BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net **Prediction Model**

Volume 7(1)

Adaptive thresholds to detect differentially expressed genes in microarray data

Yutaka Fukuoka[1]*, Hidenori Inaoka[2], Makoto Noshiro[2]

[1]Department of Biosystems Modeling, Graduate School of Biomedical Science, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo, Tokyo 113-8510, Japan; [2]School of Allied Health Sciences, Kitasato University, Kanagawa, Japan; Yutaka Fukuoka - Email: fukuoka.bsm@tmd.ac.jp; Phone: +81-3-5803-4777; Fax: +81-3-5803-4777; *Corresponding author

Received August 04, 2011; Accepted August 04, 2011; Published August 20, 2011

Abstract:
To detect changes in gene expression data from microarrays, a fixed threshold for fold difference is used widely. However, it is not always guaranteed that a threshold value which is appropriate for highly expressed genes is suitable for lowly expressed genes. In this study, aiming at detecting truly differentially expressed genes from a wide expression range, we proposed an adaptive threshold method (AT). The adaptive thresholds, which have different values for different expression levels, are calculated based on two measurements under the same condition. The sensitivity, specificity and false discovery rate (FDR) of AT were investigated by simulations. The sensitivity and specificity under various noise conditions were greater than 89.7% and 99.32%, respectively. The FDR was smaller than 0.27. These results demonstrated the reliability of the method.

Keywords: microarray, fold difference, threshold, differentially expressed genes, false discovery rate (FDR).

Background:

To detect changes in gene expression data from microarrays, a threshold for fold difference has been used widely **[1-5]**. In this approach, a small threshold value may cause many false positives and this makes interpretation of results difficult. Although dividing a number by a small number could result in a large fold difference by chance, some truly differentially expressed genes (DEGs) could be missed with a large threshold value. In this way, it is not always guaranteed that a value which is appropriate for highly expressed genes is suitable for lowly expressed genes **[6]**. Some researchers have addressed this problem. Rocke and Durbin have developed a model for measurement error in expression data as a function of the expression level **[7]**. They divide the total noise into additive and proportional components and employ replicated measurements to model the noises. They also apply the model to compare expressions between conditions, but the expressions are divided into only two levels. Colantuoni *et al*. (2002) proposed a method for local variance correction, in which a standard deviation (s.d.) was calculated locally across expression levels **[8]**. Loots *et al*. (2006) developed a similar method, in which fold differences between signal and control are binned into groups according to the expression level and local variation is calculated for each group **[9]**. These methods employ fold difference between control and signal. Consequently, it is not always guaranteed that these can model the additive and proportional noises correctly. There are some other methods which estimate variation in duplicated/replicated data. These are similar to the method proposed here and are described later. In this study, aiming at detecting true DEGs from a wide expression range, we propose adaptive thresholds for fold difference. It employs two measurements under the same condition to model the total error and divides the data into some bins according to the expression level. Based on local variance, upper and lower thresholds are calculated in each bin. The minimum requirement of the method is two measurements under the same condition. The method is designed to detect DEGs from small size data, in which there is a trade-off between suppressing false positives and achieving perfect DEG detection. This paper focuses on the former, namely a low false discovery rate (FDR), because it makes interpretation of results easier.

Methods which employ thresholds based on variations in duplicates/replicates have been proposed. Tsien *et al*. developed a method for evidence-based noise reduction **[10]**. In this method, a threshold is calculated to define a region of noise inherent in the data. They assume that corresponding pairs should have fold differences of 1.0. The method requires duplicates, in which the operating conditions, etc., are exactly controlled to be identical. Determination of the threshold involves segmental calculation of the average and s.d. in each segment. Then a candidate border point is determined as (average) + (a constant value) X (s.d.). Based on the candidates, a best-fit line/curve is calculated to define the border of the identity region of insignificant fold changes. Draghici et al. proposed a similar method, named a noise sampling method (NS) **[6, 11]**, based on an analysis of variance approach **[12, 13]**. The method employs replicate spots to estimate a distribution of noise. The measured log ratio, log R(i, s), for gene i and spot s is modeled as log $R(i, s)=\mu+G(i)+\varepsilon(i, s)$, where μ is the average log ratio over the whole microarray, G (i) is a term for differential regulation of i, and ε (i, s) is a noise term. Based on the equation, we can calculate an empirical distribution of the noise. In order to detect DEGs at a given confidence level, the deviation from the mean of the distribution is calculated. Bootstrapping is used to map the confidence level from the noise distribution to the log ratio of expression.

