Utilization of Substructure Identification through MS^n Analysis for Unknown Structure Determination Assisted with *in silico* Fragmentation Prediction

Tim Stratton, Thermo Fisher Scientific, Austin Texas

ABSTRACT

Purpose: To demonstrate the applicability of MS^n spectral library data for the determination of previously unknown molecules through substructure identification combined with *in silico* fragmentation prediction.

Methods: HRAM MS^n fragmentation data was acquired for a set of known compounds which were not present in a reference fragmentation library. The data was searched against a reference spectral library of chemically diverse compounds containing MS/MS and MS^n data on >17,000 chemical standards. Substructure information was obtained by matching portions of the query tree to reference spectral trees and assembling the proposed substructures to putative candidates which underwent fragmentation prediction to assist in rank order determination of the best candidates.

Results: Out of twenty compounds in the study, the approach used was able to propose the correct structure for one while the other nineteen compounds had structures proposed that were similar to the correct structure with some small structure variations potentially due to chemical substructure coverage of the reference library.

INTRODUCTION

A multi-spectrum match against a high quality reference spectral library is an important tool for the identification of unknown compounds in any sample. However, given the potential chemical diversity possible in a real world sample, it is unreasonable to assume that every potential unknown exists within the currently available reference spectral libraries. While those libraries will continue to grow over time, alternative techniques to propose candidate structures for unknowns are necessary. Here we demonstrate the application of high resolution accurate mass multi-stage fragmentation data (HRAM MS^n) in the determination of substructures for unknown compounds by searching the query data against a large and diverse MS^n reference library. Substructures determines through library searching were assembled into putative candidates with the candidates being ranked for their likelihood based on *in silico* fragmentation prediction annotation of the original query MS^n data.

MATERIALS AND METHODS

Sample Preparation

Chemical standards for twenty small molecule plant and fungal metabolites were obtained (AnalytiCon Discovery GmbH, Germany) and prepared both as mixes of the pure standards and spiked into an extract of St. Johns Wart to provide a background matrix, prepared by extracting 1g of dried St. Johns Wart powder with 10mL MeOH:Water followed by centrifugation. Samples were prepared so that the final concentration for the compounds was 1uM

Mass Spectrometer Acquisition Conditions

High resolution accurate mass fragmentation data was acquired on a Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer connected to a Thermo Scientific™ Vanquish™ UHPLC system. Samples were separated after injection on a 100 X 2.1mm, 1.9 um Thermo Scientific™ Hypersil GOLD™ aQ column with a gradient elution of methanol and water with 0.1% formic acid over a fifteen minute run time at a flow rate of 0.5 mL/min (Table 1). Data dependent precursor ion selected MS^n data was acquired by triggering an fragmentation event using high energy collisional dissociation (HCD) at a 50% normalized collision energy to generate the MS^2 level spectra with up to 3 productions being subsequently serially isolated for MS^3 fragmentation using trap collisional dissociation (CID) at a 45% energy with a helium collision gas.

Table 1. LC Gradient for Sample Analysis

Time (min)	% A (Water + 0.1% Formic Acid)	% B (MeOH + 0.1% Formic Acid)
0.0	99	1
1.0	99	1
10.0	1	99
11.5	1	99
11.51	99	1
15.0	99	1

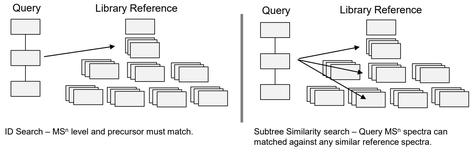
RESULTS

Search Approach

The acquisition method chosen was designed to acquire fragmentation data on substructures small in relative size to the precursor structure. The on-mass resonance excitation of trap CID leads to the creation of fewer fragment ions resulting from the most energetically favorable fragmentation mechanisms. Utilizing HCD fragmentation for the MS^2 stage provided a wider range of fragment ions, including those at a lower m/z, which were available for selection for subsequent MS^3 fragmentation. This approach gives a broad spectral tree which branches quickly resulting in access to fragmentation of small substructures (low m/z) from the unknown.

Typically, the MS^n data acquired is used to perform an identification search where the query spectra must match the library spectra both in spectral content but also in connectivity (Figure 1). In this approach, the need to match connectivity was not considered – the precursor history of the MS^n tree was not a constraint on the search and only similarity of the MS^n spectra to a library MS^n reference was used. In this way substructures can be gleaned from the library when the unknown query compound does not exist in the library.

Figure 1. Subtree Library Search Approach



ID Search – MS^n level and precursor must match.

Subtree Similarity search – Query MS^n spectra can matched against any similar reference spectra.

Data Processing Examples

The MS^n spectral trees for the unknown compounds were submitted for spectral library search in different ways. The first approach was to submit the entire MS^n spectral tree for a search using a substructure tree similarity search. The second approach was to submit individual nodes of the unknown MS^n trees for similarity search and collate the resulting library results.

Using the complete MS^n spectral tree approach, hits were sparse given that the library did not contain the test compounds. The alternative MS^n tree search approach was to perform an MS^n similarity search were the submitted unknown query tree was searched only for similarity matches. In this approach the precursor match was not required, only similarity between the query spectra and library hits. This provided information on the potential substructures as shown in Figure 2 for an unknown with m/z 305.0661.

