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.

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 Moreover, the permutation-based p-value estimation methods have been replaced by exact methods, providing faster and more accurate results.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; Kuziol, 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.

© The Author(s) 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
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.

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 specialty chemicals".

References


