Beyond monoisotopic accurate mass spectrometry: ancillary techniques for identifying unknown features in non-targeted discovery analysis

Joachim D Pleil, M Ariel Geer Wallace and James McCord
Exposure Methods and Measurements Division, National Exposure Research Laboratory, US Environmental Protection Agency, NC 27711, United States of America

1 Author to whom any correspondence should be addressed.
E-mail: pleil.joachim@epa.gov

Keywords: accurate mass, monoisotopic mass, compound identification, chemical features, high resolution MS, isotope ratios, mass fragments

Abstract
High-resolution mass spectrometry (HR-MS) is an important tool for performing non-targeted analysis for investigating complex organic mixtures in human or environmental media. This perspective demonstrates HR-MS compound identification strategies using atom counting, isotope ratios, and fragmentation pattern analysis based on ‘exact’ or ‘accurate’ mass, which allows analytical distinction among mass fragments with the same integer mass, but with different atomic constituents of the original molecules. Herein, HR-MS technology is shown to narrow down the identity of unknown compounds for specific examples, and ultimately inform future analyses when these compounds reoccur. Although HR-MS is important for all biological media, this is particularly critical for new methods and instrumentation invoking exhaled breath condensate, particles, and aerosols. In contrast to standard breath gas-phase analyses where 1 mass unit (Da) resolution is generally sufficient, the condensed phase breath media are particularly vulnerable to errors in compound identification because the larger organic non-volatile molecules can form identical integer mass fragments from different atomic constituents which then require high-resolution mass analyses to tell them apart.

Premise
High-resolution mass spectrometry (HR-MS) has introduced an additional dimension for identifying unknown organic compounds in complex mixtures. Termed ‘non-targeted analysis’, this approach is defined as agnostic with respect to analytes; the samples are processed with no preconception as to content and the results are considered to be unidentified chemical ‘features’. These features are then post-processed to assign chemical formulae based on the HR-MS results (Schymanski et al 2015, Sobus et al 2018).

This perspective is intended to explain the value and general methods for HR-MS in investigating complex organic mixtures as in human blood, breath and urine, and demonstrate the value of some compound identification strategies with specific examples. The primary HR-MS parameter, referred to as ‘exact’, or ‘accurate’ mass, allows analytical distinction among mass fragments with the same integer mass, but with different atomic constituents of the original molecules. The most basic discrimination occurs when separating ionic fragments with resolution at the fourth or fifth decimal place. This is a vast improvement over standard single-digit mass units (integer mass) instrumentation. The underlying concepts and physics principles of exact mass analysis have been described in detail in a predecessor article (Pleil and Isaacs 2016). Such ‘monoisotopic’ analysis is only the first, and most basic step. Once features are located and assigned to a chromatographic retention time or retention index, they can be further investigated by their molecular fragmentation patterns, the exact masses of the fragments, and their relative isotopic abundances (McLafferty et al 1999). This perspective...
describes how to exploit new technology for identifying unknown organic constituents within complex matrices, develop reasonable confidence in their identity, and ultimately inform future analyses when these compounds reoccur. Although HR-MS is crucial for all biological media, this is particularly important for new methods and instrumentation invoking exhaled breath condensate, particles, and aerosols as a diagnostic biological medium complementing blood and urine analysis (Ladva et al 2000, Zamuruyev et al 2016, Ghio et al 2017, Sauvain et al 2017, Winters et al 2017, Wallace and Pleil 2018a, 2018b). In contrast to standard breath gas-phase analyses where 1 Da resolution is generally sufficient, the condensed phase media are particularly vulnerable to errors in identification as the larger organic molecules contained therein have more possibilities for forming integer mass fragments.

Various forms of HR-MS are now being employed in new breath applications. A cursory search of recent articles finds that real-time and gas chromatographic (GC) instruments are employing time-of-flight (ToF) as a replacement for linear quadrupole detectors to improve discrimination of ionic fragments and that liquid chromatography HR-MS applications using ToF, MS-MS and Orbitrap instruments are becoming more prevalent (Herbig et al 2009, Sukul et al 2016, Nizio et al 2016, Peralbo-Molina et al 2017, Andra et al 2017, Li et al 2017, Bregy et al 2018, Singh et al 2018).

