Package 'MSeasy'

September 7, 2011

Type Package

Title Preprocessing of Gas Chromatography-Mass Spectrometry (GC-MS) data

Version 5.2

Date 2011-09-06

Depends amap, clValid, cluster, fpc

Suggests xcms, tcltk

Author Elodie Courtois, Yann Guitton, Florence Nicole

Maintainer <florence.nicole@univ-st-etienne.fr>

Description Package for the detection of molecules in complex mixtures of compounds. It creates an initial data matrix from several GC-MS analyses by collecting and assembling the information from chromatograms and mass spectra (MS.DataCreation). It tests for the best unsupervised clustering method to group similar mass spectra into molecules (MS.test.clust). It runs the optimal unsupervised clustering method on the initial data matrix, identifies the optimal number of clusters, produces different files for facilitating the identification of molecules and returns fingerprinting or profiling matrices (MS.clust).

License GPL-2

URL http://sites.google.com/site/rpackagemseasy/

LazyLoad yes

R topics documented:

`MSeasy-package` *Preprocessing of Gas Chromatography-Mass Spectrometry (GC-MS) data*

Description

Package for the detection of molecules in complex mixtures of compounds. It creates an initial data matrix from several GC-MS analyses by collecting and assembling the information from chromatograms and mass spectra (*MS.DataCreation*). It tests for the best unsupervised clustering method to group similar mass spectra into molecules (*MS.test.clust*). It runs the optimal unsupervised clustering method on the initial data matrix, identifies the optimal number of clusters, produces different files for facilitating the identification of molecules and returns fingerprinting or profiling matrices (*MS.clust*).

Details

Package:	MSeasy
Type:	Package
Version:	5.2
Date:	2011-09-06
License:	GPL-2
LazyLoad:	yes

Author(s)

Elodie Courtois, Yann Guitton, Florence Nicole

`Agilent_MSDataCreation` *Demonstration folder for MS.DataCreation*

Description

This demonstration folder includes 2 GC-MS analyses of Lavandula obtained from Agilent. The two analyses represent a total of 54 chromatogram's peaks. The folder can be used with the function MS.DataCreation that collects and assembles the information from chromatograms and mass spectra of the two samples in a initial data matrix with peaks in row and mass spectrum in columns.

Usage

```
Agilent_MSDataCreation
```

Format

A folder with two different sub-folders, each corresponding to one GC-MS analysis. Each sub-folder contains an export3ddata and a rteres files.

Examples

```
data(Agilent_MSDataCreation)
```

Agilent_quantF_MSclust

Demonstration dataset for MS.clust

Description

This demonstration dataset includes 2 GC-MS analyses of Lavandula, representing a total of 54 chromatogram's peaks. The file was created with MS.DataCreation (option quant=FALSE) from Agilent data. It can be used with the function MS.clust:

(i) to identify the optimal number of clusters.

(ii) to obtain the fingerprinting matrix (absence or presence of peaks for all samples)

Usage

```
data(Agilent_quantF_MSclust)
```

Format

A data frame with 54 chromatogram's peaks from 2 GC-MS analyses.

- `header line` the first row contains columns' names
- `first column` name of the sample/analysis
- `second column` retention time of the peak
- `following columns` mean relative mass spectrum of the peak (the intensity of one mass fragment (m/z) per column; Mean mass spectrum calculated by averaging 5 percent of the mass spectra surrounding the apex; The intensity of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum)

Examples

```
data(Agilent_quantF_MSclust)
```

Agilent_quantT_MSclust *Demonstration dataset for MS.clust*

Description

This demonstration dataset includes 2 GC-MS analyses of Lavandula, representing a total of 54 chromatogram's peaks. The file was created with MS.DataCreation (option quant=TRUE) from Agilent data. It can be used with the function MS.clust:

(i) to identify the optimal number of clusters.

(ii) to obtain two profiling matrices, one with the corrected peak area and one with the percent of the total corrected area

Usage

```
data(Agilent_quantT_MSclust)
```

Format

A data frame with 54 chromatogram's peaks from 2 GC-MS analyses.

