Statistical methods for identifying differentially Expressed genes in microarray data

Bindu Punathumparambath, Sebastian George, Kannan V. M.

Abstract— Microarray is a recently developed functional genomic technology that has powerful applications in a wide array of biological research areas, including the medical sciences, agriculture, biotechnology and environmental studies. One of the important problems in the analysis of microarray data is the identification of differentially expressed genes. Commonly used approaches for identifying differentially expressed genes are fold change, standard t-test, significance analysis of microarrays (SAM) and regularized t-test (Cyber-T). In the present study the generalized p-value method is used to test the differential expression of individual genes. We used the environmental microarray data set to test the proposed method and compared with existing methods considering univariate testing problem for each gene. Numerical results confirmed the superiority of the procedure based on the generalized p-value technique to identify genes with a low level of false discovery rate.

Index Terms— Environmental microarray data, Gene differential expression, Generalized p-value, Multiple hypothesis testing.

1 Introduction

DNA microarra have been used to monitor changes in gene expression during important biological processes and to study variations in gene expression across collections of related samples (for example, tumor samples from patients with cancer). These experiments compare two different samples of cDNA coloured with different dyes (red and green) to measure the intensity of fluorescence after hybridization. This method allows us to compare a large amount of data simultaneously in order to identify and quantify genes which are differentially expressed. Microarray experiments typically consist of intensity measurements of thousands of genes. The large volume of data generated from these experiments need efficient and proper statistical methods for deriving valid conclusions and has created tremendous opportunities for the statisticians to develop ppropriate statistical tools for analysing these datasets.

After normalization, gene expression distribution generally presents heavier tails than Gaussian distribution and have asymmetry of varying degrees with a sharp peak, due to the bulk of the mass at the middle. The gene expression distribution has been modelled using several densities**.** In [9], the gene expression distribution is fitted using the asymmetric Laplace distribution. In [4], the distribution of gene expression is modelled using asymmetric type II compound Laplace distribution. [3] introduced a family of skew-slash distributions generated by normal kernel, skew-slash distributions generate generated by Cauchy kernel ([7]), skew-slash t and skew-slash Cauchy distributions ([6]) and asymmetric slash Laplace distribution ([5]) for modelling microarray gene expression data.

An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. The classes may corresponds to tissues or cells. A straightforward approach to the identification of differentially expressed genes is to perform a univariate analysis of group mean differences for each gene, and then identify those genes that are most statistically significant. The large number of genes on a microarray, will lead to the identification of many genes that are not truly differentially expressed (false discoveries). Fold change is the simplest method with an arbitrary cut-off value used to determine differentially expressed genes. This method is unreliable as it does not take into account the statistical variability. In order to determine statistical significance, t–test can be performed for each gene. However, when many hypotheses are tested the probability of a type I error (false positive) occurring increases sharply with the number of hypotheses.

The problem of simultaneously testing a large number of hypotheses has generated a great amount of interest. [1] introduced the concept of False Discovery Rate (FDR). FDR is defined as the expected value of the ratio of the number of incorrectly rejected hypotheses to the total of number of rejected hypotheses. Assume a usual p–value is available for each hypothesis. Based on the p–values of the hypotheses, [1] provided a multiple testing procedure that guarantees the FDR to be less than or equal to a prefixed value q. In the present study the multiple hypotheses testing problem based on the generalized p–value is developed to identify the differentially expressed genes in the yeast Saccharomyces cerevisiae responding to diverse environmental transitions. Generalized p–value method for comparison of two exponential means has been applied to the raw dataset without log transformation to test the differential expression of individual genes. The multiple

- *Bindu Punathumparambath is currently working as Assistant Professor in Statistics, Vimala College, Thrissur, Kerla, India, PH-9947217775. E-mail: ppbindukannan@gmail.com*
- *Sebastian George is currently working as Associate Professor in Statistics, St. Thomas College, Pala, Kerala, India, PH-04822 202288. E-mail: sthottom@gmail.com*
- *Kannan V. M. is currently working as Associate Professor in Zoology, University of Calicut, Kerala, India, PH-9388157623, E-mail:kannanvm@yahoo.com*

hypotheses testing problem in microarray using the generalized p–value approach is discussed in section 2. The section 3 develop multiple hypothesis testing procedure using generalized p-value approach for microarray data analysis and discusses multiple hypotheses testing for two parameter exponential means using the generalized p–value approach. Ilustration of the generalized p-value approach for environmental microarray data is given in section 4. Our article ends with a brief concluding discussion.

