BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 12 2008, pages 1461–1462
doi:10.1093/bioinformatics/btn209

Gene expression

fdrtool: a versatile R package for estimating local and tail area-based false discovery rates

Korbinian Strimmer

Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Received on December 12, 2007; revised on January 28, 2008; accepted on April 23, 2008

Advance Access publication April 25, 2008

Associate Editor: Trey Ideker

ABSTRACT

Summary: False discovery rate (FDR) methodologies are essential in the study of high-dimensional genomic and proteomic data. The R package 'fdrtool' facilitates such analyses by offering a comprehensive set of procedures for FDR estimation. Its distinctive features include: (i) many different types of test statistics are allowed as input data, such as *P*-values, *z*-scores, correlations and *t*-scores; (ii) simultaneously, both local FDR and tail area-based FDR values are estimated for all test statistics and (iii) empirical null models are fit where possible, thereby taking account of potential over- or under-dispersion of the theoretical null. In addition, 'fdrtool' provides readily interpretable graphical output, and can be applied to very large scale (in the order of millions of hypotheses) multiple testing problems. Consequently, 'fdrtool' implements a flexible FDR estimation scheme that is unified across different test statistics and variants of FDR.

Availability: The program is freely available from the Comprehensive R Archive Network (http://cran.r-project.org/) under the terms of the GNU General Public License (version 3 or later).

Contact: strimmer@uni-leipzig.de

1 INTRODUCTION

Multiple testing is often an essential step in the analysis of high-dimensional genomic or proteomic data. In this context, false discovery rates (FDR) have proven to be reliable as statistical criteria to determine the significance of genomic features. Correspondingly, FDR methodologies are currently employed in many settings such as differential expression, spectrometric peak detection, SNP discovery, edge selection in genetic networks, to name but a few examples.

FDR theory starts with the seminal papers by Schweder and Spjøtvoll (1982) and Benjamini and Hochberg (1995). Local FDR was introduced by Efron *et al*. (2001). For a general overview over FDR methodologies see, e.g. the review of Broberg (2005) and Efron (2004, 2007).

Generally, two distinct types of FDR need be distinguished:

(1) density-based local FDR (= 'fdr'), and

(2) tail area-based FDR (= 'Fdr').

Intuitively, tail area-based FDR is simply a *P*-value corrected for multiplicity, whereas local FDR is a corresponding probability value.

More formally, consider an observed test statistic $y \geq 0$ designed such that a small y indicates an 'uninteresting' null case, and conversely, a large y an 'interesting' alternative case. It is assumed that across hypotheses $i = 1, \ldots, m$ the test statistics y_i follow a two-component mixture, with density

$$f(y) = \eta_0 f_0(y|\theta) + (1 - \eta_0) f_A(y) \quad (1)$$

and distribution function

$$F(y) = \eta_0 F_0(y|\theta) + (1 - \eta_0) F_A(y). \quad (2)$$

The local and tail area-based FDR are then defined as follows:

$$\mathrm{fdr}(y) = \Pr(\text{'uninteresting'} \,|\, Y = y) = \eta_0 \frac{f_0(y|\theta)}{f(y)} \quad (3)$$

and

$$\mathrm{Fdr}(y) = \Pr(\text{'uninteresting'} \,|\, Y \geq y) = \eta_0 \frac{1 - F_0(y|\theta)}{1 - F(y)}. \quad (4)$$

In order to estimate FDR one proceeds by fitting the above mixture model to the observed data. This involves identifying the alternative model (f_A and F_A) and finding suitable estimates for the proportion of null values η_0 and for the parameters θ. Note that this is not an easy task, and that this is precisely the point where the various FDR algorithms differ.

While both 'Fdr' and 'fdr' are defined for any arbitrary test statistic, it is common practice to use *P*-values for estimating 'Fdr', and *z*-scores for 'fdr' computations.

2 DISTINCTIVE FEATURES OF 'FDRTOOL'

In contrast to other FDR estimation schemes, in 'fdrtool' there is no unnecessary distinction between *P*-values and other test statistics. Instead, one common algorithm is used to fit the mixture distribution and to infer its parameters. Currently, null models are implemented for *P*-values, *z*-scores, correlations and *t*-scores. It is straightforward to extend 'fdrtool' to allow further types of test statistics.

A second distinguishing feature of 'fdrtool' is that, regardless of the choice of test statistic, simultaneously both local FDR as well as tail area-based FDR values are estimated. This enables, e.g. the computation of local FDR from *P*-values, and also ensures that $\widehat{\mathrm{Fdr}}(y_i) \leq \widehat{\mathrm{fdr}}(y_i)$.

