Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Order-restricted inference for ordered gene expression (ORIOGEN) data under heteroscedastic variances

Susan J. Simmons* **and Shyamal D. Peddada**

Department of Mathematics and Statistics, University of North Carolina Wilmington, Wilmington, NC 28403; Biostatistics Branch, NIEHS (NIH), RTP, NC - 27709; Susan J. Simmons* - Email: simmonssj@uncw.edu;

* Corresponding author

received August 08, 2006; accepted January 15, 2007; published online April 10, 2007

Abstract:

This article extends the order restricted inference approach for time-course or dose-response gene expression microarray data, introduced by Peddada and colleagues (2003) for the case when gene expression is heteroscedastic over time or dose. The new methodology uses an iterative algorithm to estimate mean expression at various times/doses when mean expression is subject to pre-defined patterns or profiles, known as order-restrictions. Simulation studies reveal that the resulting bootstrap-based methodology for gene selection maintains the false positive rate at the nominal level while competing well with ORIOGEN in terms of power. The proposed methodology is illustrated using a breast cancer cell-line data analyzed by Peddada and colleagues (2003).

Keywords: ordered gene expression; heteroscedastic variances; restricted inference; iterative algorithm

Background:

Increasingly, researchers are interested in understanding changes in gene expression when an animal/tissue/cell line is exposed to a chemical/treatment over time and/or dose. For instance, researchers in the U.S. National Toxicology Program are conducting numerous gene expression studies to evaluate toxicity of a variety of chemicals on various tissues/organs in rodents using dose-response studies. There are a variety of reasons for conducting a dose-response/time-course gene expression study. Sometimes a researcher may be interested in understanding the changes in gene expression at a specific time/dose relative to the control. In other situations, a researcher may be interested in understanding the time-course pattern (or profile) of gene expression. Accordingly, statistical methodology for the analysis of time-course/dose-response gene expression data has been an area of active research in recent years. Although the methodology described in this paper is equally applicable to both time- course and dose-response studies, for simplicity of exposition we shall only discuss time-course studies. However, the same methodology may be applied to dose-response studies. Further, this work is motivated by experiments where independent samples are obtained at different time points, unlike repeated measures or longitudinal studies.

Depending upon the application, one may use a variety of available statistical methods for analysis. For example, if the objective is to identify genes that have significantly different expression values between two specific doses or a dose and control, then one may use statistical procedures such as SAM (Statistical Analysis of Microarrays) **[2, 19]**, BAMarray (Bayesian Analysis for Microarrays) **[5, 6, 7]**, GA/KNN (Genetic Algorithm with K nearest neighbors) **[8, 9]**, etc. However, if the objective is to select significant genes on the basis of their pattern/profile of expression over time, then one may use procedures such as Linear/Quadratic regression based method of Liu et al., **[10]**, EDGE **[18]**, ORIOGEN (Order Restricted Inference for Ordered Gene ExpressioN) **[13, 14]** etc. Each of these procedures identifies significant genes on the basis of their pattern of expression over time. The Linear/Quadratic regression based method of **[10]** is a very quick and simple methodology that fits standard linear and quadratic regression models for each gene over time. Based on the statistical significance of various regression coefficients, genes are clustered into groups. The EDGE methodology of **[18]** may be viewed as a nonparametric version of. **[10]** EDGE exploits the smoothing spline models of **[1]** to fit gene expression over time. The machinery developed in **[1]** is specifically designed for repeated measurements on individuals. Consequently, the EDGE methodology can be used for analyzing gene expression data under repeated measures setting as well. Unlike regression procedures of **[10]** and EDGE, ORIOGEN is entirely nonparametric in the sense that no functional form and no distributional assumptions are made for gene expression over time (http://dir.niehs.nih.gov/dirbb/oriogen/index.cfm). Instead the procedure represents the mean expression over time by mathematical inequalities, known as *order restrictions*, and the P-values are determined by bootstrap methodology. Thus, the null hypothesis in ORIOGEN is that the mean gene expression is the same across all times and the alternative hypothesis is a union of all potential patterns declared of interest by the researcher. The software allows the researcher to provide a list of gene expression patterns of interest by clicking on radio buttons. The output not only selects statistically significant genes, but it also clusters genes with similar time-course profile. If a gene ontology database is available, then ORIOGEN can link the significant genes to the gene ontology database and provide further description on each selected gene. It has been demonstrated in **[13]** that ORIOGEN maintains the nominal

