Annotated Recurrent Unidentified Spectra (ARUS)

To assist in the compound identification problem, the NIST Mass Spectrometry Data Center has developed a novel type of mass spectral library, one that includes all recurrent unidentified mass spectra in a material. Unlike traditional spectral libraries, which consist of reference spectra of known compounds derived from neat standards, these libraries are derived from recurring spectra of unknown identity in the target material itself, where spectra are extracted, clustered, and where possible annotated prior to entry into a library. Building the library itself follows a similar methodological procedure to the one described for libraries of neat compounds, though with a different set of spectrum and measurement annotation.

In general, this type of library can be useful in many usual tasks of 'omics studies (i) answering where, how often, and in what conditions certain ions are observed, (ii) assigning class ID for compounds not in current tandem mass spectral libraries or not commercially available, (iii) connecting samples in an unambiguous way for control-case studies or interlaboratory comparisons (each molecular feature is represented by a spectrum in the library).

Getting Started

To run the ARUS library, download the installation program for the NIST Search Software and the library from the NIST Website (https://chemdata.nist.gov/). A copy of the library must be present in the folder "MSSEARCH" before opening the browser.

Most datasets were generated using a Fusion Lumos Orbitrap. Eventually, data from other instruments such as Agilent and Waters QTOF was also collected. Although no systematic comparison has been made between data from different instruments, the library coverage for all instruments is similar. However, mass accuracy and ranges need to be adjusted accordingly in order to yield similar library scores for the same ions.

Figure 1 shows an example of a library spectrum that was annotated using the hybrid search technique:

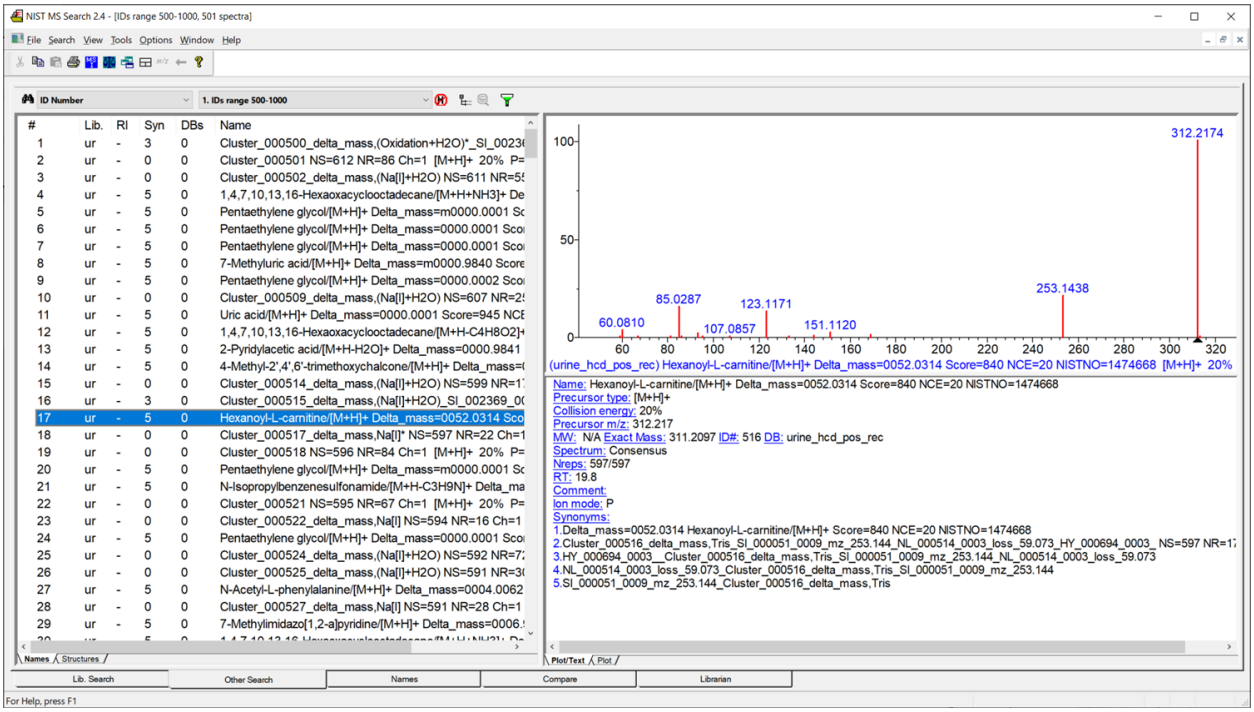

Figure1. Example of an annotated compound from the urine_hcd_pos_rec library using the NIST Hybrid Search and the NIST Tandem Library.

