Leveraging MS Search, Accurate Mass, and Retention Index Data with a Novel User Interface to Accelerate Qualitative GC/MS Sample Review

Don Kuehl, Stacey Simonoff, Yongdong Wang
Cerno Bioscience, Las Vegas, NV USA

TP 141

Introduction

GC/MS library search (Search) provides a ranked list of possible matching compounds (Hit List), but the correct hit is not always the top hit due to similar spectra, such as certain isomers, measurement variability, and/or interference from background or co-eluting species. Providing additional metrics such as retention index (RI) values and accurate mass formula identification (Formula ID) of ions can greatly improve compound ID confidence. RI is exceptionally powerful and complementary but it can be tedious to fully calibrate the GC with a typical n-alkane series on a daily basis.

In addition, the burden of evaluating these metrics still falls on the Analyst who must carefully review each eluting peak, a tedious and time-consuming process sometimes taking hours or even days for each run. In this paper we will describe a new method for automatically calibrating for RI and using it in a combined identification metric with a unique color-coding system of results to accelerate the review process.

Methodology

Complementary Metrics

GC/MS library search depends on the ability to accurately match the fragment "fingerprint" of "pure" EI spectra measured at a given source voltage. The search algorithms return a ranked Hit List of most likely matches which the user must review and evaluate to verify the correct identification. In general, match quality is limited by the purity of the peak and the variability in the IE spectra, due to differences in instrumentation and measurement conditions.

Significant separation in the Hit List values ("breakouts") between the top match and the other matches is one indicator of a correct ID. However, the spectral "fingerprints" of isomers or similar compounds can be very hard to differentiate, making the correct ID somewhat ambiguous. In some cases, knowledge of the sample chemistry can help, but there can still be significant uncertainty.

To improve library search accuracy, a logical set of complementary and orthogonal measurements and advanced processing can be applied as follows:

1. Identify mixture peaks and deconvolve to pure component spectra[1]
2. Perform a conventional library search (both forward and reverse) to produce a list of compound candidates
3. Use retention index (RI) to identify the correct match from the Hit List
4. Further validate each library match using accurate mass/spectral accuracy Formula ID[2]

Automated Retention Index Calibration

RI is a powerful tool to assist in the unknown identification of organic compounds provided the RI of the unknown compound is available. NIST has updated and maintained a database of measured RI values for almost 140,000 compounds. In addition, NIST has recently developed an artificial intelligence (AI) model which can calculate RI from the chemical structure with high accuracy, providing RI values for nearly all of the over 300,000 compounds in the Electron Ionization Mass Spectrometry (EI-MS) database. To correctly calculate RI values for unknowns, one must carefully build a calibration with known compounds, usually an n-alkane ladder, under conditions identical to the sample run. As the RI values will be calculated by interpolation between the RI standards, it is optimal to have the standards spaced relatively evenly and close together across the chromatographic run. Internal standards are sometimes used but typically provide limited coverage and are prone to interference with more complex samples.

Here we introduce a new method for automatically calibrating RI (Auto-RI). This method uses the NIST forward and reverse search plus accurate mass to identify those compounds in the sample run with the highest confidence. The library RI for each of these compounds is then used to create an RI calibration curve for the run. Once created, possible miss-identified compounds can be discovered by any significant deviation from the RI calibration curve, eliminated, and then a new more accurate calibration is calculated (Figure 1). This method can be very accurate and comparable to external calibration methods in most sample runs.

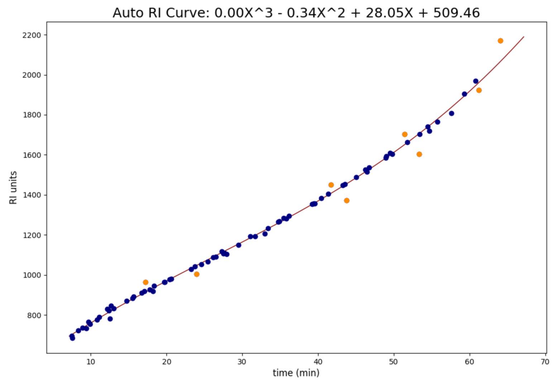

Figure 1. An Auto-RI curve generated from a complex, thermally degraded sample. The blue dots represent compounds identified with high confidence using forward and reverse search. The orange dots represent compounds with large standard deviation in retention index and are eliminated to attain a more accurate RI calibration.