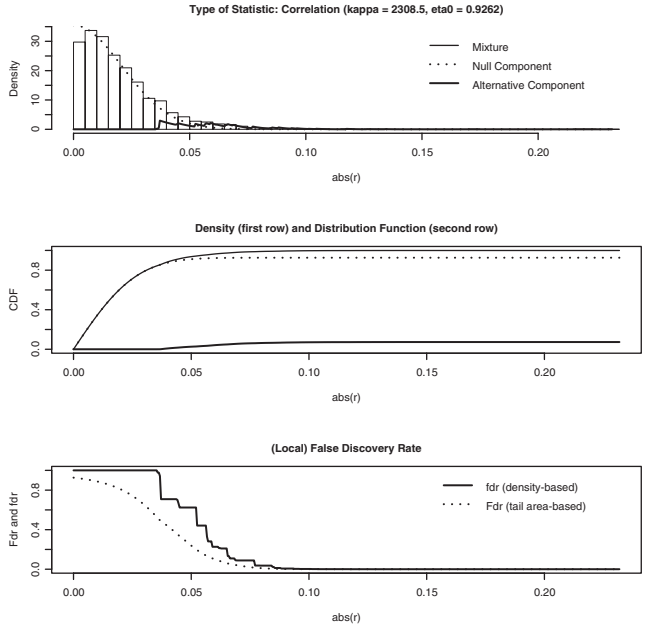

Fig. 1. Typical graphical output of the function `fdrtool`. In this example the input test statistics are correlations. Other possible test statistics are *t*-scores, *z*-scores and *P*-values. The first row shows the histogram and the density of the fitted two-component model. Also indicated are the value of estimated parameters, in this case the proportion of null values η_0 and the effective degree of freedom κ of the correlations. The second row depicts the corresponding cumulative density functions. In the third row the local FDR as well as the tail area-based FDR are shown in dependence of the value of the test statistic. Note that the default output is in color, but if desired (as in this figure) a black & white version can be produced by invoking the option `color.figure=FALSE`.

Furthermore, all null models may contain free parameters, typically related to scale or location. This implies that 'fdrtool' facilitates the use of an empirical null model (Efron, 2004; Schäfer and Strimmer, 2005). This is beneficial if hypotheses are correlated, and if there is an over- or underdispersion of the theoretical null model (Efron, 2007).

The learning algorithm employed in 'fdrtool' merges the Grenander-density approaches (Broberg, 2005; Langaas *et al.*, 2005) with empirical null modeling (Efron, 2004). Precise details of this procedure and its statistical properties will be reported elsewhere (Strimmer, 2008, manuscript in preparation).

3 AN EXAMPLE SESSION

FDR analysis with 'fdrtool' is simple: start the R application (R Development Core Team, 2007), arrange the test statistics in vector format, and run the `fdrtool` command. In the following example `r` is a vector of correlations:

```
library(''fdrtool'')
fdr.out = fdrtool(r, type=''correlation'')
```

The resulting graphical output is shown in Figure 1. The actual estimated (local) FDR values can be accessed as follows:

```
fdr.out$pval # p-values
fdr.out$lfdr # local FDR (=fdr)
fdr.out$qval # tail area-based FDR (=Fdr)
fdr.out$param # estimated parameters
```

The manual accompanying the 'fdrtool' R package documents this and a number of other procedures in more detail.

4 CONCLUSION

'fdrtool' is a flexible and simple to use software package for the R environment that allows to obtain estimates of local FDR and frequentist FDR, with a unified interface and algorithm for a diverse set of test statistics and variants of FDR.

ACKNOWLEDGEMENTS

I thank Brit B. Turnbull, Stanford University, for valuable discussion of the local FDR estimation procedure implemented in the R package 'locfdr', and for kindly sharing an unpublished preprint.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Broberg,P. (2005) A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, **6**, 199.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.*, **99**, 96–104.

Efron, B. (2007) Correlation and large-scale simultaneous significance tesing. *J. Amer. Statist. Assoc.*, **102**, 93–103.

Efron,B. *et al.* (2001) Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, **96**, 1151–1160.

Langaas,M. *et al.* (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Statist. Soc. B*, **67**, 565–572.

R Development Core Team (2007) *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schäfer,J. and Strimmer,K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Schweder,T. and Spjøtvoll,E. (1982) Plots of *p*-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.