A parser for raw and identification mass-spectrometry data

Bernd Fischer*
Steffen Neumann†
Laurent Gatto‡
Qiang Kou§

September 29, 2015

Contents

1 Introduction

The *mzR* package aims at providing a common interface to several mass spectrometry data formats, namely `mzData` [1], `mzXML` [2], `mzML` [3] for raw data, and `mzIdentML` [4], 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* [5, 6] or *MSnbase* [7].

Most importantly, access to the data should be fast and memory efficient. This is made possible by allowing on-disk random file access, i.e. retrieving specific data of interest without having to sequentially browser the full content nor loading the entire data into memory.

The actual work of reading and parsing the data files is handled by the included C/C++ libraries or "backends". The `mzRramp` 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. The C++ reference implementation for the `mzML` is the proteowizard library [8] (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. More recently, the proteowizard[1] [9] has been fully integrated using the

*bernd.fischer@embl.de
†sneumann@ipb-halle.de
‡lg390@cam.ac.uk
§qkou@umail.iu.edu
[1] http://proteowizard.sourceforge.net/

mzRpwiz backend for raw data. The mzRnetCDF backend provides support to CDF-based formats. Finally, the mzRident backend is available to access identification data (mzIdentML) through pwiz.

warning: It is anticipated to switch to the mzRpwiz backend in Bioconductor 3.1. We advise users and developers to test it and report any issues on the github issue tracker https://github.com/sneumann/mzR/issues.

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 [10].

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 Identification result access

The main access to identification result is done via psms, score and modifications. psms and score will return the detailed information on each psm and scores. modifications will return the details on each modification found in peptide.

2.3 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.

Identification metadata is available via mzidInfo(), which will return a list with information such as the software, ModificationSearched, enzymes, SpectraSource and other information for this identification result.

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

3 Example

3.1 mzXML/mzML/mzData files

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

```
library(mzR)
```

```
## Loading required package:  Rcpp
```

```
library(msdata)

mzxml <- system.file("threonine/threonine_i2_e35_pH_tree.mzXML",
                     package = "msdata")
aa <- openMSfile(mzxml) ## ramp, default backend
```

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
##
## $polarity
## [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.

