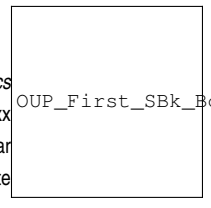
Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications note



Gene expression

RankProd 2.0: a refactored Bioconductor package for detecting differentially expressed features in molecular profiling datasets

Francesco Del Carratore [1], Andris Jankevics [2], Rob Eisinga [3], Tom Heskes [4], Fangxin Hong [5] and Rainer Breitling [1,*]

[1]Manchester Institute of Biotechnology, Faculty of Science and Engineering, University of Manchester, Manchester, M1 7DN, UK
[2]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3GJ, UK
[3]Department of Social Science Research Methods, Radboud University Nijmegen, Nijmegen, 6525 GD, Netherlands
[4]Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, 6525 EC, Netherlands
[5]Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA

*To whom correspondence should be addressed.

Associate Editor: Dr. Ziv Bar-Joseph

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The Rank Products is a statistical technique widely used to detect differentially expressed features in molecular profiling experiments such as transcriptomics, metabolomics and proteomics studies. An implementation of the Rank Product (RP) and the closely related Rank Sum (RS) statistics has been available in the RankProd Bioconductor package for several years. However, several recent advances in the understanding of the statistical foundations of the method have made a complete refactoring of the existing package desirable.

Results: We implemented a completely refactored version of the RankProd package, which provides a more principled implementation of the statistics for unpaired datasets. Moreover, the permutation-based p-value estimation methods have been replaced by exact methods, providing faster and more accurate results.

Availability: RankProd 2.0 is available at Bioconductor (https://www.bioconductor.org/packages/devel/bioc/html/RankProd.html) and as part of the mzMatch pipeline (http://www.mzmatch.sourceforge.net).

Contact: rainer.breitling@manchester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Finding differentially expressed molecular features when comparing different conditions plays a pivotal role in all kinds of molecular profiling studies ("omics"). The Rank Product (RP) and the Rank Sum (RS) are two non-parametric statistics widely used to detect variables consistently upregulated (or downregulated) in replicate experiments (Breitling *et al.*, 2004; Breitling and Herzyk, 2005). Originally developed for the analysis of gene expression microarrays, both methods are more accurate and powerful than their usual competitors in a number of different scenarios (e.g. abnormally distributed noise, heterogeneity of samples, small fraction of changed features, small sample size), as demonstrated in extensive numerical studies (Breitling and Herzyk, 2005; Jeffery *et al.*, 2006; Koziol, 2010a,b). The main identified weakness of the RP method is its sensitivity to variable-specific measurement variance. Nevertheless, this problem has been successfully addressed by a number of variance stabilizing normalization techniques (Durbin *et al.*, 2002; Huber *et al.*, 2002; Breitling and Herzyk, 2005). An R Bioconductor package implementing RP and the closely related RS has been available and widely used for several years (Hong *et al.*, 2006). However, recent improvements in our understanding of the two statistics made a refactored version of the package desirable.

In the old implementation, the p-value estimation had been performed by a permutation-based method for both statistics (Hong *et al.*, 2006). This method requires a computationally demanding number of permutations in order to obtain accurate results and, when dealing with the tails of the distribution (i.e. the most interesting molecular features), the estimates are particularly unreliable. In RankProd 2.0, this limitation has been successfully tackled. Regarding the RP, the p-value estimation is now performed by applying the fast method proposed by Heskes *et al.* (2014). This tailor-made solution calculates strict bounds and very accurate approximate p-values for RP analysis. For the RS, a new exact method for the evaluation of the p-values has been developed and implemented as described in **Section 3**. The RP was initially introduced for the analysis of gene expression in paired datasets, specifically two-color microarrays (Breitling *et al.*, 2004). Nevertheless, the old RankProd package provided an ad hoc strategy to cope with unpaired datasets. Provided that unpaired datasets are increasingly common, we developed a more principled approach described in **Section 4**, which provides a more reliable application of RP and RS in the analysis of unpaired datasets.

