prostate Cancer.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

1. Introduction

The use of DNA microarray technology has allowed for new understanding of various cancers. The hybridization of cDNA to arrays containing thousands of genes and ESTs permits a global genomewide evaluation of tumor samples. This technology has led to development of statistical methodology in various areas of microarray data analysis, such as methods for differential expression (Efron et al., 2001; Dudoit et al., 2002b), clustering (Eisen et al., 1998) and classification (Hastie et al., 2000; Dudoit et al., 2002a). However, it has been also realized that microarrays have a fundamental level of experimental variation and that global tasks, such as reconstruction of gene networks, still remains a very elusive task.

An alternative approach to analyzing these data is to incorporate available biological knowledge. The motivating example is from a microarray experiment in prostate cancer (Dhanasekaran et al., 2001). We have profiled tissue samples from various stages of prostate cancer (e.g., normal adjacent prostate, benign prostatic hyperplasia, localized prostate cancer, advanced metastatic prostate cancer). The samples are linked to a patient clinical database that has other parameters, such as Gleason score, survival time and status, and time to PSA recurrence. One of the main hypotheses of interest to scientists is that there exist distinct sets of genes and proteins dictate progression from precursor lesion, to localized disease, and finally to metastatic disease. This hypothesis is biological in nature and is focused upon learning about which genes are involved in cancer pathways. We will refer to genes satisfying this hypothesis as tumor progressor genes.

The ideal design for studying development of gene expression profiles in tumors would be a longitudinal experiment. The tumor is commonly thought to originate as a progenitor cell and goes through several stages of progression (e.g., benign hyperplasia, in situ). Such a model for tumor progression has been postulated by Fearon and Vogelstein (1990). If it were possible to sample the same tumor in these various stages of development and to generate gene expression profiles for each of the time points, then this would provide the optimal setting for studying the effect of gene expression profiles

http://biostats.bepress.com/umichbiostat/paper40

on tumor progression. While this is possible for studying tumor volume progression in mouse models (Ferrante et al., 2000), this is not feasible for humans as tumor tissue is completely resected from the patient. The data typically available are the gene expression profiles for the tumor sampled at one point of time in the tumor progression for a given patient.

One can view the gene expression profile as a high-dimensional phenotypic property of the tumor. There has been a rich literature existing on statistical models for tumor progression in which the phenotype considered was size of the tumor (Kimmel and Flehinger, 1991; Xu and Prorok, 1997). However, no such development has occurred for gene expression profile and its effects on tumor progression. By incorporating clinical information on stage of the tumor (e.g., precursor lesion, localized prostate cancer and metastatic lesion), one can utilize microarray data potentially in a more efficient fashion. However, an important feature that must be considered is that many of the genes are noninformative about tumor progression. In this article, we seek to develop statistical methods for characterizing the relationship of gene expression profile on tumor progression. The gene expression profile is treated as a phenotype of the tumor that we wish to associate with clinical progression. We develop two semiparametric methods to address this goal. The structure of this paper is as follows. In Section 2, we describe the data structures and two statistical procedures for analyzing the effects of gene expression on tumor stage. The first class of procedures is based on the proportional odds model (Agresti, 2002) and complements some of the existing methodology on multiple testing procedures (Efron et al., 2001; Tusher et al., 2001). A second class of procedures attempts to exploits constraints on the ordering of the gene expression profiles (Robertson, Wright and Dykstra, 1988; Peddada et al., 2003). The methods are compared in simulation studies and illustrated with application to the previously mentioned prostate cancer data in Section 3. We conclude with some discussion in Section 4.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

2. Systems and Methods

2.1. Notation and Preliminaries

Let D denote the stage of disease; we assume that it takes values $(1, \ldots, d)$, where increasing numbers corresponding to progressively advanced stages of disease. Thus, D will be treated as an ordinal variable here and in the sequel. We will assume that $d > 2$. Let $\mathbf{X}$ denote the G-dimensional gene expression profile. We observe the data $(D_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, iid observations from the joint distribution of $(D, \mathbf{X})$. In most situations we consider, G is typically much larger than n. We will assume throughout the paper that the gene expression data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ have been suitably preprocessed and normalized both within and across slides.

