

CompMS2miner_Workflow

WMB Edmands

June 27, 2016

Matches MS1 features to MS2 spectra (.mzXML) files based on a mass-to-charge and retention time tolerance. Composite spectra and other data can subsequently be visualized during any stage of the CompMS2miner processing workflow. Composite spectra can be denoised, ion signals grouped and summed, substructure groups identified, common Phase II metabolites predicted, calculation of a correlation network and features matched to data bases monoisotopic mass data and insilico MS2 fragmentation data. The resulting data can then be readily curated by publishing as online shiny application on shinyapps.io, as a zip file which can be shared or by sending to a local or online couchDB database.

At any stage during the compMS2 workflow a compMS2 class object can be visualized with a Shiny application *compMS2explorer*. Full usage of the *compMS2explorer* application functionality requires an internet connection. The end result of following the workflow within this document using the example data provided can be visualized using the function, comments made in the metID comments tab can be saved on closing the application by assigning the compMS2explorer output to an object:

```
library(CompMS2miner)
# assign any metabolite identification comments to a new or the same "CompMS2" object
compMS2example_commented <- compMS2explorer(compMS2example)
```

The following example illustrates the CompMS2miner workflow:

1. construct compMS2 class object.

From MS2 data (in the .mzXML file format) and an MS1feature table. The compMS2 object can also be constructed in as a parallel computation in the case of large peak tables and/ or larger numbers of MS2 data files.

```
# file path example MS1features in comma delimited csv file
# (see ?example_mzXML_MS1features for details).
MS1features_example <- system.file("extdata", "MS1features_example.csv",
                                   package = "CompMS2miner")

# mzXml file examples directory
mzXmlDir_example <- dirname(MS1features_example)
# observation MS1 feature table column names character vector for corrNetwork function
obsNames <- c(paste0(rep("ACN_80_", 6), rep(LETTERS[1:3], each=2), rep(1:2, 3)),
              paste0(rep("MeOH_80_", 6), rep(LETTERS[1:3], each=2), rep(1:2, 3)))
# use parallel package to detect number of cores
nCores <- parallel::detectCores()

# read in example peakTable
peakTable <- read.csv(MS1features_example, header=TRUE, stringsAsFactors=FALSE)
# create compMS2 object
compMS2demo <- compMS2(MS1features = peakTable,
                       mzXMLdir = mzXmlDir_example, nCores=nCores,
                       mode = "pos", precursorPpm = 10, ret = 20,
                       TICfilter = 10000)
## creating compMS2 object in positive ionisation mode
```

```
## Loading required package: foreach
## Loading required package: Rcpp
## Loading required package: shiny
## 2 MS2 (.mzXML) files were detected within the directory...
## Starting SNOW cluster with 8 local sockets...
## matching MS1 peak table features to the following MS2 files:
## DDA_ACN_80.mzXML
## DDA_MeOH_80.mzXML
##
## "obsNames" in argsCorrNetwork argument missing. Not performing correlation network calculation.

# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 202 MS1 features were matched to 2514 MS2 precursor scans
## containing 323968 ion features
##
## Average ppm match accuracy: 1.979
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 6.85 seconds
## with a retention time tolerance of (+/-) 20 seconds
##
## Memory usage: 6.53 MB
```

2. Dynamic noise filtration.

filter variable noise from the data using a dynamic noise filter.

```
# dynamic noise filter
compMS2demo <- deconvNoise(compMS2demo, "DNF")
## Applying dynamic noise filter to 327 spectra...
## Starting SNOW cluster with 8 local sockets...
## ...done
## 326 spectra contained more than or equal to 3 peaks following dynamic noise filtration
# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 202 MS1 features were matched to 2512 MS2 precursor scans
## containing 6425 ion features
##
## Average ppm match accuracy: 1.983
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 6.87 seconds
## with a retention time tolerance of (+/-) 20 seconds
##
## Memory usage: 1.74 MB
```

