Improved Ranking of Putative Candidates Through a Hybrid *In Silico* / Real Fragmentation Technique

Tim Stratton[1], Michal Raab[2], Jakub Mezey[2], Ioanna Ntai[1], Ralf Tautenhahn[1], Robert Mistrik[2]
[1]Thermo Fisher Scientific, Austin TX, USA; [2]HighChem LLC, Bratislava, Slovakia

ABSTRACT

Purpose: To demonstrate an algorithmic approach to rank putative chemical database candidates for unknown compounds by utilizing reference spectral library data.

Methods: HRAM MS/MS and MS^n data on a series of compounds not present in the reference spectral library was acquired. This data was processed using an algorithm (mzLogic™) that performed, broadly, three steps consisting of chemical database search (ChemSpider™) to obtain putative structures followed by spectral library similarity search (mzCloud™) and a ranking of putative structures based on the common substructure explained by fragments observed in the reference library.

INTRODUCTION

Identification of compounds is typically the most difficult step in many fields of small molecule analysis including metabolomics and environmental research. While matching MS/MS or MS^n query data against a reference spectral library such as NIST, MassBank, or mzCloud is often one of the best ways to provide identification information, the limited size and coverage of available libraries necessitates alternative approaches. An extremely common approach is to search the unknown molecular weight or elemental composition against chemical information databases for putative hits, which still require some additional sorting or ranking to find relevant or likely candidates. We present here an approach that leverages both reference spectral library searching (in a similarity mode) with chemical database searching, merging them together to provide a data driven ranking of putative candidates. The algorithm, mzLogic, was implemented both in Thermo Scientific™ Mass Frontier™ 8.0 software and Thermo Scientific™ Compound Discoverer™ 3.0 software.

MATERIALS AND METHODS

Sample Preparation

Standards, as either single compounds or mixes of up to ten compounds, were prepared by dissolving the test material in a suitable solvent (DMSO or MeOH) to create a stock solution of between 0.1 to 0.5 mM. These stock solutions were further diluted with MeOH:water to create the final concentration for injection (50 nM).

Mass Spectrometer Acquisition Conditions

Samples were separated on a Thermo Scientific™ Hypersil GOLD™ 100 x 5 mm, 3 µm C18 column maintained at 35 ºC. Ionization was performed by electrospray ionization in positive and negative ionization mode (separate injections). The initial high resolution acquisition obtained full MS^1 data at a resolution of 60,000 (FWHM @ *m/z* 200). Data-dependent MS^2 was triggered using higher energy collisional dissociation (HCD) with a stepped collision of 40%±20% normalized collision energy at a resolution of 30,000. MS^3 fragmentation on the top 3 MS^2 ions was performed by trap collisional dissociation (CID) at 30% normalized collision energy (NCE).

Mass spectrometer: Thermo Scientific™ Orbitrap Fusion™ Tribrid™ MS
LC: Thermo Scientific™ Vanquish™ UHPLC system

Table 1. LC gradient for sample analysis

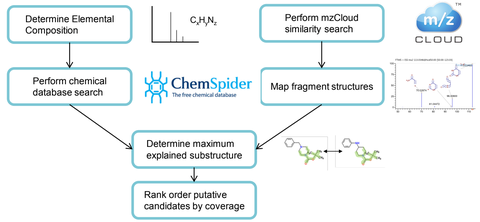
Time (min)	% A (Water + 0.1% Formic acid)	% B (ACN + 0.1% Formic acid)
0	100	0
8	50	50
9	2	98
13	2	98
13.1	100	0
15	100	0

RESULTS

The mzLogic Process Overview

The process of ranking putative candidates utilized in mzLogic is unique in that it combines streams of information previously not collectively considered. The concept of attempting to use fragmentation information to rank order database hits is not a new one. Approaches such as MAGMa[1], CDM-ID[2], and others have implemented this. These methods vary primarily in how they predict fragmentation from putative candidates *in silico* but they all do some form of prediction. For compounds where a spectral library match is not obtained, a spectral similarity search is performed to find potentially structurally related candidates. In parallel, a chemical database search (using molecular weight or elemental composition) is performed to obtain a list of putative candidates. These independent sources of data are combined by the algorithm (Figure 1). The details of each step and the critical aspects for each are discussed following.