2 MULTIPLE HYPOTHESIS TESTING IN MICROARRAY

The biological question of differential expression can be considered as a problem in multiple hypothesis testing in which m null hypotheses were simultaneously tested, where m (the number of genes whose expression levels were measured) can be considerably large. The large number of genes on a microarray, will lead to the identification of many genes that truly are not differentially expressed (false discoveries). In such situations, false discoveries (true null hypothesis declared significant) are inevitable. Thus, it is important in any multiple testing problem to control the error rate of false discoveries. Multiple testing procedures consist of choosing a vector of cutoffs for the test statistics such that a suitably defined false positive rate is controlled at an a priori specified level α. A standard approach to the multiple testing problem consists of two aspects:

1. Computing a test statistic T_j for each gene j
2. Applying multiple testing procedures to determine which hypotheses to reject while controlling a suitably defined Type I error rate .

Let $H_1, \ldots, H_m$ be m independent hypotheses to be tested. Let R_T be the number of true hypotheses that are incorrectly rejected, and let R_N be the number of not true hypotheses that are rejected. The total number of hypotheses rejected is $R = R_T + R_N$. Assume that the m_0 out of the m null hypothesis is true (m_0 genes are not differentially expressed) and $m_1 = m - m_0$, null hypothesis are false. A statistical test is performed independently, let p_i, i = 1, 2, . . . , m be the corresponding p–values. The decision to reject (or not) can be correct or false; when the null hypothesis is rejected for one of the m_0 variables for which it is actually true, is the false discovery (type I error).

The research in the area of multiple testing has generated a great deal of interest. Obtaining a multiple testing procedure when the hypotheses are not independent is an important problem. [2] propose very interesting and important results in this area. In the present study we use the generalized p–value approach for dealing the multiple hypothesis testing in environmental microarray studies.

2.1 Generalized p-value

The generalized p–value method is introduced in [10], has been used to successfully provide finite sample solutions for many hypothesis testing problems when no solutions are available using the usual approach. The generalized inference method was motivated by the fact that the small sample optimal confidence intervals (CIs) in statistical problems involving nuisance parameters may not be available. For example, for the difference between means of two exponential distributions, or two heteroscedastic normal distributions, classical small sample-inference does not provide optimal test and confidence intervals (see [11]). To overcome this problem, [10][11] introduced the concept of generalized confidence interval (GCI) and generalized p–value (GPV). These GPV and GCI can be considered as extension of the classical p–value and confidence interval. Interestingly, for some problems where the classical procedures are not optimal, GCI and GPV have performed well.

A general setup where the concepts of generalized confidence intervals and generalized p-values are defined is as follows.

Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from a distribution which depends on the parameters (θ, η), where θ is the parameter of interest and η is a vector of nuisance parameters. Consider the testing of $H_0 : \theta \le \theta_0$ vs. $H1 : \theta > \theta_0$, for a specified θ_0. A generalized test variable of the form $T(X; x, \theta, \eta)$, where x is an observed value of X, is chosen to satisfy the following three properties:

(1) The value of $T(X; x, \theta, \eta)$ at $X = x$ is free of any unknown parameters.

(2) For fixed x, the distribution of $T(X; x, \theta, \eta)$ is free of the vector of nuisance parameters η.

(3) For fixed x and η, and for all t, $P[T(X; x, \theta, \eta) > t]$ is either an increasing or a decreasing function of η.

The property (1) ensure that the sample space of possible values of subset of $T(X; x, \theta, \eta)$ can be found at a given value of the confidence coefficient with no knowledge of the parameters. This property is related to the notion of similarity in hypotheses testing ([11]). Property (2) guarantee that probability statement based on a generalized pivotal quantity will lead to confidence regions involving observed data x only. From property (3), generalized extreme region is defined as $G = \{X : T(X; x, \theta, \eta) \ge T(x; x, \theta, \eta)\}$ (or $G = \{X : T(X; x, \theta, \eta) \le T(x; x, \theta, \eta)\}$) if $T(X; x, \theta, \eta)$ is stochastically increasing (or decreasing) in θ. The generalized p–value is defined as $\sup_{H_0} P(G \mid H_0)$, where G is the extreme region defined above. See, [12] for further details.

