A Parser for mzXML, mzData and mzML files

Bernd Fischer*
Steffen Neumann†
Laurent Gatto‡

October 2, 2013

Contents

1 Introduction **1**

2 Mass spectrometry raw data **2**
2.1 Spectral data access 2
2.2 Metadata access 2

3 Example **2**

4 Future plans **5**

5 Session information **6**

1 Introduction

The mzR (Chambers et al., 2012) package aims at providing a common interface to several mass spectrometry data formats, namely `mzData` (Orchard et al., 2007), `mzXML` (Pedrioli et al., 2004) and the latest `mzML` (Martens et al., 2010), somewhat similar to the Bioconductor package affyio for affymetrix raw data. No processing is done in mzR, which is left to packages such as XCMS[1] or MSnbase[2].

Most importantly, access to the data should be fast and memory efficient. This is made possible by allowing random file access, i.e. retrieving specific data of interest without having to sequentially browser the full content.

The actual work of reading and parsing the data files is handled by the included C/C++ libraries or "backends". The RAMP parser, written at the Institute for Systems Biology (ISB) is a fast and lightweight parser in pure C. Later, it gained support for the `mzData` format.

*bernd.fischer@embl.de
†sneumann@ipb-halle.de
‡lg390@cam.ac.uk

[1] `http://www.bioconductor.org/packages/release/bioc/html/xcms.html`
[2] `http://www.bioconductor.org/packages/release/bioc/html/MSnbase.html`

The C++ reference implementation for the `mzML` is the proteowizard library (Kessner et al., 2008) (pwiz in short), which in turn makes use of the boost C++ (`http://www.boost.org/`) library. RAMP is able to access `mzML` files by calling pwiz methods.

The mzR package is in essence a collection of wrappers to the C++ code, and benefits from the C++ interface provided through the Rcpp package (Eddelbuettel and François, 2011).

2 Mass spectrometry raw data

All the mass spectrometry file formats are organized similarly, where a set of metadata nodes about the run is followed by a list of spectra with the actual masses and intensities. In addition, each of these spectra has its own set of metadata, such as the retention time and acquisition parameters.

2.1 Spectral data access

Access to the spectral data is done via the peaks function. The return value is a list of two-column mass-to-charge and intensity matrices or a single matrix if one spectrum is queried.

2.2 Metadata access

Run metadata is available via several functions such as `instrumentInfo()` or `runInfo()`. The individual fields can be accessed via e.g. detector() etc.

Spectrum metadata is available via `header()`, which will return a list (for single scans) or a dataframe with information such as the `basePeakMZ`, `peaksCount`, ... or, for higher-order MS the `msLevel` and precursor information.

The availability of this metadata can not always be guaranteed, and depends on the MS software which converted the data.

3 Example

A short example sequence to read data from a mass spectrometer. First open the file.

```
> library(mzR)
> library(msdata)
> mzxml <- system.file("threonine/threonine_i2_e35_pH_tree.mzXML",
+                      package = "msdata")
> aa <- openMSfile(mzxml)
```

We can obtain different kind of header information.

```
> runInfo(aa)
```

```
$scanCount
[1] 55

$lowMz
[1] 50.0036

$highMz
[1] 298.673

$dStartTime
[1] 0.3485

$dEndTime
[1] 390.027

$msLevels
[1] 1 2 3 4
```

> instrumentInfo(aa)

```
$manufacturer
[1] "Thermo Scientific"

$model
[1] "LTQ Orbitrap"

$ionisation
[1] "ESI"

$analyzer
[1] "FTMS"

$detector
[1] "unknown"
```

> header(aa,1)

```
$seqNum
[1] 1

$acquisitionNum
[1] 1

$msLevel
[1] 1

$peaksCount
[1] 684

$totIonCurrent
[1] 341427000
```

```
$retentionTime
[1] 0.3485

$basePeakMZ
[1] 120.066

$basePeakIntensity
[1] 211860000

$collisionEnergy
[1] 0

$ionisationEnergy
[1] 0

$lowMZ
[1] 50.3254

$highMZ
[1] 298.673

$precursorScanNum
[1] 0

$precursorMZ
[1] 0

$precursorCharge
[1] 0

$precursorIntensity
[1] 0

$mergedScan
[1] 0

$mergedResultScanNum
[1] 0

$mergedResultStartScanNum
[1] 0

$mergedResultEndScanNum
[1] 0
```

Read a single spectrum from the file.

