ROTS: Reproducibility Optimized Test Statistic

Fatemeh Seyednasrollah, Tomi Suomi, Laura L. Elo

March 3, 2016

Contents

1 Introduction

Differential expression testing is perhaps the most common approach among current omics analysies. Reproducibility optimized test statistic (ROTS) aims to rank genomic features of interest (such as genes, proteins and transcripts) in order of evidence for differential expression in two-group comparisons. Initially, ROTS was developed to test differential expression in microarray studies (Elo 2008). However, the general design of the algorithm supports the utility of the method in proteomics and count-based technologies like RNA-seq and single cell datasets (Seyednasrollah et al. 2015, Pursiheimo et al. 2015). ROTS is a data adaptive method which can optimize its parameters based on intrinsic features of input data. Also, the method aims to solve the common problem of small sample size through a resampling procedure.

The ROTS statistic is optimized among a family of *t-type* statistics $d = r/(\alpha_1 + \alpha_2 \times s)$, where r is the absolute difference between the group averages $|\bar{x_1} - \bar{x_2}|$, s is the pooled standard error, and α_1 and α_2 are the non-negative parameters to be optimized. Two special cases of this family are the ordinary *t-statistic* ($\alpha_1 = 0, \alpha_2 = 1$) and the signal log-ratio ($\alpha_1 = 1, \alpha_2 = 0$).

The optimality is defined in terms of maximal overlap of top-ranked features in group-preserving bootstrap datasets. Importantly, besides the group labels, no a priori information about the properties of the data is required and no fixed cutoff for the gene rankings needs to be specified. The user is given the option to adjust the largest top list size (K) considered in the reproducibility calculations, since lowering this size can markedly reduce the computation time. In large data matrices with thousands of rows, we generally recommend using a size of several thousands.

ROTS tolerates a moderate number of missing values in the data matrix by effectively ignoring their contribution during the operation of the procedure. However, each row of the data matrix must contain at least two values in both groups. The rows containing only a few non-missing values should be removed; or alternatively, the missing data entries can be imputed using, e.g., the K-nearest neighbour imputation, which is implemented in the Bioconductor package impute. If the parameter values α_1 and α_2 are set by the user, then no optimization is performed but the statistic and FDR-values are calculated for the given parameters. The false discovery rate (FDR) for the optimized test statistic is calculated by permuting the sample labels. The results for all the genes can be obtained by setting the FDR cutoff to 1.

2 Algorithm overview

ROTS optimizes the reproducibility among a family of modified statistics:

$$d_\alpha = \frac{r}{\alpha_1 + \alpha_2 \times s} \tag{1}$$

where r is a score, α_1 and α_2 are non-negative parameters to be optimized, and s is standard deviation.

The optimal statistic is determined by maximizing the reproducibility Z-score:

$$Z_k\left(d_\alpha\right) = \frac{R_k\left(d_\alpha\right) - R_k^0\left(d_\alpha\right)}{s_k\left(d_\alpha\right)} \tag{2}$$

over a dence lattice $\alpha_1 \in [0, 0.01, ..., 5]$, $\alpha_2 \in \{0, 1\}$, $k \in \{1, 2, ..., G\}$. Here, $R_k\left(d_\alpha\right)$ is the observed reproducibility at top list size k, $R_k^0\left(d_\alpha\right)$ is the corresponding reproducibility in randomized datasets (permuted over samples), $s_k\left(d_\alpha\right)$ is the standard deviation of the bootstrap distribution, and G is the total number of genes/proteins in the data. Reproducibility is defined as the average overlap of k top-ranked features over pairs of bootstrapped datasets.

In two-group comparisons, ROTS optimizes the reproducibility of top-ranked features in group-preserving bootstrap datasets among a family of modified t-statistics, where the score r is the absolute difference between the group averages and s is the pooled standard error:

$$r = |\bar{x_1} - \bar{x_2}| \tag{3}$$

$$s = \left[\frac{\sum_{i \in C_1}(x_i - \bar{x}_1)^2 + \sum_{i \in C_2}(x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}(1/n_1 + 1/n_2)\right]^{1/2} \tag{4}$$

where i has the indices of observations in classes C_1 and C_2, and n_1 and n_2 are the number samples in classes 1 and 2, respectively.

In multi-group comparisons, ROTS optimizes the reproducibility of top-ranked features in group-preserving bootstrap datasets among a family of modified f-statistics:

$$r = \left[\left\{\sum n_c / \prod n_c\right\} \sum_{c=1}^{C} n_c \left(\bar{x}_c - \bar{x}\right)^2\right]^{1/2} \tag{5}$$

$$s = \left[\frac{1}{\sum\left(n_c - 1\right)}\left(\sum \frac{1}{n_c}\right) \sum_{c=1}^{C} \sum_{i \in C_c} \left(x_i - \bar{x}_c\right)^2\right]^{1/2} \tag{6}$$

where c is the different classes $\{1, 2, ..., C\}$, n_c is the number samples in class c, and i has the indices of observations in class C_c.

In survival analysis, ROTS optimizes the reproducibility of top-ranked features among Cox scores:

$$r = \sum_{t=1_1}^{T} [x_{D_t} - d_t \bar{x}_t] \tag{7}$$

$$s = \left[\sum_{t=t_1}^{T} (d_t/k_t) \sum_{i \in R_t} (x_i - \bar{x}_t)^2 \right]^{1/2} \tag{8}$$

where D_t is indices of observations at the different death times $\{t_1, t_2, ..., T\}$, R_t indices of the observations at risk at these times, and d_t and k_t the number of deaths and individuals at risk, at the time, respectively.

For more detailed information about the ROTS algorithm, see Elo et al. (2008) and Seyednasrollah et al. (2015).

