cosmiq - COmbining Single Masses Into Quantities

David Fischer[‡], Christian Panse[◇], and Endre Laczko[◇]

1 Introduction

cosmiq is a tool for the preprocessing of liquid- or gas-chromatography mass spectrometry (LCMS/GCMS) data with a focus on metabolomics or lipidomics applications. The Bioconductor package [1] has been developed and has shown to be effective using liquid ultra performance capillary chromatography coupled with high accuracy mass data (*full width at half maximum* > 20000), e.g., by using TOF or Q-TOF type mass spectrometer. The data we have used consists of one hundreds files having a size of approx. 500MBytes each (see also [2]).

Because those high-resolution data are too huge for beeing included in the package we will demonstrate the usage of the *cosmiq* package using the smaller *faahKO* data set which is already available on Bioconductor.

The following code of the `cosmiq` wrapper function shows a typical usage:

```
R> library(cosmiq)
R> cdfpath <- file.path(find.package("faahKO"),
+          "cdf")
R> my.input.files <- dir(c(paste(cdfpath,
+          "WT", sep='/'),
+          paste(cdfpath, "KO", sep='/')),
+          full.names=TRUE)
R> # run cosmiq wrapper function
R> #
R> x <- cosmiq(files = my.input.files,
+          mzbin=0.25,
+          SNR.Th=0,
+          linear=TRUE)
R> # graph result
R> image(t(x$eicmatrix),
+          main='mz versus RT map')
R> head(x$xs@peaks)
```

The `cosmiq` function is composed of the following steps:

- Combining spectra
- Detecting mz peaks on master spectrum
- Quantifying masses
- RT correction
- Computing the EIC matrix
- Detecting chromatographic peaks from EIC matrix
- Quantifying mz/RT features

`cosmiq` uses the *xcms* [3] object structure for handling the data. The following pages of this vignette are indented to demonstrate how all the steps can be run manually using the *faahKO* data set.

2 LCMS feature detection step by step using cosmiq

2.1 The Input

The faah knockout dataset [4] will be used as input.

```
R> library(cosmiq)
R> cdfpath <- file.path(find.package("faahKO"),
+          "cdf")
R> my.input.files <- dir(c(paste(cdfpath,
+          "WT", sep='/'),
+          paste(cdfpath, "KO", sep='/')),
+          full.names=TRUE)
R> #
R> # create xcmsSet object
R> # todo
R> xs <- new("xcmsSet")
R> xs@filepaths <- my.input.files
```

Define the `phenoData`. This is usually done by the unexported method `xcms:::phenoDataFromPaths`.

```
R> class <- as.data.frame(c(rep("KO",6),
+          rep("WT", 6)))
R> rownames(class) <- basename(my.input.files)
R> xs@phenoData <- class
```

The *xcms* object `xs` will be used as container to keep all the data.

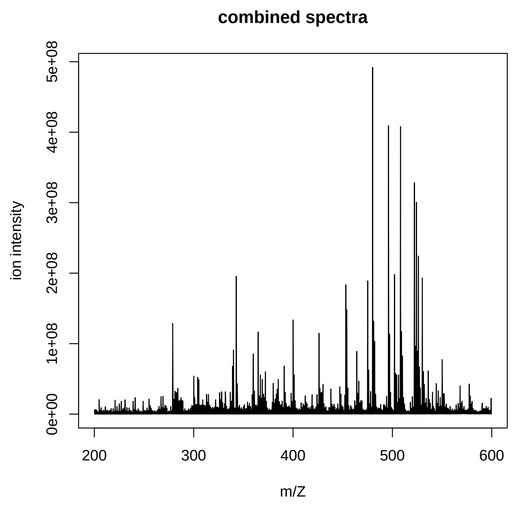
```
R> xs.attr <- attributes(xs)
R> xs.attr$phenoData
```

```
         c(rep("KO", 6), rep("WT", 6))
ko15.CDF                            KO
ko16.CDF                            KO
ko18.CDF                            KO
ko19.CDF                            KO
ko21.CDF                            KO
ko22.CDF                            KO
wt15.CDF                            WT
wt16.CDF                            WT
wt18.CDF                            WT
wt19.CDF                            WT
wt21.CDF                            WT
wt22.CDF                            WT
```

2.2 Combination of mass spectra

The first two processing steps search for relevant mass bins in the dataset. In order to select for optimal bins, we first calculate a combined spectrum. This approach of overlaying and summing intensities of single scans together is usual for applications in flow injection mass spectrometry and aims to improve ion statistics. Not only are mass spectra from all scans from a single LCMS run combined but from all acquired datasets. As a result, signal to noise ratio increases for each additional LCMS run and a master list of observed mass is generated.

