FEBS Letters 587 (2013) 677–682

FEBS Letters

journal homepage: www.FEBSLetters.org

The exact probability distribution of the rank product statistics for replicated experiments

Rob Eisinga [a,*], Rainer Breitling [b], Tom Heskes [c]

[a] Department of Social Science Research Methods, Radboud University Nijmegen, Netherlands
[b] Manchester Institute of Biotechnology, University of Manchester, United Kingdom
[c] Institute for Computing and Information Sciences, Radboud University Nijmegen, Netherlands

ARTICLE INFO

Article history:
Received 11 October 2012
Revised 16 January 2013
Accepted 17 January 2013
Available online 8 February 2013

Edited by Paul Bertone

Keywords:
Rank product method
Microarray
Permutational inference
Gamma approximation
Exact inference

ABSTRACT

The rank product method is a widely accepted technique for detecting differentially regulated genes in replicated microarray experiments. To approximate the sampling distribution of the rank product statistic, the original publication proposed a permutation approach, whereas recently an alternative approximation based on the continuous gamma distribution was suggested. However, both approximations are imperfect for estimating small tail probabilities. In this paper we relate the rank product statistic to number theory and provide a derivation of its exact probability distribution and the true tail probabilities.

1. Introduction

The rank product method is a popular non-parametric technique introduced by Breitling et al. [1] for identifying differentially expressed genes using data from replicated microarray experiments. It has also been widely applied to other post-genomic datasets that generate replicated rankable scores, for example in proteomics and metabolomics [3–5]. The method entails ranking genes according to their differential expression within each replicate experiment and subsequently calculating the product of the ranks across replicates. An important next step is to compare the observed rank products to their sampling distribution under the null hypothesis that the differential expression values are identically distributed (i.e., statistically exchangeable) within each of the independent experiments. Breitling et al. [1] proposed a permutation sampling procedure to approximate this distribution, whereas Koziol [2] recently suggested an alternative approximation based on the continuous gamma distribution. The latter cautions, however, that both permutation re-sampling and the gamma approximation fail to provide accurate estimates of the extreme tail probabilities of the rank product statistics.

This note provides a combinatorial exact expression for calculating the probability mass function of the rank product statistic and the exact *P*-values based on the fundamental theorem of arithmetic. The underlying method has previously been suggested by Lehner et al. [6] in a different research area, but their expression is exact only for the restricted case that the rank product is not larger than the number of genes in the array. In this paper, we remove this restriction, making the resulting counting method generally applicable to the analysis of microarray and other data. Our numerical example shows that the exact probability mass function offers an improvement over the continuous gamma approximation, which tends to understate the evidence against the null hypothesis, and permutation. This improvement is important for the application of the rank product method in all areas of biological data analysis, as the main interest is typically directed towards the tail of the distribution, that is, the detection of "significantly changed" genes, proteins or metabolites.

2. Rank product analysis

Suppose we have differential expression data for a total of n genes from k replicated experiments, with all replicates measuring the same number of genes. The underlying distribution of the differential expression values themselves is unknown, prohibiting the calculation of the probability distribution of the raw expression data. For this reason, each measurement of the differential

* Corresponding author. Fax: +31 24 3612351.
E-mail addresses: r.eisinga@maw.ru.nl (R. Eisinga), rainer.breitling@manchester.ac.uk (R. Breitling), t.heskes@science.ru.nl (T. Heskes).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2013.01.037

expression for the ith gene in the jth replicate is replaced with its rank, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant k$. The most strongly up-regulated gene in each replicate is assigned rank 1 and the most strongly down-regulated gene is assigned rank n, giving k sets of ranks, denoted r_{ij}, $1 \leqslant r_{ij} \leqslant n$. Assuming no ties, each rank occurs once and only once in each replicate.

For each gene i we have a rank tuple $\{r_{i1}, \ldots, r_{ik}\}$, and rank product analysis intends to examine those tuples where all of the ranks are sufficiently small. The individual rank scores r_{ij} can be used as a test statistic for the null hypothesis that a gene is not significantly regulated against the alternative that it is differentially expressed, yielding a P-value given by $P(R \leqslant r_{ij}) = r_{ij}/n$. Rank product analysis aims to integrate the evidence from k independent biological replicates to provide a P-value for the overall test that all k single null hypotheses are true.

In line with Fisher's [7] method, the rank product approach to combining the individual P-values is to obtain the product of the ranks for gene i over the independent replicates k, i.e., $rp_i = \prod_{j=1}^{k} r_{ij}$. The observed rank product is then compared to the sampling distribution of the rank product values under the overall null hypothesis that the expression levels are identically distributed within each of the k independent replicates. Assessing the statistical significance, or P-value, of the observed expression changes therefore relies on the ability to obtain this null distribution accurately. In the original publication, Breitling et al. [1] proposed to obtain an approximate distribution under the condition that all the null hypotheses are true by permutation re-sampling. This strategy requires a computationally demanding large number of permutations to get reliable estimates of the P-values at the tails of the distribution, that is, for the most significantly changed genes. Therefore, an analytical approach for calculating the distribution without requiring permutations was desirable. Hereafter, for notational convenience, we will drop all reference to the symbol i and consider how to make probability calculations using the gamma approximation and exact calculation.

