A current perspective on using R and Bioconductor for proteomics data analysis

S. Gibb[1,3], LM. Breckels[1,2], T. Lin Pedersen[4], VA. Petyuk[5], KS. Lilley[2] and L. Gatto[1,2,*]

[1]Computational Proteomics Unit and [2]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, UK
[3]Department of Anesthesiology and Intensive Care, Medical Faculty Carl Gustav Carus, Technical University Dresden, Germany
[4]Chr. Hansen A/S, Hørsholm / Technical University of Denmark, Kgs. Lyngby, Denmark
[5]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

*lg390@cam.ac.uk http://cpu.sysbiol.cam.ac.uk

Introduction

The R statistical environment and programming language is a key player in many domains that require robust data analysis. The Bioconductor project offer a wide range of R packages dedicated to the analysis and comprehension of high throughput biology. Originally focused on genomics, R/Bioconductor are gaining increasing attention in the proteomics, metabolomics and mass spectrometry communities, as reflected by the download statistics and package contributions.

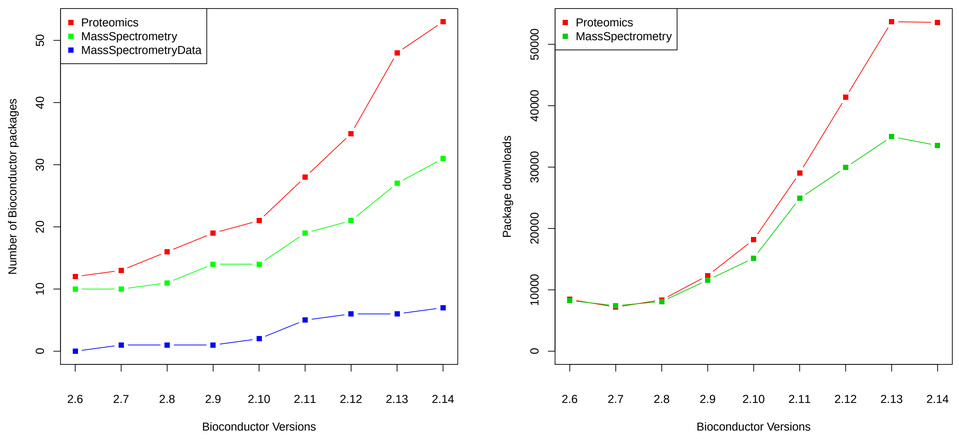

Figure 1 : The left figure shows the number of Bioconductor packages dedicated to *Proteomics*, *MassSpectrometry* and *MassSpectrometryData* packages on Bioconductor in the BiocViews . On the right figure, we show the number of distinct package downloads. Note that the current version, 2.14 currently encounters downloads. (NB: the data for Bioc 2.12 are interpolated due to massive scripted downloads).

Here, we present an overview of current Bioconductor infrastructure dedicated to proteomics and mass spectrometry. These software packages **tightly interoperate**, providing **well defined pipeline workflows** and a **flexible and in-depth development environment**.

Working with raw data

The proteomics community has developed a range of data standards and formats for MS data (e.g. mzML, mzIdentML) to overcome the shortcomings of closed, binary vendor-specific formats.

One of the main projects that implement parsers for the XML-based open formats is the C++ proteowizard project, which is interfaced by the mzR Bioconductor package using the Rcpp infrastructure for fast raw and (starting with Bioconductor version 3.0) identification data. mzIdentML files can also be parsed with the mzID package.

```
library("mzR")
ms <- openMSfile("raw_data.raw")
id <- openIDfile("msgf-res.mzid")
library("mzID")
id2 <- mzID("msgf-res2.mzid")
```

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

Identification

R/Bioconductor provide software to parse mzIdentML files (mzID and mzR, see above), directly run identification R (rTANDEM, MSGFplus) and optimise FDR calculations based on decoy searches. Below, we use MSnID to define identification searches filters using precursor mass error, identification scores and missed cleavage to minimise false discovery rates.

```
(id <- apply_filter(id, filter))
: MSnID object
: Working directory: "."
: #Spectrum Files:  24
: #PSMs: 514638 at 0.042 % FDR
: #peptides: 56546 at 0.11 % FDR
: #accessions: 30944 at 0.56 % FDR
```

Quantitation

Several quantitation pipelines are supported: MSnbase and isobar for isobaric tagging such as iTRAQ and TMT, MALDIquant for MALDI data, MSnbase for spectra counting and other MS^2 label free methods and synapter supports a DIA complete pipeline, including ion mobility separation on Waters Synapt instruments.

MS data processing

MS spectrum processing is available in multiple packages, optimised for specific use cases and pipelines: MSnbase, xcms, MALDIquant.

