Research Article

Rapid Communications in Mass Spectrometry

Received: 28 July 2013 Revised: 26 September 2013 Accepted: 6 October 2013 Published online in Wiley Online Library

Rapid Commun. Mass Spectrom. **2014**, *28*, 83–95
(wileyonlinelibrary.com) DOI: 10.1002/rcm.6759

Statistical approach to establish equivalence of unabbreviated mass spectra

Melissa A. Bodnar Willard[1,2], Ruth Waddell Smith[2]* and Victoria L. McGuffin[1]*

[1]Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA
[2]Forensic Science Program, School of Criminal Justice, Michigan State University, East Lansing, MI 48824, USA

RATIONALE: In many legal and regulatory applications, mass spectral comparison of an unknown or questioned sample to a reference standard or database is used for identification; however, no statistical confidence level or error rate is determined. Therefore, a simple and rapid method to establish the statistical equivalence of mass spectra is needed.
METHODS: The standard deviation of the abundance at each *m/z* ratio was determined from replicate measurements or from a statistical model. These standard deviations were used in an unequal variance *t*-test to compare two spectra at every *m/z* ratio over the entire scan range. If determined to be statistically indistinguishable at every *m/z* ratio, the random-match probability (RMP) that the specific mass spectral fragmentation pattern occurred by chance was calculated.
RESULTS: *n*-Alkane and alkylbenzene standards of varying concentrations were analyzed on the same instrument at different ionization voltages. Using the proposed method, replicate spectra were successfully associated at the 99.9% confidence level, with RMP values less than 10^{-29}. Despite the similarity in fragmentation patterns, spectra were distinguished from others in the homologous series. Moreover, the *n*-alkane spectra were appropriately associated to and discriminated from those in a standard reference database at the 99.9% confidence level.
CONCLUSIONS: A simple and rapid method to assign statistical significance to the comparison of mass spectra was developed and validated. This method may be useful for legal and regulatory applications, such as the identification of controlled substances, environmental pollutants, and food and drug contaminants.

In many legal and regulatory applications, evidence must be presented with statistical assessment of its validity. Statistical methods are well established for the comparison of deoxyribonucleic acid (DNA) samples, which are routinely used in court testimony.[1] For other types of evidence, statistical assessment is not yet available, as highlighted in a report published by the National Academy of Sciences, National Research Council (NRC).[2] In particular, mass spectrometry is used extensively for the identification of controlled substances, ignitable liquid residues, and other types of chemical evidence in forensic science.[2] In addition, the Environmental Protection Agency and the Food and Drug Administration use mass spectrometry for the identification of contaminants in the environment, food, pharmaceuticals, tobacco, etc.[3,4] Yet, in current methods, this identification is not supported by statistical assessment of the veracity by means of confidence levels or error rates. Such an assessment would address the NRC recommendations and be a timely advance not only for legal and regulatory applications, but for any application in which independent and objective validation is desired.

Figures of merit to describe the similarity of mass spectra are well established. For example, indices based on the dot product, composite similarity, probability-based matching, Hertz similarity, Euclidean or absolute value distances, and other methods have been developed.[5–11] These indices, particularly the dot product and probability-based matching, are widely used in commercial software to search mass spectral databases.[6,12–16] Because of their simplicity, these indices can rapidly determine the most likely identity of an unknown compound by comparison to reference mass spectra in the database. However, for legal or regulatory purposes, a further statistical test is needed to establish whether this tentative identification is objectively correct. Similarly, when an unknown or questioned compound is compared to an authentic standard, rather than a database, the same type of statistical test is needed.

