BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2278–2280
doi:10.1093/bioinformatics/bts427

Systems biology

Advance Access publication July 10, 2012

MSeasy: unsupervised and untargeted GC-MS data processing

Florence Nicolè[1,*], Yann Guitton[2], Elodie A. Courtois[3], Sandrine Moja[1], Laurent Legendre[1,5] and Martine Hossaert-McKey[4]

[1]Université de Lyon, F-69003, Lyon; Université de Saint-Etienne, F-42000, Saint-Etienne; Laboratoire BVpam, EA 3061; 23 rue du Dr Michelon, F-42000, Saint-Etienne, [2]INRA USTL UMR SADV 1281 Stress Abiotiques et Différenciation des Végétaux Cultivés, Université Lille Nord de France, Lille 1, SN2, F-59655 Villeneuve D'Ascq, [3]Station d'écologie expérimentale du CNRS à Moulis, 2 route du CNRS, 09200 Moulis, [4]Centre d'Ecologie Fonctionnelle et Evolutive UMR 5175, 1919, route de Mende - 34293 Montpellier cedex 5, France and [5]Université de Lyon, F-69622, Lyon, France; Université Lyon 1, Villeurbanne, France; CNRS, UMR5557, Ecologie Microbienne, Villeurbanne, France

Associate Editor: Trey Ideker

ABSTRACT

Summary: `MSeasy` performs unsupervised data mining on gas chromatography–mass spectrometry data. It detects putative compounds within complex metabolic mixtures through the clustering of mass spectra. Retention times or retention indices are used after clustering, together with other validation criteria, for quality control of putative compounds. The package generates a fingerprinting or profiling matrix compatible with NIST mass spectral search program and ARISTO webtool (Automatic Reduction of Ion Spectra To Ontology) for molecule identification. Most commonly used file formats, NetCDF, mzXML and ASCII, are acceptable. A graphical and user-friendly interface, `MSeasyTkGUI`, is available for R novices.

Availability: `MSeasy` and `MSeasytkGUI` are implemented as R packages available at http://cran.r-project.org/web/packages/MSeasy/index.html and http://cran.r-project.org/web/packages/MSeasyTkGUI/index.html

Contact: florence.nicole@univ-st-etienne.fr

Supplementary information: Additional information, self-guided tutorials and demonstration data are available on the web site: http://sites.google.com/site/rpackagemseasy/home. Workflow of MSeasy is available in supplementary material

Received on May 3, 2012; revised on May 3, 2012; accepted on June 29, 2012

1. INTRODUCTION

Unveiling metabolic profiles of biological systems and their interface with their environment are being studied increasingly in an untargeted holistic manner. Indeed, global metabolic profiling or fingerprinting (Fiehn, 2002) provides a very powerful means of classifying, comparing and discriminating groups of samples. Gas chromatography–mass spectrometry (GC-MS) is one of the leading analytical platforms for this approach (Katajamaa and Oresic, 2007). This hyphenated mass spectrometric technique produces large multidimensional data with retention times and peak quantification from gas chromatography and mass-to-charge ratio (m/z) and fragment intensity from mass spectrometer. Handling such complex datasets requires efficient bioinformatics processing tools.

*To whom correspondence should be addressed.

In a recent review of bioinformatics tools in metabolomics, Sugimoto *et al.* (2012) have shown that the alignment (i.e. the elimination of retention time shifts between datasets) is still a challenging step associated with numerous technical difficulties. Here, we present an alternative approach for GC-MS data processing, insensitive to shift in retention time. `MSeasy` works directly on the raw mass spectra (MS) rather than on extensively corrected chromatograms. Unsupervised clustering algorithms group MS into putative compounds. The optimal number of putative compounds in the dataset is identified when the total number of molecules is unknown. `MSeasy` accelerates the data processing and helps to interpret complex GC-MS datasets by extracting human-understandable structure and supplying quality control criteria. The method was developed as an R package offering substantial flexibility and opportunities for further developments.

2. DESCRIPTION

2.1 Requirements

`MSeasy` and `MSeasyTkGUI` depend on the following R packages: `fpc`, `clValid` and `amap`. To read netCDF or mzXML files and to activate GUI, the `XCMS` and `tcltk` packages are, respectively, needed.

2.2 Workflow

GC-MS analysis generates two components: the chromatogram, where each peak corresponds to the elution of a distinct molecule, and the MS obtained by breaking each molecule into ionized fragments and represented by a histogram displaying the intensity of each fragment depending on its m/z. In Step 1 of `MSeasy` (function `MS.DataCreation`), the raw data from chromatograms and MS of all samples are collected into a global matrix called *initial_DATA.txt* with one line for each detected chromatographic peak. Each line contains the analysis name, retention time or retention index (RT/RI) and relative MS. The intensity (in counts) of each mass fragment is transformed into a relative percentage of the highest mass fragment per spectrum in order to compare among MS from different analyses with variable peak intensities.

The three input formats in `MSeasy` include the most common raw data format for hyphenated mass spectrometry methods

(netCDF, mzXML and ASCII) and can be obtained from many providers.

When `DataType=''CDF''`, netCDF or mzXML are the input formats of MS, a peak list named peaklist.txt should be added in each sample folder. When `DataType=''ASCII''`, data have to go through the `trans.ASCII` function to obtain compatible sample folders. Then the `MS.DataCreation` function smoothes the chromatograms (`N_filt` option) and detects peaks by the succession of three points of increasing intensity directly followed by three points of decreasing intensity.

When working with data from Agilent Technologies (`DataType = ''Agilent''`), the peak list is directly extracted from the rteres.txt file in the .D directory and MS are extracted from the AIA/ANDI files generated with the Chemstation software.

