Re-weighted ODP for differential gene expression analysis

Hiroto Nomura*, Shigeyuki Oba*, and Shin Ishii**.

*: *NARA Institute of Science and Technology Graduate school of information science*
8916-5, Takayama, Ikoma, Nara, JAPAN 630-0192
(Tel : 81-743-72-5986; Fax : 81-743-72-5989) (Email h-nomura@is.naist.jp)
**: *Graduate school of Informatics. Department of systems science Graduate school of informatics.*
Kyoto University
Gokasho, Uji, Kyoto, JAPAN 611-0011

Abstract: Statistical hypothesis testing is often applied to detection of differentially expressed (DE) genes whose expression levels are different between two or more different conditions of samples, where measurement is done by, for example, DNA microarrays. In any statistical testing, selection of an appropriate "statistic", a significance measure, is important for increasing the detection performance. Storey *et al.* (2005) previously proposed a statistic, optimal discovery procedure (ODP), based on the Neyman-Pearson lemma, and proved that the ODP is the most powerful statistic in multiple testing situation in which multiple hypothesis testings are performed simultaneously, as done in analyses of DNA microarray data.

In order to exploit the power of the ODP, we need the binary label which indicates the true condition of each gene, DE or non-DE, which is unknown in practical situations. Then, we need to estimate the label based on a usual individual statistical test, such as t-test.

To boost up the power of the practical ODP, we propose in this study a new method "re-weighted ODP (RODP)" which calculates the ODP statistic (second ODP) by using the label estimated by another ODP statistic (first ODP). We compare our new statistic with some conventionally-used statistics using artificial data sets, and show that the new one exhibits a stronger power not only than the existing statistics but also a standard usage of the ODP statistic. Furthermore, even in a cancer outlier gene detection problem where DE genes are highly expressed in a part of tumor samples, the new statistic shows in many cases a stronger power than other statistics in particularly designed to detect cancer outlier genes.

Keywords: ODP, multiple testing, statistics, cancer outlier gene detection

I. Introduction

Recent developments in simultaneous measurement techniques of a large number of proteins and genes have enabled us to examine behaviors of molecules in various cellular states. For example, DNA microarrays, which measure expressions of thousands of genes simultaneously, are often used for detecting a set of differentially expressed (DE) genes showing different mean expression levels between two or more different cellular states. When statistical hypothesis testing is used to identify DE genes from observed gene expression data, selection of the test statistic is an important issue. The test statistic is defined as a criterion of significance of genes based on each observation. For example t-statistic, is a generally used statistic which is defined as

$$T_i = \frac{\overline{\mathrm{X}}_{i2} - \overline{\mathrm{X}}_{i1}}{S_i},$$

where $\overline{\mathrm{X}}_{ij}$ is the mean expression level of gene i in samples of class j, and

$$s_i^2 = \frac{(n_1 - 1)s_{i1}^2 + (n_2 - 1)s_{i2}^2}{n_1 + n_2 - 1}$$

is the pooled variance. s_{ij}^2 is the variance of i th gene's expression in the sample's class j. A statistic is said most powerful if the statistic with a certain threshold shows the strongest power, where a test is said most powerful test if it achieves a larger number of true positives than any statistical test with the same number of false positives. Neyman and Pearson (1933) proved that the likelihood ratio statistic, the proportion of the likelihood of a null model to that of an alternative model, is the most powerful statistic in an individual testing, if both of the alternative and null likelihood functions are known. In a multiple testing situation, on the other hand, the likelihood ratio is no longer the most powerful, suggesting there is still some room for improvement in detection power.

II. The ODP statistics

The optimal discovery procedure (ODP) was proposed by Storey *et al.*[1] according to the Neymann-Pearson lemma, which is theoretically the most powerful statistic for the multiple hypothesis testing. The ODP statistic is a ratio of the sum of null likelihoods and the sum of alternative likelihoods, where the null and alternative likelihood functions are evaluated for genes whose null and alternative hypotheses are true, respectively. When there are m genes, the ODP statistic for the i th gene is defined as

$$S_{ODP}(X_i) = \frac{\sum_{j=1}^{m}(1-w_j)g_j(X_i)}{\sum_{j=1}^{m}w_j f_j(X_i)},$$

where X_i is an observation vector of the i th gene, $g_j(\cdot)$ and $f_j(\cdot)$ are the alternative and null likelihood functions corresponding to the j th gene, respectively, and w_j is a weight parameter which takes a binary value; 0 if the alternative hypothesis is true and 1 if the null hypothesis is true for the j th gene. Storey proved that if $g_j(\cdot)$, $f_j(\cdot)$ and w_j are known for all genes, the ODP statistic exhibits the strongest power at any fixed significance level.