Type I error rates when the variances are homoscedastic. Throughout this paper the terms "Type I error" and "power" refer to the standard false positive and true positive rates for a given test. They are not adjusted for multiple testing. Recently, several authors (c.f. **[3, 15]**) have discussed methods for analyzing gene expression data that control for false discovery rates. An important development in this field is the work of Datta and Datta. **[3]** They develop an empirical Bayes methodology for screening P-values so that the overall sensitivity of multiple testing is increased with a modest increase in false discovery rates.

Most procedures described above, are based on the assumption that for each gene, the expression values are homoscedastic (i.e., have equal variance) across times. In practice this assumption may not be true. Heteroscedasticity (i.e., unequal variances over time) may arise for a number of reasons. For instance, variability in gene expression could depend upon the mean expression value, or dose and/or duration of exposure. A potential consequence of heteroscedasticity is an increased false positive (and false discovery) rate and decreased power. Hence it is important to adjust for heteroscedasticity while analyzing gene expression data.

In section 2 we provide a step by step description of the new methodology for selecting statistically significant genes and clustering genes with similar time-course profiles. As in **[12, 13, 14]**, all profiles are described by mathematical inequalities between the unknown parameters. We also compare the performance of the new procedure with ORIOGEN in terms of Type I error and power using a small simulation study. In section 3 we illustrate the proposed methodology using a data set described in Lobenhofer et al., **[11]** which was previously analyzed in. **[13]** Concluding remarks are provided in section 5 and in the Appendix we sketch the details of the proposed estimation and testing procedures.

Throughout this paper we use the terms "profiles", "patterns" and "order-restrictions" synonymously. Similarly, we use the terms "dose-response" and "time-course" interchangeably.

Methodology:

For a given gene *g*, *g = 1, 2, …, G*, let y_{gij} denote its expression in the j^{th} sample, $j = 1, 2, \ldots, n_{gi}$, at the i^{th} time period, $i = 1, 2, \ldots, T$, with $E(y_{gij}) = \mu_{gi}$, $Var(y_{gij}) = \sigma^2_{gi}$. In the following steps we describe the proposed methodology for evaluating the statistical significance of gene *g* and clustering genes with similar expression profiles over time. Throughout this paper $\boldsymbol{\mu}$ denotes the mean expression vector of suitable order. The order of $\boldsymbol{\mu}$ would be clear from the context.

Algorithm 1:

Step 1 (Profiles specification)

As in **[13, 14]**, the researcher pre-selects all the time-course profiles of interest in the study in terms of *k* sets of inequalities between the mean expressions. Thus, the desired parameter space of interest is $\Theta = \bigcup_{p=1}^{k} \Theta_p \subseteq R^T$ where each subset Θ_p represents a time course profile of interest.

Two common profiles of interest are: (1) *Increasing* profile (*simple order* restriction), where $\Theta_1 = \{\boldsymbol{\mu} \in R^T \mid \mu_1 \le \mu_2 \le .... \le \mu_T\}$, with at least one strict inequality. (2) *Up-down* profile (*umbrella order* restriction), where $\Theta_2 = \{\boldsymbol{\mu} \in R^T \mid \mu_1 \le \mu_2 \le .... \le \mu_s \ge \mu_{s+1} \ge ... \ge \mu_T\}$, with at least one strict inequality.

Similar to ORIOGEN, the proposed methodology tests the null hypothesis of no difference in mean expression over time, i.e $H_0 : \mu_{g1} = \mu_{g2} = ... = \mu_{gT}$, against the alternative that the mean expression has one of the forms described by $\Theta = \bigcup_{p=1}^{k} \Theta_p \subseteq R^T$.

Step 2 (Profile fitting)

For each gene *g* we fit the observed expression values y_{gij} against each profile Θ_p in the alternative hypothesis using the estimation procedure described in Appendix A1. For each fitted profile Θ_p, we compute a goodness-of-fit statistic as described in Appendix A1 and select the profile with the largest goodness-of-fit statistic.

