Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

An adaptive alpha spending algorithm improves the power of statistical inference in microarray data analysis

Jacob P.L. Brand[1, 2, 3*], Lang Chen[3], Xiangqin Cui[3], Alfred A. Bartolucci[3], Grier P. Page[3], Kyoungmi Kim[3, 4], Stephen Barnes[5], Vinodh Srinivasasainagendra[3], Mark T. Beasley[3] and David. B. Allison[3, 6]

[1]Genomic Technologies Section - Research Technology Branch, NIH / NIAID, 50 South Drive, Room 5505, Bethesda, MD 20892-8005; [2]Pennington Biomedical Research Center, Human Genomics Laboratory, Louisiana State University System, 6400 Perkins Rd, Baton Rouge, La 70808; [3]Department of Biostatistics, 1665 University Boulevard, Ryals Public Health Building ,Birmingham, AL 35294-0022; [4]Department of Public Health Sciences, One Shields Avenue, TB-168, University of California, Davis, Davis, California 95616-8638; [5]Department of Pharmacology and Toxicology, 452 McCallum Research Building, University of Alabama at Birmingham,1918 University Boulevard, Birmingham, Alabama; [6]Center for Research on Clinical Nutrition, University of Alabama at Birmingham, Birmingham, Alabama; Jacob P.L. Brand* - Email: brandj@niaid.nih.gov; * Corresponding author

received October 10, 2006; accepted October 21, 2006; published online April 10, 2007

Abstract:

The adaptive alpha-spending algorithm incorporates additional contextual evidence (including correlations among genes) about differential expression to adjust the initial p-values to yield the alpha-spending adjusted p-values. The alpha-spending algorithm is named so because of its similarity with the alpha-spending algorithm in interim analysis of clinical trials in which stage-specific significance levels are assigned to each stage of the clinical trial. We show that the Bonferroni correction applied to the alpha-spending adjusted p-values approximately controls the Family Wise Error Rate under the complete null hypothesis. Using simulations we also show that the use of the alpha spending algorithm yields increased power over the unadjusted p-values while controlling FDR. We found the greater benefits of the alpha spending algorithm with increasing sample sizes and correlation among genes. The use of the alpha spending algorithm will result in microarray experiments that make more efficient use of their data and may help conserve resources.

Keywords: microarray data; contextual evidence; adaptive alpha spending

Background:

Microarray technology has become a widely used and effective research tool in modern molecular biology. It can produce a snapshot of the expression levels of thousands of genes simultaneously at a very low cost per data point. However, researchers are often more interested in how biological pathways respond to experimental condition changes rather than in changes in expression levels of individual genes. The total flux through a pathway can change dramatically through subtle changes in expression levels of genes involved in that pathway. **[1]** Thus, the prevalence of microarray technology in the research of complex metabolic disorders makes the problem of identifying genes with subtle differential expression increasingly important. Unfortunately, the identification of genes with subtle differential expression is challenging due to the huge number of genes involved, the noisiness of the data, and the very small sample sizes (often not more than 5 observed expression levels per gene and/or per treatment group).

Most approaches for identifying differentially expressed genes may be of limited power because they neither take into account nor capitalize on dependencies among genes. As an alternative, we propose an adaptive alpha-spending algorithm that takes into account the dependencies of expression levels among genes explicitly by assigning gene-specific significance levels to each gene. The alpha-spending algorithm is named so because of its similarity with alpha-spending algorithms in interim analysis in clinical trials. **[2]** Interim analysis is often carried out at multiple times in a clinical trial for reasons such as checking adherence to the protocol, economic and ethical reasons. Because in interim analysis the same null-hypothesis is tested multiple times, not correcting for multiple testing will inflate the type 1 error. Multiplicity is controlled in the alpha-spending algorithm by assigning stage specific significance levels to each stage in the clinical trial such that the sum of stage specific significance levels is equal to the overall significance level, i.e., $\sum_{i=1}^{k} \hat{\alpha}_i = \alpha$ with k the number of stages, $\hat{\alpha}_i$ the stage-specific significance level for the *i*-th stage and α the global significance level. The stage-specific significance level is given by $\hat{\alpha}_i = \alpha(t_i) - \alpha(t_{i-1})$, where $\alpha(.)$ is a monotonic non-decreasing function with $\alpha(0) = 0$ and $\alpha(1) = \alpha$ called the alpha-spending function and t_i is the fraction of information accrued in the clinical trial at stage *i*, a quantity between 0 and 1, which is often defined as a function of accrued and planned sample sizes in the clinical trial. **[3]** If, for instance, between stages $i-1$ and i many subjects entered the clinical trial, the resulting significance level $\hat{\alpha}_i = \alpha(t_i) - \alpha(t_{i-1})$ assigned to stage *i* will be relatively high, resulting in a relatively high power for that stage. That $\sum_{i=1}^{k} \hat{\alpha}_i = \alpha$ follows directly from