In application of the ODP to practical problems, however, there are some difficulties:

1. True values of the weight parameters are unknown
2. Exact shapes of alternative and null distributions are unknown.

Then, Storey proposed the following method for application of the ODP to a realistic situation.

a) Estimate w_j approximately by a (possibly, individual) hypothesis testing using a certain statistic.

b) Estimate parameters of the alternative and null distributions based on the assumption that the both distribution are normal distributions.

If these estimations are far from the truth, however, the ODP statistic does not work well, indicating better estimation of these parameters is important. In this study, then, we focus on the weight estimation problem.

III. Re-weighted ODP statistic

To improve the ODP statistic, weight values were determined by the other estimation of the ODP, which is expected to provide more accurate weight and to lead to more accurate estimation of the ODP. More concrete procedure is as follows:

1) Perform an individual t-test with a p-value threshold $\lambda = 0.05$ to give an initial estimate to w_j for each gene j.

2) Estimate the proportion of null genes

$$\pi_0 = \frac{n_{p>\lambda}}{n\lambda},$$

where $n_{p>\lambda}$ is the number of genes whose p-value is smaller than the threshold λ, and n is the number of all genes.

3) Estimate model parameters of the null and alternative likelihood functions for genes with w_j=1 and w_j=0, respectively.

4) Calculate the ODP statistic (ODPt) by using the likelihood functions estimated above.

5) Assign w_j=0 for top $(n-\pi_0)$ significant genes based on the ODPt statistic, and assign w_j=1 to the other genes.

6) Calculate the re-weighted ODP statistic (RODP) based on the re-estimated weight values in step 5).

IV. Simulation study and comparison to existing statistics

To compare the new statistic with the existing statistics, we applied them to some detection tasks with artificial data sets.

1. Simulation under normal situation

Artificial data set 1 was designed to describe a standard situation. It consists of 500 genes, including 400 non-DE genes and 100 DE genes, i.e., $\pi_0 = 0.8$ and $N = 40$ samples consisting of 20 samples from a control group and 20 samples from a disease group. All gene expression levels were sampled from normal distribution with mean zero and variance 1.0, except that for DE genes in the disease group, the mean was set at $C = 0.3$. We compared the detection power of the RODPt statistic to these of t-statistic, regularized t-statistic, ODPt and correctly weighted ODP statistic (TODP). Note here that the TODP is not applicable in practice, because there is no way to accurately know the true weight.

The regularized t-statistic is a regularized variation of the t-statistic, used in SAM [2], which is defined as

$$T_i = \frac{\overline{X}_{i2} - \overline{X}_{i1}}{s_i + s_0},$$

where s_0 is a regularization constant and set at the lower 10th percentile of pooled standard deviations of all genes. Figure 1 shows receiver operating characteristic (ROC) curves. The horizontal axis denotes the false discovery rate (FDR), i.e., the proportion of non-DE genes in the detected genes, and the vertical axis denotes the true positive ratio, i.e., the proportion of detected genes to the actually DE genes. Each curve is drawn by changing the threshold for the corresponding statistic. The ROC curve approaches the

upper left corner, when the statistic is powerful. RODPt shows a superior performance not only to existing statistics but also to the ODP calculated by the procedure described in the Storey's original paper.

