prot2D : Statistical Tools for volume data from 2D Gel Electrophoresis

Sébastien Artigaud *

April 30, 2018

Contents

1 Introduction

This document briefly describes how to use the *prot2D* package. An R package designed to analyze (i.e. Normalize and select significant spots) data issued from 2D SDS PAGE experiments. *prot2D* provides a simple interface for analysing data from 2D gel experiments. Functions for normalization as well as selecting significant spots are provided. Furthermore, a function to simulates realistic 2D Gel volume data is also provided.If you use this package please cite :

- Artigaud, S., Gauthier, O. & Pichereau, V. (2013) "Identifying differentially expressed proteins in two-dimensional electrophoresis experiments: inputs from transcriptomics statistical tools." Bioinformatics, vol.29 (21): 2729-2734.

*Laboratoire des Sciences de l'Environnement Marin, LEMAR UMR 6539, Université de Bretagne Occidentale, Institut Universitaire Européen de la Mer, 29280 Plouzané, France

Table 1: Example of input data for prot2D with 3 replicates gels in each condition

	Replicates Condition 1			Replicates Condition 2		
	$Gel1$	$Gel2$	$Gel3$	$Gel1'$	$Gel2'$	$Gel3'$
$Spot_1$	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{1',1}$	$X_{2',1}$	$X_{3',1}$
$Spot_2$	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{1',2}$	$X_{2',2}$	$X_{3',2}$
$Spot_i$	$X_{1,i}$	$X_{2,i}$	$X_{3,i}$	$X_{1',i}$	$X_{2',i}$	$X_{3',i}$

2 Input data for *prot2D*

prot2D uses raw 2D volume data intensities as input, datasets must be exported from specialized Image Software in the form of a dataframe of volume data $X_{j,i}$ with gels j as columns and spots i as rows. Note that the name of columns should therefore corresponds to the names of the gels and the names of the rows to the name of the spots. The replicates for each condition should be ordered in the following columns (see Table 1). Furthermore, another dataframe is needed to describe the experiment with the names of gels as rownames and a single column giving the two level of condition for data.

3 Getting started

To load the *prot2D* package into your R envirnoment type:

```
> library(prot2D)
```

In this tutorial we will be using the `pecten` dataset obtained from a 2-DE experiment performed on proteins from the gills of *Pecten maximus* subjected to a temperature challenge. 766 spots were identified with 6 replicates per condition, therefore the dataset is a dataframe of 766 rows and 12 columns (for details, see Artigaud et al., 2013). `pecten.fac` describe the data by giving the names of the gels (as rownames) and the condition for the temperature challenge (15C = control vs 25C) in the "Condition" column, load by typing:

```
> data(pecten)
> data(pecten.fac)
```

4 *prot2D* workflow

4.1 Vizualize and Normalize volume data

Vizualization

Dudoit et al. (2002) proposed a method for the visualization of artifacts in microarray datasets, called the MA-plot, which was transposed for proteomics data as the Ratio-Intensity plot (Meunier et al., 2005; R-I plot).

It consists in plotting the intensity log_2-ratio (R) against mean log_{10} intensity (I):

$$R = log_2 \frac{mean(V_{Cond2})}{mean(V_{Cond1})} \quad (1)$$

The intensity (I) is the log10 of the mean of volume data in condition 2 by the mean of volume data in condition 1 :

$$I = log_{10}(mean(V_{Cond2}) \times mean(V_{Cond1})) \quad (2)$$

Where V_{Cond1} and V_{Cond2} are spot volumes for conditions 1 and 2, respectively. In *prot2D*, R-I plot can be easily displayed with `RIplot`:

```
> RIplot(pecten, n1=6, n2=6)
```

`RIplot` requires data supplied as a dataframe or a matrix as well as the number of replicates for each conditions.

