J. Am. Soc. Mass Spectrom. (2011) 22:348–359
DOI: 10.1007/s13361-010-0034-3

RESEARCH ARTICLE

Identification of "Known Unknowns" Utilizing Accurate Mass Data and Chemical Abstracts Service Databases

James L. Little,[1] Curtis D. Cleven,[1] Stacy D. Brown[2]

[1]Building 150, Eastman Chemical Company, Kingsport, TN 37662, USA
[2]East Tennessee State University, College of Pharmacy, Johnson City, 37614, USA

Abstract

In many cases, an unknown to an investigator is actually known in the chemical literature. We refer to these types of compounds as "known unknowns." Chemical Abstracts Service (CAS) Registry is a particularly good source of these substances as it contains over 54 million entries. Accurate mass measurements can be used to query the CAS Registry by either molecular formulae or average molecular weights. Searching the database by the web-based version of SciFinder is the preferred approach when molecular formulae are available. However, if a definitive molecular formula cannot be ascertained, searching the database with STN Express by average molecular weights is a viable alternative. The results from either approach are refined by employing the number of associated references or minimal sample history as orthogonal filters. These approaches were shown to be successful in identifying "known unknowns" noted in LC-MS and even GC-MS analyses in our laboratory. In addition, they were demonstrated in the identification of a variety of compounds of interest to others.

Key words: Accurate mass spectrometry, SciFinder, CAS Registry, STN Express, Known unknowns, Identification, LC-MS, GC-MS, Electrospray ionization, ESI, Chemical Abstracts Service

Introduction

Accurate mass measurement employing a high resolution mass spectrometer is a powerful approach for the characterization of organic compounds. We define the classes of compounds to be characterized as either "known knowns," "known unknowns," or "unknown unknowns." These terms originated from a quote by Donald Rumsfeld, Secretary of Defense, with regard to weapons of mass destruction in Iraq [Department of Defense News Briefing, Feb. 12, 2002]:

> ". . . there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones."

Though in a much different context, these terms aptly describe the classes of organic compounds to be characterized in mixtures by hyphenated techniques such as liquid chromatography-mass spectrometry (LC-MS)

Electronic supplementary material The online version of this article (doi:10.1007/s13361-010-0034-3) contains supplementary material, which is available to authorized users.

Correspondence to: James L. Little; *e-mail:* jameslittle@eastman.com

Received: 24 September 2010
Revised: 6 November 2010
Accepted: 9 November 2010
Published online: 28 January 2011

and gas chromatography-mass spectrometry (GC-MS). For example, we refer to a compound suspected to be present in a mixture whose identity is to be confirmed by mass spectrometric analyses as a "known known." A "known unknown" is a compound that is unknown to the investigator, but is cited in the chemical literature or mass spectrometry reference databases. Lastly, an "unknown unknown" is a compound that is not previously cited. The approaches for characterizing these three classes of components are much different.

The most difficult ones are "unknown unknowns." Their identification normally requires extensive sample history and/or further characterization by techniques such as nuclear magnetic resonance (NMR) or infrared (IR) spectroscopies.

Conversely, "known unknowns" can be routinely identified by mass spectrometry data with minimal sample history. The most successful approach, by far, is electron impact (EI) GC-MS employing computer-searchable reference databases. EI mass spectra are very reproducible and a large combined version, containing both the National Institute of Standards and Technology (NIST 08) and Wiley 9 collections of ~796,000 spectra and ~670,000 structures, is commercially available. In addition, many specialty databases are available from Wiley for classes of compounds such as flavors, fragrances, designer drugs, pesticides, and volatiles in food. Many private collections of spectra also exist. For example, Eastman Chemical Company has a database of ~51,000 spectra linked to ~30,000 structures collected over the past 32 years.

Many organic compounds are not amenable to GC-MS [1–3] either due to their low volatility, thermal instability, or high polarity. LC-MS is capable of analyzing these types of compounds [3] utilizing soft ionization sources such as electrospray (ESI) in conjunction with a variety of high performance liquid chromatographic (HPLC) separation modes including reversed phase, normal phase, and ion exchange. Examples of compounds analyzed [1–3] include drugs, drug metabolites, pesticides, endocrine disruptors, personal care products, natural toxins, etc.

There has been a phenomenal increase in the availability and use of accurate mass instrumentation for structure elucidation including time-of-flight (TOF), quadrupole TOF (Q-TOF), and Orbitrap mass spectrometers [1, 2]. Unfortunately, the collision-induced dissociation (CID) spectra obtained by ESI and atmospheric pressure chemical ionization (APCI) vary significantly with instrument designs and parameters and are not as reproducible as EI spectra. Thus, the availability of computer searchable CID references databases is somewhat limited [3]. Nevertheless, we find the commercial NIST (~15,000 spectra) database and our corporate database (~3,000 spectra) and their associated structures somewhat useful in characterizing unknowns.

Eastman has extensively employed "spectraless" databases for many years [4, 5], created from Eastman's manufacturing material database and the Toxic Substances Control Act (TSCA) listing, to identify unknowns. The databases included molecular weight (accurate and nominal), molecular formulae, CAS numbers, and chemical names fields. They were searched primarily by molecular weight (MW) or molecular formula, but data evaluation could be very tedious since no orthogonal filters were employed to prioritize the candidate list and structures were not directly associated with the entries.

Others [1–3, 6–8] have also employed "spectraless" databases for the identification of "known unknowns." A wide variety of data sources were utilized including user-created databases, ChemFinder (www.chemfinder.com), ChemSpider (www.chemspider.com), Merck Index, Sigma-Aldrich Catalog, Farm Chemicals Handbook, ChemIndex, Chemical Abstracts Service Database, Beilstein, KEGG (Kyoto Encyclopedia of Genes and Genomes), and others.

Orthogonal filters are very critical in minimizing the number of possible molecular formulae candidates obtained with accurate mass data. It was demonstrated [9] that even at mass accuracies of less than 1 part per million (ppm) or even 0.1 ppm, accurate mass data alone is not adequate to determine unique molecular formulae. Thus, orthogonal filters such as [9–13] isotopic ratio abundances and a variety of heuristic and chemistry rules for constraining molecular formula generators must be employed to limit the result to one or a few molecular formulae. All the major mass spectrometer manufacturers' software packages include various proprietary algorithms for minimizing molecular formulae candidates using a variety of these approaches.

The number of literature citations for compounds [14] was found to be useful in the identification of unknowns in mixtures. The citation count was obtained from the CAS Chemical Substance Index (6 month collective index, printed version). It was shown that high values for literature citations and "cocitations" (patents, literature articles, etc.), used in conjunction with nominal mass EI spectra and retention indices, significantly increased the confidence in the identification of minor impurities in samples of *n*-hexane and naphthalene and of polycyclic aromatic hydrocarbons in waste gas.