3 Input data

ROTS expects the input data to be in form of a matrix with genomic features as rows and samples as columns. It is recommended to use normalized data as the input for ROTS. The matrix can be either of integer numbers, e.g. for RNA-seq and single cells, or float numbers, e.g. microarray intensities.

4 Preprocessing

For count-based data, we recommend the widely used preprocessing techniques like TMM (Trimmed Mean of M-values) normalization available in edgeR Bioconductor package or TMM normalization plus Voom transformation available in Limma Bioconductor package. For microarray and proteomics data, standard normalization techniques are recommended.

5 Differential expression testing

We use here a proteomics dataset as an example for differential expression testing. The overall approach is the same for other omics data along with recommended preprocessing strategies.

The analysis starts by loading the ROTS package and the example dataset, which contains two sample groups each having three replicates:

```
> library(ROTS)
> data(upsSpikeIn)
> input = upsSpikeIn
```

In the next step we determine the experimental design for differential expression analysis. Please note that the order of the samples in the data matrix should be exactly the same as the groups vector defined.

```
> groups = c(rep(0,3), rep(1,3))
> groups

[1] 0 0 0 1 1 1
```

The ROTS function performs the final differential expression testing. The user can set the function parameters before running the analysis:

```
> results = ROTS(data = input, groups = groups , B = 100 , K = 500 , seed = 1234)
> names(results)

 [1] "data"   "B"      "d"      "logfc"  "pvalue" "FDR"    "a1"     "a2"
 [9] "k"      "R"      "Z"      "ztable" "cl"
```

In this example, we set the number of bootstrapping (B) and the number of top-ranked features for reproducibility optimization (K) to 100 and 500 respectively, to reduce running time of the example. In real analysis it is preferred to use a higher number of bootstraps (e.g. 1000). The optimization parameters a1 and a2 should always be non-negative. The output of ROTS function includes test statistic (d), estimated p-value (pvalue), False Discovery Rate (FDR), optimized test statistic parameters and top list size (a1, a2, k), optimized reproducibility value (R) and Z-score (Z). In general, the Z-score and reproducibility are the main indicators to decide the success of differential expression analysis. As a rule of thumb, reproducibility Z-scores below 2 indicate that the data or the statistics are not sufficient for reliable detection.

Finally, it is possible to summarize the results based on criteria selected by the user. For instance, the following code lists the top ranked differentially expressed features with FDR below 0.05:

```
> summary(results, fdr = 0.05)

ROTS results:

Number of resamplings:  100

a1:                     1.4
a2:                     1
Top list size:          20
```

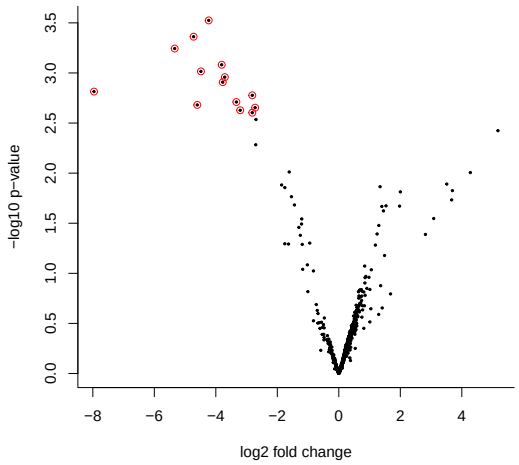
```
Reproducibility value:  0.7945
Z-score:                4.530883

14 rows satisfy the condition. Only ten first rows are
 displayed, see the return value for the whole output.
          Row ROTS-statistic       pvalue FDR
P51965ups  13       2.598928 0.0002989130   0
P00441ups  24       2.315927 0.0004347826   0
P08758ups  16       2.176674 0.0005706522   0
O00762ups  19       2.139412 0.0008288043   0
P01375ups  22       2.059381 0.0009646739   0
P06396ups  27       2.001048 0.0011005435   0
P07339ups  11       1.958776 0.0012364130   0
P15559ups  80       1.864312 0.0015353261   0
P08263ups   1       1.802017 0.0016711957   0
P08311ups   9       1.698728 0.0019497283   0
...
```

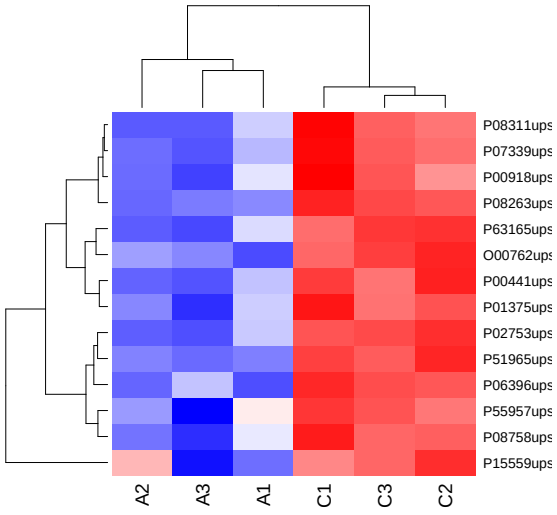
6 Visualization

Results can also be visualized using the standard plot command:

```
> plot(results, fdr = 0.05, type = "volcano")
```




```
> plot(results, fdr = 0.05, type = "heatmap")
```



7 References

Elo, L.L. et al., *Reproducibility-optimized test statistic for ranking genes in microarray studies.* IEEE/ACM transactions on computational biology and bioinformatics, 5, 423-431, 2008.

Elo, L.L. et al. *Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets.* Briefings in bioinformatics, 10, 547-555, 2009.

Seyednasrollah et al. *ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer.* Nucleic Acids Research, 2015.

Pursiheimo et al. *Optimization of Statistical Methods Impact on Quantitative Proteomics Data.* J. Proteome Res., 2015.