RMassBank: Automatic Recalibration and Processing of Tandem HR-MS Spectra for MassBank

Eawag: Swiss Federal Institute of Aquatic Science and Technology

Michael Stravs, Emma Schymanski, Heinz Singer and Juliane Hollender

Eawag – Department of Environmental Chemistry
Dübendorf (Zurich), Switzerland

RMassBank Questions: massbank@eawag.ch
Overview

- Building a spectral library: challenges
  - clean-up
  - annotation
- RMassBank workflow
  - data processing
  - record creation
- Results and examples
## Building a spectral library - Challenges

### Processing spectra efficiently
- Manual entry
  - Tedious
  - Available tools not suited for mass processing
- Metadata collection / annotation
- Reproducible procedure

### Ensuring high quality
- Noise removal / clean-up
- Mass accuracy
- Curation
Building a spectral library

Previous approaches to «clean-up» and annotation*
- Simple noise cutoff – e.g. WA001201
  - Cut-off at 5 ‰ (MassBank reporting can go down to 1 ‰)
- Minimal or no processing – e.g. CE000143
  - Many low intensity noise peaks, including peaks above [M+H]^+
- Minimal annotation – e.g. JEL00007

The result
- Inconsistent and varying quality of spectra in MassBank
  → “spectral dump” **

* These examples are demonstrations only and no offense intended! 
** Thanks to Oliver Fiehn for this apt description!
Spectral Clean-up

MassBank Record: WA001201

MS$DATA_PROCESSING: FIND_PEA$k ignore rel.int. < 5

PK$NUM_PEA$k: 19
PK$PEAK: m/z int. rel.int.
102 63 63
105 8 8
130 133 133
131 8 8
139 20 20
162 12 12
163 12 12
166 834 834
167 59 59
178 12 12
180 63 63
182 24 24
184 999 999
185 63 63

* These examples are demonstrations only and no offense intended!
Spectral Clean-up

MassBank Record: CE000143

MS$FOCUSED_ION: PRECURSOR_M/Z 268.10404
MS$FOCUSED_ION: PRECURSOR_TYPE [M+H]+

PK$NUM_PEAK: 20
PK$PEAK: m/z int. rel.int.

<table>
<thead>
<tr>
<th>m/z</th>
<th>int</th>
<th>rel.int</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.423012</td>
<td>1164.474609</td>
<td>2</td>
</tr>
<tr>
<td>79.144173</td>
<td>972.286438</td>
<td>2</td>
</tr>
<tr>
<td>88.860359</td>
<td>1310.202271</td>
<td>2</td>
</tr>
<tr>
<td>106.183556</td>
<td>1195.565674</td>
<td>2</td>
</tr>
<tr>
<td>135.887161</td>
<td>3993.299072</td>
<td>7</td>
</tr>
<tr>
<td>135.992874</td>
<td>3979.442627</td>
<td>7</td>
</tr>
<tr>
<td>136.034454</td>
<td>2340.972656</td>
<td>4</td>
</tr>
<tr>
<td>136.041092</td>
<td>1992.728271</td>
<td>4</td>
</tr>
<tr>
<td>136.061371</td>
<td>568236.125</td>
<td>999</td>
</tr>
<tr>
<td>136.08461</td>
<td>4257.030762</td>
<td>7</td>
</tr>
<tr>
<td>136.130478</td>
<td>2980.912598</td>
<td>5</td>
</tr>
<tr>
<td>136.228577</td>
<td>1192.598145</td>
<td>2</td>
</tr>
<tr>
<td>136.234634</td>
<td>1166.915161</td>
<td>2</td>
</tr>
<tr>
<td>197.613419</td>
<td>1193.385498</td>
<td>2</td>
</tr>
<tr>
<td>203.694138</td>
<td>1189.134888</td>
<td>2</td>
</tr>
<tr>
<td>219.079987</td>
<td>21971.183594</td>
<td>39</td>
</tr>
<tr>
<td>237.09053</td>
<td>180033.921875</td>
<td>317</td>
</tr>
<tr>
<td>292.21759</td>
<td>1154.94812</td>
<td>2</td>
</tr>
<tr>
<td>293.092346</td>
<td>1334.223755</td>
<td>2</td>
</tr>
<tr>
<td>293.587189</td>
<td>1281.828003</td>
<td>2</td>
</tr>
</tbody>
</table>

These examples are demonstrations only and no offense intended!
Spectral Clean-up – Orbitrap spectra

Garbage in = garbage out!