We have also employed citations, i.e., references, in a different manner, for the identification of "known unknowns" utilizing computer searches of the CAS Registry and associated databases [15]. The databases are queried by either molecular formula or average molecular data obtained from accurate mass spectrometry data. The initial candidate list is then refined by orthogonal filters, including the number of associated references or minimal sample history. This paper will demonstrate our approach with examples from our laboratory and with compounds of interest to others.

Experimental

Instrumentation

The accurate mass ESI LC-MS data was obtained on a LCT time-of-flight mass spectrometer (Waters Corporation, Milford, MA, USA) equipped with a LockSpray secondary ESI

probe. An Agilent 1100 Series liquid chromatograph, autosampler, degasser, and ultraviolet/visible (UV-VIS) diode array spectrophotometer (Agilent Technologies, Santa Clara, CA, USA) were employed for the analyses at a flow rate of 1.5 mL/min.

A Waters 510 pump was used for post-column addition of various reagents at a flow rate of 0.1 mL/min via a stainless steel tee and PEEK (polyetheretherketone) tubing. The total flow to the ESI primary probe was decreased to 0.120 mL/min employing a stainless steel tee to perform the split with 0.005 in. PEEK tubing to the ESI probe and 0.020 in. PEEK tubing to waste. A Harvard Apparatus Pump-11 (Holliston, MA, USA) was used to infuse the LockSpray solution at 10-25 μL/min to the secondary ESI probe. The LockSpray solution was a 5 ng/μL solution of leucine enkephalin acetate salt hydrate (cat. no. L9133, ≥95%; Sigma Aldrich, St. Louis, MO, USA) in 50/50 vol/vol acetonitrile/water.

The GC-MS data was obtained on either a Waters GCT time-of-flight mass spectrometer equipped with an Agilent 5890 gas chromatograph and autosampler or a DSQ-II GC-MS equipped with a Trace GC and TriPlus Autosampler (ThermoFisher Scientific, Waltham, MA, USA). A custom chemical ionization manifold [16] replaced the standard one on the DSQ-II mass spectrometer for introduction of reagent gases.

Solutions Added Post Column and Ion Adduct Determinations

Additives were routinely added post-column to enhance ESI ionization or determine the identity of the molecular ion adduct. Normally a 25 mM solution of ammonium acetate in methanol was employed; however, in some cases when the actual ion adduct observed was in doubt, a 700 μM solution of potassium acetate in methanol was employed. Another means to determine the identity of the molecular ion adduct is to increase the cone voltage from 25 to 75 V when employing either ammonium acetate or organic acid as ionization reagents. The absolute intensity of the sodium adduct ($M + Na^+$), for the molecular ion (M) normally increases significantly with a corresponding decrease in the protonated ($M + H^+$) or ammonium ion ($M + NH_4^+$) adducts.

Chromatographic Separations

The typical LC-MS separation employed a Hypersil ODS column, 5 μ×4.6 mm i.d.×100 mm (Hewlett Packard, Palo Alto, CA, USA), at a flow rate of 1.5 mL/min and a column temperature of 30 °C. The LC gradient was 5%–100% organic in 15 min with a total analysis time of 30 min. Acetonitrile, UV grade, was the organic solvent (Honeywell Burdick and Jackson). The aqueous solvent was prepared by mixing 192 mg of ammonium acetate in 1000 mL of water (Milli-Q Water System, Millipore Corp., Billerica, MA, USA) and adding 30 mL of acetonitrile. Addition of the organic solvent to the aqueous solvent retards bacterial growth and the small amount of ammonium acetate stabilizes (minimizes effect of CO_2 in air on pH) the retention times of acidic species.

GC-MS separations were typically performed on a DB5-MS, 30 m×0.25 μm film×0.25 mm i.d. column (Agilent J&W) employing helium as the carrier gas at a constant flow rate of either 1.0 or 1.5 mL/min. The temperature was programmed from 40 to 320 °C at 15 °C/min after an initial hold time of 2 min. The total run time was 40 min.

Determination of Exchangeable Protons

Three different methods were employed to determine the number of exchangeable protons in the mixtures or extracts. In the first method, the sample is infused at 10 μL/min into the ESI source dissolved in a solution of 50/50 vol/vol acetonitrile/D_2O containing 2 mM ammonium acetate or 700 μM potassium acetate. The trace amounts of non-deuterated reagents, such as ammonium acetate used to enhance ionization, do not normally hamper the counting of exchangeable protons. In the second method [16], ND_3 is utilized as the chemical ionization reagent for GC-MS analyses. The third method employs N,O-*bis*(trimethylsilyl) trifluoroacetamide (BSTFA) as the derivatization reagent for GC-MS analyses [17].

Calculation of Average Molecular Weights and Molecular Formulae

The data was obtained in continuum mode and then centroided for molecular formulae and average molecular weight (MW) determinations. Molecular formulae were generated with the Waters Elemental Composition Program (ver. 4.0), which included i-FIT for numerically ranking the observed isotopic pattern to the theoretical one. Average molecular weights were calculated from centroided intensities (see Figure 1) using Excel (Microsoft Corp., Redmond, WA, USA).

We found it much more difficult to measure the average MW than the monoisotopic MW. Failing to include a significant isotope intensity in the calculation due to signal/noise limitations or the presence of chemical noise will significantly bias the results. Particular attention was paid to obtain good signal/noise determinations for the molecular ion and its significant isotopes, while avoiding mass shifts from dead time corrections [18]. Measurements were frequently taken by averaging the "scans" on the sides of chromatographic peaks attempting to obtain approximately 400–500 counts per "scan" per s in continuum mode.

The standard deviations for the monoisotopic and average molecular weight (MW) determinations were approximately 6 and 25 ppm, respectively, in chromatographic determinations. These values could be improved by about a factor of two if the samples were infused employing larger average numbers of scans for the measurement and carefully

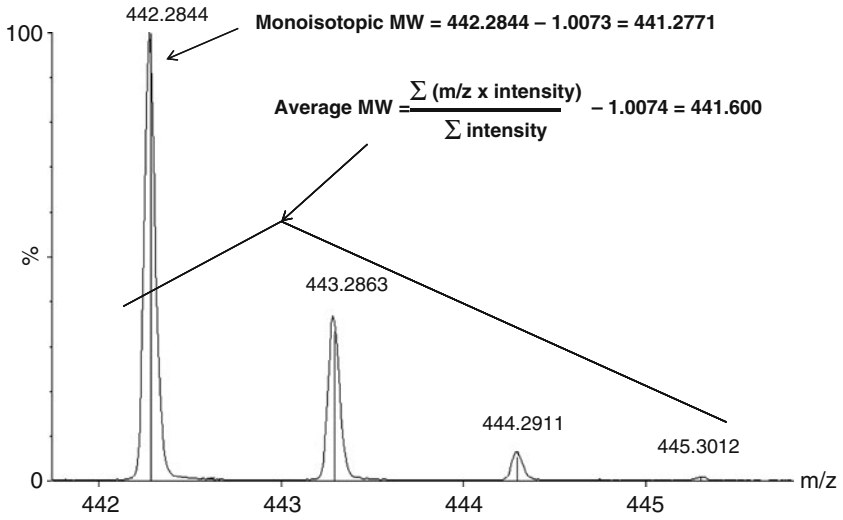


Figure 1. Determining the average MW of "known unknown" from centroided data for M + H species

matching the response of the lock reference mass to that of the unknown. The range of values noted for the average MW measurements was –50 to 50 ppm, thus a window of ±70 ppm was employed in all searches to ensure no compounds of interest were excluded from the initial search results.

