Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

C. R. Acad. Sci. Paris, Ser. I 337 (2003) 343–346

Comptes Rendus Mathematique

Statistics/Probability Theory

An iterative procedure for differential analysis of gene expression

Avner Bar-Hen [a], Stéphane Robin [b]

[a] *Université Aix-Marseille III, FST Saint Jérôme, case 451, 13397 Marseille cedex 20, France*
[b] *INA-PG/INRA Biométrie, 16, rue Claude Bernard, 75005 Paris, France*

Received 20 April 2003; accepted after revision 16 July 2003

Presented by Paul Deheuvels

Abstract

Microarrays are a popular technology to study genes that are differentially expressed between two conditions. In this Note, we propose an iterative procedure to determine the biggest subset of non-differentially expressed genes. We prove a pseudo Markov relationship that allows practical computations. We obtain explicit expressions for FDR and the level of the proposed test at each step. ***To cite this article: A. Bar-Hen, S. Robin, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Résumé

Procédure itérative pour l'analyse différentielle d'expression de gènes. Les biopuces constituent une technologie très utilisée pour étudier si des gènes s'expriment différemment entre deux conditions. Dans cette Note nous proposons une méthode itérative pour rechercher le plus grand sous-ensemble de gènes non-différentiellement exprimés. Nous prouvons une relation de type chaîne de Markov d'ordre deux qui simplifie fortement les calculs. Nous obtenons de manière explicite le FDR associé à notre procédure ainsi que le niveau du test à chaque étape. ***Pour citer cet article : A. Bar-Hen, S. Robin, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

1. Aim

Microarrays are part of a new class of biotechnologies that allow the monitoring of the expression level of thousands of genes simultaneously. It is a powerful methodology for identifying differentially expressed genes. However, when thousands of genes in a microarray data set are evaluated simultaneously by fold changes and significance tests, the probability of detecting false positives rises sharply. Basically, various procedures have been proposed in the literature to test the null hypothesis

$$\mathbf{H}_0(i) = \{\text{gene } i \text{ is not differentially expressed}\}.$$

For example, in the case of balanced analysis of variance, if we denote X_{ir} the differential expression of gene i during the r-th replicate, we have $\mathbf{H}_0(i) = \{\forall r,\ \mathbb{E}(X_{ir} = 0)\}$.

E-mail addresses: avner@bar-hen.net (A. Bar-Hen), robin@inapg.inra.fr (S. Robin).

1631-073X/$ – see front matter © 2003 Académie des sciences. Published by Éditions scientifiques et médicales Elsevier SAS. All rights reserved.
doi:10.1016/S1631-073X(03)00365-0

In this “one-by-one gene” approach, multiple testing problems immediately arise and lead to many false positive genes. Several solutions have been proposed in the statistical literature to control the global type I error rate (see for example [2] or, more recently, the false discovery rate (FDR, see [1] or [3]).

In this paper, we aim to determine the largest set of genes having the same differential expression. From a biological point of view, it is known that, in many experiments, most of the genes are non-differentially expressed. The set detected by our procedure will hence be considered as the set of non-differentially expressed genes.

2. Iterative procedure

We consider a statistic T of the form $T = \sum_{i=1}^{n} Z_i$ where i denotes the gene and where Z_i's are positive individual scores, taking low values for nondifferentially expressed genes and high values for others. This kind of statistic is encountered in many statistical methods such as analysis of variance, or Kruskall–Wallis test.

For example in ANOVA, $Z_i \propto (X_{i\bullet} - X_{\bullet\bullet})^2$ and statistic T is proportional to the sum of squares associated with the gene effect.

Let $Z_{[1]} < Z_{[2]} < \cdots < Z_{[n]}$ denote the order statistics associated with $(Z_1, Z_2, \ldots, Z_n)$, assuming that no tie exists. Conversely, let i_k denote the index of the gene having the k-th score Z: $Z_{i_k} = Z_{[k]}$.

We propose the following procedure:

Step 0: We test

$$\mathbf{H}_0^n = \bigcap_{i=1}^{n} \mathbf{H}_0(i)$$

using the complete statistic $T_n = \sum_{i=1}^{n} Z_i$. If T_n is greater than a threshold (associated with some predefined risk), $\mathbf{H}_0^n$ is rejected and we conclude that there exists at least one differentially expressed gene.