Magic Highlighter

Providing additional metrics for compound ID can significantly improve the accuracy of unknown Identification. However, the process of confirming the compound ID means digesting a lot of data, which can be tedious and time consuming. To accelerate the process we can blend the five metrics (forward search, reverse search, Retention Index and accurate mass Formula ID for both molecular and fragment ions) to produce a confidence value for each compound on the "Hit" list. These can be categorized as High, Medium, and Low confidence. The peaks identified are then color coded to Blue (High), Yellow (Medium), and Red (Low) confidence. This allows the user to easily visualize the results and focus on compounds that need more careful review (Yellow and Red). Table 1 shows a summary of the best search hits for each peak in a run of a plant based meat product. The user can easily ID results that are problematic and click to review the detailed search Hit list.

Foods using Auto RI and Accurate Mass

Peak	Time	Quant	Mix	Lib	NIST	RevS	SA-M	SA-F	RIFit	RI	RILib	MW	Formula	Name
1	7.84	489977.44	1/1	NM+C	949	953	99.53	99.73	95.20	936.80	932	136.125	C10H16	(1R)-2,6,6-Trimethylbicyclo[3.1.1]hept-2-e
2	8.82	387965.78	1/1	NM+C	945	949	99.50	99.65	99.20	974.80	974	136.125	C10H16	Bicyclo[3.1.0]hexane, 4-methylene-1-(1-met
3	8.88	624171.25	1/1	NM+C	941	942	99.44	99.79	95.28	977.28	982	136.125	C10H16	Bicyclo[3.1.1]heptane, 6,6-dimethyl-2-meth
4	8.91	64706.17	1/2	NM+C	935	942	98.50	99.75	96.78	978.78	982	136.125	C10H16	Bicyclo[3.1.1]heptane, 6,6-dimethyl-2-meth
5	8.97	51315.36	2/2	NM+C	940	940	-78.88	99.65	98.71	981.29	980	128.120	C8H16O	1-Octen-3-ol
6	9.22	743908.50	1/1	NM+C	933	946	97.81	99.77	98.97	992.03	991	136.125	C10H16	á-Myrcene
7	9.32	539856.06	1/1	NM+C	926	935	-94.08	99.71	96.53	996.47	993	130.136	C8H18O	3-Octanol
8	9.76	321430.44	1/1	NM+C	931	946	99.57	99.75	99.39	1016.39	1017	136.125	C10H16	1,3-Cyclohexadiene, 1-methyl-4-(1-methylet
9	9.94	401759.00	1/1+1	NM+C	946	946	99.70	99.66	97.53	1024.47	1022	134.110	C10H14	o-Cymene
10	10.03	614093.25	1/1+1	NM+C	952	952	99.61	99.85	97.70	1028.70	1031	136.125	C10H16	D-Limonene
11	10.04	-983896.94	1/2	NM+C	863	863	24.92	95.21	98.55	1029.55	1031	136.125	C10H16	D-Limonene
12	10.09	4378477.50	2/2	NM+C	937	937	99.25	99.64	99.55	1031.55	1032	154.136	C10H18O	Eucalyptol
13	10.21	502621.88	1/1	NM+C	943	944	98.37	99.56	88.58	1037.58	1049	136.125	C10H16	trans-á-Ocimene
14	10.41	120735.48	1/1+1	NM+C	943	943	99.30	99.81	90.50	1047.50	1038	136.125	C10H16	1,3,6-Octatriene, 3,7-dimethyl-, (Z)-
15	10.63	521796.66	1/1	NM+C	962	963	99.60	99.66	98.23	1058.23	1060	136.125	C10H16	ç-Terpinene
16	10.80	1308363.12	1/1	NM+C	949	950	99.57	99.73	92.03	1067.03	1075	154.136	C10H18O	5-Isopropyl-2-methylbicyclo[3.1.0]hexan-2-
17	11.19	520186.09	1/1	NM+C	656	816	-68.20	99.82	0.00	1087.59	1624	236.178	C15H24O2	Pinocamphyl angelate, iso-
18	11.39	441010.03	1/1	NM+C	894	896	-98.09	99.43	99.46	1098.46	1099	154.136	C10H18O	Linalool
19	11.53	116981.83	1/1	NM+C	864	911	-90.55	99.44	98.93	1105.93	1107	172.146	C10H20O2	Butanoic acid, 3-methyl-, 2-methylbutyl es
20	11.80	105396.55	1/1	NM+C	839	865	98.77	98.53	81.11	1121.11	1140	154.136	C10H18O	2-Cyclohexen-1-ol, 1-methyl-4-(1-methyleth
21	12.24	285938.78	1/1	NM+C	924	934	98.84	99.19	64.90	1145.90	1181	154.136	C10H18O	(1R,2R,5S)-5-Methyl-2-(prop-1-en-2-yl)cycl
22	12.43	5798956.00	1/2	NM+C	758	760	19.85	92.44	93.03	1157.03	1164	154.136	C10H18O	Cyclohexanone, 5-methyl-2-(1-methylethyl)-
23	12.43	10385700.00	2/2	NM+C	824	840	90.04	26.16	99.97	1157.03	1157	154.136	C10H18O	Cyclohexanone, 5-methyl-2-(1-methylethyl)-