The accurate masses calculated by the Waters software are not corrected for the mass of an electron [19]. The errors cancel from within the Elemental Composition program since the reference calibration tables are also not corrected. However, all data exported into other applications for further data processing were corrected manually. Furthermore, the isotope ratios generated in the Waters Isotope modeling program utilizes the pre-1995 value for the average atomic molecular weight for carbon, $A_r(C)$, of 12.011 instead of the currently accepted $A_r(C)$ value of 12.0107 [20]. The average molecular weights calculated by Advanced Chemical Laboratories for CAS Registry are based on the currently accepted value.

Preparation of Samples for Analysis

Polymer samples for analysis of additives are prepared in several different ways. In one method, the polymer sample is extracted overnight in a Soxhlet extractor utilizing diethyl ether as solvent. The extract is then concentrated to dryness and redissolved in acetone. In the second method, the polymer (5–25 mg) is dissolved in 1 mL of acetone, methylene chloride, or tetrahydrofuran. The polymer solution is injected directly into either the GC-MS or LC-MS systems. The dissolved polymer normally causes no significant problems for the HPLC column, but the GC injector is changed more frequently than normal. In a variation of the second method, the polymer is dissolved in a strong solvent such as methylene chloride and methanol is added to precipitate the bulk of the polymer. The methanol/methylene chloride solution is then decanted or filtered and concentrated for analysis.

Results and Discussion

Introduction to Two Approaches

The Chemical Abstracts Service[sm] (CAS) Database contains over 54 million organic and inorganic substances and 32 million document records. Thus it is by far the largest readily accessible curated database [7]. The database can be queried by a variety of inputs including research topics, molecular formula, average molecular weight, structure, etc. We employ both STN (Science and Technical Information Network) Express and the web-based version of SciFinder for searching the database. SciFinder employs a much more intuitive graphics interface while STN Express utilizes a command language interface. The latter is intended primarily for the information professional, thus it requires more initial user training. There is a fee associated for accessing the databases and the cost can be significant depending on the user contract and the interface employed.

The molecular formula is the best search parameter for identifying "known unknowns." However, if a unique molecular formula cannot be determined, the average molecular weight can also be searched. Others have noted [3] that searching databases with the use of molecular weight instead of the standard molecular formula search to be very effective. Indeed, they demonstrated that even unit mass resolution can be used for higher molecular weight compounds. In theory, as the molecular weight of an unknown increases the number of possible molecular formulae will increase exponentially [9, 11]. However, in practice, as shown in Figure 2, the number of compounds accessed by STN Express maximized around ~400 Da and falls precipitously at higher mass.

We have successfully employed our approaches to identify a large number of unknowns in additives to

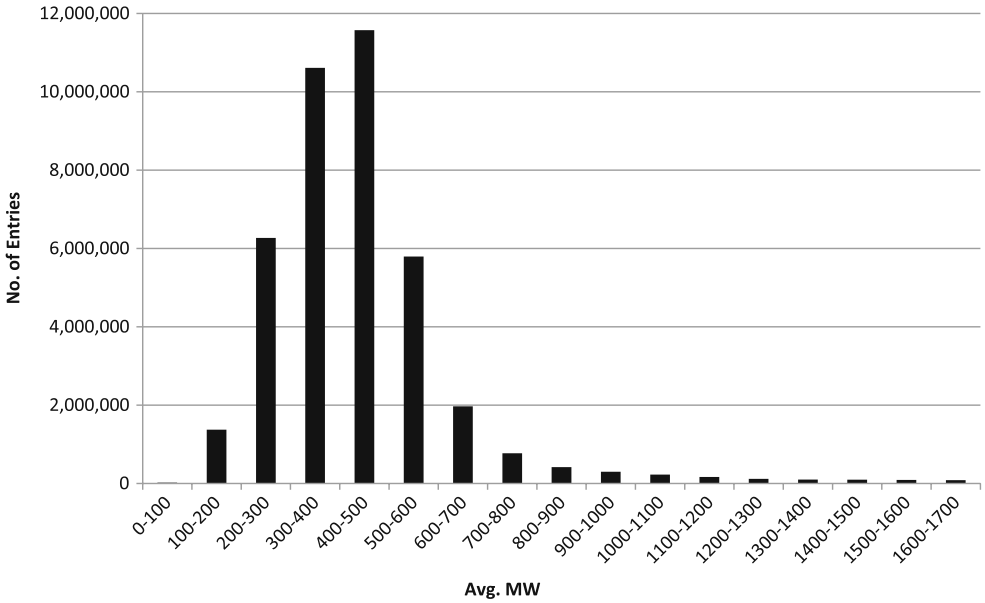


Figure 2. Average MW distribution of compounds in CAS Registry (Dec. 2009) determined by s /mw command in STN Express

commercial products; extracts from natural products; and impurities in research and developmental products. Candidate structures from the approach are always further substantiated with additional information such as MS/MS and electron impact fragmentation patterns, nuclear magnetic resonance data, relative retention times, UV-VIS diode array spectra, number of exchangeable protons [10, 16, 17], presence of similar compounds in the mixture, chemical derivatization [10, 17], etc. Of course, the ideal confirmation is comparison of the data of the unknown to that of either a commercial or prepared standard of the material.

Several examples are discussed that illustrate the nuances of our approaches in the identification of "known unknowns" from our laboratory. In addition, we have validated our approaches in "identifying" other components including pharmaceuticals, toxins, and a variety of polymer additives noted in the literature, in technical conferences, and on the internet.

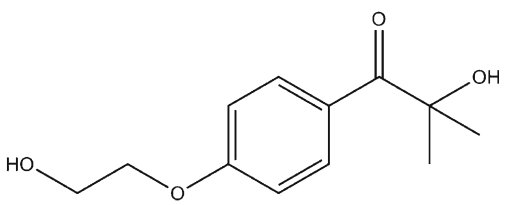
SciFinder Molecular Formula Search Refined by Number of References

A search of the CAS Registry with the web-based version of SciFinder using molecular formula data is the preferred approach when attempting to identify components with essentially no sample history. The initial candidate list is refined using the number of associated references for the compounds in descending order as an orthogonal filter. This approach is demonstrated in the identification of a component in a diethyl ether Soxhlet extract of a polyester label.

