Journal of Data Science **5**(2007), 00-00(JDS-341)

Detection of Differentially Expressed Genes In Small Sets of cDNA Microarrays

Simon Rosenfeld
National Cancer Institute

Abstract: Methods for testing the equality of two means are of critical importance in many areas of applied statistics. In the microarray context, it is often necessary to apply this kind of testing to small samples containing no more than a dozen elements, when inevitably the power of these tests is low. We suggest an augmentation of the classical t-test by introducing a new test statistic which we call "bio-weight." We show by simulation that in practically important cases of small sample size, the test based on this statistic is substantially more powerful than that of the classical t-test. The power computations are accompanied by ROC and FDR analysis of the simulated microarray data.

Key words: Differential expression, false discovery rate, hypotheses testing, microarray data, Monte Carlo simulation, t-test.

1. Introduction

A central problem in microarray data analysis is identification of differentially expressed genes, i.e., those genes whose expression intensities are significantly associated with the group label or covariate (Simon *et al.*, 2004). In the univariate case, a generally accepted way of statistical reasoning for solving this problem would be through hypothesis testing. In general, this consists of formulating an appropriate null hypothesis, constructing the criterion for discriminating between the null and alternative, and computing the probability of committing Type I error, i.e., erroneously rejecting the null hypothesis when in fact it is true. In microarray data analysis, this method of statistical reasoning stumbles over a fundamental impediment which is termed in the literature the problem of multiple testing (Dudoit *et al.*, 2003). Because there is simultaneous testing of about 10,000 null hypotheses, it is quite possible by pure chance that a large number of genes would be declared differentially expressed even when, in fact, the null hypothesis (of no differential expression) is true. Furthermore, the genes erroneously declared significant may substantially outnumber the modest group of genes which are actually differentially expressed, thus providing a biologist with numerous false leads for his follow up investigations.

Classical statistics usually deals with the situation when the number of parameters to be estimated is smaller, or even substantially smaller, than the number of subjects in the study. This is not the case in microarray data analysis where the expression profiles of many thousands of genes are usually estimated from experiments with only several dozen subjects. This new statistical paradigm, often referred to as "curse of dimensionality" (Donoho, 2000), stimulated p-values development of a wide range of innovative ideas whose pros and cons are currently being assessed to determine their practical usability (Efron *et al.*, 2001; Kerr *et al.*, 2000). Quite frequently, the real-life situation is even worse than simply the "curse of dimensionality." Oftentimes, it is the "ultimate curse of dimensionality" because only two or three microarrays are available for the analysis. Having at our disposal a couple dozen samples, we may reasonably expect at least a vague resemblance between our statistical estimates and their asymptotic counterparts. However, we have to completely give up this hope with the sample size of only two or three. We believe that this case of very small number of microarrays requires special statistical treatment.

In this paper, we suggest a modification of the classical t-test which is specifically designed to enhance the sensitivity for small number of microarrays. This modification was first introduced by the author in (Rosenfeld *et al.*, 2004) and was based on the subject matter considerations which are briefly reproduced here. Figure 1 represents the so called volcano plot which is frequently used in microarray data analysis. In this plot, computed from the cDNA microarray experiments with dyallil disulfide (Rosenfeld *et al.*, 2004), the horizontal axis represents the average (across replicates) $\log_2(I^{red}/I^{green})$, where I^{red} and I^{green} are the fluorescent intensities corresponding to the treatment and control groups, respectively. The vertical axis corresponds to the negative decimal logarithm of the p-values for the gene specific t-tests. This plot vividly illustrates a conflict between the notions of the "biological" and "statistical" significance. The most statistically significant genes (i.e., those corresponding to the smallest p-values) are located at the top of the volcano plot. The leftmost and rightmost genes correspond to a large absolute fold change. While "biologically" the genes with the largest fold changes are of major interest, "statistically" these genes are usually not the most significant, and vice versa. Application of both of these criteria usually produces very few, if any, significant genes. In practice, the number of replicates is small, often no more than five. In such cases, the power of the t-test is very low and there is no compelling reason to follow the recommendations based solely on the t-test p-values. To reconcile conflicting meanings of significance suggested by the smallest p-values and the largest fold changes, we introduce a new test statistic that we call "bio-weight." This test statistic, $\hat{b}$, is defined as the product of the absolute fold change and negative decimal logarithm of the t-test p-value. The concept of bio-weight is illustrated in Figure 1 where the pair of solid hyperbolic lines represents the 99% quantile of the bio weight. Application of the bio weight resolves, or at least mitigates, the conflict between the requirements of statistical and biological significance. In contrast to the above mentioned two alternative approaches, i.e., scoring significance by either the smallest p-values or the largest fold changes, it pays attention to both. Extensive Monte Carlo simulation shows that performance of the bio weight, measured by its sensitivity to presence of the differentially expressed genes, is noticeably higher than that of the standard t-test.