2.2. Proportional odds model

Define $Pr(A)$ to be the probability of the event A. One simple model for associating gene expression with stage of disease is the proportional odds model (Agresti, 2002, §7.2.2): for $r = 0, 1, \ldots, d$,

$$\log\left\{\frac{Pr(D_i \leq r)}{Pr(D_i > r)}\right\} = \alpha_{rg} + \beta_g X_{ig}, \tag{1}$$

where $(\alpha_{0g}, \ldots, \alpha_{dg})$ are gene-specific cutpoints, β_g is a gene-specific regression coefficient, and X_{ig} is the gth component of $\mathbf{X}_i$ $(i = 1, \ldots, n; g = 1, \ldots, G)$. Note that α_{rg} is increasing in r since $Pr(D_i \leq r | X_{ig})$ is increasing in r. Positive values of β_g indicate that higher values of gene expression are associated with increased odds that D is small, while negative values of β_g demonstrate the converse.

An alternative motivation of model (1) is to use a latent underlying random variable Z_{ig}, where $Z_{ig} - \beta_g X_{ig}$ has a standard logistic distribution, i.e. $Pr(Z_{ig} - \beta_g X_{ig} \leq u) = \exp(u)/\{1 + \exp(u)\}$. Then the event $\{D_i = d\}$ corresponds to the event $\alpha_{d-1} < Z_{ig} \leq \alpha_d$. This implies that

$$\begin{aligned} Pr(D_i \leq d) &= Pr(Z_i \leq \alpha_d) \\ &= Pr(Z_i - \beta_g X_{ig} \leq \alpha_d - \beta_g X_{ig}) \\ &= \frac{\exp(\alpha_d - \beta_g X_{ig})}{1 + \exp(\alpha_d - \beta_g X_{ig})}. \end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/umichbiostat/paper40

The proportional odds model can be fit using many standard software packages, such as SAS or S-Plus.

In most microarray studies, G is much larger than n. The model in (1) is univariate and does not incorporate dependence between genes. One method of doing this is to incorporate a second stage in which $\beta_1, \ldots, \beta_G$ is a random sample from a mixture distribution:

$$\beta_g \overset{iid}{\sim} \pi_0 F_1 + (1 - \pi_0) F_2. \tag{2}$$

In model (2), π_0 represents the proportion of genes that are noninformative about tumor progression, while the remaining percentage, $1 - \pi_0$ are indicators of gene progression. F_1 and F_2 are the distribution functions for the noninformative and informative tumor progressor genes, respectively. The two-stage formulation (1) and (2) implies that gene expression measurements are dependent.

It turns out that the model (1)-(2) has a connection with multiple testing procedures based on the false discovery rate (Benjamini and Hochberg, 1995; Storey, 2002). We consider the G univariate null hypotheses $H_{0g} : \beta_g = 0,\ g = 1, \ldots, G$. Mimicking the arguments of Theorem 1 in Storey (2002), we have that based on the two-stage model (1)-(2), the gene-specific false-discovery rate is given by $FDR_g = Pr(H_{0g}|T_g \in R_g)$, where R is the rejection region for the gth test statistic T_g, $g = 1, \ldots, G$. We have the following algorithm for the estimation of gene-specific false-discovery rates:

1. Fit (1) for each gene g using maximum likelihood for $g = 1, \ldots, G$.

2. Calculate a p-value using $|\hat{\beta}_{1g}|/\hat{SE}(\hat{\beta}_{1g}),\ g = 1, \ldots, G$.

3. Let $p_1, \ldots, p_G$ denote the G p-values. Estimate π_0, the proportion of differentially expressed genes and $F_P(x)$, the cdf of the p-values, by

$$\hat{\pi}_0 = \frac{W(\lambda)}{(1 - \lambda)G}$$

and

$$\hat{F}_P(x) = \frac{\min\{R(\gamma), 1\}}{G},$$

where $R(\gamma) = \#\{p_i \leq \gamma\}$ and $W(\lambda) = \#\{p_i > \lambda\}$.