The molecular formula was determined to be $C_{12}H_{16}O_4$ from ESI LC-MS data. A search using SciFinder (Explore Substances/molecular formula) yielded 4486 substances. This list was sorted in descending order by number of references (pull down menu, sort by: Number of References↓). Screen displays of the search can be found in the Supplementary Material. The top substance in the list, Scheme 1, had 849 references with the next highest hit having 193 references. The ability to sort by number of references is only available on the web-based version of SciFinder and not the client-based version.

The proposed structure was consistent with the sample history since photoinitiators are utilized in UV-curable inks for labels. However, additional data was needed to confirm the proposed structure.

Two exchangeable protons were observed by ND_3 chemical ionization [16] GC-MS, which is consistent with the presence of two hydroxyl groups in the candidate structure. We routinely determine the number of exchangeable protons in compounds by this approach. However, two other approaches can be employed for either thermally labile or nonvolatile compounds. The first involves ESI analyses by infusing the extracts in solvents mixtures containing D_2O. The second approach is to convert the compounds to their trimethylsilyl (TMS) derivatives [10, 17]. The number of exchangeable protons is determined by the increase of 72 units in the molecular weight for each group derivatized. Furthermore, this latter approach offers an additional benefit



Scheme 1. Photoinitiator identified in diethyl ether extract of polyester label

since many compounds are present in computer-searchable EI databases as their TMS derivatives.

The base peak for the in-source [21] CID spectrum at *m/z* 179 in the positive ion ESI spectrum has a molecular formula of $C_{11}H_{15}O_2$. This could be due to an unexpected rearrangement of the protonated molecular ion to give the benzyl stabilized cation (Scheme 2).

However, no model compounds could be found in our limited MS/MS reference databases to support this supposition. Therefore, the sample was analyzed by EI GC-MS to obtain further data. EI fragmentation data is normally much better understood [22, 23] than that of MS/MS and many more reference model spectra are available.

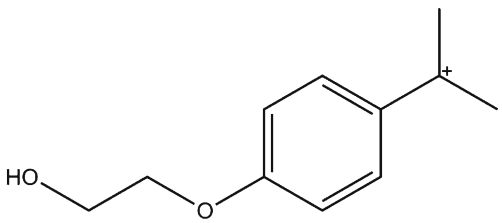
We routinely analyze mixtures by both GC-MS and LC-MS since the techniques often offer complimentary and supplementary data. In this case, neither the EI mass spectrum of the candidate structure or its *bis*-(TMS) derivative was found in any of our computer-searchable databases. However, a model compound, Scheme 3, was found, which showed the distinctive loss of 59 from the molecular ion and a fragment ion at m/z 59. This same behavior is noted in the EI mass spectrum of the unknown.

Ultimately, the identity of the photoinitiator was confirmed by the purchase of a reference sample. SciFinder is a very useful resource for locating commercial sources of reference compounds. For example, 42 commercial sources of the photoinitiator were listed.

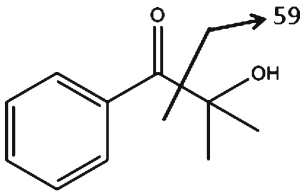
The same type of approach can be used to search the CAS registry employing STN Express instead of SciFinder by substituting/MF for /MW in the search command in Figure 3. However, we generally find SciFinder to be much easier to use compared to STN Express and less expensive for the individual user.

SciFinder Molecular Formula Search Refined with Minimal Sample History

If some sample history is available, a SciFinder search of the CAS Registry employing the molecular formula is the preferred approach. The summed list of references for all candidates is further refined using a key word as an orthogonal filter. This approach is demonstrated in three examples that illustrate the minimal sample history needed for success.



Scheme 2. Fragment ion noted at *m/z* 179 for in-source ESI MS/MS spectrum of photoinitiator



Scheme 3. Fragments noted in EI mass spectrum for model compound for photoinitiator

The first is a detailed example for the identification of a compound found in the extract of a can coating. Polymers used in food contact applications must meet criteria for extractables when an appropriate food simulating solvent is employed. A polyester coating was applied to a metal can at a contract laboratory. Extraction of the can was found to yield the expected linear and cyclic polyesters routinely extracted from the polymer coating. However, an additional UV-absorbing material was noted by LC-MS.

Accurate mass data indicated a molecular formula of $C_{36}H_{40}O_6$. A search using SciFinder (Explore Substances/molecular formula) yielded 181 substances. Since some limited sample history was available, the search was further refined. All the references for all the 142 substances (Get References/All Substances) were selected. The list of 168 references was then refined (Refine/Research Topic) with the phrase "can coating." The first four of the five references showed the highlighted structure for cyclo-DiBADGE (cyclo-di-bisphenol A diglycidyl ether, Scheme 4) as the identity of the "known unknown." Screen displays of the search can be found in the Supplementary Material.

The material is a common low molecular weight cyclic monomer noted in the extracts from BADGE type epoxy based resins used in can coatings. The contract lab had contaminated our coating with the material. The identification of the unknown was further confirmed by proton NMR data of the mixture showing its characteristic aromatic proton resonances [24].

The second example is the identification of an acetone extract of cotton linters. Cotton linters are acetylated to form cellulose triacetate utilized in liquid-crystal displays (LCD). A search of the molecular formula $C_{22}H_{44}O_3$ in SciFinder yielded 284 substances with 697 combined references. Refining the list of references with "cotton or linter" (logical operator utilized) yielded several references indicating the material was the omega-hydroxy C22 fatty acid (see Scheme 5). The number of exchangeable protons was determined to be two by formation of the trimethylsilyl derivative for GC-MS analysis. A purchased sample of the material confirmed the identification.

The third example was the identification of a colored component noted in the application of a hydrocarbon resin as an adhesive in a disposable diaper. The LC-UV-VIS-MS analysis targeted a yellow species with a strong absorbance in the 375–450 nm range with a molecular formula of $C_{30}H_{42}O_2$. The SciFinder search yielded 257 substances

```
=> file registry   [move to registry file]

=> s 441.57-441.63/mw   [search average MW range]

L1       11898 441.57-441.63/MW   [11,898 candidates found]

=> sort L1  [sort entries by No. of references in descending order]
SORT ENTIRE ANSWER SET? (Y)/N:y
ENTER SORT FIELDS AND SORT DIRECTION (?):ref d

6687 ANSWERS DID NOT HAVE 'REF' SORT FIELD
PROCESSING COMPLETED FOR L1
L2       11898 SORT L1 REF D

=> d L2 6688-6692  [display first 5 registry entries in list with references, entries with no references
                    occur at top of list!]

=> file home   [move to home, minimize charges to account]
```

Figure 3. STN Express search of CAS Registry by average MW range, comments added by authors in brackets after command for explanation of process

with 472 combined references. Refining the list of references with "yellow" yielded eight references indicating the compound was a highly conjugated quinone. The color problem was due to the oxidation (Scheme 6) of excess BHT added to the adhesive as an antioxidant.

