The Fragmentarium: A Universal Query Service Enabling Partial Matching of Unidentified Spectra across the Full Gamut of NIST MS Spectral Libraries

Manor Askenazi[1], Stephen E. Stein[2]

[1]Biomedical Hosting LLC, Arlington, MA, [2]National Institute of Standards and Technology, Gaithersburg, MD

NIST National Institute of Standards and Technology U.S. Department of Commerce — Mass Spectrometry Data Center

Overview

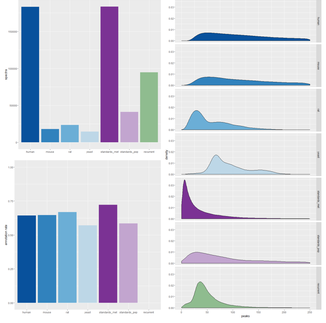
The Fragmentarium is a spectral library search (http://fragmentarium.biomedical.hosting) based on the premise that:

1) A diverse **"Library of Libraries"** including both metabolites *and* peptides, generated from pure standards *and* naturally occurring mixtures,
2) using only **Accurate Mass Fragment Data** (with errors usually much lower than 100ppm),
3) and applying a **Precursors Independent** similarity score

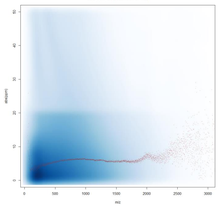
can help elucidate unidentified spectra by suggesting co-occurring fragment sets which match large parts of the query spectrum regardless of precursor value.

Introduction

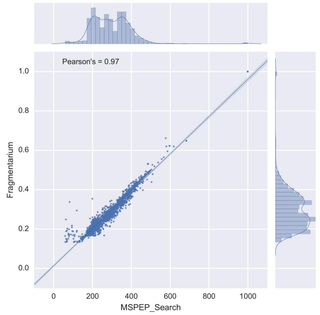
Matching fragment spectra against a spectral library constitutes the only truly universal method of mass spectral identification, since it can be performed across the entire range of chemistry, separation, ionization and detection techniques. As the curated set of NIST spectral libraries continues to cover increasingly diverse regions of chemical space, and as these libraries shift to accurate mass, the value of the individual curated fragment information generated by even a partial spectral match will become increasingly useful. We have therefore implemented a website enabling the user to search all the available NIST accurate mass libraries (excluding iTRAQ data), using only fragment peak-list data (the user need not provide intact precursor mass, instrument type nor sample preparation details). The system then generates a report including a **Fragment-Oriented Visualization** of the resulting matches.



"Library of Libraries" The Fragmentarium is composed of 7 libraries: 4 peptide libraries generated from naturally occurring protein samples (in shades of blue), 2 libraries generated from pure standards of both metabolites and peptides (purples), as well as one library (green) of challenging recurrent spectra recently elucidated using a novel search technique [1]. Despite differences in raw spectral count (top left histogram) and peak number distribution (right) – note in particular the differences between rat/yeast and mouse/human which reflect differences in the instrumentation (QTOF vs. Orbitrap, respectively) the individual peaks in these libraries are *annotated at a consistent and high rate* (bottom left histogram) with the exception of the very recent "recurrent" library which, while providing identifications at the spectral level, does not yet offer individual peak annotation.



Accurate Mass Fragment Data The libraries are all comprised of Accurate Mass spectra, as demonstrated by this density plot (above) of fragment accuracy by m/z value for the human peptide library (the red points represent a running median estimate at 1Da intervals).

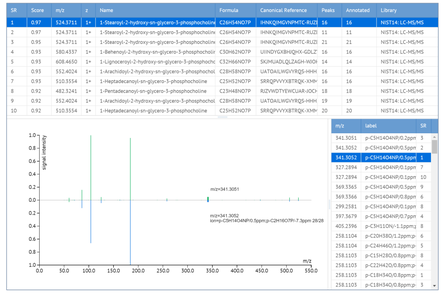


Precursor Independent Similarity The Fragmentarium applies a similarity score to the complete library, without filtering by precursor. This is achieved using a dot-product approximation implemented entirely in Python and standard SQL using efficiently indexed range queries. The system produces scores approximating the dot-products reported by NIST's MSPepSearch (Windows-based) program, at roughly equivalent speeds while remaining OS independent. The correlation plot (above) shows the score-pairs generated by shared matches among the top 400 hits returned by MSPepSearch and the Fragmentarium when applied to a random sample of 8 spectra from the human peptide library. The discrepancy between the scores is due primarily to a difference in policy regarding the matching of peaks where, in the interest of speed and simplicity, the Fragmentarium does not enforce a one-to-one mapping between pairs of peaks across the query and candidate spectra.

