Poster Reprint

ASMS 2020
TP 576

Identifying Food and Environmental Contaminants using the New NIST High-Res MS/MS Library Search Algorithms and Publicly Available LC/MS/MS Spectral Libraries

Emma E Rennie[1], Frank Kuhlmann[1], James S Pyke[1], Stephen Madden[1] and O. David Sparkman[2].

[1]Agilent Technologies Inc., Santa Clara

[2]University of the Pacific, Stockton, CA

Introduction

Crowd Sourced Publicly Available Libraries

- Are growing at an increasingly fast rate, a rate that no single vendor or academic group could possibly achieve.
- Are globally utilized in both high resolution accurate mass (HRAM) LC/MS/MS and GC/MS suspect screening, non-target screening and unknown compound identification workflows.

Unlike GCMS EI 70eV spectra, LC/MS/MS spectra are not reproducible across different instrument platforms because their fragmentation patterns, or the relative abundance of fragment ions, are highly dependent on the:

- Analyzer or instrument type.
- Ion source parameters.
- Collision energy (CE).

In fact, fragment ions which are present in a Q-TOF spectrum may not be present, or are of such low abundance, that they are barely seen in a linear or orbital trap spectrum.

Crowd Sourced HRAM LC/MS/MS Spectral Libraries

- Contain spectra from multiple vendors and instrument types.
- Can contain high quality curated content which has accompanying reference information along with un-curated spectra with little metadata.
- Can contain spectra collected with a variety of experimental and curation protocols.

Library search algorithms which can provide highly confident compound identifications from these public crowd sourced HRAM spectral libraries are essential for a successful data analysis workflow.

Experimental

Food and Environmental Contaminants Data

Typical known contaminants are relevant to food safety and environmental applications were spiked into solvent and measured in Auto MS/MS mode on Agilent Q-TOF LC/MS instruments. The data were analyzed using the Find by Auto MS/MS compound mining algorithm in the MassHunter Qualitative Analysis Software 10.0 (Qual) and sent to the NIST MS Search Program v.2.3 (MS Search) for identification.

This study has concentrated on the library search performance with 3rd party crowd sourced libraries and not on the library content.

Experimental

HRAM LC/MS/MS Spectral Libraries

Almost 100,000 LC/MS/MS spectra were downloaded as NIST MS Search compatible msp files from the Mass Bank of North America (MoNA)[2] and converted into NIST MS user libraries. These user libraries contained experimental spectra from multiple data repository sources, such as the Vaniya/Fiehn Natural Products Library, MassBank EU, ReSpect, HMDB, MetaboBASE and GNPS.

The performance of the NIST Library Search algorithms with these libraries was assessed by calculating the top ranked hits and receiver operator characteristic (ROC) curves for a number of compound classes, including pesticides, veterinary drugs and pharmaceuticals and personal care products (PPCPs).

NIST MS Search Program Library Search Results

Each MS/MS spectrum for a known contaminant was searched in MS Search using MS/MS library search options provided by NIST and optimized for the Agilent LC/Q-TOF. Search results for the pesticide imazalil are shown below. The Q-TOF MS/MS sample spectrum with a CE of 22 V is shown on top in red and the top ranked hit, an orbital trap spectrum with a CE of 35 (nominal), is shown on the bottom in blue.

