MetFamily Input Specification

Hendrik Treutler
hendrik.treutler@ipb-halle.de

Karin Gorzolka
karin.gorzolka@ipb-halle.de

Leibniz Institute for Plant Biochemistry,
Dept. of Stress and Developmental Biology,
Weinberg 3, 06120 Halle, Germany

May 2016

Abstract

The MetFamily web application is designed for the identification of regulated metabolite families [2]. This is possible on the basis of metabolite profiles for a set of MS^1 features as well as one MS/MS spectrum for each MS^1 feature. Group-discriminating MS^1 features are identified using a principal component analysis (PCA) of metabolite profiles and metabolite families are identified using a hierarchical cluster analysis (HCA) of MS/MS spectra. Regulated metabolite families are identified by considering group-discriminating MS^1 features from corporate metabolite families. The MetFamily web application is available at `http://msbi.ipb-halle.de/MetFamily/`.

1 Introduction

Here, we specify the input file format for the MetFamily web app version 1.0. Currently, MetFamily is able to process the output of MS-DIAL [7]. This guide enables researchers to reformat data from different sources to a format which is processible by MetFamily as well.

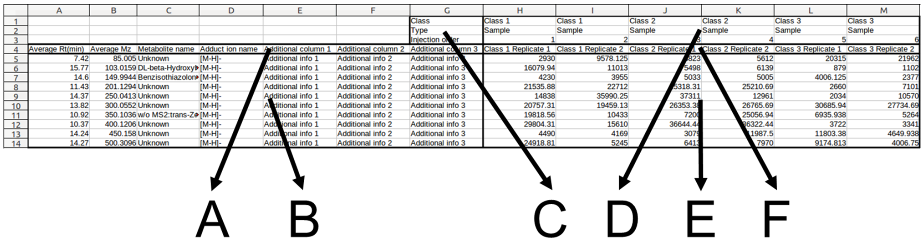
MetFamily processes a metabolite profile comprising a set of MS^1 features as well as an MS/MS library comprising MS/MS spectra for these MS^1 features. The metabolite profile contains m/z / retention time features from MS^1 scans plus the associated abundances and the MS/MS library contains MS/MS spectra of the MS^1 features with m/z and the intensities of the fragment ions. In the following we describe the file format of the metabolite profile and the file format of the MS/MS library.

2 Metabolite profile format

The metabolite profile is comprised in a table which is stored in a tab-delimted file (see Figure 1). This table contains a left part (A and B in Figure 1) comprising specifications of the MS^1 features (i.e. the precursor ions) and a right part (C, D, E, and F in Figure 1) comprising the abundances of the MS^1 features in the individual samples.

The left part comprises four columns with retention time (`Average Rt(min)`), m/z (`Average Mz`), the putative metabolite name (`Metabolite name`), the precursor ion species (`Adduct ion name`), and any number of additional columns. The combination of retention time and m/z constitutes unique IDs for the individual MS^1 features in MetFamily and the additional columns are not processed by MetFamily. The column header of the left part (A in Figure 1) is located in the forth row preceded by three blank rows. The column data of the left part (B in Figure 1) contains properties of the MS^1 features and starts in the fifth row downwards.