Overview

As discussed in recent conceptual articles, chemical toxicity testing and human disease diagnosis have evolved beyond simple targeted analysis of chemicals of exposure, their chemical biomarkers, and certain endogenous response metabolites (Krewski et al 2010, Ala-Korpela et al 2012, Teeguarden et al 2016, Vineis et al 2017). Basically, samples are analyzed for as many compounds as feasible within a particular laboratory’s capability, and subsequently subjected to statistical analysis to identify ‘features’ characteristic of the behavior under investigation. The simplest version of this process is designating ‘case-control’ pairs, where one sample is treated or exposed in some fashion. At this point, the results are compared and unknowns are differentiated between cases and controls by identity, relative concentration, or correlation and are subsequently further explored. The methodology has been implemented for a wide range of sample types including human blood, breath, and urine, as well as for in vitro systems investigating chemical changes in cell lines, tissue biopsies, and bacterial/fungal cultures (Vorst et al 2005, Aura et al 2008, Croley et al 2012, Kerian et al 2015).

Until recently, discrimination among sample groups has relied on targeted compounds, or those that were readily identified by existing methodologies. Data post-processing for such targeted experiments is relatively straightforward, and has for the most part been streamlined with computational tools. The advent of HR-MS, along with extraordinary advances in sensitivity, has resulted in an explosion of available data, and a concomitant burden on researchers in deciphering their meaning. Currently, the newest non-targeted (discovery) analyses require detailed supervision from subject matter experts to provide defensible results. A non-targeted experiment refers to one aiming to observe as many chemicals as possible from a complex sample mixture; in practice complete detection of all chemicals in a mixture is not possible with a single technique. Non-targeted approaches allow for sophisticated data mining and multivariate analysis to tease apart individual compounds associated with sample groups, but appropriate choice of techniques and experimental design is non-trivial. The state-of-the-art technology for interpreting HR-MS data is still under development; instrument manufacturers each have proprietary software/firmware and numerous open-source data analysis packages exist, which further complicates the task of creating a single data analysis workflow.

As HR-MS technologies are becoming more commonplace in the analytical laboratory, we have developed some general guidance for the analyst community as to how to implement compound identification techniques beyond rudimentary library searches of exact match candidates. Specifically, we present a series of examples where uncertainty is reduced for assigning chemical structure and formula by implementing exact mass fragments and isotope ratios. We further suggest how software products could assist in automating aspects of decision-making for identifying unknown chemicals.

High-resolution mass spectrometry: what is ‘high-res’?

When comparing mass spectrometry instrumentation, there are several critical parameters of merit. The most central is the mass resolving power of the platform defined as the ability to successfully distinguish two closely separated masses. The International Union of Pure and Applied Chemistry (IUPAC) definition of MS resolution (R) is $R = M/\Delta M$ where $M$ is the measured mass, and $\Delta M$ is the separation required to distinguish two peaks at a certain height from the baseline, similar to the definition of peak resolution in chromatography (IUPAC 1997). An instrument might be discussed in terms of the minimum separation required to resolve two equal height mass peaks (e.g. mass resolution of 0.001 Da at 250 Da). IUPAC further defines a single peak measurement methodology, also called mass resolving power, as $M/\Delta M$ where $\Delta M$ is the peak width at a specified height (figure 1). Because resolving power is an instrument performance parameter, it is the most
commonly quoted value, and is traditionally measured at the full width at half maximum height (FWHM) of the MS peak (figure 1). Mass resolution can be measured at any degree of peak separation but most frequently at 10% of the maximum peak height, equivalent to the full peak width at 5% peak height for an isolated gaussian peak. It is worth noting that in common usage the terms resolution and resolving power are used interchangeably but should always specify the means of determination and the target mass (e.g. $R_{\text{FWHM}} @ m/z 200 = 100 000$) because resolution/resolving power values are dependent on the mass measured and the height of peak measurement. For the remainder of this manuscript R will refer to FWHM resolving power at the discussed mass.

An example calculation for theoretical peak pairs is shown in figure 1. A peak with a mass of 400.0000 and an observed peak width at half height of 0.5 Da has a single peak resolving power calculated as:

$$R_{\text{FWHM}} = \frac{M}{\Delta M} = \frac{400}{0.5} = 800.$$ 

The resolution of this peak from another peak at 401.0000 at 10% of the chromatographic peak width is 1 Da and has an R value calculated as:

$$R_{10\%} = \frac{m_1}{m_2 - m_1} = \frac{400}{401 - 400} = 400.$$ 

At a significantly higher resolving power the peak widths decreased substantially (figure 1, right) and the effective mass resolution likewise decreased (0.02 Da for the figure shown).