- `header line` the first row contains columns' names
- `first column` name of the sample/analysis
- `second column` retention time of the peak
- `third column` corrected peak area (corrArea)
- `fourth column` percent of the total corrected area (PercTotal)
- `following columns` mean relative mass spectrum of the peak (the intensity of one mass fragment (m/z) per column; Mean mass spectrum calculated by averaging 5 percent of the mass spectra surrounding the apex; The intensity of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum)

Examples

```
data(Agilent_quantT_MSclust)
```

ASCII_MSclust *Demonstration dataset for MS.clust*

Description

This demonstration dataset includes 2 GC-MS analyses of Petrel, representing a total of 67 chromatogram's peaks. It can be used with the function MS.clust:

(i) to identify the optimal number of clusters.

(ii) to obtain the fingerprinting matrix (absence or presence of peaks for all samples)

Usage

```
data(ASCII_MSclust)
```

Format

A data frame with 67 chromatogram's peaks from 2 GC-MS analyses.

- `header line` the first row contains columns' names
- `first column` name of the sample/analysis
- `second column` retention time of the peak
- `following columns` mean relative mass spectrum of the peak (the intensity of one mass fragment (m/z) per column; Mean mass spectrum calculated by averaging 5 percent of the mass spectra surrounding the apex; The intensity of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum)

Examples

```
data(ASCII_MSclust)
```

ASCII_MSDataCreation

Demonstration folder for MS.DataCreation

Description

This demonstration folder includes 2 transformed GC-MS analyses of Petrel obtained from trans.ASCII. The two analyses represent a total of 67 chromatogram's peaks. The folder can be used with the function MS.DataCreation that collects and assembles the information from chromatograms and mass spectra of the two samples in a initial data matrix with peaks in row and mass spectrum in columns.

Usage

```
ASCII_MSDataCreation
```

Format

A folder with two different transformed ascii files, each corresponding to one GC-MS analysis.

Examples

```
data(ASCII_MSDataCreation)
```

ASCII_TransASCII *Demonstration folder for trans.ASCII*

Description

This demonstration folder includes 2 raw GC-MS analyses of Petrel in ASCII format. The data in ASCII format have to be transformed with the function trans.ASCII for further analyses with MS.DataCreation. The folder can be used with the function trans.ASCII to transform the raw ascii GC-MS data in the format suitable for MS.DataCreation.

Usage

```
ASCII_TransASCII
```

Format

A folder with two different raw ascii files corresponding to the two different GC-MS analyses.

Examples

```
data(ASCII_TransASCII)
```

Data_testclust *Demonstration dataset for MS.test.clust*

Description

To test for the best unsupervised clustering method, a dataset where molecules are already identified is created. Each molecule is represented by several samples' mass spectra. Here, the dataset contains 10 molecules obtained in different samples (84 Lavandula GC-MS analyses). In the function MS.test.clust, different clustering methods are tested for their abilities to find the correct structure of the dataset. Three different cluster validity indices are calculated to evaluate the results: the matching coefficient, the silhouette width and the Dunn index (see MS.test.clust for details)

Usage

```
data(Data_testclust)
```

Format

A data frame with 10 molecules from 84 GC-MS analyses.

- `header line` the first row must contains the columns' names
- `first column` name of the molecule
- `second column` sample name
- `third column` retention time
- `following columns` mean relative mass spectrum of the molecule (the intensity of one mass fragment (m/z) per column; Mean mass spectrum calculated by averaging 5 percent of the mass spectra surrounding the apex; The intensity of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum)

Examples

```
data(Data_testclust)
```

`MS.clust` *Mass spectra clustering and creation of a fingerprinting or profiling matrix*

Description

MS.clust runs unsupervised clustering methods on mass spectra. It can identify the optimal number of clusters using a cluster validity index (silhouette width), produces different files for facilitating the identification of molecules within a complex dataset of numerous mass spectra, and returns a fingerprinting or profiling matrix for homogeneous clusters (see details below for the definition of homogeneous clusters).