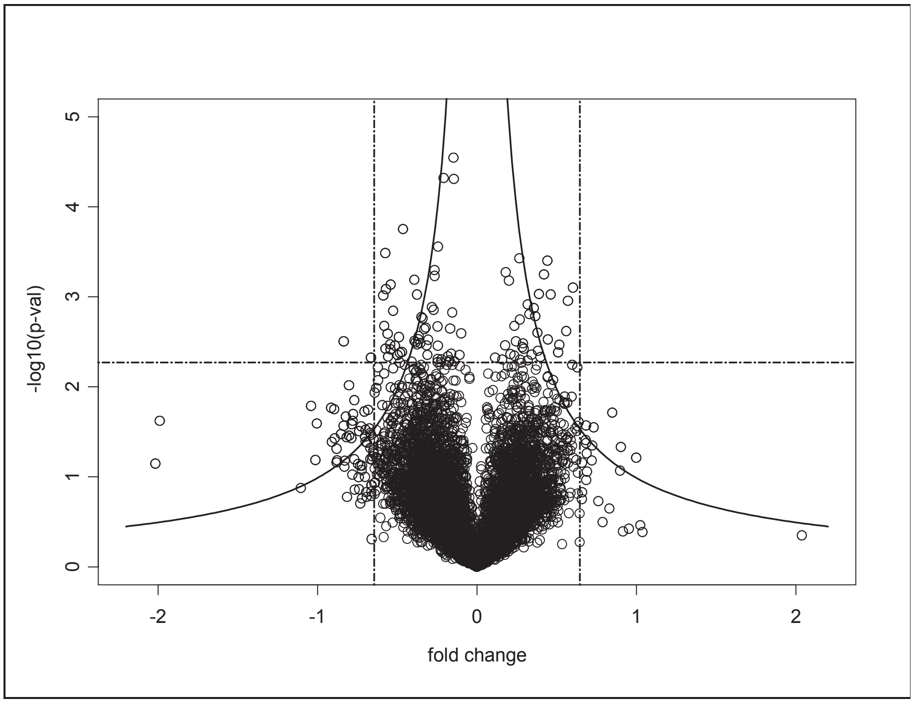

Figure 1: Illustration of the concept of bio-weight

In this paper, we introduce the bio-weight test statistic in a more formal way and present the results of comprehensive simulation study comparing its power with the power of the classical t-test.

2. Statistical Model and Simulation Framework

As a starting point for our analysis, we adopt the Random Variance Model (RVM) introduced in (Wright and Simon, 2003). In our experience, this model provides a flexible, mathematically tractable tool, capable of realistically capturing various aspects of the microarray signal structure. We apply the RVM to a class discovery problem where the only goal is to determine the set of genes differentially expressed in the treatment versus control groups of arrays. Following the RVM, we assume that the ratios of fluorescent intensities in these two groups are lognormally distributed, thus producing the normal distribution of log-ratios, $x_{ij} = \log(I_{ij}^{ref}/I_{ij}^{green})$, where i and j are the indices of arrays and genes, respectively. We further assume that the gene-specific variances of these normal distributions, σ_j^2, are not identical for all the genes but are considered themselves as random variables drawn from a suitably selected probabilistic distribution. Although many ways are conceivable to characterize variability of the gene-specific variances, we follow Wright and Simon, 2003, in their choice of the "inverse gamma" model. That is, we assume that σ_j^{-2} are gamma-distributed with parameters α ("shape") and θ (inverse "rate"). Under the above formulated assumptions, joint distribution of any randomly selected x from x_{ij} and y from $y_j = \sigma_j^{-2}$ is expressed as follows

$$H(x,y) = \sqrt{\frac{y}{1\pi}}e^{-\frac{x^2 y}{2}}\frac{1}{\Gamma(\alpha)\theta^{\alpha}}y^{\alpha-1}e^{-\frac{y}{\theta}} \tag{2.1}$$

A nice property of distribution (2.1) (apparently not noticed by Wright and Simon, 2003) is that marginal distribution of x, obtained by integration of $H(x,y)$ over y, is reduced to the Student' s t-distribution. More specifically, the marginal distribution of $x\sqrt{\alpha\theta}$ is t-distribution with 2α degrees of freedom. This finding may be effectively used for estimating parameters α and θ from the data at hand. Figuratively speaking, here the "blessing of dimensionality" (Donoho, 2000) comes into play because now tens of thousands of gene expressions may be lumped together for estimating only two parameters. To this end, we first recall that the variance, μ_2, and kurtosis , γ_4, of the t-distribution with 2α degrees of freedom are expressed as (Armitage and Colton, 1998, p.4396):

$$\mu_2 = [(\alpha-1)\theta]^{-1}; \quad \gamma_4 = 3(\alpha-2)^{-1}\mu_2 \tag{2.2}$$

Substituting the estimates $\hat{\mu}_2$ and $\hat{\gamma}_4$ into equations (2.2), we obtain

$$\hat{\alpha} = 2 + \frac{3}{\gamma_4}; \quad \hat{\theta}^{-1} = \left(1 + \frac{3}{\gamma_4}\hat{\mu}_2\right) \tag{2.3}$$

Simulation experiments with typically 10,000 genes, 3 to 10 replicates and 1,000 Monte Carlo repetitions reveal very high accuracy of these estimates. It is of particular importance to have an unbiased estimate for the combination $\hat{\alpha}\hat{\theta}$ because, if the RVM indeed provides an adequate description of the microarray data, then $x\sqrt{\hat{\alpha}\hat{\theta}}$ is expected to be t-distributed with $2\hat{\alpha}$ degrees of freedom. As

Table 1: Estimation of parameters of parental gamma distribution. α is the shape parameter used in simulation, $\hat{\alpha}$ and $\hat{\theta}$ are the parameters estimated through equations (2.3)

	α						
	2.5	3	4	5	8	12	20
$\hat{\alpha}$	2.65	3.08	4.06	5.04	8.06	12.2	20.5
$(\hat{\alpha}\hat{\theta}/\alpha\theta)^{1/2}$	0.984	0.995	0.998	1.000	1.000	1.000	1.000

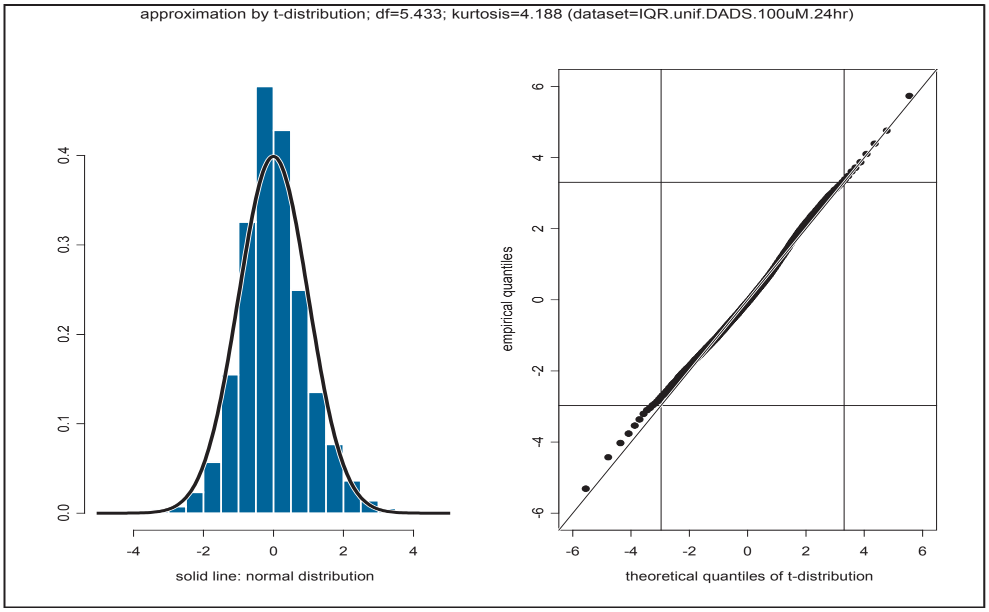

Figure 2: Left panel: empirical distribution of pooled log-ratios. Right panel: QQ-plot empirical vs. theoretical t-distributions.