Given the incremental progress of mass spectrometry instrumentation over the years, there is a constantly moving goalpost for describing when an instrument or spectrum is 'high-resolution.' This is further complicated by instrument manufacturers changing target masses for quoted resolution, and intermixed usage of FWHM and 10% valley definitions. Nevertheless, there are broad thresholds of resolution at which increasing amount of chemical information can be gained. (Marshall et al 2002, Marshall and Hendrickson 2008) At $R \sim 1000$ nominal masses are distinguishable, (i.e., isotopic peaks of a single spectrum) and at $R \sim 100 000$ isotopologues differing only in the presence of nominally identical isotopes begin to separate (e.g., $^{13}\text{C}_2\text{H}_4^{14}\text{NH}_3$ and $^{12}\text{C}_3^{15}\text{NH}_3$, see figure 2).

For the purposes of this commentary, any measurements capable of resolving the isotopic fingerprint of molecules can apply the compound identification strategies discussed. We also note that the mass described by significant digits beyond integer mass may be referred to as ‘exact mass’ or as ‘accurate mass’. In general, ‘exact mass’ is the fragment mass calculated from a known chemical formula, whereas the term ‘accurate mass’ refers to a measurement with high precision; however, they are used interchangeably for the purposes of identifying compounds.

High-resolution mass spectrometry: a brief history

Exact (or accurate) mass spectrometers are not new; they trace their origins to ‘one of a kind’ cyclotron and magnetic sector instruments at major research centers that were used to separate inorganic radio-isotopes

Figure 1. Mass resolution definitions. Simulated FWHM and 10% valley $\Delta M$ measurements for a pair of isotopes at $R_{\text{FWHM}} \sim 1000$ (left) and $R_{\text{FWHM}} \sim 40 000$ (right).
and organic molecules (e.g., Beynon 1956, Beynon 1959, Henning et al. 1981, De Laeter and Kurz 2006). A timeline of the history of exact mass measurement is available from the archives of the American Society for Mass Spectrometry; the Society attributes the initial achievements of high resolving power mass spectrometry to E O Lawrence and M S Livingston in 1932, and lists some of the major technical advances for Time of Flight (ToF-MS) in 1956, double-magnetic sector geometries from 1957 to 1960, and Fourier transform ion cyclotron resonance from 1965 to 1968 (Grayson 2008).

The history of commercial HR-MS products began in the 1940's. The Chemical Electrodynamics Corporation entered the MS commercial market in 1943 with the 21-101 Mass Spectrometer based on magnetic sector technology; it was used to assess petroleum hydrocarbons and had an estimated mass resolution of 0.05 Da at 250 Da (R ~ 4000) (Carlson et al. 1960). The Omegatron, based on vacuum tube technology, was developed by University of Minnesota in 1949 as an MS instrument designed primarily as a residual gas analyzer separating unit Da gases for vacuum applications (Zdanuk et al. 1960); a patent filed in 1957 indicates that it was improved to achieve resolution to ~0.01 Da @ 250 Da (R ~ 25 000) (McNarry and Hobson 1957). Subsequently other instruments entered the commercial HR-MS market including the Bendix ToF-MS (Wiley 1956), the Finnigan MAT Sector in 1978 (Huebschmann 2011), and the Bruker FTICR in 1983, based on research by Mel Comisarow and Alan Marshall (Comisarow and Marshall 1974). These legacy commercial instruments had resolving powers up to 10 000 (~1 ppm).

Contemporary HR-MS systems almost always incorporate a high-performance gas or liquid chromatography platform for analyte separation coupled to a TOF or FT detector. Some exceptions are the use of ionization sources (e.g., MALDI, SELDI, APCI, PTR, etc.) without a separation step that are directly coupled to the MS (Byrdwell 2001, Petricoin and Liotta 2003, Jordan et al. 2009, Gaugg et al. 2016, Ruhaak et al. 2016). High-performance TOF instruments routinely generate resolutions of 10 000+, and research platforms exceeding 40 000, with very fast cycle times for fragmentations scans. While cutting edge research on FTICR continues, producing multi-million resolving power instruments (Hendrickson et al. 2015), an electrostatic trapping FT instrument, the Orbitrap (Makarov 2000), has become the most prevalent FT-MS platform. Orbitrap instruments likewise can perform rapid MS/MS fragmentation with scalable MS and MS/MS resolutions from 10 000 to 100 000+.