This method is called a “hybrid” because it combines matching both ion m/z and mass losses from the precursor ion (neutral losses for singly charged ions). It is used to identify tandem mass spectra of compounds that differ from library compounds by a single chemical group. Peaks containing this group will, of course, be shifted by the mass of this group, as will the mass of the molecule that contains this group. This difference, termed DeltaMass, is used to shift the product ions in the library spectrum that contain the modification, thereby allowing library product ions that contain the unexpected modification to match the query spectrum. Peaks that match before or after shifting are treated equally, and if a single peak matches both before and after shifting, the abundance is partitioned. The hybrid search is implemented in the MS Search Program and can be used interactively or in a batch run.

Compounds that do not match a library spectrum using the NIST Hybrid Search are true unknown and named after the cluster number of the consensus spectrum (see Figure 2).

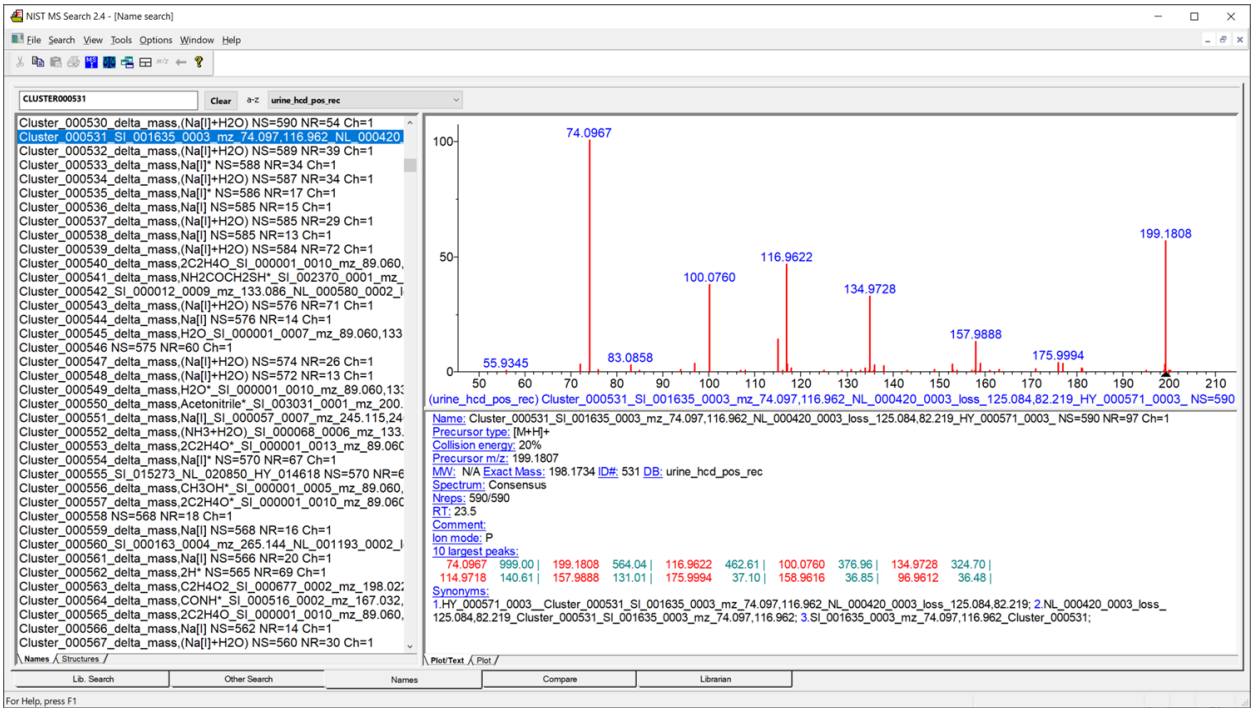

Figure 2. Example of an unknown compound from the urine_hcd_pos_rec library.

The following nomenclature was included within compound names in the library: Name_Adduct Type_Score_DeltaMasss_Formula_LibID, where Name, DeltaMass, Score and Formula are derived from the best hit in a hybrid search. LibID is a sequential number assigned to spectra in the archive. In case of not matching, known-unknown “Names” were simply given as cluster numbers (see Figure 2).

Relevant information about the spectrum, such as name, retention time, the number of spectra used for building the consensus spectrum and others can be retrieved by using the Options tab > Comments Field Display (see Figure 3).