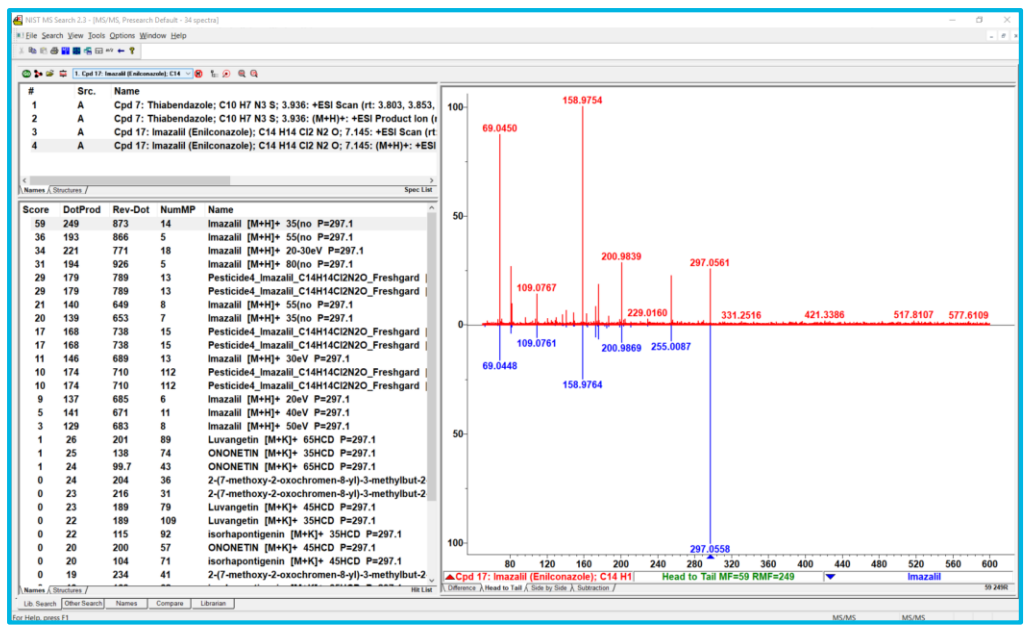

Figure 1. Typical NIST MS/MS library search results.

Library Search Assessment Calculations

The Hit list for each spectrum was copied into Excel® and each hit was manually inspected and assigned as being a correct or incorrect compound ID ($Comp_{ID}$). The $Comp_{ID}$ for ~1200 library hits for 47 contaminants was assigned.

The top ranked hits were calculated using the $Comp_{ID}$, Score, DotProd and Rev-Dot.

ROC curves were calculated, using protocols[3] established for when many identification hits are returned for each spectrum, to asses the sensitivity and specificity of the DotProd and Rev-Dot scores.

Top Ranked Hits

The top ranked hits (table 1 below) were calculated for all 47 contaminants as well as the 3 compound classes:

- Top1*$_{(RD)}$, where the correct Comp$_{ID}$ ranks in first place based on the Rev-Dot and Rev-Dot ≥ 700.
- Top3*$_{(RD)}$, where the correct Comp$_{ID}$ is amongst the top 3 entries based on the Rev-Dot and Rev-Dot ≥ 700.
- Top3$_{(RD)}$, where the correct Comp$_{ID}$ is amongst the top 3 entries based on the Rev-Dot.
- Top1*$_{(S)}$, where the correct Comp$_{ID}$ ranks in first place based on the Score and Rev-Dot ≥ 700.
- Top3*$_{(S)}$, where the correct Comp$_{ID}$ is amongst the top 3 entries based on the Score and Rev-Dot ≥ 700.
- Top3$_{(S)}$, where the correct Comp$_{ID}$ is amongst the top 3 entries based on the Score.

Compound Class	Top1*$_{(RD)}$	Top3*$_{(RD)}$	Top3$_{(RD)}$	Top1*$_{(S)}$	Top3*$_{(S)}$	Top3$_{(S)}$
Pesticides	95%	95%	100%	90%	95%	100%
Vet. Drugs	100%	100%	100%	100%	100%	100%
PPCPs	85%	100%	100%	100%	100%	100%
All Contaminants	93%	98%	100%	95%	98%	100%

Table 1. Correct Comp$_{ID}$ Identifier in Top Ranked Hits (%).

The Score and Rev-Dot ranking were shown to be excellent indicators of the correct Comp$_{ID}$.

When calculating the top ranked hits with the DotProd (not shown) instead of the RevDot, the Top3($_{DP}$) was found to be 100%, while the other ranking hits had values of ~50% showing that the DotProd alone is not a useful identification indicator for the correct Comp$_{ID}$.