Usage

```
MS.clust(data_tot, quant=FALSE, clV, ncmin, ncmax, Nbc, varRT = 0.1,
disMeth, linkMeth, clustMeth)
```

Arguments

`data_tot`	R object data frame as returned by *MS.DataCreation* (`initial_DATA.txt`), or a *user made file* (.txt, .csv...) with the first row containing columns' names; first column contains sample/analysis name; second column contains retention time of the peak (or retention index); optionally third and fourth columns may contain respectively corrected peak area and percent of the total corrected area, and following columns contains the mean relative mass spectrum (the intensity of one mass fragment (m/z) per column; each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum)
`clV`	TRUE indicates that the function clValid will be used to identify the optimal number of clusters. FALSE, when the number of clusters is already known, escapes the clValid step and goes directly to the clustering.
`ncmin`	If clV = TRUE, a numeric value giving the minimum number of clusters to be evaluated.
`ncmax`	If clV = TRUE, a numeric value giving the maximum number of clusters to be evaluated.
`Nbc`	If clV = FALSE, a numeric vector giving the number(s) of clusters to be evaluated. e.g., Nbc=c(20,25) would evaluate the number of clusters 20 and 25.
`varRT`	RT or RI's range to define homogeneous clusters, i.e. the accepted range of variation of RT/RI for a given molecule. Default value is set to 0.1. If the varRT is baseless (analyses from different GC columns for example), set the varRT to a high value (1000 for example).
`clustMeth`	A character vector giving the clustering methods. Available options are *hierarchical*, *diana*, *kmeans* and *pam*
`disMeth`	The metric used to determine the distance matrix. Possible choices are *euclidean*, *manhattan*, and *correlation*. For pam and diana, only euclidean and manhattan are available.

`linkMeth` For hierarchical clustering, the agglomeration method used. Available choices are *ward*, *single*, *complete*, *centroid* and *average*. For all others `clustMeth`, `linkMeth=NULL`

`quant` `TRUE` if columns 3 and 4 of the input file contains corrected peak area (CorrArea) and percent of the total corrected area (PercTot). CorrArea is used for absolute quantification when associated with the use of external and/or internal standards. PercTot is used for relative quantification (no external or internal standard needed). This option generates two distinct profiling matrices in outfiles, one with CorrArea and one with PercTot. `FALSE` if these two columns are absent. Then, a fingerprinting matrix (absence or presence of each molecule) is generated

Format

- `header line` the first row must contains columns' names
- `first column` name of the sample/analysis
- `second column` retention time (RT) of the peak (or retention index (RI))
- *optionally* `third column` corrected peak area
- *optionally* `fourth column` percent of the total corrected area
- `following columns` relative mass spectrum of the peak (mass spectrum at the apex or obtained by averaging 5 percent of the mass spectra surrounding the apex; Each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum); The intensity of one mass fragment (m/z) per column

Details

MS.clust runs several unsupervised clustering methods on a dataset composed of numerous mass spectra from different samples/analyses. When the total number of molecules in the dataset is unknown, MS.clust can first identify the optimal number of clusters with a cluster validity index (silhouette width) after running the clustering on a range of numbers of clusters (clValid procedure, `clV=TRUE`).

A graphic window displays the mean silhouette width as a function of the number of clusters. A red line indicates the optimal number of clusters with the highest silhouette width. The values of silhouette width for the different numbers of clusters are summarized in a table named *res_clValid.txt* and saved in a folder called *Output_resultdate_time*.

Following the graphic display, the user is asked *How many clustering separations ?*, i.e. how many times should the dataset be cut into clusters. The answer is an integer. If the graph indicates a unique and clear optimal at the apex of a curve, only one cut at the optimal number of clusters is expected. If the graph display an optimal value located on a plateau, the user might be interested to perform different cut, one at the optimal number of clusters together with one at the minimum and one at the maximum numbers of clusters delimiting the plateau.