The right part comprises one column for each sample. The first three rows of the right part (D in Figure 1) comprise information about the individual samples, namely the sample-group (`Class`), sample-type (`Type`), and injection order (`Injection order`) of the samples, where only the sample-groups (e.g. Genotypes, Tissues, time points, ...) are used by MetFamily. The header of the sample information (`Class`, `Type`, `Injection order`; C in Figure 1) is located in the first three rows above the last column of the left part. The column header of the right part (F in Figure 1) contains the sample names and is located in the forth row preceded by three rows with sample information. The column data of the right part contains the abundances of the MS^1 features in the individual samples (E in Figure 1) and starts in the fifth row downwards.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1							Class	Class 1	Class 1	Class 2	Class 2	Class 3	Class 3
2							Type	Sample	Sample	Sample	Sample	Sample	Sample
3							Injection order	1	2	3	4	5	6
4	Average Rt(min)	Average Mz	Metabolite name	Adduct ion name	Additional column 1	Additional column 2	Additional column 3	Class 1 Replicate 1	Class 1 Replicate 2	Class 2 Replicate 1	Class 2 Replicate 2	Class 3 Replicate 1	Class 3 Replicate 2
5	7.42	85.005	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	2930	9578.125	[illegible]823	5612	20315	21962
6	15.77	103.0159	DL-beta-Hydroxyb	[M-H]-	Additional info 1	Additional info 2	Additional info 3	16079.94	11013	5498	6139	879	1102
7	14.6	149.9944	Benzisothiazolon	[M-H]-	Additional info 1	Additional info 2	Additional info 3	4230	3955	5033	5005	4006.125	2377
8	11.43	201.1294	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	21535.88	22712	[illegible]5318.31	25210.69	2660	7101
9	14.37	250.0413	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	14838	35990.25	37311	12961	2034	10570
10	13.82	300.0552	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	20757.31	19459.13	26353.38	26765.69	30685.94	27734.69
11	10.92	350.1036	w/o MS2:trans-Ze	[M-H]-	Additional info 1	Additional info 2	Additional info 3	19818.56	10433	7200	25056.94	6935.938	5264
12	10.37	400.1206	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	29804.31	15610	36644.44	36322.44	3722	3341
13	14.24	450.158	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	4490	4169	3079	[illegible]1987.5	11803.38	4649.938
14	14.27	500.3096	Unknown	[M-H]-	Additional info 1	Additional info 2	Additional info 3	24918.81	5245	6413	7970	9174.813	4006.75

Figure 1: Screenshot of a metabolite profile in a spreadsheet program with annotations for the different data sections. This example comprises ten MS^1 features which were measured in three sample groups with two replicates each. A, B: Header and values of various MS^1 feature properties. C, D: Header and values of sample information. E, F: Values and header of MS^1 feature abundances in the individual samples.

3 MS/MS library format

The MS/MS library is comprised in a text file in the MSP format (see Figure 2). This MS/MS library contains a set of MS/MS spectra which correspond to the MS^1 features in the metabolite profile. The MS/MS spectra in the MS/MS library are matched to the MS^1 features in the metabolite profile by matching retention time and m/z.

The MS/MS spectrum records are separated by blank lines. Each MS/MS spectrum record starts with spectrum meta data (A in Figure 2) followed by the spectrum data (B in Figure 2).

The spectrum meta data contains a set of properties each with a property name and a property value separated by colon and space / tab. Required properties are the name of the record (`NAME`, may be 'Unknown'), the retention time of the precursor ion (`RETENTIONTIME`), the m/z of the precursor ion (`PRECURSORMZ`), the name of the precursor ion (`METABOLITENAME`, may be 'Unknown' or empty), the ion species of the precursor ion (`ADDUCTIONNAME`), and the number of peaks in the MS/MS spectrum (`Num Peaks`).

The spectrum data contains a set of fragment peaks. Each fragment peak is characterized by m/z and intensity separated by space / tab.

In addition to the described format we also support differing syntax. Thus, also MSP files from MassBank and MassBank of North America (MoNA) can be imported into MetFamily.