Influence of Precursor *m/z* Tolerance

With unknown library content it could be argued that, for the vast majority of the library hits only one possible candidate exists for each precursor ion and a tight tolerance window. The chart below shows the number of unique compounds for each precursor ion *m/z* for 21 pesticides in the libraries. A 10 ppm precursor *m/z* tolerance was used throughout this study.

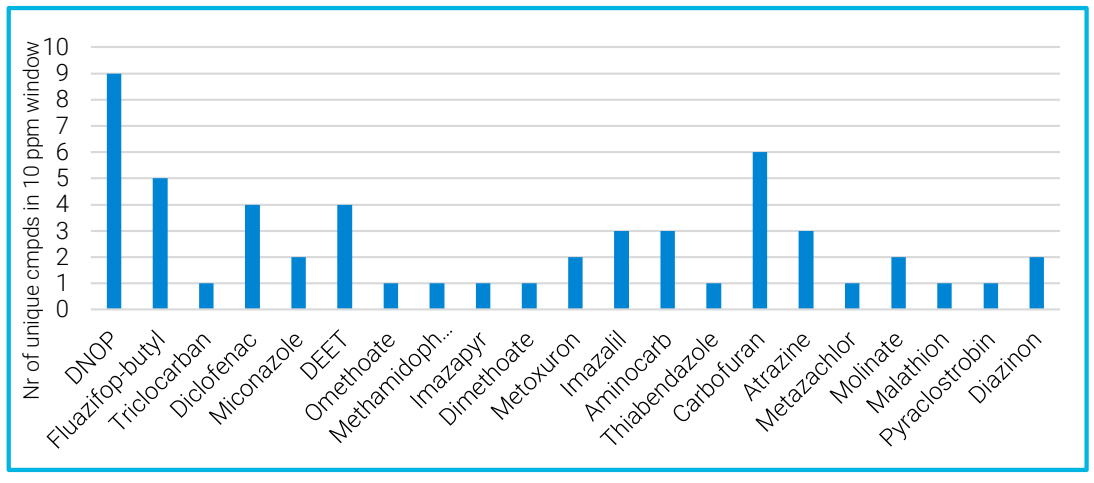

Figure 2. Precursor *m/z* search with 10 ppm tolerance.

ROC Calculations

The performance of DotProd and Rev-Dot as true Comp$_{ID}$ indicators is seen in the shape and position of the ROC curves. Generally, poor models have lines close to the rising diagonal, whereas perfect models produce curves that coincide with the left and top sides of the plot, where both the sensitivity and the specificity are 1. The area under the curve (AUC) represents the degree of separability, the higher the AUC, the better the model.