As a specific example, consider when a controlled substance, such as cocaine, is seized and is sent to a forensic laboratory for identification. The questioned sample is generally analyzed by gas chromatography/mass spectrometry (GC/MS) and the retention time and mass spectrum are visually compared to those of an authentic standard or a library database. The Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG), under the auspices of the U.S. Department of Justice, Drug Enforcement Administration, has established general recommendations for analysis of such questioned samples.[17,18] However, specific details and criteria for mass spectral identification of the controlled substance are not defined in these recommendations and, hence, are left to the individual forensic laboratories and their jurisdictions. As a consequence, different laboratories may have vastly different

* *Correspondence to:* R. Waddell Smith and V. L. McGuffin, Michigan State University, East Lansing, MI 48824, USA.
E-mail: rwsmith@msu.edu; mcguffin@msu.edu

criteria in their standard operating procedures (SOPs). Some SOPs require that certain ions be present in both the questioned and standard spectra for that controlled substance (e.g., the molecular ion at *m/z* 303 for cocaine),[19] whereas others require that all ions greater than 10% of the base peak in the questioned spectrum correspond to those in the standard spectrum.[20] Still other SOPs allow the analyst to determine acceptable criteria for identification.[21] If the criteria in the SOP are fulfilled, the questioned sample is considered to be positively associated to the authentic standard. The analyst may then be called as an expert witness to explain the procedures used to identify the controlled substance and the corresponding results. Unfortunately, visual comparison of mass spectra in this manner can be subjective and inconsistent, which may result in rejection of scientific testimony in court. Moreover, neither visual examination nor the similarity indices mentioned above provide an unequivocal measure of the statistical confidence in the identification of the controlled substance, as required by the *Daubert* standard for the admissibility of evidence.[22] A confidence level and error rate associated with the identification would be beneficial, as they would meet the requirements of the *Daubert* standard[22] and address the recommendations set forth by the NRC report.[2]

In the present work, we describe a statistical approach to establish the equivalence of unabbreviated mass spectra. This approach is composed of two phases. Initially, statistical hypothesis testing, in the form of an unequal variance *t*-test, is applied at each mass-to-charge (*m/z*) value in the scan range for the two mass spectra. This test is used to determine if the abundance at each and every *m/z* value is statistically indistinguishable at a given confidence level. Then, if the two spectra are wholly equivalent, the random-match probability (RMP) is used to estimate the error rate. The RMP is calculated based on the frequency of ion occurrence at each *m/z* value in a selected database, similar to the approach used for DNA profiling in forensic science.[1] In the present case, the RMP assesses the probability that the characteristic fragmentation pattern of the two mass spectra would occur by random chance alone.

The proposed method utilizes the entire mass spectrum, rather than an abbreviated spectrum consisting of the most abundant ions, to establish the identity of the unknown or questioned sample. Accordingly, low-abundance ions, including the more characteristic high-mass ions, can provide vital information to discriminate spectra.[23] Compounds that have similar mass spectra pose the greatest challenge for identification. For this reason, normal (*n*) alkanes and alkylbenzenes were used for method development and validation. These compounds have mass spectra that are visually similar, containing common fragment ions such as *m/z* 43, 57, 71, 85, *etc*. These compounds, therefore, provide a rigorous test of the effectiveness of the proposed method.

STATISTICAL THEORY OF THE PROPOSED METHOD

Unequal variance *t*-test

The mean abundance and associated standard deviation at each *m/z* value in the mass scan range are required prior to calculating the statistical confidence of the association between two mass spectra, **1** and **2**. The mean abundance and standard deviation can be calculated from replicate mass spectra (sample replicates, solution replicates, and/or instrumental replicates, as appropriate). Alternatively, standard deviations can be predicted by using the counting statistics of the mass spectrometer detector. In this approach, the statistical response of the electron multiplier provides an independent means of estimating the variance inherent in the ion abundance and, for this purpose, may prove to be more robust and accurate than the traditional means of calculating standard deviations.

The statistical confidence of the association between spectra **1** and **2** can then be determined using hypothesis testing for each ion in the spectra. Statistically, the null hypothesis (H_0) is stated as:

$$H_0 : \left|\mu_{\mathbf{1}j} - \mu_{\mathbf{2}j}\right| = 0 \quad (1)$$

where the mean abundance, $\mu_{\mathbf{1}j}$ and $\mu_{\mathbf{2}j}$, of ion j in spectra **1** and **2** are statistically indistinguishable. The alternative hypothesis (H_a) is stated as:

$$H_a : \left|\mu_{\mathbf{1}j} - \mu_{\mathbf{2}j}\right| \neq 0 \quad (2)$$

where $\mu_{\mathbf{1}j}$ and $\mu_{\mathbf{2}j}$ are statistically distinguishable.

