MS data processing

GC/MS data processing

Yann GUITTON

2018/10/04

v 3.0.0
Why using a specific tool for processing GC-MS data?

- GC-MS data have their own specificity
  - First one is higher amount of ions in a mass spectra with EI compared to ESI
Processing GC-MS data in Galaxy

**Downstream tools**

<table>
<thead>
<tr>
<th>Name</th>
<th>Output file</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine Vdk or Lowess</td>
<td>dataMatrix.tsv</td>
<td>Tabular</td>
</tr>
<tr>
<td>Normalization Vdk/Lowess</td>
<td>dataMatrix.tsv</td>
<td>Tabular</td>
</tr>
<tr>
<td>Anova</td>
<td>dataMatrix.tsv</td>
<td>Tabular</td>
</tr>
<tr>
<td>PCA</td>
<td>dataMatrix.tsv</td>
<td>Tabular</td>
</tr>
<tr>
<td>Hierarchical Clustering</td>
<td>dataMatrix.tsv</td>
<td>Tabular</td>
</tr>
<tr>
<td>Golm Metabolome Search</td>
<td>peakspectra.msp</td>
<td>Text</td>
</tr>
</tbody>
</table>

General schema of the metabolomic workflow for GCMS

**Workflow Position:**
- **Current tool**
- **Downstream tools**

**Preprocessing**
- metams.runGC: GC-MS data preprocessing using metams package

**Galaxy library**
- clean.png: Search GCMS spectra from msp files into Golm Metabolome database

**Statistical Analysis**
- ACP
- Others...

**Annotation**
- Golm Metabolome
  - Or NIST on your local computer
Processing GC-MS data in Galaxy

- **In fine** that tool use `xcms/CAMERA` under `metaMS` R package from R. Wehrens
  
  `metaMS` : Wehrens, R.; Weingart, G.; Mattivi, F. Journal of Chromatography B.

- with adapted parameters and dedicated outputs
Processing GC-MS data in Galaxy

Peakpicking/grouping quality control
**Processing GC-MS data in Galaxy**

A MSP export of pseudospectra readable by NIST or Golm Metabolome

http://web11.sb-roscoff.fr/download/w4m/howto/w4m_HowToUseNIST_V01.pdf

### Table Example

<table>
<thead>
<tr>
<th>No. Spectra</th>
<th>Spectrum</th>
<th>Analyte</th>
<th>Retention Index</th>
<th>Retention Discrepancy</th>
<th>Dot product</th>
<th>Euclidean</th>
<th>Jaccard</th>
<th>Hamming</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Sorbitol (6TMS) [A153001-ambient-O-21]</td>
<td>Sorbitol (6TMS)</td>
<td>1919.73633</td>
<td>419.736328</td>
<td>0.007266</td>
<td>0.004080</td>
<td>0.561433</td>
<td>329</td>
</tr>
<tr>
<td>2</td>
<td>Mannitol (6TMS) [A163002-ambient-na-16]</td>
<td>Mannitol (6TMS)</td>
<td>1913.16931</td>
<td>413.1693</td>
<td>0.009170</td>
<td>0.008272</td>
<td>0.533582</td>
<td>143</td>
</tr>
<tr>
<td>2</td>
<td>Galactose (1MFOX) (6TMS) BP [A191002-ambient-DL-10]</td>
<td>Galactose (1MFOX) (6TMS) BP</td>
<td>1902.42212</td>
<td>402.422119</td>
<td>0.011220</td>
<td>0.009001</td>
<td>0.606560</td>
<td>183</td>
</tr>
<tr>
<td>2</td>
<td>Galactose (1MFOX) (5TMS) MP [A188001-ambient-DL-7]</td>
<td>Galactose (1MFOX) (5TMS) MP</td>
<td>1876.05653</td>
<td>376.056528</td>
<td>0.014903</td>
<td>0.009820</td>
<td>0.495342</td>
<td>149</td>
</tr>
<tr>
<td>2</td>
<td>Glucose (1MFOX) (6TMS) BP [A191001-ambient-na-26]</td>
<td>Glucose (1MFOX) (6TMS) BP</td>
<td>1899.05493</td>
<td>399.054962</td>
<td>0.023291</td>
<td>0.009331</td>
<td>0.540187</td>
<td>289</td>
</tr>
<tr>
<td>2</td>
<td>Ribitol (6TMS) [A173001-ambient-na-7]</td>
<td>Ribitol (6TMS)</td>
<td>1712.74365</td>
<td>212.743668</td>
<td>0.025201</td>
<td>0.013441</td>
<td>0.473118</td>
<td>132</td>
</tr>
</tbody>
</table>
Processing GC-MS data in Galaxy