an example, Table 1 shows the results of computation of $\hat{\alpha}$ and $\sqrt{\hat{\alpha}\hat{\theta}}$ as functions of the shape parameter used in simulation (10,000 genes, 3 replicates). As seen from this Table, parameters estimated from the model are always very close to those used in simulation, thus corroborating high accuracy of estimating equations (2.3).

An important question associated with RVM is this: “How successful is this model in representing the real-world data?” Figure 2 shows an example of fitting RVM to the dataset from the above cited DADS study (Rosenfeld *et al.*, 2004). In this example, the kurtosis and number of degrees of freedom of the approximating t-distribution are found to be 5.3 and 4.2, respectively. The Q-Q plot in Figure 2 compares empirical quantiles of the pooled log-ratios (i.e., all the x_{ij}) and theoretical quantiles of the t-distribution with the parameters estimated as in (2.3). As seen from this plot, there is a good agreement between the RVM model and the data observed in the DADS experiment.

Another useful property of the RVM is its ability to correctly capture heteroscedasticity of the microarray signal, i.e., dependence of the between-replicates variability on the strength of the microarray signal. Figure 3 shows the scatter plots and regression lines in the model $\sigma_i = A + B \log_2(I_i^{red}/I_i^{green})$ for the observed (left panel) and simulated (right panel) data. Again, a remarkably good agreement between them serves as additional evidence of adequacy of the RVM.

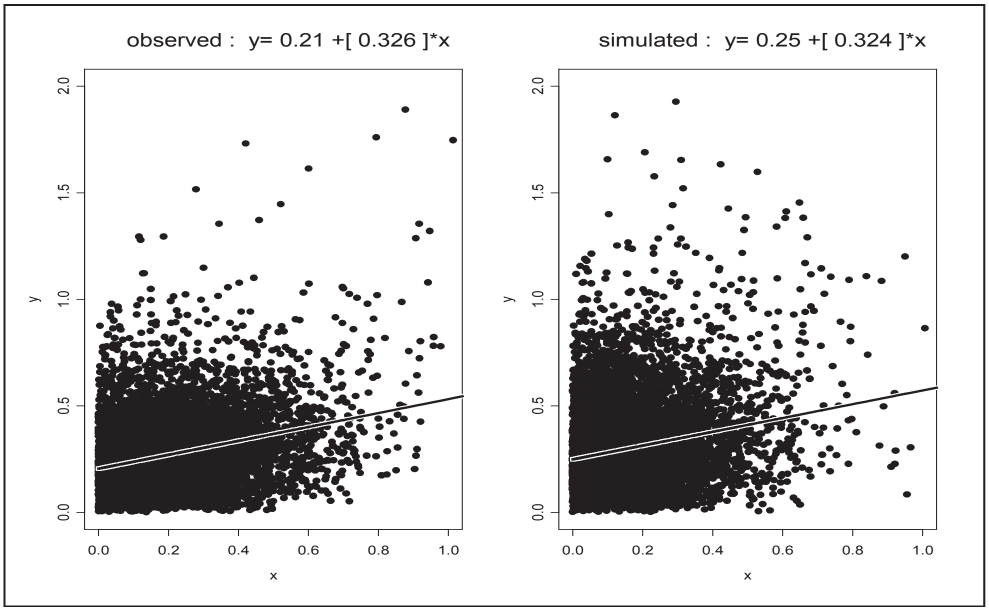

Figure 3: Heteroscedasticity of microarray signal. Left panel: observed in DADS experiment. Right panel: RVM model for DADS data

In this work, the RVM has been used as the basis for simulation experiments with the bio-weight test statistic. The simulation model is controlled by three parameters: number of genes, N, number of replicates, n, and shape paramete, α, in the "inverse gamma" model for the variance variability. The scale parameter, θ, is of no particular importance and without loss of generality may be conveniently set to α^{-1}, in which case expectation of the inverse variance is always equal to 1. Because the expectation in the gamma model is $\alpha\theta$ and the variance is $\alpha\theta^2$, the condition $\alpha \gg 1$ corresponds to low gene-to-gene heterogeneity of variances. We will further refer to this case as "weak fluctuations." So far, our experience with the RVM model has been limited to fitting the DADS data, where degrees of freedom of the pooled t-distributions were always between 4 and

5 (corresponding to $2 \leq \alpha \leq 2.5$), thus covering the range of weak fluctuations. It is our perception that the case of weak fluctuations is the most representative in practical perspective. However, in simulation it is quite possible to consider also the case of "moderate fluctuations," $1 \leq \alpha \leq 2$, when standard deviation of variance is comparable to its expectation, and even the case of "strong fluctuations," $\alpha < 1$, when standard deviation of variance is much greater than its expectation.

3. Definition and Power of the Bio-weight Test Statistic

To introduce a formal definition of the bio-weight test statistic, let $x_{ij} = \log_2(I_{ij}^{red}/I_{ij}^{green})$, where the indices i and j denote the microarray and gene labels, respectively. We first perform the sequence of gene-by-gene one-sample two-sided t-tests and obtain the corresponding fold changes, $\hat{R}_j$, and p-values, $\hat{p}_j$. At this point, we assume that the t-statistics are t-distributed with $(n-1)$ degrees of freedom, where n, as before, denotes the number of microarrays. Next, we pool all the x_{ij} together, compute the variance, $\hat{\mu}_2$, and kurtosis, $\hat{\gamma}_4$, for the pooled sample, estimate $\hat{\alpha}$ and $\hat{\theta}$ from equations (2.3), and create the global scale parameter, $\hat{\sigma}_g = (\hat{\alpha}\hat{\theta})^{-1/2}$. Having performed all these preparatory steps, we compute the set of gene-specific statistics, $\hat{b}_j = -|\hat{R}_j|\log(p_j)/\sigma_g$, which we call "bio-weight." This test statistic is used as a new score of gene significance.