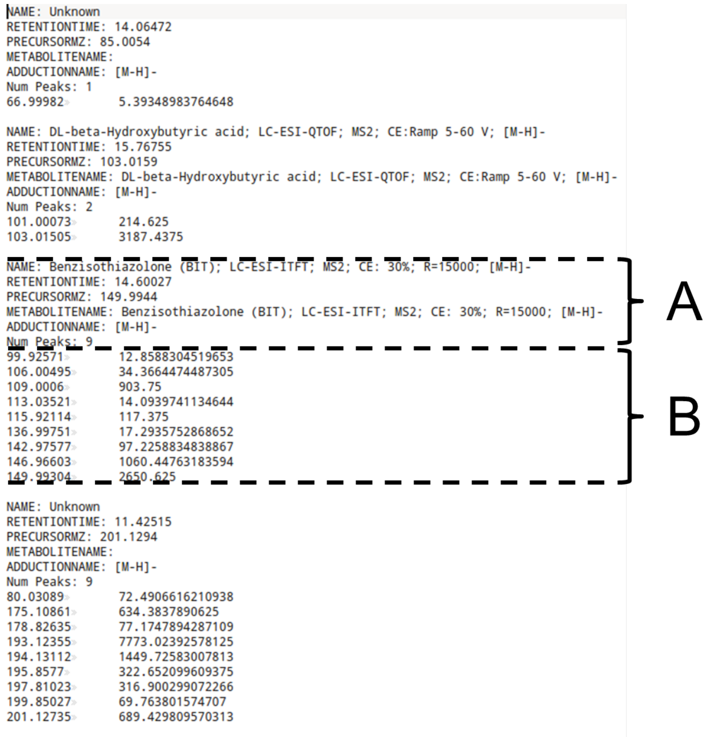
4 Data sources for MetFamily

In [2], we recommend SWATH data processed with MS-DIAL as input for MetFamily. However, we successfully applied MetFamily to various other kinds of MS data such as GC-EI-MS, idMSMS spectra, and LC-MS/MS with DDA. In the following, we exemplarily describe the conversion of GC-EI-MSS data to a MetFamily compatible format in detail.

4.1 Detailed example for GC-EI-MS data

We provide an exemplary R script for the processing of GC-EI-MS raw data and the conversion of the processed data into a MetFamily compatible format. We demonstrate the usage of this R script using a GC-EI-MS data set which was acquired in conventional EI ionization mode without any special MS settings for MetFamily. Consequently, this R script can be applied to all GC-EI-MS data in theory. Electron impact ionization causes characteristic fragmentation of the eluting ions, which resembles to MS/MS of the compounds. Please feel free to use or modify this R script in any way.

First, a set of GC-EI-MS raw data files are processed by standard functions of the R packages *xcms* [6] (version 1.44.0) and *CAMERA* [5] (version 1.27.0) (see Listing 1). Both *xcms* and *CAMERA* are freely available on Bioconductor [1]. The user is required to specify the folder with the GC-EI-MS raw data files



```
NAME: Unknown
RETENTIONTIME: 14.06472
PRECURSORMZ: 85.0054
METABOLITENAME:
ADDUCTIONNAME: [M-H]-
Num Peaks: 1
66.99982	5.39348983764648

NAME: DL-beta-Hydroxybutyric acid; LC-ESI-QTOF; MS2; CE:Ramp 5-60 V; [M-H]-
RETENTIONTIME: 15.76755
PRECURSORMZ: 103.0159
METABOLITENAME: DL-beta-Hydroxybutyric acid; LC-ESI-QTOF; MS2; CE:Ramp 5-60 V; [M-H]-
ADDUCTIONNAME: [M-H]-
Num Peaks: 2
101.00073	214.625
103.01505	3187.4375

NAME: Benzisothiazolone (BIT); LC-ESI-ITFT; MS2; CE: 30%; R=15000; [M-H]-
RETENTIONTIME: 14.60027
PRECURSORMZ: 149.9944
METABOLITENAME: Benzisothiazolone (BIT); LC-ESI-ITFT; MS2; CE: 30%; R=15000; [M-H]-
ADDUCTIONNAME: [M-H]-
Num Peaks: 9
99.92571	12.8588304519653
106.00495	34.3664474487305
109.0006	903.75
113.03521	14.0939741134644
115.92114	117.375
136.99751	17.2935752868652
142.97577	97.2258834838867
146.96603	1060.44763183594
149.99304	2650.625

NAME: Unknown
RETENTIONTIME: 11.42515
PRECURSORMZ: 201.1294
METABOLITENAME:
ADDUCTIONNAME: [M-H]-
Num Peaks: 9
80.03089	72.4906616210938
175.10861	634.3837890625
178.82635	77.1747894287109
193.12355	7773.02392578125
194.13112	1449.72583007813
195.8577	322.652099609375
197.81023	316.900299072266
199.85027	69.763801574707
201.12735	689.429809570313
```