To determine which hypothesis is verified, an unequal variance *t*-test is used at each *m/z* value to determine if that ion in spectra **1** and **2** is statistically indistinguishable (H_0 accepted) or if it is statistically distinguishable (H_0 rejected).[24] If, at every *m/z* value, H_0 is accepted, then the two spectra are considered statistically equivalent or associated. Alternatively, the spectra are statistically differentiated if H_a is accepted for any ion in the spectra.

Two types of errors can arise in hypothesis testing. Type I errors arise if H_0 is accepted when it is false (i.e., spectra **1** and **2** are considered equivalent when, in truth, they are not the same compound) and Type II errors arise if H_0 is rejected when it is true (i.e., spectra **1** and **2** are considered differentiated when, in truth, they are the same compound). The confidence level at which the statistical test is performed relates to the probability of these errors occurring; for example, a two-tailed *t*-test at the 99.9% confidence level indicates the analyst is 99.9% confident there is not a Type I error.

Random-match probability

In order to evaluate the RMP, it is necessary to estimate the probability of random occurrence for each ion. It is commonly known that some fragment ions occur with greater frequency than others, where the less common ions may be more characteristic.[7] The frequency with which these ions occur in an extensive set of known mass spectra can be used to estimate the random probability. For this study, the frequency was determined using the National Institute of Standards and Technology (NIST) Mass Spectral Search Program, which contains approximately 150,000 electron ionization (EI) mass spectra, as the reference database. Other databases, including those generated in house, could be used as the reference.

Rapid Communications in Mass Spectrometry

Basic rules of probability theory are then used to calculate the likely occurrence of each ion in the mass spectrum.[24] If an ion is present in the mass spectrum, the probability P_j of occurrence is calculated by:

$$P_j = \frac{N_j}{N_T} \tag{3}$$

where N_j is the number of compounds containing ion j in a database of N_T total spectra. Conversely, if ion j is not present in the mass spectrum, the probability of non-occurrence is calculated by:

$$P_j = 1 - \frac{N_j}{N_T} \tag{4}$$

The total random-match probability of a particular sequence of ions appearing in the mass spectrum is calculated using the multiplicative rule[24]:

$$\mathrm{RMP} = \prod_{j=(m/z)_i}^{(m/z)_f} P_j = P_{(m/z)_i} \times P_{(m/z)_{i+1}} \times \ldots \times P_{(m/z)_f} \tag{5}$$

where $(m/z)_i$ and $(m/z)_f$ are the initial and final mass-to-charge ratios, respectively, in the mass scan range. This equation involves the simplistic assumption that the presence or absence of an ion at each m/z value is a statistically independent event. Independence has been commonly assumed when calculating probabilities of mass spectral data, most notably in the probability-based matching algorithm developed by McLafferty *et al.*[7] This algorithm is the default search engine for Agilent's ChemStation software, widely used in both gas and liquid chromatography with mass spectrometry.[12,13] Independence is also central to probability calculations in peptide-scoring algorithms, used in tandem mass spectrometry.[25–27] As noted by McLafferty *et al.*, this assumption is not rigorously true, but the probabilities provide a useful upper limit for these applications.[7] In our future studies, these probabilities will be further examined and refined.

EXPERIMENTAL

GC/MS analysis

Standards containing the *n*-alkanes decane (C_{10}), undecane (C_{11}), dodecane (C_{12}), tridecane (C_{13}), tetradecane (C_{14}), and hexadecane (C_{16}), as well as the alkylbenzenes propylbenzene, butylbenzene, amylbenzene, and hexylbenzene (Sigma, St. Louis, MO, USA), were prepared at different concentrations in dichloromethane (99.9% purity, Honeywell Burdick and Jackson, Morristown, NJ, USA). All compounds were present at the same concentration (0.05, 0.1, 0.5, 1.0, 5.0 mM) in each standard.