Measurement artefacts include

○ Systematic increase in ppm error with low $m/z$ values
  ○ 5 ppm accuracy $=>$ 15 ppm at $m/z < 100$
○ Satellite or shoulder peaks – result of FT instrument processing
○ Consistent electronic or measurement noise peaks
○ Noise peaks at a ~fixed level hiding real peaks of lower intensity
  ○ A strict noise cut-off will result in a loss of information

High mass accuracy standard spectra have advantages:

○ Meringer et al. (2011) showed a missing subformula indicates instrument noise or interfering peaks
  ○ Use subformula assignment to perform spectral clean-up!

Annotation

Mandatory
- Name, structure, InChI, SMILES
- Minimal analytical information
  - MS instrument
  - MS polarity and spectrum type

Optional and useful!
- Links to databases
  - CAS, KEGG, ChEBI, PubChem...
- Analytical conditions
  - Chromatography, RT [...]
  - MS2: precursor, conditions [...]
- Peak annotation
  - chemical formula, substructure

more is better, but manual work is tedious
Building a spectral library - RMassBank

- Automate what can be automated
- Ensure high quality
- Workflow from raw LC-MS file to annotated spectrum
  - spectrum extraction
  - processing and clean-up
  - automated annotation using internet resources
RMassBank – Example Clean-up

C\text{\textsubscript{10}}H\text{\textsubscript{13}}O\text{\textsubscript{2}}\textsuperscript{+}
165.0910 Da
2.1 ppm

Verapamil
C\text{\textsubscript{27}}H\text{\textsubscript{39}}N\text{\textsubscript{2}}O\text{\textsubscript{4}}\textsuperscript{+}
[M+H]\textsuperscript{+} = 455.2904 Da

if a peak doesn’t fit a subformula, it is probably noise.
Spectral Annotation
Matching Spectrum and Compound Information

User needs to contribute a bare minimum of information
- Only the user knows what compound has been measured
- At least one form of unambiguous compound identifier is required
  - e.g. internal ID, name, SMILES and retention time
- Measurement parameters / methods / settings are relatively consistent
  - These can be added in batch form, not individually

Internet Services: Let search engines do the work for you!
- CACTUS Chemical Identifier Resolver\(^1\)
  - SMILES \((c1cccccc1)\) to InChI Key \((UHOVQNZJYSORNB-UHFFFAOYSA-N)\)
- Chemical Translation Service (CTS)\(^2\) to do the rest
  - Names, CAS #, InChI and Identifiers (IDs, if available): PubChem CID, ChemSpider, ChEBI, HMDB, KEGG, LipidMaps

\(^1\) [http://cactus.nci.nih.gov/chemical/structure](http://cactus.nci.nih.gov/chemical/structure)
RMassBank – Clean-up and Recalibration

LC-MS/MS raw data

Extract MS2 spectra

Subformula assignment: large tolerance

Recalibrate raw data with assigned fragments

Subformula reassignment: 5 ppm tolerance

Multiplicity filtering

noise peaks

intense unmatched peaks

discarded noise

Filtered peaks

Automated extraction by accurate m/z and RT. FT-satellite removal (± 0.5 Da; I < 5 % of peak)

Elements of molecular formula ± adduct ppm ≤ 15 (m/z < 120) or ≤ 10 (m/z > 120)

Use only peaks with unique formula calculated

Formulas recalculated post calibration. For "Fail Peaks", recalculate with +2N & O

Peaks are retained if they occur ≥2x per compound (in e.g. 14 recorded spectra)

Filtered peaks go through to Records

Intense unmatched peaks for manual check
RMassBank – Spectrum Annotation

- compound list: SMILES, name
- online resources: CTS, CACTUS
- Chemical annotation: InChI, ChemSpiderID, PubChem CID...
- Experimental annotation: CE, ionization, RT...
- Record generation
  - Recalibrated, cleaned-up
  - Annotated
  - Ready for upload!