Table 2: Powers of the t-test and bio-weight (b) for normal population. n is the sample size.

n	R_0									
	1		2		3		4		5	
	t	b	t	b	t	b	t	b	t	b
2	0.053	0.137	0.179	0.632	0.262	0.932	0.336	0.995	0.437	1.000
3	0.104	0.235	0.466	0.857	0.732	0.995	0.921	1.000	0.972	1.000
4	0.164	0.458	0.724	0.949	0.968	1.000	0.999	1.000	1.000	1.000
5	0.399	0.546	0.910	0.986	0.998	1.000	1.000	1.000	1.000	1.000
6	0.505	0.635	0.971	0.997	1.000	1.000	1.000	1.000	1.000	1.000
7	0.586	0.690	0.991	0.999	1.000	1.000	1.000	1.000	1.000	1.000
8	0.677	0.774	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
9	0.739	0.811	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	0.796	0.856	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

In order to demonstrate the superior power the $\hat{b}$-statistic, we first consider performance of the $\hat{b}$-statistic applied to each gene independently, much in the same manner as we usually use in the gene-by-gene t-testing. The only difference we set up is that in each Monte Carlo repetition we generate our samples with different variances independently drawn, in accordance to RVM, from the inverse gamma population. In each Monte Carlo cycle, we test the null hypothesis $R = 0$ against the alternative$|R| \geq R_0$. Table 2 shows power of the bio-weight test as a function of R_0 in comparison with the power of the standard one sample t-test, each at significance level 5%. As seen from this Table, power of the bio-weight is substantially higher than that of the t-test if the sample size is small and/or the shift parameter R_0 is small. Typically, 1,000 repetitions have been used in these simulations. Examples in Tables 2 and 3 are computed with $\alpha = 2$.

Table 3: Powers of the t-test and bio-weight (b) for gamma population with shape=2. n is the sample size.

n	R_0									
	1		2		3		4		5	
	t	b	t	b	t	b	t	b	t	b
2	0.050	0.138	0.149	0.675	0.217	1.000	0.295	1.000	0.379	1.000
3	0.079	0.220	0.344	0.992	0.641	1.000	0.802	1.000	0.938	1.000
4	0.099	0.345	0.652	1.000	0.917	1.000	0.981	1.000	0.997	1.000
5	0.181	0.492	0.880	1.000	0.990	1.000	0.999	1.000	1.000	1.000
6	0.318	0.644	0.967	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7	0.443	0.728	0.993	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	0.532	0.801	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
9	0.692	0.874	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	0.798	0.925	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

As mentioned above, it is generally accepted in microarray data analysis that the gene-specific log-ratios of fluorescent intensities are distributed normally. It may be noted, however, that this assumption is not self-evident, and often is difficult to validate due to lack of replicates and possible heterogeneity of variances. For this reason, it is of some interest to investigate the robustness of the bio-weight test statistic with respect to deviations from normality. To this end, we no longer assume that the gene-specific t-statistics are t-distributed; however, we still use the same expression for the $\hat{b}$- statistic. Table 3 shows the results of power computations for the case when the null and alternative samples are both drawn from a highly skewed gamma distribution with shape parameter 2 (shifted on R_0). Comparison of Tables 2 and 3 shows that generally the power of the $\hat{b}$- statistic not only does not decrease due to non-normality but in some cases becomes even slightly greater. On the contrary, the power of the standard t-test is somewhat diminished by the non-normality, thus making the comparative benefits of the bio-weight even more valuable.

4. Application of the Bio-weight to Detection of Differentially Expressed Genes

Higher power of the bio-weight test naturally translates into a higher sensitivity in detection of differentially expressed genes. After selection of the significance score, two main strategies in the search for such genes are usually employed. In the first strategy, we fix a certain number of genes with top scores and try to evaluate the error rate associated with this selection. The second strategy consists of fixing the error rate and declaring significant all the genes compatible with this rate. Obviously, success of both strategies depends on actual presence and magnitude of the "genomic signal" in the microarray data at hand, i.e., on the number of differentially expressed genes and their average fold change. In (Rosenfeld *et al.*, 2004) we have developed a simulation model mimicking a real-life experiment which takes into consideration non-normality of the cDNA fluorescent log-intensities. Despite this non-normality, we found that the difference between the red and green log-intensities, i.e., log-ratio, turned out to be very close to normal, in agreement with what is commonly assumed in microarray data analyses. To illustrate the core advantages of the bio-weight, we adopt this somewhat simplified way of reasoning and accept the assumption of normality of log-ratios, and the RVM as the simulation platform. Suppose that there are $m/2$ of over- and $m/2$ of under-expressed genes among the total of N genes printed on the microarrays we analyze. We will use the abbreviation AFC to denote the Average (absolute) Fold Change. As an error control, we use the true positive fraction (TPF) and false positive fraction (FPF) enabling us to display the results in the form of receiver operating characteristic (ROC) curves. Figure 4 depicts one such curve computed for AFC=3 and sample size 3. As shown in this Figure, if we fix the FPF on the level 5%, then the probability to detect significant genes, measured by TPF, is almost twice as large for the bio-weight test as compared to the t-test. Summary of all results of this kind, computed for multiple AFC and sample sizes, is presented in Table 4. These results suggest that in the case of a small number of replicates, the bio-weight test is tangibly more powerful than the t-test and provides a higher probability to detect differentially expressed genes. For example, if the AFC of the differentially expressed genes is about 3 and the number of replicates is 3 then the t-test selection offers only $\sim 29\%$ probability of recovering those genes, whereas the bio-weight test increases this probability to 54%.