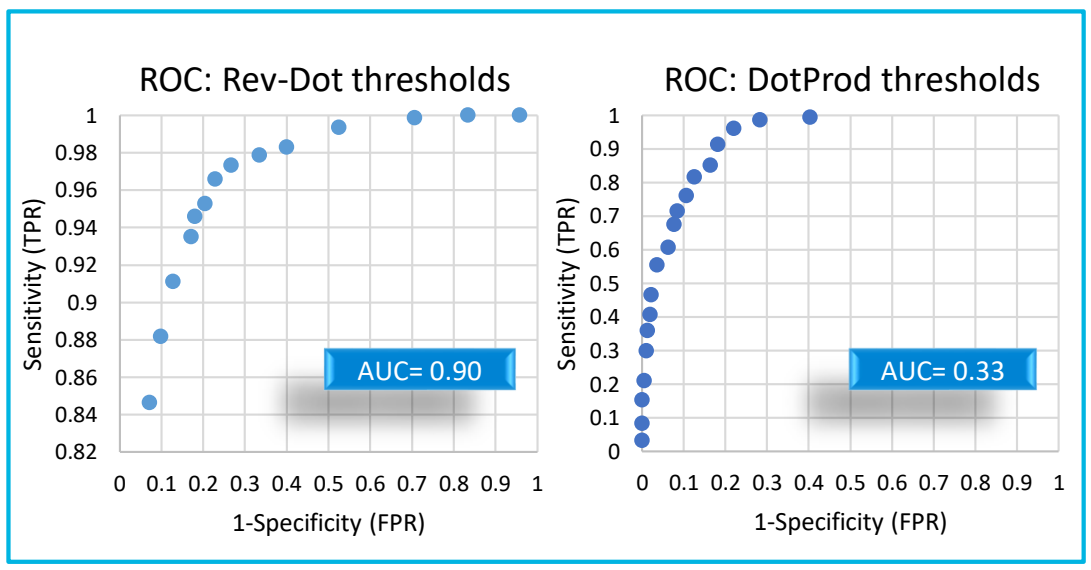

Figure 3. ROC Curves generated using Rev-Dot and DotProd thresholds.

From the two models tested, the best performance was obtained for Rev-Dot, which is in very good agreement with the results from the top ranked hits calculations.

It is not possible to calculate a ROC curve based on the Score, which is not normalized between Hit lists, and should only be used as a rough measure of identification confidence.

The Rev-Dot ROC curve can be used to determine the minimum Rev-Dot threshold that should be used to classify a library hit match as being correct. A good compromise between sensitivity and specificity is obtained by using a Rev-Dot threshold of 650, where the sensitivity, or the true positive rate (TPR), is 0.91 and the specificity is 0.87, equaling a false positive rate (FPR) of 0.12.

The result for the Rev-Dot when using public crowd sourced HRAM libraries is in very good agreement with the guidance published by NIST for the DotProd and Rev-Dot scores[4] when using MS Search with the curated NIST library content.

However, from both the top ranked hits and the ROC curves it can be seen that the DotProd score is a much less useful indicator than the Rev-Dot or the Score when using public crowd sourced HRAM libraries instead of the curated NIST HRAM Tandem Mass Spectral Library.

Results and Discussion

The NIST HiRes MS/MS Hybrid Library Search

A new addition to the MS Search v.2.3 is the HiRes MS/MS Hybrid Search. This is a similarity search which can be performed when it is believed that a spectrum of the unknown compound is not in any of the searched libraries. This option finds compounds that differ from a library compound by an 'inert' chemical group[3]. Spectra similar to the searched spectrum make up the Hit list, with a DeltaMass column showing the library spectrum accurate mass subtracted from the search spectrum accurate mass.

