BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 17 2010, pages 2202–2203
doi:10.1093/bioinformatics/btq347

Systems biology Advance Access publication July 15, 2010

RPPanalyzer: Analysis of reverse-phase protein array data

Heiko A. Mannsperger[1,*], Stephan Gade[1], Frauke Henjes[1], Tim Beissbarth[2] and Ulrike Korf[1]

[1]Department of Molecular Genome Analysis, German Cancer Research Center Heidelberg and [2]Department of Medical Statistics, Georg-August-University Goettingen, Germany

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: RPPanalyzer is a statistical tool developed to read reverse-phase protein array data, to perform the basic data analysis and to visualize the resulting biological information. The R-package provides different functions to compare protein expression levels of different samples and to normalize the data. Implemented plotting functions permit a quality control by monitoring data distribution and signal validity. Finally, the data can be visualized in heatmaps, boxplots, time course plots and correlation plots. RPPanalyzer is a flexible tool and tolerates a huge variety of different experimental designs.

Availability: The RPPAanalyzer is open source and freely available as an R-Package on the CRAN platform http://cran.r-project.org/

Contact: h.mannsperger@dkfz.de

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

Received on February 25, 2010; revised on May 31, 2010; accepted on June 26, 2010

1 INTRODUCTION

In systems biology as well as in biomarker discovery reverse-phase protein arrays (RPPAs) have emerged as a useful tool for the large-scale analysis of protein expression and protein activation (Paweletz *et al.*, 2001). The method follows the basic principle of printing large numbers of crude protein extracts in parallel on a solid phase carrier to form a single array. Multiple slides are printed in parallel and each (sub)array is probed with a different monospecific antibody. To quantify protein expression or protein activation detectable signals are generated via fluorescence, dye precipitation or chemoluminescence. Different methods for calculating and normalizing sample concentrations are available. We implemented this package in the free and open-source R programming environment for statistical computing regarding the fact that R is a widely accepted standard platform for statistical and bioinformatics method exchange (R Development Core Team, 2005).

2 PACKAGE FEATURES

RPPA data analysis requires three sources of information: signal intensities gained from the image analysis software GenePix, sample annotation and slide-specific information. After optional background correction, the relative concentration of each sample is calculated. To ensure comparability of individual samples a signal normalization method matching the experimental design has to be chosen. Finally, data can be visualized using different plotting methods, can be exported as text file, or R-object. The data analysis workflow is described in detail in the package vignette.

2.1 Data input and annotation

The information describing the experiment is saved in a single folder:

(1) The sample description is summarized as a tab-delimited text file containing the sample phenodata and the corresponding position in the source plate.

(2) The slide description file describes array features and contains antibody information.

(3) Raw signal intensities have to be provided as GenePix result files (gpr files).

Data import automatically generates a list with four components. The first data matrix contains foreground signal intensities and a second matrix an estimate of the local background. Each row of the matrices corresponds to a vector of spot intensities and columns link the information to individual arrays. The third and fourth components are matrices with information on samples and arrays. The four lists are kept throughout the analysis and can be exported as text files after each step of the data analysis procedure.

2.2 Data processing

Background correction can be performed as an optional first step using the limma package (Ritchie *et al.*, 2007). The median of replicate spots is used to determine the relative concentration of each detected protein in each sample. In case of serially diluted samples, the relative protein concentrations are calculated using a linear or three parameter logistic model for each sample. The most recent methods use one model to describe the dynamic of all samples of one array. Hu *et al.* (2007) suggest a non-parametric model implemented in the SuperCurve package (Coombes *et al.*, 2009). Another approach is the serial dilution curve algorithm introduced by Zhang *et al.* (2009). Both are accessible in the package.

2.3 Data normalization

Each step during the preparation and spotting of samples for RPPA experiments is a potential error source which has an influence on the total protein amount of the individual spot. Thus, a mandatory step in RPPA data processing is the adjustment of

*To whom correspondence should be addressed.