Hosted by The Berkeley Electronic Press

4. For any rejection region of interest $[0, \gamma]$, estimate the gene-specific FDR as

$$\widehat{FDR}(\gamma) = \frac{\hat{\pi}_0 \gamma}{\hat{F}_P(\gamma)\{1 - (1-\gamma)^G\}}.$$

5. Estimate FDR as

$$\widehat{FDR} = \frac{\hat{\pi}_0 \gamma}{\hat{F}_P(\gamma)}.$$

There are two issues in this algorithm that need to be resolved. The first is method of calculating the p-value in step 2 of the algorithm. We use permutation methods where the sample labels $D_1, \ldots, D_n$ are permuted. Note the validity of the method depends on the assumption that under the global null hypothesis of no difference in progression groups for any of the genes, the data are exchangeable. The second issue is the choice of λ. Observe that there is a bias-variance tradeoff in the choice of λ. It turns out that the bias of π_0 is minimized when $\lambda = 1$. This leads to the following algorithm to determine π_0, described by Storey and Tibshirani (2003):

1. Order the G p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(G)}$.

2. Construct a grid of L λ values, $\lambda_1, \ldots, \lambda_L$ and calculate

$$\hat{\pi}_0(\lambda_l) = \frac{\#\{p_j > \lambda\}}{G(1-\lambda)},$$

$l = 1, \ldots, L$.

3. Fit a cubic smoothing spline to the values $\{\lambda_l, \hat{\pi}_0(\lambda_l)\}$, $l = 1, \ldots, L$.

4. Estimate π_0 by the interpolated value at $\lambda = 1$.

Given this algorithm, one can then estimate gene-specific q-values (Tusher et al., 2001; Storey and Tibshirani, 2003) for the individual genes. For the gene with the largest p-value the q-value is given by

$$q(p_{(G)}) = \min_{t \geq p_{(G)}} \frac{\hat{\pi}_0 G t}{\#\{p_j \leq t\}} = \hat{\pi}_0 p_{(G)},$$

and for $i = G-1, G-2, \ldots, 1$, $q(p_{(i)}) = \min(\hat{\pi}_0 G p_{(i)}/i, \hat{p}_{(i+1)})$. This guarantees that the q-values will be monotonically increasing as a function of p-values.

http://biostats.bepress.com/umichbiostat/paper40

While the proportional odds model is a popular model to fit to ordinal data, there is little biological basis for such a model to hold in this setting. In the next section, we will describe an approach to finding tumor progressor genes that does not rely on model (1) and instead incorporates constraints on mean expression profiles. This will lead to estimation of mean profiles under order restrictions (Robertson et al., 1988).

2.3. Order-restricted methods

The biological concept that underlies a tumor progressor gene is that in normal tissue, this gene functions normally, but as a normal cell progresses to precursor lesion to localized cancer to metastatic cancer, the expression of the gene shows a trend. One example of a tumor progressor gene might be an oncogene, which is activated when the normal cell function becomes dysregulated. Such a gene will tend to show higher expression as the tumor develops. Another example of a tumor progressor gene is a tumor suppressor gene. Such genes are inactivated in tumors so that their expression tends to decrease with increasing levels of tumor progression.

Let $\mu_g \equiv (\mu_{g1}, \mu_{g2}, \ldots, \mu_{gd})$ denote the population gene expression means at each of the d stages of progression. Their empirical estimate is given by $\bar{X}_{gl} = n_l^{-1} \sum_{i=1}^{n_l} X_{gi}$, where n_d is the number of tumor samples in disease stage l, $l = 1, \ldots, d$. The idea is to estimate $(\mu_{g1}, \ldots, \mu_{gd})$ under two sets of constraints:

$$C_{MI} = \{\mu_g \in R^d : \mu_{g1} \leq \mu_{g2} \leq \cdots \leq \mu_{gd}\} \tag{3}$$

and

$$C_{MD} = \{\mu_g \in R^d : \mu_{g1} \geq \mu_{g2} \geq \cdots \geq \mu_{gd}\}. \tag{4}$$

It is implicitly assumed that there is at least one strict inequality in constraints (3) and (4). We will refer to (3) and (4) as monotonically increasing and monotonically decreasing candidate profiles for the gth genes, $g = 1, \ldots, G$. Thus, we have specified *a priori* that our interest is in finding genes whose expression profiles are consistent with (3) and (4). The goal is to develop a statistical framework for classifying genes into one of these profiles.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

If we wish to estimate μ_g subject to the constraints (3) and (4), then order-restricted techniques (Robertson et al., 1988) must be used. For estimation subject to (3), the optimization problem is to find $\mu_g \in C_{MI}$ that minimizes

$$\sum_{l=1}^{d} w_l (\bar{X}_{gl} - \mu_{gl})^2,$$

where $w_l = (\hat{\sigma}_{gl}^2 / n_l)^{-1}$, and

$$\hat{\sigma}_{gl}^2 = (n_l - 1)^{-1} \sum_{i=1}^{n_l} (X_{gi} - \bar{X}_{gl})^2.$$

An analogous optimization problem can be written for estimation subject to $\mu_g \in C_{MD}$. Note that we have incorporated weights $w_1, \ldots, w_d$ into the optimization problem because of the variation in sample sizes across various tumor progression stages and difference in gene-specific variability. The estimation problem is an isotonic regression problem and can be solved using the pooled adjacent violators algorithm (Robertson et al., 1988).