Figure 2: Screenshot of a MS/MS library in a text file editor. This example comprises four MS/MS spectra which correspond to the first four MS^1 features in the metabolite profile example (see Figure 1).

(`rawData_dir`). The file format of the raw data must be supported by *xcms*, i.e. one of AIA/ANDI NetCDF, mzXML, mzData, and mzML. Upon execution, the script performs peak picking, feature grouping, and the annotation of pseudospectra and the output is an xsAnnotate object from the *CAMERA* package. Please note that the user will almost certainly have to adopt the individual parameters in order to meet the characteristics of the processed data.

Second, we define a function for the conversion of an xsAnnotate object to a metabolite profile and an MS/MS library which are usable as input for MetFamily (see Listing 2). Here, we specify an R function for the conversion, which requires an xsAnnotate object (the output of Listing 1) and the target files for the metabolite profile and the MS/MS library. The extracted metabolite profile and the extracted MS/MS library are automatically written to the specified target files.

Third, we exemplify the usage of the conversion function (see Listing 3). Here, we generate file names for the metabolite profile and the MS/MS library and pass these target files together with the xsAnnotate object from Listing 1 to the conversion function specified in Listing 2. The extracted metabolite profile and the extracted MS/MS library are automatically written to the specified target files.

We exemplified the usage of the described R script using data from the MetaboLights repository with accession ID 'MTBLS288' [3]. This data was published and analyzed in detail in [4]. Specifically, we selected a subset of the data comprising one rice cultivar (Qingfengai) during germination with all sampling time points (7, 10, 14, 28, and 48 days) and replicates provided in MetaboLights. We processed the raw data files using the R script from Listing 1 and extracted a metabolite profile and an MS/MS library using the R script Listing 3. We imported both files to MetFamily using the import parameters in Listing 4. We found clusters of sugars, amino acids, and lipid constituents using hierarchical cluster analysis and we found a clear separation of the sample groups using principal component analysis in agreement with the results in the original manuscript.

Listing 1: Processing of GC-EI-MS raw data with xcms and CAMERA.

```
##################################################################
## preprocess GC-EI-MS raw data with xcms/CAMERA

## Load libraries
library("xcms") # Peak picking, feature grouping across samples
library("CAMERA") # Metabolite Profile Annotation

## input files
rawData_dir <- "[/path/to/my/data]"
rawData_files <- list.files(rawData_dir, full.names=TRUE,
    recursive=TRUE)

## perform peak picking
xset <- xcmsSet(files=rawData_files, mzdiff=0.1, fwhm=5, step
    =0.1, snthresh=5)

## group samples
xset<-group(xset, bw=4, minfrac=0.1, mzwid=0.1, max=100)
## fill missing data
xset <- fillPeaks(xset)

## convert to xsAnnotate CAMERA object
xcam <- xsAnnotate(xset)
## extract pseudospectra according to FWHM
xcam <- groupFWHM(xcam, sigma=2, perfwhm=0.6, intval="maxo")
```