To verify that the ion abundance followed a normal (Gaussian) distribution, the 1.0 mM standard was analyzed with a large number of replicates (n = 30). For the statistical tests, two sets of replicate (n = 3) standards at all five concentrations, hereafter designated as Set 1 and Set 2, were analyzed sequentially on the same day. All standards were analyzed using a gas chromatograph (model 6890N, Agilent Technologies, Santa Clara, CA, USA) equipped with a DB-5MS column (30 m × 0.25 mm i.d. × 0.25 µm film thickness, Agilent Technologies) and an automatic liquid sampler (model 7683B, Agilent Technologies). Ultra-high purity helium (Airgas Great Lakes, Independence, OH, USA) was used as the carrier gas at a nominal flow rate of 1 mL/min. The inlet was maintained at 250 °C and 1 µL of the standard was injected in splitless mode. The oven temperature program was as follows: 40 °C for 2 min, 15 °C/min to 280 °C, with a final hold at 280 °C for 2 min. The transfer line to the mass-selective detector (model 5975C, Agilent Technologies) was maintained at 300 °C. Electron ionization (70 eV) was used and the quadrupole mass analyzer was operated in the full scan mode (m/z 40–550) with a scan rate of 2.86 scans/s and an instrumental peak threshold of 150. The 1.0 mM standard was also analyzed using ionizing voltages of 50 and 90 eV under the same conditions.

Several considerations involving the sample compound and GC/MS instrument are essential to the accuracy of the statistical association and discrimination of mass spectra. The sample compound must be both chemically and thermally stable, as well as sufficiently concentrated to produce a representative mass spectrum. The column and GC temperature program should be chosen such that sample compounds are baseline resolved. To insure that the mass spectra are reproducible, the instrument must be clean and well maintained. The septum and chromatographic column should be low bleed to minimize extraneous background ions. Constant instrumental parameters, for example the electron ionization energy and tune conditions, should be used throughout the duration of data collection to minimize instrumental contributions to variance.

Data analysis

The mass spectra were exported from ChemStation Software (version E01.02.16, Agilent Technologies) to Microsoft Excel (version 2007, Microsoft Corp., Redmond, WA, USA). All calculations and logical functions were performed in Microsoft Excel. The exported data from ChemStation contain only the abundances of the ions present in the spectrum above the instrumental peak threshold. Therefore, to create a complete mass spectrum, the m/z value for each ion was rounded to its integer value and the corresponding abundance was tabulated for the entire mass scan range. For any ion not present in the mass spectrum, an abundance of 0 was entered.

Both the traditional method of calculating standard deviations and the predicted standard deviations previously mentioned were investigated. For the traditional method, the mean abundance and standard deviation were calculated for each ion in the triplicate mass spectra for each compound at each concentration. For the predicted standard deviation method, a logarithmic graph of standard deviation vs mean abundance for all mass spectra was created in Microsoft Excel. In so doing, each ion abundance was represented by a total of 90 mass spectra, thereby providing a more robust statistical approach than the traditional method. The linear least-squares regression line was then calculated and used to predict the standard deviations for all ions. For any ion at or below the instrumental threshold (150 counts), the standard deviation was predicted at an abundance of 150. The mean abundances and standard deviations were then normalized to the base peak for both methods.

All comparisons, unless otherwise stated, were based on the triplicate mass spectra in Sets 1 and 2. Comparisons were made between the same compound (e.g., C_{10} in Set 1 compared to C_{10} in Set 2) to examine statistical association. Comparisons were also made between different compounds (e.g., C_{10} in Set 1 compared to C_{11} in Set 2) to examine statistical differentiation.