The prevalence of these high resolving power instruments, coupled with the power of accurate mass analysis and achievable MS/MS fragmentation duty cycles allows for analysts to achieve an unprecedented degree of information about unknown chemical species.

**Interpreting high-resolution features: isotope ratios**

The power of HR-MS measurement is the ability to resolve combinations of elements that differ in weights much less than $1.67 \times 10^{-27}$ kg, or 1 Da. This enables very fine measurements of the exact mass, from which much information can be extracted (Pleil and Isaacs 2016), but also enables elemental composition analysis based on elemental isotopes.

Each signal in a mass spectrometer is the measurement of a single type of ion composed of a linear combination of elemental isotopes. The mass of a molecule composed entirely of the most abundant stable isotope of each constituent element is known as the mono-isotopic mass (A) and is but one of many masses that...
can be measured for a molecule. Because isotopes of common elements such as carbon, sulfur, and many halogen species occur at appreciable levels in nature, molecules incorporating one or more higher mass stable isotopes are encountered whenever a molecular sample is measured. The variation in the number of neutrons contained in elemental isotopes creates atomic combinations with masses that differ nominally by a single Da, and these peaks are identified based on the nominal mass shift relative to the monoisotopic mass (i.e. $A-1$, $A+1$, $A+2$ etc.). The theoretical spectrum for a given molecule is thus the combination of all the peaks for all the isotopic combinations of atoms making up that molecule. Although rudimentary assignments of an atom can be made using isotope ratios using single Da resolution, it requires high-resolution to confirm which atoms are actually responsible for the isotopic ratio. For example, figure 3 shows the resolved structure for the two different $A+2$ possibilities within the perfluorinated compound $C_{12}H_{12}F_{17}NO_{5}S$, wherein the sulfur $^{34}S$ and one carbon $^{13}C$'s could each contribute to the $A+2$ isotopic peak. Note that with sufficient resolution, one can distinguish the contribution from $^{34}S$ versus that from $^{13}C_2$, which could not be done with single Da resolution:

$$\text{Exact Mass} (^{34}S^{32}S) = 33,967,867,004 - 31,972,071,174 = 1.995,795,83$$

and from $^{13}C_2$:

$$\text{Exact Mass} (^{13}C_2 - ^{12}C_2) = 2 \times (13,003,354,84 - 12,000,000,00) = 2.006,709,68,$$

as shown in the inset centered a 607 Da.

The exact mass and relative abundance of atomic isotopes in nature is well characterized (table 1, Berglund and Wieser 2011). This means that calculating a theoretical mass spectrum is a straightforward, but computationally taxing problem of combinatorics, which has many implemented solutions.
(Valkenborg et al. 2012). For the purposes of molecular formula generation, it is therefore possible to compare empirical chemical spectra against theoretical distributions for molecular formula generation. This is necessary for the assignment of molecules of middling complexity, because even an excellent <1 ppm mass accuracy is often insufficient to uniquely resolve the majority of chemicals currently known (Kind and Fiehn 2006). Note that the average molecular weight that appears in the standard Periodic Table of the Elements and is familiar to most chemists or biologists, is, in-fact, the weighted average of these many isotopic combinations.

The number of formula matching and scoring algorithms is extensive and constantly evolving, thanks in no small part to entries in the yearly Critical Assessment of Small Molecule Identification (CASMI) contest (http://casmi-contest.org/), which specifically invites the development of new formula generation software and techniques. Nevertheless, some of the basics for molecular assignment based on elemental composition are amenable to manual inspection of isotopic patterns and can be useful in the assignment of molecular formulae. It is worth noting that many of these strategies can be applied even to low resolution full-scan data, but this is difficult in complex samples where isotope peaks are convoluted with other compounds.

**Example 1. Carbon Counting**

For a typical organic molecule, the major impact on the A+1 abundance is the amount of carbon contained in the molecule. This is because carbon is generally the most prevalent individual atom in an organic structure, and the relative abundance of Carbon-13 (1.07%) is substantially higher than that from $^2$H (0.1%), $^{15}$N (0.3%), and $^{17}$O (0.03%). Under this enormous simplifying assumption, the abundance of the A+1 peak is simply the probability of having only one $^{13}$C in the molecule. For example, in a molecule with 8 carbon atoms, this can be approximated by the multinomial expansion (Valkenborg et al. 2012) as having a likelihood of 8.3% and allows rough estimation of the carbon content of a molecule. Figure 4 shows the A+1 ion for four different configurations of organic molecule with 8 carbons. Although the relative prevalence of the A+1 varies slightly due to the other atoms, it maintains a value around 8%, indicating the presence of 8 carbon atoms in each chemical regardless of the other atoms. We note that the exact mass difference of 1.003 35 Da helps confirm that the isotope is $^{13}$C, rather than from $^2$H’s, $^{15}$N’s, or $^{17}$O’s.