$$\sum_{i=1}^{k}\hat{\alpha}_i = \sum_{i=1}^{k}\{\alpha(t_i)-\alpha(t_{i-1})\} = \alpha(1)-\alpha(0) = \alpha .$$

The key assumption underlying our alpha-spending algorithm is: if the expression levels of two genes are positively/negatively correlated, then one of the two genes is an activator/repressor of the other gene. This assumption is incorporated into the alpha-spending algorithm by computing the gene-specific significance levels in such a way that they are proportional to the linear regression predictor computed from the correlation matrix of the observed differential expression levels and the observed differential expression levels of other genes. For instance, if a particular gene *A* is highly positively correlated to many up-regulated genes, then this provides additional contextual evidence that gene *A* is also up-regulated. This additional contextual evidence is fed back into the alpha-spending algorithm by assigning a higher significance level to gene *A*. Similar to alpha-spending in clinical trials, the gene-specific significance levels $\hat{\alpha}_1,\ldots\hat{\alpha}_k$ are computed such that they satisfy the condition $\sum_{i=1}^{k}\hat{\alpha}_i = k\alpha$ in order to provide a mechanism for controlling the number of false positives. It can be seen that the alpha-spending algorithm controls the FWER in the weak sense. By this we mean that, under the global null-hypothesis that all genes are non-differentially expressed, the Bonferroni correction applied to the alpha-spending adjusted p-values controls the FWER. This approximate weak control of the FWER follows directly from Bonferroni's inequality $\text{FWER} \le \sum_{i=1}^{k}\alpha_i/k = \alpha$ where α_i is the population quantity from which $\hat{\alpha}_i$ can be regarded as an estimate. The alpha-spending adjusted p-values will be derived in the next section. The alpha-spending adjusted p-values will be derived in the next section.

Methodolgy:

The gene-specific significance levels are based on a prediction equation similar to the linear regression prediction equation $\hat{y}_{y|x} = \bar{y} + \hat{R}_{y,x} * \hat{R}_{x,x}^{-1}\hat{\sigma}_y\hat{D}_x^{-1}(x-\bar{x})$ of an outcome variable y and a vector of predictor variables x, in the case that multivariate normality can be assumed for y and x. In this prediction equation $\hat{y}_{y|x}$ and $\bar{y}$ are the predicted outcome and sample average of y, $\hat{R}_{y,x}$ the row-vector containing the observed correlations between y and x, $\hat{R}_{x,x}^{-1}$ the inverse correlation matrix of x, $\hat{\sigma}_y$ the standard error of $\bar{y}$ and $\hat{D}_x^{-1}$ the diagonal matrix containing on its diagonal the reciprocals of the standard errors of the sample mean vector $\bar{x}$ of x. Because $\bar{y}$ can be interpreted as the predicted outcome in case the values of the predictor variables are unknown, the term $\hat{R}_{x,x}^{-1}\hat{D}_x^{-1}\hat{\sigma}_y(x-\bar{x})$ can be interpreted as the predictive information from the observed values of x for y.