Afterward, the user is asked *How many clusters?* The answer is an integer. If several values, each integer should be entered and followed by Enter key. When the number(s) of clusters to be analyzed is defined, the clustering is performed. The arguments of clustMeth includes hierarchical, diana, kmeans and pam. For disMeth and linkMeth, the arguments are similar to those of the clValid package. See arguments below and the documentation of this package for more details. Following the clustering, the function identifies homogeneous and inhomogeneous clusters. The homogeneous clusters are defined by a variation in retention time lower than varRT (0.1 by default). Homogeneous clusters may correspond to a well-defined molecule, with clear mass spectra. Inhomogeneous clusters usually need manual investigations to be further classified as molecules.

Value

MS.clust produces different files in folder *Output_MSclust_resultdate_time* for facilitating the identification of molecules within a dataset composed of numerous mass spectra:

`Output_cluster.txt`
: in column, the number of the cluster, quality of the cluster based on the variation of retention time (0 if inhomogeneous, 1 if homogeneous), number of distinct individuals within the cluster and total number of peaks in the cluster (to check for unique occurence of each given analysis in the cluster), mean retention time (RT), range of retention time (max(RT)-min(RT)), mean silhouette width. Follow the 8 highest mass fragments (m/z) and the complete mean relative mass spectrum.

`Output_peaks.txt`
: in column, the number of the cluster, the sample name, the retention time, the silhouette width, the neighbor cluster, *optionally* if `quant=TRUE` corrArea and PercTotal, the 8 highest mass fragments and the complete mean relative mass spectrum.

`Hist_cluster_ok_RT.pdf`
: a pdf file displaying the histogram of the distribution of retention times for each homogeneous cluster.

`Hist_cluster_ok_silhouette.pdf`
: a pdf file displaying the histrogram of the distribution of silhouette width for each homogeneous cluster.

`Hist_cluster_problem_RT.pdf`
: a pdf file displaying the histrogram of the distribution of retention times for each inhomogeneous cluster.

`Hist_cluster_problem_silhouette.pdf`
: a pdf file displaying the histrogram of the distribution of silhouette width for each inhomogeneous cluster.

Depending on the quant option

`Output_fingerprintingmatrix.txt`
: a fingerprinting matrix (0 for absence, 1 for presence) with samples' names in the first column, retention time in the second column and presence or absence for homogeneous clusters in the following columns.

or

`Output_profilingmatrix_CorrArea.txt`
: a profiling matrix (0 for absence, corrected area if present) with samples' names in the first column, retention time in the second column and corrected area for homogeneous clusters in the following columns.

`Output_profilingmatrix_PercTot.txt`
: a profiling matrix (0 for absence, percent of the total corrected area if present) with samples' names in the first column, retention time in the second column and percent of the total corrected area for homogeneous clusters in the following columns.

Author(s)

Elodie Courtois, Yann Guitton, Florence Nicole

See Also

cluster, kohonen, class, mclust, amap, ClValid, fpc, flexmix

Examples

```
data(Agilent_quantT_MSclust)
MS.clust(Agilent_quantT_MSclust, quant=TRUE, clV=TRUE, ncmin=10, ncmax=50,
 varRT = 0.1, disMeth="euclidean", linkMeth="ward", clustMeth="hierarchical")
1
21
## 21 clusters have been determined as the optimal number of cluters.
##with the option quant=TRUE, generate profiling matrices in output

data(Agilent_quantF_MSclust)
MS.clust(Agilent_quantF_MSclust, quant=FALSE, clV=FALSE, Nbc=21,
varRT = 0.1, disMeth="euclidean", linkMeth="ward", clustMeth="hierarchical")
##with clV=FALSE, if you already know the number of molecules in the dataset
##with the option quant=FALSE, generate a fingerprinting matrix in output

data(ASCII_MSclust)
MS.clust(ASCII_MSclust, quant=FALSE, clV=TRUE, ncmin=10, ncmax=50,
varRT = 0.1, disMeth="euclidean", linkMeth=NULL, clustMeth="kmeans")
3
26
28
30
## output files are generated for three different numbers of clusters.
## with 3 as the number of clustering separations
## 26 # First number of clusters
## 28 # Second number of clusters
## 30 # Third number of clusters
```

MS.DataCreation *Create an initial data matrix from GC-MS analyses by collecting and assembling the information from chromatograms and mass spectra*

Description

This function constructs an initial data matrix by collecting and assembling the information from chromatograms and mass spectra from several GC-MS analyses. It performs peak detection if the input file is an ASCII. For all input files, peak retention times (or retention indices) are retrieved from the chromatograms and associated to their respective mass spectrum. Each row of the output data matrix represent one peak in one analysis and give the sample name in first column, the peak retention time (or retention index) in second column and the mass spectrum of the peak in the following columns. If the input file is in Agilent format, it is possible to add quantification information by reporting percent of the total corrected area and corrected area.