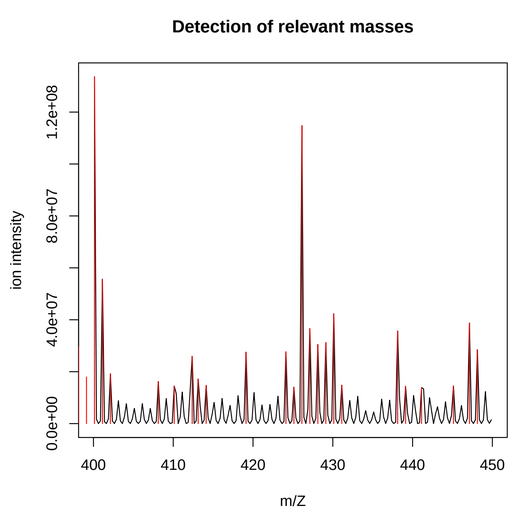
```
R> x <- combine_spectra(xs=xs, mzbin=0.25,
+          linear=TRUE, continuum=FALSE)
R> plot(x$mz, x$intensity, type='l',
+          main='combined spectra',
+          xlab='m/Z', ylab='ion intensity')
R>
```



2.3 Detection of relevant masses

Based on this combined master mass spectrum we then determine location and boundaries of each observed mass. A modified peak detection algorithm based on continuous wavelet transformation (CWT) is used for this step [5]. Peak detection based on CWT has the advantage that a sliding scale of wavelets instead of a single filter function with fixed wavelength is used. This allows for a flexible and automatic approximation of the peak width. As a result, it is possible to locate both narrow and broad peaks within a given dynamic range. The CWT algorithm was modified in order to consider overlapping peaks [2].

```
R> xy <- peakdetection(x=x$mz, y=x$intensity,
+       scales=1:10,
+       SNR.Th=1.0,
+       SNR.area=20, mintr=0.5)
R> id.peakcenter<-xy[,4]
R> filter.mz <- 400 < x$mz & x$mz < 450
R> plot(x$mz[filter.mz],
+       x$intensity[filter.mz],
+       main='Detection of relevant masses',
+       type='l',
+       xlab='m/Z',
+       ylab='ion intensity')
R> points(x$mz[id.peakcenter],
+       x$intensity[id.peakcenter],
+       col='red', type='h')
R>
```



2.4 Generation and combination of extracted ion chromatograms

Until now only the mz information was considered. In the following processing steps, the chromatographic information will be added. For the comparison of different LCMS datasets, it is important to consider RT shifts. These shifts are typically caused by technical variations and need to be corrected before chromatographic peaks between different LCMS runs are aligned. For this purpose, cosmiq implements `xcms` retention time alignment using the obiwarp algorithm. For each detected mass in step 2.3 we calculate an extracted ion chromatogram (EIC). In order to determine the elution time for each detected mass, the EICs of every mass are combined between all acquired runs. Again, this combination approach aims for an improvement of the signal-to-noise ratio (SNR).

```
R> # create dummy object
R> xs@peaks <- matrix(c(rep(1,
+          length(my.input.files) * 6),
+          1:length(my.input.files)),
+          ncol=7)
R> colnames(xs@peaks) <- c("mz",
+          "mzmin", "mzmax", "rt",
+          "rtmin", "rtmax", "sample")
R> xs <- xcms::retcor(xs,
+          method = "obiwarp", profStep=1,
+          distFunc="cor", center=1)
R>
```

2.5 Detection of chromatographic peaks

Based on the combined EICs there is another peak detection step to be performed. The algorithm as described for the peak picking of m/z signals in Step 2.3 is used also for peak picking in the retention time domain. The final result is a peak table with location and boundaries of each mz/RT feature. This information will be further used to locate the relevant position in every single LCMS dataset in order to quantify sample specific feature intensities. Because the mz/RT features were detected on the combined mass spectra or EICs of all samples it is not necessary to align features between different LCMS runs as for a typical raw data processing workflow. Instead, a data matrix with intensity values for every mz/RT feature and every sample can be immediately calculated. An example can bee seen in Figure 1.

```
[1] 136 143
```

```
[1] 2501.378 4499.824
```

```
[1] 475.125 483.125
```

2.6 Localisation and quantification of detected peaks

With the information about their position in the combined datasets, each individual mz/RT feature is then located in the raw data. Due to the retention time correction, each feature is expected at the same RT position as in the combined EIC. However small shifts in retention time still occur for most of the peaks. In order to locate the correct position of each feature, the EIC of the selected mass is calculated for the whole retention time. This EIC is filtered with CWT using only the scale where the feature was optimally located on the combined EIC in step 3. Local maxima are calculated on this transformed data and the maximum with the closest position to the expected retention time is chosen.

```
R> xs <- create_datamatrix(xs = xs,
+          rxy = rxy)
```

2.7 The Output

The output is a xcmsSet object including all necessary information (peak location and peak area), for further data analysis (statistics, metabolite database information). The output is visible in Table 1 and can be extracted by using the following command.

```
R> peaktable <- xcms::peakTable(xs)
```

References

[1] David Fischer, Christian Panse, and Endre Laczko. *cosmiq: cosmiq - COmbining Single Masses Into Quantities*, 2017. R package version 1.10.0. URL: http://www.bioconductor.org/packages/devel/bioc/html/cosmiq.html.