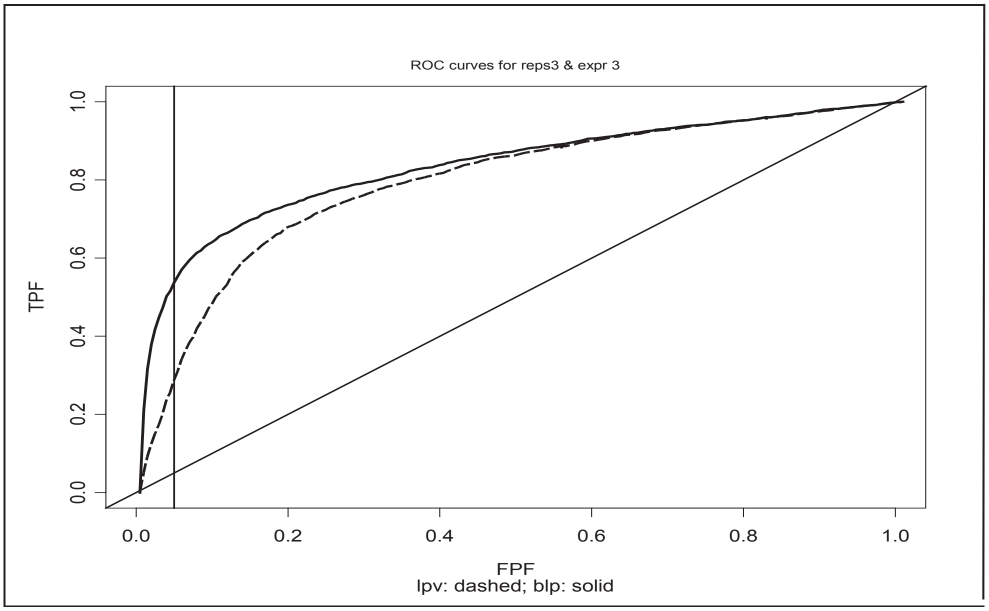

Figure 4: ROC curves for selecting significant genes: dashed line: t-test; solid line: bio-weight

We now consider the second of the aforementioned strategies for detecting the differentially expressed genes, i.e., the one where the error rate is *a priori* fixed but the number of genes to be discovered is unknown. In this situation, an appropriate measure for the error control is False Discovery Rate (FDR), i.e., proportion of the genes erroneously declared significant among *all* the genes declared significant (Dudoit *et al.*, 2003; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Hsueh *et al.*, 2003; Kwong *et al.*, 2002). Among several techniques for implementing the FDR principle, the Benjamini-Hochberg procedure (BH/FDR) has recently won the widest popularity due to its computational simplicity and ease of interpretation. Transplanted into the genomic context, the BH/FDR procedure controls *expectation* of the proportion of false discoveries among the genes declared significant. According to this procedure, in order to obtain the subset of significant genes from the experiment at hand, several simple steps should be performed: a) obtain the gene-specific p-values; b) arrange the p-values in ascending order; c) find the subset of smallest p-values satisfying the inequality, $p_{(i)} \leq FDR(i/N)$, where $p_{(\cdot)}$ are the ordered p-values and FDR is the pre-specified desirable false discovery rate; d) declare all the corresponding genes significant. Performance of the BH/FDR procedure strongly depends on the strength of genomic signal (i.e., AFC), on the number of replicates available for the analysis, and on the power of the test employed for the p-value computation. Figures 5 and 6 illustrate the BH/FDR procedure graphically. In both cases, 100 genes were generated to be truly differentially expressed. In Figure 5, we show

Table 4: True positive fractions (TPF) for two methods of selecting significant genes: t-test (first entries) and bio-weight test (second entries). False Positive Fractions (FPF) fixed at the 5% level. Total number of genes 1,000 with 50 truly differentially expressed.

AFC	Number of arrays								
	2	3	4	5	6	7	8	9	10
1.2	0.042	0.052	0.069	0.073	0.085	0.093	0.101	0.119	0.126
	0.049	0.063	0.081	0.085	0.098	0.107	0.114	0.133	0.138
1.5	0.057	0.100	0.132	0.199	0.228	0.268	0.329	0.350	0.386
	0.105	0.170	0.195	0.270	0.228	0.323	0.380	0.390	0.424
2.0	0.085	0.204	0.286	0.367	0.451	0.497	0.540	0.571	0.609
	0.225	0.366	0.434	0.480	0.528	0.567	0.593	0.612	0.646
2.5	0.099	0.244	0.370	0.466	0.574	0.634	0.638	0.669	0.696
	0.323	0.453	0.536	0.578	0.648	0.685	0.679	0.704	0.724
3.0	0.134	0.288	0.436	0.555	0.645	0.657	0.714	0.728	0.743
	0.421	0.537	0.592	0.655	0.699	0.707	0.748	0.754	0.766
5.0	0.161	0.404	0.631	0.707	0.739	0.778	0.796	0.808	0.820
	0.563	0.673	0.737	0.768	0.790	0.807	0.824	0.827	0.838
10.0	0.245	0.600	0.710	0.769	0.811	0.854	0.857	0.866	0.886
	0.710	0.793	0.808	0.823	0.848	0.875	0.878	0.884	0.897

the ordered p-values and volcano plot for the case of t-testing. In this case, due to low sensitivity of the t-test for small sample size, the number of genes declared significant is small and contamination by false discoveries is high. In addition, because the slant straight line separating the rejection and acceptance regions cleaves the manifold of p-values at a low angle, there is a large uncertainty in the number of the rejected null hypotheses. In the example in Figure 5, 17 genes are declared significant with only 12 of them being true discoveries. Because the acceptance-rejection borderline is blurred, the standard error of the number of true discoveries is fairly large, as high as 8. A drastically different picture is presented in Figure 6 where the results of computations with the bio-weight test are shown (see Appendix for technical details of the p-value computation). High sensitivity of the bio-weight testing results in 89 genes declared differentially expressed with 74 of them being true discoveries. The separation between the rejection and acceptance regions is well defined, and as a result, the standard error of the number of discoveries is as low as 4.6. Table 5 summarizes all the findings of this kind for multiple sample sizes and fold changes. Upon examination of the results presented in this Table, we may draw the conclusion that for small sample sizes, the bio-weight test is vastly superior to the *t*-test. Even for a comparatively large sample size, such as 20, there is still some advantage of using the bio-weight test rather than the *t*-test.