Step 1: Assuming that $\mathbf{H}_0^n$ is rejected, we test $\mathbf{H}_0^{n-1} = \{$there exists a subset of $n-1$ non-differentially expressed genes$\}$. $\mathbf{H}_0^{n-1}$ can be tested with the minimum of the n statistics defined on $n-1$ genes:

$$T_{n-1} = \min_i (T_n - Z_i) = T_n - \max_i (Z_i) = \sum_{i=1}^{n-1} Z_{[i]}.$$

T_{n-1} is obtained by removing the gene that most contributes to T_n.

An alternative procedure can be directly based on the order statistics $Z_{[i]}$. The idea is the same and we will not be developed it in this note.

Step k: Assuming that $\mathbf{H}_0^{n-k+1}$ is rejected, we follow the same principle as in Step 1: to test $\mathbf{H}_0^{n-k} = \{$there exists a subset of $n-k$ non-differentially expressed genes$\}$, we use the statistic

$$T_{n-k} = \sum_{i=1}^{n-k} Z_{[i]} = T_{n-k+1} - Z_{[n-k+1]}. \qquad (1)$$

$\mathbf{H}_0^{n-k}$ is rejected if T_{n-k} is greater than a threshold.

The general idea behind this procedure is to remove at each step the gene that most contributes to the statistic. It is directly given by the order statistics. The aim is to only keep the non-differentially expressed genes in the statistic and to accept the null hypothesis at the final step.

To derive the statistical properties of the procedure, we need the joint distribution of statistics $T_n,\ T_{n-1}, \ldots, T_1$. To make the dependency between successive steps tractable, we conserve the values of the scores Z_i all along the procedure. This leads to a recurrence formula for the statistics: $T_{k-1} = T_k - Z_{[k]}$. The price for this trick is the non-standard form of the null distributions of the T_k's.

For example, in the ANOVA context, the grand mean $X_{\bullet\bullet}$ is not recomputed at each step and, under $\mathbf{H}_0^{n-k}$, $Z_i = (X_{i\bullet} - X_{\bullet\bullet})^2$ has a truncated non-central chi-square distribution. Recomputing $X_{\bullet\bullet}$ would lead to more classical distributions but to intractable dependency between T_k's.

3. Statistical properties of the test statistics

Notations. In the following, we denote f_X (resp. F_X) the probability density function (resp. cumulative distribution function, cdf) of the random variable (rv) X. We use the special notation ϕ (resp. Φ) for the rv Z and ϕ_n (resp. Φ_n) for the sum of n iid rv's with density ϕ. For example, under $\mathbf{H}_0^n$ we have $\Pr\{T_n \leqslant b\} = \Phi_n(b)$. The distribution ϕ is assumed to be known.

Joint distribution. As shown in previous section, we are interested in the statistics T_k, which are sums of order statistics $\{Z_{[i]}\}$. To control the overall risk associated with the test procedure, we need to derive the joint distribution of $(T_n, T_{n-1}, \ldots, T_1)$.

Proposition 3.1. *Let $Z_1, \ldots, Z_n$ be n independent positive rv's with common cdf Φ and density ϕ. The joint density of $(T_n, \ldots, T_1)$, where T_k denote the sum defined in Eq.* (1), *is*

$$f_{T_n \ldots T_1}(t_n, \ldots, t_1) = n! \prod_{k=1}^{n} \phi(t_k - t_{k-1}) \mathbb{I}\{\forall k\colon\ t_k - t_{k-1} > t_{k-1} - t_{k-2}\},$$

where $\mathbb{I}\{A\} = 1$ if A is true and 0 *otherwise, and with the convention $t_0 = 0$.*

Proof. The joint density of $(T_n, \ldots, T_1)$ can be directly expressed in terms of order statistics $Z_{[i]}$. To get the result, we use the reciprocal transformation of Eq. (1): $z_k = t_k - t_{k-1}$, for $1 \leqslant k \leqslant n$. The Jacobian of this transform is equal to one. We get $\phi(z_k) = \phi(t_k - t_{k-1})$. Condition $z_1 < \cdots < z_n$ becomes $t_k - t_{k-1} > t_{k-1} - t_{k-2}$ for all k. □

3.1. Conditional distributions

Since we propose an iterative procedure, we will also need the conditional distribution of the statistics given the results of the preceding steps.