Step 3 (Bootstrap significance)

We evaluate the statistical significance of the largest goodness-of-fit statistic obtained in Step 2 using the bootstrap methodology. Since the data are heteroscedastic, the bootstrap methodology used in **[13]** is not appropriate; instead we use the bootstrap procedure described in Appendix A2. To keep the false positive and false discovery rates small, we advise the user to test the significance of each gene at a very small level of significance. Further, since the level of significance is small, we run a large number of bootstraps.

Genes with a P-value less than the pre-selected level of significance are selected as the significant genes. All significant genes with the same selected profile are clustered together.

We compared the performance of the above methodology with ORIOGEN using a small simulation study. The goal

is to compare the two procedures in terms of Type I error rate and the power. In our simulation study we considered G = 1000 genes, *T* = 6 time points with 10 independent normally distributed random samples per time point. For each gene *g* and time *i* the mean and variance patterns considered are as follows: A. Null hypothesis $(\mu_{gi} = 0)$ with various patterns of variances:

(1) Homoscedastic: $\sigma_{gi}^2 = 16,$

(2) Heteroscedastic: $\sigma_{gi}^2 = i^2,$ (3) Strongly Heteroscedastic: $\sigma_{gi}^2 = i^3$. B. Ordered alternative hypothesis with various patterns of variances:

(4) Homoscedastic: $\mu_{gi} = i, \sigma_{gi}^2 = 16,$

(5) Heteroscedastic: $\mu_{gi} = i, \sigma_{gi}^2 = i^2,$ (6) Strongly Heteroscedastic: $\mu_{gi} = i,\ \sigma_{gi}^2 = i^3,$ (7) Umbrella profile: $\mu_{gi} = 0, \sigma_{gi}^2 = 16$ for *i* = 1, 2, 4, 5, 6 and $\mu_{g3} = 3, \sigma_{g3}^2 = 9.$

The funnel shaped heteroscedastic patterns considered above can be viewed as an "extreme" pattern in the sense that we expect this variance pattern to have greater impact on the false positive rate of test procedures based on homoscedastic variances than if the variance pattern has, for instance, an umbrella-shaped order restriction. We recognize that this is a small simulation study, but it conveys the drawbacks of procedures which do not account for heteroscedasticity and demonstrates that the modification proposed in this paper performs well. It is also important to note that the amount of variation in the data considered in patterns (6) and (7) are very extreme compared to the differences among the means and hence in this case neither of the methods is expected to have good power.

The results of our simulation study, based on 1000 bootstrap samples at a level of significance of 0.05, are reported in Table 1. Patterns (1), (2) and (3) provide the Type I errors of the two procedures, whereas patterns (4), (5), (6) and (7) provide the power of the procedure. As seen from Table 1, the new procedure (denoted as ORIOGEN-*Hetero*) never exceeds the nominal level of 0.05, whereas ORIOGEN can be very liberal (larger Type I error than the nominal levels) as the amount of heteroscedasticity increases. For instance, in the case of patterns (2) and (3) the Type I error of ORIOGEN-*Hetero* is at most 0.03, whereas the ORIOGEN had a Type I error as high as 0.12. Not only does the new procedure have a Type I error rate within the nominal level of 0.05, it actually performs very well in terms of power when compared to ORIOGEN as seen in patterns (5), (6) and (7). Further, in the case of homoscedastic variances, pattern (4), the proposed procedure competes very well with ORIOGEN in terms of power.