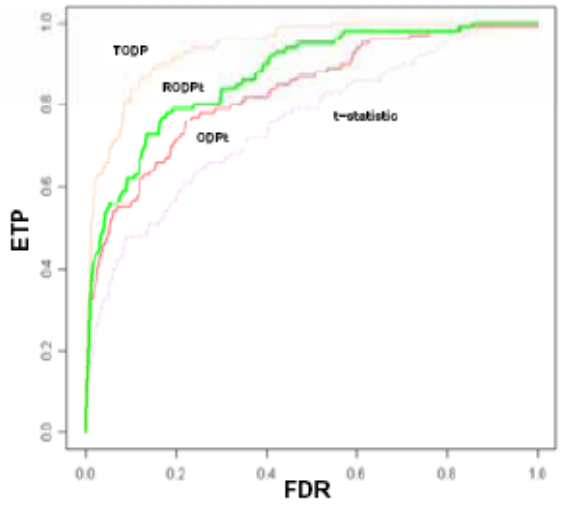


Figure 1 ROC curves of four statistics, (from lower right to upper left) t-statistic, ODPt, RODPt, and TODP applied to the artificial data set 1.
Horizontal and vertical axes denote FDR and ETP, respectively. Regularized-t statistic is omitted because it was worse than the t-statistic.

2. Cancer outlier gene situation

Artificial dataset 2 was designed to examine the detection of so-called "cancer outlier genes" [3] which are differentially expressed only in a part of disease samples. We set the proportion of samples in which DE genes are expressed differentially at $K = 0.3$. The number of genes, the proportion of non-DE genes, and the distribution of expression levels were the same as those in the previous experiment except that the mean expression of DE genes in 6 samples out of 20 samples in the disease group was set at $C = 2.0$.

2.1 Statistics specified for detecting cancer outlier genes

The outlier sum statistic [4] and outlier robust t-statistic (ORT) [5] were proposed in particularly to detect cancer outlier genes. The outlier sum statistic is defined as

$$\mathrm{W}_j = \frac{\sum_{i \in R_j} (x_{ij} - \mathrm{med}_j)}{\mathrm{mad}_j},$$

where R_j is the set of samples in which the j th gene is highly expressed;

$$R_j = \{i \mid x_{ij} > 2q_{75}(\mathrm{x}_j) - q_{25}(\mathrm{x}_j)\}.$$

$q_r(\mathrm{x}_j)$ is the r th percentile of expressions of gene j, med_j is the median of the j th gene's expression levels, and $\mathrm{mad}_j \equiv \mathrm{median}_{i=1,\ldots,n}(|x_{ij} - \mathrm{median}_j|)$ is the mean absolute difference.
On the other hand, ORT is defined as

$$\mathrm{T}_j^* = \frac{\sum_{i \in R_j} (x_{ij} - \mathrm{med}_{1\mathrm{j}})}{\mathrm{mad}_{\mathrm{j}}^{1,2}},$$

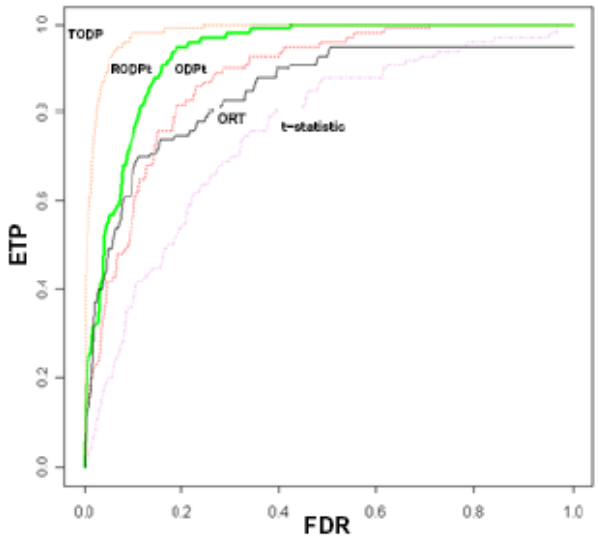
where $\mathrm{med}_{1\mathrm{j}}$ is the median of expression levels of the j th gene in samples in the disease group, and $\mathrm{mad}_{\mathrm{j}}^{1,2}$ is the pooled median of absolute differences in the same group:

$$\mathrm{mad}_{\mathrm{j}}^{1,2} = \mathrm{median}(y_1, \ldots, y_n),$$

where $y_i = \left| x_{ij} - \mathrm{med}_{1\mathrm{j}} \right|$ if the i th sample is in the disease group, or $y_i = \left| x_{ij} - \mathrm{med}_{2\mathrm{j}} \right|$, otherwise.