Normalization

2D Gel Volume data must be normalized in order to remove systemic variation prior to data analysis. Two widely used methods are provided, the "Variance Stabilizing Normalization" (vsn) and the "Quantiles, Normalization". The principle of the "quantiles normalization" is to set each quantile of each column (i.e. the spots volume data of each gels) to the mean of that quantile across gels. The intention is to make all the normalized columns have the same empirical distribution. Whereas the vsn methods relies on a transformation h, of the parametric form *h(x)= arsinh(a+bx)* (Huber et al., 2002). The parameters of h together with those of the calibration between experiments are estimated with a robust variant of maximum-likelihood estimation. Both methods recentered the data around a zero log ratio, nevertheless for low values of intensities the vsn normalized data seems to be less efficient in order to recentered the cloud of points. Users are thus advised to use the quantiles normalization via a call to `Norm.qt`.

```
> pecten.norm <- Norm.qt(pecten, n1=6, n2=6, plot=TRUE)
```

```
>
```

4.2 Coerce data into an ExpressionSet

Prior to analysis for finding differentially expressed proteins, data must be coerced into an `ExpressionSet`. This can be done easily with `ES.prot`, which requires a matrix of normalized volume data, the number of replicates in each condition and a dataframe giving the condition for the experiment.

```
> ES.p <- ES.prot(pecten.norm, n1=6, n2=6, f=pecten.fac)
```

The matrix of spots intensities (i.e. Volume) is log2 transformed and stored in the `assayData` slot of the `ExpressionSet`. Furthermore, the log2-ratio is computed and stored in the `featureData` slot.

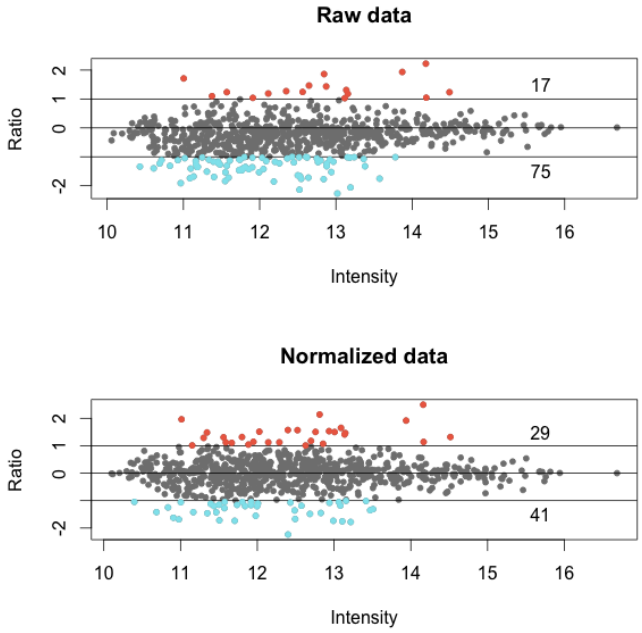

Figure 1: Ratio-Intensity plot showing Quantiles normalization

4.3 Find differentially expressed proteins

As described in Artigaud et al (2013) *prot2D* provide functions adapted from microarray analysis (from the *st, samr, limma, fdrtool* package). 2-DE experiments analysis require a variant of the t-statistic that is suitable for high-dimensional data and large-scale multiple testing. For this purpose, in the last few years, various test procedures have been suggested. *prot2D* provides:

- the classical Student's t-test (in `ttest.Prot` function)
- two tests especially modified for micro-array analysis : Efron's t-test (2001; `efronT.Prot` function) and the modified t-test used in Significance Analysis for Microarray (Tusher et al, 2001; `samT.Prot` function).
- two methods that take advantage of hierarchical Bayes methods for estimation of the variance across genes: the "moderate t-test" from Smyth (2004; `modT.Prot` function) and the "Shrinkage t" statistic test from Opgen-Rhein and Strimmer (2007; `shrinkT.Prot` function).

As statistical tests allowing the identification of differentially expressed proteins must take into account a correction for multiple tests in order to avoid false conclusions. *prot2D* also provide different methods to estimate the False Discovery Rate :

- the classical FDR estimator of Benjamini and Hochberg (1995).

– the local FDR estimator of Strimmer (2008).
– the "robust FDR" estimator of Pounds and Cheng (2006).