	Method	
Pattern $(i = 1, 2, ..., 6)$	**ORIOGEN**	**ORIOGEN -Hetero**
1. $\mu_{gi} = 0,\ \sigma_{gi}^2 = 16$	0.04	0.04
2. $\mu_{gi} = 0,\ \sigma_{gi}^2 = i^2$	0.12	0.05
3. $\mu_{gi} = 0,\ \sigma_{gi}^2 = i^3$	0.11	0.03
4. $\mu_{gi} = i,\ \sigma_{gi}^2 = 16$	0.66	0.64
5. $\mu_{gi} = i,\ \sigma_{gi}^2 = i^2$	0.72	0.77
6. $\mu_{gi} = i,\ \sigma_{gi}^2 = i^3$	0.24	0.25
7. $\mu_{gi} = 0,\ \sigma_{gi}^2 = 16, i =$ $\mu_{g3} = 3,\ \sigma_{g3}^2 = 9$	0.32	0.39

Table 1: Power and Type-I Error rate comparisons between ORIOGEN and ORIOGEN-*Hetero*

Illustration:

Lobenhofer et al., **[11]** conducted a microarray experiment to evaluate the effects of 17-β estrodial on the gene expression of MCF-7 breast cancer cells. Microarrays were obtained after 1, 4, 12, 24, 36 and 48 hours of treatment. There were 8 cDNA chips per time point, and each chip had 1900 probes. As done in **[13]**, the gene expressions are log transformed. For each gene the null hypothesis was that the mean expression did not change over the 6 time points and the alternative was the union of 10 hypotheses as

follows: (1) mean expression is non-decreasing with time, (2) mean expression is non-increasing with time, (3,4,5,6) mean expression has an umbrella shape with peaks 4, 12, 24, and 36 hours and (7,8,9,10) mean expression has an inverted umbrella with troughs at 4, 12, 24, and 36 hours. Before implementing the new procedure, we applied Hartley's test for heteroscedasticity of variances. The P-values for the Hartley's test statistic was computed by bootstrapping the residuals since the null distribution of the Hartley's test is sensitive to normality assumption and gene expression data are not necessarily normally distributed. Using the usual level of significance of 0.05, we found that 367 genes out of 1900 were heteroscedastic. At 0.10 level of significance, this number jumps up to 610 genes. Thus there appears to be some amount of heteroscedasticity in the data which motivates us to apply the new methodology on this data.

According to ORIOGEN, which assumes homoscedasticity of variances, 197 out of 1900 genes were statistically significant at a level of significance α= 0.005. When we re-analyzed the data using the new methodology ORIOGEN-*Hetero*, we found 140 out of 1900 genes were significant at a level of significance α= 0.005. Of these 140, 115 were also selected by ORIOGEN. These common genes are listed in the attached spreadsheet. Thus 82 genes were selected only by ORIOGEN while 35 were selected only by ORIOGEN-*Hetero*. The discrepancy between these two procedures is possibly due to the amount of heteroscedasticity present in the data.

Conclusions and Discussion:

In this article we extended the order restricted inference procedure ORIOGEN of **[13, 14]** for the case when the gene expressions may be subject to unequal variance across time. The new methodology, ORIOGEN-*Hetero*, uses an iterative algorithm to estimate the mean expression values subject to a given profile and statistical inferences are conducted by suitably bootstrapping the residuals. ORIOGEN and ORIOGEN-*Hetero* differ in both the method of estimation of parameters subject to order restrictions as well as the bootstrap methodology used in determining the P-values. While ORIOGEN directly uses the point estimators developed in **[4]** under the assumption of equal variance across time for a given gene, ORIOGEN-*Hetero* uses an iterated version of **[4]** where the unknown variances are estimated along with the means subject to order restrictions. Further, by bootstrapping the residuals, ORIOGEN-*Hetero* allows heteroscedasticity, whereas in ORIOGEN resampling was performed by mixing samples from all time points for a given gene.

A simulation study reported in this paper reveals that the new methodology performs well in controlling the Type I errors and hence is expected to perform well in controlling the overall false discovery rates when the gene expression data are subject to unequal variances across time. Further, our modest simulation study suggests that the new method improves the power of the test as well when the variances are heteroscedastic. However, as seen in our simulation study, when the variances are homoscedastic, the new method may lose power relative to ORIOGEN. One way to get around this problem is to perform a test procedure such as Hartley's test for homoscedasticity of variances. Since Hartley's test is not robust against non-normality and gene expression data are not necessarily normally distributed, P-values for the Hartley's test may be determined by bootstrapping appropriate residuals. If the null hypothesis of homoscedasticity of variances is not rejected at some pre-specified level of significance of α, then one may implement ORIOGEN for such genes. For genes where the null hypothesis of homoscedasticity of variances is rejected by Hartley's test, then in such cases one may use the new method proposed in this paper. Such a pre-testing strategy might increase the power while protecting the Type I error and false discovery rates.