The concept of False Discovery Rate in testing simultaneously several independent null hypotheses is introduced in [1]. They propose a procedure that is based on the p–value for testing each individual hypothesis. Their procedure guarantees that the FDR cannot exceed a prefixed rate q. In this section, we

extend the multiple testing problem based on the generalized p–value. Which can be used for the case when ordinary p–values are not available, one such example is simultaneously testing several independent Behrens–fisher problems, which appear to be useful in many applications.

The false discovery proportion Q is ratio of R_T and R ($R \neq 0$). The value of Q is zero for R = 0 and FDR is defined as the expected value of Q. The procedure for controlling the FDR in testing several independent null hypotheses based on generalized p-values is given as follows.

Let $H_1, \ldots, H_m$ are independent null hypotheses to be tested and $p_1, \ldots, p_m$ be the corresponding generalized p–values. For a given $q > 0$, suppose that there exists a cumulative distribution function Fi such that:

$$P(p_i \leq r|H_i) \leq F_i(r), r \leq r_0, \qquad (1)$$

for some r_0 satisfying $q \leq F_i(r_0)$, assuming that H_i is a true hypothesis. Let $p_i^* = F_i(p_i)$, for i = 1, . . . , m and $p_{(1)}^* \leq \ldots \leq p_{(m)}^*$ be the ordered values of p_i^*. Let $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypotheses. Define $q_i = iq/m$ and

$$k^* = \max\{i: p_{(i)}^* \leq q_i\}. \qquad (2)$$

Then, the procedure that rejects $H_{(i)}$ for $i \leq k^*$ guarantees that FDR ≤ q.

3 MICROARRAY ANALYSIS USING GENERALIZED P-VALUE

In [9] log ratio of red and green expession values are modeled using asymmetric Laplace distribution. Asymmetric Laplace distribution is the difference of two exponential distributions. Hence the problem of testing the differential expression in red and green expressions values reduces to the problem of testing the difference in means of two exponential distributions. We developed the generalized testing procedure for the comparison of two exponential means. Comparison of the means of transformed data in two samples can produce a different conclusion as opposed to comparing the means of the original data. The large sample test is too liberal where as the test based on the generalized p–values controls the type–I error quite satisfactorily. The procedure to be applied on untransformed data is summarized as follows:

Let X_{ijg} , i = 1, 2; j = 1, 2..., n_i and g = 1, 2..., n_g denote the random samples of gene expression data. Where i is the red and green intensities, j is the number of replications and g is the number of genes. Let X_{gij} 's follow two exponential distributions with parameters μ_i and σ_i, for i = 1, 2. Then the problem of testing the differential expression reduces to the testing of the equality means of two exponential means.

Let X follows a two parameter exponential family, then the probability density function (pdf) is given by

$$f(x, \mu, \sigma) = \frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}}, X > \mu, \mu \geq 0, \sigma > 0, \qquad (3)$$

where μ is the location parameter and σ is the scale parameter. Let $X_1, \ldots, X_n$ be a sample of observations from an exponential distribution with pdf in (3). The maximum likelihood estimators of μ and σ are given by

$$\hat{\mu} = X_{(1)}, \qquad \hat{\sigma} = \bar{X} - X_{(1)} \qquad (4)$$

where $X_{(1)}$, is the smallest of the X and $\hat{\mu}$ and $\hat{\sigma}$ are independent. Then

$$\frac{2n(\hat{\mu} - \mu)}{\sigma} \sim \chi^2_2 \ and \ \frac{2n\hat{\sigma}}{\sigma} \sim \chi^2_{2n-2}. \qquad (5)$$

Let $\widehat{\mu_0}$ and $\hat{\sigma}_0$ be the observed values of $\hat{\mu}$ and $\hat{\sigma}$ respectively. Then the generalize pivotal variable for μ is given by

$$T_\mu = \widehat{\mu_0} - \frac{2n(\hat{\mu}-\mu)}{2n\sigma}\frac{\sigma}{\hat{\sigma}}\widehat{\sigma_0} = \widehat{\mu_0} - \frac{\chi^2_2}{\chi^2_{2n-2}}\widehat{\sigma_0}. \qquad (6)$$