`ttest.Prot` function, `modT.Prot` function, `samT.Prot` function, `efronT.Prot` function or `shrinkT.Prot` function can be used to find differentially expressed proteins and the different FDR mode of calculation are implemented with the `method.fdr` argument. However, the moderate t-test with the FDR correction of Benjamini and Hochberg was find to be the best combination in terms of sensitivity and specificity. Thus, users are advised to use this combination by typing :

```
> ES.diff <- modT.Prot(ES.p, plot=TRUE, fdr.thr=0.1,
+                      method.fdr="BH" )

Number of up-regulated spots in Condition 2
[1] 0
Number of down-regulated spots in Condition 2
[1] 1
```

The function returns an `ExpressionSet` containing only the spots declared as significant. A plot can also be generated to vizualize the FDR cut-off. Additionally, it can be usefull to select only the spots with an absolute ratio greater than 2, as they are often considered as the most biologically relevant proteins, this can be done by adding the command `Fold2=T`. The names of the selected spots can be retrieved with :

```
> featureNames(ES.diff)

[1] "2607"
```

Displaying fold change (as log2(ratio)) for selected spots

```
> head(fData(ES.diff))

         ratio
2607 -1.912657
```

Volume normalized data for selected spots

```
> head(exprs(ES.diff))

     Br_23865 Br_23883 Br_23884 Br_23728 Br_23729 Br_23730 Br_23731 Br_23732
2607 23.21036   23.454 20.22889 22.34141 23.07774  23.0274 22.63096 20.07119
     Br_23733 Br_23875 Br_23876 Br_23877
2607 21.07323 20.52232 20.02905 19.53711
```

5 Simulation of 2D Volume data

In order to compare FDR and the responses of the different tests as well as the influence of the number of replicates, simulated data ca be used. `Sim.Prot.2D` simulates realistic 2D Gel volume data, based on parameters estimate from real dataset. Volume data are computed following these steps (see Smyth, 2004 and Artigaud et al., 2013 for details) :

- Log2 mean volumes from data are computed for each spot.
- Means are used as input parameters in order to simulate a normal distribution (with no differential expression between conditions) for each spot with standard deviations computed as described by Smyth (2004).
- A define proportion `p0` of the spots are randomly picked for introducing differential expression in both conditions (`p0`/2 in each condition).

The `Sim.Prot.2D` returns an `ExpressionSet` of simulated volume data (log2 transformed) with 2 conditions ("Cond1" and "Cond2" in `phenoData`) slot of the `ExpressionSet`.The spots differentially generated can be retrieve with `notes`. Simulate data based on "pecten"

```
> Sim.data <- Sim.Prot.2D(data=pecten, nsp=700,
+                         nr=10, p0=0.1, s2_0=0.2, d0=3)
```

Compare different methods for finding differentially expressed proteins

```
> res.stud <- ttest.Prot(Sim.data, fdr.thr=0.1, plot=FALSE)

Number of up-regulated spots in Condition 2
[1] 14
Number of down-regulated spots in Condition 2
[1] 14

> res.mo <- modT.Prot(Sim.data, fdr.thr=0.1, plot=FALSE)

Number of up-regulated spots in Condition 2
[1] 29
Number of down-regulated spots in Condition 2
[1] 28
```

Names of the spots selected by student's t-test with an FDR of 0.1

```
> featureNames(res.stud)

 [1] "15"  "20"  "21"  "39"  "58"  "93"  "153" "155" "193" "199" "213" "217"
[13] "239" "295" "304" "312" "346" "366" "428" "471" "478" "491" "500" "544"
[25] "608" "617" "641" "669"
```

Names of the spots selected by modT-test with an FDR of 0.1

```
> featureNames(res.mo)

 [1] "17"  "20"  "21"  "32"  "39"  "41"  "58"  "93"  "101" "106" "155" "159"
[13] "193" "199" "217" "224" "238" "258" "259" "283" "295" "312" "329" "334"
[25] "346" "354" "355" "361" "363" "364" "375" "376" "417" "428" "436" "454"
[37] "457" "471" "476" "478" "481" "491" "500" "544" "557" "559" "561" "570"
[49] "580" "587" "608" "617" "627" "641" "669" "691" "692"
```