The resampling procedure used in ORIOGEN and ORIOGEN-*Hetero* does not allow for dependence in the samples across time as typically observed in a repeated measure study design. Estimation and testing for order restrictions under repeated measures design is a nontrivial generalization of the method described here. In an ongoing project we are generalizing ORIOGEN to allow for repeated measures data.

Acknowledgement:

The authors thank David Umbach, Grace Kissling, the reviewer and the editor for their careful reading of this manuscript and for numerous suggestions which improved the presentation of the manuscript substantially. The second author's research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Appendix:

Throughout the Appendix we shall use the notations introduced in the main text.

A1. Estimation of means subject to order restrictions under heteroscedasticity

As in **[12]**, we express each order restriction of interest on a parameter vector $\mu = (\mu_1, \mu_2, ..., \mu_T)'$, by a graph where two components μ_i and $\mu_j, i \neq j,$ are said to be linked if the inequality between the two is specified by the order restriction. For instance in the case of a simple order restriction all parameters are linked, whereas in the case of umbrella order

$\Theta_2 = \{\boldsymbol{\mu} \in R^T \mid \mu_1 \leq \mu_2 \leq .... \leq \mu_s \geq \mu_{s+1} \geq ... \geq \mu_T\}$

, with at least one strict inequality, all parameters $\mu_i, \mu_j, \forall\, 1 \leq i < j \leq s,$ are linked and $\mu_i, \mu_j, \forall\, s \leq i < j \leq T,$ are linked but $\mu_i, \mu_j, \forall\, 1 \leq i < s, s < j \leq T$ are not linked. A subgraph M formed by the subvector of $\boldsymbol{\mu}$ is said to be a

linked subgraph if all parameters in the subvector are linked. Thus in the case of simple order Θ_1 every subvector of $\boldsymbol{\mu}$ is a linked subgraph whereas in the case of umbrella order Θ_2, the subgraph formed by $\boldsymbol{\mu}$ is not a linked subgraph since μ_1, μ_T are not linked. A linked subgraph M is said to be a *maximally linked subgraph* if for any linked subgraph N, if M $\subseteq$ N then M = N. The two extreme parameters of a maximally linked subgraph are said to be the *farthest linked parameters* of the maximally linked subgraph. In the case of umbrella order Θ_2, the two maximally linked subgraphs are:

$\Theta_{21} = \{\boldsymbol{\mu} \in R^s \mid \mu_1 \le \mu_2 \le .... \le \mu_s\}$ and

$\Theta_{22} = \{\boldsymbol{\mu} \in R^{T-s+1} \mid \mu_T \le \mu_{T-1} \le .... \le \mu_s\}$.

In Θ_{21} the two farthest linked parameters are μ_1 and μ_s and in Θ_{22} they are μ_s and μ_T.

Hwang and Peddada **[4]** introduced a general methodology for estimating the mean vector $\boldsymbol{\mu}_g$ of gene *g*, for any arbitrary set of linear inequalities between the components of $\boldsymbol{\mu}_g$ when the corresponding population variances $\sigma_{gi}^2, i = 1,2,...,T,$ are known. Motivated by **[17]**, for a given profile Θ_p, in the following we propose a simple iterative scheme to estimate the mean vector $\boldsymbol{\mu}_g \in \Theta_p$ when the population variances $\sigma_{gi}^2, i = 1,2,...,T,$ are unknown. The basic idea is to invoke methodology in **[4]** at each iteration by using the estimates of the variances from the previous iteration as weights.

