A Simple Rank Product Approach for Analyzing Two Classes

Tae Young Yang

Department of Mathematics, Myongji University, Yongin, Kyonggi, Korea.

ABSTRACT: The rank product statistic has been widely used to detect differentially expressed genes in replicated microarrays and a one-class setting. The objective of this article is to apply a rank product statistic to approximate the *P*-value of differential expression in a two-class setting, such as in normal and cancer cells. For this purpose, we introduce a simple statistic that compares the *P*-values of each class's rank product statistic. Its null distribution is straightforwardly derived using the change-of-variable technique.

KEYWORDS: change of variable, chi-squared approximation, log-transformation, rank product statistic, two-class setting

CITATION: Yang. A Simple Rank Product Approach for Analyzing Two Classes. *Bioinformatics and Biology Insights* 2015:9 119–123 doi: 10.4137/BBI.S26414.

TYPE: Methodology

RECEIVED: March 31, 2015. **RESUBMITTED:** June 14, 2015. **ACCEPTED FOR PUBLICATION:** June 21, 2015.

ACADEMIC EDITOR: J.T. Efird, Associate Editor

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 833 words, excluding any confidential comments to the academic editor.

FUNDING: This research was supported by Basic Science Research Program (NRF-2010-0009461) through the National Research Foundation of Korea funded by the Ministry of Education. The author confirms that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: tyang@mju.ac.kr

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

The rank product statistic[1] is a robust nonparametric approach that has been proposed to detect differentially expressed genes in replicated microarrays with just one class or condition. Because the rank product statistic transforms expression intensity into ranks, it has several advantages, including fewer assumptions and easy handling of noisy data or few microarrays.[2] Although the rank product statistic has been used mainly for microarrays, it is also applicable to meta-analyses[3,4] and proteomics.[5]

The rank product statistic ranks genes according to expression intensities within each microarray and calculates the product of these ranks across multiple microarrays. This technique can identify genes that are consistently detected among the most differentially expressed genes in a number of replicated microarrays. However, a very large number of permutations and a substantial amount of computation time are required to accurately calculate the *P*-value to test for differential expression. Alternatively, Koziol[6] proposed a log-transformed rank product statistic and used a continuous gamma distribution to approximate its *P*-value. The computation time to calculate the *P*-value for testing differential expression is negligible compared with that required to calculate the permutations.

To extend the rank product statistic to approximate the *P*-value of differential expression under a two-class setting, such as in cancer cells and normal cells, Koziol[7] used the difference between two averaged gamma variables. However, calculating the null density of the difference is mathematically complicated. In contrast, this article proposes a simple variable for comparing the *P*-values of each class's log-transformed rank product statistic and describes its null distribution, which is easily derived by a change-of-variable technique.

Background of One-Class Rank Product Statistic

Assume that we have m replicate microarrays representing one class, with each microarray measuring expression of n genes. For each microarray j ($j = 1,\ldots,m$), Koziol[6] ranked the expression levels $X_{1j},\ldots,X_{nj}$, and denoted $R_{ij} = \text{rank}(X_{ij})$ in a way such that the most highly expressed gene is assigned rank 1 and the least expressed gene is assigned rank n, then R_{ij} $\{1,\ldots,n\}$. For each gene i, we have a rank tuple of $\{R_{i1},\ldots, R_{im}\}$. The original rank product statistic for gene i is

$$RP_i = \prod_{j=1}^{m} R_{ij},$$

which is the product of the ranks i over m independent microarrays. Assuming that each rank occurs only once with independent samples, RP_i takes discrete values of 1, 2,…, n^m. When ($R_{i1},\ldots, R_{im}$) is small, RP_i is small, indicating that gene i is expressed differentially.

To calculate the *P*-value for the test that gene i is differentially expressed, RP_i is compared with its permutation distribution under the null hypothesis that R_{ij} for $i = 1,\ldots,n$ are exchangeable within each microarray j.[1] However, to accurately approximate the distribution, a very large number of permutations is required, which becomes very time-consuming computationally. Thus, a simpler approximation approach is needed to calculate the *P*-value of RP_i.