Methodology:

Algorithm:

In the proposed method, thresholds which have different values for different expression levels are calculated before comparing expressions from different conditions. The adaptive threshold method (AT) requires two measurements under the same condition to evaluate local variations. First, the data are normalized by the median. Second, a ratio of the expression in the first measurement to the second is calculated for each gene. The ratios are plotted against the logtransformed expression levels in the first measurement. Then, the data are divided into bins, whose width is d (d=0.2 in this study). In each bin, 50 genes are randomly selected and the maximum and minimum ratios are determined. This process is repeated 50 times. The upper and lower thresholds are determined based on a confidence interval (CI) of the population mean of the 50 maximum and minimum ratios in the bin, respectively. The lower threshold is calculated as (the lower limit of the CI of the minimum ratios)/g, where g is a constant. The upper threshold is (the upper limit of the CI of the maximum ratios)Xg. Thus false positives are expected to be suppressed as g increases. The thresholds are used to compare the control to signal data. If a fold difference between the control and signal is smaller/larger than the lower/upper threshold, the gene is considered down-/up-regulated.

Simulations:

The sensitivity, specificity and FDR of AT were investigated. The performances of NS were also calculated using the same data. The simulation data were constructed as follows. The average of the 13 measurements from normal aged hippocampus in GSE5281 **[14]** (http://www.ncbi.nlm.nih.gov/geo) was used as the true value for the control data, which is denoted as c. Each measurement included 48403 probes. For the sake of simplicity, we assumed that a measurement included expressions of 48403 genes. Both additive and proportional noises were employed in the simulation. A value ep from N (0, vp) was added to 1 and that was used as the proportional noise. The additive noise, ea, was generated from N (0, va). The expression value with the noises was thus denoted as c (1+ ep) + ea. Among the 48403 probes, 500 were randomly selected as up-regulated genes and another 500 as down-regulated. The effects of up-/down-regulation were represented with random values from N (4, 0.8)/N (-4, 0.8), respectively. After the true values, c, of the regulated genes were natural logarithmical transformed, the random values were added to the transformed values. Then, the noises were applied in the same manner to the control data. The first simulation investigated the effect of noises on the performance. In AT g=2 was used while the confidence level in NS was 99.5%. The variances of noises, (va, vp), examined were (0.01, 0), (0.01, 0.05), (0.01, 0.1), (1, 0), (1, 0.05), (1, 0.1), (10, 0), (10, 0.05), (10, 0.1), (20, 0), (20, 0.05), (20, 0.1), (50, 0), (50, 0.05) and (50, 0.1). The adaptive thresholds were calculated based on two control measurements. In NS, the noise distribution was obtained using the same data. Ten different realizations of the regulated data were generated. These steps were repeated 10 times with different sets of control data and thus 100 trials were performed in total. The sensitivity, specificity and FDR were evaluated under each noise condition. The definition of FDR is FDR=E [F/ (F+T)], where F and T are the numbers of the false and true positives and E [] denotes an expected value **[15]**. The second simulation compared receiver operating characteristic (ROC) curves among AT, NS and the fixed threshold. We investigated g between 1 and 7 in AT, confidence levels between 99.0 and 99.98% in NS and fixed thresholds between 2 and 20. The variances va and vp were fixed at 20 and 0.05, respectively, because this combination best reproduced the average and s.d. in the bins of GSE5281.