For each gene, we can estimate a profile subject to (3) and (4). Denote these profiles as $(\hat{\mu}_{g1}^{MI}, \ldots, \hat{\mu}_{gd}^{MI})$ and $(\hat{\mu}_{g1}^{MD}, \ldots, \hat{\mu}_{gd}^{MD})$, $g = 1, \ldots, G$. We then want to classify the observed gene profile as being more consistent with the former or latter pattern. To do this, we will calculate $L_g^{MI} \equiv \hat{\mu}_{gd}^{MI} - \hat{\mu}_{g1}^{MI}$ and $L_g^{MD} \equiv \hat{\mu}_{gd}^{MD} - \hat{\mu}_{g1}^{MD}$ for the gth gene, $g = 1, \ldots, G$. Suppressing dependence on g, the quantities L^{MI} and L^{MD} are special cases of the l_∞ norms used in other order-restricted inference procedures (Dunnett, 1955; Williams, 1977). If $L_g^{MI} > L_g^{MD}$, then we classify the gene as having a tumor expression pattern consistent with a monotone increasing candidate profile; otherwise, we classify the profile gene as monotone decreasing profile.

Define the null candidate profile $C^N \equiv \{\mu_g \in R^d : \mu_{g1} = \mu_{g2} = \cdots = \mu_{gd}\}$. For the gth gene ($g = 1, \ldots, G$), we wish to test the null hypothesis $H_0 : \mu_g \in C^N$ versus the alternative $H_1 : \mu_g \in C^{MI} \cup C^{MD}$. We employ the following algorithm:

1. Estimate μ_g under constraints C^{MD} and C^{MI} using the weighted pooled adjacent violators algorithm.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/umichbiostat/paper40

2. For the gth gene, calculate $L_g = \max(L_g^{MI}, L_g^{MD})$, $g = 1, \ldots, G$.

3. Resample the dataset by sampling n_l samples with replacement for the lth stage. Repeat steps 1 and 2 for the genes in the bth bootstrapped dataset; calculate $L_{g,b}^{MI}$ and $L_{g,b}^{MD}$ for the gth gene in the bth dataset ($g = 1, \ldots, G; b = 1, \ldots, B$). Calculate $L_{g,b} = \max(L_{g,b}^{MI}, L_{g,b}^{MD})$, $g = 1, \ldots, G$, $b = 1, \ldots, B$.

4. Calculate a critical value for L_g based on the empirical distribution of $L_{g,b}$, $b = 1, \ldots, B$. If L_g is bigger than the upper αth percentile, then classify it into the observed profile; otherwise do not classify into a profile.

We denote the proportion of bootstrapped datasets with $L_{g,b} \leq L_g$ as the r-value. A high r-value corresponds to increased confidence that the observed profile is inconsistent with the null hypothesis; note that this interpretation is inverse that of a p-value or q-value. The procedures developed in this section attempt to use information on the shape of the gene expression profile over tumor progression stages in order to identify tumor progression genes and does not rely on an assumption such as proportional odds in model (1).

3. Numerical Examples

3.1. Simulation studies

To examine the finite-sample properties of the proposed methodologies, several simulation studies were conducted. We generated data for $d = 3$ classes of tumors; we considered 10 and 30 samples in each group. Data on $G = 1000$ genes were generated; two models were considered. For the first model, the tumor sample label and gene expression measurement, (D_i, X_{ig}), were generated from a logistic distribution so that

$$Pr(D_i < d) = \frac{\exp(\alpha_d - \beta_g X_{ig})}{1 + \exp(\alpha_d - \beta_g X_{ig})}.$$

At the second stage of this model, β_g were generated from (2) with F_1 being the cdf of a N(0,1) r.v., and F_2 being that of a N(4,1) r.v. Different values of π_0 were considered;