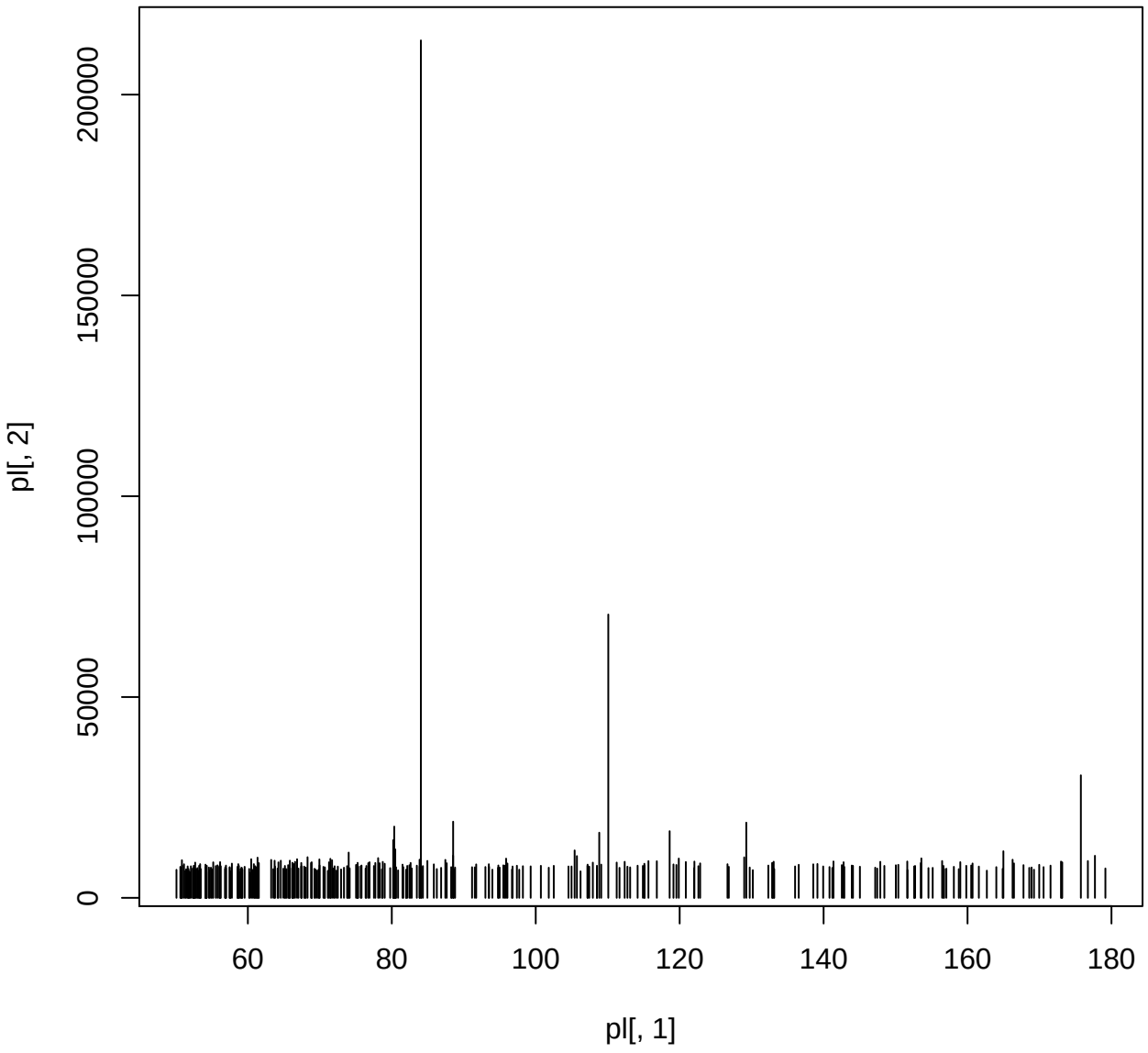
```
pl <- peaks(aa,10)
peaksCount(aa,10)
```

```
## [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 if you are using RAMP backend. This will release the memory of cached content.

```
close(aa)
```

3.2 mzIdentML files

You can use openIDfile to read a mzIdentML file (version 1.1), which use the pwiz backend.

```
library(mzR)
library(msdata)
```

```
file <- system.file("mzid", "Tandem.mzid.gz", package="msdata")
x <- openIDfile(file)
```

`mzidInfo` function will return general information about this identification result.

```
mzidInfo(x)
```

```
## $FileProvider
## [1] "researcher"
##
## $CreationDate
## [1] "2012-07-25T14:03:16"
##
## $software
## [1] "xtandem x! tandem CYCLONE (2010.06.01.5) "
## [2] "ProteoWizard MzIdentML 3.0.6239 ProteoWizard"
##
## $ModificationSearched
## [1] "Oxidation"       "Carbamidomethyl"
##
## $FragmentTolerance
## [1] "0.8 dalton"
##
## $ParentTolerance
## [1] "1.5 dalton"
##
## $enzymes
## $enzymes$name
## [1] "Trypsin"
##
## $enzymes$nTermGain
## [1] "H"
##
## $enzymes$cTermGain
## [1] "OH"
##
## $enzymes$minDistance
## [1] "0"
##
## $enzymes$missedCleavages
## [1] "1"
##
##
## $SpectraSource
## [1] "D:/TestSpace/NeoTestMarch2011/55merge.mgf"
```

`psms` will return the detailed information on each peptide-spectrum-match, include `spectrumID`, `chargeState`, `sequence`. `modNum` and others.