3. Intra-spectrum ion grouping and inter-MS2 file spectra grouping with signal summing.

group and sum ions from different scans and then combine summed ion composite spectra across multiple files. This create a single composite spectra for each MS1 EIC matched to MS2 precursor scans. Optionally, a minimum inter-spectrum similarity filter (dot product) can be applied where only spectra with a minimum spectral similarity (see argument `specSimFilter` in `?combineMS2.Spectra`) will be combined in the resultant composite spectrum. However, if none of the spectra of a group have any similarity to one another the spectrum with the highest mean precursor intensity will be returned.

```
# intra-spectrum ion grouping and signal summing
compMS2demo <- combineMS2(compMS2demo, "Ions")
## Grouping ions in 326 spectra...
## Starting SNOW cluster with 8 local sockets...
## ...done
## 311 spectra contained more than or equal to 3 peaks following ion grouping
## The range of interfragment differences less than 0.1 m/z in the spectra is min : 1 max : 4
## The average number of interfragment differences less than 0.1 m/z in the spectra is 1
# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 198 MS1 features were matched to 2347 MS2 precursor scans
## containing 2166 ion features
##
## Average ppm match accuracy: 1.997
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 6.83 seconds
## with a retention time tolerance of (+/-) 20 seconds
##
## Memory usage: 1.9 MB
#inter spectrum ion grouping and signal summing if the argument specSimFilter
# is supplied (values 0-1) then any spectrum below this spectral similarity
# threshold (dot product) will not be included in the composite spectrum generated.
compMS2demo <- combineMS2(compMS2demo, "Spectra", specSimFilter=NULL)
## Combining 311 spectra by MS1 feature number...
## Starting SNOW cluster with 8 local sockets...
## ...done
## 198 composite spectra contained more than or equal to 3 peaks following ion grouping
## The range of interfragment differences less than 0.1 m/z in the composite spectra is min : 1 max : 5
## The average number of interfragment differences less than 0.1 m/z in the composite spectra is 1
# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 198 MS1 features were matched to 2347 MS2 precursor scans
## containing 1576 ion features
##
## Average ppm match accuracy: 2.084
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 6.83 seconds
## with a retention time tolerance of (+/-) 20 seconds
```

```
##  
## Memory usage: 1.83 MB
```

4. Possible substructure identification.

Characteristic neutral losses/ fragments of electrospray adducts and metabolites from literature sources (?Substructure_masses for details).

```
# annotate substructures  
compMS2demo <- subStructure(compMS2demo, "Annotate")  
## matching Precursor to fragment and interfragment neutral losses and fragments in 198 composite spectra  
# identify most probable substructure annotation based on total relative intensity  
# explained  
compMS2demo <- subStructure(compMS2demo, "prob")  
## Identifying likely substructure type...  
# summary of most probable substructure annotation  
mostProbSubStr <- subStructure(compMS2demo, "probSummary")  
## Substructure annotation summary :  
## 198 composite spectra  
## 183 substructures identified above minimum sum relative intensity of 30  
##  
## SubStr.table  
##  
## phosphatidylcholine  
## 99  
## methionine side chain  
## 43  
## polysiloxane  
## 30  
## phthalate  
## 4  
## carboxylic acids, aldehydes, ester  
## 2  
## nitroaromatics, hydroxyaldehydes  
## 2  
## aromatic n-methylamines, methoxy derivative, tert. butyl  
## 1  
## methyl esters  
## 1  
## triazines  
## 1
```