Log-Transformed Rank Product Statistic

An individual P-value given by $R_{ij}/(n+1)$ is approximately uniformly distributed on the unit interval (0, 1), with the approximation improving as n (the number of genes) increases. If $R_{ij}/(n+1)$ is continuously uniform on (0, 1), the transformation of $-2\ln(R_{ij}/(n+1))$ has a chi-squared distribution with two degrees of freedom, denoted as $\chi^2(2)$. In contrast, Koziol[6] used the transformation $-\ln(R_{ij}/(n+1))$, which has an exponential density Exp(1). Chi-squared tables are readily available, so the advantages of chi-squared favor the approach proposed here.

We can combine individual chi-squared variables as follows

$$-2\sum_{j=1}^{m} \ln(R_{ij}/(n+1)) = -2\ln(RP_i/(n+1)^m), \qquad (1)$$

which has a $\chi^2(2m)$ density. Because the monotonicity of the log function ensures that significance levels of RP_i and $\ln RP_i$ are identical, the chi-squared density provides a simple calculation to obtain the P-value of RP_i.

Let $(r_{i1},\ldots,r_{im})$ and $rp_i = \Pi_{j=1}^{m} r_{ij}$ be the observed values of $(R_{i1},\ldots, R_{im})$ and RP_i, respectively. The P-value of rp_i for testing the differential expression of gene i is

$$P(\chi^2\,(2m) > -2\ln(rp_i/(n+1)^m)).$$

When $(r_{i1},\ldots,r_{im})$ is small, rp_i and its P-value are also small, indicating that gene i is differentially expressed.

A New Statistic for Analyzing Two Classes

Suppose we extend the analyses to two classes, with m_1 independent microarrays in class 1 and m_2 independent microarrays in class 2. Each microarray measures n genes. Going forward, for simplicity, the i gene label is omitted. Let $RP_1 = \Pi_{j=1}^{m_1} R_{ij}$ and $RP_2 = \Pi_{j=1}^{m_2} R_{ij}$ be the rank product statistics of classes 1 and 2, respectively. Note that rp_1 and rp_2 are the observed values of RP_1 and RP_2, respectively.

Let X_1 and X_2 be

$$X_1 = -2\sum_{j=1}^{m_1} \ln(R_{ij}/(n+1)) = -2\ln(RP_1/(n+1)^{m_1})$$
$$X_2 = -2\sum_{j=1}^{m_2} \log(R_{ij}/(n+1)) = -2\ln(RP_2/(n+1)^{m_2}).$$

Note that the two independent random variables X_1 and X_2 have $\chi^2(2m_1)$ and $\chi^2(2m_2)$, respectively, under the null hypothesis that R_{ij} for $i = 1,\ldots,n$ are exchangeable within each microarray j.

To calculate the P-value of differential expression of gene i under a two-class setting, we define a new statistic

$$V = \frac{P\,(\chi^2(2m_1) > x_1)}{P(\chi^2(2m_1) > x_1) + P(\chi^2(2m_2) > x_2)}, \qquad (2)$$

where $x_1 = -2\ln(rp_1/(n+1)^{m_1})$ and $x_2 = -2\ln(rp_2/(n+1)^{m_2})$ are the observed values of X_1 and X_2, respectively. Genes associated with sufficiently small V would be differentially expressed for testing H_0: class 1 = class 2 vs. H_a: class 1 > class 2.

The distributions of $P(\chi^2(2m_1) > x_1)$ and $P(\chi^2(2m_2) > x_2)$ are uniform (0, 1) under the null hypothesis. Then, the density of V is

$$f(V) = \begin{matrix} \dfrac{1}{2(1-V)^2}, 0 < V < \dfrac{1}{2} \\ \dfrac{1}{2V^2}, \dfrac{1}{2} < V < 1. \end{matrix}$$

The proof is presented in the Appendix. The P-value for testing H_0: class 1 = class 2 vs. H_a: class 1 > class 2 can be obtained by

$$P\left(V < \frac{p_1}{p_1 + p_2}\right) = \int_0^{\frac{p_1}{p_1+p_2}} f(V)dV, \qquad (3)$$

where $p_1 = P\,(\chi^2(2m_1) > x_1)$ and $p_2 = P\,(\chi^2(2m_2) > x_2)$.