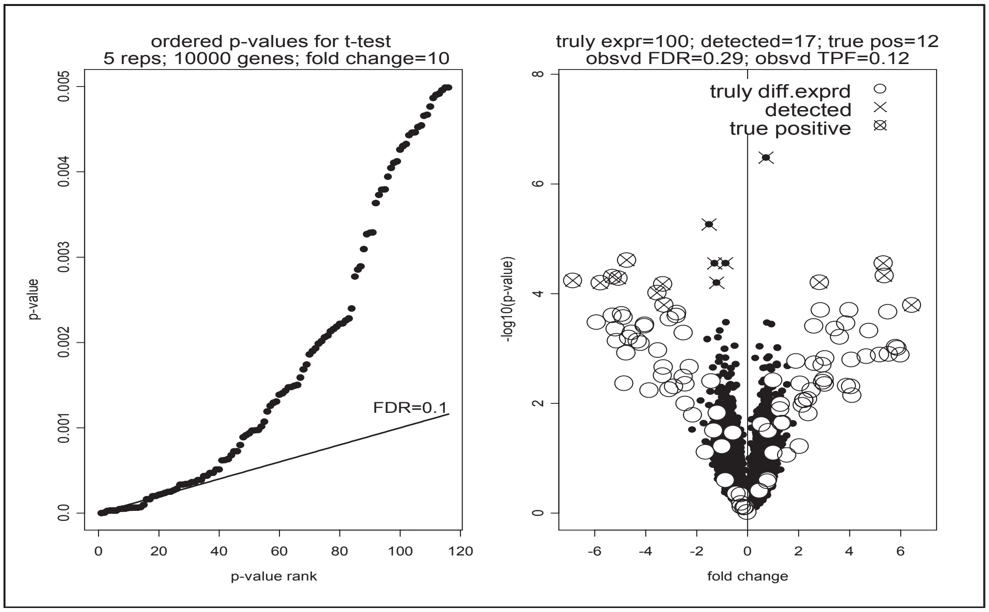

Figure 5: Performance of BH/FDR using *t*-test. Left panel: ordered p-values; right panel: truly differentially expressed genes (open circles), BH/FDR discoveries (crosses), and true discoveries (crossed circles)

5. Discussion

Apart from purely statistical considerations, we may also suggest some subject matter arguments in favor of using the bio-weight as a measure of significance of differentially expressed genes. These arguments come from the consideration of validation. Any instrument used for the purpose of validation always has its own limitations. For example, the genes that are declared highly significant in a microarray experiment may turn out to be intractable for an independent validation in a different experimental setting due to a very small over- or under-expression. On the other hand, the validation instrument may not have certain drawbacks inherent in the microarray technology, for example gene-to-gene variations of the binding affinity (Held *et al.*, 2003), thus being able to reveal the features not observed in the microarray setting. Currently, the most common experimental methodology for validating the microarray data is the quantitative reverse

Table 5: Numbers of true discoveries vs. sample size and fold change. Total number of genes 10,000 with 100 of truly differentially expressed. Two methods of selecting significant genes: *t*-test (first entries) and bio-weight test (second entries).

# replicates	Fold change							
	2		3		4		5	
	#true	st.err	#true	st.err	#true	st.err	#true	st.err
2	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0
	0.00	0.0	1.00	0.9	4.80	2.8	16.3	7.6
3	0.00	0.0	0.00	0.0	0.00	0.0	0.10	0.3
	0.50	0.9 1	3.3	5.2 2	9.5	6.8	41.0	6.6
4	0.00	0.0	0.10	0.3	0.10	0.3	0.30	0.7
	0.80	1.1	32.9	4.9	48.0	4.6	52.9	4.5
5	0.00	0.0	0.80	1.3	1.50	2.9	3.90	4.0
	5.10	4.2	41.1	4.6	52.5	5.4	61.9	4.6
6	0.00	0.0	3.10	2.5	14.2	11.	29.7	6.7
1	4.4	4.7	47.0	5.4	61.7	6.2	66.0	2.9
7	0.70	0.8	11.4	5.8	37.2	5.5	46.1	8.6
	28.8	5.9	52.8	4.8	61.8	6.0	69.1	6.5
8	1.40	2.0	31.2	8.4	45.8	6.1	56.4	3.5
	30.7	5.2	56.3	6.0	65.8	5.5	71.9	4.1
9	5.40	4.7	37.4	6.6	56.5	3.3	64.6	5.2
	33.4	4.8	58.3	4.7	67.5	3.9	74.5	3.8
10	14.3	5.5	48.2	3.6	64.7	5.4	65.3	4.5
	38.0	3.3	61.5	2.7	74.0	4.1	73.7	4.8
20 5	7.0	6.4	70.1	6.5	77.6	5.0	79.3	5.3
	64.0	4.3	73.3	5.6	82.3	4.0	82.3	4.6

transcriptase polymerase chain reaction (q-RT-PCR) (Rajeevan, 2001). A major limitation of the q-RT-PCR is the necessity to keep the amplification curve within the exponential growth limits (Tichopad *et al.*, 2003). These limits, as well as the PCR amplification efficiency, are generally unknown and often require sophisticated algorithms and additional calibration efforts to reduce the errors in quantitation (Meijerink *et al.*, 2001). Importantly, the genes with relative abundances too close to 1 are not reliably discernable by q-RT-PCR. Therefore, the genes declared significant solely because of smallness of their t-test p-values in the microarray setting have little chance to be independently validated due to too low over- or under-expression. As a result, an experimental biologist faces a difficult choice between the sole reliance on the microarray data on one hand, and relaxing the requirements of statistical significance in exchange for higher probability of independent validation on the other hand. The bio-weight criterion attempts to reconcile these conflicting requirements and detect those genes which are the most likely candidates for follow up validation. These intuitive considerations are supported by direct numerical simulation. A simple stochastic model for simulating PCR amplification and subsequent quantitation of the initial (i.e., before amplification) log ratios has been suggested by the author in (Rosenfeld *et al.*, 2004.) This model is constucted in such a way as to produce realistic behavior of the amplification curves and PCR amplification efficiency. Numerical simulation shows that for small number of arrays, there is a greater overlap between the PCR and bio-weight predictions in comparison to that of the PCR and t-test.