STN Express Average Molecular Weight Search Prioritized by Number of References

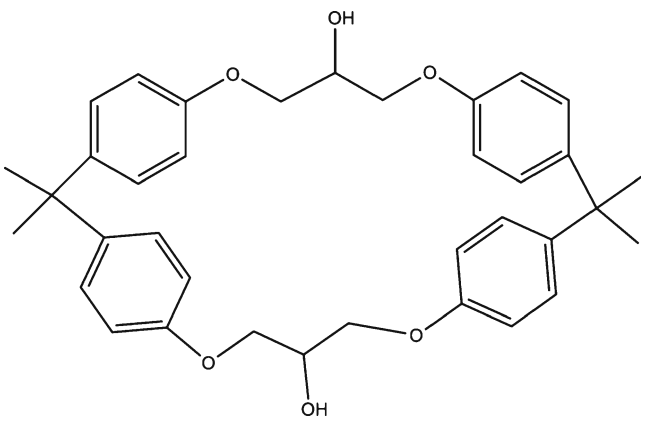
If a definitive molecular formula cannot be determined, the CAS Registry can be searched by average molecular weight and the candidate list ordered by the number of references. In theory, this approach should not work as well at higher MW ranges since the number of possible compounds increases exponentially as a function of MW. However, in practice, the number of compounds present in the CAS Registry falls precipitously as the MW increases (see Figure 2).

Unfortunately, STN Express can only be searched with the average molecular weight (MW) and not the monoisotopic MW. The limitations of employing average MW versus monoisotopic MW are further discussed in a later section. The web-based version of SciFinder cannot be searched at a higher level by average MW. It can only be used as a variable to further refine an initial search.

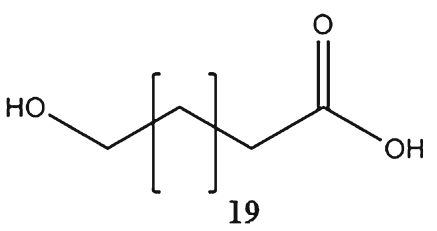
This approach was employed to identify an additive in a competitive polyester resin. The average MW for the compound was determined to be 441.600 (see Figure 1). All values for average MW in the CAS Registry are rounded to the hundredths place. The determined MW was rounded to 441.60 and then a window of ±0.03 (±70 ppm) Da was used for the search window. This wide error window is employed to ensure that no reasonable candidates are excluded from the search results.

The command sequence for searching by a MW range is shown in Figure 3. There were 11,898 candidate structures and sorting the references in descending order yielded 6687 candidates that *did not* have reference fields at the *top* of the list. This is somewhat counterintuitive since entries with no entries logically should be found at the bottom of the list when sorted in descending order. Nevertheless, the top five hits from the search are displayed using a range of 6688–6692. The first entry in the displayed results was Tinuvin 928, Scheme 7, which had 112 references. The next entry had only 19 associated references. Tinuvin 928, from references in the CAS database and internet searches, is a light stabilizer which is a reasonable component to be present in a commercial polyester sample.

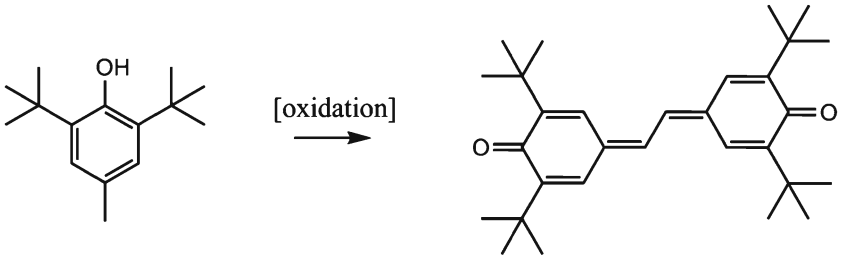
The in-source CID spectrum for the compound (see Figure 4) was consistent with the proposed structure and the number of exchangeable protons was determined to be one by ESI infusion. In addition, the observed monoisotopic



Scheme 4. Cyclic dimer noted in extract from contamination in can coating process



Scheme 5. Structure for component in acetone extract of cotton linters



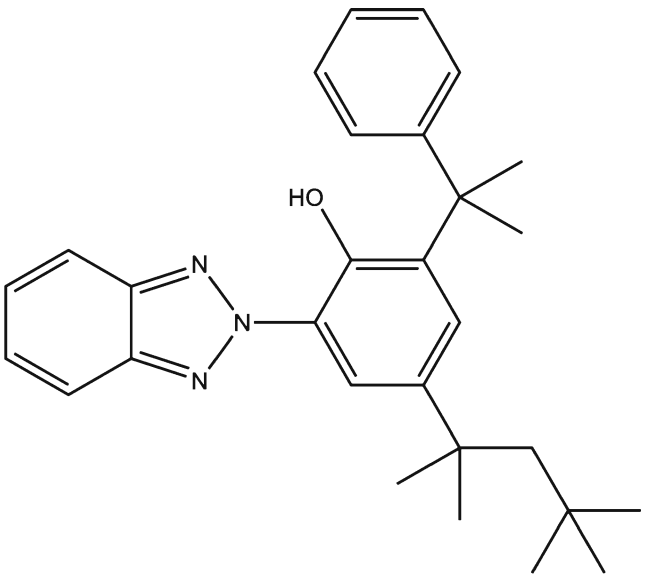
Scheme 6. Conjugated quinone dimer formed from oxidation of BHT leading to yellow color in diaper

MW (within 2.0 ppm) and isotopic pattern for the molecular ion cluster were consistent with the proposed molecular formula. The initial data was so definitive that the results could be reported to the customer. At a later date, a reference sample of the material was obtained which confirmed the identification.

In retrospect, it is somewhat surprising that the correct structure was noted with this approach employing such a wide average MW search window. The MW of the compound is in the middle of the MW range with the largest number of structures (see Figure 2). In theory, this approach should work best with higher MW compounds which are present in the database at much lower frequency.

SciFinder Average MW Search Refined with Minimal Sample History

If a definitive molecular formula cannot be determined, the CAS Registry can be searched by average MW and further refined using a key word as an orthogonal filter. The initial step in this approach is the same as that employed in the first step in the previous section.