- Then go to other databases with the .MSP file
- Or go to statistical analysis with

<table>
<thead>
<tr>
<th></th>
<th>Metadata Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>sampleMetadata.tsv</td>
</tr>
<tr>
<td>4</td>
<td>variableMetadata.tsv</td>
</tr>
<tr>
<td>5</td>
<td>dataMatrix.tsv</td>
</tr>
</tbody>
</table>
Processing GC-MS data in Galaxy

HOW TO
How to: Processing GC-MS data in Galaxy

Workflows for metabolomics

Workflow Position:
- Current tool
- Downstream tools

Downstream tools:
- Name          | Output file   | Format
- Determine Vdk Lowess | dataMatrix.tsv | Tabular
- Normalization Vdk/Lowess | dataMatrix.tsv | Tabular
- Anova          | dataMatrix.tsv | Tabular
- PCA            | dataMatrix.tsv | Tabular
- Hierarchical Clustering | dataMatrix.tsv | Tabular
- Golm Metabolome Search | peakspectra.msp | Text

General schema of the metabolomic workflow for GCMS

Preprocessing
- GC-MS data preprocessing using metaMS package

Normalisation
- Quality Control
- Statistical Analysis
- Annotation

NMR
- Preprocessing
- Normalisation
- Quality Control
- Statistical Analysis

LC-MS
- Format Conversion
- Preprocessing
- Normalisation
- Quality Control
- Statistical Analysis
- Annotation

Upload File from your computer
Export Data

Workflow library
- xcmsSet
- runGC
- Golm Metabolome
- OCP
- ACP
- Statistical Analysis
- Annotation

Preprocessing menu:
- Metabolomics runGC GC-MS data preprocessing using metaMS package

Integrated search tool:
- dbsearch.golm Search GCMS spectra from msp files into Golm Metabolome database

Statistical Analysis
- Or NIST on your local computer
How to : Processing GC-MS data in Galaxy

- Can be tested with GCMS_Idealg_FWS_SWS.zip

Input your data as zip, or as library (data should be in CDF, mzXML or mzData) and organised in sub folders

- GC default settings, give it a try
- RT range: cut a part of your chromatogramme
- Can import a DB in MSP format
- Calculate Ris need a file with alcane RT

Only the five first EIC by default use 0 to draw all

Author(s) Ron Wehrens (ron.wehrens@gmail.com), Georg Weingart, Fulvio Mattivi
Galaxy wrapper and scripts developers Guittion Yann IDEALG Project, CNRS-IRISA/LINA, Rennes, France, yann.guittion@irisa.fr

Please cites
How to : Processing GC-MS data in Galaxy

In fact all are xcms/CAMERA parameters
How to: Processing GC-MS data in Galaxy

Galaxy / 4 / Metabolomics

Would you use a file

YES

If 'NO' is selected then one or more mass(es) must be entered manually.

File of masses (format: msp)

6: peakspectra.msp

.msp output file from metaMS.runGC function, or any msp file.

Column

- VAR5
- MDN35

5%-phenyl-95%-dimethylpolysiloxane (VAR5), 35%-phenyl-65%-dimethylpolysiloxane (MDN35).

Alkane Retention Index

1899

If neither an alkane RIs for VAR5 nor MDN35 is available in your setup, please select 'none' in the input field above.