2 P-values estimation for the Rank Product

The p-value estimation for the RP has been intensely studied in the last few years. Koziol (2010a, 2016) approximated the distribution of the RP with a gamma distribution. Such approximation resulted to be imprecise when dealing with the tails of the distribution (Eisinga *et al.*, 2013). Eisinga *et al.* (2013) derived the exact probability distribution of the RP statistic. Unfortunately, this is time-demanding and impractical to use with large datasets. For this reason, we chose the method proposed by Heskes *et al.* (2014), which allows a very accurate approximation of the p-values in a computationally fast manner. This method allows us to calculate strict bounds for the exact p-values and extremely accurate estimates by considering the geometric mean of the upper and lower bounds. This approach significantly speeds up the RP analysis. When considering a typical paired dataset ($N = 1000$ and $K = 10$), the computation time is now reduced by a factor of ~ 500, when compared with the analysis performed with the previous approach (using $10,000$ permutations).

3 P-values estimation for the Rank Sum

Previously, the only method available to estimate the p-values for the RS statistic was the permutation-based approach already implemented in the RankProd package (Hong *et al.*, 2006). Here we introduce a method for the exact calculation of the RS p-values. This is derived from the simple observation that under the null hypothesis, the probability distribution of the RS, in an experiment with N variables and K replicates, is exactly the same as the probability distribution of the sum of the outcomes obtained by rolling K dice with N faces (`http://mathworld.wolfram.com/Dice.html`). The implementation of this approach notably speeds up the RS analysis. When considering a typical paired dataset ($N = 1000$ and $K = 10$), the computation time is now reduced by a factor of ~ 1200, when compared with the analysis performed with the previous approach (using $10,000$ permutations). When the size of the dataset is such that the time needed to evaluate the exact p-values becomes unacceptable, the new package uses the exact distribution for the tails of the distribution only, whereas all the other p-values are evaluated through a very accurate Gaussian approximation. The extent of the tails and the threshold used to switch between the two strategies are determined by the heuristic rule described, together with the details of the calculation, in the **Supplementary material**.

4 Application to unpaired datasets

The previous version of the RankProd package provided an ad hoc approach to analyse unpaired datasets. This approach consists in considering all the possible pairs that can be obtained from the unpaired samples. Conversely, our new approach computes a user-defined number of random paired datasets and evaluates the RP (or RS) statistic per each of them. Each of these randomly paired datasets has the same size as if the experiment had originally been performed in a paired design. For each variable, the final RP (or RS) value returned is the median of all the values found during the random pairing process. The p-values are then computed as in the case of a paired experiment. A detailed description of this new approach can be found in the **Supplementary material**.

5 Conclusion

The RankProd 2.0 package provides a robust and reliable implementation of the Rank Products methods. Unpaired datasets are now handled through a new approach that significantly improves the performance of the methods. The p-value estimation for the RP is now faster and much more accurate, while for the RS we introduced a new and fast method able to evaluate the exact p-values. Full backward compatibility has been kept despite the complete refactoring. This improved implementation allows a more reliable application of these methods across the full spectrum of modern molecular profiling technologies. The new implementation of the method has also been integrated in the mzMatch pipeline (Scheltema *et al.*, 2011).

Funding

The authors acknowledge funding from the BBSRC under grant BB/M017702/1, “Centre for synthetic biology of fine and speciality chemicals”.

References

Breitling, R. and Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology*, **3**(5), 1171–1189.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, **573**(1), 83–92.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**(Suppl. 1), S105–S110.

Eisinga, R., Breitling, R., and Heskes, T. (2013). The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Letters*, **587**(6), 677–682.

Heskes, T., Eisinga, R., and Breitling, R. (2014). A fast algorithm for determining bounds and accurate approximate p-values of the rank product statistic for replicate experiments. *BMC Bioinformatics*, **15**(1), 367.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**(22), 2825–2827.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.

Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, **7**(1), 359.

Koziol, J. A. (2010a). Comments on the rank product method for analyzing replicated experiments. *FEBS Letters*, **584**(5), 941–944.

Koziol, J. A. (2010b). The rank product method with two samples. *FEBS Letters*, **584**(21), 4481–4484.

Koziol, J. A. (2016). A cautionary note on the rank product statistic. *FEBS Letters*, **590**(11), 1586–1591.

Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A., and Breitling, R. (2011). Peakml/mzmatch: a file format, java library, r library, and tool-chain for mass spectrometry data analysis. *Analytical Chemistry*, **83**(7), 2786–2793.