[2] David Fischer. *Analysis of Metabolic Changes during Colorectal Cancer Development*. PhD thesis, University Zurich, Switzerland, May 2014. URL: http://opac.nebis.ch/F/?local_base=NEBIS&CON_LNG=GER&func=find-b&find_code=SYS&request=010268246.

[3] Ralf Tautenhahn, Christoph Boettcher, and Steffen Neumann. Highly sensitive feature detection for high resolution lc/ms. *BMC Bioinformatics*, 9:504, 2008. URL: https://doi.org/10.1186/1471-2105-9-504.

[4] Colin A. Smith. *faahKO: Saghatelian et al. (2004) FAAH knockout LC/MS data*, 2012. R package version 1.2.17. URL: http://dx.doi.org/10.1021/bi0480335.

[5] Pan Du, Warren A. Kibbe, and Simon M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22:2059–2065, 2006. URL: https://doi.org/10.1093/bioinformatics/btl355.

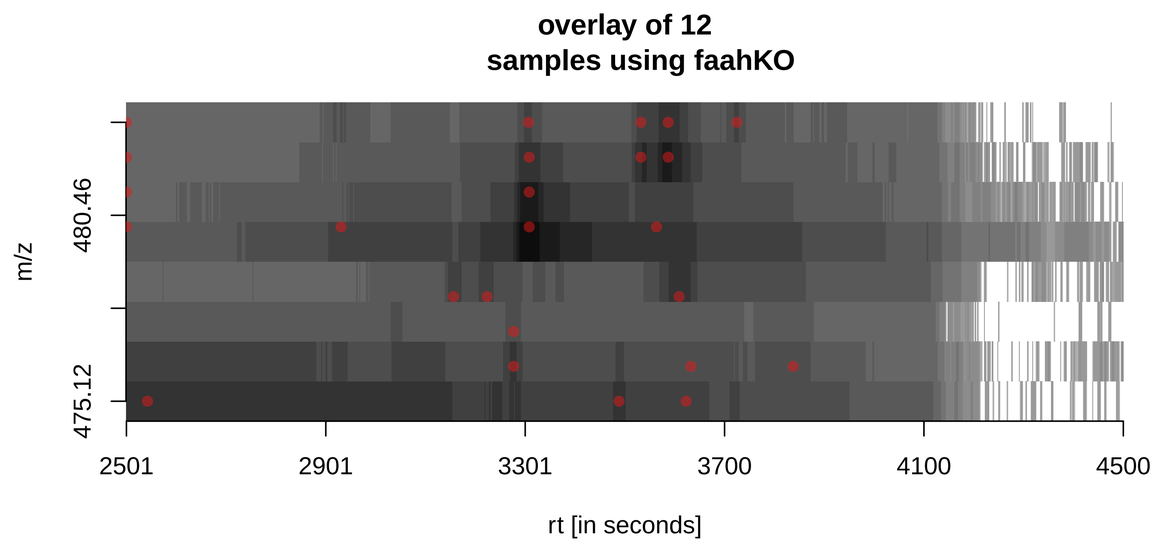


Figure 1: A "feature map", generated by using *cosmiq*, of the *faahKO* data is shown.