Results and Discussion:

A typical distribution of the ratios between two control measurements (va=20, vp=0.05) is shown in Figure 1a. The figure also shows the upper and lower adaptive thresholds (solid lines) and the confidence levels in NS (dashed lines). A dot represents the ratio of a gene and the vertical, dashed lines indicate the bins. Less than 0.5% of the 48403 genes were greater than the upper adaptive threshold while almost no genes were smaller than the lower boundaries of AT and NS. The upper confidence level of NS was smaller than the upper adaptive threshold. Accordingly, more genes were greater than the upper confidence level. The relationship between the expression level and the number of false detections was investigated (**Figure 2**). There were many false positives in a low expression range (<0). The upper adaptive threshold steeply changed around this range (**Figure 1**) and this contributed for suppressing false positives. The upper confidence level in NS also changed in this range. However, it was smaller than the upper adaptive threshold and therefore, more false detections occurred. This explains a larger FDR with NS. **Figure 1b** compares the shapes of the adaptive thresholds obtained under (0.01, 0), (0.01, 0.1), (20, 0) and (20, 0.1). When the noises were (0.01, 0), the thresholds were almost constant for all expression levels. Without the noises, g=2 corresponds to a fixed threshold of 2 whereas the upper and lower confidence levels in NS were 1. The thresholds under (20, 0) indicate that the additive noise had minor influence on highly expressed genes. The influence of the proportional noise appeared to be large for all levels because the upper and lower thresholds were similar under (0.01, 0.1) and (20, 0.1). In this way, the shape of the adaptive thresholds varies according to va and vp, demonstrating that the method can model both additive and proportional noises.

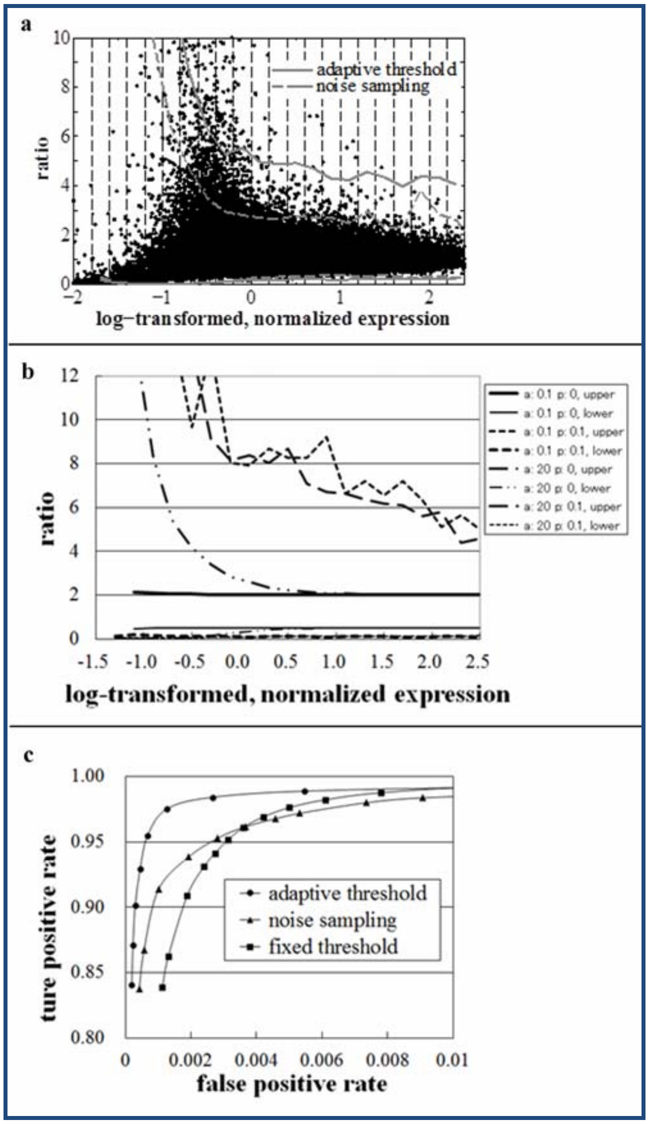

Figure 1: Simulation results with g=2 in the adaptive threshold method and the confidence level of 99.5% in the noise sampling method. a) An example of distribution of the ratio between the two control measurements and the upper and lower adaptive thresholds (solid lines). The upper and lower confidence levels in the noise sampling method were also displayed (dashed lines). The variances of the additive and proportional noises were 0.05 and 20, respectively. A dot represents a ratio of the two expression values of a gene. The bins used to calculate the adaptive thresholds are illustrated by the dashed lines. Only few genes were greater/lower than the upper/lower adaptive thresholds. b) The upper and lower thresholds obtained with different noise conditions: (0.01, 0), (0.01, 0.1), (20, 0) and (20, 0.1). c) The ROC curves for the three methods. The horizontal and vertical axes represent the false and true positive rates, respectively. The false positive rate equals to 1-(specificity), while the true positive rate is equivalent with the sensitivity.