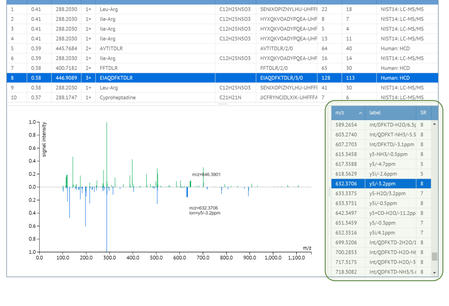
Fragment-Oriented Visualization The Fragmentarium emphasizes access to, and navigation through fragment information usually considered secondary to spectral match details. This is achieved by two primary mechanisms: A table showing all the annotated fragments (bottom visualization, green transparent box). Clicking on any of these fragments loads the entire associated spectrum (indicated by the SR column which stands for Spectral Rank) and zooms the viewer to the fragment in question. The second mechanism is fragment sorting by absolute distance to a target query peak: consider the topmost visualization where the user has clicked the green query peak at m/z=341.3051. This action caused a sorting of the fragment table by absolute distance to the this target query value. The user can immediately verify that only the top 3 spectra (SR=1,2 and 3) contain annotated matches to the peak of interest.

Results

The utility of the precursor independent search offered by the Fragmentarium is immediately obvious in the following visualization, where the user can clearly verify that spectra generated from a wide-range of precursor masses all match the query spectrum (in green) with high scores and all constitute various forms of phospholipids:



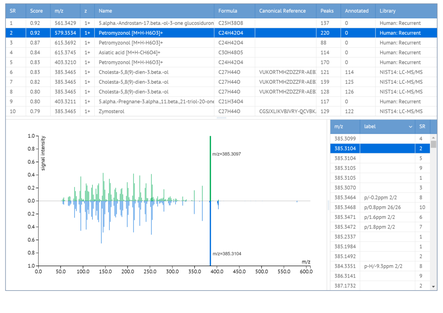
The second report (below) illustrates the elucidation of a challenging iTRAQ spectrum which was not identified by the default peptide spectral matching software due to an incomplete iTRAQ labeling displaced by a Lysine methylation. The spectrum corresponding to the reference (unlabeled form) of the peptide was returned by the Fragmentarium and can be seen along with the tell-tale delta between the y5 ion and its putatively methylated pair in the query spectrum as highlighted by the user.



Conclusions

Precursor independent searches have been used internally at NIST both for proteomics and metabolomics related investigations. In particular, partial matching has helped elucidate peptides with post-translational modifications in proteins such as Collagens and Histones and has proven useful when dealing with challenging iTRAQ labeled spectra (as shown in the bottom-most report). In the realm of metabolomics and small molecules we have benefitted from precursor-free search results when elucidating spectra from many chemical classes including Carnitine-like compounds and phospholipids (an seen in the top-most report).

We are currently updating the interface to accommodate the recent inclusion of the recurrent spectral library. Though its fragments are essentially unannotated, this novel library is extremely useful since it can provide indirect matches (via its Delta Mass identification) impossible to achieve by direct dot-product similarity search or even precursor tolerant search. The Fragmentarium can already search and display these recurrent spectra but the chemical formula displayed may not match the reported library m/z value as there is a Delta Mass component which is currently not displayed (see [1] and the example below for more details):



References

[1] *Tandem mass spectral libraries of recurrent unidentified spectra for urine and plasma: A new kind of library for metabolomics applications*, Yamil Simon, Ramesh Marupaka, Xinjian Yan, Yuri Mirokhin, Kelly H. Telu, William E Wallace; Stephen E Stein, **MOF pm 3:10**, ASMS 2016