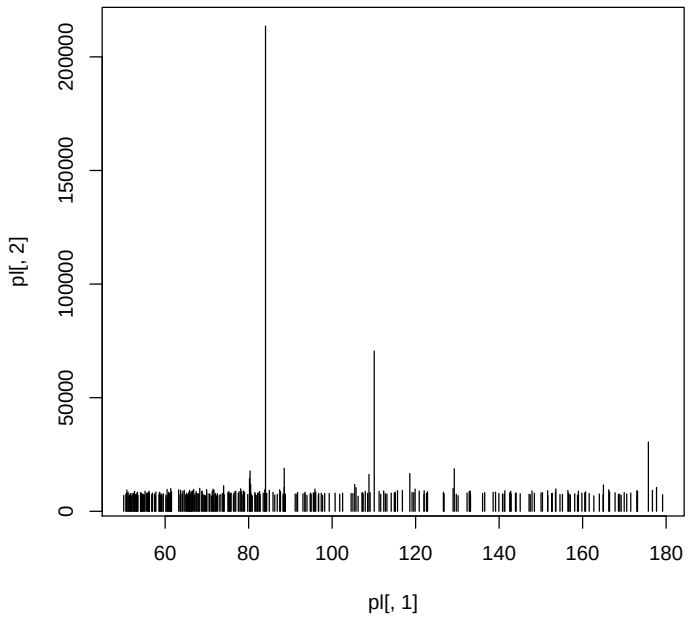
```
> pl <- peaks(aa,10)
> peaksCount(aa,10)
```

```
[1] 317

> head(pl)

         [,1]     [,2]
[1,] 50.08176 6984.858
[2,] 50.62267 7719.419
[3,] 50.70530 7185.290
[4,] 50.73298 7509.140
[5,] 50.83848 9366.624
[6,] 50.88303 8012.808

> plot(pl[,1], pl[,2], type="h", lwd=1)
```



One should always close the file when not needed any more. This will release the memory of cached content.

```
> close(aa)
```

4 Future plans

Right on the heels of the initial RAMP backend release, supporting little metadata beyond e.g. the instrument model, is support to the full pwiz functionality to access the full metadata stored in an `mzML` files, or the chromatograms (which store e.g. MRM data). Other file formats supported by pwiz, such as `mzIdentML` for protein identification results are also possible in the future.

5 Session information

- R version 3.0.1 (2013-05-16), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Rcpp 0.10.5, msdata 0.1.13, mzR 1.6.3
- Loaded via a namespace (and not attached): Biobase 2.20.1, BiocGenerics 0.6.0, codetools 0.2-8, parallel 3.0.1, tools 3.0.1

References

Matthew C. Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L. Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina A. Baker, Mi-Youn Brusniak, Christopher Paulse, David Creasy, Lisa Flashner, Kian Kani, Chris Moulding, Sean L. Seymour, Lydia M. Nuwaysir, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Suckau Detlev, Tina Hemenway, Andreas Huhmer, James Langridge, Brian Connolly, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W. Deutsch, Robert L. Moritz, Jonathan E. Katz, David B. Agus, Michael MacCoss, David L. Tabb, and Parag Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotech*, 30(10):918–920, October 2012. doi: 10.1038/nbt.2377. URL `http://dx.doi.org/10.1038/nbt.2377`.

Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. URL `http://www.jstatsoft.org/v40/i08/`.

Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–6, 2008. doi: 10.1093/bioinformatics/btn323.

Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Rompp, Steffen Neumann, Angel D Pizarro, Luisa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Puneet Souda, Henning Hermjakob, Pierre-Alain Binz, and Eric W Deutsch. mzml - a community standard for mass spectrometry data. *Molecular and Cellular Proteomics : MCP*, 2010. doi: 10.1074/mcp.R110.000133.

Sandra Orchard, Luisa Montechi-Palazzi, Eric W Deutsch, Pierre-Alain Binz, Andrew R Jones, Norman Paton, Angel Pizarro, David M Creasy, JÃl'rÃt'me Wojcik, and Henning Hermjakob. Five years of progress in the standardization of proteomics data 4th annual spring workshop of the hupo-proteomics

standards initiative april 23-25, 2007 ecole nationale supérieure (ens), lyon, france. *Proteomics*, 7(19):3436–40, 2007. doi: 10.1002/pmic.200700658.

Patrick G A Pedrioli, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, Brian Raught, Brian Pratt, Erik Nilsson, Ruth H Angeletti, Rolf Apweiler, Kei Cheung, Catherine E Costello, Henning Hermjakob, Sequin Huang, Randall K Julian, Eugene Kapp, Mark E McComb, Stephen G Oliver, Gilbert Omenn, Norman W Paton, Richard Simpson, Richard Smith, Chris F Taylor, Weimin Zhu, and Ruedi Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, 22(11):1459–66, 2004. doi: 10.1038/nbt1031.