Listing 2: Function for the conversion of GC-EI-MS data processed with xcms and CAMERA.

```
#################################################################
## conversion-function: xsAnnotate object to MetFamily files
convertXsAnnotateCAMERAtoMetFamily <- function(xcam,
    fileMetaboliteProfile, fileMsMsLibrary){
  ######################################
  ## basic information
  xcam_report <- getPeaklist(xcam)
  phenoData <- xcam@xcmsSet@phenoData
  ## peaks on the black list
  peakMzOfDerivative <- c(73.1, 147.1)
  mzAbs <- 0.1
  ## replicates and sample groups
  numberOfUniqueSampleGroups <- length(unique(phenoData$class))
  sampleGroupNames <- as.character(phenoData$class)
  numberOfSamples <- nrow(phenoData)
  replicateNames <- rownames(phenoData)
  ## measurements
  numberOfPseudoSpectra <- max(as.numeric(xcam_report$pcgroup))
  sampleColumnMin <- c(3+3+1+numberOfUniqueSampleGroups + 1)
  sampleColumnMax <- sampleColumnMin + numberOfSamples - 1

  ######################################
  ## results: metabolite profile and MS/MS library
  metaboliteProfile <- data.frame(matrix(nrow =
      numberOfPseudoSpectra + 4, ncol = 4 + numberOfSamples),
      stringsAsFactors = FALSE)
  ## column names
  columnNames <- c("Average␣Rt(min)", "Average␣Mz",
       "Metabolite␣name", "Adduct␣ion␣name", replicateNames)
  colnames(metaboliteProfile) <- columnNames
  ## first three rows
  metaboliteProfile[1,] <- c("", "", "", "Class",
      sampleGroupNames)
  metaboliteProfile[2,] <- c("", "", "", "Type", rep(x = "Sample"
      , times = numberOfSamples))
  metaboliteProfile[3,] <- c("", "", "", "Injection␣order", 1:
      numberOfSamples)
  metaboliteProfile[4,] <- columnNames

  msmsLibrary <- vector(mode = "character")

  ######################################
  ## extract pseudospectra as MS/MS spectra
  for(pseudoSpectrumIndex in 1:numberOfPseudoSpectra){
```

```
## get pseudospectrum
peakIndeces <- which(xcam_report$pcgroup ==
    pseudoSpectrumIndex)
pseudoSpectrum <- xcam_report[peakIndeces, ]

## clean spectrum from ion from the derivative
peakIndecesToExclude <- unlist(lapply(X = peakMzOfDerivative,
    FUN = function(x){which(abs(pseudoSpectrum$mz - x) <=
    mzAbs)}))
if(length(peakIndecesToExclude) > 0)
  pseudoSpectrum <- pseudoSpectrum[-peakIndecesToExclude, ]
numberOfFragments <- nrow(pseudoSpectrum)

## get MS1 and MS/MS data
pseudoSpectrum$FeatureOverAllMean <- apply(X = pseudoSpectrum
    [, sampleColumnMin:sampleColumnMax], MARGIN = 1, FUN =
    mean)
quantifierIonIndex <- which.max(pseudoSpectrum$
    FeatureOverAllMean)
precursorAbundanceEstimate <- pseudoSpectrum[
    quantifierIonIndex, sampleColumnMin:sampleColumnMax]
precursorIndex <- which.max( pseudoSpectrum$mz)
if(pseudoSpectrum$isotopes[[precursorIndex]] != ""){
  clusterIndex <- gsub(x = pseudoSpectrum$isotopes[[
      precursorIndex]], pattern =
  "\\[(\\d+)\\]\\[M\\+\\d\\]\\+", replacement = "\\1")
  precursorIndex <- grep(x = pseudoSpectrum$isotopes, pattern
      = paste("\\[", clusterIndex, "\\]\\[M\\]\\+", sep = ""))
}
precursorMz <- pseudoSpectrum$mz[[precursorIndex]]
precursorRt <- mean(pseudoSpectrum$rt)
precursorLabel <- paste(precursorMz, precursorRt, sep = "␣/␣")

## record data
metaboliteProfile[4+pseudoSpectrumIndex,] <- c(
  precursorRt,
  precursorMz,
  "Unknown",
  "Unknown", #"[M+H]"
  precursorAbundanceEstimate
)

msmsLibrary <- c(msmsLibrary,
  paste("NAME", precursorLabel, sep = ":␣"),
  paste("RETENTIONTIME", precursorRt, sep = ":␣"),
  paste("PRECURSORMZ", precursorMz, sep = ":␣"),
      paste("METABOLITENAME", "Unknown", sep = ":␣"),
      paste("ADDUCTIONNAME", "Unknown", sep = ":␣"),
      paste("Num␣Peaks", numberOfFragments, sep = ":␣"),
      paste(pseudoSpectrum$mz, pseudoSpectrum$FeatureOverAllMean,
          sep = "\t"),
      ""
    )
  }

  ######################################
  ## write files
  write.table(x = metaboliteProfile, file = fileMetaboliteProfile
      , sep = "\t", row.names = FALSE, col.names = FALSE, quote =
      FALSE)
  writeLines(text = msmsLibrary, con = fileMsMsLibrary)
}
```