**Example 2. Sulfur Counting**

As previously mentioned, the use of isotopic abundances can be of substantial value in narrowing the possible molecular formulae for a given mass. For example, a compound with an accurate mass of 307.083 906 has five possible formula matches within 5 ppm, and two within 1 ppm instrument resolution, when searched against the US Environmental Protection Agency’s (EPA) CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard), a publicly available database containing over 760 000 chemicals (Williams et al. 2017). The A+1 peaks for these molecules have theoretical values which closely coincide, as is expected given the similar number of carbon atoms in these molecules. The A + 2 peak however, shows significant variation depending on the inclusion of sulfur or chlorine. Sulfur is common in biomolecules and has an A + 2 isotope with
abundance $\sim 5\%$, meaning that for a small molecule the $A + 2$ peak is similar or higher in abundance to the $A+1$ peak. Given an empirical spectrum with a relative $A+1$ and $A + 2$ abundance of 14% and 6% we could safely conclude this molecule is likely $C_{10}H_{17}N_{3}O_{6}S$, Glutathione as shown by the red curve in figure 5.

**Example 3. Halogen Counting**

Several halogens offer very distinct isotopic patterns that are apparent to the naked eye. Both Bromine and Chlorine have major $A + 2$ isotopes with large abundances at $\sim 25\%$ for $^{37}\text{Cl}$ and $\sim 50\%$ for $^{81}\text{Br}$. Consequently, an $A + 2$ peak of substantial

![Figure 5](image-url)

**Figure 5.** Theoretical mass spectral patterns ($R \sim 15\,000$) for five molecular formulae with exact mass 307.083 906 (within 5ppm instrument resolution). Insets show the $A+1$ and $A + 2$ isotope peaks. In the second extracted panel, the relative abundance of the $A + 2$ peaks for the five candidate compounds range from 2% to 30%. The 30% value indicates chlorine, and the entries below 3% eliminate sulfur (at 4.25%). Given that the true empirical spectrum has $\sim 6\% A + 2$, and that sulfur and oxygen have $A + 2$ abundances of 4.25% and 0.25%, respectively, the relative $A + 2$ abundance for ($O_6 + S$) should match 5.75%. Furthermore, the spacing between sulfur $A$ and $A + 2$ is 1.995 795 83 whereas the shift for oxygen is 2.004 244 99, so the slight shift of the red trace confirms the sulfur atom, and the confirmed candidate is $C_{10}H_{17}N_{3}O_{6}S$, Glutathione.

![Figure 6](image-url)

**Figure 6.** Using isotopic abundance patterns to count the number of chlorine atoms in a $C_{12}$ compound: at each individual mass fragment, exact mass can be further investigated to confirm the chemical formula. Chlorine has an abundance of $A + 2$ isotopes of $\sim 25\%$, so a spacing at 1.997 375 78 Da with relative abundance of $\sim 25\%$ indicates 1 chlorine, 63% indicates 2 chlorines, 100% indicates 3 chlorines, etc. The spacing of $A+1$ at 1.003 354 84 Da, and the relative abundance at $\sim 12\%$ indicates 12 carbons.
magnitude can be representative of a compound containing one or more Cl or Br species. Inclusion of multiple halogen atoms yields complex splitting patterns in the MS spectrum, with spacing every two Daltons (figures 6, 7).

**Interpreting high-resolution features: fragmentation patterns**

When organic compounds are fragmented, whether in-source by electron ionization (EI) or chemical ionization, or through intentional MS/MS, it is possible to form charged fragments, each with their characteristic isotopic features. Even at low resolution, the patterns of these fragments serve as an additional dimension for identifying compounds, especially when the molecular ion is missing or uncertain. High resolution MS can help distinguish between multiple possibilities for the atomic composition of a particular fragment just as easily as a molecular species. For example, a mass fragment at 85 Da could be either C₆H₁₃ or C₅H₉O at 1 Da resolution. However, at high resolution the quandary is resolved, as the accurate masses 85.101725 and 85.065339, respectively, are easily separated.