Depending on the `apex` option, the relative MS of a peak is obtained by either averaging several MS around the apex or directly extracting the MS at the apex. If `quant=TRUE`, one or two quantification measures of peak size are added in *initial_DATA*. This option generates one or two distinct profiling matrices after `MS.clust`. If the two quantification columns are absent, then a fingerprinting matrix (absence or presence of each putative compound) is generated. From Agilent Technologies constructor files, the two measures are corrected peak area and percentage of the total corrected area.

The user can skip Step 1 and go directly to Step 2 by entering a file corresponding to *initial_DATA.txt*.

In Step 2, `MS.clust` runs unsupervised clustering methods on MS from the *initial_DATA* matrix or equivalent.

Prior to `MS.clust`, an optional function `MS.test.clust` can be used to identify the best clustering algorithm. User should create a training dataset where molecules are already well identified and represented by several sample MS. Since the total number of true molecules is usually unknown in untargeted metabolic approaches, the use of unsupervised clustering algorithms is required. These include partitional and hierarchical algorithms with various combinations of distance metrics and link methods (Steinbach *et al*., 2004). The results of clustering algorithms are evaluated with three cluster validity indices that assess which clustering scheme best fits the data. The matching coefficient computes for correct assignment of each MS to the expected molecules. Silhouette Width (Rousseeuw, 1987) and Dunn's Index (Dunn,1974), based on cluster compactness and isolation, assess for the quality of clustering.

`MS.test.clust` was run on various datasets: lavender species on Agilent GC-MS (Guitton, 2010), tropical trees on Varian GC-MS (Courtois, *et al*., 2009), Petrel birds on Varian GC-MS (Mardon, 2010) and Mandrills on Shimadzu GC-MS (Charpentier, pers. com.). Perfect clustering and best performances were always obtained with the hierarchical agglomerative clustering with Euclidean distance and Ward link. This method is recommended by default in `MS.clust`.

Since the best clustering method is established, the function `MS.clust` can be used. When the total number of molecules in the dataset is unknown, `MS.clust` can first identify the optimal number of clusters. After running the clustering on a user-defined range of numbers of clusters (`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. The user can define one or several optimal numbers of clusters (multimodal distribution or limits of a plateau).

For each user-defined number of clusters, a set of output files is generated to control the quality and identify putative compounds: *Output_cluster.txt* and *Output_peak.txt* summarize, respectively, cluster and peak information. They provide different quality criteria for manual investigation: among others, peak and cluster Silhouette Width, the eight more abundant mass fragments per peak, peak redundancy (indicate if a cluster contains several peaks from the same chromatogram), the range of retention time and homogeneous cluster status. Homogeneous clusters are defined by a shift in RT/RI lower than `varRT` (`varRT` $= \mathrm{RT}_{max} - \mathrm{RT}_{min} <$ 0.1 by default but can be fixed to high value for omission).

For molecule identification, automatic assignment is performed if a set of commercial standards or manually assigned reference compounds are added in the *initial_DATA* matrix. In addition, `MSeasy` output files can be used directly for molecule identifications with NIST mass spectral search program (`MSeasytoNIST`, `SeachNIST`) and ARISTO webtool (`MSeasyToARISTO`).

Finally, depending on the `quant` option, `MS.clust` returns a fingerprinting matrix (0 for absence and 1 for presence) or one or two profiling matrices for homogeneous clusters. These matrices of processed data can be directly used for data analysis.

3. CONCLUSION

The novel R package `MSeasy` is a free, easy-to-use and powerful raw data processing pipeline that extracts essential information from raw GC-MS data with minimal effort. It is insensitive to shift in retention time and does not require chromatogram alignment.

`MSeasy` speeds up greatly GC-MS data processing, allows handling large amount of data and limits labor-intensive and error-prone tasks.

ACKNOWLEDGEMENTS

We thank the contributions of all beta-testers, members of the GDR d'écologie chimique and GDR BioChiMar, Frédéric Hache, Cyrille Conord, Jérôme Chave, Sylvie Baudino, Bernard Pasquier and the CNPMAI.

Funding: GDR d'écologie chimique, Region Rhône-Alpes, University Jean Monnet Saint Etienne, Region Nord-Pas-de-Calais.

Conflict of Interest: none declared.

REFERENCES

Courtois,E.A. *et al.* (2009) Diversity of the volatile organic compounds emitted by 55 species of tropical trees: a survey in French Guiana. *J. Chem. Ecol.*, **35**, 1349–1362.

Dunn,J.C. (1974) Well separated clusters and fuzzy partitions. *J. Cybern.*, **4**, 95–104.

Fiehn,O. (2002) Metabolomics–the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Guitton,Y. (2010) Diversité des composés terpéniques volatils au sein du genre Lavandula: aspects évolutifs et physiologiques. *PhD Thesis*. LBVpam, University Jean Monnet, Saint Etienne, France.

Katajamaa,M. and Oresic,M. (2007) Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A*, **1158**, 318–328.

Mardon,J. (2010) Olfaction chez les pétrels : caractérisation chimique d'une signature olfactive et implications évolutives. *PhD Thesis*. University Montpellier II, Montpellier, France, CEFE UMR 5175.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Steinbach,M. *et al.* (2004) The challenges of clustering high dimensional data. In Wille,L.T. *et al.* (ed.) *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition*. Springer-Verlag, Berlin Heidelberg, pp. 273–309.

Sugimoto,M. *et al.* (2012) Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr. Bioinform.*, **7**, 96–108.