Table 1 (see Supplementary material) summarizes relationships between the noise variances and the performances. The numbers represent the average±s.d. over the 100 trials. The sensitivity became lower for larger noises and the sensitivities of AT and NS were comparable to each other (**Table 1a**). The influence of the noises was weaker for the specificity and it was greater than 99.3% for AT while it was about 99% for NS (**Table 1b**). The specificity greater than 99.3% leads to a smaller FDR in AT (**Table 1c**). **Figure 1c**, which compares the ROC curves, indicates that AT achieved the best performance among the three. AT was developed by expanding a fixed threshold so that local variations are considered in calculation of thresholds. This is the reason for employing a parameter g, which makes it easier to adjust the width between the upper and lower thresholds. A wider width is a key feature to suppress false positives. Tsien *et al*., (2002) proposed a similar threshold method, in which a certain noise distribution is assumed **[10]**. In contrast, AT and NS assume no distribution and accordingly, these are more flexible.

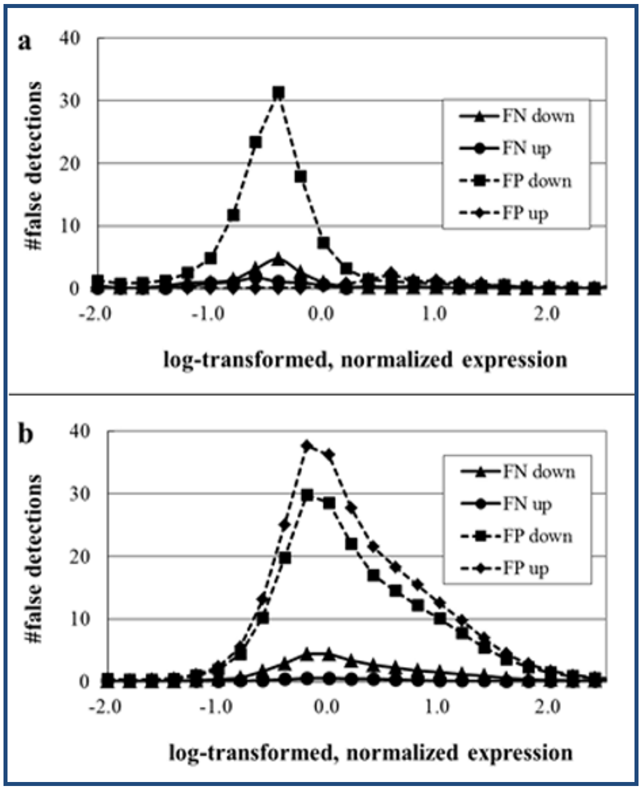

Figure 2: The number of false detections for different expression levels: a) the adaptive threshold method and b) the noise sampling method. The horizontal axis represents the log-transformed, normalized expression level while the vertical axis

indicates the number of false detections. Each mark represents the number of false positives (FP) and negatives (FN) in the bins shown in Figure 1. FN down/up represents the number of FN for down-/up-regulated genes.

Conclusion:

In this study, aiming at detecting true DEGs from a wide expression range, we proposed an adaptive threshold method. Simulations were conducted to investigate the performance of the method. The sensitivity and specificity under various noise conditions were greater than 89.7% and 99.3%, respectively. The method achieved a low FDR, indicating that it can suppress false positives and make the interpretation of results easier. The minimum requirement of the method is two measurements, under the same condition, and it is applicable to small size data.

Acknowledgments:

This study was in part supported by a grant from Kitasato University School of Allied Health Sciences (No. 2007-10014) to H. Inaoka and a scientific research grant from the Ministry of Education, Culture, Sports, Science and Technology of Japan (#20510184) to Y. Fukuoka.

