RankProd

Hong, Fangxin; Breitling, Rainer; McEntee, Connor W.; Wittner, Ben S.; Nemhauser, Jennifer L.; Chory, Joanne

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btl476

Publication date:
2006

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 22 2006, pages 2825–2827
doi:10.1093/bioinformatics/btl476

Gene expression

RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis

Fangxin Hong[1,2,*], Rainer Breitling[3], Connor W. McEntee[1], Ben S. Wittner[4], Jennifer L. Nemhauser[5] and Joanne Chory[1,2]

[1]Plant Biology Laboratory, [2]Howard Hughes Medical Institute, The Salk Institute, La Jolla, CA, USA, [3]Groningen Bioinformatics Centre, University of Groningen, Haren, The Netherlands, [4]Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA and [5]Department of Biology, University of Washington, Seattle, WA, USA

Received on March 18, 2006; revised on August 22, 2006; accepted on September 3, 2006
Advance Access publication September 18, 2006
Associate Editor: Satoru Miyano

ABSTRACT

Summary: While meta-analysis provides a powerful tool for analyzing microarray experiments by combining data from multiple studies, it presents unique computational challenges. The Bioconductor package RankProd provides a new and intuitive tool for this purpose in detecting differentially expressed genes under two experimental conditions. The package modifies and extends the rank product method proposed by Breitling *et al.*, [(2004) *FEBS Lett.*, **573**, 83–92] to integrate multiple microarray studies from different laboratories and/or platforms. It offers several advantages over *t*-test based methods and accepts pre-processed expression datasets produced from a wide variety of platforms. The significance of the detection is assessed by a non-parametric permutation test, and the associated *P*-value and false discovery rate (FDR) are included in the output alongside the genes that are detected by user-defined criteria. A visualization plot is provided to view actual expression levels for each gene with estimated significance measurements.

Availability: RankProd is available at Bioconductor http://www.bioconductor.org. A web-based interface will soon be available at http://cactus.salk.edu/RankProd

Contact: fhong@salk.edu

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

1 INTRODUCTION

Microarray-based expression profiling has become a routine procedure in biological/medical studies. There are an increasing number of publicly available databases that provide a wealth of under-analyzed data from a wide variety of sources and treatments. For example, the Genevestigator database (Zimmermann *et al.*, 2004) includes data from nearly 2000 *Arabidopsis* microarrays, and public repositories like Gene Expression Omnibus and ArrayExpress (Parkinson *et al.*, 2005) are growing rapidly. Therefore, meta-analysis for combining data from multiple microarray experiments appears to be a good and practical idea. However, direct comparison among heterogeneous datasets is not possible as a result of the complicated experimental variables embedded in microarray experiments (Choi *et al.*, 2003; Irizarry *et al.*, 2005). Array datasets produced by two different laboratories using the same platforms have been shown to retain 'lab-effects' even after the normalization process (Vert *et al.*, 2005). Moreover, simultaneous normalization of heterogeneous datasets often violates the underlying assumptions of the very normalization method.

While meta-analysis can be adapted to various types of microarray analysis, the comparison of gene expression levels under two experimental conditions is the most widely used application. Recently, several meta-analysis applications have appeared in the literature (Choi *et al.*, 2003; Rhodes *et al.*, 2004). Most of these focused on combining results of individual studies rather than combining datasets into one analysis, thus they provide no overall estimates of the magnitude of differential expression. Moreover, those methods often involve sophisticated statistical models which lack biological intuition. Recently, we have contributed a RankProd package to the Bioconductor site, in which we presented a simple but powerful meta-analysis tool to detect differentially expressed genes by integrating multiple array datasets from various experimental platforms/settings across laboratories.

The RankProd package was developed from the rank product method which was initially proposed to detect differentially expressed genes in a single experiment (Breitling *et al.*, 2004). It is a non-parametric statistic derived from biological reasoning that detects items that are consistently highly ranked in a number of lists, for example genes that are consistently found among the most strongly unregulated (or down-regulated) genes in a number of replicate experiments. It offers several advantages over linear modeling, including the biologically intuitive of fold-change (FC) criterion, fewer assumptions under the model, and increased performance with noisy data and/or low numbers of replicates (Breitling and Herzyk, 2005). Moreover, the new method implemented in RankProd offers a natural way to overcome the heterogeneity among multiple datasets and therefore to extract, compare and integrate information from them. Since it transforms the actual expression values into ranks, the algorithm can integrate datasets produced by a wide variety of platforms, such as Affymetrix oligonucleotide arrays, two-color cDNA arrays and other custom-made arrays. It has the ability to handle variability among datasets

*To whom correspondence should be addressed.

and generates a single significance measurement for each gene. Therefore, it provides scientists a powerful tool to utilize the existing wealth data resources.