Much like the isotopic mass patterns discussed in the previous section, accurate mass of fragments serves as identifiers of molecular substructures. For larger molecules with many fragments, this allows a plausible, but complicated process to reassemble the original chemical structure. Instead, patterns are most frequently compared against library spectra to generate a list of potential structures and/or substructures from the detected fragments. A standard library search (at unit resolution) generally gives a long list of potential chemical formulae, but high-resolution fragment spectra avoids many assignment issues regardless of the fragmentation method or structural assignment approach.

**Example 1. Precursor discrimination in MS/MS**

MS analysis with a triple-quadrupole mass spectrometer (QqQ) requires the ionization and selection of a precursor molecule with a single quadrupole, which is passed into a second collision cell and fragmented, before fragments are isolated by a third and final quadrupole section. The process of selecting both precursor and fragment at a known collision energy is intended to be very specific, even with low resolution quadrupoles. Nevertheless, false positives can occur between compounds of similar precursor mass with non-specific fragment transitions.

Consider the following three chemicals, taurodeoxycholic acid (TDCA), perfluorooctanesulfonic acid (PFOS), and 8-(acetyloxy)-1,3,6-pyrenesulfonic acid (1,3,6-PSA) which might coelute in food and fish samples contaminated with fluorinated compounds (table 2).

Each of the three compounds exhibit a similar precursor mass within a 1 Da isolation window and share a common mass fragment transition at 499 > 80 due to the production of an SO₃⁻ ion fragment. The QqQ
transition is thus non-specific at low resolution and requires further comparison to reference standards or the inclusion of secondary transitions to ensure the identity of the species and the purity of the transition observed. Using a higher resolution instrument, the transitions are obviously distinct. PFOS and TDCA transitions of 498.9 > 80.0 and 498.2 > 80.0 are resolvable at R ∼ 1000, while PFOS and 1,3,6-PSA require a slightly higher resolution of R ∼ 30 000 to distinguish the respective transitions of 498.9375 > 79.9568 and 498.9463 > 79.9568. Note that a typical HR-MS instrument still uses a low resolution quadrupole for precursor isolation, so coeluting species would not have isolated precursors, but accurate assignment of distinct precursor masses in a precursor scan allows for recognition of false assignments.

Example 2. Substructure Fragments

Distinguishing sub-structural fragments of an unknown compound allows for unequivocal
identification of the structure. Consider a compound with the monoisotopic mass 270.0892, assigned empirical formula of C_{16}H_{14}O_{4}, and an EI fragment spectrum as shown in figure 8.

A search of the US EPA CompTox Chemistry Dashboard yields 228 chemicals with this empirical formula, but significant variation in structure that can be investigated by fragmentation. Even at low resolution, the observable 193 fragment corresponds to the loss of a phenyl ring (loss of 77), allowing significant narrowing of the search space. The 139 fragment is more difficult to assign given only a low-resolution spectrum, as it could correspond to the loss of either C_{9}H_{7}O (131.0497) or C_{10}H_{11} (131.0861). With high resolution mass measurement, the exact mass differential of 131.0492 is consistent with the C_{9}H_{7}O loss. The fragmentation is thus consistent with a chalcone backbone with dihydroxy and methoxy functionalization on a single phenyl ring, such as Cardamomin shown in figure 8. Further details on structural elucidation of complex structures based on MS/MS are beyond the scope of this manuscript, but comparison with reference spectra allows for validation even in the absence of sufficient expertise for de novo elucidation.

**Interpreting high-resolution features: automated data reduction and high-resolution search algorithms**

So far, the identifications schemes have been implemented manually, that is, each analytical feature was investigated individually using the expertise of the researcher with assistance from mass look-up tables and database searches. With the advent of higher sensitivity instrumentation, we are now faced with thousands of features per sample for hundreds or more samples from any given study. As such, the major HR-MS manufacturers, such as Agilent Technologies (Santa Clara, CA, USA), ThermoFisher Scientific (Waltham, MA, USA), SCIEX (Framingham, MA, USA), LECO (Saint Joseph, MI, USA), Shimadzu (Nakagyo-ku, Kyoto, Japan), and Waters (Milford, MA, USA), have all developed their own proprietary software specific for their platforms to automate identifications of unknowns. In addition, academic researchers have been working on more generic algorithms that can receive input from different instrumental data streams. Regardless of the exact implementation, the overarching goal is to employ the subject matter expertise described above in an automated fashion to expedite compound identification.

