A Full-Featured Data Curation Solution for Comprehensive and High Quality HRAM MS/MS and MS^n Library Building

Caroline Ding[1], Jakub Mezey[2], Michal Raab[2], Michal Gamblika[2], Melissa Montoya[3], Tim Stratton[3], Robert Mistrik[2]
[1]Thermo Fisher Scientific, San Jose CA, USA; [2]HighChem LLC, Bratislava Slovakia; [3]Thermo Fisher Scientific, Austin TX, USA

ABSTRACT

Purpose: To demonstrate a complete software toolset for the curation of HRAM MS/MS and MS^n data to create high quality reference spectral libraries.

Methods: HRAM MS^n data was acquired on a range of reference standards using automated acquisition software (QETool, TreeROBOT). The data was subjected to either semi-automated or full-automated curation using a multi-step process in data curation software (Curator™ software).

Results: A local spectral library of HRAM MS/MS and MS^n data was created from reference data acquired on 50 compounds. The average curation time per compound with semi-automated curation was 1.5 minutes, which included spectral noise removal, spectral averaging, formula and fragment prediction, and spectral recalibration.

INTRODUCTION

Reference spectral libraries are a useful tool for the confirmation of previously identified chemical compounds and also for the elucidation of structure of new compounds. The quality of the reference spectral library has a direct impact on its utility. Of course, curation allows the detection of potential impurities in reference material that would otherwise pollute the reference spectra; however, the impact of full curation can extend beyond this obvious benefit. The impact on the quality of the spectral matches can be seen in the tighter allowed tolerance in mass accuracy that can be used for query data. A high quality reference spectra, which has been curated to remove spectral noise and recalibrated for superior mass accuracy, is of greater value.

MATERIALS AND METHODS

Sample Preparation

Reference standards were prepared for acquisition following a standard operating procedure in which initial stock solutions were prepared by dissolving a small amount of reference material in ACN:MeOH (1:1) followed by dilution (minimum 100-fold) with water:ACN:MeOH (2:1:1). Standards were individual pure compounds, not mixed with any other compound.

Mass Spectrometer Acquisition Conditions

Data was acquired using automated software for library spectra acquisition. Briefly, data on Thermo Scientific™ Q Exactive™ platforms was acquired by flow infusion using an automated approach to acquire replicate spectra (3) at each energy level (NCE, normalized collision energy) starting from 10% and ending at 200% in increments of 10%. Acquisition was only begun after the signal for the precursor was observed and sufficient signal to achieve ≥20% of the AGC (automatic gain control) level was present (AGC = 2e5). On the hybrid platforms (Thermo Scientific™ Orbitrap Elite™, Thermo Scientific™ Orbitrap Fusion™), data was acquired by nanoinfusion using an automated tool (TreeROBOT), which acquired data on all observed adducts with each adduct taken to an individual MS^n depth. The MS^n depth was set to n=10 for the M+H/M-H adduct and n=2 for all other adducts with the exception of compounds where no M+H/M-H was observed. In these cases, the most abundant non-metal adduct was instead taken to MS^{10}. HCD (higher energy collisional dissociation) fragmentation was performed on the Q Exactive system while trap CID (collision induced dissociation) was performed in an automatically optimized fashion.

Mass spectrometers: Q Exactive, Orbitrap Elite, Orbitrap Fusion systems

Sample introduction: Thermo Scientific™ Accela™ 1250 LC pump with Open AS autosampler (Q Exactive system) or Advion™ Triversa Nanomate® nano-ESI system (Orbitrap Elite and Orbitrap Fusion systems).

Table 1. LC gradient for sample analysis

Time (min)	% A (Water + 10 mM ammonium formate)	% B (ACN)
0	50	50
2	50	50
2.2	5	95
2.5	50	50
3	50	50

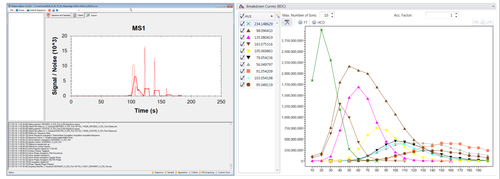
RESULTS

Acquisition Conditions – Representative Reference Data

During acquisition or real world, real time LC-MS data, the extent of MS^n data is limited by the complexity of the sample (number of targets) and width of the peak (time to acquire data). Depending on the specifics of the acquisition of the query data, it is possible that different MS^2 ions are selected as precursors for MS^3 acquisition from injection to injection. Given the query data will often come from a single injection, it is important that the reference spectral library data against which the ID will be performed contains all the likely MS^n spectra that a query acquisition would gather. In other terms, the reference spectral library should always be bigger than the possible query acquisition to assure the best possible match for confirmation of ID. There is an additional benefit for this approach – a large reference spectral library also contains a greater amount of chemical substructure fragmentation information that can be useful for substructure identification of true unknowns. To ensure that the reference data acquired was of sufficient size and coverage, automated tools were developed to facilitate acquisition.