2 APPROACHES AND IMPLEMENTATION

The software RankProd is implemented in the statistical programming language R (http://www.r-project.org) as a package of the open-resource Bioconductor project (Gentleman *et al*., 2004). It accepts a pre-processed expression dataset in matrix format and provides functions to perform meta-analysis as well as the analysis of a single experiment.

Here we describe the meta-analysis algorithm implemented in RankProd using two datasets with different origins as the example. Let T and C stand for two experimental conditions (treatment versus control), and there are n_T and n_C replicates in the first dataset, m_T and m_C replicates in the second dataset.

1. For one-channel array, compute pair-wise ratios FC within each dataset $T_{n_1}/C_{n_1}, T_{n_1}/C_{n_2}, \ldots, T_{n_T}/C_{n_C} \Rightarrow n_T \times n_C$ comparisons $T_{m_1}/C_{m_1}, T_{m_1}/C_{m_2}, \ldots, T_{m_T}/C_{m_C} \Rightarrow m_T \times m_C$ comparisons. (For two-channel array, $T_{m_1}/C_{m_1}, \ldots, T_{m_T}/C_{m_C}$, $m_T = m_C$).
2. Rank ratio within each comparison (largest $\Rightarrow$ rank 1) $\Rightarrow r_{gi}$: rank of gth gene under ith comparison. $i = 1, \ldots, K$, where $K = (n_T \times n_C) + (m_T \times m_C)$.
3. Determine rank product for each gene as $RP_g = (\Pi_i\, r_{gi})^{1/K}$.
4. Independently permute expression value within each single array relative to gene ID, repeat step (1)–(3) $\Rightarrow RP_g^{(l)}$.
5. Repeat step (4) L times, form reference distribution with $RP_g^{(l)}$ ($l = 1, \ldots, L$), determine P-value and false discovery rate (FDR) associated with each gene.

One-channel experiments include Affymetrix gene-chip and two-color cDNA arrays with reference design; direct two-color cDNA arrays are usually two-channel experiments. The algorithm results in the identification of putative up-regulated genes within the treatment group compared with the control group. It then swaps the two groups to identify genes with opposite expression changes. The function `RPadvance` is used to perform such analysis.

```
RP.out <-RPadvance(data, cl, origin, rand = 123)
```

The function `topGene` outputs identified genes with the associated statistics into two tables, up-regulated and down-regulated genes in class 1 compared with class 2 (for details see package vignette). The package also provides a visualization tool `plotGene` to check the expression values and statistical results for each individual gene based on a user query (Fig. 1).

```
plotGene('245265_at', RP.out, data, cl, origin)
```

From Figure 1, it is clear that the FCs is similar across the two datasets although the actual expression values are quite different. The statistics printed on the plot indicate that probeset '245265_at' is potentially up-regulated in class 2 compared with class 1 with an estimated P-value < 0.001 and pfp(or FDR) = 0.001. There are additional functions available in the RankProd package for performing the modified rank sum test (Breitling and Herzyk, 2005). These are described in the package vignette.