Similar to the predictive information $\hat{R}_{x,x}^{-1}\hat{\sigma}_y\hat{D}_x^{-1}(x-\bar{x})$ above, we will derive predictive information for the unknown population differential expression level δ_i for gene *i*, from the observed differential expression levels $\hat{\delta}_{(-i)}$ of the genes other than gene *i*, the row-vector $\hat{\phi}(-i)$ containing the correlations between the observed expression level $\hat{\delta}_i$ and $\hat{\delta}_{(-i)}$, the inverse correlation-matrix $\hat{\Phi}(-i)^{-1}$ of $\hat{\delta}_{(-i)}$, the standard error $\hat{\psi}_i$ of $\hat{\delta}_i$ and the standard errors of $\hat{\delta}_{(-i)}$. Under the assumption that the k-vector $\hat{\delta} = (\hat{\delta}_1,\ldots,\hat{\delta}_k)^T$ has a multivariate normal distribution with the unknown population differential expression levels $\delta = (\delta_1,\ldots,\delta_k)^T$ as mean vector and Σ as variance-covariance matrix, we can write a similar prediction equation for the predicted differential expression level $\tilde{\delta}_i$ as $\tilde{\delta}_i = \delta_i + \hat{\phi}(-i)\hat{\Phi}(-i)^{-1}\hat{\psi}_{ii}\hat{\Psi}(-i)^{-1/2}(\hat{\delta}_{(-i)}-\delta_{(-i)})$. In this equation $\delta_{(-i)}$ is the $(k-1)$ vector containing the population differential expression levels of all genes other than gene *i*, and $\hat{\Psi}(-i)^{-1/2}$ is a diagonal matrix containing the reciprocals $\sqrt{1/\hat{\psi}_{jj}^2}$ of standard errors $\hat{\psi}_{jj}$ of $\hat{\delta}_{(-i)}$. Because $\delta_{(-i)}$ is an unknown vector, we will rewrite this prediction equation into $\tilde{\delta}_i = \delta_i + \hat{\phi}(-i)\hat{\Phi}(-i)^{-1}\hat{\psi}_{ii}\hat{\Psi}(-i)^{-1/2}\hat{\delta}_{(-i)} - \hat{\phi}(-i)\hat{\Phi}(-i)^{-1}\hat{\psi}_{ii}\hat{\Psi}(-i)^{-1/2}\delta_{(-i)}$ and will use $\hat{\pi}_i = \hat{\phi}(-i)\hat{\Phi}(-i)^{-1}\hat{\psi}_{ii}\hat{\Psi}(-i)^{-1/2}\hat{\delta}_{(-i)}$ as our predictive information. This predictive information $\hat{\pi}_i$ for gene *i* can be interpreted as the inproduct between the correlations $\hat{\phi}(-i)$ between the differential expression level of gene *i* and the differential expression level of all genes other than gene *i*, and the differential expression levels $\hat{\delta}_{(-i)}$ of gene *i* adjusted for the standard errors and correlation matrix of $\hat{\delta}_{(-i)}$, and the standard error of $\hat{\delta}_i$. If, for

instance, gene *i* is strongly positively correlated to highly up-regulated genes then there is contextual evidence that gene *i* is also up-regulated and this contextual evidence is incorporated by a relatively large positive value of $\hat{\pi}_i$. Because the p-value for the differential expression of a gene depends on the absolute value of its differential expression, our gene-specific significance level will be defined as $\hat{\alpha}_i = c * |\hat{\pi}_i|$, where c is a constant. In order to provide a mechanism for controlling the number of false positives we will chose $c = 1/\overline{|\hat{\pi}|}$ with $\overline{|\hat{\pi}|} = \frac{1}{k}\sum_{i=1}^{k}|\hat{\pi}_i|$ so that $\frac{1}{k}\sum_{i=1}^{k}\hat{\alpha}_i = \alpha$, i.e., the average gene-specific significance level is equal to the global significance level α. The alpha-spending adjusted p-value p_i^+ for differential expression of gene *i* will be defined as $p_i^+ = (\alpha/\hat{\alpha}_i)\, p_i$ with p_i the original p-value for differential expression of gene *i*. It can be easily verified that $p_i^+ \le \alpha$ is equivalent with $p_i \le \hat{\alpha}_i$ so that indeed $\hat{\alpha}_i$ is the gene-specific significance level for gene *i*. To the alpha-spending adjusted p-values any method for multiplicity control may be applied.

Discussion:

We have proposed an adaptive alpha-spending algorithm for finding differentially expressed genes in microarray data sets in which observed dependencies among genes are incorporated by assigning gene specific significance levels to each gene. We think this procedure may increase the power in finding differentially expressed genes. The constraint $\frac{1}{k}\sum_{i=1}^{k}\hat{\alpha}_i = \alpha$ provides a mechanism for controlling the number of false positives. We have shown that the alpha-spending algorithm provides approximately weak control of the FWER.