In this example, the query spectra provided a match to a substructure observed in several library compounds. In addition, the subsequent MS^n spectra for this unknown provided additional match information against other compounds in the reference library (Figure 3) including matching deep into some reference trees. Scrutinizing these similarity matches provides information about potential substructures in the unknown compound.

Figure 2. Partial MS^n Tree Spectra Match – MS^2 Sim Match for Unknown 305.0661

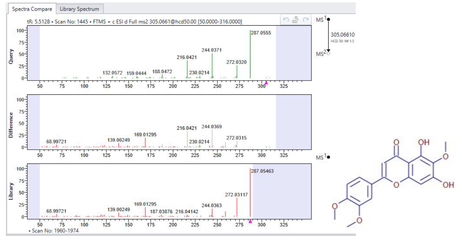
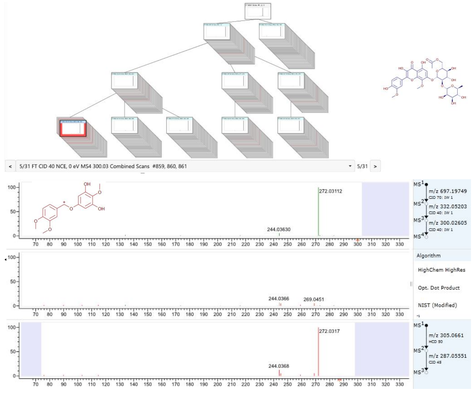


Figure 3. Partial MS^n Tree Spectra Match – MS^3 Sim Match for Unknown 305 → 287 →

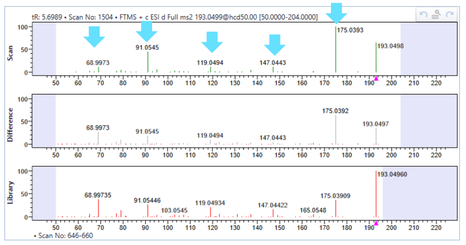


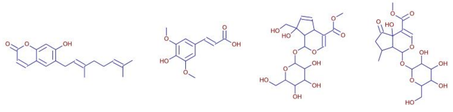
Utilizing this search approach, collections of potential substructures were created for each of the unknown compounds. These where used in combination with an understanding of the nature of the compounds (plant biochemical space) to begin to propose structures for each candidate with the proposed elemental composition acting as an upper limit and a guide on elements and atom counts. In addition to this approach, candidate substructures were also searched in chemical databases of relevant plant biochemistry to further provide candidates for each unknown.

This was a time consuming process and a good potential step for automation in the future to provide a relevant set of potential candidates based on chemical substructure to supplement the manual assembly of the proposed substructures obtained from the library search.

An example of the approach can be demonstrated with unknown 193.0499. Subtree searching resulted in multiple matches for the acquired query MS^n data which provided a number of structure candidates (Figure 4). The matching fragment ions, and the precursor fragment structures for matching MS^n spectra, formed the basis for constructing a candidate for this unknown.

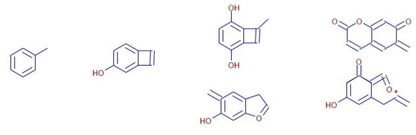
Figure 4. One of Multiple Subtree Search Similarity Matches for Unknown 193.0499





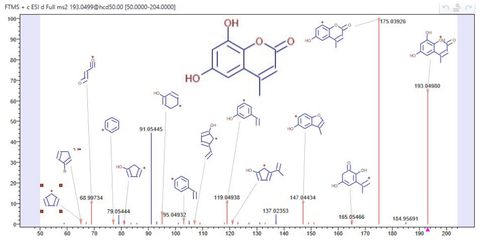
Utilizing the structure candidates from the subtree search as the starting point, the matching fragment spectra could be studied in the reference library to provide structures for the precursors of matching spectra as well as matching fragment ion structures. For unknown 193.0499 a set of substructures (Figure 5) was proposed and used to derive putative candidates.

Figure 5. Substructures Obtained from Subtree Searching of Unknown 193.0499



Each of the putative candidate structures underwent *in silico* fragmentation and annotation of the acquired MSn tree on each unknown (Figure 6). The putative candidate providing the best annotation coverage of the query spectral tree was assumed to be the best / correct candidate. These proposals were then compared to the real structures for the unknowns used in this study. In general, the proposed candidate structures were very close to the real structure however the positions of some functional groups (hydroxylation position on aromatic rings for example) were difficult to resolve absolutely with this approach which resulted in proposing correct Markush style structures but not absolute structures with the exception of one unknown where the proposed candidate was the correct structure.

Figure 6. Highest Ranking Putative Candidate for Unknown 193.0499



CONCLUSIONS

- The demonstrated approach was able to provide valid and high quality putative structures for the unknowns in the experiment
- Absolute structure and the localization of some functional groups was difficult using this approach even with MS3 and MS4 data indicating that perhaps a more advanced acquisition is necessary.
- Further automation of several time and labor intensive steps is still required.
- Extensive MSn reference data can be leveraged to give substructure information on unknowns that are not present in the reference library provided the library contains compounds with representative common substructures to the unknown.

TRADEMARKS/LICENSING

PO65537-EN0519S

ThermoFisher SCIENTIFIC