Enabling analysis of large scale metabolomics/proteomics data: on-the-fly data access in MSnbase

Johannes Rainer

December, 2017

Introduction

EuroBioc2017, December 2017, Cambridge.

- Updates to mzR, MSnbase and xcms packages:
  - MS data write support.
  - *on-the-fly* data access.
  - towards a common MS infrastructure: xcms re-uses classes from MSnbase.

onDisk *vs* inMem data; why?

- Keeping all data in memory prevents analysis of large experiments.

- <u>Example</u>: load a MS file

```
library(MSnbase)
library(pryr)

## Read a single MS level 1 file
msd <- readMSData("data/150616_POOL_IntraP_S_POS_6.mzML", mode = "onDisk")

## Object size?
object_size(msd)
```

```
602 kB
```

- Memory footprint: 977 MB (inMem) *vs* 600 kB (onDisk).

- Reading data on demand: *relatively* fast for indexed mzML files.

- Profile MS data?
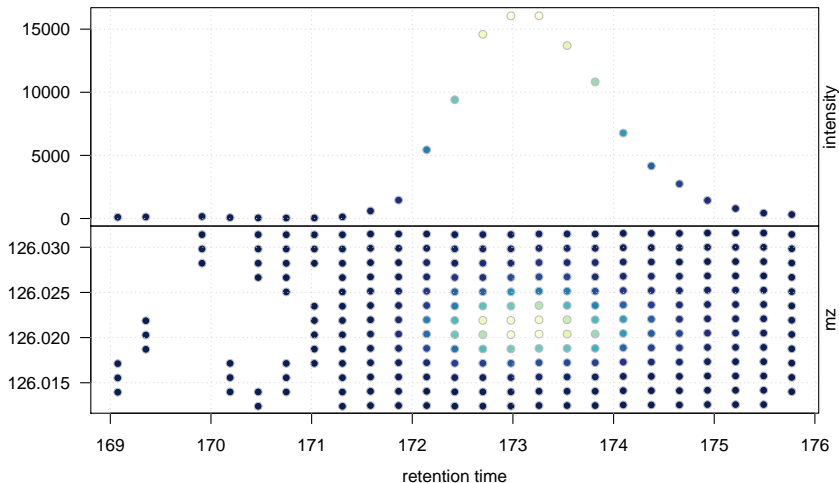
- <u>Example</u>: plot MS data for a metabolite:

```r
library(xcms)

## m/z for Taurine [M+H]+
taur_mz <- 126.02194
taur_rt <- c(169, 176)

## Extract the MS data for this ion
taur_msd <- extractMsData(msd, mz = c(taur_mz - 0.01, taur_mz + 0.01),
                          rt = taur_rt)
plotMsData(taur_msd[[1]])
```

Application example: *centroiding* of *profile* MS data

- Profile-mode MS data for Taurine [M+H]+ ion.

- *centroiding*: represent mass peaks by their *centroid*.

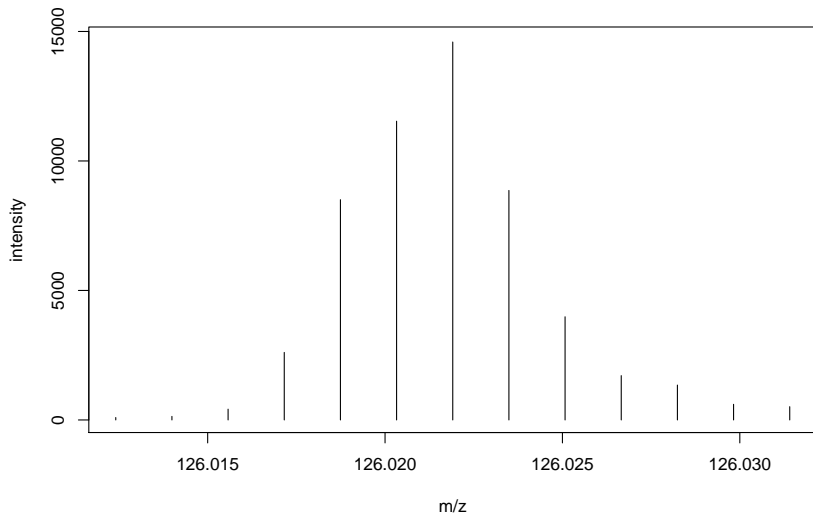- <u>Example</u>: plot mass peak for taurin from one spectrum.

```r
library(magrittr)

## Extract one spectrum for the metabolite
spct <- msd %>% filterAcquisitionNum(n = 619L) %>%
    filterMz(mz = c(taur_mz - 0.01, taur_mz + 0.01))

par(mfrow = c(1, 1), mar = c(4.5, 4, 1, 1))
plot(mz(spct[[1]]), intensity(spct[[1]]), type = "h", ylab = "intensity",
    xlab = "m/z")
```