The method based on estimating equations (2.3) may be viewed as an alternative method for fitting the RVM. In a sense, it is complimentary to that proposed in (Wright and Simon, 2003) where model fitting was based on the distributional properties of σ_j^2 . Although the cautionary notes in (Wright and Simon, 2003) toward utilization of the method of moments for fitting RVM are quite reasonable, we found that the estimates (2.3) are highly reliable and literally indistinguishable from those produced by the maximum likelihood method. In addition, they provide an additional insight into the model structure and simple rules for controlling the simulation process.

As mentioned above, all the simulations summarized in this paper have been performed for the case of “weak fluctuations.” We believe that it is the most important case from the practical standpoint. However, we have also performed limited simulations to explore the cases of “moderate” and “strong” fluctuations, i.e., the cases where the standard error of the gene-to-gene variability of σ_j^2 is comparable or greater than its expectation. A surprising result yet to be properly understood is that sensitivity of the bio-weight test statistic to the presence of differentially expressed genes remains generally intact. A preliminary explanation is that the misclassification error caused by an additional source of noise moves significant and insignificant genes in both directions across the borderline separating the rejection and acceptance regions, thus approximately compensating for their totals. A detailed account of these analyses is planned to be published in future.

6. Summary

We have suggested a modification of the classical t-test specifically designed to enhance sensitivity of detection of differentially expressed genes in microarray data analysis. This modification, termed as "bio-weight," is shown to have higher statistical power compared to the standard t-test in the microarray setting. Application of the bio-weight is particularly useful in the situations when the number of microarrays available for the analysis is below a dozen. It is shown that the bio-weight testing is more robust with respect to deviation of samples from normality compared to the standard t-test. ROC analysis shows that for a fixed false positive rate, the bio-weight approach offers a noticeably higher true positive rate than that of the standard t-test. It is also demonstrated that application of the bio-weight testing allows one to substantially improve performance of the Benjamini-Hochberg False Discovery Rate procedure for detection of the differentially expressed genes.

Acknowledgments

The author expresses his gratitude to Drs P.Prorok, D.Corle and P.Pinsky for useful discussions of the statistical issues and valuable help in refining the text of the paper.

Appendix. Computation of p-values for the Bbio-weight Test Statistic

It does not seem possible to analytically derive distribution of the bio-weight test statistic, $\hat{b}$; however, it is not a difficult task for computational methods. Examination of the distributional shapes of $\hat{b}$ for various sample sizes suggests that they may be closely approximated by the gamma distribution upon appropriate selection of the parameters. We apply a computational approach in which we estimate these parameters by minimizing the mean least square differences between exact quantiles of the bio-weight and theoretical quantiles of the gamma distributions. Although any set of quantiles may be used in our procedure, we are especially concerned with precise representation of the distributional tails, i.e., the domain where the cumulative distribution function is greater than, say, 80%. The so called "smooth nonlinear local minimizer subject to bound constrained parameters" (*nlminb*) of S-PLUS is used for parameter fitting (Gay, 1983). In these computations, the sample size (number of replicates) varies from 2 to 10, and the shape parameter, α, in the parental gamma distribution varies from 2 to 15, thus covering the entire range of weak fluctuations. The results of these computations are the shape, A , and rate, Θ^{-1}, of the gamma distribution best representing the tails of the bio-weight distribution. Table 6 shows parameters

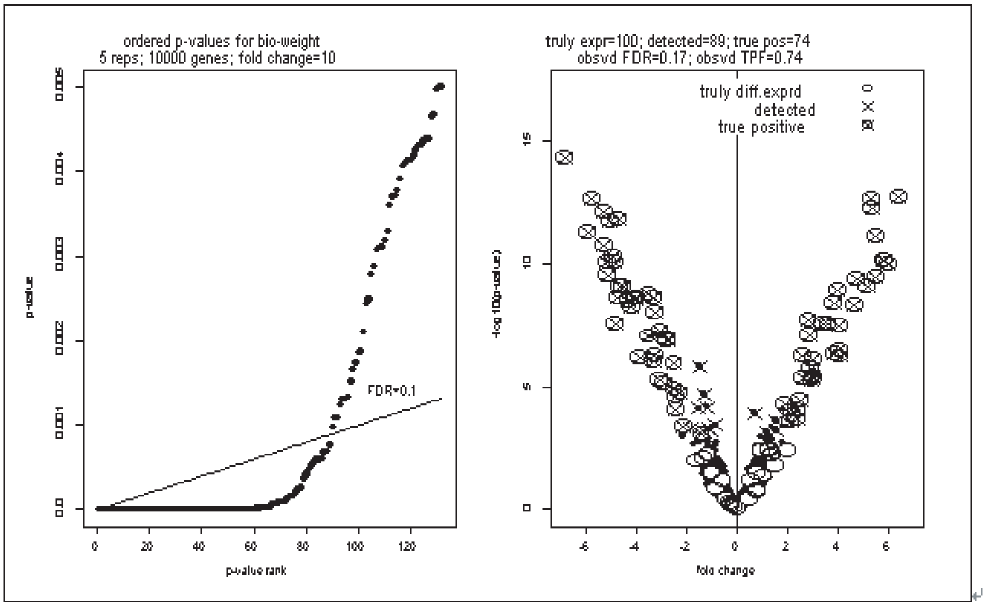

Figure 6: Performance of BH/FDR using the bio-weight. Left panel: ordered p-values; right panel: truly differentially expressed genes (open circles); BH/FDR discoveries (crosses) and true discoveries (crossed circles)

Table 6: Parameters shape, $A(n, \alpha)$, and rate, $\Theta^{-1}(n, \alpha)$ of the gamma distribution approximating the tail distribution of bio-weight.