Similarly, the P-value for testing H_0: class 1 = class 2 vs. H_a: class 1 < class 2 can be obtained by

$$P\left(V < \frac{p_1}{p_1 + p_2}\right) = \int_0^{\frac{p_2}{p_1+p_2}} f(V)dV.$$

Numerical Examples

Simulation study. We evaluated the performance of the proposed statistic V in Equation (2) by comparing its specificity (or 1 false-positive rate) and sensitivity (or power) in detecting differential expression to the Wilcoxon rank-sum statistic, which is widely used for nonparametrical testing to calculate the P-value of differential expression under a two-class setting. For the following specifications, we conducted 1,000 simulation experiments to assess the specificity and sensitivity of the statistic.

To assess the specificity of the proposed statistic, we simulated 10,000 genes such that the gene expression in 40 microarrays for each gene was simulated independently from a standard normal distribution, where the first 20 samples ($m_1 = 20$) were the control group and the second 20 were ($m_2 = 20$) the treatment group. This specification represents a situation in which no genes are differentially expressed. The false-positive rate was then calculated as follows: the number of genes found to be differentially expressed at nominal level α were counted and divided by 10,000 (the number of genes).

Table 1 presents the false-positive rates of the proposed statistic for various α, m_1, and m_2. As can be seen from the table, the statistic maintained appropriate α-levels.

To assess the power of the proposed statistic, 10,000 genes were simulated such that the gene expression for each gene in

Table 1. False-positive rates of the proposed statistic for various nominal α-levels and numbers of samples, where m_1 and m_2 are the sample numbers of the control group and the treatment group, respectively.

m_1, m_2	α-LEVEL			
	0.01	0.05	0.10	0.25
10,10	0.0097	0.0494	0.0997	0.2501
20,20	0.0099	0.0496	0.0993	0.2495
30,30	0.0095	0.0492	0.0994	0.2491
10,20	0.0097	0.0496	0.0994	0.2493
20,10	0.0095	0.0496	0.0997	0.2499

Note: The numbers denote the rates of genes that were identified by the proposed statistic as differentially expressed at α.

40 microarrays was simulated independently from a standard normal distribution and where the first 20 samples were the control group and the second 20 were the treatment group. Next, 5% of genes were randomly selected, and a constant of 0.25 was added to their treatment group. These selected genes had a higher average expression in the treatment group; however, there was no difference between the two groups for the remaining 95% genes. We repeated the same procedure by adding larger constants: 0.5, 1.0, and 1.5. In Table 2, the numbers represent the percentages of the selected 5% differentially expressed genes that were found to be differentially expressed at various significance levels α. The results of the proposed statistic were compared with those obtained from the Wilcoxon rank-sum test statistic. The table clearly shows that the proposed statistic is more powerful than the Wilcoxon statistic and that it was able to accurately detect the differentially expressed genes.

Table 2. Power of the proposed statistic for various nominal α-levels.

α-LEVEL	ADDED CONSTANT			
	0.25	0.5	1.0	1.5
0.01	0.08 (0.06)	0.32 (0.20)	0.91 (0.74)	1.0 (0.98)
0.05	0.24 (0.19)	0.56 (0.44)	0.97 (0.91)	1.0 (1.0)
0.1	0.35 (0.30)	0.69 (0.58)	0.98 (0.96)	1.0 (1.0)
0.2	0.57 (0.54)	0.83 (0.80)	0.99 (0.99)	1.0 (1.0)

Notes: We simulated 10,000 genes such that the gene expression in 40 microarrays for each gene was simulated independently from a standard normal distribution, and where the first 20 samples were the control group and the second 20 were the treatment group. We randomly selected 5% of genes and added a constant of 0.25 to their treatment group. These selected genes had a higher average expression in the treatment group; however, there was no difference between the two groups for the remaining 95% of genes. We repeated the same procedure by adding larger constants: 0.5, 1.0, and 1.5. The numbers denote the percentages of differentially expressed genes that were identified by the proposed statistic as differentially expressed. For comparison, the numbers inside parentheses denote the percentages of differentially expressed genes identified by the Wilcoxon rank-sum statistic.