Usage

```
MS.DataCreation(path, mz, DataType, N_filt, apex, quant = FALSE)
```

Arguments

`path`	Name of the folder containing all the GC-MS analyses
`mz`	Range of mass fragments delimiting the mass spectrum, e.g. 30:250
`DataType`	Indicate the type of input files: *Agilent* when sample folders are obtained with Agilent Technologies machines (extension .D) or *ASCII* when sample folders contains files as returned by trans.ASCII
`N_filt`	When selecting *ASCII* data type, N_filt must be informed for chromatogram smoothing before peak detection. For more details about smoothing, please refer to the documentation of the function *filter* with method=*convolution*. If N_filt is lower than 3, there will be no smoothing of the profile. A high N_filt will lower the noise in the chromatogram but can result in the loss of low concentrated peaks
`apex`	`TRUE` indicates that the mass spectrum is considered at the apex of the peak and `FALSE` indicates that a mean mass spectrum is obtained by averaging 5 percent of the mass spectra surrounding the apex (apex included) for Agilent and by averaging the mass spectrum before, the mass spectrum after and the mass spectrum in the apex for ASCII files
`quant`	If DataType= *Agilent*, the option quant indicates if quantification information should be extracted from rteres.txt and added to the initial data matrix. `TRUE` indicates that the two quantification columns corr.area (corrected peak area) and

Details

After a GC-MS analysis, a folder is created and contains different files from the chromatograph and from the mass spectrometer. The input files in the sample folder can be of different origins:

(i) For Agilent Technologies providers (using the default parameters): each analysis returns a folder .D that contains a file rteres.txt with summary information of the chromatogram. A second file (with information of the mass spectra) is needed and can be generated by the user with the Chemstation dataanalysis software (Menu/Tools/Export3-D...), by default the generated file is Export3d.CSV and is placed in the .D folder.

You should then select the option DataType=*Agilent*. The function first checks if all samples folders (.D) within the folder *path* have both types of file rteres.txt and Export3d.CSV. If one file is missing, the analysis stops and indicates the name of the problematic sample. The analysis should be restarted after correction or removal. In a second time, the function collects the peak's retention time (or retention index) in rteres.txt and look for corresponding mass spectra in Export3d.CSV. Depending on the Apex option, the mean mass spectrum per each peak is calculated or the mass spectrum at the apex is extracted. The intensity, in counts, of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum. If quant = TRUE, the two quantification columns CorrArea (corrected peak area) and PercTot (percent of the total corrected area) are extracted for each peak from rteres.txt and placed respectively in columns 3 and 4 of the output data matrix.

(ii) For other providers: data should be transformed into the international ASCII format. All files (one per analysis) should be grouped in the folder *path* and then pass through the trans.ASCII function. The first step includes a smoothing of chromatogram depending on the option N_filt (see the documentation of the function *filter*, method=*convolution*). Afterward, peak are detected by the succession of 3 points with increasing intensity directly followed by three points of decreasing intensity (all points should have an intensity higher than 10 kilocounts). The first and last peaks of the chromatogram are removed if incomplete. In a third time, depending on the Apex option, the function calculates the mean mass spectrum per each peak or extracts the mass spectrum at the apex

and the intensity (in counts) of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum.

During the analysis, a temporary file called save_list_temp.rda is automatically generated in folder *path*.

The final output file called initial_DATA.txt is saved in folder *Output_MSDataCreation_resultdate_time*. The output data matrix contains the relative mass spectrum of each peak of all samples. The first column contains sample name (the name of the folder containing the GC-MS analysis), the second column is the peak retention time (or retention index) and the following columns correspond to the relative mass spectrum of the peak (within the range of the mass spectrum).

