Suitable reporting for the reproducible research: an added value in the analysis of proteomics data

Eugenio Del Prete[1,2], Angelo Facchiano[2], Aldo Profumo[3], Claudia Angelini[4] and Paolo Romano[3]

[1] Dipartimento di Scienze, Università della Basilicata, Viale dell'Ateneo Lucano 10, 85100, Potenza (Italy)
[2] Istituto di Scienze dell'Alimentazione, CNR, Via Roma 64, 83100 Avellino (Italy)
[3] IRCCS AOU San Martino IST, Largo Rosanna Benzi 10, 16132 Genova (Italy)
[4] Istituto per le Applicazioni del Calcolo, CNR, Via Pietro Castellino 111, 80131 Napoli (Italy)

Introduction

The reproducibility and the transparency of a scientific experiment should be an integral part of the research work itself, so that the results can be better assessed and validated, and the proposed methods and procedures can be re-used, thus becoming a kind of protocol for similar experiments. In this context, computational reproducibility [1] refers to the possibility of reconstructing all the steps of a workflow that connects raw data, processed data and results.
Computational reproducibility is a fundamental issue in the omic studies because of the complex and high-dimensional nature of the involved data. The analysis of omics data needs to exploit multi-step workflows including pre-processing, elaboration, statistical validation, interpretation and presentation. Although some analysis platforms are able to ensure computational reproducibility for different omics studies, they do not provide explicit information about the executed code. Clearly, the importance of knowing the actual code depends on the confidence level and/or on the interest of the user for this level of details. However, the availability of the code increases the quality of research in terms of transparency and knowledge transfer. Moreover, it allows other researchers to reproduce the results in a local system (using the same or another programming language), make a comparison among the results and re-use computer code for analyzing different dataset.
MALDI-ToF mass spectra constitute one important type of proteomics data. Spectrometers provide raw mass spectra, while most of elaborations, such as peak detection and spectra alignment, are performed by specialized software [2]. Geena 2 [3] is a web-platform that allows the pre-processing of MALDI-ToF mass spectra by means of a user-friendly interface that guides users through the entire analysis. Its prospected extension GeenaR [4] is aimed at providing additional functionalities based on the R environment.
Here, we describe how GeenaR is going to incorporate some tools supporting reproducible research by combining statistical programming, good computational practice and user-friendly web-interface.

Methods

Geena 2 is a robust web tool for MALDI-ToF mass spectra pre-processing. Its main output is the list of common peaks identified by aligning average spectra originated from groups of replicates from different samples. Intermediate results are also made available. GeenaR is an extension of Geena 2 still under development. Its objective is the integration in the platform of some R libraries, which may provide advanced statistical analyses, thus enriching the current output.
It is noteworthy that many R packages follow the reproducible research philosophy. Since R users and developers often overlap, the curse of reproducibility is well-known and taken into account. The importance of reproducible research and the ways how many R packages can cope with this issue are clearly presented in [5]. For the aims of GeenaR, the following R packages and tools have been considered: *R-Markdown*, *knitr* and *spin*. *R-Markdown* is able to integrate R and the Markdown language; it allows to create documents containing R code, which is then evaluated as an embedded part of the Markdown processing [6]. *Knitr* is an R package for the treatment of particular kinds of documentation, recognizing R-Markdown, LaTex and HTML as documentation language and with the possibility of converting documents from these formats into PDF. Two of the main *knitr* features are its ability to flexibly manage portions of code, called 'chunks', and to cache data and results for a faster elaboration of concatenated subroutines. The *spin* function accelerates the conversion of R code directly into HTML and R-Markdown [7]. The implementation of these resources on an existing

web platform can be an added value for its reporting features, since it improves the creation of a report about the work carried out, especially with reference to the code.

Results and Discussion

One of the aims of both Geena 2 and GeenaR is facilitating the users in analyzing MALDI-ToF mass spectra by providing a web-interface that allows to upload data, select different algorithms and parameters, execute the analysis in order to obtain results according to a specific demand. Both expert and non-expert users can take advantage of such platforms.
In particular, non-expert-users can take advantage from a simplified interface with an automatic choice for most of the parameters. On the contrary, expert users can make access to a more detailed interface that allows a fine-tuning of most of the parameters.
In both cases, thanks to the novel reproducible research module implemented in GeenaR, the system generates a report containing all the steps performed. The report can be customized according to the detail of interest or to the expertise of the user. Therefore, minimal information about the used functions and the parameters to ensure computational reproducibility will be provided.
More in details, the report will provide: date and time of the execution, the R libraries used for the process, chunks of code for main elaborations, selected parameters (either by the users or by the system), uploaded data in MALDIquant 'Mass Spectrum' class type, numerical and graphical results, short explanation about the workflow, version of the system and of the packages.
GeenaR generates the results in a compressed archive, with separated log and graphical results, and a report, both in R-Markdown and in HTML format. An example is provided as supplementary material in a ZIP compressed archive [8]. The archive includes three files: a) an R script, which generates the other two files, b) the generated Markdown code, which is a plain text file with a name extension ".md", and c) the output in HTML, with parameters, data and some results, which can be open by any browser.
It is important to underline strongly that reproducible research is not an optional, but a fundamental component of a good computational practice, which becomes essential in computational biology. The chance to reproduce exactly an experiment, from the beginning to the end, improves the robustness of results and it leaves a trail about how a particular result can be produced, with a view to simplify the knowledge transfer, even among researchers with different backgrounds.

References

1. Peng RD. Reproducible research in computational science. Science, 334:6060, 1226-1227, 2011.
2. Del Prete E, d'Esposito D, Mazzeo MF, et al. Comparative analysis of MALDI-ToF mass spectrometry data in proteomics: a case study. Computational Intelligence Methods for Bioinformatics and Biostatistics, Lecture Notes in Bioinformatics, 9874, 154-164, 2016.
3. Romano P, Profumo A, Rocco M, et al. Geena 2, improved automated analysis of MALDI/TOF mass spectra. BMC Bionformatics, 17(Suppl 4):61, 2016.
4. Del Prete E, Facchiano A, Profumo A, et al. GeenaR: a flexible approach to pre-process, analyse and compare MALDI-ToF mass spectra. Conference Proceedings, V Congresso Gruppo Nazionale di Bioingegneria, 2016.
5. Russo F, Righelli D, Angelini C. Advantages and limits in the adoption of reproducible research and R-tools for the analysis of omic data. Computational Intelligence Methods for Bioinformatics and Biostatistics, Lecture Notes in Bioinformatics, 9874, 245-258, 2016.
6. Allaire JJ, Cheng J, Xie Y, et al. rmarkdown: dynamic documents for R. R package version 0.6.1, 2015. *http://CRAN.R-project.org/package=rmarkdown*
7. Xie Y. knitr: a general-purpose package for dynamic report generation in R. R package version 1.10.5, 2015.
8. Supplementary materials for this document: *http://bioinformatics.hsanmartino.it/geenar/docs/nettab2016_suppl_materials.zip*