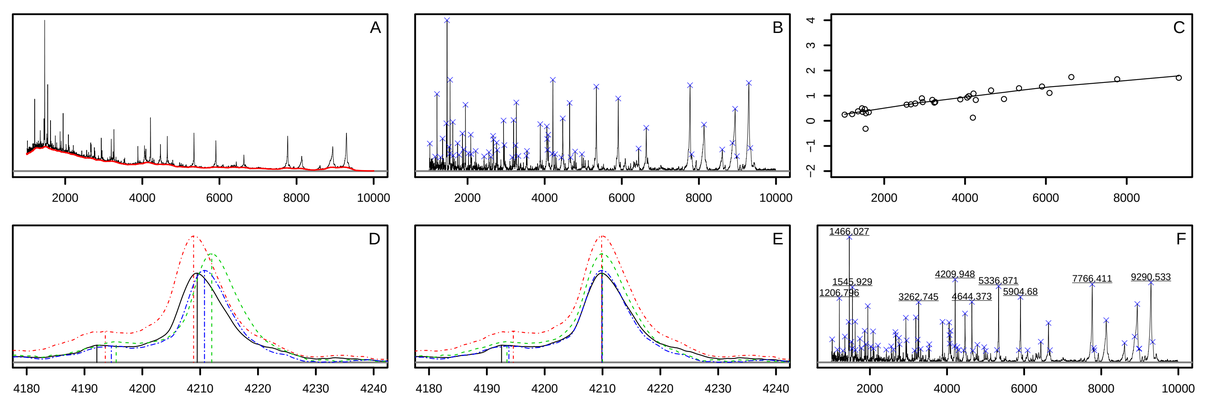

Figure 2 : Illustration of the MALDIquant preprocessing pipeline. The first figure (A) shows a raw spectrum with the estimated baseline. In the second figure (B) the spectrum is variance-stabilised, smoothed, baseline-corrected and the detected peaks are marked with blue crosses. The third figure (C) is an example of a fitted warping function for peak alignment. In the next figures, four peaks are shown before (D) and after (E) performing the alignment. The last figure (F) represents a merged spectrum with discovered and labelled peaks.

Visualisation

Using R for high quality programmable figures and interactive graphical user interfaces such as https://lgatto.shinyapps.io/shinyMA/.

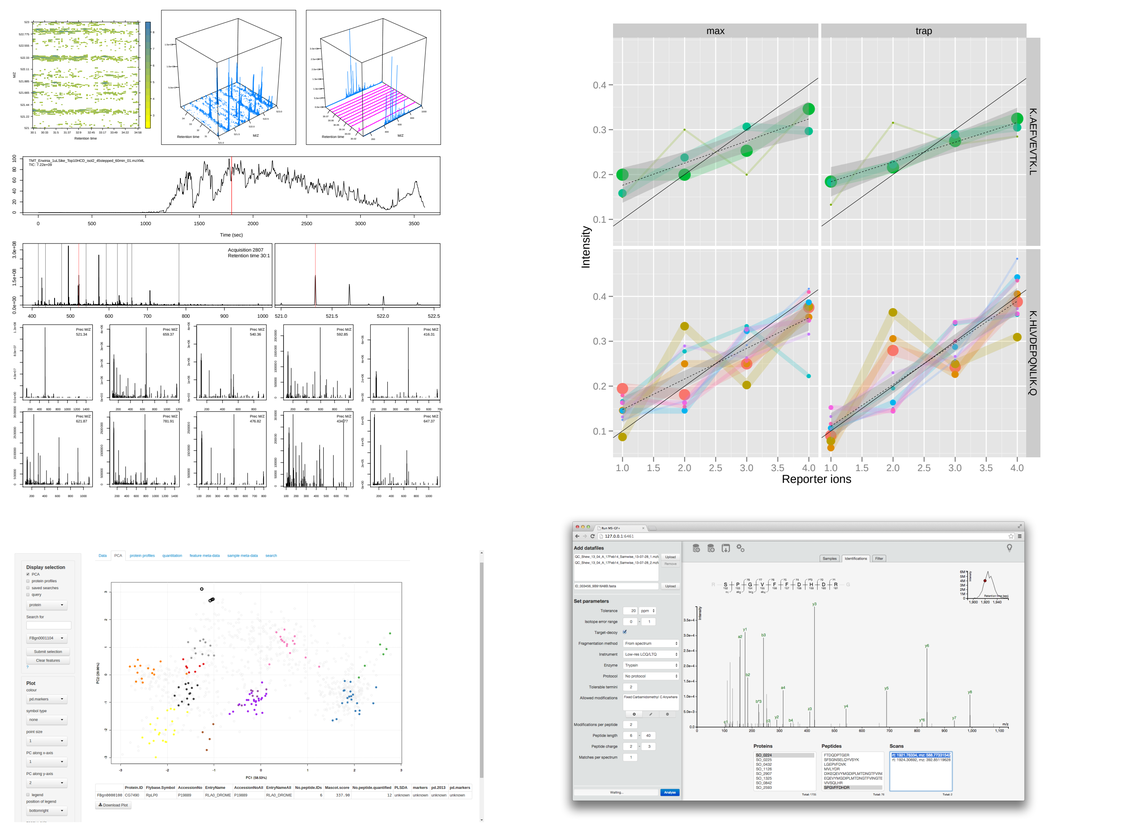

Figure 3 : Visualising raw data and process quantitative data (top) and interactive graphical interfaces (bottom) from the pRolocGUI and MSGFgui packages.

Data processing, statistics and machine learning

As an environment for statistical computing, the very best of data processing, statistical modelling and machine learning is readily available in packages such as MSnbase, isobar, MSstats, msmsTests.

MSstats offers set of tools for statistical relative protein significance analysis in DDA, SRM and DIA experiments as well as power calculation. msmsTests leverages powerful statistical modelling from the edgeR package to analyse spectral counting data.

Conclusion

A wide range of other applications such as post-translational modifications (isobar), quality control (qcmetrics and proteoQC) or spatial proteomics (pRoloc) are available as Bioconductor packages. A complete list is available on the Bioconductor page and in the RforProteomics package. Contributions and discussions are welcome on our Google group and GitHub wiki.

The R project http://www.r-project.org/
Bioconductor http://bioconductor.org/
RforProteomics http://is.gd/R4Proteomics
Google group https://is.gd/rbioc_sig_proteomics
Support site https://support.bioconductor.org

This work was supported by the **European Union 7^{th} Framework Program PRIME-XS project** and a **BBSRC Tools and Resources Development Fund**.