3. Gamma approximation for rank products

In [2], Koziol argues that under the null hypothesis $r_j/(n+1)$ is approximately uniformly distributed on the interval [0,1] and he uses this argument to propose a continuous gamma distribution approximation for the log-transformed rank products, $z = -\log(rp/[n+1]^k)$.

If the P-values $r_j/(n+1)$ are uniform and continuous on the unit interval [0,1], the probability distribution of $w_j = -\log(r_j/[n+1])$ is given by the exponential distribution $p(w_j) = e^{-w_j}$ with scale parameter 1, denoted as Exp(1). Given that w_j is distributed as Exp(1), the sum of w_j over k independent replicates has a gamma $(k,1)$ distribution, i.e., $p(z) = \Gamma(k)^{-1} z^{k-1} e^{-z}$, where $z = \sum_{j=1}^{k} w_j$ [see 2,6,8]. Koziol [2] shows that the empirical distribution of the log-transformed rank product values is well-approximated by the continuous gamma $(k,1)$ distribution over the (almost) entire range of support. He urges, however, that estimation of small tail probabilities of the rank products from the gamma approximation is imprecise.

The reason for the deviation is that the rank products take discrete values on the real number line, i.e., $1, 2, 3, \ldots, n^k$, whereas the continuous gamma distribution allows all non-negative real numbers. The deviations are most prominent if the rank products are small, hence at the right tail of the distribution. Below we will give an example that illustrates the difference between the true P-value and the approximate P-value based on the gamma $(k,1)$ probability density function.

4. Exact distribution of rank products

To overcome the limitations of the approximation strategies, recall that the rank products have a probability mass function. This function gives the probability that a discrete random variable RP is exactly equal to some value rp. This probability, denoted $P(RP = rp)$, can be obtained by calculating the total number of ways to get rp by multiplying k integers (number of replicates) between 1 and n (number of genes), and dividing the result by n^k. One approach to this counting problem is using a for loop. That is, run k nested loops from 1 to n, most efficiently by the divisors of rp, and count the number of times the resulting product equals rp. This brute-force search performs well, but it becomes computationally time consuming if either n or k or both are large. The more so, if in addition to the probability the P-value of large rank products is required.

An alternative calculation relies on the fundamental theorem of arithmetic also known as the unique-prime-factorization theorem [9,10]. This theorem states that every positive integer (except the number 1) has a unique prime-factorization implying that it can be presented in exactly one way as a product of powers of primes. For the problem at hand, this implies that every rank product rp greater than 1 is either prime itself or is the product of primes, i.e., $rp = p_1^{a_1} \ldots p_m^{a_m}$, where $p_1 < p_2 < \ldots p_m$ are distinct primes and the prime exponents a_t are non-negative integers. Obviously, the same goes for the divisors d of rank product rp.

We denote by $H(rp;k,n)$ the total number of representations of rank product rp as an ordered product of k ranks smaller than or equal to n. That is, two representations of rp are identical only if they contain the same ranks in the same order. We also assume by definition that $H(1;k,n) = 1$.

In their discussion of rank statistics, Lehner et al. [6] have shown that we can enumerate the number of ordered k-tuples such that their product equals rp, using

$$H(rp;k,n) = \prod_{t=1}^{m} \binom{a_t + k - 1}{k - 1} \text{ if } rp \leqslant n.$$

The computation of $H(rp;k,n)$ is an application of the so-called Piltz divisor function [11, see also Sloane's (A007425) at http://oeis.org/A007425], and intimately related to the study of ordered factorizations of integers [12]. For a proof see Nathanson [10], Theorem 7.5, and Lehner et al. [6]. The above expression for $H(rp;k,n)$ is a valid method for counting the representations of rp as long as the rank product is less than or equal to the number of genes. The function is then independent of n, and it offers the total number of ways of writing rp as an ordered product of k ranks.

This counting formula may occasionally be appropriate for examining top-lists of most up-regulated genes, if n is large and the number of replicates is small. But in many biological applications, with several replicates and noisy data, for many genes rp will be larger than n, possibly even for strongly differentially expressed genes. If that is the case, the above expression for $H(rp;k,n)$ is invalid, as it includes rank tuples with rank values that are larger than n. Obviously, such rank tuples are impossible in replicates with n genes.

Let d_g be a divisor of rp that is larger than n, where $g = 1, \ldots, v$. To obtain a generic formula that is valid for all possible rank product values, we express $H(rp;k,n)$ in terms of functions $H(\cdot;\cdot,\infty)$ as

$$H(rp;k,n) = \sum_{s=0}^{k} \sum_{\beta:\sum_g \beta_g = s} (-1)^s \binom{k}{s} \binom{s}{\beta_1, \ldots, \beta_v} \times H\left(rp/\prod_g d_g^{\beta_g}; k-s, \infty\right),$$