2.2 Result

In figure 2 (upper), we compare RODP with the other statistics, when applied to artificial data set 2. RODPt shows the most powerful performance among the statistics with $C = 2.0$.
In figure 2 (lower), we set the number of differentially expressed samples at 4, i.e. $K = 0.2$. Area under the ROC curve (AUC) of t-statistic shows low value of 0.683, whereas RODPt shows comparable AUC value to ORT, actually the best performance among practically available statistics.



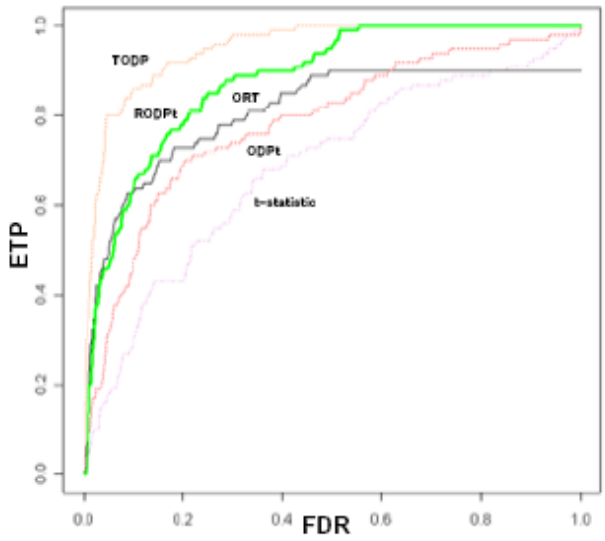


Figure 2 ROC curve of each statistics applied to the cancer outlier gene dection (upper $C = 2$, lower $K = 1$) Both figures show TODP, RODPt, ODPt, ORT, and t-statistic.
Outlier sum statistic is not shown because its power was lower than ORT.

We also examined various settings, and found that in many cases, RODPt shows better performance than the outlier sum statistic and ORT. In addition, we have found the following tendencies:

- As the number of samples increases, RODPt exhibits good performance in comparison to the others.
- If the mean expression difference in a DE gene is small, e.g., $C < 0.1$, and/or the proportion of samples in which the gene is differentially expressed is small, e.g., $K < 0.2$, RODPt does not perform well. Because the situation is so difficult that the detection power of ODPt is almost at the chance level, it is difficult for ODP employing it to estimate the weight parameter to attain high detection performance.
- When the number of non-DE genes is as small as $\pi_0 < 0.1$ ($N = 40$), or as large as $\pi_0 > 0.9$ ($N = 40$), the RODPt no longer exhibits better performance than the others, because the ODP statistic basically boosts its power by utilizing commonality within multiple tests.

V. Discussion

In this study, we proposed a re-weighted ODP estimator, an ODP with better estimation of weight parameters, for applications to bioinformatics studies. When applied to practical problems, we showed that re-weighted ODP estimator improved the detection power of significant genes. In a cancer outlier gene detection problem, to which some specialized statistics were formerly proposed, our RODPt showed a stronger detection power.
In a difficult situation in which the t-statistic performed almost at the chance level, on the other hand, our RODP did not perform well; to devise a method to deal with such a case remains as a future study.

Acknowledgements

This work was supported by a Grant-in-Aid for Young Scientists 19710172 from the MEXT.

REFERENCES

[1] Storey JD, Dai JY, Leek JT (2005), **The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments.** Biostatistics 2007 8(2):414-432.
[2] Tusher V, Tibshirani R, Chu Goss GV (2001), **Significance analysis of microarrays applied to the ionizing radiation response.** PNAS 2001 98 5116-5121.
[3] Tomlins SA, Rhodes DR, Perner S, et al. (2005), **Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer.** Science 310, 644-8
[4] Tibshirani R, Hastie T, (2006), **Outlier sums for differential gene expression analysis.** Biostatistic 2007 8(1):2-8
[5] Baolin Wu (2006), **Cancer outlier differential gene expression detection.** Biostatistics 2007 8(3):566-575