For these comparisons, an unequal variance *t*-test was performed in which the Welch *t*-statistic, t_{calc}, and the associated degrees of freedom were calculated at each *m/z* value.[24] Over the scan range of *m/z* 40–550, this corresponded to 510 individual *t*-tests for each spectral comparison. The t_{calc} values were compared to the critical values, t_{crit}, at various confidence levels using a two-tailed table. Using an IF function in Excel, a value of 1 was returned if the mean abundance for each *m/z* value was statistically indistinguishable ($t_{calc} \leq t_{crit}$), and a value of 0 was returned if statistically different ($t_{calc} > t_{crit}$) at the specified confidence level. The two spectra were considered statistically associated if the product of these values was 1 (i.e., if values of 1 were returned for every *m/z* value) and were considered statistically different if the product was 0 (i.e., if a value of 0 was returned for any *m/z* value).

Random-match probability calculations

If the two spectra were statistically differentiated, the random-match probability (RMP) is, by definition, equal to zero. On the other hand, if the two spectra were statistically associated, the RMP was calculated in the following manner. A binary consolidated array, ***C***, of the two spectra was created, in which a value of 1 was returned at that *m/z* value if an ion was present in both spectra. Conversely, a value of 0 was returned if the ion was absent in both spectra. Although the spectra are statistically equivalent, it is still possible for an ion to be present in one spectrum (i.e., near the threshold) but not in the other (i.e., below the threshold). In such cases, to be conservative, the ion was not included in the RMP calculation.

The NIST Mass Spectral Search Program (version 2.0d, Gaithersburg, MD, USA), which contains 147,198 mass spectra collected using electron ionization at 70 eV, was used as a representative database to determine the frequency of fragment ions. The number of spectra in the database containing each ion in the mass scan range above 1% threshold (the lowest threshold allowed) was tabulated using the search function in the NIST program. The probability of each ion in ***C*** with a value of 1 was calculated by Eqn. (3), while the probability of each ion in ***C*** with a value of 0 was calculated by Eqn. (4). The total probability of array ***C*** was then calculated by Eqn. (5) to obtain the RMP. This represents the probability that the pattern of ions occurs by random chance alone.

Ions that are known to be chemically irrelevant were removed from the RMP calculations in two ways. The mass scan range *m/z* 40–550 was chosen to avoid common atmospheric compounds (such as H_2O, N_2, O_2, etc.) that are not removed completely by the vacuum pump. In addition, common contaminant ions, such as those from column or septum degradation (e.g., *m/z* 73, 147, 207, 221, 281, 295, 355, 429) and fluorinated hydrocarbons used for mass tuning (e.g., *m/z* 69, 219, 502), were ignored in the RMP calculations if below a user-defined value (e.g., 5% of the base peak). If above this value, these ions were assumed to be chemically relevant to the compound and were included in the calculation.

RESULTS AND DISCUSSION

As noted previously, homologous series of *n*-alkanes and alkylbenzenes were chosen to develop and validate this method. Each series has very similar mass spectra, as evidenced by their Pearson product-moment correlation coefficients.[24] The coefficients for pair-wise comparisons of the same *n*-alkane (Set 1 and Set 2) ranged from 0.9611 to 1.000, whereas those for different *n*-alkanes ranged from 0.9189 to 0.9973. The coefficients for pair-wise comparisons of the same alkylbenzene (Set 1 and Set 2) ranged from 0.9679 to 1.000, whereas those for different alkylbenzenes ranged from 0.7580 to 0.9542. These values indicate a strong correlation for *n*-alkane spectra, and a moderate to strong correlation for alkylbenzene spectra.[24,28] Hence, these compounds provide a rigorous test of the ability of the proposed method to associate and discriminate mass spectra.