To further investigate power of alpha-spending procedure and its ability to control the number of false positives we have conducted a simulation study with a relatively small number of genes ($k = 700$) with two treatment groups of equal sample sizes. The alpha-spending algorithm was applied to the equal variances t-test for comparing the two groups using the within group correlation among genes as contextual information. We assessed the Per Comparison Error Rate (PCER) under the complete null, i.e. all genes are non-differentially expressed, as well as the partial null, some genes are non-differentially expressed but not all. We also evaluated the False Discovery Rate (FDR) defined in **[4]** under the partial null only and the power improvement in special circumstances. The PCER is the expected number of false positives divided by the number of truly differentially expressed genes. The FDR is defined as $E\left[\frac{V}{R} \mid R > 0\right] P(R > 0)$, where V is the number of false positives and R the number of genes declared significant. The FDR can be loosely interpreted as the proportion of false positives among all genes declared significant. For all simulation parameter settings, simulated data sets were generated from which 20% of the genes were correlated with the same correlation coefficient ρ and the remaining 80% of the genes were not correlated and not correlated with the group of correlated genes either. The ρ parameter was varied ($\rho = 0.3, 0.5, 0.7$) and the group size n was also varied ($n = 4, 6, 10$). Under the partial null, the population mean difference $\Delta = \Delta(1-\beta)$ of the 20% correlated genes was varied such that the corresponding power of the ordinary t-test was varied by $1-\beta = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ and the remaining 80% other genes were non-differentially expressed. All simulated microarray data sets were generated from a multivariate normal distribution.

Our simulation study confirms that the alpha-spending algorithm controls the PCER and FDR in many practical situations. Under the complete null, the PCER was controlled with respect to all genes overall as well as for the group of uncorrelated genes. For the group of correlated genes, the PCER tended to be inflated (Table 1). Under the partial null, the PCER was controlled in all simulation parameter settings and the FDR was controlled in most of the simulation parameter settings (Figure 1). The observed PCER decreases for increasing group-size and correlation, but this relationship was not seen in the observed FDR. On average the alpha-spending algorithm improves the power and this power improvement increased for increasing group size or increasing correlation. The power improvement can be up to 47% for $\rho = 0.7$ and $n = 6$ (Figure 2). However the power improvement varied substantially across individual simulated data sets. For lower values of ρ and n power decreased for some simulated data sets and this decrease in power was up to 15% for $\rho = 0.3$ and $n = 4$. For $n \ge 6$ the alpha-spending algorithm seemed to have added value. We also increased the number of genes in the simulation to 2000 for some cases; the results are very similar to what was obtained for the simulations with 700 genes.

		Correlated genes			Uncorrelated genes			All genes		
ρ	n	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
0.3	4	0.0092	0.0506	0.1044	0.0099	0.0483	0.0966	0.0098	0.0487	0.0982
0.3	6	0.0136	0.0689	0.1362	0.0095	0.0466	0.0938	0.0103	0.0510	0.1023
0.3	10	0.0117	0.0660	0.1316	0.0098	0.0463	0.0928	0.0102	0.0502	0.1006
0.5	4	0.0111	0.0663	0.1333	0.0091	0.0466	0.0932	0.0095	0.0505	0.1012
0.5	6	0.0175	0.0864	0.1664	0.0085	0.0421	0.0849	0.0103	0.0510	0.1012
0.5	10	0.0238	0.1006	0.1849	0.0081	0.0437	0.0875	0.0112	0.0551	0.1070
0.7	4	0.0326	0.1078	0.1908	0.0088	0.0450	0.0897	0.0136	0.0575	0.1099
0.7	6	0.0126	0.0794	0.1723	0.0088	0.0433	0.0864	0.0096	0.0505	0.1036
0.7	10	0.0353	0.1265	0.2249	0.0079	0.0389	0.0813	0.0134	0.0564	0.1101

Table 1: Observed PCER for the alpha-spending post-processed p-values estimated for correlated genes, uncorrelated genes, and all genes under the complete null hypothesis that all genes are non-differentially expressed. The number of genes in each simulation was 700 and the nominal alpha levels of 0.01, 0.05, and 0.1 were used for identifying differential genes. In each simulation parameter setting (ρ, n) the observed PCER was estimated from 100 simulated data sets