Let X be an exponential random variable with pdf f(x; μ_1, σ_1) and Y be an exponential random variable with pdf f(y; μ_2, σ_2), with pdf given in (3). Let $X_1, \ldots, X_{n1}$ be a sample of observations from X and $Y_1, \ldots, Y_{n2}$ be a sample of observations from Y . From equation (4), the maximum likelihood estimators of μ_1, μ_2, σ_1 and σ_2 are given by

$$\widehat{\mu_1} = X_{(1)}, \widehat{\mu_2} = X_{(2)}, \widehat{\sigma_1} = \bar{X} - X_{(1)} \ and \ \widehat{\sigma_2} = \bar{X} - X_{(2)}.$$

The the generalized pivotal variable for the difference of means of two exponential distributions is

$$T^* = T_{\mu_1} - T_{\mu_2}. \qquad (7)$$

Where $T_{\mu_i} = \widehat{\mu_{i0}} - \frac{2n(\hat{\mu_i}-\mu_i)}{2n\sigma}\frac{\sigma_i}{\widehat{\sigma_i}}\widehat{\sigma_{i0}}$, i=1, 2. From (6), $T_{\mu_i} = \widehat{\mu_{i0}} - \frac{\chi^2_2}{\chi^2_{2n-2}}\widehat{\sigma_{i0}}$. Where $\widehat{\mu_{i0}}$ and $\widehat{\sigma_{i0}}$ be the observed values of $\hat{\mu_i}$ and $\hat{\sigma_i}$ respectively. Consider the problem of testing,

$$H_0: \mu_1 = \mu_2 \ vs \ H_1: \mu_1 \neq \mu_2. \qquad (8)$$

From equation (7) the generalized test variable for testing $H_0: \mu_1 = \mu_2$ is

$$T_g = T^* - (\mu_1 - \mu_2). \qquad (9)$$

The hypothesis will be rejected for a small p-value for comparison of the either of the two alternative hypothesis. Hence the generalized p-value for the two sided test can be obtained as:

$$2\min\{P(T_g \le 0), P(T_g \ge 0)\}. \qquad (10)$$

3.1 Estimation of FDR

Let N be the total number of genes, G be the total number of significant genes and p_i is the generalized p-value of the i^{th} most significant gene as estimated from generalized p-value technique. Then the FDR can be estimated using the following formula.

$$FDR_G = Min_{i \ge G} \left\{\frac{N\, p_i}{i}\right\}. \qquad (11)$$

4 APPLICATIONS

In this section we illustrate the application of the multiple hypothesis testing problem in microarray gene expression studies based on the concept of FDR and generalized p-values. For this purpose we used the yeast Saccharomyces cerevisiae microarray dataset [8]. We have downloaded the microarray data set from http://www-genome.stanford.edu/yeast stress. In the present work we explored the data on genomic expression patterns in the yeast Saccharomyces cerevisiae responding to heat shock out of the different environmental parameters on the expression levels of 6200 genes studied in [8]. The yeast cells subjected to a larger shift in temperature responded with larger and more prolonged alterations in gene expression before adapting to their new steady-state expression levels, relative to cells exposed to smaller temperature changes. One group consisted of genes whose transcript levels increased in abundance in response to the environmental changes, and the other group was comprised of genes whose transcript levels decreased following environmental stress. The genes whose transcript levels increase in response to environmental change will be referred to as induced, while genes whose transcript levels decrease will be referred to as repressed. A large set of genes (~ 900) showed a similar drastic response to these environmental changes.

We have performed testing for differentially expressed genes using the generalized method for the downloaded datasets from [8]. After getting the raw p–values for individual gene by using either the t–test (unequal variance) or by the generalized p–value method, p–values were adjusted for the multiplicity problem by fixing different q values using the equation (2). The estimated false discovery rate were computed using the equation (11) and results are given in Table 1. For computation we used the R package.