Names of the differentially generated spots

```
> notes(Sim.data)$SpotSig
```

```
 [1] "2"   "17"  "20"  "24"  "32"  "58"  "101" "106" "155" "199" "217" "258"
[13] "283" "295" "361" "363" "417" "428" "471" "481" "491" "500" "504" "557"
[25] "560" "570" "575" "610" "637" "641" "700" "4"   "21"  "39"  "41"  "93"
[37] "193" "224" "238" "312" "313" "346" "355" "364" "376" "426" "457" "476"
[49] "478" "544" "561" "580" "587" "608" "610" "617" "669" "691" "692"
```

6 Session Info

```
> sessionInfo()

R version 3.5.0 (2018-04-23)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.4 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.7-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.7-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] prot2D_1.18.0       qvalue_2.12.0       MASS_7.3-50
 [4] Mulcom_1.30.0       fields_9.6          maps_3.3.0
 [7] spam_2.1-4          dotCall64_0.9-5.2   limma_3.36.0
[10] Biobase_2.40.0      BiocGenerics_0.26.0 samr_2.0
[13] matrixStats_0.53.1  impute_1.54.0       st_1.2.5
[16] sda_1.3.7           corpcor_1.6.9       entropy_1.2.1
[19] fdrtool_1.2.15

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.16     magrittr_1.5     splines_3.5.0    statmod_1.4.30
 [5] munsell_0.4.3    colorspace_1.3-2 rlang_0.2.0      stringr_1.3.0
 [9] plyr_1.8.4       tools_3.5.0      gtable_0.2.0     lazyeval_0.2.1
[13] tibble_1.4.2     reshape2_1.4.3   ggplot2_2.2.1    stringi_1.1.7
[17] pillar_1.2.2     compiler_3.5.0   scales_0.5.0
```

7 References

Artigaud, S., Gauthier, O. & Pichereau, V. (2013) "Identifying differentially expressed proteins in two-dimensional electrophoresis experiments: inputs from transcriptomics statistical tools." Bioinformatics, vol.29 (21): 2729-2734.

Benjamini, Y. & Hochberg, Y. (1995) "Controlling the false discovery rate: a practical and powerful approach to multiple testing" Journal of the Royal Statistical Society. Series B. Methodological.: 289-300.

Dudoit, S., Yang, Y.H., Callow, M.J., & Speed, T.P. (2002) "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments" Statistica Sinica, vol. 12: 111-139.

Efron, B., Tibshirani, R., Storey, J.D., & Tusher, V. (2001) "Empirical Bayes Analysis of a Microarray Experiment" Journal of the American Statistical Association, vol. 96 (456): 1151-1160.

Huber, W., Heydebreck, von, A., Sultmann, H., Poustka, A., & Vingron, M. (2002) "Variance stabilization applied to microarray data calibration and to the quantification of differential expression" Bioinformatics, vol. 18 (Suppl 1): S96-S104.

Meunier, B., Bouley, J., Piec, I., Bernard, C., Picard, B., & Hocquette, J.-F. (2005) "Data analysis methods for detection of differential protein expression in two-dimensional gel electrophoresis" Analytical Biochemistry, vol. 340 (2): 226-230.

Opgen-Rhein, R. & Strimmer, K. (2007) "Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach" Statistical Applications in Genetics and Molecular Biology, vol. 6 (1).

Pounds, S. & Cheng, C. (2006) "Robust estimation of the false discovery rate" Bioinformatics, vol. 22 (16): 1979-1987.

Smyth, G.K. (2004) "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Statistical Applications in Genetics and Molecular Biology, vol. 3: Article3.

Strimmer, K. (2008) "A unified approach to false discovery rate estimation." BMC Bioinformatics, vol. 9: 303.

Tusher, V.G., Tibshirani, R., & Chu, G. (2001) "Significance analysis of microarrays applied to the ionizing radiation response" Proceedings of the National Academy of Sciences of the United States of America, vol. 98 (9): 5116-5121.