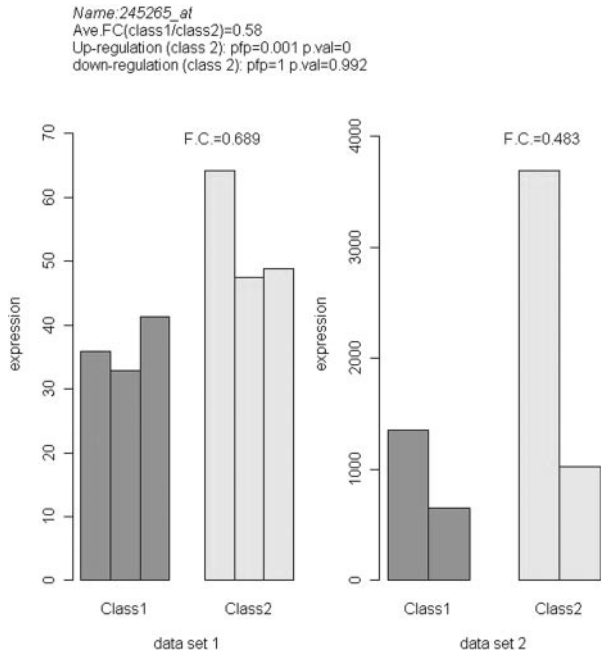

Fig. 1. Expression levels and identification statistics for a probeset named '245265_at' in *Arabidopsis* Affymetrix ATH1 chip. The FCs in the two datasets are 0.689 and 0.483; the combined FC is 0.58. The gene is detected as up-regulated in class 2 with P-value < 0.0001 and pfp = 0.001.

3 APPLICATION

RankProd has been used for detecting differentially expressed genes in various studies (Gurvich *et al*., 2005; Vert *et al*., 2005; Wilson *et al*., 2006). Indeed, the theory behind the method is easily understood and the results have been shown to be more biologically relevant than those of other methods, especially in studies with a low number of replicates (Breitling *et al*., 2004). We have employed RankProd for various meta-analyses, such as the study of the effect of a plant hormone using two datasets produced in two different laboratories (data used in Fig. 1; Vert *et al*., 2005). Two laboratories treated plants with the same hormone but at different concentrations and time intervals (Fig. 1). The analysis was able to identify many more genes by combining two datasets into one analysis than by analyzing each dataset individually (Package vignette). Moreover, the genes identified by the meta-analysis tend to have more overlap with genes identified in other studies, suggesting an increased reliability (See Supplementary Figure 1).

4 DISCUSSION

RankProd provides a simple, yet powerful meta-analysis tool for detecting differentially expressed genes between two experimental conditions. The approach overcomes the heterogeneity among multiple datasets and naturally combines them to achieve increased sensitivity and reliability. It is worth pointing out that it does not require the simultaneous normalization of multiple datasets, which solves a frequently encountered dilemma in microarray pre-processing step. Therefore, this new tool provides researchers

a way to take advantage of the rapidly growing amount of publicly available array data. This can even be extended across species by using ortholog identification approaches. RankProd can also be applied to proteomic and metabolomic studies where ranked lists of changed proteins or metabolites are produced by 2D-gels or mass spectrometry. To further increase the versatility of our approach, we are currently constructing a web-based tool to perform rank product analyses.

ACKNOWLEDGEMENTS

The authors would like to thank Todd C. Mockler and Todd P. Michael for critical discussion and useful comments. Our studies are supported by the National Science Foundation and the Howard Hughes Medical Institute.

Conflict of Interest: none declared.

REFERENCES

Breitling,R. *et al*. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*., **573**, 83–92.

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. *J. Bioinf. Comp. Biol*., **3**, 1171–1189.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Gurvich,N. *et al*. (2005) Association of valproate-induced teratogenesis with histone deacetylase inhibition *in vivo*. *FASEB J*., **19**, 1166–1168.

Nembauser,L.J. *et al*. (2006) Plant hormones regulate similar processes through largely non-overlapping transcriptional responses. *Cell*, **126**, 467–475.

Irizarry,R.A. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Meth*., **2**, 345–349.

Parkinson,H. *et al*. (2005) ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **33**, D553–D555.

Rhodes,D.R. *et al*. (2004) Large-scale meta-analysis of cancer microarray data identified common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Meth. Genet. Mol. Biol*., **3**, 3.

Vert,G. *et al*. (2005) Molecular mechanisms of steroid hormone signalling in plants. *Annu. Rev. Cell. Dev. Biol*., **21**, 177–201.

Wilson,C.L. *et al*. (2006) Effects of oestrogen on gene expression in epithelium and stroma of normal human breast tissue. *J. Endocrinol*., in press.

Zimmermann,P. *et al*. (2004) GENEVESTIGATOR. *Arabidopsis* Microarray Database and Analysis Toolbox. *Plant Physiol*., **136**, 2621–2632.