TABLE 1
THE ESTIMATED FALSE DISCOVERY RATES BASED ON T–TEST AND GENERALIZED P-VALUE TECHNIQUE

No. of genes selected	Estimated False Discovery Rate			
	Induced		Repressed	
	t-test	generalized p value	t-test	generalized p value
200	0.009	0.001	0.012	0.004
250	0.011	0.004	0.023	0.009
300	0.012	0.006	0.035	0.014
350	0.015	0.008	0.045	0.024
400	0.020	0.011	0.056	0.032
450	0.023	0.012	0.067	0.040
500	0.026	0.015	0.078	0.046
550	0.031	0.020	0.089	0.058
600	0.038	0.024	0.102	0.071
700	0.049	0.029	0.125	0.095
800	0.058	0.035	0.143	0.109
900	0.075	0.043	0.161	0.123

5 CONCLUTIONS

In the present study we developed multiple testing procedure that can control the FDR using generalized p-value. From the Table 1 we can see that the generalized p–value method could find more number of genes at a particular level of q from the repressed dataset as well as in induced dataset. By reducing the level of q we can go for stringent selection of genes but will increase the chance of not detecting truly expressed genes. Thus there is a trade off between the level of false discovery rate and the false negative rate. In general the generalized p value method could identify more number of truly expressed genes with a fewer chance of accepting false discoveries. We computed expected false discovery rate by selecting 200 to 900 genes from the control dataset and test dataset based on different temperature. From Table 1, it can be noted that by selecting top 500 genes from the induced based on t–test (unequal variance), the estimated false discovery rate was 0.026, whereas it was only 0.015 while genes selection is carried out using the generalized p–value technique. Also out of 500 repressed genes, based on t– test (unequal variance), the estimated false discovery rate was 0.078, whereas by generalized p–value technique the estimated false discovery rate was 0.046. Hence for the generalized p–value approach the expected false discoveries were only 27 and 473 are expected to be true positives.

ACKNOWLEDGMENTS

The first author thank the Department of Science and Technology, Government of India, New Delhi, for financial assistance underWomen Scientist Scheme (WOS-A) Project No:SR/WOSA/MS-09/2008. Also we express sincere gratitude to Dr. Samaradasa Weerahandi, Pfizer, New York, NY 10017, for valuable suggestions and motivation.

REFERENCES

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to

multiple testing". *Journal of the Royal Statistical Society B*, vol. 57, 289-300, 1995.

[2] Y. Benjamini, and D. Yekutieli, "The Control of the False Dis covery Rate in Multiple Testing under Dependency, " Annals of Statis tics, vol. 29, pp. 1165-1188, 2001.

[3] P. Bindu, "A new family of skewed slash distribution generated by the normal kernel," *STATISTICA*, vol. 71, no. 3, pp. 345- 353, 2011.

[4] P. Bindu, K. Sangita and G. Sebastian, "Asymmetric type II compound Laplace distribution and its application to microarray gene expression." *Computational Statistics and Data analysis*, vol. 56, pp. 1396-1404, 2012a.

[5] P. Bindu, "The multivariate asymmetric slash Laplace distribution and its applications," *STATISTICA*, vol. 72, no. 2, pp. 235-249, 2012c.

[6] P. Bindu, "The multivariate skew-slash t and skew-slash Cauchy Distributions," Model Assisted Statistics and Applications, 7, 33-40, 2012b.

[7] P. Bindu, "A new family of skewed slash distribution generated by the Cauchy kernel." Communications in Statistics- Theory and Me thods, vol. 42, pp. 2351-2361, 2013.

[8] A. P. Gasch, T. S. Paul, M. K. Camilla, Orna Carmel-Harel, B. E. Mi chael, S. Gisela, S., David B. and P. O. Brown, "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Molecular Biology of the Cell*, vol. 11, pp. 4241-4257, 2000.

[9] E. Purdom and S. Holmes, "Error Distribution for Gene Ex pression Da ta," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, Article 16, 2005.

[10] K.W. Tsui, S. Weerahandi, " Generalized P-values in signific ance testing of hypotheses in the presence of nuisance parameters," Journal of the American Statististical Association, 84, pp. 602-607, 1989.

[11] S. Weerahandi, "Generalized confidence intervals," Journal of the American Statististical Association, vol. 88, pp.899-905, 1993.

[12] S. Weerahandi, "Exact Statistical Methods for Data Analysis." Springer-Verlag, 1995.