Listing 3: Conversion of GC-EI-MS data processed with xcms and CAMERA.

```
##################################################################
## convert xsAnnotate object to MetFamily files

## files to write
dataSetName <- gsub("␣", "_", gsub(":", ".", Sys.time()))
fileMetaboliteProfile <- paste(dataSetName,
       "_Metabolite_profile.tsv",sep="")
fileMsMsLibrary <- paste(dataSetName, "_MSMS_library.msp",sep="")

convertXsAnnotateCAMERAtoMetFamily(xcam, fileMetaboliteProfile,
    fileMsMsLibrary)
```

Listing 4: File content of the import parameter file from MetFamily.

```
# This is the set of parameters which have been used for the
# initial data import to MetFamily 1.0
# Exported with MetFamily 1.0
projectName=Subset of MTBLS288
projectDescription=Subset of MTBLS288 published in Hu et al 2016
toolVersion=MetFamily 1.0
minimumIntensityOfMaximalMS2peak=20
minimumProportionOfMS2peaks=0.03
mzDeviationAbsolute_grouping=0.1
mzDeviationInPPM_grouping=150
doPrecursorDeisotoping=TRUE
```

```
mzDeviationAbsolute_precursorDeisotoping=0.1
mzDeviationInPPM_precursorDeisotoping=100
maximumRtDifference=0.02
doMs2PeakGroupDeisotoping=TRUE
mzDeviationAbsolute_ms2PeakGroupDeisotoping=0.1
mzDeviationInPPM_ms2PeakGroupDeisotoping=100
proportionOfMatchingPeaks_ms2PeakGroupDeisotoping=0.9
mzDeviationAbsolute_mapping=0.01
neutralLossesPrecursorToFragments=FALSE
neutralLossesFragmentsToFragments=FALSE
```

References

[1] Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80–16, 2004.

[2] A. Porzel A. Tissier S. Neumann H. Treutler, H. Tsugawa and G. Balcke. Discovering regulated metabolite families in untargeted metabolomics studies. *Submitted to Anal Chem*, 2016.

[3] Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, Eamonn Maguire, Alejandra González-Beltrán, Susanna-Assunta A. Sansone, Julian L. Griffin, and Christoph Steinbeck. MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(Database issue):D781–D786, January 2013.

[4] Chaoyang Hu, Takayuki Tohge, Shen-An A. Chan, Yue Song, Jun Rao, Bo Cui, Hong Lin, Lei Wang, Alisdair R. Fernie, Dabing Zhang, and Jianxin Shi. Identification of conserved and diverse metabolic shifts during rice grain development. *Scientific reports*, 6, 2016.

[5] Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R. Larson, and Steffen Neumann. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry*, 84(1):283–289, January 2012.

[6] Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing Mass Spectrometry Data for Metabolite

Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.*, 78(3):779–787, February 2006.

[7] Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Meth*, 12(6):523–526, June 2015.