Algorithm (Estimation of parameters for gene g under profile Θ_p)

Step 1 (initial estimates): Let

$$\hat{\mu}_{gi}^{(0)} = \bar{y}_{gi} = \frac{\sum_{j=1}^{n_{gi}} y_{gij}}{n_{gi}},$$

$$\hat{\sigma}_{gi}^{2(0)} = s_{gi}^2 = \frac{\sum_{j=1}^{n_{gi}} (y_{gij} - \bar{y}_{gi})^2}{n_{gi} - 1}, \quad w_{gi}^{(0)} = \frac{n_{gi}}{\hat{\sigma}_{gi}^{2(0)}}.$$

Step 2 (r^{th} iterate, $r = 1, 2, ...$): Apply the methodology in **[4]** on the estimates, $\hat{\mu}_g^{(r-1)} = (\hat{\mu}_{g1}^{(r-1)}, \hat{\mu}_{g2}^{(r-1)}, ..., \hat{\mu}_{gT}^{(r-1)})'$, with weights $w_g^{(r-1)} = (w_{g1}^{(r-1)}, w_{g2}^{(r-1)}, ..., w_{gT}^{(r-1)})'$, obtained in the $(r\text{-}1)^{th}$ iterate. Denote the resulting estimates by

$$\hat{\mu}_g^{(r)} = (\hat{\mu}_{g1}^{(r)}, \hat{\mu}_{g2}^{(r)}, ..., \hat{\mu}_{gT}^{(r)})',$$

$$\hat{\sigma}_{gi}^{2(r)} = \frac{\sum_{j=1}^{n_{gi}} (y_{gij} - \hat{\mu}_{gi}^{(r)})^2}{n_{gi} - 1}, \quad i = 1, 2, ..., T$$

and

$$w_g^{(r)} = (w_{g1}^{(r)}, w_{g2}^{(r)}, ..., w_{gT}^{(r)})' = \left(\frac{n_{g1}}{\hat{\sigma}_{g1}^{2(r)}}, \frac{n_{g2}}{\hat{\sigma}_{g2}^{2(r)}}, ..., \frac{n_{gT}}{\hat{\sigma}_{gT}^{2(r)}}\right)'.$$ If any of the denominators is zero, as in **[16]** we replace it by an arbitrarily small positive real number.

Step 3 (Convergence): Repeat Step 2 until $\| \hat{\mu}_g^{(r+1)} - \hat{\mu}_g^{(r)} \|_2^2 < \delta,$ where $\| . \|_2^2$ is the square of the usual L_2 norm and δ is some small positive constant. In the simulations contained herein, δ is chosen to be 0.0001. Upon convergence, the estimates are denoted by $\hat{\boldsymbol{\mu}}_g = (\hat{\mu}_{g1}, \hat{\mu}_{g2}, ..., \hat{\mu}_{gT})'$

and $\hat{\sigma}_{gi}^2 = \frac{\sum_{j=1}^{n_{gi}} (y_{gij} - \hat{\mu}_{gi})^2}{n_{gi} - 1}$.

For normally distributed data, with Θ_p satisfying the simple order restriction, Shi and Jiang **[17]** discussed the convergence of the above algorithm. Although in this paper we do not discuss the convergence of the above algorithm for the general linear inequality restrictions, extensive simulation studies, using an umbrella order restriction, suggest that the above algorithm converges rapidly. On average it took less than 10 iterations in the simulations we performed.

Step 4 (Computation of goodness-of-fit statistic): For profile Θ_p, identify all maximally linked subgraphs. Within each maximally linked subgraph identify the farthest linked parameters. Then the goodness-of-fit statistic for Θ_p is defined as the maximum studentized difference between the estimates of two farthest linked parameters, where maximum is taken over all maximally linked subgraphs. Denote the statistic by $l_{p(g)}^{\infty}$. This can be viewed as the standard infinity norm of a vector.

Examples: For simple order $\Theta_1 = \{\boldsymbol{\mu} \in R^T \mid \mu_1 \leq \mu_2 \leq .... \leq \mu_T\}$ the only maximally linked subgraph is Θ_1 itself and the two farthest linked parameters are μ_1 and μ_T. Thus here

$$l^{\infty}_{1(g)} = \frac{\hat{\mu}_{gT} - \hat{\mu}_{g1}}{\sqrt{\frac{\hat{\sigma}^2_{gT}}{n_{gT}} + \frac{\hat{\sigma}^2_{g1}}{n_{g1}}}}.$$