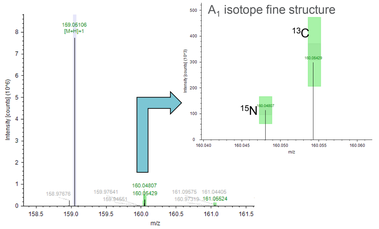
Figure 1. Workflow for mzLogic



Elemental Composition Determination and Database Searching

Although the molecular weight of the unknown can be used for chemical database searching, using the elemental composition can often give fewer putative candidates – reducing the complexity. The difficulty lies in accurate elemental compositions determination for the unknown compound. The application of very high resolution, accurate mass data allows for elemental composition determination using fine isotopic information (Figure 2), which can be combined with MS/MS fragment spectra coverage to provide a refined elemental composition calculation.

Figure 2. Fine isotopic pattern and predicted elemental compositions



Putative elemental compositions, calculated from the accurate mass MS^1 A_0 can be refined through analysis of the fine isotopes present in other isotopes. These putative elemental compositions must also be able to explain the fragment ions observed as the fragments are substructures of the precursor. Scoring the ability of putative elemental compositions to predict fragment compositions provides an enhanced composition prediction (Figure 3).

Figure 3. Improved determination of elemental composition

Formula	Molecular Weight	ΔMass [Da]	ΔMass [ppm]	RDBE	H/C	Rank ▲	MS Cov. [%]	# Matched Iso.	# Missed Iso.	MSMS Cov. [%]
C9 H13 N2 O9 P	324.03587	-0.00011	-0.35	5.0	1.4	1	100.00	4	0	99.36
C11 H8 N4 O8	324.03421	0.00154	4.75	10.0	0.7	2	100.00	4	0	97.90
C10 H9 N5 O5 P	324.03720	-0.00145	-4.48	10.0	0.9	3	100.00	4	0	99.36
C12 H4 N8 O4	324.03555	0.00020	0.63	15.0	0.3	4	100.00	4	0	97.42
C7 H16 O12 S	324.03625	-0.00049	-1.52	0.0	2.3	5	100.00	4	1	91.16
C11 H18 O5 P2 S	324.03502	0.00074	2.27	4.0	1.6	6	100.00	4	2	92.63
C13 H15 N2 O2 P3	324.03464	0.00112	3.44	9.0	1.2	7	98.00	2	1	97.63
C6 H15 N8 P3 S	324.03532	0.00043	1.33	5.0	2.5	8	98.00	2	1	0.00

Elemental compositions predicted for uridine monophosphate ($C_9H_{13}N_2O_9P$) when using a large elemental composition prediction set. Mass accuracy, isotopic pattern, and MS/MS coverage all combined to provide a final ranking.

Similarity Search Results

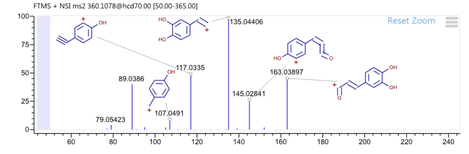
The next step in the mzLogic process is to obtain results from a similarity search in mzCloud. Query spectra for the unknown are searched against the library without constraining for the precursor mass or limiting it to MS/MS (Figure 4). Matches may come from anywhere in the MS^n tree of reference compounds and represent potential structure overlap between the similarity hit and the unknown compound. In addition, queries are run for similarity both forward and reverse to determine the most representative similarity candidate to use for subsequent substructure matching.

