CompMS2miner_Workflow

WMB Edmands

May 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:

```
library(CompMS2miner)
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 = 10,
                       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
##
## 163 MS1 features were matched to 1784 MS2 precursor scans
## containing 220161 ion features
##
## Average ppm match accuracy: 1.77
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 3.93 seconds
## with a retention time tolerance of (+/-) 10  seconds
##
## Memory usage: 4.57 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 257 composite spectra...
## Starting SNOW cluster with 8 local sockets...
## ...done
## 256 composite spectra contained more than or equal to 5 peaks following dynamic noise filtration
# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 163 MS1 features were matched to 1782 MS2 precursor scans
## containing 5030 ion features
##
## Average ppm match accuracy: 1.775
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 3.94 seconds
## with a retention time tolerance of (+/-) 10  seconds
##
## Memory usage: 1.32 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.

```
# intra-spectrum ion grouping and signal summing
compMS2demo <- combineMS2(compMS2demo, "Ions")
## Grouping ions in 256 composite spectra...
## Starting SNOW cluster with 8 local sockets...
## ...done
## 250 composite spectra contained more than or equal to 5 peaks following ion grouping
## The range of interfragment differences less than 0.1 m/z in the composite spectra is min : 1 max : 12
## The average number of interfragment differences less than 0.1 m/z in the composite spectra is 5
## Warning in .local(object, ...): The average number of interfragment
## differences less than 0.1 m/z is greater than 2: please consider increasing
## the mzError parameter above 0.001
# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 162 MS1 features were matched to 1738 MS2 precursor scans
## containing 2820 ion features
##
## Average ppm match accuracy: 1.789
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 3.93 seconds
## with a retention time tolerance of (+/-) 10  seconds
##
## Memory usage: 1.64 MB
#inter spectrum ion grouping and signal summing
compMS2demo <- combineMS2(compMS2demo, "Spectra")
## Combining 250 composite spectra by MS1 feature number...
## Starting SNOW cluster with 8 local sockets...
## ...done
## 162 composite spectra contained more than or equal to 5 peaks following ion grouping
## The range of interfragment differences less than 0.1 m/z in the composite spectra is min : 1 max : 14
## The average number of interfragment differences less than 0.1 m/z in the composite spectra is 6
## Warning in .local(object, ...): The average number of interfragment
## differences less than 0.1 m/z is greater than 2: please consider increasing
## the mzError parameter above 0.001
# View summary of compMS2 class object at any time
compMS2demo
## A "CompMS2" class object derived from 2 MS2 files
##
## 162 MS1 features were matched to 1738 MS2 precursor scans
## containing 2304 ion features
##
## Average ppm match accuracy: 1.817
## with a ppm mass accuracy tolerance of (+/-) 10
##
## Average retention time match accuracy: 3.77 seconds
## with a retention time tolerance of (+/-) 10  seconds
```

```
##
## Memory usage: 1.57 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 162 composite spect
# 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 :
## 162 composite spectra
## 151 substructures identified above minimum sum relative intensity of 30
##
## SubStr.table
##              phosphatidylcholine            methionine side chain
##                               81                               32
##                     polysiloxane                        phthalate
##                               32                                3
## nitroaromatics, hydroxyaldehydes                    nitrotoluenes
##                                1                                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 ReSpect databases can also be queried).
#Warning: this may take 2-3 mins as large number of query masses
compMS2demo <- metID(compMS2demo, "dbAnnotate")
## matching 162 unknowns to 8,902,538 possible ESI artefacts and substructure mass shifts...

# select most probable annotations based on substructures detected
compMS2demo <- metID(compMS2demo, "dbProb")
```

```
# predict Phase II metabolites from SMILES codes
compMS2demo <- metID(compMS2demo, "predSMILES")

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

MetMSLine data-preprocessing and correlation network calculation.

```
##############################################################
### 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
# calculate coefficient of variation for Acetonitrile extraction replicates
peakTable <- cvCalc(peakTable, obsNames[grep('ACN', obsNames)], thresh=Inf)
## calculating coefficient of variation...
## No features were found to be below the CV% threshold of Inf
# rename coeffVar column
colnames(peakTable)[ncol(peakTable)] <- 'coeffVar_ACN'
# calculate coefficient of variation for Methanol extraction replicates
peakTable <- cvCalc(peakTable, obsNames[grep('MeOH', obsNames)], thresh=Inf)
## calculating coefficient of variation...
##
## No features were found to be below the CV% threshold of Inf
# rename coeffVar column
colnames(peakTable)[ncol(peakTable)] <- 'coeffVar_MeOH'
# all features less than 20% cv either ACN replicates or MeOH replicates
peakTable <- peakTable[peakTable$coeffVar_ACN <= 20 | peakTable$coeffVar_MeOH <= 20, ]
# deconvolute data with RamClust modified for MetMSLine
wMeanTable <- ramClustMod(peakTable, obsNames)
```

```
##
##  calculating ramclustR similarity: nblocks =  1
##  finished:1
##
## RAMClust feature similarity matrix calculated and stored: 0 minutes
## Converting RAMClustR similarity matrix to a distance object...this
##         is a potentially computationally intensive step please wait...
##
##
## RAMClust distances converted to distance object: 0 minutes
##
## fastcluster based clustering complete: 0 minutes
## cutting dendrogram dynamically...
## Calculating weighted mean for 296 pseudospectra accounting for 1694 of 1855 total features
peakTable <- wMeanTable$wMeanPspec
# log transform
peakTable <- logTrans(peakTable, obsNames)
## log transforming to the base 2.718...

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

# move eic nos to first column for corr network function
peakTable <- cbind(peakTable$EICno, peakTable)

# add correlation network using pre-processed peak table
compMS2demo <- metID(compMS2demo, method='corrNetwork', peakTable, obsNames,
                           corrMethod='pearson', corrThresh=0.9, MTC='none')
## 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 correlation matrix for 296 features
## less than 300 nodes using Fruchterman-Reingold layout. see ?igraph::with_fr()
## 259 nodes with 1504 edges identified at a correlation threshold >= 0.9 (pearson, p/q value <= 0.05,
```

6. Optional curate data in CouchDB/ publish results as a shiny App.

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 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 acco
# quick guide here for setting up your account: http://shiny.rstudio.com/articles/shinyapps.html
publishApp(compMS2demo, appName='compMS2demo')
```