References:

[1] DeRisi JL *et al. Science* 1997 **278**: 680 [PMID: 9381177]

[2] Moreno-Bueno G *et al. Cancer Res.* 2003 **63**: 5697 [PMID: 14522886]

[3] Welle S *et al. PLoS One.* 2008 **3**: e1385 [PMID: 18167544]

[4] McGrath MF & de Bold AJ. *BMC Genomics.* 2009 **10**: 254 [PMID: 19486520]

[5] Shi WD *et al. Cancer Lett.* 2009 **283**: 84 [PMID: 19375852]

[6] Draghici S. *Drug Discov Today.* 2002 **7**: S55 [PMID: 12047881]

[7] Rocke DM & Durbin B. *J Comput Biol.* 2001 **8**: 557 [PMID: 11747612]

[8] Colantuoni C *et al. Bioinformatics* 2002 **18**: 1540 [PMID: 12424128]

[9] Loots GG *et al. BMC Bioinformatics* 2006 **7**: 307 [PMID: 16780584]

[10] Tsien CL *et al. Proc AMIA Symp.* 2002 **810**: 4 [PMID: 12463937]

[11] Draghici S *et al. Bioinformatics* 2003 **19**: 1348 [PMID: 12874046]

[12] Kerr MK *et al. J Comput Biol.* 2000 **7**: 819 [PMID: 11382364]

[13] Kerr MK & Churchill GA. *Proc Natl Acad Sci U S A.* 2001 **98**: 8961 [PMID: 11470909]

[14] Liang WS *et al. Physiol Genomics.* 2007 **28**: 311 [PMID: 17077275]

[15] Storey JD & Tibshirani R. *Proc Natl Acad Sci U S A.* 2003 **100**: 9440 [PMID: 12883005]

Edited by P Kangueane
Citation: Fukuoka *et al.* Bioinformation 7(1): 33-37 (2011)

Supplementary material:

Table 1: The sensitivity, specificity and FDR by the adaptive threshold method (g=2) and the noise sampling method (confidence level 99.5%).

a) Sensitivity (%)

v_a	adaptive threshold			noise sampling		
	v_p			v_p		
	0	0.05	0.1	0	0.05	0.1
0.01	99.97±0.06	99.60±0.21	96.73±0.68	100.0±0.0	99.93±0.09	98.54±0.45
1	99.99±0.30	99.98±0.22	97.0±0.63	100.0±0.0	99.90±0.08	98.57±0.41
10	99.93±0.01	97.89±0.57	95.75±0.86	99.18±0.32	98.87±0.39	95.58±0.78
20	98.67±0.33	97.52±0.45	93.29±0.93	98.11±0.51	96.76±0.62	91.73±1.44
50	96.52±0.52	94.21±0.79	89.70±0.84	94.72±0.83	92.02±0.93	86.16±1.34

b)Specificity (%)

v_a	adaptive threshold			noise sampling		
	v_p			v_p		
	0	0.05	0.1	0	0.05	0.1
0.01	100.0±0.0	99.94±0.02	99.60±0.03	98.24±1.33	99.11±0.06	99.09±0.04
1	100.0±0.0	99.94±0.01	99.58±0.03	99.05±0.06	99.13±0.06	99.10±0.05
10	100.0±0.0	99.82±0.02	99.45±0.03	99.02±0.06	99.02±0.08	99.08±0.05
20	99.84±0.04	99.75±0.03	99.39±0.05	99.08±0.07	99.08±0.07	99.11±0.07
50	99.60±0.05	99.56±0.09	99.32±0.02	99.05±0.08	99.11±0.07	99.10±0.07

c) FDR

v_a	adaptive threshold			noise sampling		
	v_p			v_p		
	0	0.05	0.1	0	0.05	0.1
0.01	0.0±0.0	0.03±0.01	0.16±0.01	0.39±0.21	0.30±0.01	0.31±0.01
1	0.0±0.0	0.03±0.01	0.17±0.01	0.31±0.01	0.29±0.01	0.30±0.01
10	0.03±0.01	0.09±0.01	0.21±0.01	0.32±0.01	0.32±0.02	0.31±0.01
20	0.07±0.02	0.11±0.01	0.24±0.02	0.31±0.02	0.29±0.02	0.32±0.01
50	0.17±0.02	0.18±0.03	0.27±0.01	0.32±0.02	0.32±0.02	0.33±0.02