Scheme 7. UV light stabilizer identified in competitive polyester sample

This approach is illustrated in the identification of an additive in a polypropylene polymer whose average MW was determined to be 784.065. The CAS Registry was searched with an average MW range of 784.01–784.12 (±70 ppm) as shown in Figure 5. The 824 candidate structures were then refined in HCAPLUS (Chemical Abstracts Plus, database with bibliographic references) using "polymer and additive" as orthogonal filters to yield 121 records.

There are two ways to view the results. The first way is to sort (sort command, Figure 5) by the number of references each record contained in descending order. This tends to bring the most pertinent records to the top of the list. When displaying the first 10 hits in the sorted list, all but two had Irganox 3114, Scheme 8, highlighted. The large fragment ion at *m/z* 219 in the in-source CID spectrum was consistent with the presence of the butylated-phenol groups and a reference sample of the material confirmed its identity.

The second way is to analyze the 121 records (analyze command, Figure 5) by the frequency of the number of registry numbers occurring within these records. The first hit from the analyze command was again Irganox 3114. Normally the most likely candidates will be the top 1–3 hits in the displayed table. The other candidates do not necessarily have the average MW searched initially, since they just reflect the frequency of the compounds in the selected records. However, the results will likely reveal what other additives are often used with or instead of the compound of interest. In this case the other CAS numbers in the table were for polypropylene, polyethylene, and seven other antioxidants. None of the other antioxidants had molecular weights in the same range as that specified for Irganox 3114. The presence of polyethylene and polypropylene indicated these antioxidants are used in these polymers as stabilizers.

Comparison of Approaches for Additional Compounds

A group of 90 compounds was assembled from literature sources [2, 3, 6, 8], internet sites, and American Society for Mass Spectrometry Conference presentations to compare the effectiveness of our two approaches. The results are summarized in Tables 1 and 2. As expected, searching by

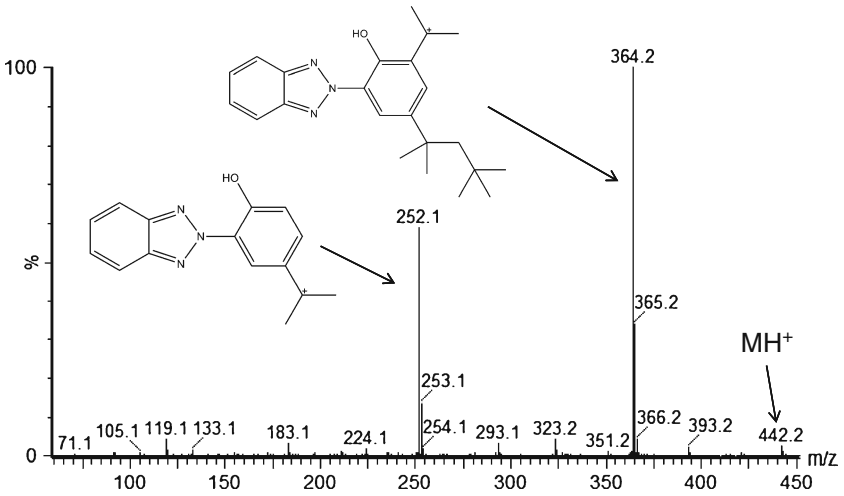
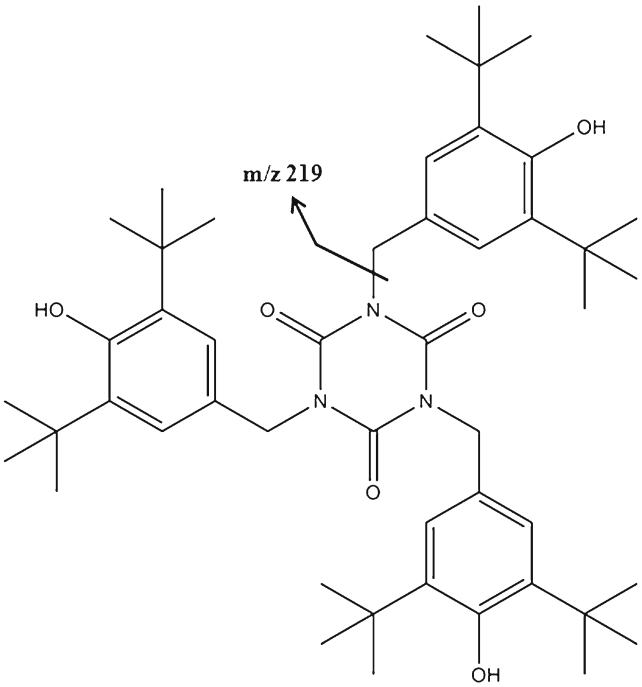


Figure 4. In-source CID spectrum of unknown showing fragment ions consistent with structure

```
=> file registry   [move to registry file]

=> s 784.01-784.12/mw  [search average MW range]

L1       824 784.01-784.12/MW  [824 candidates found]

=> file hcaplus   [move to bibliographic database]

=> s L1 and polymer? and additive?   [searching candidates and polymer and additive, expansion of
                                      characters beyond stem term]

L2       121 L1 AND POLYMER? AND ADDITIVE?   [found 121 matches]

=> sort L2   [method 1, sort hits descending order by No. of references in record]
SORT ENTIRE ANSWER SET? (Y)/N:y
ENTER SORT FIELDS AND SORT DIRECTION (?):rec d
         71 ANSWERS DID NOT HAVE 'REC' SORT FIELD
PROCESSING COMPLETED FOR L2
L3        121 SORT L2 REC D

=> d L3 1-10 hit   [display the top ten results, CAS numbers of interest are highlighted, ones with no
                    references are found at end of list!]

=> analyze L2 rn   [method 2, determine the highest frequency of registry number, RN, components in records]
ENTER ANSWER NUMBER OR RANGE (1-):1-121
L4        ANALYZE L2 1-121 RN :    950  TERMS

=> d L4 rn  [display table of frequency of RN components in records]
L5        ANALYZE L4 1-121 RN :    950 TERMS

TERM #   #OCC  #DOC  %DOC    RN
------ ------- ------ ------ -----------------------
    1       116    116  95.87   27676-62-6 [Irganox 3114, correct MW]
    2        70     70  57.85    6683-19-8  [antioxidant]
    3        52     52  42.98    2082-79-3  [antioxidant]
    4        48     48  39.67    1709-70-2  [antioxidant]
    5        45     45  37.19   31570-04-4  [antioxidant]
    6        43     40  33.06    9003-07-0  [polypropylene]
    7        32     32  26.45     128-37-0  [antioxidant]
    8        27     27  22.31   26741-53-7  [antioxidant]
    9        26     26  21.49    9002-88-4  [polyethylene]
   10        24     24  19.83     693-36-7  [antioxidant]

=> file home  [move to home, minimize charges to account]
```

Figure 5. STN Express Search of CAS Registry by average MW range, candidates limited by sample history, comments added by authors in brackets after command for explanation of process



Scheme 8. Irganox 3114, antioxidant identified in competitive polypropylene polymer showing characteristic ESI in-source MS/MS fragment ion

molecular formulae, then sorting by the number of references, yielded more target compounds correctly ranked as number 1 compared with the same approach using average molecular weights. However, the overall number with rankings less than or equal to 5 was comparable. Thus, the use of average molecular weight is a viable approach when a unique molecular formula is not readily determined.