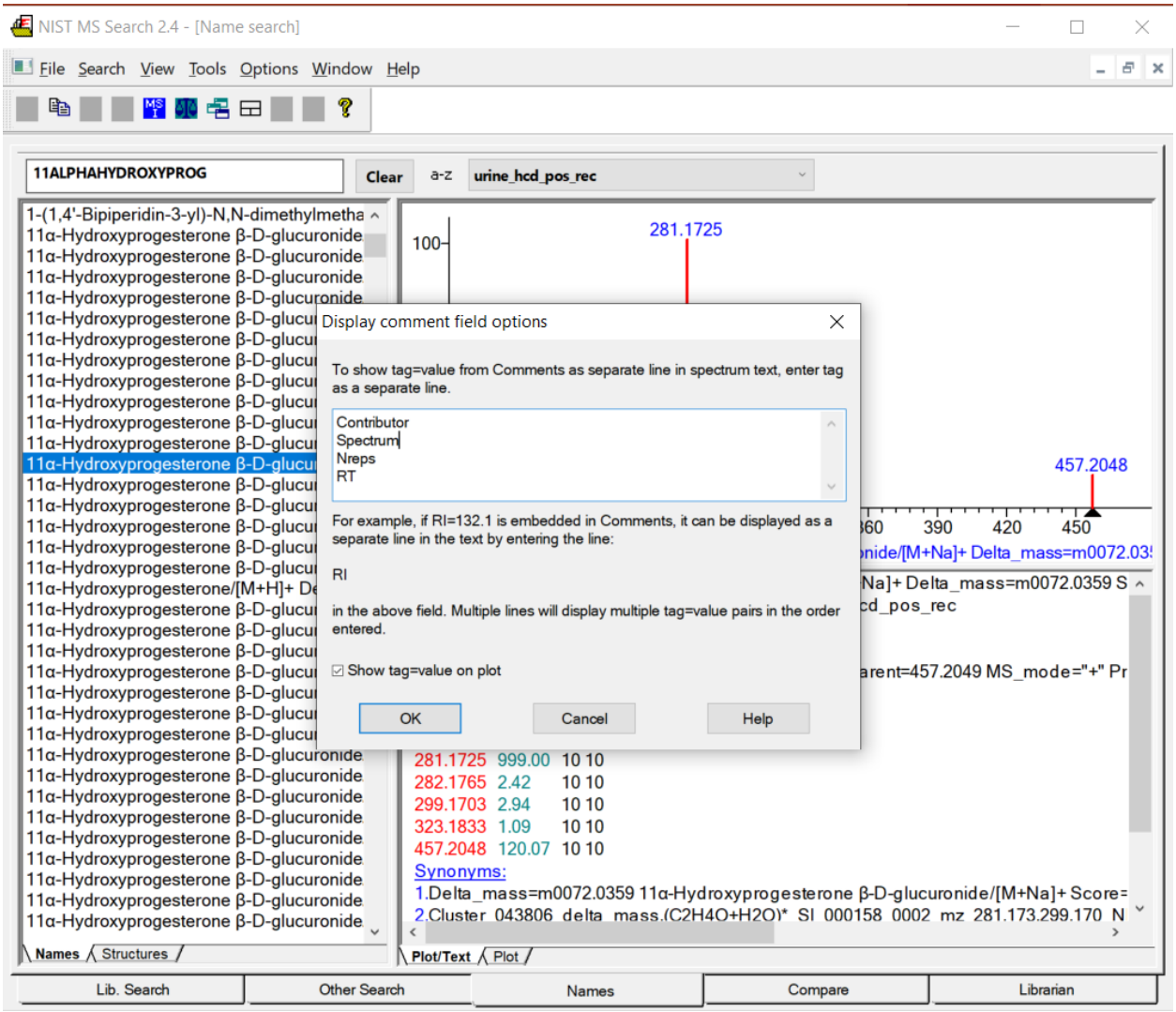

Figure 3. Comment field options.

Prerequisites

Downloaded the NIST Search Software and libraries from the NIST Website (https://chemdata.nist.gov/).

Installing

The installation of the search program is straightforward; however, a detailed manual can be found on the website accompanying the software.

Running the tests

For testing the installation and library performance take the following these steps:

1.- Download and install the NIST20 browser.

2.- Download one or more raw datafiles. A human plasma data-dependent dataset can be found at https://nvlpubs.nist.gov/nistpubs/jres/121/jres.121.022.pdf and used for this purpose.

3.- Use ProteoWizard or proprietary software to extract the tandem mass spectra in MGF format.

4.- Import the MGF files into the NIST20 browser by using the FILE tab > Open.

5.- Adjust searching options (Figure 3) by using the Options tab > Library searching options.