Hosted by The Berkeley Electronic Press

$\pi_0 = 0.7, 0.8, 0.9$. In this situation, the proportional odds model holds. The second scenario consisted of generating gene expression measurements from a normal model, where for a fraction of genes, π_0, mean values did not increase across groups, while for the remaining $(1 - \pi_0)$, mean values increased across group. We set $\pi_0 = 0.7, 0.8, 0.9$ and $\mu_1 = -2$, $\mu_2 = 0$ and $\mu_3 = 2$; all gene expression variances were taken to be 1. For this setting, the proportional odds assumption is not satisfied. 2000 simulation samples were considered for each setting; for both the proportional odds and order-restricted methodology described earlier, 1000 permutations were used to calculate q-values or p-values, as appropriate. Because we knew in this simulation which genes were differentially expressed and which genes were not, we used a sensitivity and specificity measure to summarize each simulation study. For the FDR-based proposed methodology, we defined sensitivity as having a q-value $<= 0.05$ among the differentially expressed genes and specificity as having a q-value > 0.05 among the non-differentially expressed genes. For the order-restricted method, we defined sensitivity as being above the 95th percentile of the bootstrap distribution for differentially expressed genes and specificity as being below the 95th percentile of the bootstrap distribution among nondifferentially expressed genes. The results are summarized in Tables 1 and 2. Based on these results, we find that the proportional odds method tends to be more powerful than the order-restricted inference method in both situations. We also note that there are substantial increases in power of detection due to sample size.

3.2. Prostate cancer data

The dataset we will be using to illustrate the ideas in the paper is from a molecular profiling study in prostate cancer (Dhanasekaran et al., 2001). The benign and malignant prostate tissues were analyzed using a 9984 element (10K) human cDNA microarray. A two-channel (Cy5/Cy3) scheme was utilized. While there are 9984 genes on the original array and 101 samples from $d = 3$ tumor classes: benign precursor, localized prostate cancer and metastatic prostate cancer. We did some preprocessing to reduce the number of genes considered; namely, we filtered out genes that are reported

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/umichbiostat/paper40

as missing in more than 10% of the samples. This left a total of $G = 7910$ genes for analysis.

We first performed the analysis based on fitting the proportional odds model described in Section 2.2. A spreadsheet containing gene names, estimated regression coefficients and associated Wald statistics can be downloaded as the file pgenes1.csv from the following website:

http://www.sph.umich.edu/∼ghoshd/COMPBIO/TPROG/.

Based on permutation methods, we calculated p-values and then applied the false discovery rate estimation procedure of Storey and Tibshirani (2003). The results are summarized in Figure 1. Based on the graphs, 1582 genes have q-values less than 0.001, 753 have those less than 0.0001, and 313 less than 0.00001.

We next applied the order-restricted methodology described in Section 2.3. A list of genes, statistics and r-values can be downloaded as the file pgenes3.csv at the URL given above. Based on the analysis, we find 1141 genes that have an r-value greater than 0.9, 956 with an r-value greater than 0.95, 663 greater than 0.99, 426 greater than 0.999, and 309 with an r-value greater than 0.9999.

In studies such as these, investigators are typically interested in developing a gene list of candidate biomarkers that they would be interested in performing further validation analyses, such as immunohistochemistry or quantitative RT-PCR. Because we do not know the true underlying model for the data, we used a consensus approach combining the results of the two analyses described in Section 2.2 and 2.3. We first considered only genes that had an r-value greater than 0.9999. Then, genes were ranked on the t-statistic from the proportional odds model and then based on the estimated coefficient. Here, we focus on genes that show decreased expression with increasing tumor progression. We focus on three results from such an analysis. The first is the identification of homologs of mammalian transcription factors. Among the top 200 genes are a homolog of a yeast transcription factor (Sec23 - Hs. 753381), a homolog of the FAT tumor suppressor in Drosophila (Hs. 591266), a homolog of a transcription

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

factor in Xenopus laevis (Hs. 760299), a homolog of the snail transcription factor in Drosophila (Hs. 293339) and another Drosophila transcription factor homolog, frizzled, (Hs. 298122). Given the recent discovery (Varambally et al., 2002) of a prostate cancer biomarker that is a homolog of a Drosophila transcription factor, EZH2, it is of interest to the investigator to identify other homologs of mammalian transcription factors that might be involved in cancer dysregulation. Another finding is the decreased expression of cell surface and cell adhesion genes and products in the top 200 list. This includes genes such as catenin (Hs. 364921), moesin (Hs. 131362), integrin (Hs. 502527), and integrin, beta 1 (Hs. 343072). Given that a hallmark of metastatic tumors is the lack of cell differentiation and loss of adhesion to epithelial cells, it is worthwhile to follow these genes up further.