Proposition 3.2. *Under $\mathbf{H}_0^k$,*

$$f_{Z_{[k]}\,|Z_{[n]},\ldots,Z_{[k+1]}} = f_{Z_{[k]}\,|Z_{[k+1]}}.$$

Proof. Under $\mathbf{H}_0^k$, we have

$$\{Z_k, \ldots, Z_1\} \sim \phi(z) \text{ iid}, \quad \text{and} \quad \{Z_n, \ldots, Z_{k+1}\} \sim \psi(z_n, \ldots, z_{k+1}).$$

The joint distribution is

$$f_{Z_n,\ldots,Z_1}(z_n, \ldots, z_1) = \psi(z_n, \ldots, z_{k+1}) \prod_{j=k+1}^{n} \mathbb{I}\{z_j \geqslant z_{[k]}\} \prod_{i=1}^{k-1} \phi(z_i) \mathbb{I}\{z_{[i+1]} \geqslant z_{[i]}\}.$$

The first product of $\mathbb{I}\{\cdot\}$ terms comes from the previous steps of the iterative procedure. We have

$$f_{Z_{[k+1]},Z_{[k]}}(z_{k+1}, z_k) = f_{\min\{Z_n,\ldots,Z_{k+1}\}}(z_{k+1}) f_{\max\{Z_k,\ldots,Z_1\}}(z_k) \mathbb{I}\{z_{[k+1]} \geqslant z_{[k]}\}$$

and the result is direct. □

Proposition 3.3. *We have*

$$f_{T_k\ |T_n \ldots T_{k+1}} = f_{T_k\ |T_{k+2}, T_{k+1}}$$

Proof. The proof follows the same principle as the proof of Proposition 3.2. The second order is a consequence that, according to Eq. (1), Z_k's are the increment of T_k's. □

These two propositions have a strong flavour of Markov process of order 1 and 2 but the recurrence relations do not hold since the null hypothesis is varying along the steps.

The p-value $P_k = \Pr\{T_k \geqslant t_k | t_{k+2}, t_{k+1}\}$ can be calculated according to Proposition 3.3. It can be noted that this calculation does not rely on any assumption (such as independence or distributional hypothesis) about differentially expressed genes.

Proposition 3.4. *When testing $\mathbf{H}_0^k$ with statistic T_k and conditional to the preceding steps, the rejection rule providing a Type I error rate α is $T_k > c_k$ where c_k satisfies*

$$1 - \alpha = G_k^*(c_k; Z_{[k+1]}),$$

where $G_k^*(\cdot; Z_{[k+1]})$ denotes the k times convolution of $\phi^*(x; Z_{[k+1]}) = \phi(x)/\Phi(Z_{[k+1]})\mathbb{I}\{x \leqslant Z_{[k+1]}\}$, that is of distribution ϕ truncated at $Z_{[k+1]}$.

Proof. Under $\mathbf{H}_0^k$, $Z_{[k+1]}$ contains all the relevant information contained in the preceding steps. The result is obtained by remarking that, under $\mathbf{H}_0^k$ and given $Z_{[k+1]}$, $(Z_{[1]}, \ldots, Z_{[k]})$ have the same distribution as the order statistics of an iid sample $(Z_1^*, \ldots, Z_k^*)$ with density ϕ truncated at $Z_{[k+1]}$. □

Proposition 3.4 allows us to control the type I error at a desired value α at each step, conditional to the preceding steps (thanks to the Markovian properties 3.2 and 3.3).

We now consider the false discovery rate (FDR) introduced by [1]. Denoting R the number of rejected null hypothesis and V the number of nondifferentially expressed genes among these R. The FDR is $\mathbb{E}(V/R)$. In our iterative procedure, we define V_k as the number of false discoveries up to step k. Thus, at step k, we have $R = k$ and $FDR(k) = \mathbb{E}(V_k)/k$ that can be calculated thanks to the following proposition.

Proposition 3.5. *The expected number of false discoveries up to step k is $\mathbb{E}(V_k) = \sum_{i=k}^{n} P_i$.*

Proof. We clearly have $\Pr(V_n = 1) = P_n$. Moreover, under $\mathbf{H}_0^k$, the rejection of $\mathbf{H}_0^k$ leads to one more false discovery with probability P_k. So we have

$$\Pr(V_k = j) = P_k \Pr(V_{k-1} = j - 1) + (1 - P_k)\Pr(V_{k-1} = j).$$

The result is given by convolution of binomial laws with different probabilities. □

References

[1] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, JRSSB 57 (1) (1995) 289–300.

[2] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Statist. 6 (1979) 65–70.

[3] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (2001) 5116–5121.