MSnbase, efficient R-based access and manipulation of raw mass spectrometry data

Laurent Gatto,[*,†] Sebastian Gibb,[‡] and Johannes Rainer[¶]

†*Computational Biology Unit, de Duve Institute, Université catholique de Louvain, Brussels, Belgium*

‡*Department of Anaesthesiology and Intensive Care of the University Medicine Greifswald, Germany*

¶*Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy*

E-mail: laurent.gatto@uclouvain.be

Abstract

We present version 2 of the `MSnbase` R/Bioconductor package. `MSnbase` provides infrastructure for the manipulation, processing and visualisation of mass spectrometry data. Here we present how the new *on-disk* infrastructure allows the handling of hundreds on commodity hardware and present some application of the package.

Introduction

Mass spectrometry is a powerful technology to assays chemical and biological samples. It is used routinely, with well characterised protocol, as well a development platform, to improve on existing methods and devise new ones to analyse ever more complex sample in greater details. The complexity and diversity of mass spectrometry yields data that is itself complex

and often times of considerable size, that requires non trivial processing before producing interpretable results. This is particularly relevant, and can constitue a significant challenge for method developers that, in addition to the development of sample processing and mass spectrometry methods, need to process and analyse these new data to demonstrate the improvement in their technical and analytical work.

There exists a very diverse catalogue of software tools to explore, process and interpret mass spectrometry data. These range from low level software libraries such as vendor libraries, jmzML (ref), proteowizard (ref), ... that are aimed at programmers to develop new applications, to user-oriented applications, such as ProteomeDiscoverer, MaxQuant, ... that provide a limited and fixed set of functionality. The former are used through application programming interfaces exclusively, while the latter generally featuring graphical user interfaces (GUI).

TODO: Give examples of libraries re-used in user/gui focused application...

In this software note, we present version 2 of the `MSnbase`[1] R/Bioconductor software package. `MSnbase` offers a platform that lies between low level libraries and end-use software. It provides a flexible command line environment for metabolomics and proteomics mass spectrometry-based application, that allows a detailed step-by-step processing, analysis and exploration of the data and development of novel computational mass spectrometry methods.

Software functionality

On-disk backend

The main feature in version 2 of the `MSnbase` package was the addition of different backends for raw data storage, namely *in-memory* and on-disk. The following code chunk demonstrates how to create two `MSnExp` objects, tailored to manage mass spectrometry experiments, storing data in-memory or on-disk.

```
library("MSnbase")

raw_mem <- readMSData("file.mzML", mode = "inMemory")

raw_dsk <- readMSData("file.mzML", mode = "onDisk")
```

The former is the legacy storage mode, implemented in the first version of the package, that loads all the raw data and the metadata in memory. This solution doesn't scale for modern large dataset, and was complemented by the on-disk backend, that only load metadata in memory and accesses the spectra in the original files when needed. There are two direct benefits using the on-disk backend. Figure 1 shows 4-fold faster reading times (left) and 10-fold smaller memory footprint (right).