```
p <- psms(x)
colnames(p)
```

```
##  [1] "spectrumID"               "chargeState"              "rank"
##  [4] "passThreshold"            "experimentalMassToCharge" "calculatedMassToCharge"
##  [7] "sequence"                 "modNum"                   "isDecoy"
## [10] "post"                     "pre"                      "start"
```

```
## [13] "end"                      "DatabaseAccess"           "DBseqLength"
## [16] "DatabaseSeq"              "DatabaseDescription"      "acquisitionNum"
```

The modifications information can be accessed using `modifications`, which will return the `spectrumID`, `sequence`, `name`, `mass` and `location`.

```
m <- modifications(x)
head(m)
```

```
##   spectrumID                sequence            name    mass location
## 1   index=12  LCYIALDFDEEMKAAEDSSDIEK Carbamidomethyl 57.0215        2
## 2   index=12  LCYIALDFDEEMKAAEDSSDIEK       Oxidation 15.9949       12
## 3  index=285   KDLYGNVVLSGGTTMYEGIGER       Oxidation 15.9949       15
## 4   index=83   KDLYGNVVLSGGTTMYEGIGER       Oxidation 15.9949       15
## 5   index=21 VIDENFGLVEGLMTTVHAATGTQK       Oxidation 15.9949       13
## 6  index=198          GVGGAIVLVLYDEMK       Oxidation 15.9949       14
```

Since different software will use different scoring function, we provide a `score` to extract the scores for each psm. It will return a data.frame with different columns depending on software generating this file.

```
scr <- score(x)
colnames(scr)
```

```
## [1] "spectrumID"          "X.Tandem.expect"     "X.Tandem.hyperscore"
```

4 Future plans

Other file formats provided by HUPO, such as `mzQuantML` for quantitative data [11] are also possible in the future.

5 Session information

- R version 3.2.2 (2015-08-14), `x86_64-pc-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=en_US.UTF-8`, `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 0.12.1, msdata 0.6.0, mzR 2.2.2
- Loaded via a namespace (and not attached): Biobase 2.28.0, BiocGenerics 0.14.0, BiocStyle 1.6.0, ProtGenerics 1.0.0, codetools 0.2-14, evaluate 0.8, formatR 1.2.1, highr 0.5.1, knitr 1.11, magrittr 1.5, parallel 3.2.2, stringi 0.5-5, stringr 1.0.0, tools 3.2.2

References

[1] Sandra Orchard, Luisa Montechi-Palazzi, Eric W Deutsch, Pierre-Alain Binz, Andrew R Jones, Norman Paton, Angel Pizarro, David M Creasy, Jrme 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.

[2] 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.

[3] 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.

[4] A R Jones, M Eisenacher, G Mayer, O Kohlbacher, J Siepen, S J Hubbard, J N Selley, B C Searle, J Shofstahl, S L Seymour, R Julian, P A Binz, E W Deutsch, H Hermjakob, F Reisinger, J Griss, J A Vizcano, M Chambers, A Pizarro, and D Creasy. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*, 11(7):M111.014381, Jul 2012. doi:10.1074/mcp.M111.014381.

[5] C A Smith, E J Want, G O'Maille, R Abagyan, and G Siuzdak. Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, 78(3):779–87, Feb 2006. doi:10.1021/ac051437y.

[6] R Tautenhahn, C Bttcher, and S Neumann. Highly sensitive feature detection for high resolution lc/ms. *BMC Bioinformatics*, 9:504, 2008. doi:10.1186/1471-2105-9-504.

[7] L Gatto and K S Lilley. MSnbase – an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–9, Jan 2012. doi:10.1093/bioinformatics/btr645.

[8] 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.

[9] 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. URL: http://dx.doi.org/10.1038/nbt.2377, doi:10.1038/nbt.2377.

[10] 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/.

[11] M Walzer, D Qi, G Mayer, J Uszkoreit, M Eisenacher, T Sachsenberg, F F Gonzalez-Galarza, J Fan, C Bessant, E W Deutsch, F Reisinger, J A Vizcano, J A Medina-Aunon, J P Albar, O Kohlbacher, and A R Jones. The mzquantml data standard for mass spectrometry-based quantitative studies in proteomics. *Mol Cell Proteomics*, 12(8):2332–40, Aug 2013. doi:10.1074/mcp.O113.028506.