For the Q Exactive platform, an automated tool that allowed for batch acquisition of large numbers of standards was created (Figure 1). The tool allowed for sequence setup of one or more target standards per injection with acquisition performed by flow infusion analysis. During each injection, each standard was detected by its target *m/z* provided in the sequence. For target *m/z* values meeting acquisition criteria for ion flux and stability, data was acquired from 10 to 200% NCE in 10% increments. Replicate scans (n=3) were acquired at each energy level. In addition, the tool automatically re-checked the ion flux of the target every 15 scans (every five energy levels).

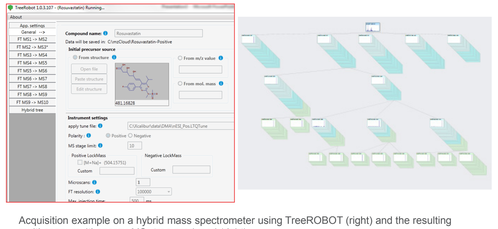
Figure 1. Q Exactive platform data example



Real time acquisition of HCD MS^2 for 4-methoxy-a-pyrrolidinopropiophenone (left) and the resulting breakdown curves for the top 10 fragment ions (right)

For hybrid acquisition, the process is more complex due to multiple fragment techniques and greater depth (up to MS^{10}). For this, another acquisition tool, TreeROBOT, was created. Specific tools were created both for legacy hybrid systems (Orbitrap Velos Pro MS and Orbitrap Elite MS) and the new generation platform (Orbitrap Fusion MS and Orbitrap Fusion Lumos MS). User values are entered for depth and width of spectral tree, adducts to consider, minimum signal intensity for MS to MS/MS and for subsequent MS^n to n+1. Detection of selected adducts in the full MS1 signal, confirmation of mass accuracy and minimum signal intensity, and selection of adducts to acquire are all automated. HCD is acquired as on the Q Exactive platform with replicate scans acquired at each energy level as the energy level was incremented from 10% to 200% NCE. Trap CID is acquired in a smart fashion to find the narrow range of the breakdown curve and acquire sufficient data across this region.

Figure 2. Data acquisition in TreeROBOT for hybrid platforms

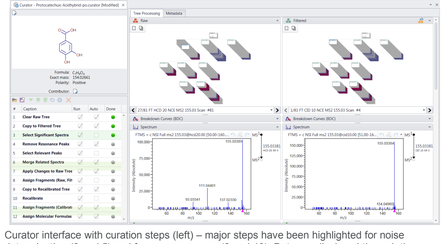


Acquisition example on a hybrid mass spectrometer using TreeROBOT (right) and the resulting multi-scan, multi-energy MS^n tree produced (right).

Data Curation for Library Creation

Acquisition of the reference spectral library data is only the first step. There are a number of aspects of raw mass spectral data that can impact the quality of a spectral match. The software application developed allows for manual, semi-automatic, or automatic processing through a series of curation steps (Figure 3). The major steps in curation involve addressing sources of variability in mass spectrometric data. Two major aspects of variability that can impact spectral library matching include spectral noise and mass inaccuracy.

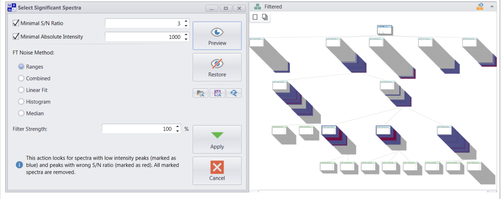
Figure 3. Major Curator steps and interface



Curator interface with curation steps (left) – major steps have been highlighted for noise determination (3 and 5) and for mass accuracy (3 and 10). Data are displayed through the process moving from Raw (middle) through Filtered (right) and finally to Recalibrated (not shown).

Noise can arise from multiple sources, the most prevalent of which are electronic and chemical noise. In the case of electronic noise, it is typically non-reproducible low-level signal observed in acquired mass spectra. Chemical noise may arise from co-isolated species – either observable resolved signals or unresolved 'hidden' co-isolated components. Another potential source for chemical 'noise' – better considered as interference – comes from an impure standard, which is a mix of two isomers. In any of these cases, fragment ions may be observed in the spectra that are not directly formed from the expected isolated precursor. These fragment ions will have a negative effect on attempts to match the reference as the query spectra may not contain these ions. During curation, these signals are detected and filtered out in one of two ways. The first is through analysis of the replicate spectra obtained for a given set of conditions. Signals that are not consistent across replicate scans are likely to be noise and are flagged as such. Second, signals that cannot be explained as a formula or a fragment ion structure of the isolated parent are also possibly noise. The first step is to establish the spectral noise level, and optionally to set an absolute signal threshold (Figure 4), which can be applied immediately to remove spectra which contain no significant signal.

