simExTargId (simultaneous experiment - MS/MS target identification)

WMB Edmands

June 12, 2017

The following illustrates the simExTargId workflow with example (.RAW) data files of human plasma obtained on a Thermo FT-ICR-MS.

First the warning email notification function (*emailNotifier*) must be initiated in a seperate R session to monitor the raw data directory on the mass spectrometer workstation.

For the purposes of this vignette the main function (*simExTargId*) can be illustrated by simulating a real-time data acquisition worklist. First a dummy raw data directory must be created on your computer's hard drive into which example raw data files can be moved. After moving the raw data files in to the directory their last write time can be adjusted (*Sys.setFileTime*) to initiate the **simExTargId** autonomous metabolomic data-analysis process. This allows the full functionality of **simExTargId** to be demonstrated with example data rather than when needed for a metabolomic profiling data collection.

1. Set the real-time email notifier

The email notifier must be run in a seperate R session to ensure the raw data directory on the mass spectrometer workstation can be closely monitored and any stoppage/error quickly detected. The `emailNotifier` function can send warning emails to a character vector of multiple email addresses (e.g. within a laboratory group) if an experimental run has stopped unexpectedly or if there are spray quality issues for example leading to files which are much smaller in size (default: >= 3 median absolute deviations) than the files already collected. If this function as part of the main **simExTargId** function there might be a delay in informing the user of an unexpected instrument stoppage or spray quality issues for example when file-conversion, peak-picking or another lengthy process is taking place.

```
# select directory where example directories for raw files can be created and also results
# output saved
# For example the C: directory
# create a dummy raw MS1 profiling data directory
# for example
studyName <- paste0(gsub(" ", "", format(Sys.time(), "%Y %m %d")), "_simExTargId_Example")
dummyRawDir <- paste0("C:/", studyName)
# create directory
dir.create(dummyRawDir)
# replace email address with a vector of one or more email addresses of
# people in your laboratory group
emailNotifier(rawDir=dummyRawDir, emailAddress='johndoe@emailprovider.com',
              emailTime=5)
```

2. Set the main simExTargId function running

Set the main `simExTargId` function running in a different R session to the `emailNotifier` function. This function is essentially a giant wrapper function for a multitude of other R packages and softwares and utilizes a while loop to continually monitor the raw data directory on the mass spectrometer computer until the last

data file acquired is beyond a maximum waiting time (see maxTime argument). The function automatically generates a recommended and organized sub-directory structure in to which all output from `simExTargId` is continually saved.

A co-variates/worklist table supplied at initiation will inform both the xcms sub-directory grouping, the pooled QC-based signal adjustment and CV% filtering, PCA analysis and also the automatic statistical analysis.

As new data files are collected they will be automatically converted to the mzXML open file format (MSConvert, ProteoWizard) and peak-picked (`xcmsSet`). After a minimum number of files has been collected subsequent stages of retention time alignment (`retcor.obiwarp`), grouping (`xcms::group`) and missing peak imputation (`xcms::fillPeaks`). Following xcms CAMERA identifies pseudospectra, isotopes and ESI adducts/in-source fragments. The resulting peak table is then pre-processed, outliers identified by PCA and finally automatic co-variate based statistical analysis is performed using various functions from the MetMSLine package. A function peakMonitor internal to the simExTargId function is able to monitor signal-drift/attenuation of a database (.csv) of previously identified metabolites supplied at initiation.

N.B. Make sure that you have the **MSConvert** software in your path if using Windows and that you are able to successfully run the command `> MSConvert` in your system shell. This is necessary for the automatic mzXML/mzML file conversion process and **simExTargId** will not work successfully without it.

```
# simExTargId extdata directory
extdataDir <- system.file("extdata", package="simExTargId")
# list blank files
blanksRaw <- list.files(extdataDir, pattern="blank", full.names=TRUE)
# list plasma IPA extract files
samplesRaw <- list.files(extdataDir, pattern="sample", full.names=TRUE)
# covariates file
coVariates <- paste(extdataDir, "coVariates.csv", sep="/")
# illustrative metabolite database table
metabDb <- paste(extdataDir, 'exampleMetabDatabase.csv', sep='/')
# identify number of virtual cores for parallel processing using parallel package
nCores <- parallel::detectCores()

# as no qc files then use sample files twice to illustrate the
# peakMonitor function

# move the blank files into the temporary directory to start the process
blankRawCopies <- paste(dummyRawDir, basename(blanksRaw), sep="/")
file.copy(from=blanksRaw, to=blankRawCopies)
# set the file time to simulate the files having been acquired at least 5 mins
# since last modification
setTheTime <- function(fileCopy, time){
  Sys.setFileTime(fileCopy, Sys.time() - time)
}
# apply to newly copied files
sapply(blankRawCopies, setTheTime, 240)

# move the plasma samples twice first time rename as QC
# and set the file time less than 5 mins
samplesRawCopies <- paste(dummyRawDir, basename(samplesRaw), sep="/")
file.copy(from=samplesRaw, to=samplesRawCopies)
qcFiles <- gsub('sample', 'qc', samplesRawCopies)
file.rename(from=samplesRawCopies, to=qcFiles)
file.copy(from=samplesRaw, to=samplesRawCopies)
```

```
# apply to newly copied files
sapply(c(samplesRawCopies, qcFiles), setTheTime, 300)

# Start simExTargId function
simExTargId(rawDir=dummyRawDir, studyName = studyName, analysisDir='C:/',
            coVar=coVariates, metab=metabDb, nCores=nCores, ionMode='nega',
            minFiles=3)
```

simExTargId will wait until at least five minutes after the raw data file was last modified, to ensure that the file acquisition has fully completed.

After at least 3 samples of each class found in the second column of the co-variates table, retention time alignment, grouping, zero-filling, then pre-processing, PCA analysis, stats analysis and data-deconvolution will occur.

3. peakMonitor (shiny application)

A database table of previously identified metabolites can be monitored in real-time and analytical CV% and signal attenuation affects monitored using the shiny application `peakMonitor`

```
peakMonitor(analysisDir=paste0(dummyRawDir, "_analysis/NEG/output/peakMonitor"))
```

4. targetId (shiny application)

During a run the output of the statistical analyses can be viewed using the shiny application `targetId` a zip file containing containing a .csv file for each statistical test after setting the thresholds can be downloaded and used to guide and plan further MS/MS experiments.

```
# this command will open the application in your web-browser
targetId(analysisDir=paste0(dummyRawDir, "_analysis/NEG/output/04.stats"))
```