In silico MS/MS fragmentation spectra for identifying chemical unknowns: applications and performance validation

Antony Williams1, Andrew McEachran2, Alex Chao1, Tom Transue3, Tommy Cathey3, and Jon Sobus1

1Ctr. for Comput. Toxi. & Exposure, ORD, U.S. EPA; 2Agilent Technologies, Santa Clara, USA. 3GDIT, RTP, USA

OBJECTIVES

• Demonstrate identification of unknown chemicals using high resolution mass spectrometry (MS) utilizing workflows with relevant data and software analysis tools [1-3]
• Examine whether the comparison of experimental MS fragmentation data with predicted fragmentation data can increase confidence in compound identification [4]
• Demonstrate whether predicted fragmentation data, coupled with relevant metadata, helps identify unknowns

APPROACH

• Use “MS-Ready” forms of structures from US-EPA CompTox Chemicals Dashboard [5] as input files: ~800,000 structures
• Use CFM-ID package (https://cfmid.wishartlab.com/) to generate mass spec. fragmentation spectra for +ve and –ve ion LCMS and EI GCMS spectra. 7 spectra per chemical.
• Combine rich Dashboard metadata with fragmentation matching of experimental spectra to rank candidate hit lists

MAIN RESULTS

• The identification of “known-unknowns” using non-targeted analysis benefits from the use of CFM-ID as an in silico fragmentation prediction tool
• Combining metadata candidate ranking of hits based on mass or formula searches gives improved results
• CFM-ID predicted spectra are available as FAIR Open Data
• Proof-of-concept web applications are in testing

IMPACT

• The free availability of the CompTox Chemicals Dashboard for the community, coupled with MS-Ready structures to generate in silico MS/MS fragmentation data, and metadata for candidate ranking, is a basis for the development of structure identification software tools at EPA
• For more information, contact: Antony Williams, williams.antony@epa.gov
In silico MS/MS fragmentation spectra for identifying chemical unknowns: applications and performance validation

MAIN RESULTS

Validation of performance of combined approach with 5 years of CASMI contest data [8]. Percentage of compounds from each dataset ranked in the top (number 1) position by in silico MS/MS match only, Data Source count (DS) only, and the combined score of in silico MS/MS data with Data Source counts.

~800,000 MS-Ready structures were used to predict fragmentation [7]. The dataset is available as a FAIR dataset for repurposing: https://doi.org/10.23645/epacomptox.7776212.v1

MS-Ready Structures [6] are the inputs to in silico fragmentation. This approach removes stereobonds, desalts and splits multicomponent chemicals but maps back to the original substances in the CompTox Chemicals Dashboard. This mapping provides association with substance
Summary

In silico MS/MS fragmentation is highly beneficial for the identification of unknowns and supporting non-targeted analysis

- Our multiple studies [1-3, 7-9] demonstrate the benefit of in silico prediction especially when coupled with metadata for candidate ranking of hits
- MS-Ready structure generation [6] is an essential step to the production of input structures for processing

References

8. McEachran et al., Revisiting Five Years of CASMI Contests with EPA Identification Tools, Metabolites 2020, 10(6), 260;

Future Plans

Following testing and performance validation the software applications described here will be released.

- Public access to the CFM-ID experimental search tool
- A new non-targeted analysis web application (NTA WebApp) reading instrument data and using both in silico fragmentation data and metadata for candidate ranking will be made available for community use [9]
- Public access to MS-Ready structure set processing

Ongoing updates to CFM-ID fragmentation predictions will be provided as FAIR data for reuse and repurposing