This is somewhat surprising since the average number of entries (11,126) is much larger when employing average molecular weights with a window of ±70 ppm compared with the average number of entries (1265) when employing molecular formulae. It is expected that if the CAS Registry could be searched by monoisotopic molecular weights with a window of ±5 ppm instead of average molecular weights, the results from the former would be comparable to those obtained with molecular formulae. This would be advantageous since no subjective restrictions on the elements, the range of elements, and the double bond equivalents would be required prior to the initial search to limit candidate compounds.

Limitations in Approaches

There are several limitations to our approaches including (i) employing average MW instead of monoisotopic MW; (ii) the availability of average MW values for certain registry entries; (iii) lack of average MW data for incompletely defined substances; (iv) the listing and searching of charged species; and (v) the inefficiencies in processing complex samples. These limitations are discussed in the following paragraphs.

The monoisotopic MW of a compound can be determined much more precisely than the average MW. The precision is much lower in the latter case since the *m/z* and intensity of all the ions in the molecular ion cluster must be measured (see Figure 1). Failing to include a significant isotope in the calculation due to signal/noise limitation or the presence of chemical interferences can significantly bias the results.

Not all entries in the Registry include average molecular weight data. The formula weight (FW, nominal MW) field is generated and indexed by CAS for all entries with structural connection tables. On the other hand, the average MW field is generated for only single-component substances and is not generated for polymers, coordination compounds, metal-containing species, ionic species, and radicals. An STN Express search of the Registry with FWs and average MWs between 400 and 405 yielded ~900,000 entries with FW fields, but only approximately 20% of these entries included average MW fields. A random sampling of the latter substances showed that the average MW field for many compounds appeared to have been inadvertently omitted from these records. CAS personnel plan to evaluate these omissions to determine whether or not MW values should be included.

There are many incompletely defined substances in the CAS database, which include molecular formulae fields. However, average molecular weight fields are not calculated

Table 1. SciFinder (web-based version) approach searching with molecular formula and sorting by number of references descending

Class of compounds	Number compounds in class	Position of compound sorted in descending order by number of references					
		#1	#2	#3	#4	#5	>#5
Drugs	45	44	1				
Pesticides	8	8					
Toxins	2	2					
Polymer antioxidants	15	15					
Polymer UV stabilizers	10	8	1	1			
Polymer clarifying agent (Irgaclear DM)	1	1					
Polyurethane additives	4	2	1		1		
Natural products	3	3					
Herbicide (clofibric acid)	1		1				
Artificial sweetener (sucralose)	1	1					
Total Compounds	90	84	4	1	1		

Table 2. STN Express approach searching with average molecular weight and sorting by number of references descending

Class of compounds	Number compounds in class	Position of compound sorted in descending order by number of references					
		#1	#2	#3	#4	#5	>#5
Drugs	45	34	7	2		1	1 (#7)
Pesticides	8	6	2				
Toxins	2	1	1				
Polymer antioxidants	15	14		1			
Polymer UV stabilizers	10	6		1	1	2	
Polymer clarifying agent (Irgaclear DM)	1	1					
Polyurethane additives	4	1	2				1(#17)
Natural products	3	2					1(#9)
Herbicide (clofibric acid)	1		1				
Artificial sweetener (sucralose)	1	1					
Total compounds	90	66	13	4	1	3	3

for this class of entries. Two examples are tris(nonylphenyl) phosphite (CAS No. 26523-78-4, polymer antioxidant) and isomers of diethyl-methyl-1,3-benzenediamine (CAS No. 75389-89-8, polyurethane additive). Thus, the identifications of these types of materials would be limited to molecular formulae searches.

The listing for organic ions is not straightforward in the database, and it is important to be aware of the format for effectively searching their molecular formulae in SciFinder. As a simple example, the molecular formula of sodium acetate is listed as $C_2H_4O_2$.Na. An additional proton is added to the molecular formula of the organic anion and it is separated from the cation with a period. For many compounds, there will be significant entries in the database for the neutral acid, the salt, and even the free anion. Thus, depending on the particular species, the database should be searched by one or more of the following molecular formulae: $C_2H_4O_2$, $C_2H_4O_2$.X (where X = Na, K, NH_3, etc.), and $C_2H_3O_2$.

A similar situation is observed for quaternary amines. A simple example is tetramethylammonium hydroxide listed in the CAS database as C_4H_{12}N.HO. Thus, one or more of the following molecular formulae should be searched: C_4H_{12}N.X [where X = HO, $C_2H_3O_2$ (acetate), Cl, Br, etc.]. Several different criteria are useful in recognizing the presence of a quaternary ammonium species in ESI analyses [3].

Amphoteric (inner salt, zwitterionic) species are listed in the database with no modification to their molecular formulae. For example, $(CH_3)_3{}^+NCH_2CO_2{}^-$ is listed as $C_5H_{11}NO_2$. This type of compound would yield $(M + H)^+$ and $(M + \text{acetate})^-$ ions, respectively, in positive and negative ion ESI analyses using acetate in the LC eluent. Thus, the molecular formula observed for the species would need to be corrected before searching.

The process of analyzing a relatively complex mixture by our approaches can be very time consuming. The processing efficiency could be significantly increased if the observed molecular formulae could be automatically transferred into and searched by SciFinder. The results could then be sorted by number of references and the 5–10 best candidates reported. The structures should be copied from the report as connection tables and transferred to a commercial drawing program through the computer "clipboard." The structures could then be manually fragmented by the user with a "lasso" to determine if the fragment ions in the MS/MS spectrum were consistent with the proposed structures. Alternatively, the structures could be combined with the spectrum and transferred to a program such as NIST MS Interpreter, which automatically correlates ions with sub-structural features.

Conclusions

Searching the CAS database with SciFinder (web-based version) and STN Express by molecular formulae and by average molecular weights, respectively, were demonstrated to be very useful approaches for the identification of "known unknowns." The approaches have been utilized to identify a variety of components, including polymer additives, extracts from natural products, and impurities in research and developmental products. The two approaches were also evaluated with a variety of compounds of interest to others.

The web-based version of SciFinder was the preferred approach when a unique molecular formula could be determined. However, STN Express using the average molecular weight was a viable approach when the accurate mass data did not yield a unique molecular formula. In both approaches, sorting the candidate structures from the initial search by the number of associated references was an effective orthogonal filter for prioritizing the list of candidate structures. Also, minimal sample history was found to be useful in further optimizing the list.