Briefly, some of the major software packages for feature extraction, compound identification and annotation, as well as statistical analysis and visualization include the following:

- MassHunter Profinder (Agilent Technologies)
- Mass Profiler Professional (Agilent Technologies)
- Unknowns Analysis (Agilent Technologies)
- BioConfirm Software (Agilent Technologies)
- Mass Frontier Spectral Interpretation Software (ThermoFisher Scientific)
- Compound Discoverer Software (small molecule identification) (ThermoFisher Scientific)
- TargetQuan3 Software (identify persistent organic pollutants) (ThermoFisher Scientific)
- ToxFinder (ThermoFisher Scientific)
- PeakView Software (SCIEX)
- MarkerView Software (SCIEX)
- XCMSPlus Software (SCIEX)
- ChromaTOF (LECO)
- LabSolutions Insight (Shimadzu)
- ChromaLynx (Waters)
- MarkerLynx (Waters).

Each vendor has their own preferred software packages for data reduction and analysis, and no two packages provide all of the same features and capabilities. Software selections depend upon the instrumentation being used for analysis as well as the needs of the researcher. Additionally, computer-based spectral fragmentation software methods, including CFM-ID, MAGMA, and MaxQuant are available to assist with data analysis (Cox and Mann 2008, Allen et al 2014). MaxQuant consists of a series of algorithms that can be used for peak detection and quantification as well as identification of fragmentation spectra (Cox and Mann 2008). CFM-ID is a web server that can be used for peak annotation, spectra prediction, and metabolite identification (Allen et al 2014). XCMS software is also now available as an online tool that can be used to analyze non-targeted LC/MS metabolomics data through feature detection, retention time correction, and alignment as well as providing a platform for statistical analyses and data visualization tools (Tautenhahn et al 2012).

General mass spectral libraries such as the National Institute of Standards and Technology library can be used to search non-targeted mass spectral data. Additional databases, such as the U.S. EPA CompTox Chemistry Dashboard can be implemented into the data processing workflow to help characterize identified compounds of interest. Compounds can be searched by chemical name, CASRN, DSSTox ID, MS-ready formulae, exact formulae, monoisotopic mass, or InChIKey. A list of tentative compound identifications based on the results of MS data post-processing can be retrieved from the Dashboard. Additional compound information, including presence in lists, number of data sources, National Health and Nutrition
Examination Survey predicted exposure, and number of PubMed articles can also be downloaded from this site, as well as specific properties of compounds of interest. Online databases, including the U.S. EPA CompTox Chemistry Dashboard, US EPA DSSTox, and ChemSpider can be used to rank order unknown compounds based on monoisotopic masses and chemical formulae as well as retrieve important chemical identification information, properties, and structural images (Rager et al 2016, McEachran et al 2017, Sobus et al 2017). Metabolomics databases, such as the human metabolome database (http://hmdb.ca/), METLIN (https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage), and the Golm Metabolome Database (http://gmd.mpimp-golm.mpg.de/) can be used to search high resolution tandem MS/MS, GC/MS, and LC/MS spectra for known and unknown human metabolites (Rathahao-Paris et al 2016).

Summary and recommendations

Accurate (or exact) mass from HR-MS brings an additional dimension to the identification of complex organic molecules. The use of monoisotopic exact mass alone is a great improvement over standard single Da unit resolution, and often narrows compound identity from hundreds of possibilities to a dozen or so. Herein, we discussed the value of going beyond this single monoisotopic focus and incorporated additional knowledge to narrow the possibilities even further.

There are two basic techniques that help confirm the suspected identities of unknown features based on their monoisotopic mass library search:

- count major constituent atoms such as carbon, sulfur, or halogens based on their isotopic abundances and exact mass differences between isotopes;
- identify the various molecular fragments from ionization by their exact mass and gain additional information to reconstruct the original molecule based on its building blocks.

These methods can be implemented on a molecule-by-molecule basis by the researcher and are generally a first approach for initial evaluation of analytical features thought to be of importance. For example, in a case-control evaluation, a handful of analytical features thought to be of importance. For example, in case-control evaluation, a handful of features may be differentiated between the groups, and a hands-on identification effort is called for.

However, this is not practical for processing hundreds (possibly thousands) of features from many samples. As described briefly, efficient data reduction requires specialized software. There are many entries in the mass spectrometry software arena that are constantly being updated and improved. The primary focus currently is building reference data used to confirm the calculations based on empirical spectra.