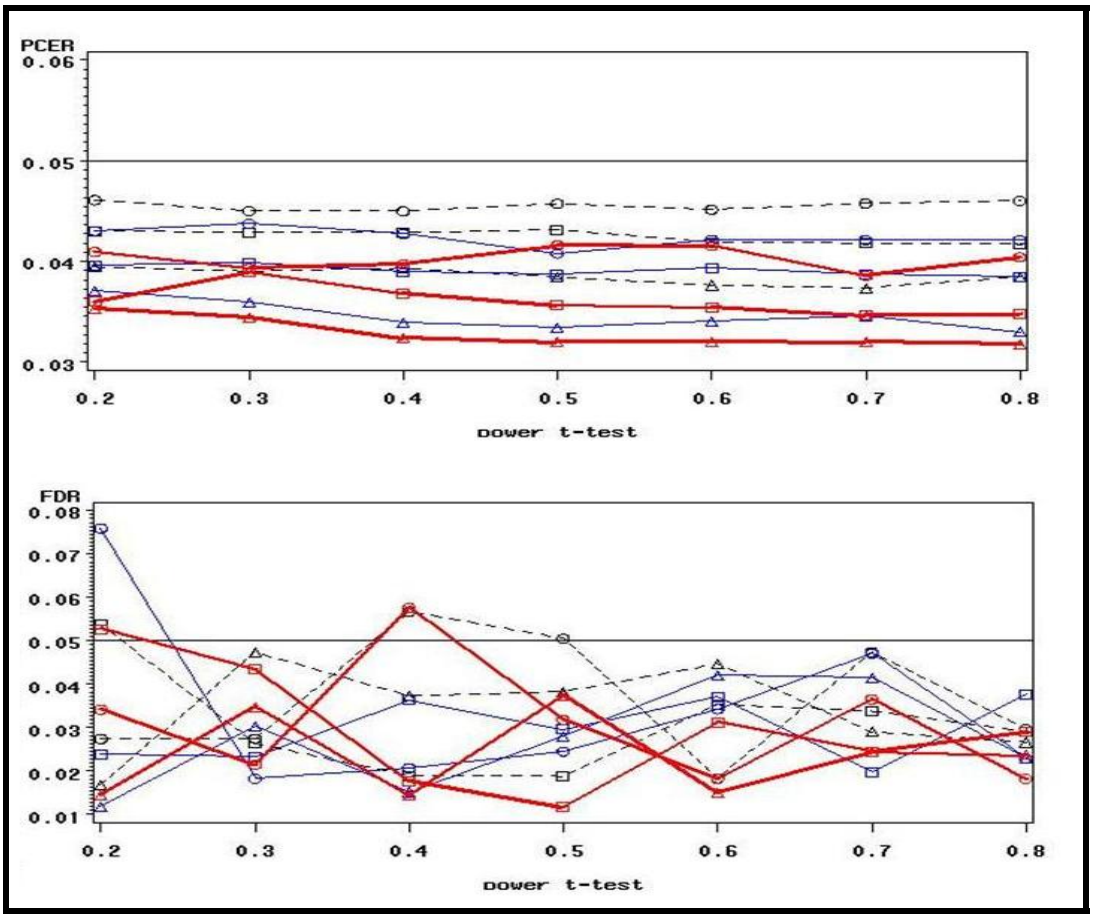

Figure 1: Observed PCER and observed FDR of the alpha-spending algorithm as a function of power of the ordinary t-test for different correlations $\rho = 0.3, 0.5, 0.7$ and different group sizes $n = 4, 6, 10$ for $k = 700$. The number of genes in each simulation was 700 and the nominal alpha levels of 0.05 was used for identifying differential genes. A thin dashed black line, a solid blue line, and a thick red line refer to a correlation ρ of 0.3, 0.5, and 0.7, respectively. The group sizes of 4, 6, and 10 are represented by circles, squares, and triangles, respectively. In each simulation parameter setting (ρ, n) the observed PCER was estimated from 100 simulated data sets

