BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 2 2008, pages 276–278
doi:10.1093/bioinformatics/btm556

Gene expression

RReportGenerator: automatic reports from routine statistical analysis using R

Wolfgang Raffelsberger[1,*,†], Yannick Krause[1,†], Luc Moulinier[1], David Kieffer[1], Anne-Laure Morand[2], Laurent Brino[2] and Olivier Poch[1]

[1]Laboratoire de Bioinformatique et Génomique Intégratives, IGBMC, UMR 7104, 67404 Illkirch, France and [2]Plate-forme 'Puces à Cellules Transfectées', LBGS, CEBGS-IGBMC, 67404 Illkirch, France

Received on June 28, 2007; revised and accepted on November 2, 2007

Advance Access publication November 24, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Summary: With the establishment of high-throughput (HT) screening methods there is an increasing need for automatic analysis methods. Here we present RReportGenerator, a user-friendly portal for automatic routine analysis using the statistical platform R and Bioconductor. RReportGenerator is designed to analyze data using predefined analysis scenarios via a graphical user interface (GUI). A report in pdf format combining text, figures and tables is automatically generated and results may be exported. To demonstrate suitable analysis tasks we provide direct web access to a collection of analysis scenarios for summarizing data from transfected cell arrays (TCA), segmentation of CGH data, and microarray quality control and normalization.

Availability: RReportGenerator, a user manual and a collection of analysis scenarios are available under a GNU public license on http://www-bio3d-igbmc.u-strasbg.fr/~wraff

Contact: wolfgang.raffelsberger@igbmc.u-strasbg.fr

1 INTRODUCTION

The sequencing of the human genome has opened the way for numerous high-throughput (HT) analysis and high content screening (HCS) techniques. Among the programs and platforms capable of performing statistical analyses, 'R Development Core Team, 2005' (http://www.r-project.org/) has gained much popularity since many active partners are further developing this open-source language and its additional libraries at CRAN (http://cran.r-project.org/) and Bioconductor (Gentleman *et al.*, 2004). R itself provides a command line interface, which is very powerful but rather difficult to approach for the inexperienced user seeking automated solutions for routine analyses.

Several graphical user interfaces (GUIs) for R have been created, e.g. Simple-R (http://www-sre.wu-wien.ac.at/SimpleR), R-pad (http://www.rpad.org/Rpad/) and iPlots (http://rosuda.org/iPlots/). However, these GUIs were not designed specifically for generating reports from routine analysis. In this context we have created RReportGenerator, a GUI giving inexperienced users the possibility to perform automatic routine analysis while benefitting from the advantages of R and its libraries.

2 SOFTWARE OVERVIEW

RReportGenerator allows calling R and executing the code from a user-selected pre-defined 'Analysis Scenario' for automatically generating reports via a simple GUI. Using the 'Library' button, a steadily growing collection of validated scenarios dedicated to biological research can be directly accessed from our website. Furthermore, this web service guarantees to always work with the most recent versions of the scenarios. More information about the selected scenario can be displayed via the 'Infos' button. At report generation the intermediary .tex file is passed to LaTeX or MikTex (Windows version), transforming the report into pdf format. Finally, RReportGenerator deletes all temporary files.

Besides, it is possible to keep the intermediary .tex file and separate (post-script) files for plots when selecting the 'save .tex file' option or to generate a .dvi version of the analysis report. Providing a filename in the field 'Supplemental Data Output File' activates the option to generate an additional file (if part of the scenario-code) designed for exporting data (e.g. to spreadsheet programs like Excel). If the input data is not available as a single file, a supplemental input file may be specified via the GUI. Furthermore, the scenario code can be designed to read all files from the directory of a selected input file, e.g. in Affymetrix microarray quality control (QC) scenarios. The default path for searching input data, scenarios or for saving output can be customized through a configuration window. Internal messages and those created from Sweave are displayed in the 'Session Window', allowing to monitor the progression of the data analysis. RReportGenerator was written in the TCL-TK language and compiled for use under Linux and Windows OS (a Mac version will be released soon).

3 ANALYSIS SCENARIOS

While our collection of analysis scenarios is growing permanently experienced users can write and use their own

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

```
%@RRG_INFO
%This scenario will read a tab-delimited file with
  input data and draw a scatter plot.
%@RRG_INFO_END

\documentclass[a4paper]{article}
\title{minimal test scenario for RReportGenerator}
\SweaveOpts{echo=FALSE}
\usepackage{a4wide,Sweave}

\begin{document}
\maketitle

The file selected as 'Data Input File' is read :
<<chunk_read,echo=FALSE,print=TRUE>>=
mydata <- read.table("<DATA_IN_FILE>")
t.test(mydata[,1],mydata[,2],paired=TRUE)
@

The selected data have \Sexpr{nrow(mydata)} rows
and \Sexpr{ncol(mydata)} columns. \\

Scatter-plot of the first 2 columns of the data :
\begin{center}
<<chunk_plot1,echo=FALSE,fig=TRUE>>=
plot(mydata[,2]~mydata[,1])
@
\end{center}

% If a name for 'Supplemental Data Output File'
% was given, a summary of the data is saved ...

<<chunk_save, echo=F, print=F, results=hide>>=
write.table(summary(mydata), "<DATA_OUT_FILE>",
  row.names=F)
@
\end{document}
```

