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.

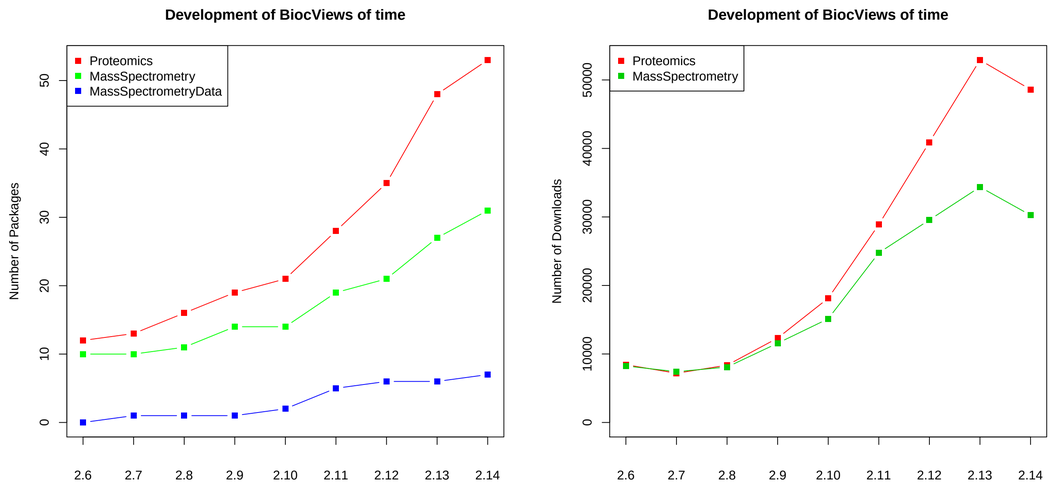

Figure : The left figure shows the number of Bioconductor packages dedicated to *Proteomics*, *MassSpectrometry* and *MassSpectrometryData* packages on Bioconductor in the BiocViews . On the right figure, we show the number of distinct package downloads. Note that the current version, 2.14 currently 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 [**?**], which is interfaced by the mzR Bioconductor package using the Rcpp infrastructure for 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 <- openIDfile("msgf-res2.mzid")
```

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.

Identification

Running search engines in R and parsing identification files.

mzID, (mzR), MSnID, rTANDEM, (MSGFplus)

Quantitation

Quantitative proteomcs in R .

MSnbase, MALDIquant, synapter

Visualisation

R excels in graphics. It also allows to easily program interactive visualisation interfaces.

A figure from the vis paper, a shiny GUI example.

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.

MSnbase, isobar, MSstats, msmsTests,

Misc

Any other applications: PTMs, spatial proteomics, ...

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

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