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net

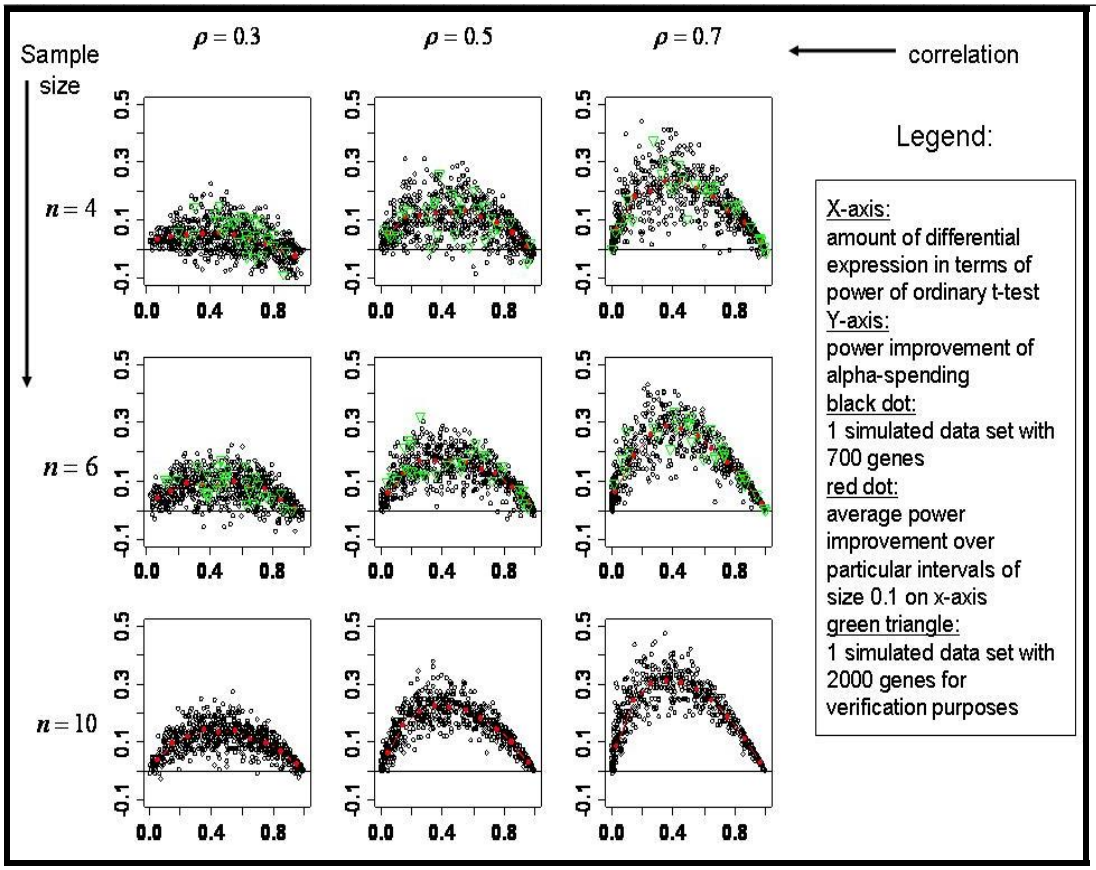

Figure 2: Power improvement of alpha-spending p-values with respect to the ordinary t-test. The results are from the partial null hypothesis simulations with 20% of the genes differentially expressed and correlated with the same correlation coefficient ρ and 80% of the genes non-differentially expressed and uncorrelated. For $k = 700$, the $700 = 7 \times 100$ simulated data sets per plot were obtained by independently generating 100 data sets for each of seven different values of the population mean differential expression Δ. These seven values of $\Delta = \Delta(1-\beta)$ were obtained such that the corresponding power of the ordinary t-test in detecting the differentially expressed genes was varied by $1-\beta = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$. For $k = 2000$ the 30 simulated data sets correspond to $1-\beta = 0.5$ only. The situation $k = 2000$ is simulated for $n = 4, 6$ but not for $n = 10$