	mz	mzmin	mzmax	rt	rtmin	rtmax	npeaks	ko15.CDF	ko16.CDF	ko18.CDF	ko19.CDF	ko21.CDF	ko22.CDF	wt15.CDF	wt16.CDF	wt18.CDF	wt19.CDF	wt21.CDF	wt22.CDF
13	480.12	479.62	480.62	3308.89	3269.77	3346.45	12.00	50056574.74	49188673.99	42604200.43	32851699.17	32167083.04	28232603.93	50843961.77	53491143.06	44170197.70	31030987.86	33940727.49	26817322.53
16	481.12	480.62	481.62	3308.89	3269.77	3346.45	12.00	13099238.89	12892410.39	11239240.82	8722475.72	8416403.21	7386527.17	13329889.37	13761043.95	11389325.40	8178813.42	8871975.10	7162358.69
20	482.12	481.62	482.62	3587.46	3556.16	3625.01	12.00	9236048.74	9808221.16	9208820.69	5944352.74	6643532.65	4673246.58	6213538.11	9478365.28	8806738.76	7065545.40	6784720.21	5304007.67
1	475.12	474.62	475.62	2543.63	2501.38	2620.32	12.00	5186949.19	487408.49	5335581.42	6752927.55	1041640.40	325511.86	5199329.83	390509.41	4955607.87	7390373.01	1091444.77	332135.57
14	480.12	479.62	480.62	3563.98	3517.03	3599.97	12.00	3793218.13	3609194.92	3262888.26	2278957.11	2281625.97	1919592.19	3233217.08	3710765.24	3778748.06	2791484.26	2661819.89	1700321.22
18	482.12	481.62	482.62	3308.89	3269.77	3346.45	12.00	2436222.04	2411748.96	2178509.21	1850595.04	1760884.74	1513305.56	2430863.62	2558690.87	2246852.34	1680230.77	1852048.69	1488797.21
24	483.12	482.62	483.62	3587.46	3556.16	3625.01	12.00	2453488.79	2644783.20	2465869.41	1600305.36	1788174.64	1259131.65	1710400.41	2486024.63	2350213.82	1842232.12	1838208.47	1479050.87
10	478.12	477.62	478.62	3609.36	3557.72	3659.44	12.00	1781011.37	1530799.56	2339941.43	1746024.01	2002111.10	2306808.53	2315119.37	1184549.17	2596885.26	1610089.96	2152341.25	1936846.77
19	482.12	481.62	482.62	3532.68	3507.64	3556.16	12.00	2894511.13	2223288.19	1837688.94	1028426.47	1058758.71	937052.39	2817059.31	2387335.47	1801587.69	1226877.24	1316071.33	1132986.62
3	475.12	474.62	475.62	3623.45	3582.76	3676.66	12.00	1645767.46	327646.28	2111210.41	3635674.44	313337.16	134904.64	266162.14	264337.89	1828129.46	3371455.60	300460.78	116042.21
2	475.12	474.62	475.62	3488.86	3456.00	3523.29	12.00	1365292.03	207845.48	1974452.50	2501026.19	328727.49	138709.97	1562604.41	130589.92	1736168.07	3149059.84	288418.70	106937.94
12	480.12	479.62	480.62	2931.74	2891.05	2970.86	12.00	320143.15	2244074.91	970810.56	708970.12	453703.41	265494.04	1273170.72	530927.03	1143132.91	681184.49	878490.94	459179.04
5	476.12	475.62	476.62	3632.84	3556.16	3707.96	12.00	1326705.40	223602.11	991497.73	1793404.15	408466.66	219508.80	267728.22	214672.89	953058.81	1552155.54	330294.32	199937.41
4	476.12	475.62	476.62	3277.59	3250.99	3304.20	12.00	1017097.52	566738.36	938016.24	1006652.05	322057.40	277767.96	1014075.61	530283.76	776078.79	957879.55	330335.47	261446.90
23	483.12	482.62	483.62	3532.68	3506.08	3556.16	12.00	824724.45	717630.28	501639.82	300146.18	302852.18	259151.41	789391.93	633351.60	488747.06	340751.93	363169.24	318034.35
6	476.12	475.62	476.62	3837.85	3726.74	3938.01	12.00	75128.82	981658.22	713584.23	696291.39	128107.82	119574.53	61482.87	724052.74	901890.74	621936.43	124351.98	117733.72
9	478.12	477.62	478.62	3224.39	3189.96	3260.38	12.00	593890.64	624494.43	434788.25	274664.24	264986.46	230329.59	620885.22	589401.87	411366.45	300202.21	304203.29	238878.11
8	478.12	477.62	478.62	3157.09	3125.79	3189.96	12.00	496893.92	530543.66	454704.22	300974.64	313026.09	273014.41	569016.99	492274.06	405468.26	313791.05	331575.42	267200.70
22	483.12	482.62	483.62	3307.33	3274.47	3343.32	12.00	338584.18	350909.53	309209.05	294198.03	280731.04	238100.47	354019.82	359693.84	367188.74	254850.40	287734.63	263301.99
25	483.12	482.62	483.62	3725.17	3693.87	3758.04	12.00	106202.28	1157183.66	162057.02	50980.25	50969.98	97685.46	77838.50	449676.22	185229.53	45302.75	45332.99	191087.01

Table 1: The spreadsheet shows the top 20 most intense rows (`order(rowSums(peaktable[,8:19]), decreasing=TRUE)`) of the `peakTable` result.

Contact: ◇ ETHZ|UZH, Functional Genomics Center Zurich, Winterthurerstr. 190, CH-8057 Zurich, SWITZERLAND, Phone: +41 44 635 39 12, Fax: +41 44 635 39 22, EMail: cp@fgcz.ethz.ch, URL: http://bioconductor.org/packages/cosmiq/; ‡ David Fischer, Evolva SA, Basel, SWITZERLAND.