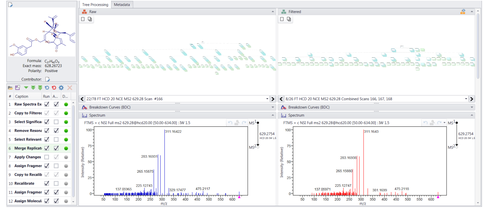
Figure 4. Establishing noise threshold and impact on spectral tree



First spectral noise curation step where both S/N and absolute threshold can be set (left) and the effect on the spectral tree seen (right, highlighted spectra in blue or red will be filtered out as containing no signal above the noise level)

To meet the approach of having a reference library spectral tree that is sufficiently complete to account for variations in routine acquisition of query data, very large reference spectral trees were acquired and tested through the curation process. Although the largest MS^n trees (n=9) could contain up to and beyond 5,000 individual spectra, the curation tool was able to process the spectra. Curation times were not significantly longer for the large spectral trees as the automatic data handling and curation UI allowed for quick review of the data (Figure 5). On average, while a relatively small MS^4 spectral tree of 300–600 spectra would take between 2 to 5 minutes to curate, the larger spectral trees reaching MS^8 and beyond with spectral counts of 2,500 and more took between 4 to 8 minutes.

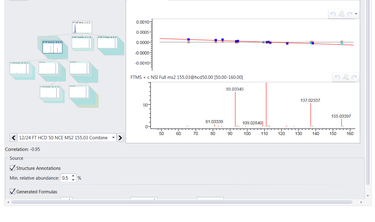
Figure 5. Large scale MS^n spectral tree handling



Curation of acquired data for resiniferatoxin containing 4,295 spectra.

During the process of determining relevant scans and spectral peaks, formulas and fragment structures were predicted for observed peaks. These formulas and structures aided in determining which peaks may be noise and which were relevant signals. In addition, these also provide theoretical values for the *m/z* of observed spectral peaks. It is possible to use these theoretical values as a guide in recalibration of individual spectra. While using a theoretical value alone would be risky, carrying the chance of error if the theoretical prediction was incorrect, we can use a "preponderance of evidence" approach where the predicted formulas and fragment *m/z* values for all spectral peaks in a spectra are used collectively. Prior to this step in the curation, the multiple replicate scans acquired at each condition were averaged to give a single *m/z* value for each fragment ion as an average of three replicate acquisitions. If multiple formulas or fragment structures can be predicted for a single spectral peak, all of them are used in the creation of the recalibration (Figure 6). The recalibration is automatically constructed and applied to all spectra with the MS^n tree.

Figure 6. Recalibration of spectra

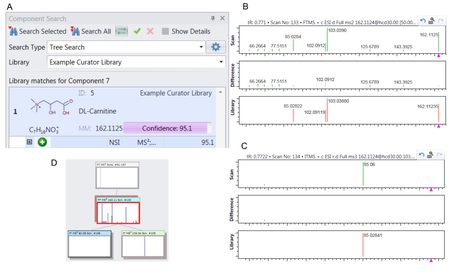


After curation, entries were exported to a local library, which can be stored either on a single PC or stored on a local network for shared use. Spectral entries were exported so that all three types of data were retained, original raw, filtered for spectral noise, and fully curated with recalibration, although it is an option to only export curated data as well.

Spectral Library Searching

The constructed library was used to perform MS/MS and MS^n spectral library matches for compounds either as pure standards or in matrix to demonstrate the utility of the library. The ability to match MS^n spectral data against a curated library provides for a greater degree of confidence in a match by using both higher quality reference data and matching sequential fragmentation events in the MS^n tree branches (Figure 7).

Figure 7. Example of MS^n Tree Identification Search using library



(Clockwise from top left):
A. Search result for MS^n tree similarity search.
B. Library result of query (green) vs library (red) for MS^2 of unknown (*m/z* 162.1124)
C. Library result of query (green) vs library (red) for MS^3 of unknown (*m/z* 162.1124 → 103.0390)
D. MS^n tree of query unknown compound

CONCLUSIONS

- Curation of reference spectral data is an important aspect of improving the quality of a reference spectral library.
- Higher quality reference spectral data can allow tighter query mass tolerance, which can improve confidence in spectral library ID hits.
- Automated software tools, like Curator, can speed up the curation of even complex large MS^n spectral trees, streamlining the reference spectral library creation process.

TRADEMARKS/LICENSING

PO65283-EN 0618S