If quant = TRUE for DataType= Agilent, the first column contains sample name, the second column is the peak retention time (or retention index), the third column contains corrected area (CorrArea), the fourth column contains percent of the total corrected arera (PercTot) and the following columns correspond to the relative mass spectrum of the peak (within the range of the mass spectrum).

Value

MS.DataCreation returns a data matrix as an object in R and this data matrix, called initial_DATA.txt, is also saved in folder *Output_MSDataCreation_resultdate_time*. It contains one row per peak and per individual with the information in column of the sample name, the retention time (or retention index) and the relative mass spectrum. If quant =TRUE for DataType = Agilent, two supplementary columns corrArea and PercTot are added after the column retention time. A temporary list is generated during the process. It allows recovering temporary informations if the function stopped before ending because of errors.

Author(s)

Elodie Courtois, Yann Guitton, Florence Nicole

Examples

```
## Not run:
##not run
##For Agilent GC-MS files (rteres.txt and Export3d.CSV)
pathAgilent<-system.file("doc/Agilent_MSDataCreation",
package="MSeasy")
MS.DataCreation(path=pathAgilent,mz=30:250,DataType="Agilent",apex=FALSE)

##For ASCII GC-MS files
pathASCII<-system.file("doc/ASCII_MSDataCreation",
package="MSeasy")
MS.DataCreation(path=pathASCII,mz=30:250,DataType="ASCII",apex=TRUE, N_filt=3)

## End(Not run)
```

`MS.DataCreationCDF` *Same as 'MS.DataCreation' but with the capability to read AIA/ANDI NetCDF, mzXML, mzData and mzML files. Create an initial data matrix from GC-MS analyses by collecting and assembling the information from chromatograms and mass spectra from AIA/ANDI NetCDF, mzXML, mzData and mzML files. Warning: the use of xcms package is necessary.*

Description

This function constructs an initial data matrix by collecting and assembling the information from chromatograms and mass spectra from several GC-MS analyses. For all input files, peak retention times (or retention indices) are retrieved from the chromatograms (from rteres.txt file) and associated to their respective mass spectrum (from CDF file). Each row of the output data matrix represent one peak in one analysis and give the sample name in first column, the peak retention time (or retention index) in second column and the mass spectrum of the peak in the following columns. If the input file is in Agilent format, it is possible to add quantification information by reporting percent of the total corrected area and corrected area.

Usage

```
##xcms R package needed
##copy paste this to download xcms.  Remove the comment # signs
##source("http://bioconductor.org/biocLite.R");biocLite("xcms")

MS.DataCreationCDF(path, pathCDF="", mz, apex, quant = FALSE)
```

Arguments

`path`	Name of the folder containing all the GC-MS analyses (e.i The Agilent .D folders with at least the rteres.txt file).
`pathCDF`	Name of the folder containing all the CDF files, You can write the path (`pathCDF="c:/Myfolder/"`) or keep this value empty. By default `pathCDF=""`. If `pathCDF=""` the function require the `tcltk` R package to be installed. If codepathCDF="" an interactive window will help you browse your computer for the folder containing all the CDF files.
`mz`	Range of mass fragments delimiting the mass spectrum, e.g. 30:250
`apex`	`TRUE` indicates that the mass spectrum is considered at the apex of the peak and `FALSE` indicates that a mean mass spectrum is obtained by averaging 5 percent of the mass spectra surrounding the apex (apex included) for Agilent and by averaging the mass spectrum before, the mass spectrum after and the mass spectrum in the apex for ASCII files
`quant`	The option quant indicates if quantification information should be extracted from rteres.txt and added to the initial data matrix. `TRUE` indicates that the two quantification columns corr.area (corrected peak area) and

Details

After a GC-MS analysis with Agilent apparatus, a .D folder is created and contains different files from the chromatograph and from the mass spectrometer. The input files in the sample folder can be of different origins:

(i) For Agilent Technologies providers (using the default parameters): each analysis returns a folder .D that contains a file rteres.txt with summary information of the chromatogram. A second file (with information of the mass spectra) is needed and can be generated by the user with the Chemstation dataanalysis software (Menu/File/AIA ANDII...), by default the generated file is in *.CDF and is placed in a user defined folder.