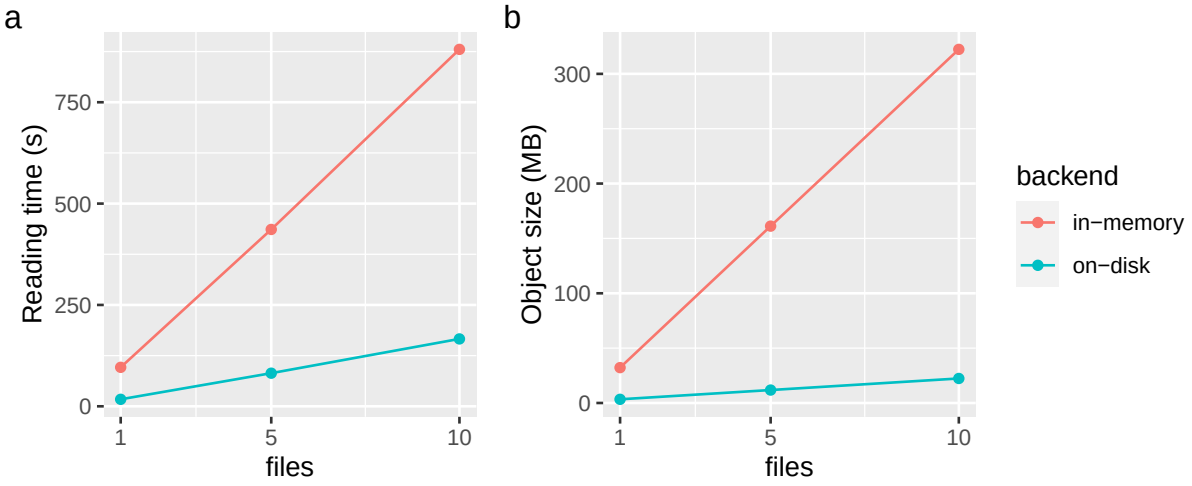

Figure 1: (a) Reading time (in seconds) and (b) data size in memory (in MB) to read/store 1, 5 and 10 files containing 1431 MS1 (on-disk only) and 6103 MS2 (on-disk and in-memory) spectra.

The on-disk backend also offers efficient data manipulation by way of *lazy processing*. Operations on the raw data are stored in a processing queue and only effectively applied when raw data is accessed on disk. As an example, the following short analysis pipeline, that can equally be applied to on in-memory or on-disk data retains MS2 spectra acquired between 1000 and 3000 seconds, extract the M/Z range corresponding to the TMT 6-plex range and focuses on the MS2 spectra with a precursor intensity greater than 11×10^6 (the

median precursor intensity).

```
x <- x_dsk %>%
    filterRt(c(1000, 3000)) %>%
    filterMz(120, 135)
x[precursorIntensity(x) > 11e6, ]
```

As shown on figure 2, this lazy mechanism si significantly faster than its application on in-memory data.

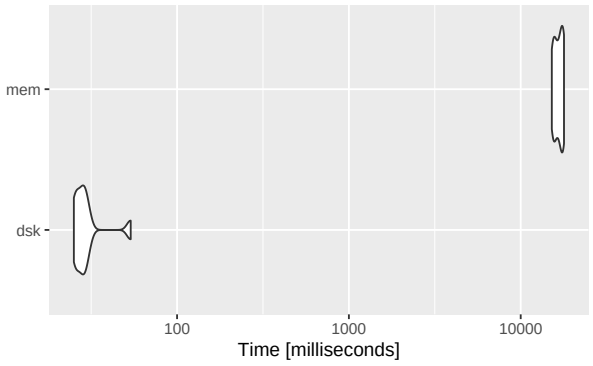

Figure 2: Filtering benchmark assessed over 10 interations on in-memory (top) and on-disk (bottom) data containing 6103 MS2 spectra.

The advantageous reading and execution times and memory footprint of the on-disk backend are possible by avoiding unnecessary access to the raw data. Once access to the spectra M/Z and intensity values become mandatory (for example for plotting), then the in-memory backend becomes more efficient, as illustrated on figure 3. This gain is maximal when the whole dataset is the be accessed (i.e. all spectra are already in memory) and negligeable when large fractions of the data need to be subset.

This new on-disk infrastructure enables large scale data analyses using `MSnbase` (metabolomics example, see Johannes).

Prototyping

See https://github.com/lgatto/msnbase_boxcar.

Figure 3: Access time to spectra for the in-memory (left) and on-disk (right) backends for 1, 10, 100 1000, 5000 and all 6103 spectra.

Visualisation

Examples: 3D MSmap, boxcar, centroiding vignette

Discussion

To address (from guidelines):

- potential for reuse: see[2–4] for examples.
- general limitations
- system limitations
- end-user documentation
- developer documentation
- sample data
- benchmark data set

- availability
- license information
- system requirements

Collaborative development, 11 contributors since creation (see blog post).

Count packages depending on `MSnbase`.

Future developments.

The version of `MSnbase` used in this manuscritp is version 2.10.0. The main features presented here were available since version 2.0.

Acknowledgement

The authors thank the various contributors and users who have provided constructuve input and feedback that have helped, over the years, the improvement of the package. The authors declare no conflict of interest.

References

(1) Gatto, L.; Lilley, K. S. MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **2012**, *28*, 288–9.

(2) Wieczorek, S.; Combes, F.; Lazar, C.; Giai Gianetto, Q.; Gatto, L.; Dorffer, A.; Hesse, A. M.; Couté, Y.; Ferro, M.; Bruley, C.; Burger, T. DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* **2017**, *33*, 135–136.

(3) Griss, J.; Vinterhalter, G.; Schwämmle, V. IsoProt: A Complete and Reproducible Workflow To Analyze iTRAQ/TMT Experiments. *J Proteome Res* **2019**, *18*, 1751–1759.

(4) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* **2006**, *78*, 779–87.