A caveat of these results is that they are not confirmatory but rather hypothesis-generating. Further computational and/or biological experiments would be needed to validate these findings.

4. Discussion

In this article, we have described a model-based and model-free procedure for identifying genes that associate with tumor progression in cancer studies using microarray data. These techniques complement the multiple testing methods currently available for microarray data (Efron et al., 2001; Dudoit et al., 2002b).

A crucial assumption in the methods developed so far is that there is no confounding of gene expression by other clinical factors. One could imagine that sample characteristics, such as age of the patient or tissue heterogeneity, could confound the association between gene expression and tumor progression. If these characteristics have been measured, then we would extend the proportional odds model approach by including them as covariates. As discussed in Ghosh and Chinnaiyan (2003), the validity of p-values derived from permutation testing in this setting would be questionable. How to incorporate covariates in the order-restricted procedure remains an open question.

We have attempted to treat gene expression as a phenotypic property of the tumor

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/umichbiostat/paper40

sample and correlate it with tumor progression. An alternative approach, not considered in this paper, would be to formulate a stochastic modelling approach in which a mechanistic model for gene expression development is postulated. This has precedents in the mathematical modelling literature (Yakovlev and Tsodikov, 1996). This is an area that is currently under exploration.

Acknowledgments

The research of the first author was supported by grant NIH 1R01GM72007-01 from the Joint DMS/DBS/NIGMS Biological Mathematics Program.

References

Agresti, A. A. (2002). *Categorical Data Analysis, 2nd edition.* New York: Wiley.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *JRSS-B* **57**, 289–300.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A. and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826.

Dudoit, S., Fridyland, J. F. and Speed, T. P. (2002a). Comparison of discrimination methods for tumor classification based on microarray data. *Journal of the American Statistical Association* **97**, 77 – 87.

Dudoit, S., Yang, Y.–H., Callow, M. J. and Speed, T. P. (2002b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111 – 140.

Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096 – 1121.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

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

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–14868.

Fearon, E. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* **61**, 759 – 767.

Ferrante, L., Bompadre, S., Possati, L. and Leone, L. (2000). Parameter estimation in a Gompertzian stochastic model for tumor growth. *Biometrics* **56**, 1076 – 1081.

Ghosh, D. and Chinnaiyan, A. M. (2003). Covariate adjustment in the analysis of gene expression data. Technical report, Department of Biostatistics, University of Michigan.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D. and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**, RESEARCH0003.

Kimmel, M. and Flehinger, B. J. (1991). Nonparametric estimation of the size-metastasis relationship in solid cancers. *Biometrics* **47**, 987 – 1004.

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

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

Storey, J. D. (2002). A direct approach to false discovery rates. *JRSS-B* **64**, 479 – 498.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/umichbiostat/paper40

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA* **100**, 9440 – 9445.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to ionization radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116–5121.

Williams, D. (1977). Some inference procedures for monotonically ordered normal means. *Biometrika* **64**, 9 – 14.

Xu, J. L. and Prorok, P. C. (1997). Nonparametric estimation of solid cancer size at metastasis and probability of presenting with metastasis at detection. *Biometrics* **53**, 579 – 591.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications.* Singapore: World Scientific Press.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1: Summary of simulation results for proportional odds scenario.

		Proportional Odds		Order-restricted	
N	π_0	Sens.	Spec.	Sens.	Spec.
10	0.70	0.32	0.78	0.14	0.75
10	0.80	0.30	0.85	0.12	0.81
10	0.90	0.28	0.91	0.11	0.86
30	0.70	0.82	0.88	0.55	0.85
30	0.80	0.74	0.93	0.61	0.87
30	0.90	0.66	0.95	0.63	0.89

Table 2: Summary of simulation results for non-proportional odds scenario.

