A Parser for mzXML, mzData and mzML files

Bernd Fischer, Laurent Gatto and Steffen Neumann

June 1, 2011

Contents

1 Introduction

The mzR package aims at providing a common interface to several mass spectrometry data formats: `mzData` (Orchard et al., 2007), `mzXML` (Pedrioli et al., 2004) and the latest `mzML` (Martens et al., 2010).

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.

- Proteowizard and Ramp, fast random access
- Rcpp

2 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] 2.584685e+161

$highMZ
[1] 9.932698e+247

$startMZ
[1] 6.528011e+35

$endMZ
[1] 9.798654e+58

$dStartTime
[1] 0.3485

$dEndTime
[1] 390.027
```

```
> 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.

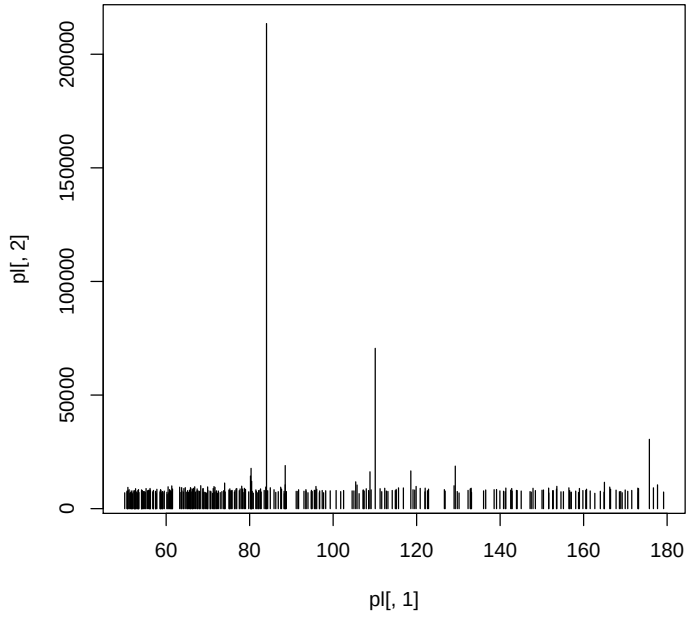
```
> pl <- peaks(aa,10)
> peaksCount(aa,10)

[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)
```



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

```
> close(aa)
```

3 Future plans

- pwiz's metadata
- mzIdentML

4 Session information

- R version 2.14.0 Under development (unstable) (2011-05-30 r56024), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_GB.utf8`, `LC_NUMERIC=C`, `LC_TIME=en_GB.utf8`, `LC_COLLATE=en_GB.utf8`, `LC_MONETARY=en_GB.utf8`, `LC_MESSAGES=en_GB.utf8`, `LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_GB.utf8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: msdata 0.1.5, mzR 0.4.3, Rcpp 0.9.4.2
- Loaded via a namespace (and not attached): Biobase 2.13.2, codetools 0.2-8, tools 2.14.0

References

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 supÃl'rieure (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.