The function first checks if all samples folders (.D) within the folder *path* have file rteres.txt . If one file is missing, the analysis stops and indicates the name of the problematic sample. The analysis should be restarted after correction or removal. In a second time, the function ask the path to the folder with AIA/ANDI NetCDF, mzXML, mzData or mzML files by a prompt window and then

collects the peak's retention time (or retention index) in rteres.txt and look for corresponding mass spectra in AIA/ANDI NetCDF, mzXML, mzData or mzML from the second directory. Depending on the Apex option, the mean mass spectrum per each peak is calculated or the mass spectrum at the apex is extracted. The intensity, in counts, of each mass fragment is transformed to a relative percentage of the highest mass fragment per spectrum. If quant = TRUE, the two quantification columns CorrArea (corrected peak area) and PercTot (percent of the total corrected area) are extracted for each peak from rteres.txt and placed respectively in columns 3 and 4 of the output data matrix.

(ii) For other providers: data should be transformed using codeMS.DataCreation function.

During the analysis, a temporary file called save_list_temp.rda is automatically generated in folder *path*.

The final output file called initial_DATA.txt is saved in folder *Output_MSDataCreation_resultdate_time*. The output data matrix contains the relative mass spectrum of each peak of all samples. The first column contains sample name (the name of the folder containing the GC-MS analysis), the second column is the peak retention time (or retention index) and the following columns correspond to the relative mass spectrum of the peak (within the range of the mass spectrum).

If quant = TRUE for DataType= Agilent, the first column contains sample name, the second column is the peak retention time (or retention index), the third column contains corrected area (CorrArea), the fourth column contains percent of the total corrected arera (PercTot) and the following columns correspond to the relative mass spectrum of the peak (within the range of the mass spectrum).

Value

MS.DataCreationCDF returns a data matrix as an object in R and this data matrix, called initial_DATA.txt, is also saved in folder *Output_MSDataCreation_resultdate_time*. It contains one row per peak and per individual with the information in column of the sample name, the retention time (or retention index) and the relative mass spectrum. If quant =TRUE for DataType = Agilent, two supplementary columns corrArea and PercTot are added after the column retention time. A temporary list is generated during the process. It allows recovering temporary informations if the function stopped before ending because of errors.

Author(s)

Elodie Courtois, Yann Guitton, Florence Nicole

Examples

```
## Not run:
##not run
##require xcms package
##For Agilent GC-MS files (You have to create 2 folders:
## One folder with all the rteres.txt placed in divers .D sub-folders
## and one folder with all CDF or XML files )
## CDF files have to be downloaded from MSeasy web site
##  http://sites.google.com/site/rpackagemseasy/downloads/ExempleCDF.zip


## url1<-"http://sites.google.com/site/rpackagemseasy/downloads/ExempleCDF.zip"
## download.file(url=url1, destfile="AgilentCDF.zip")
## unzip(zipfile="AgilentCDF.zip", exdir=".")
## a folder is created in your current working directory
## unlink("AgilentCDF.zip")  ##delete the zip files
```

```
pathAgilent<-system.file("doc/Agilent_MSDataCreation", package="MSeasy")

#with pathCDF
MS.DataCreationCDF(path=pathAgilent, pathCDF=getwd(), mz=30:250,apex=FALSE)

# without pathCDF
MS.DataCreationCDF(path=pathAgilent, mz=30:250,apex=FALSE)


## A box appears and ask for the path to the ExempleCDF folder
## downloaded and unziped from MSeasy website



## End(Not run)
```

`MS.test.clust` *Test for the best clustering method*

Description

This function tests the efficiency of several unsupervised clustering methods to group similar mass spectra from mass spectrometry (MS) data. Using a dataset where molecules are already well-identified and represented by several samples/individuals' mass-spectra, the clustering algorithms are tested for their ability to find the correct structure of the dataset (correctly assign the different mass spectra to the pre-defined number of molecules).