		Proportional Odds		Order-restricted	
N	π_0	Sens.	Spec.	Sens.	Spec.
10	0.70	0.24	0.71	0.16	0.76
10	0.80	0.22	0.79	0.14	0.84
10	0.90	0.18	0.86	0.12	0.88
30	0.70	0.72	0.82	0.95	0.80
30	0.80	0.76	0.88	0.98	0.85
30	0.90	0.82	0.91	0.99	0.89

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/umichbiostat/paper40

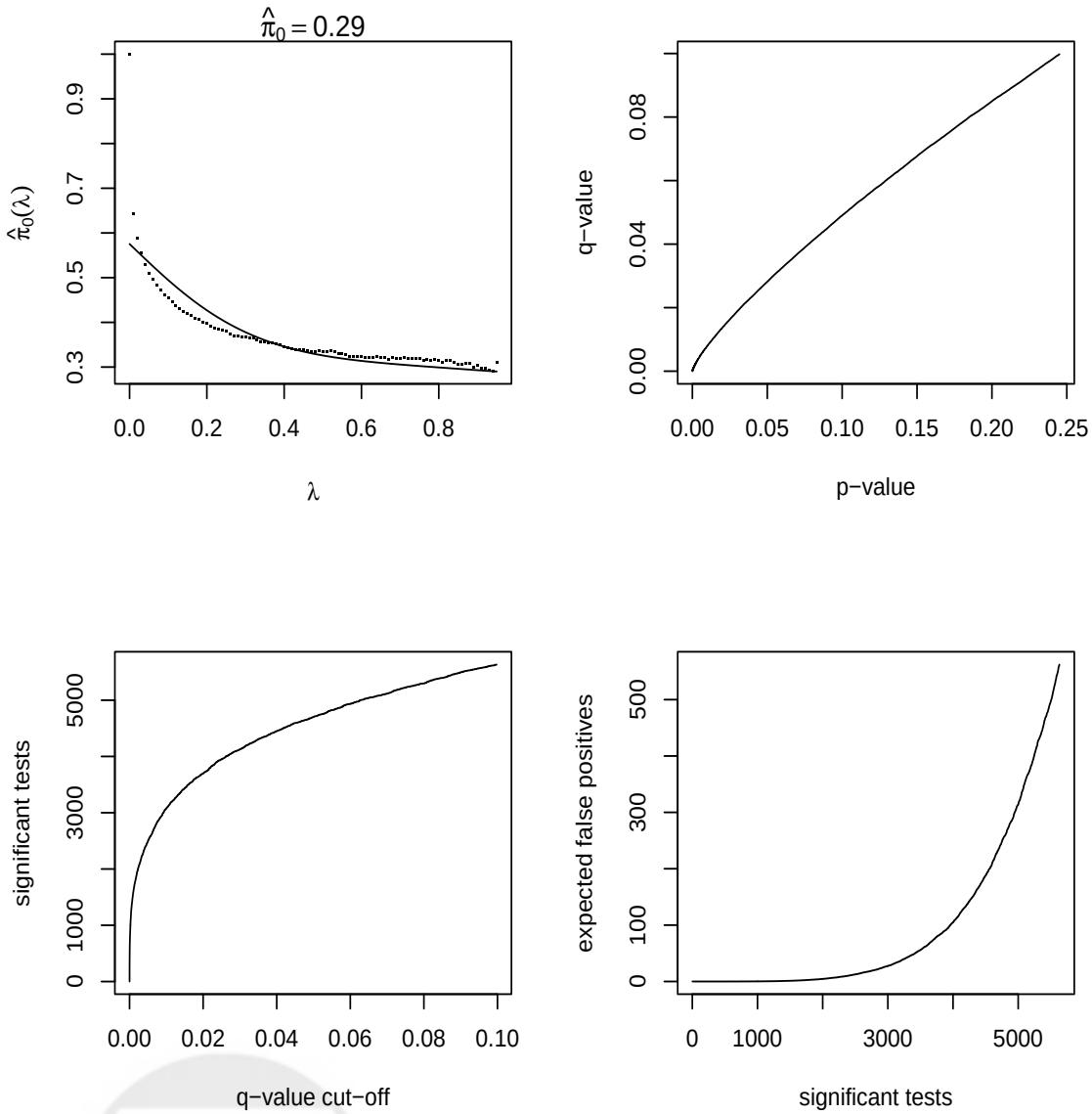

Figure 1: Output of proportional odds method combined with false discovery rate estimation procedures. The plot in the upper left-hand corner shows the estimated false discovery rate using the method of Storey and Tibshirani (2003). The upper right-hand plot shows the conversion of p-values to q-values as discussed in Section 2.2. The graph on the lower left-hand side shows the number of significant tests as a function of q-value cut-off. The lower right-hand graph displays the expected false positives as a function of number of significant tests; the estimated false discovery rate is the ratio of these quantities.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press