As HR-MS instruments improve in sensitivity and mass accuracy, the combination of exact mass, knowledge of relative isotope abundances and the increasing size of confirmatory databases will continue to improve identification of unknowns in complex biological samples. The primary importance of this process is to document and develop subsets of biological exposome data as possible to develop a comprehensive understanding of human systems biology. As we better understand the connections between health-related stressors and the human response through investigations of subtle biochemical perturbations, we will be able to develop interventions to preserve public health.

Acknowledgments

The authors are grateful for expert advice from Mark Strynar, Seth Newton, and Jon Sobus of US EPA. This article was reviewed in accordance with the policies of the National Exposure Research Laboratory, US Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Theoretical spectra presented throughout the manuscript were prepared using the enviPat in R (DOE: 10.1021/acs.analchem.5b00941) at the presented resolving power.

ORCID iDs

Joachim D Pleil @ https://orcid.org/0000-0001-8211-0796
M Ariel Geer Wallace @ https://orcid.org/0000-0002-3966-5101
James McCord @ https://orcid.org/0000-0002-1780-4916

References

Ala-Korpela M, Kangas AJ and Soininen P 2012 Quantitative high-throughput metabolomics: a new era in epidemiology and genetics Genome Medicine 4 36

Allen F, Pon A, Wilson M, Greiner R and Wishart D 2014 CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra Nucleic Acids Research 42 W94–9

Andra S S, Austin C, Patel D, Dolios G, Awawda M and Arora M 2017 Trends in the application of high-resolution mass spectrometry for human biomonitoring: an analytical primer to studying the environmental chemical space of the human exposome Environment International 100 52–61

Aura A M, Mattila I, Seppälä-Laakso T, Miettinen J, Oksman-Caldentey K M and Oreil M 2008 Microbial metabolism of catechin stereoisomers by human faecal microbiota: comparison of targeted analysis and a non-targeted metabolomics method Phytochemistry Letters 1 18–22

Berglund M and Wieser M 2011 Isotopic compositions of the elements (IUPAC technical report) Pure Appl. Chem. 83 397–410
Singh K D, del Miguel G V, Gaugg M T, Ibañez A J, Zenobi R, Kohler M, Frey U and Sinues P M 2018 Translating secondary electrospray ionization–high-resolution mass spectrometry to the clinical environment J. Breath Res. 12 027113
Sobus J R et al 2018 Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA Journal of Exposure Science & Environmental Epidemiology 28 411–26
Sukul P, Treff P, Kamysek S, Schubert J K and Miekisch W 2015 Instant effects of changing body positions on compositions of exhaled breath J. Breath Res. 9 047105
Tautenhahn R, Patti G J, Rinehart D and Siuzdak G 2012 XCMS Online: a web-based platform to process untargeted metabolomic data Anal. Chem. 84 5035–9
Teeguarden J G et al 2016 Completing the link between exposure science and toxicology for improved environmental health decision making: The aggregate exposure pathway framework Environ. Sci. Technol. 50 4579–86
Valkenborg D, Mertens I, Lemière F, Witters E and Burzykowski T 2012 The isotopic distribution conundrum Mass Spectrometry Reviews 31 96–109
Vineis P et al 2017 The exposome in practice: design of the EXPOsOMICS project International Journal of Hygiene and Environmental Health 220 142–51
Vorst O, De Vos C H, Lommen A, Staps R V, Visser R G, Bino R J and Hall R D 2005 A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles Metabolomics 1 169–80
Wallace M A and Pleil J D 2018a Evolution of clinical and environmental health applications of exhaled breath research: review of methods and instrumentation for gas-phase, condensate, and aerosols Analytica Chemica Acta 1024 18–38
Wallace M A and Pleil J D 2018b Dataset of breath research manuscripts curated using PubMed search strings from 1995–2016 Data in Brief 18 1711–24
Wiley W C 1956 Bendix time-of-flight mass spectrometer Science 124 817–20
Williams A J et al 2017 Journal of Cheminformatics 9 61
Winters B R, Pleil J D, Angrish M M, Stiegel M A, Risby T H and Madden M C 2017 Standardization of the collection of exhaled breath condensate and exhaled breath aerosol using a feedback regulated sampling device J. Breath Res. 11 047107
Zamuruyev K O et al 2016 Human breath metabolomics using an optimized non-invasive exhaled breath condensate sampler J. Breath Res. 11 016001
Zdanuk E J, Bierig R, Rubin L G and Wolsky S P 1960 An omegatron spectrometer, its characteristics and application Vacuum 10 362–9