遗 传 学 报 *Acta Genetica Sinica*, December 2006, 33 (12)：1132–1140 ISSN 0379−4172

Powers of Multiple-Testing Procedures for Identification of Genes Significantly Differentially Expressed in Microarray Experiments

TAN Yuan-De, YAN Heng-Mei①

College of Life Science, Hunan Normal University, Changsha 410081, *China*

Abstract: Because of the high operation costs involved in microarray experiments, the determination of the number of replicates required to detect a gene significantly differentially expressed in a given multiple-testing procedure is of considerable significance. Calculation of power/replicate numbers required in multiple-testing procedures provides design guidance for microarray experiments. Based on this model and by choice of a multiple-testing procedure, expression noises based on permutation resampling can be considerably minimized. The method for mixture distribution model is suitable to various microarray data types obtained from single noise sources, or from multiple noise sources. By using the biological replicate number required in microarray experiments for a given power or by determining the power required to detect a gene significantly differentially expressed, given the sample size, or the best multiple-testing method can be chosen. As an example, a single-distribution model of *t*-statistic was fitted to an observed microarray dataset of 3 000 genes responsive to stroke in rat, and then used to calculate powers of four popular multiple-testing procedures to detect a gene of an expression change *D*. The results show that the B-procedure had the lowest power to detect a gene of small change among the multiple-testing procedures, whereas the BH-procedure had the highest power. However, all multiple-testing procedures had the same power to identify a gene having the largest change. Similar to a single test, the power of the BH-procedure to detect a small change does not vary as the number of genes increases, but powers of the other three multiple-testing procedures decline as the number of genes increases.

Key words: Multiple-testing procedure; microarrays; gene; power

Microarray is a novel and promising biotechnology that provides significant insight into the differential expression profiles of thousands of genes in cells[1,2] and is becoming very popular in biological and medical studies in addressing a wide range of problems, such as classification of tumors and host genomic response to bacterial infections[3]. An important and common goal in microarray experiments is to identify differentially expressed genes, that is, the identification of genes whose expression levels may be associated with a response to a complicated disease, such as stroke and cancer, presents useful information of important biological processes or functions[4,5].

However, as many authors[6–11] have pointed out, many noises occur in microarray expression profiles[1,2] because of multiple sources of variation that arise in producing the measurements, such as variation in the preparation of mRNA samples and in the incorporation of fluorescent tags, optical noise, and cross-hybridization, which lead to false positive results that are obtained merely based on differential expression analysis. Therefore, an existent challenge is how to detect those genuine changes from noisy microarray data. In practice, a common way to distinguish true changes from noises is to design an experiment in which replicates are required for statistical analyses of gene expression[6]. The key question is, however, how many replicates will be enough to detect differentially expressed genes in microarray

Received: 2006−06−02; Accepted: 2006−07−07

This work was supported by the National Natural Science Foundation of China (No. 30370208, 30570226).

① Corresponding author. E-mail: yanhm03@126.com

experiments. This problem then seems to fall into the traditional two-sample comparison in statistics. For microarray experiments, however, it is impossible and unnecessary to set as large a sample size as that required in a traditional two-sample test because of some limitations of experimental conditions and the fact that gene expression variations derived from noises on arrays may be assumed to be random, independent, and follow one or several distributions, which makes it possible to decrease sample sizes that must be required in the traditional experiments for the powers to detect systemic or treatment effects[8].

Black and Doerge[9] studied powers/replicates for the detection of differentially expressed genes using a parametric method. As is known, parametric methods depend on an assumption that noises in microarray experiments follow lognormal or gamma distributions[6]. However, true null distributions in microarray experiments are unknown. An alternative approach that avoids this assumption is to construct a distribution model consisting of either single or multiple distributions. Such a distribution model is very useful and necessary for the reasonable estimation of the null distribution because if experimental noises come from a single source, the null distribution based on this model is reliably estimated by a single empirical distribution, or if there are multiple noise sources in microarray experiments and these noises follow different distributions, the model will be composed of multiple components at different proportions, and thus the estimate of the null distribution is neither biased toward one of these different distributions nor made with an assumed distribution. However, a critical need for such a model is to be able to obtain an empirical distribution of the expression noises for the observed microarray data. Currently, there have already been several approaches to this task, of which permutation indeed is a preferred approach to construction of a reasonable empirical distribution of expression noises for the microarray data. Although housekeeping gene is one of the approaches, it is unacceptable in practice because it is difficult to find thousands of housekeeping genes for a given microarray experiment[8].

In this study, the permutation approach is used to construct the mixture distribution model for calculation of the required power/replicate number in different multiple-testing procedures to detect the genes that are differentially expressed in microarray experiments.

1 Materials and Methods

1.1 Statistical model

It is generally assumed that there are two systems for differential expression of genes in microarray experiments where the systems may be treatment and control[9], conditions[8], different tissues or genotypes, depending on experimental objectives. Let $x=(x_{ki1},x_{ki2},\text{L},x_{kim_{ki}})$ be a sample of m_{ki} observed values that are randomly drawn from system k (k = 1, 2) for gene i (i = 1, L , N), where N is the number of genes detected on arrays. It is noteworthy that only biological replication such as tissue (or individual) replication is taken into account for this study. For two-experiment design, only one sample is generally drawn from one system. Therefore, the sample from system k is also called sample k for convenience. The jth observed expression value of gene i in sample k is expressed by

$$x_{kij}=\mu_i+\tau_{ki}+e_{kij}, \tag{1}$$

where μ_i is the overall mean of all observed expression values of gene i; τ_{ki}, effect of system k contributing to expression variation of gene i and e_{kij}, random error or expression noise with mean $E(e_{kij})=0$ and variance σ_{ki}^2. The number of replicated observed values for expression of gene i in system k is specified by m_{ki} and N genes may have different replicates, i.e., m_{ki} may be unequal to m_{kl} ($i\neq l$, i, l = 1,L , N). In other words, any data with replicate number $\geqslant$ 2, including incomplete data, are taken into account in this study. From equation (1), the expectation of expression level x_{kij} is denoted by $E(x_{kij})=\mu_i+\tau_{ki}$. $E(x_{kij})$ is estimated by mean $\bar{x}_{ki}$. Thus a null hypothesis for the systemic effects