Figure 4. Similarity Search Result Display – Mass Frontier 8.0 software

Combining Data Sources – Deriving the Final Ranking

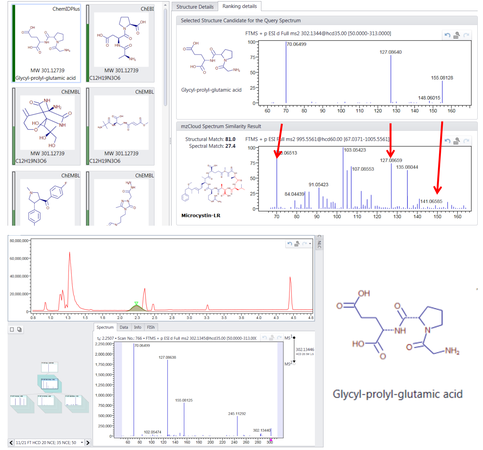
With the results from both searches, the algorithm determines the coverage – the maximum substructure – between each chemical database hit and the relevant similarity hit from the mzCloud spectral library search. This alone would not be sufficient, as the overlapping substructure may not be in common with observed fragmentation. Because of this, the implementation also considers the known fragmentation structure from the mzCloud library, which has extensive annotation of fragments (Figure 5) as a confirmation that the overlapping substructure observed in the database candidate does come from the real world fragment in the library.

Figure 5. Fragmentation annotation available on mzCloud



The approach presented can help to separate a large set of putative structures easily based on the comparison against real world representative fragmentation data. A first example (Figure 6) shows the ranking of all chemical database hits (total of 36) for the molecular weight 302.1344. The query compound was glycyl-prolyl-glutamic acid, a compound that is not in the mzCloud reference library. In this case the similarity hit selected was a natural toxin, the matching explained fragments are highlighted in the image. The mzLogic algorithm screens through the putative chemical database hits looking for common substructures from the similarity hit, considering which ones also match to the known fragment structure. In this way we develop two different scores, the spectral similarity and also the structural match – how much of the compound database hit is matching to the known reference fragmentation. This is used, along with the proportion of the chemical database candidate explained, to derive a final ranking for each putative candidate. Additionally, the second candidate is a close structural analogue of the correct result, varying in the position of a methyl group. The third and fourth ranked structures are somewhat peptide-like in their structure but are sorted to a much lower overall score by the algorithm.

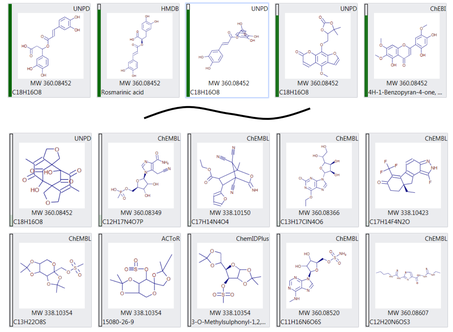
Figure 6. Example of mzLogic result for glycyl-prolyl-glutamic acid



Top: mzLogic sorted results from the processing of the acquired data from glycyl-prolyl-glutamic acid.
Bottom: Query data acquired showing the HCD MS^2 spectra and structure of the compound.

Another example of the mzLogic sorting capability can be seen with a more complex example. The data was from rosmarinic acid, another compound not present in the reference spectral library. Chemical database searching returned over 250 putative candidates. After sorting with mzLogic, the correct result was the second candidate (Figure 7), while the first and third ranked candidates were structurally similar. The range of structures, and resulting range of scores from the algorithm, are also shown with the top five candidates compared to the bottom ten. While the correct result was not the first, there is still a significant improvement in reducing the complexity of putative hits (from more than 250).

Figure 5. Base peak and XIC of potential flavonoid conjugates in orange



Results for mzLogic search of data from rosmarinic acid. Top of image shows the first five hits, the bottom are the lowest ten hits from the chemical database search after sorting by mzLogic.

CONCLUSIONS

- mzLogic combines chemical database searching for putative candidates with reference spectral library similarity searching in a new unique approach.
- Utilizing reference spectral library data removes the limitations and potential risk of purely *in silico* approaches by making use of real world observed fragmentation information.

REFERENCES

1. Ridder, L; van der Hooft, J; Verhoeven, S; de Vos, R; van Schaik, R; Vervoort, J "Substructure-based annotation of high-resolution multistage MS^n spectral trees", RCMS 2012, https://doi.org/10.1002/rcm.6364
2. Allen, F; Pon, A; Wilson, M; Greiner, R; Wishart, D. "CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra." Nucleic Acids Res. June 2014.

TRADEMARKS/LICENSING

PO65317-EN0518S