Several limitations in the current approaches were detailed, which could be overcome by modifications in the CAS Registry and associated search engines. The changes would involve inclusion of a monoisotopic MW field, searching partial molecular formulae fields for charged species, and the calculation of molecular weights for incompletely defined substances. In addition, the ability to search the average MW at a higher level in the web-based SciFinder software would be useful until the average MW field could be added to the CAS Registry.

More complex changes would be needed to improve the efficiency of our approaches for complex samples. Molecular formulae and/or monoisotopic accurate masses would need to be inputted automatically from the mass spectrometer's data processing software. A report of the 5–10 best hits would then need to be automatically generated with structure connection tables easily exported via the personal computer's "clipboard" to other programs for further processing.

Acknowledgments

The authors thank Jean D. Coffman and Mike Ramsey of Eastman Chemical Company for assistance in developing the STN Express approaches for identifying unknowns and Adam Howard of Eastman Chemical Company for assistance in acquiring and interpreting mass spectral data. They thank Bill Tindall and Kent Morrill (retirees from Eastman Chemical Company) for their initial work on "spectraless" databases using the Eastman Corporate Plant Material and TSCA Databases. A special thanks to Jim Lekander from Waters Corporation for advice in optimizing accurate mass measurements using LockSpray and to Anthony Machosky from CAS for reviewing the article and offering suggested changes. The artwork for the Graphical Abstract was drawn by a dear sailing friend, Minta Fannon.

References

1. Richardson, S.D.: Environmental mass spectrometry: emerging contaminants and current issues. *Anal. Chem.* **80**, 4373–4402 (2008)
2. Hogenboom, A.C., van Leerdam, J.A., de Voogt, P.: Accurate mass screening and identification of emerging contaminants in environmental samples by liquid chromatography-hybrid linear ion trap orbitrap mass spectrometry. *J. Chromatogr. A* **1216**, 510–519 (2009)
3. Liao, W., Draper, W.M., Perera, S.K.: Identification of unknowns in atmospheric pressure ionization mass spectrometry using a mass to structure search engine. *Anal. Chem.* **80**, 7765–7777 (2008)
4. Little, J.L: Identification of Unknowns with LC/MS/MS Data Using TSCA, Chemical Inventory, or SciFinder. Proceedings of the ASMS, Chicago, IL (May 2001)
5. Little, J.L.: Identification of Surfactants by Electrospray Liquid Chromatography-Mass Spectrometry. Proceedings of the ASMS, Nashville, TN (May 2004)
6. Garcia-Reyes, J.F., Ferrer, I., Thurman, E.M., Molina-Diaz, A., Fernandez-Alba, A.R.: Searching for non-target chlorinated pesticides in food by liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **19**, 2780–2788 (2005)
7. Kind, T., Scholz, M., Fiehn, O.: How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS One* **4**(5), e5440 (2009). doi:10.1371/journal.pone.0005440
8. Ojanperä, S., Pelander, A., Peizing, M., Krebs, I., Vuori, E., Ojanperä, I.: Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **20**, 1161–1167 (2006)
9. Kind, T., Fiehn, O.: Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 PPM. BMC Bioinformatics. 7(234) (2006). doi:10.1186/1471-2105-7-234
10. Kind, T., Fiehn, O.: Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**(105), 105 (2007). doi:10.1186/1471-2105-8-105
11. Suzuki, S., Ishii, T., Yasuhara, A., Sakai, S.: Method for the elucidation of the elemental composition of low molecular mass chemicals using exact masses of product ions and neutral losses: application to environmental chemicals measured by liquid chromatography with hybrid quadrupole/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **19**(23), 3500–3516 (2005)
12. Grange, A.H., Zumwalt, M.C., Sovocool, G.W.: Determination of ion and neutral loss compositions and deconvolution of product ion mass spectra using an orthogonal acceleration time-of-flight mass spectrometer and an ion correlation program. *Rapid Commun. Mass Spectrom.* **20**, 89–102 (2006)
13. Kaufmann, A.: Determination of the elemental composition of trace analytes in complex matrices using exact masses of product ions and corresponding neutral losses. *Rapid Commun. Mass Spectrom.* **21**, 2003–2013 (2007)
14. Milman, B.L.A.: Procedure for decreasing uncertainty in the identification of chemical compounds based on their literature citation and co-citation. Two case studies. *Anal. Chem.* **74**, 1484–1492 (2002)
15. Little, J.L., Cleven, C.D., Brown, S.D.: Accurate Mass Measurements: Identifying Known Unknowns Using Publicly Accessible Databases. Proceedings of the ASMS, Salt Lake City, UT (May 2010)
16. Parees, R.M., Kamzelski, A.Z., Little, J.L.: Deuterium Ammonia Chemical Ionization: Use in Counting Exchangeable Hydrogen Sites on Organic Compounds. In: Gross, M.L., Caprioli, R.M., Nibbering, N. M.M. (eds.) The Encyclopedia of Mass Spectrometry, vol. I, Fundamentals of an Applications to Organic (and Organometallic) Compounds, pp. 772–780. Elsevier, Amsterdam (2005)
17. Little, J.L.: Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. *J. Chromatogr. A* **844**(1/2), 1–22 (1999)
18. Chernuschevich, I.V., Loboda, A.V., Thomson, B.A.: Special feature tutorial: an introduction to quadrupole-time-of-flight mass spectrometry. *J. Mass Spectrom.* **36**, 849–865 (2001)
19. Mamer, O.A., Lesimple, A.: Common shortcomings in computer programs calculating molecular weights. *J. Am. Soc. Mass Spectrom.* **14**, 626 (2004)
20. de Laeter, J.R., Böhlke, J.K., De Biévre, P., Hidaka, H., Peiser, H.S., Rosman, K.J.R., Taylor, P.D.P.: Atomic weights of the elements: review 2000. *Pure Appl. Chem.* **75**(6), 683–800 (2003)
21. Bristow, A.W.T., Nichols, W.F., Webb, K.S., Conway, B.: Evaluation of protocols for reproducible electrospray in-source collisionally induced dissociation on various liquid chromatography/mass spectrometry instruments and the development of spectral libraries. *Rapid Commun. Mass Spectrom.* **16**(24), 2374–2386 (2002)
22. Budzikiewicz, H., Djerassi, C., Williams, D.H.: Mass Spectrometry of Organic Compounds. University Microfilms International Books on Demand, Holden Day, Inc., San Francisco (1967)
23. McLafferty, F.W., Tureček, F.: Interpretation of Mass Spectra, 4th edn. University Science Books, Sausalito (1993)
24. Tanaka, S., Yokoyama, K., Takashima, M.: Isolation of cyclic dimer from polyhydroxyether (I) (1). *J. Polym. Sci. Part B* **6**(6), 385–388 (1968)