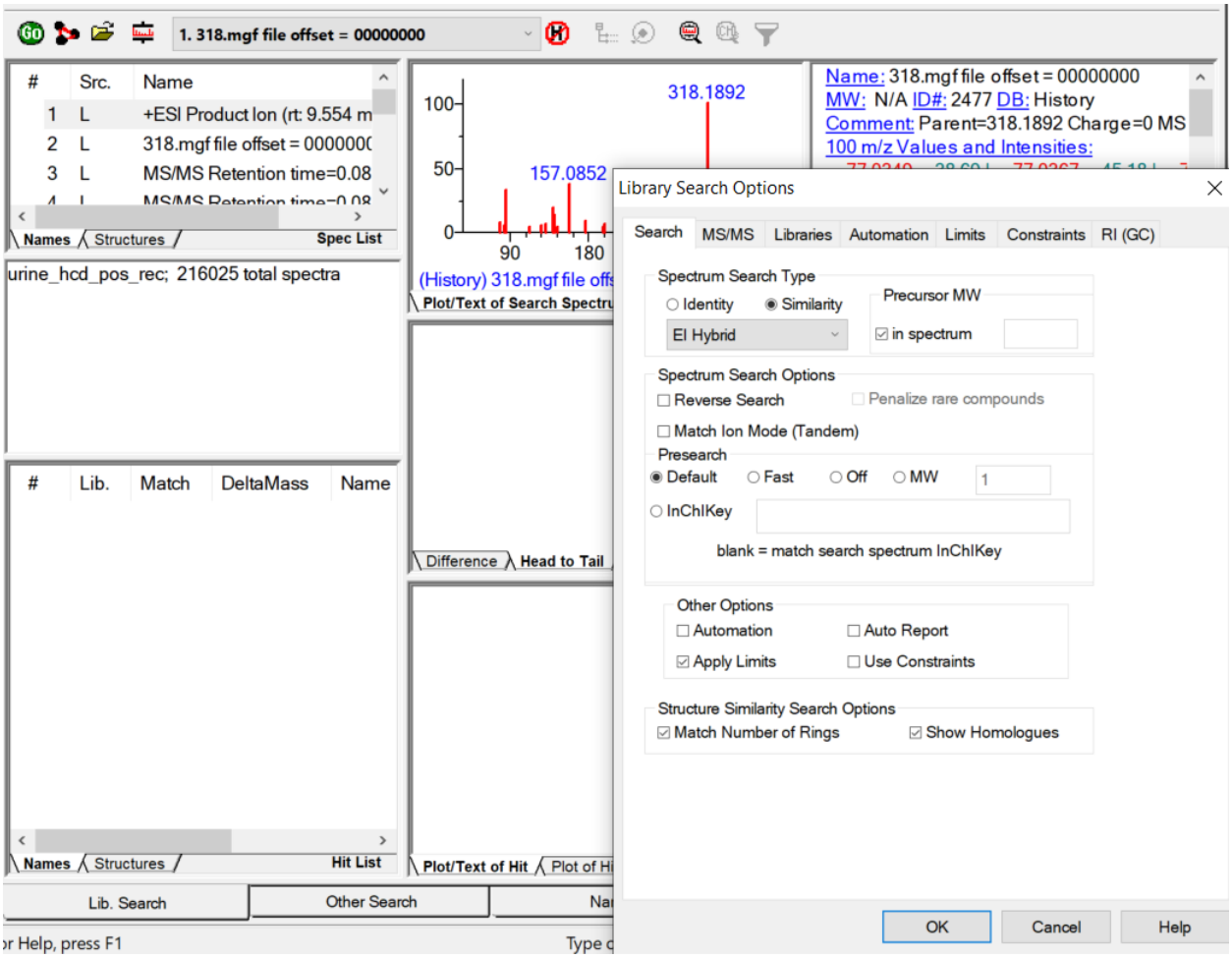

Figure 3. Library searching options

6.- For searching a single query spectrum, simply double-click on its name. For searching batches using MSPepSearch in a command line, consult the documentation of the NIST browser and the MSPepSearch program (help tab). A MS Excel file containing the library identifications in a single run of a plasma sample can be found with the ARUS libraries. The first cell of the second raw contains the command line used for the library searching.

Built With

- NIST pipeline
- NISTms and MSpepSearch

Contributions

All software and datasets were developed at the NIST Mass Spectrometry Data Center. For submitting requests or external contributions, please contact:

massspec@nist.gov

Versioning

Versions 3.0, other versions are available upon request.

Authors

NIST Mass Spectrometry Data Center.

License

NIST Licensing Policy. The library is offered without warranty and will be in continuous development in the future (http://chmdatafnx.nist.gov/dokuwiki/doku.php?id=chemdata:arus).