5. Metabolite identification methods.

A variety of metabolite identification methods are implemented through the function metID (see ?metID) for further details.

```
# annotate composite MS2 matched MS1 features to metabolomic databases (default  
#is HMDB, also DrugBank, T3DB and ReSpec databases can also be queried).  
#Warning: this may take 2-3 mins as large number of query masses  
compMS2demo <- metID(compMS2demo, "dbAnnotate")  
## matching 198 unknowns to 8,902,538 possible ESI artefacts and substructure mass shifts...
```

```
# match composite spectra to spectral databases as .msp files (e.g. lipidBlast, source http://prime.psc.riken.jp/Metabolomics\_Software/MS-DIAL/LipidBlast\_PositiveIonization)
compMS2demo <- metID(compMS2demo, 'matchSpectralDB',
                      mspFile='http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/LipidBlast_PositiveIonization')

# massBank .msp
compMS2demo <- metID(compMS2demo, 'matchSpectralDB',
                      mspFile='http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/MassBank_MSMS_PositiveIonization')

# ReSpect .msp
compMS2demo <- metID(compMS2demo, 'matchSpectralDB',
                      mspFile='http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/Respect_20120925')
```

```
# metFrag insilico fragmentation.
compMS2demo <- metID(compMS2demo, "metFrag")
```

```
# calculate spectral similarity network (dot product >= 0.8 default)
compMS2demo <- metID(compMS2demo, 'specSimNetwork')
## Loading required package: igraph
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:stats':
##
##      decompose, spectrum
## The following object is masked from 'package:base':
##
##      union
## Calculating dot product matrix 198 spectra
##
## Calculating dot product matrix 198 spectra
## 182 nodes with 2431 edges identified at a spectral similarity (dot product score) >= 0.8
## Of the edges:
## 2286 were based on a shared ion fragment pattern
## 145 were based on a shared neutral loss pattern
#####
### data pre-processing MS1features with MetMSLine package ###
#####

if(!require(MetMSLine)){
  # if not installed then install from github
  devtools::install_github('WMBEdmands/MetMSLine')
  require(MetMSLine)
}
```

```

## Loading required package: MetMSLine
## Loading required package: tcltk2
## Loading required package: tcltk
##
## Attaching package: 'tcltk2'
## The following object is masked from 'package:shiny':
##
##     tag
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Loading required package: fpc
## Loading required package: fastcluster
##
## Attaching package: 'fastcluster'
## The following object is masked from 'package:stats':
##
##     hclust

# zero fill
peakTable <- zeroFill(peakTable, obsNames)
## zero filling with half the minimum non-zero value
# log transform
peakTable <- logTrans(peakTable, obsNames)
## log transforming to the base 2.718...

#####
##### end MetMSLine pre-processing #####
#####

# add correlation network using pre-processed MS2 matched peak table
compMS2demo <- metID(compMS2demo, method='corrNetwork', peakTable, obsNames,
                     corrMethod='pearson', corrThresh=0.95, MTC='none', MS2only=3)
## Calculating correlation matrix for 198 features
## less than 300 nodes using Fruchterman-Reingold layout. see ?igraph::with_fr()
## 141 nodes with 595 edges identified at a correlation threshold >= 0.95 (pearson, p/q value <= 0.05, 1

```

6. Optional curate data as zip file, in CouchDB and/or publish results as a shiny application.

CompMS2 class objects can be sent to either a local or online couchDB database. Furthermore, the results of the CompMS2miner workflow can be published on the web as a shiny application to share metabolite identification information or as a self-contained zip file using the publishApp function.

```

# publish your app to shinyapps.io see ?publishApp for more details
# you may need to install the rsconnect and shinyapps packages and also sign up for a shinyapps.io account
# quick guide here for setting up your account: http://shiny.rstudio.com/articles/shinyapps.html
publishApp(compMS2demo, appName='compMS2demo')

```

sessionInfo

```
sessionInfo()
```

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] tcltk      stats      graphics  grDevices  utils      datasets  methods
## [8] base
##
## other attached packages:
## [1] MetMSLine_1.0.0    fastcluster_1.1.20 fpc_2.1-10
## [4] zoo_1.7-13         tcltk2_1.2-11      igraph_1.0.1
## [7] shiny_0.13.2       Rcpp_0.12.5        foreach_1.4.3
## [10] CompMS2miner_1.2.4 knitr_1.13
##
## loaded via a namespace (and not attached):
## [1] DEoptimR_1.0-4     compiler_3.3.0     formatR_1.4
## [4] plyr_1.8.4         class_7.3-14       iterators_1.0.8
## [7] tools_3.3.0        prabclus_2.2-6     digest_0.6.9
## [10] mclust_5.2         evaluate_0.9        lattice_0.20-33
## [13] doSNOW_1.0.14      yaml_2.1.13        parallel_3.3.0
## [16] mvtnorm_1.0-5      trimcluster_0.1-2  stringr_1.0.0
## [19] cluster_2.0.4      diptest_0.75-7     stats4_3.3.0
## [22] grid_3.3.0         nnet_7.3-12        robustbase_0.92-5
## [25] R6_2.1.2           snow_0.4-1         flexmix_2.3-13
## [28] rmarkdown_0.9.6    kernlab_0.9-24     reshape2_1.4.1
## [31] magrittr_1.5        codetools_0.2-14   htmltools_0.3.5
## [34] modeltools_0.2-21  MASS_7.3-45        mime_0.4
## [37] xtable_1.8-2       httpuv_1.3.3       stringi_1.1.1
```