Real data analysis. The widely used data set of Golub et al.[8] came from a study of gene expression in two classes of acute leukemia: acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide microarrays containing 6,817 human genes. Three preprocessing procedures were applied to the gene expression levels and are available at http://www.genome.wi.mit.edu/MPR. These preprocessing procedures included (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with (max/min) ≤5 or (max-min) ≤500, where max and min refer, respectively, to the maximum and minimum levels for a particular gene across mRNA samples; and (iii) $\log_{10}$ transformation.[9] The data were then summarized by a 3,051 × 38 matrix, which is implanted in the multitest package from http://www.bioconductor.org/biocLite.R.

Table 3 presents the top 25 AML significant genes from Equation (3). Eleven genes marked with * are also reported among the top 25 AML-specific genes in Golub et al. We also compared *P*-values of the proposed statistic to those of the Wilcoxon rank-sum statistic. The proposed *P*-values were obtained under the overall null hypothesis that the expression levels are exchangeable within each of the independent microarrays. Eleven genes marked with * were also reported among the top 25 AML-specific genes in Golub et al.

Conclusion

To approximate the *P*-value of differential expression under a two-class setting, Koziol[7] derived the density of the difference between two averaged gamma variables, which is mathematically complex. In contrast, we provided a simple, nonparametric statistic V in Equation (2). Its null distribution was easily derived by the change-of-variable technique. In the sensitivity analysis presented in the Simulation study section, the proposed statistic was more powerful than the Wilcoxon statistic. In the specificity analysis, it also maintained appropriate α-levels. We developed an R program for this statistic, available at http://home.mju.ac.kr/home/index.action?siteId=tyang.

Koziol[6] noted that the *P*-values of $\ln RP_i$ in Equation (1) were well approximated by the corresponding continuous gamma approximation (or in our case, chi-squared) over most of the data range; however, the estimation of extremely small *P*-values was rather imprecise. Specifically, the gamma approximation is conservative in that it tends to overestimate extremely small *P*-values, leading to false-negative results, which is due to the fact that the discrete rank products take values of 1, 2,…, n^m, whereas the continuous chi-squared distribution uses positive, real numbers.[10] Because p_1 and p_2 in Equation (3) are based on gamma approximation, the *P*-value of the proposed statistic V may be imprecise, particularly when both p_1 and p_2 are extremely small.

Table 3. Our *P*-values obtained under the overall null hypothesis that the expression levels are exchangeable within each of the independent microarrays.

