

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

Johannes Rainer

December, 2017

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.
- Example: load a MS file

```
library(MSnbase)
library(pryr)

## Read a single MS level 1 file
msd <- readMSData("data/150616_P00L_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.

Application example: *centroiding of profile MS data*

- Profile MS data?
- Example: plot MS data for a metabolite:

```
library(xcms)
library(magrittr)

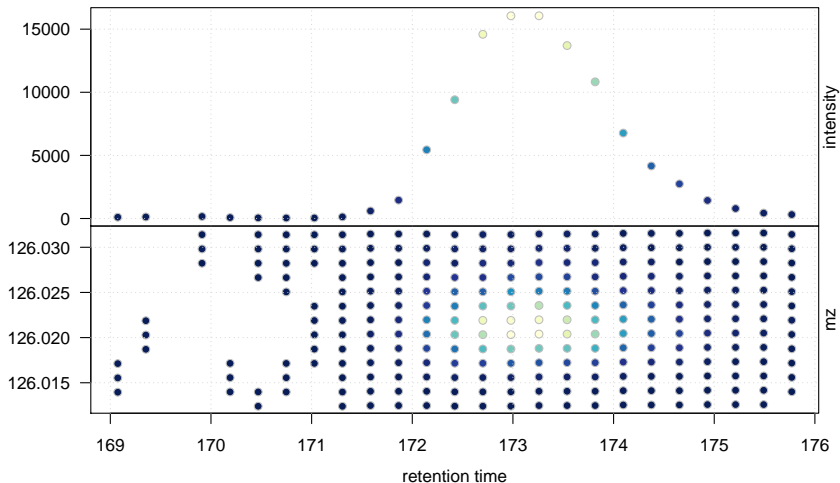
## m/z for Taurine [M+H]+
taur_mz <- 126.02194
taur_rt <- c(169, 176)

## Extract the MS data for this ion
taur_msd <- msd %>%
  filterRt(rt = taur_rt) %>%
  filterMz(mz = c(taur_mz - 0.01, taur_mz + 0.01)) %>%
  extractMsData

plotMsData(taur_msd[[1]])
```

Application example: *centroiding of profile MS data*

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



Application example: *centroiding* of profile MS data

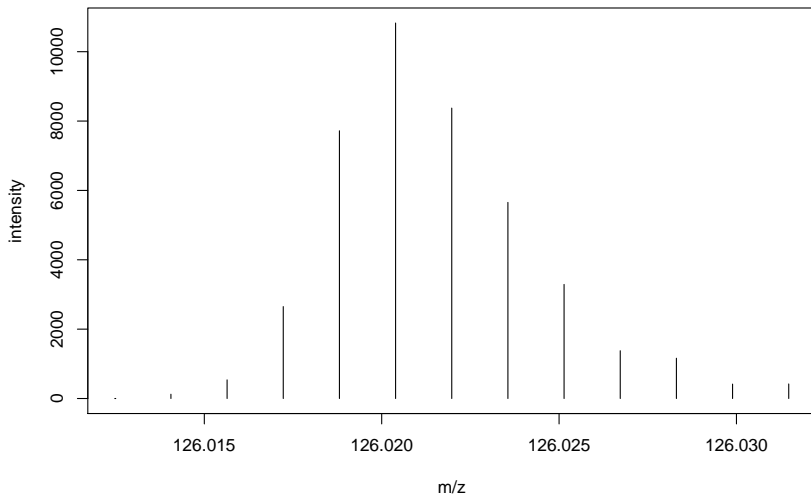
- *centroiding*: represent mass peaks by their *centroid*.
- Example: plot mass peak for taurin from one spectrum.

```
## Identify values corresponding to one spectrum
idx <- grep("F1.S0623", rownames(taur_msd[[1]]))

## Plot the m/z and intensity
par(mfrow = c(1, 1), mar = c(4.5, 4, 1, 1))
plot(taur_msd[[1]]$mz[idx], taur_msd[[1]]$i[idx],
     type = "h", xlab = "m/z", ylab = "intensity")
```

Application example: *centroiding of profile MS data*

- Taurin ion mass peak in one spectrum.



Application example: *centroiding of profile MS data*

- onDisk data: lazy execution of data manipulations. Example: 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
```

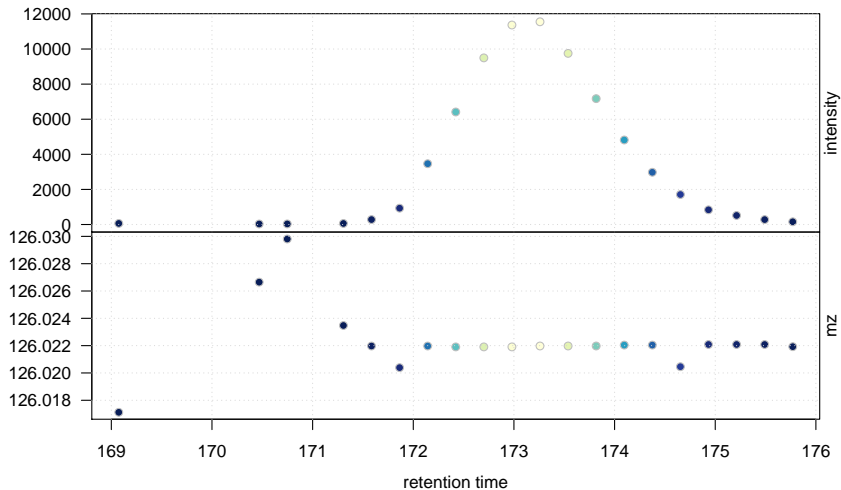

Application example: *centroiding of profile MS data*

- 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.



Application example: *centroiding of profile MS data*

- *Make persistent*: write to MS data file. Example: write the centroided data to disk.

```
writeMSData(cnr, 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>