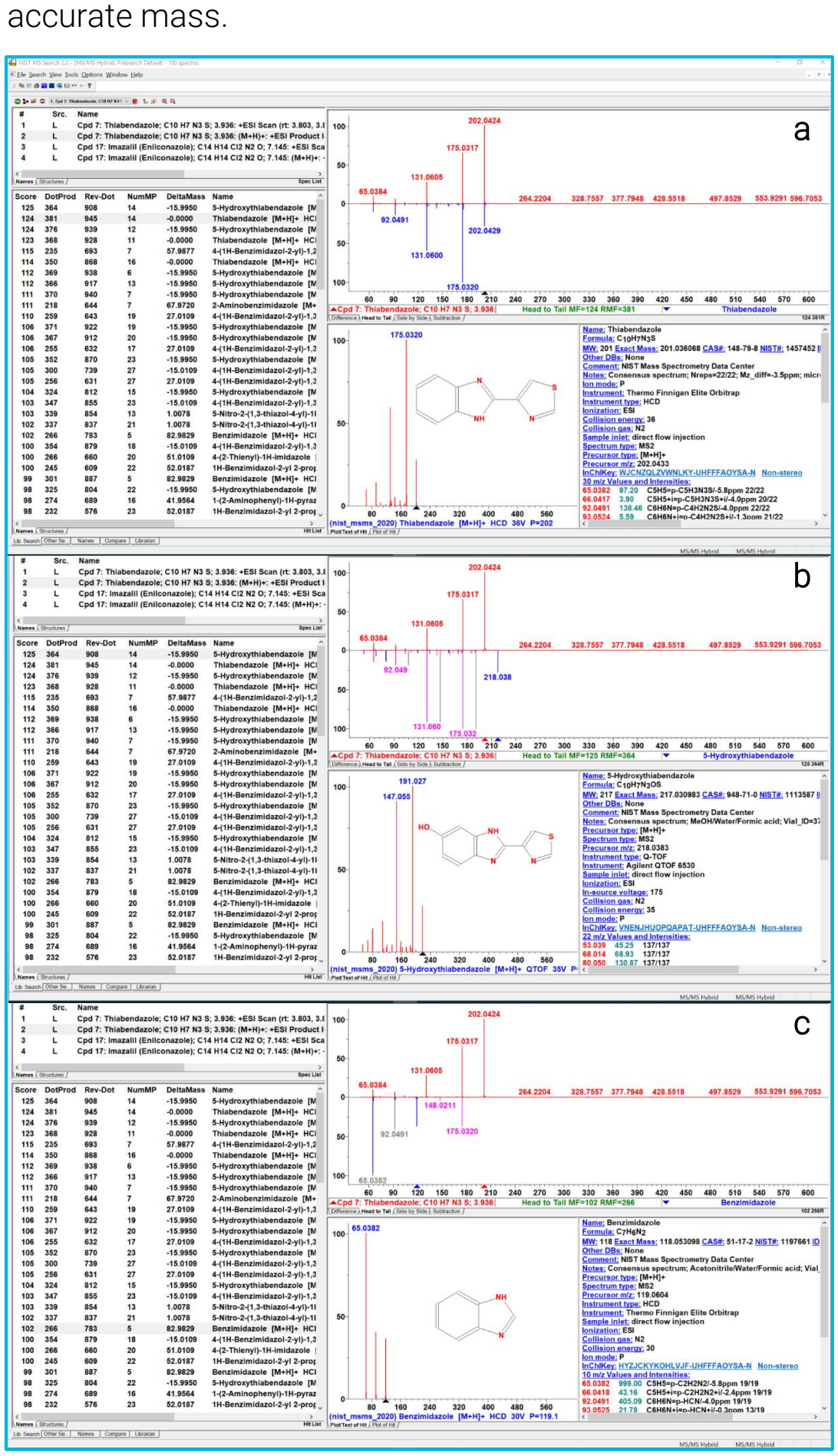

Figure 4. MS/MS hybrid search for thiabendazole.

The results from the MS/MS Hybrid Similarity Search against the NIST 20 Tandem library for thiabendazole are shown in Figure 4. The NIST 20 Tandem library contains structural information which is not present in a msp file. The chemical structures are extremely useful when using the MS/MS hybrid search option for unknown compound identification.

Three Hit list entries have been shown; a) thiabenadazole, b) 5-hydroxythiabendazole and c) benzimidazole. The head-to-tail plots show matching unshifted product ions in blue, unmatching unshifted product ions in gray, and matching shifted product ions in pink. Browsing through the hit list, it is very easy to correlate the shifted pink peaks to chemical sub structures providing a wealth of information for unknown compound chemical structure elucidation.

Conclusions

- The NIST MS/MS Identity Library Search provides highly confident compound identifications from public crowd sourced HRAM spectral libraries.
- NIST MS/MS Identity Search:
  - Compounds present in the library return multiple spectra with a good to excellent Rev-Dot or Score.
  - Compounds not present in the library return isomer hits, no hits or very poor scoring hits.
- NIST MS/MS Hybrid Similarity Search:
  - Compounds not present in the library return a list of the most closely related compounds with sub-structure details.

We would like to thank Dr. Stephen E. Stein (Mass Spectrometry Data Center, NIST) for his help and for providing the MS/MS library search options parameters.

References

[1]https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start

[2]https://mona.fiehnlab.ucdavis.edu/

[3]Alex Chao et al., Anal Bioanal Chem 412, 1303 (2020).

[4] NIST Mass Spectral Search Program (Version 2.3) User's Guide.

Published in USA, June 1, 2020

Agilent
Trusted Answers