Using R and Bioconductor for proteomics data analysis

L. Gatto*,[1], L.M. Breckels[1], S. Gibb[2], A. Christoforou[1] and K. S. Lilley[1]

[1]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, UK
[2]Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Germany
*lg390@cam.ac.uk – http://www.bio.cam.ac.uk/proteomics/

Introduction

R and Bioconductor have had a tremendous impact on the quality of genomics data analysis [1]. The members of the project have demonstrated that if life scientists wanted to extract relevant information from expensive data, it was absolutely necessary to value data analysis by investing the required time and energy. Nowadays, the Bioconductor user base comprises a variety of users, including non-bioinformaticians that are have overcome the initial hurdle of the command-line interfaces and non-trivial data analysis to explore and comprehend high-throughput data.

Even well-known and respected leader in proteomics agree that it lies 10 years behind genomics. There are several valid reasons for this, including the chemical complexity of proteins, the technical complexity of the instrumentation (in particular mass-spectrometers - MS) and the vast possiblilties in the study of proteins. An often overseen albeit essential component of this failure is the quality of the scientific software that is used and valued inside the community. Computational proteomics researcher, who value quality software, comprehensive data analysis and reproducible research ought to demonstrate that better results can be achieved with better tools to invite the proteomics community to embrace quality, open-source and fexible tools. Here, we illustrate some examples of proteomics data analysis in R .

Working with raw data

The proteomics community has developed a range of data standards and formats for MS data (the latest being mzML) to overcome the shortcomings or closed, binary vendor-specific formats. One of the main projects that implement parsers for the XML-based open formats is the C++ proteowizard project [2], which is interfaced by the mzR Bioconductor package using the Rcpp package.

```
library("mzR")
library("RforProteomics")
fname <- getPXD000001mzXML()
ms <- openMSfile(fname)
```

The resulting ms object is a file handle that allows fast direct and random access to the individual spectra. mzR is used by a variety of other packages like xcms, MSnbase, RMassBank and TargetSearch.

Challenges Further improve support of raw MS data and develop the range of supported formats, in particluar identification (mzIdentML) and quantitation (mzQuantML) formats.

MS^2 labeled quantitation

Label-free quantitation

MS^e data independent acquisition

Peptide identification

This work has been supported by the PRIME-XS project, grant agreement number 262067, funded by the European Union 7^{th} Framework Program.

UNIVERSITY OF CAMBRIDGE

Figure 1: A figure.

Conclusions and perspectives

The felxibility of the R environment and the breath of available packages is sometimes daunting for newcomers and introductionary points of entry are welcome. The RforProteomics package [3] [https://github.com/lgatto/RforProteomics] ought to assume such roles. For this, RforProteomics should be a collaborative project and contribution through the github repository are encouraged.

Despite well known advantages in terms of statistical analyses of data and some unique software for proteomics and mass-spectrometry data analysis, there remains a lot of efforts and work to be done for R/Bioc to become a complete framework for proteomics data processing. These efforts should be tackled by a group of developers. It is our hope that the RforProteomics will be a helpful targetted introduction to new users and motivate collaborative development of package developers.

References

Gentleman *et al. Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol. 2004 PMID: 15461798.

Chambers *et al. A cross-platform toolkit for mass spectrometry and proteomics.* Nat Biotechnol. 2012 PMID: 23051804.

Gatto L, Christoforou A. *Using R and Bioconductor for proteomics data analysis.* Biochim Biophys Acta. 2013 PMID: 23692960.