Usage

```
MS.test.clust(data_tot, nclust)
```

Arguments

`data_tot`	data matrix with the name of the molecule in the first column, the name of the sample in the second column, the retention time (or retention index) in the third column and the mass spectrum displayed in the other column.
`nclust`	number of molecules in the dataset

Details

This function tests the efficiency of several unsupervised clustering methods to group similar mass spectra from mass spectrometry data. Using a dataset where molecules are already well-identified and represented by several samples/individuals mass-spectra, the clustering algorithms are tested for their ability to correctly assign the different mass spectra to the pre-defined number of molecules. The clustering algorithms tested are partition around medoid (pam), hierarchical divisive clustering (Diana), hierarchical agglomerative clustering (hclust), with various combinations of distance metrics and link methods. Distance metrics include euclidean, correlation and manhattan. Link methods include single, average, complete, centroid and ward.

The results of clustering algorithms are evaluated with three quality indices that assess which clustering scheme best fits the data. The matching coefficient computes for correct assignment of each

mass spectrum to the expected molecules. When one cluster groups the mass spectra corresponding to the same molecule, then 1 is attributed and when one cluster contains mass spectra of different molecules, then 0 is attributed. The sum is then divided by the total expected number of molecules/clusters. The value of the matching coefficient varies from 0 to 1 and 1 indicates perfect clustering. Matching coefficient= Number of clusters grouping mass spectra of the same molecule divided by the total number of clusters.

The second cluster validity index is called silhouette width and described by Rosseeuw (Rousseeuw, 1987). This index is based on two criteria: cluster compactness and isolation.

Silhouette width s(i) is defined as: s(i)=(b-a)/max(a,b)

where a is the average distance of a point from the other points of the same cluster (variation intracluster / compactness) and b represents the minimum of the average distances of the point from the points of the other clusters (cluster separation)

Another quality index, the Dunn index D, is defined as:

D=[mink,l-numbers of clustersdist(Ck, Cl)]/[maxm-cluster numberdiam(Cm)]

k,l,m - numbers of clusters which come from the same partitioning, dist(Ck,Cl) - inter cluster distance between clusters Ck and Cl, diam(Cm) - intra cluster diameter computed for cluster Cm.

Value

This function will return three matrices with the distance metric in column and the clustering algorithms in row.

`Dunn.test` display the Dunn index

`silhouette.test`
display the silhouette index

`matching.coef`
display the matching coefficient

This function produces a pdf file *Graph_MStestClust.pdf* displaying graphics in the folder *Output_Date_time* to help identifying the best clustering method.

Author(s)

Elodie Courtois, Yann Guitton, Florence Nicole

Examples

```
## Not run:
data(Data_testclust)
MS.test.clust(Data_testclust,10)

## End(Not run)
```

`trans.ASCII` *Transform GC-MS data in ASCII format to suitable data matrix for MS.DataCreation*

Description

This function transform each ASCII file (i.e. each GC-MS analysis in ASCII format) into a new file with a format suitable for MS.DataCreation.

Usage

```
trans.ASCII(path, mz)
```

Arguments

`path`	Name of the folder containing all the GC-MS analyses in ASCII format
`mz`	Range of mass fragments delimiting the mass spectrum (each mass fragment is characterized by its mass-to-charge ratio m/z)

Details

When coming from other providers than Agilent, data should be transformed into the international ASCII format (.txt) with the software GCMS file translator pro or any other softwares. The data in ASCII format have to be transformed with the function trans.ASCII for further analyses with MS.DataCreation.

Value

trans.ASCII creates a folder named *output_transASCII_Date_Hour* which contains the same number of files than `path`.

Author(s)

Elodie Courtois, Yann Guitton, Florence Nicole

Examples

```
## Not run:
##not run
##For ASCII GC-MS files
path<-system.file("doc/ASCII_TransASCII",package="MSeasy")
trans.ASCII(path=path,mz=30:250)

## End(Not run)
```

Index