Fig. 1. Example code of a mimimal scenario for RReportGenerator. Analysis scenarios use the syntax of Sweave, an R package allowing to weave R with LaTeX markup. Besides, we have added specific terms like '%@RRG_INFO' and '@RRG_INFO_END' for extracting information about the scenario or '<DATA_IN_FILE>' allowing to automatically address a file selected through the GUI from within R commands.

scenarios, too. Novel analysis scenarios contributed from other researchers will also be made available as part of the web service collection.

Internally, analysis scenarios make use of the R-package Sweave (Leisch, 2002) allowing to combine LaTeX markup and R language to integrate text, figures and tables in a pdf report. An example for the code of a very simple analysis scenario is shown in Figure 1. As shown, the R-code of scenarios is typically organized in different chunks dedicated to tasks like performing calculations within R (e.g. 'chunk_read'), printing figures (e.g. 'chunk_plot1') or writing files. Only a few special items of Sweave code need to be adopted for use with RReportGenerator: In order to display a brief summary about the scenario when clicking the 'Infos' button from the GUI, the text between the marks '%@RRG_INFO' and '%@RRG_INFO_END' is extracted for display. An input file selected via the GUI can be accessed using the variable '<DATA_IN_FILE>' (containing the complete path and name). Similarly the term '<DATA_OUT_FILE>' is used for automatically inserting the 'supplemental data output file'-name provided through the GUI.

4 USAGE AND APPLICATION

RReportGenerator can be used in a wide range of routine analysis cases in clinical and biological research where the aim and structure of the experiments change rarely, like automatic quality control (e.g. of Affymetrix arrays) or analyzing HCS plates.

This program was designed for (i) biologists (performing HT or HCS experiments) who prefer to avoid command line code and who are primarily interested in automated solutions to obtain well-documented reports. Besides, (ii) bioinformaticians familiar with R can write novel analysis scenarios for automating their analysis tasks. (iii) The reports generated (and the supplementary data output) may be used to standardize the exchange of data at the interface from institutional facilities (e.g. from microarray facility to biostatistics facility performing analysis). Our growing collection of validated analysis scenarios covers applications dedicated to performing multiple segmentation methods on comparative genomic hybridization (CGH) data (Marioni *et al*., 2006), scenarios for conveniently combining multiple QC methods available for Affymetrix gene expression arrays (Bolstad *et al*., 2003; Gautier *et al*., 2004; Wilson and Miller, 2005), and a scenario for normalizing and combining technical replicates from two-colour microarrays analysed using MAIA (Novikov and Barillot, 2007). Furthermore, we have developed scenarios for the analysis of transfected cell array (TCA) data with different input formats (tabulated text or multi-sheet Excel files from GE Healthcare IN Cell Analyzers).

In our experience the use of RReportGenerator has given biologists and operators of HCS projects the means of easily performing immediate quality control and basic data analysis tasks, speeding up the overall performance. In consequence, the quality of our routine documentation has improved, biologists can focus quicker on project-specific interpretation and bioinformaticians have more time available for project-specific tasks or writing novel scenarios.

As we are dedicated to improving automatic analysis procedures, in particular for HT testing and HCS data, the list of applications will be further expanded and our tools further developed. In conclusion, we hope that this program and the increasing collection of available analysis scenarios will represent a valuable tool for laboratories performing routine HT experiments.

ACKNOWLEDGEMENTS

The authors thank Guillaume Berthommier (IGBMC) for his assistance, L. Vallar (CRP-Santé Luxembourg) for sharing data, A. Carles (IGBMC) and B. Tiwari (NERC, Oxford, UK) for critical discussions. This project was supported through the European Retinal Research Training Network 'RETNET' (MRTN-CT-2003-504003), EVI-Genoret (LSHG-CT-2005-512036), INSERM, CNRS, ULP, Region Alsace, Conseil Général du Bas Rhin, CUS and Cancéropôle Grand Est.

Conflict of Interest: none declared.

REFERENCES

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance. *Bioinformatics*, **19**, 185–193.

Gautier,L. *et al*. (2004) affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

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

Leisch,F. (2002) Sweave: dynamic generation of statistical reports using literate data analysis. In Härdle,W. and Rönz,B. (eds.) *Proceedings in Computational Statistics*. Physica Verlag, Heidelberg, pp. 575–580.

Marioni,J.C. *et al*. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Novikov,E. and Barillot,E. (2007) Software package for automatic microarray image analysis (MAIA). *Bioinformatics*, **23**, 639–640.

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.

Wilson,C.L. and Miller,C.J. (2005) Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*, **21**, 3683–3685.