The above mentioned opposite relationships between ρ and n on one side and power improvement and observed PCER on the other side, can be explained by $k\alpha = E\left[\sum_{i=1}^{k} \hat{\alpha}_i\right] = E\left[\sum_{i \in \mathrm{DE}} \hat{\alpha}_i\right] + E\left[\sum_{i \in \mathrm{NDE}} \hat{\alpha}_i\right]$, where DE represent the set of truly differentially expressed genes and NDE represent the set of truly non-differentially expressed genes. An increase in power means an increase of $E\left[\sum_{i \in \mathrm{DE}} \hat{\alpha}_i\right]$, which results in a decrease of $E\left[\sum_{i \in \mathrm{NDE}} \hat{\alpha}_i\right]$ and thus a decrease in PCER. This relationship was not found for FDR, which was possibly due to simulation error in estimating FDR as a consequence of the large variation in the number of genes declared differentially expressed across different simulated data sets (see Figure 2). The inflation of the PCER among correlated genes under the global null may be explained by the fact that $|\pi_i|$ increases for increasing correlations between the i-th genes and other genes. Because of $\frac{1}{k}\sum_{i=1}^{k} \hat{\alpha}_i = \alpha$, the type 1 error is

inflated for the correlated genes and deflated for the uncorrelated genes. This situation highlighted that the alpha-spending algorithm may be more likely to detect spurious findings in case of strong correlations among many non-differentially expressed genes. A topic of future research is to investigate whether this situation can be ameliorated by developing adjustment procedures for the gene-specific significance levels in which lower gene-specific significance levels are assigned to genes with lower observed differentially expression levels. Another topic of future research is the improvement of the power of the alpha-spending algorithm by the application of Empirical Bayes techniques **[5]** to the estimation of differential expression levels **[6]**, correlations among differential expression levels **[7]**, and the standard error of differential expression levels. **[8]** A simulation study reported that the mean squared error of EB estimates of differential expression levels is as low as 0.05 times that of the ordinary least squares estimators. **[6]** Finally, because the number of genes often runs in the ten thousands, the inversion of the correlation matrix $\hat{R}_{x,x}^{-1}$ is extremely computationally intensive and may require a super-computer. The approximation of $\hat{R}_{x,x}^{-1}$ by easier to invert block-diagonal matrices based on clustering of the genes might be investigated for the purpose of more practical use of the alpha-spending algorithm.

Conclusion:

We have proposed an adaptive alpha-spending algorithm for finding differentially expressed genes in microarray data sets in which observed dependencies among genes are incorporated by assigning gene specific significance levels to each gene. We have shown that the alpha-spending algorithm approximately controls the FWER under the complete null. In a simulation study we have illustrated that the alpha-spending algorithm controls the PCER and FDR and improves the power when applied to the ordinary t-test under special circumstances within the two group comparisons with equal group sizes. However, there may be situations in which the PCER is inflated as was shown for the correlated genes under the complete null.

Acknowledgement:

We thank Dr. Gary Gadbury for helpful contributions to earlier versions of this paper and Mr. Jelai Wang for help in developing early versions of the simulation code and the online supplement. We thank Dr. Purushotham Bangalore from the Department of Computer and Information Sciences (CIS) at the University of Alabama at Birmingham (UAB) for allowing us to use CPU cycles on the Everest cluster of UAB CIS and Dr. Alan Shih for allowing us to use CPU cycles at the Cahaba cluster at The Enabling Technology Lab, which part of the Mechanical Engineering Dept at UAB. This research was supported in part by NIH grants P30DK56336, P01AG11915, R01AG018922, P20CA093753, R01AG011653, R01DK56366, R01ES09912, U24DK058776, and U54CA100949; NSF grant 0090286; and a grant from the University of Alabama Health Services Foundation.

References :

[01] V. K. Mootha, *et al., Nat. Genet*., 34:267 (2003) [PMID: 12808457]

[02] K. K. G. Lan & D. L. Demets, *Biometrika*, 70:659 (1983)

[03] J. M. Lachin, *Stat Med*., 24:2747 (2005)

[04] Y. Benjamini & Y. Hochberg, *JRSS-B*, 57:289

[05] C. N. Morris, *J Am Stat Assoc*., 78:47 (1983)

[06] J. W. Edwards, *et al., Funct. Integr. Genomics,* 5:32 (2005) [PMID: 15455262]

[07] V. Cherepinsky, *et al., Proc. Natl. Acad. Sci.,* 100:9668 (2003) [PMID: 12902543]

[08] X. G. Cui, *et al., Biostatistics*, 6:59 (2005) [PMID: 15618528]

Edited by Susmita Datta

Citation: Brand *et al.,* Bioinformation 1(10): 384-389 (2007)