α	$n = 2$		$n = 10$	
	shape	rate	shape	rate
2	0.164	0.57	0.144	0.89
3	0.266	0.87	0.194	1.19
4	0.307	0.99	0.215	1.31
5	0.325	1.05	0.225	1.38
6	0.336	1.08	0.231	1.42
7	0.343	1.10	0.236	1.45
8	0.349	1.12	0.241	1.47
9	0.355	1.14	0.244	1.50
10	0.361	1.16	0.247	1.53
11	0.366	1.18	0.251	1.55
12	0.372	1.20	0.255	1.58
13	0.378	1.22	0.259	1.60
14	0.383	1.24	0.263	1.63
15	0.389	1.26	0.266	1.66

$A(n, \alpha)$ and $\Theta(n, \alpha)$ as the functions of α for $n = 2$ and $n = 10$. It has been found that the dependencies $A(n, \alpha)$ and $\Theta(n, \alpha)$ are perfectly linear in n; hence, to save page space, we provide these parameters only for $n = 2$ and $n = 10$. Table 6 is intended to be used as follows. If the task is simulation then n and α serve as input parameters, therefore the approximating gamma distribution for representing the $\hat{b}$statistic is $\gamma(x|A(n, \alpha), \Theta(n, \alpha))$. If the task is the analysis of empirical data, then we first estimate $\hat{\alpha}$ according to the first equation (2.3) and then use Table 6 for selecting an appropriate $\gamma(x|A(n, \alpha), \Theta(n, \alpha))$. Figure 7 shows an example of the histograms and QQ-plots comparing quantiles of the empirical distribution of bio-weight to the theoretical quantiles of the gamma distribution fitted using the above described procedure. As seen from this figure, gamma distribution provides a very good approximation, especially on the tail, what is particularly important for accurate computation of p-values.

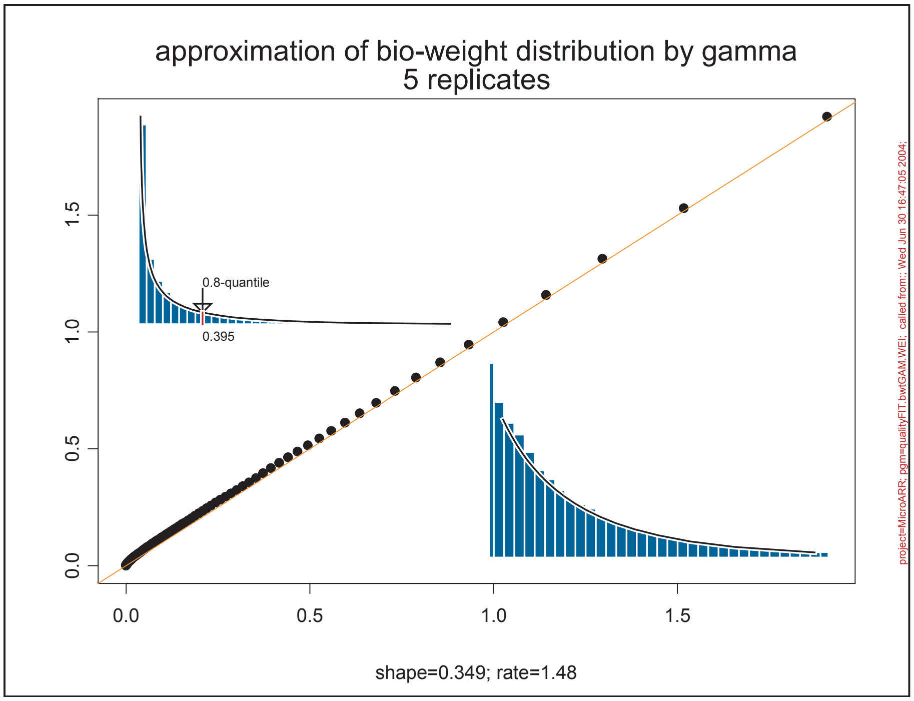

Figure 7: Histogram of the bio-weight distribution. QQ-plot compares empirical and theoretical quantiles of bio-weight distributions

References

Armitage, P., Colton, T, ed. (1998). *Encyclopedia of Biostatistics.* Wiley.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc., B* **57**, 289-300.

Benjamilni, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics **29**, 1165-1188.

Donoho, D. L. (2000). High-dimensional data analysis: the curses and blessings of dimensionality. Aide-Memoir Report, Department of Statistics, Stanford University.

Dudoit, S., Shaffer, J. and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **49**, 71-103.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Asso.* **96**, 1151-1160.

Gay, D. M. (1983). Algorithm 611: Subroutines for unconstrained minimization using a model/trust region approach. *ACM Transactions on Mathematical Software* **9**, 503 524.

Held, G., Grinstein, G. and Tu, Y. (2003). Modeling of DNA microarray data by using physical properties of hybridization. *Proc. National Academy of Science* **100**, 7575-7580.

Hsueh, H., Chen, J., Kodell, R. (2003). Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J. Biopharm. Statist.* **13**, 675-689.

Kerr, M., Martin, M., Churchhill, G. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**, 819-837.

Meijerink, J., Mandigers, C., van de Locht, L., Tonissen, E., Goodsaid, F. and Raemaekers, J. (2001). A novel method to compensate for different amplification efficiencies between patient DNA samples in quantitative real-time PCR. *J. Mol. Diagn.* **3**, 55-61.

Rajeevan, M., Ranamukhaarachchi, D., Vernon, S., Unger, E. R. (2001). Use of real-time quantitative pcr to validate the results of cdna array and differential display pcr. *Technologies Method* **25**, 443-451.

Rosenfeld, S., Wang, T., Kim, Y. and Milner, J. (2004). Numerical deconvolution of microarray signal. Simulation study. *Annals of the New York Academy of Sciences* **1020**, 110-123.

Rosenfeld, S. (2004). Performance of benjamini-hochberg false discovery rate for small sets of cDNA microarrays. *Proceedings of the Interface 2004*, Baltimore, May 26-29.

Simon, R., Korn, E., McShane, L., Radmacher, M., Wright, G., Zhao, Y. (2004). *Design and Analysis of DNA Microarray Investigations.* Springer.

Tichopad, A., Dilger, M., Schwarz, G. and Pfaffi, M. (2003). Standardized determination of real-time PCR efficiency from a single reaction set-up. *Nucleic Acids Res.* **31**, e122.

Wright, G. and Simon, R. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448-2455.

Received January 17, 2006; accepted February 26, 2006.

Simon Rosenfeld
DHHS/National Cancer Institute
EPN 3108, 6130 Executive Blvd
Rockville, MD, 20904, USA
sr212a@nih.gov