Table 1. A list of the top NIST match for each compound in a run. Note the metrics NIST = forward search, RevS = reverse search, SA-M is the spectral accuracy for the hit formula, RI = peak RI, RILib is the compound RI and RIFit is 100 minus the difference.

For example, Peak 21 in Table 1 is marked Red (low confidence). Clicking on the peak table row for Peak 21 brings up the detailed Hit list as shown in Figure 2. The top hit (sorted by forward search) is highlighted "Red". Even though the forward search has the highest score, the retention index match (RIFit with 100 being a perfect RI match) is poor at 64.9, making the hit highly suspect. The second hit is highlighted "Blue" at RIFit = 82.9, but hits 4-7 (as we are including the NIST replica library) not only have excellent forward and reverse search values but also an almost perfect RIFit at 99.9. The analyst can now promote the correct compound as the top hit along with any relevant notes.

Note that in this example, all the top hits have the same molecular formula, a common theme in natural products, and the spectral accuracy value that confirms the molecular formula does not help differentiate the isomers. In this case only RI matching can help confirm the correct hit.

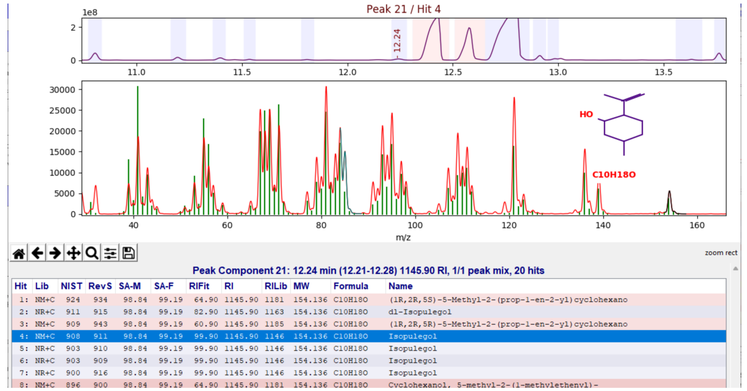

Figure 2. Color coded hit list for peak 21 show a bad 1st hit as the RIFit is over 35 iu off. The 2nd hit is reasonable but hits 4-7 are more confident due to the excellent RIFit among the group of replica hits.

Conclusion

The combined metrics of forward and reverse search, RI, and accurate mass formula ID can significantly improve the accuracy and confidence in identifying unknown compounds by GC/MS. The additional burden of performing the accurate calibration of the GC to obtain RI data can be eliminated in many cases by using compounds in the run that exhibit both highly confident search results and are highly correlated with a RI trend line (Auto-RI).

However, there is still a burden on the analyst to review and validate the results. By using a combined weighting metric and color coding the hits, the Analyst can quickly identify which compounds need more detailed review. The review process is further accelerated by highlighting the most likely hit in the hit list which oftentimes is not the compound with highest library search score.

References

1. Y. Wang and S. Simonoff. ASMS 2018 TP816.
2. Y. Wang, M. Gu., Anal. Chem., **2010**, *82*, 7055.