AFFYMETRIX ID	DESCRIPTION	OUR TOP 25 P-VALUES	WILCOXON'S *P*-VALUE
Y00787*	interleukin-8 precursor	6.07×10^{-11}	3.39×10^{-06}
M27891*	CST3 Cystatin C	9.69×10^{-09}	3.32×10^{-09}
M96326*	Azurocidin gene	6.69×10^{-08}	8.28×10^{-06}
M28130*	Interleukin 8 gene	2.85×10^{-07}	2.67×10^{-06}
M63438	glutamine synthase	7.17×10^{-07}	1.10×10^{-04}
X17042*	PRG1 Proteoglycan 1, secretory granule	2.51×10^{-06}	2.74×10^{-05}
U01317	Delta-globin gene	4.47×10^{-06}	6.42×10^{-04}
M19507	mpo myeloperoxidase	5.95×10^{-06}	1.53×10^{-05}
M91036	G-gamma globin	8.83×10^{-06}	1.37×10^{-03}
M87789	hybridoma H210	1.00×10^{-05}	2.06×10^{-04}
X95735*	Zyxin	1.14×10^{-05}	8.31×10^{-10}
M19045*	LYZ	1.27×10^{-05}	2.67×10^{-06}
X14008	Lysozyme gene	1.81×10^{-05}	6.67×10^{-06}
X64072	SELL Leukocyte adhesion protein beta subunit	2.09×10^{-05}	2.74×10^{-05}
J04990	cathepsin g precursor	2.38×10^{-05}	1.53×10^{-05}
J03801	LYZ	2.59×10^{-05}	1.63×10^{-06}
X62320	GRN Granulin	4.60×10^{-05}	4.16×10^{-07}
X04085*	Catalase 5'flank and exon 1 mapping to chr 11	5.59×10^{-05}	2.67×10^{-06}
M21119	LYZ	7.99×10^{-05}	9.49×10^{-04}
M84526*	DF D component of complement	1.09×10^{-04}	3.30×10^{-05}
M57710*	galectin 3	1.11×10^{-04}	9.37×10^{-05}
L09209	APLP2 Amyloid beta (A4) precursor-like protein 2	1.33×10^{-04}	5.56×10^{-08}
L08246*	induced myeloid leukemia cell differentiation protein mcl1	1.53×10^{-04}	2.67×10^{-06}
X62654	ME491	2.21×10^{-04}	1.15×10^{-07}
X65965	manganese superoxide dismutase	3.26×10^{-04}	7.78×10^{-03}

Notes: The top 25 *P*-values for AML-specific genes from the leukemia data of Golub et al from Equation (3). Among them, 11 genes marked with * were reported among the top 25 AML-specific genes in Golub et al. Our *P*-values are compared with *P*-values of Wilcox rank-sum test. Ten genes of the Wilcoxon rank-sum statistic were reported among the top 25 AML-specific genes in Golub et al.

Author Contributions

Conceived and designed the experiments: TYY. Analyzed the data: TYY. Wrote the first draft of the manuscript: TYY. Contributed to the writing of the manuscript: TYY. Agree with manuscript results and conclusions: TYY. Jointly developed the structure and arguments for the paper: TYY. Made critical revisions and approved final version: TYY. The author reviewed and approved of the final manuscript.

REFERENCES

1. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*. 2004;573:83–92.
2. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006;22:2825–7.
3. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008;24:374–82.
4. Campain A, Yang YH. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*. 2010;11:408.
5. Hoefsloot H, Smit S, Smilde A. A classification model for the Leiden proteomics competition. *Stat Appl Genet Mol Biol*. 2008;7:8.
6. Koziol JA. Comments on the rank product method for analyzing replicated experiments. *FEBS Lett*. 2010;584:941–4.
7. Koziol JA. The rank product method with two samples. *FEBS Lett*. 2010;584:4481–4.
8. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–7.
9. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Ass*. 2002;97:77–87.
10. Eisinga R, Breitling R, Heskes T. The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Lett*. 2013;587:677–82.

Appendix A. Proof

The density of $W = U_1/(U_1 + U_2)$ with $U_i \sim \text{Uniform}(0,1)$ $(i = 1,2)$ is

$$f(W) = \begin{array}{l} \dfrac{1}{2(1-W)^2}, 0 \ < \ W \ < \ \dfrac{1}{2} \\ \dfrac{1}{2W^2}, \dfrac{1}{2} \ < \ W \ < \ 1. \end{array}$$

Proof. Let $Z = U_1 + U_2$ and $W = U_1/(U_1 + U_2)$. Then, $U_1 = WZ$ and $U_2 = (1 - W)Z$. The Jacobian is Z. The joint density of Z and W is $f(Z, W) = Z$. Then, the marginal density of W is obtained by

$$f(W) = \begin{array}{l} \displaystyle\int_0^{1/(1-W)} Z \ dZ = \frac{1}{2(1-W)^2}, \ 0 < W < \frac{1}{2} \\ \displaystyle\int_0^{1/w} Z \quad dZ = \frac{1}{2W^2}, \frac{1}{2} < W < 1. \end{array}$$