- manual curation
- filtered peaks
- user-defined additional peaks
- MassBank records
  - structure files
Example MassBank Record

MassBank Record: EA015612

Asulam; LC-ESI-ITFT; MS2; 75%; R=15000; [M+H]^+

ACCESSION: EA015612
RECORD_TITLE: Asulam; LC-ESI-ITFT; MS2; 75%; R=15000; [M+H]^+
DATE: 2012.03.16
AUTHORS: Strav M, Schymanski E, Singer H, Department of Environmental Chemistry, Eawag
LICENSE: http://massbank.ufz.de/MassBank/files/license.html
COPYRIGHT: Copyright (C) 2011 Eawag, Duebendorf, Switzerland
COMMENT: CONFIDENCE standard compound
COMMENT: EAWAG_UCHEM_ID 156
RMassBank with 70 Eawag Pesticide Spectra

Experimental

- Individual Injection; routine chromatography; Orbitrap XL with ESI + / -
- Daily vendor-recommended calibration

Screened for [M+H]^+ precursor within RT ± 0.3 min
MS/MS retrieved from MS with highest intensity
RMassBank with 70 Eawag Pesticide Spectra

Processing: The Numbers

- 68 of 70 pesticides with sufficient [M+H]+ for processing
- 55,594 peaks present following satellite removal
  - 14,699 with at least one subformula post-calibration
  - 13,305 of these present in two or more spectra per compound
- 454 peaks with subformula only when adding 2N + O
  - 256 of these occurred at least twice per compound => reproducible
- Only 44 peaks remained for “manual inspection”
- No difference observed between spectra with different resolutions

~76 % of peaks are noise!

Additional modes (results not shown here)


Effect of Recalibration

- Shown in the next few slides…

Formation of N₂ and H₂O adducts is relevant in MS/MS!
RMassBank with 70 Eawag Pesticide Spectra

Recalibration Curve: Relative mass deviation over m/z

before recalibration

recalibration by massive subformula assignment

after recalibration
Recalibration: Relative Mass Deviation Distribution

(a) before recalibration

(b) after recalibration
RMassBank with 70 Eawag Pesticide Spectra

Frequency of Occurrence of Peaks by Intensity: Multiplicity Filtering

if a peak doesn’t occur at least twice (in 2x7 spectra) it is probably noise
RMassBank with 70 Eawag Pesticide Spectra

944 MS/MS spectra from [M+H]$^+$ of 70 pesticides

- [www.massbank.jp](http://www.massbank.jp)
- [www.massbank.eu](http://www.massbank.eu)

Total Number of “RMassBank Spectra”

RMassBank records in NORMAN MassBank:

- 6,106 records (364 compounds) from Eawag Orbitrap XL
- 1,030 records (216 compounds) from UFZ Orbitrap XL
- Q-Exactive spectra (not yet on NORMAN MassBank)
RMassBank «advanced usage»

For R workflow developers (xcms, nontarget, CAMERA…)

- dd-MS2 processing
- recalibration, filtering
- Interface to CTS, CACTUS
- Chemical formula calculations (C6H5 + H2O = C6H7O)
- fragment formula assignment
- (database search)

... if you are fluent in R 😊
Conclusions: RMassBank

The RMassBank Workflow

- Reduces much manual work associated with bulk creation of many records
- Creates high quality MS/MS spectra
- Annotation with formula adds value to the spectra
- Works very well for the spectra it was developed on (Orbitrap)
- BUT: Every mass spectrometer is different:
  - Processing and measurement steps will probably need adjusting

Benefit for Contributors

- We have learnt a lot about our spectra and compounds (e.g. MS/MS adducts!)
- MassBank is being used within our department
- If you want to know more about what recalibration can do for your data:
Acknowledgements

- Coauthors: Emma Schymanski, Heinz Singer, Juliane Hollender
- R & MassBank help: Steffen Neumann, Michael Gerlich, Carsten Kuhl, (IPB)
- Discussions: Matthias Ruff, Martin Loos (Eawag); Tobias Schulze, Martin Krauss, Werner Brack (UFZ)
- Programs: Markus Meringer (MOLGEN-MS/MS), Florian Rasche (Sirius), Miguel Rojas-Cherto & Egon Willighagen (MEF)
- MassBank & Naming Rights: Prof. Takaaki Nishioka
- NORMAN Association
Any Questions?

Michael Stravs, Emma Schymanski, Heinz Singer and Juliane Hollender

massbank@eawag.ch (ES, MS)
stravsmi@eawag.ch
emma.schymanski@eawag.ch

DOI: 10.1002/jms.3131
All Details Contained Within!