However, for the umbrella order, $\Theta_2 = \{\boldsymbol{\mu} \in R^T \mid \mu_1 \leq \mu_2 \leq .... \leq \mu_s \geq \mu_{s+1} \geq ... \geq \mu_T\}$, the two maximally linked subgraphs are Θ_{21} and Θ_{22} with farthest linked parameters μ_1, μ_s and μ_s, μ_T, respectively. Thus in this case

$$l^{\infty}_{2(g)} = \max\left(\frac{\hat{\mu}_{gs} - \hat{\mu}_{g1}}{\sqrt{\frac{\hat{\sigma}^2_{gs}}{n_{gs}} + \frac{\hat{\sigma}^2_{g1}}{n_{g1}}}}, \frac{\hat{\mu}_{gs} - \hat{\mu}_{gT}}{\sqrt{\frac{\hat{\sigma}^2_{gs}}{n_{gs}} + \frac{\hat{\sigma}^2_{gT}}{n_{gT}}}} \right).$$

A2. Bootstrap procedure for testing under heteroscedasticity

For each gene *g* and time point *i* obtain the residuals $e_{gij} = y_{gij} - \bar{y}_{gi}, j = 1, 2, \ldots, n_{gi}, \; i = 1, 2, \ldots, T.$ Next within the i^{th} time point draw a simple random sample of size n_{gi} (with replacement) from $\{e_{gi1}, e_{gi2}, ..., e_{gin_{gi}}\}$. Denote the resampled residuals by $\{e^*_{gi1}, e^*_{gi2}, ..., e^*_{gin_{gi}}\}$. Then the bootstrap data y^*_{gij} are obtained by $y^*_{gij} = \bar{y}_g + e^*_{gij}$, where $\bar{y}_g = \frac{\sum_{i=1}^{T} n_{gi} \bar{y}_{gi}}{\sum_{i=1}^{T} n_{gi}}$. Using this bootstrap data apply Step 2 of Algorithm 1. This process is repeated a large number of times to derive the null distribution of the test statistic required for testing the significance of a gene *g* in Step 3 of Algorithm

References:

[01] B. Brumback & J. Rice, *J Am Stat Assoc.,* 93:443 (1998)

[02] G. Chu, *et al.*, *Users Guide and Technical Document,* (2002)

[03] S. Datta & S. Datta, *Bioinformatics,* 21:9 (2005) [PMID: 15691856]

[04] J. Hwang & S. Peddada, *Annals of Statistics,* 22:1 (1994)

[05] H. Ishwaran & J. Rao, *J Am Stat Assoc.,* 98:462 (2003)

[06] H. Ishwaran & J. Rao, *J Am Stat Assoc.,* 100:471 (2005)

[07] H. Ishwaran & J. Rao, *Annals of Statistics,* 33:2 (2005)

[08] L. Li, *et al.*, *Comb Chem High Throughput Screen,* 4:8 (2001) [PMID: 11894805]

[09] L. Li, *et al., Bioinformatics,* 17:12 (2001) [PMID: 11751221]

[10] H. Liu, *et al.*, *BMC Bioinformatics,* 6:106 (2005) [PMID: 15850479]

[11] E. Lobenhofer, *et al.*, *Mol Endocrinol.,* 16:6 (2002) [PMID: 12040010]

[12] S. Peddada, *et al.*, *Biometrika,* 92:3 (2005)

[13] S. Peddada, *et al.*, *Bioinformatics,* 19:7 (2003) [PMID: 12724293]

[14] S. Peddada, *et al.*, *Bioinformatics,* 21:20 (2005) [PMID: 16109745]

[15] S. Pounds & S. Morris, *Bioinformatics,* 19:10 (2003) [PMID: 12835267]

[16] J. N. K. Rao & K. Subrahmaniam, *Biometrics,* 27:4 (1971)

[17] N. Shi & H. Jiang, *Journal of Multivariate Analysis,* 64:2 (1998)

[18] J. Storey, *et al.*, *Proc Natl Acad Sci.,* 102:36 (2005) [PMID: 16141318]

[19] V. Tusher, *et al.*, Proc *Natl Acad Sci.,* 98:9 (2001) [PMID: 11309499]

Edited by Susmita Datta

Citation: Simmons & Peddada., Bioinformation 1(10): 414-419 (2007)