- Taurin ion mass peak in one spectrum.

Application example: *centroiding* of *profile* MS data

- onDisk data: lazy execution of data manipulations. <u>Example:</u> Data smoothing (smooth) and centroiding (pickPeaks).

```
## 1) smooth the spectrum data
## 2) perform centroiding
cntr <- msd %>% smooth(method = "MovingAverage", halfWindowSize = 2L) %>%
    pickPeaks()

## ... not executed yet ...
cntr@spectraProcessingQueue
```

```
[[1]]
Object of class "ProcessingStep"
 Function: smooth
 Arguments:
  o method = MovingAverage
  o halfWindowSize = 2

[[2]]
Object of class "ProcessingStep"
 Function: pickPeaks
 Arguments:
  o method = MAD
  o halfWindowSize = 3
  o SNR = 0
  o ignoreCentroided = TRUE
  o refineMz = none
```

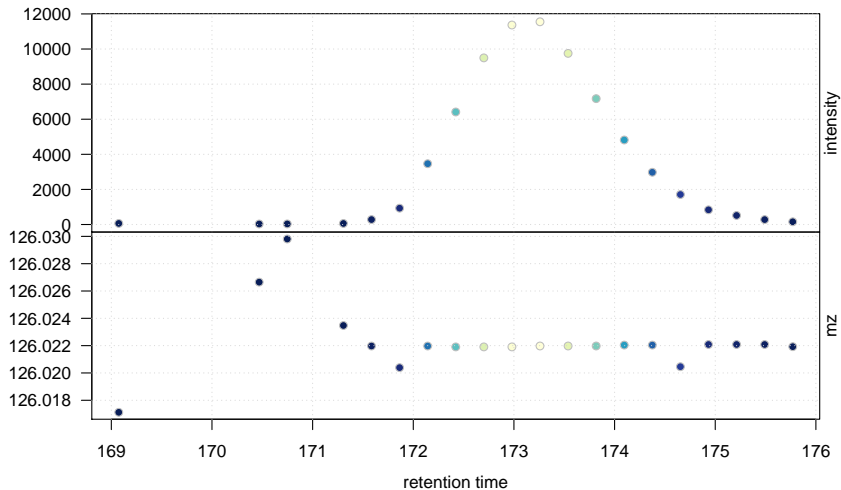- Data manipulations are applied *on-the-fly*.

```
## Extract the MS data for our example ion;
## smoothing and centroiding performed on-the-fly
taur_cntr <- extractMsData(cntr, mz = c(taur_mz - 0.01, taur_mz + 0.01),
                           rt = taur_rt)
plotMsData(taur_cntr[[1]])
```

Application example: *centroiding* of *profile* MS data

- MS data for taurin, centroided.

- *Make persistent*: write to MS data file. <u>Example</u>: write the centroided data to disk.

```
writeMSData(cntr, file = "centroided.mzML", copy = TRUE)

## 1) All processings (smoothing and peak picking) are applied
## 2) Data is exported as mzML
```

Finally

Thank you for your attention!

Collaborative work:

- Laurent Gatto (CPU Cambridge, UK)
- Sebastian Gibb (University Medicine Greifswald, Germany)
- Johannes Rainer (Eurac Research, Italy)
- Steffen Neumann (IPB Halle, Germany)

clone me!
https://github.com/jotsetung/EuroBioc2017-MSnbase.git