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

L. Gatto*,[1,2], S. Gibb[1,3]

[1]Computational Proteomics Unit and [2]Cambridge Centre for Proteomics, Department of Biochemistry, Tennis Court Road, University of Cambridge, CB2 1QR
[3]Department of Anesthesiology and Intensive Care, Medical Faculty Carl Gustav Carus, Technical University Dresden, Fetscherstr. 74, 01307 Dresden
*lg390@cam.ac.uk – http://www.bio.cam.ac.uk/proteomics/

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.

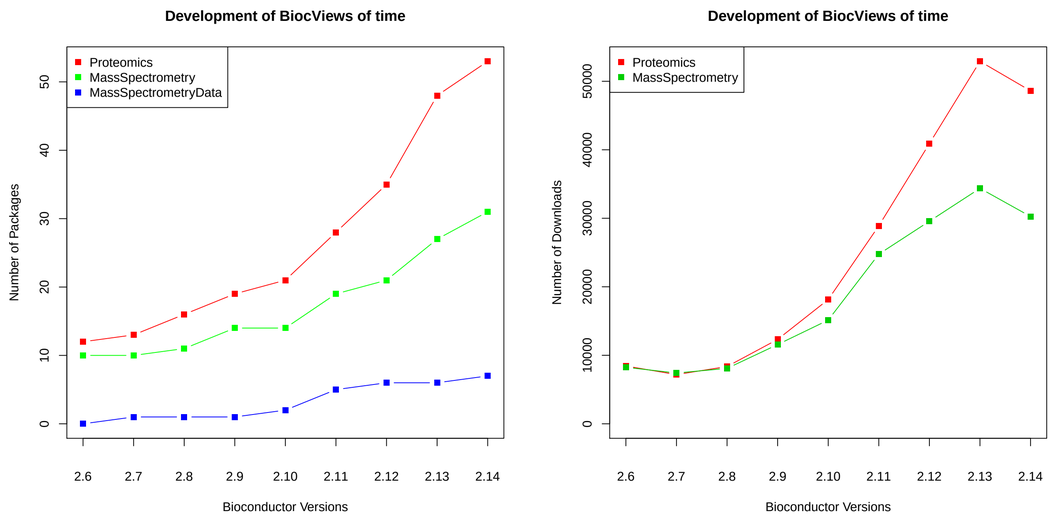

Figure : The left figure shows the number of Bioconductor packages dedicated to *Proteomics*, *MassSpectrometry* and *MassSpectrometryData* packages on Bioconductor in the BiocViews . On the right figure, we show the number of distinct package downloads. Note that the current version, 2.14 currently 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.

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 [2], which is interfaced by the mzR Bioconductor package using the Rcpp infrastructure.

```
library("mzR")
fname <- dir(system.file(package = "MSnbase", dir = "extdata"),
    full.name = TRUE, pattern = "mzXML$")
ms <- openMSfile(fname)
```

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

Labelled quantitation

The same raw data file can be imported in a convenient higher level container and directly processed, plotted, quantified and normalised with the MSnbase [5] software.

```
exp <- readMSData(fname, verbose = FALSE)
plot(exp[["X3.1"]], full = TRUE, reporters = iTRAQ4)
set <- quantify(exp, method = "trap", reporters = iTRAQ4,
    verbose = FALSE, parallel = TRUE)
head(exprs(set), n = 3)
```

Label-free quantitation

Support for data dependent label-free quantitation is available, among others, in the MSnbase [5], xcms [7] and MALDIquant [6] packages.

The latter provides a complete pipeline, including baseline subtraction, smoothing, peak detection and alignment using warping functions, handling of replicated measurements as well as allowing spectra with different resolutions.
A complete pipeline for MS^E data independent acquisition, including support for ion mobility separation is available in the synapter package [1] that, among other features, transfers identification between acquisitions to substantially reduce missing values.

Peptide identification

The recently released rTANDEM package encapsulates the X!Tandem [**?**] search engine in R.

It uses the same XML-based parameter files as the native application or dedicated R parameter object. Result files can be directly parsed and mined in R .

```
xmlres <- rtandem(spectra.mgf, taxon = "yeast",
                  taxonomy = "taxonomy.xml",
                  default.parameters = "default-params.xml")
## or xmlres <- tandem(param)
res <- GetResultsFromXML(xmlres)
proteins <- GetProteins(res) ## data.table objects
peptides <- GetPeptides(res)
```

The latest mzR and the mzID packages offer support for mzIdentML files and allow to import identification data from most commonly used search engines.

```
library("mzR")
ident <- openIDfile("identification.mzid")
# or
library("mzID")
ident <- mzID("identification.mzid")
```

Statistics

While there exists a lot of excellent packages for differential expression tests in R MSstats [3] and msmsTests offer tests that are well integrated into the MSnbase ecosystem.

```
## MSstats/msTests example?
```

[1] Bond NJ *et al. Improving Qualitative and Quantitative Performance for MS^E-based Label-free Proteomics.* J Proteome Res. 2013 PMID: 23510225.

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

[3] Choi M *et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.* Bioinformatics. 2014 PMID: 24794931.

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

[5] Gatto L and Lilley KS *MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation.* Bioinformatics. 2012 PMID: 22113085.

[6] Gibb S and Strimmer K *MALDIquant: a versatile R package for the analysis of mass spectrometry data.* Bioinformatics. 2012 PMID: 22796955.

[7] Smith CA *et al. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification.* Analytical Chemistry 2006 PMID: 16448051.

R http://www.r-project.org/
Bioconductor http://bioconductor.org/
RforProteomics http://is.gd/R4Proteomics