Retention Index Window

5

This value is for the library search used only. A larger window size will increase the number of matches. At the same time the identification becomes less reliable due to false matching spectra without RI consensus.
How to: Processing GC-MS data in Galaxy

A MSP export of pseudospectrurus readable by NIST or Golm Metabolome

http://web11.sb-roscoff.fr/download/w4m/howto/w4m_HowToUseNIST_V01.pdf
How to: Processing GC-MS data in Galaxy

Use Retention Indices options

Your RI.csv file

RI shift between DATABASE std.RI
And also
RI shift between samples

Excel sheet save as csv
How to: Processing GC-MS data in Galaxy

Annotate MSP file to create Database

For retention time use: rt

For retention Index use: std.RI
How to: A good start with shared histories

Referenced W4M histories

<table>
<thead>
<tr>
<th>WOI</th>
<th>Name &amp; DOI</th>
<th>Technology</th>
<th>Species</th>
<th>Matrice</th>
<th>Factor</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>W4M00001</td>
<td>&quot;Sacurine-statistics&quot; 10.15454/1.4611121736910142E12</td>
<td>LC-MS</td>
<td><em>H. sapiens</em></td>
<td>Urine</td>
<td>age, BMI, gender</td>
<td>184</td>
</tr>
<tr>
<td>W4M00002</td>
<td>&quot;Sacurine-comprehensive&quot; 10.15454/1.461114233733302E12</td>
<td>LC-MS</td>
<td><em>H. sapiens</em></td>
<td>Urine</td>
<td>age, BMI, gender</td>
<td>184</td>
</tr>
<tr>
<td>W4M00003</td>
<td>&quot;Diaplasma&quot; 10.15454/1.4611165052113186E12</td>
<td>LC-MS</td>
<td><em>H. sapiens</em></td>
<td>Plasma</td>
<td>diabetic type 2</td>
<td>60</td>
</tr>
<tr>
<td>W4M00004</td>
<td>&quot;GCMS Algae&quot; 10.15454/1.4611272313071519E12</td>
<td>GC-MS</td>
<td><em>E. siliculosus</em></td>
<td>Algae</td>
<td>Salinity</td>
<td>12</td>
</tr>
<tr>
<td>W4M00005</td>
<td>&quot;Ractopamine&quot; 10.15454/1.4611287270056958E12</td>
<td>LC-MS</td>
<td><em>S. scrofa</em></td>
<td>Serum</td>
<td>Ractopamine</td>
<td>124</td>
</tr>
<tr>
<td>W4M00006</td>
<td>&quot;BPA M.Musculus&quot; 10.15454/1.4621555812795176E12</td>
<td>NMR</td>
<td><em>M. musculus</em></td>
<td>Brain</td>
<td>BPA</td>
<td>24</td>
</tr>
<tr>
<td>W4M00007</td>
<td>&quot;Coffea leaves&quot; 10.15454/1.4985472277774025E12</td>
<td>LCMS</td>
<td>Coffea sp.</td>
<td>Leaves</td>
<td>N/A</td>
<td>169</td>
</tr>
</tbody>
</table>

Galaxy / 4 / Metabolomics

Published Histories | yquitton | W4M000004 GCMS_Algae

**W4M00004 GCMS_Algae**

Study: Characterization of the physiological variations of the metabolome in algae exposed to 3 different abiotic stress (salt concentration). @w4m

Dataset: The dataset contains 12 mzXML files created from Agilent RAW GC-MS file. 4 replicates per conditions:

i) Condition 01: Fresh water strain in 1.6 ppt NaCl;

ii) Condition 02: Fresh water strain in 32 ppt NaCl;

iii) Condition 03: Sea water strain in 32 ppt NaCl

Workflow: The workflow consists of the following steps: preprocessing with metaMS (runGC function) then Golm Metabolom Databases online search for annotation, locally against NIST database with MSsearch software (see "how to" section on W4M portal).

330.82 MB
Thanks!