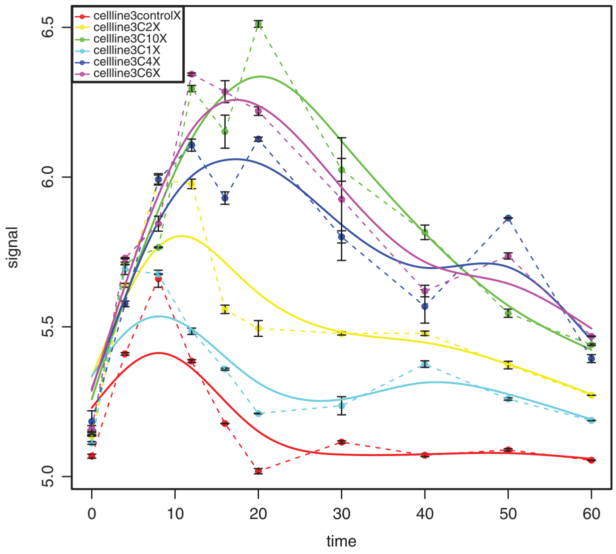

Fig. 1. Rawdata and splines of six STC experiments showing the dynamics of protein phosphorylation in cell lines between 0 and 60 min after adding the stimulus.

the antibody signals according to the total protein content of a spot. Five different normalization approaches are implemented to ensure sample comparability which are described in detail in the Supplementary Material.

2.4 Data visualization

The data robustness can be visualized in quality control plots revealing the distribution of all individual data points. Signal validity can be analyzed by comparing blank signals to target-specific signals in scatter plots. Correlation plots are implemented to analyze the correlation of protein expression and any numerical sample attribute. The significance is tested and corrected for multiple testing. Boxplots can be used to visualize the empirical distribution of the data points in order to assess differential expression of proteins in comparison to a control group. Heatmaps are tools for the visualization of protein expression in large sample sets. It is the most common method to show structures of the data with the use of hierarchical clustering. Stimulation time courses (STC) are cellular assays to analyze the time-dependent response of a cellular system to exogenic stimuli such as growth factors or UV radiation. Figure 1 shows a time course plot from a STC experiment with raw data summarized as smoothing splines. For additional analysis, the RPPA data list can be exported as tab-delimited text file or bioconductor expression set.

3 SUMMARY

We developed an R package addressing the needs for a fast and comprehensive analysis as well as visualization of RPPA data. Raw data are obtained by using GenePix and are supplemented with additional information in a text file format. Previously published software solutions for RPPA data analysis (Stanislaus *et al.*, 2008) are not compatible with R and restricted to very limited array designs. Thus, the functionalities offered by our package can easily be integrated into any type of R-based analysis and accepts any array format.

ACKNOWLEDGEMENTS

Thanks to Marc Johannes for providing the plot.heatmap function, Jitao Zhang for helping with rppaList2expressionset, and Maria Fälth and Christian Bender for support and motivation. The authors are responsible for the contents of this publication.

Funding: The German Federal Ministry of Education and Science in the framework of the program for medical genome research (FKZ: 01GS0890); the project BreastSys in the platform Medical Systems Biology (FKZ: 03153968); network IG Cellular Systems Genomics (FKZ: 01GS0864) in the platform NGFNplus; the Clinical Research Group 179 through the DFG.

Conflict of Interest: none declared.

REFERENCES

Coombes,K.R. *et al.* (2009) *SuperCurve: SuperCurve Package*. R package version 1.3.3. Available at http://bioinformatics.mdanderson.org/Software/OOMPA/#supercurve (last accessed date July 8, 2010).

Hu,J. *et al.* (2007) Non-parametric quantification of protein lysate arrays. *Bioinformatics*, **23**, 1986–1994.

Paweletz,C.P. *et al.* (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Ritchie,M.E. *et al.* (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700–2707.

R Development Core Team (2005) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Stanislaus,R. *et al.* (2008) Rppaml/rims: a metadata format and an information management system for reverse phase protein arrays. *BMC Bioinformatics*, **9**, 555.

Zhang,L. *et al.* (2009) Serial dilution curve: a new method for analysis of reverse phase protein array data. *Bioinformatics*, **25**, 650–654.