Verification of normal (Gaussian) distribution

For any hypothesis test (e.g., Eqns. (1) and (2)) to be valid, the data must conform to the probability distribution used to establish the critical values of the test statistic. Although there is no reason to suspect that ion abundances in mass spectrometry are not normally distributed, it is beneficial to demonstrate this fundamental fact. To do so, the 1.0 mM standard was analyzed with a large number of instrumental replicates (n = 30). From these data, the mean abundance (μ) and standard deviation (σ) were calculated for representative ions at *m/z* 43, 57, 71, and 85 for each of the *n*-alkanes. The standard deviate ($z = (x - \mu)/\sigma$) was calculated for each experimentally measured abundance, which was then compared to the theoretical value in a normal probability plot.[24] If the data are normally distributed, this graph is linear with a slope of unity, an intercept of zero, and a high correlation coefficient. A representative example for C_{10} is shown in Fig. 1(A). For the individual ions (n = 30), Fig. 1(A) shows linear plots with slopes ranging from 0.9805 to 0.9895 (all statistically indistinguishable from unity at the 95% confidence level), intercepts ranging from -3.1×10^{-17} to 5.5×10^{-16} (all statistically indistinguishable from zero at the 95% confidence level), and correlation coefficients (r^2) ranging from 0.9694 to 0.9872. Similar results were obtained for the other *n*-alkanes, as summarized in Supplementary Table SI.1 (see Supporting Information). Slightly better statistics were obtained when the ions were combined for all *n*-alkanes. A representative example for *m/z* 43 is shown in Fig. 1(B). For the combined ions (n = 180), Fig. 1(B) shows a linear plot with slope of 1.0103 (statistically indistinguishable from unity), intercept of 5.5×10^{-16} (statistically indistinguishable from zero), and r^2 of 0.9938. In all cases, these results are consistent with a normal (Gaussian) distribution.

As a further test of normality, the Shapiro-Wilk (W), Anderson-Darling (A^2), Lilliefors (D), and Jarque-Bera (JB) tests may be applied to these ion abundance data (XLSTAT, version 2013.4.07, Addinsoft, New York, NY, USA). As each of these tests is sensitive to different aspects of the distribution (small vs large populations, deviations in the center vs wings, etc.), they provide a thorough and detailed verification of normality.[29] The results of these tests are summarized in Supplementary Table SI.2 (see Supporting Information). For ions at *m/z* 43, 57, 71, and 85 for each of

Rapid Communications in Mass Spectrometry

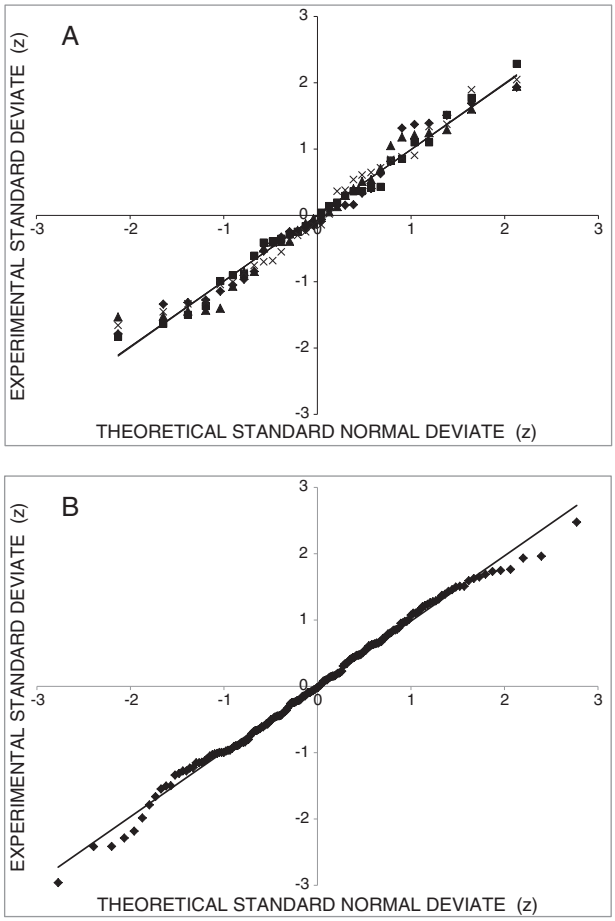

Figure 1. Experimental standard deviate vs theoretical standard normal deviate for mass spectral ion abundance. (A) C_{10}, *m/z* 43 (◆, slope = 0.9808, intercept = 5.5×10^{-16}, $r^2 = 0.9701$), *m/z* 57 (■, slope = 0.9895, intercept = 1.7×10^{-16}, $r^2 = 0.9872$), *m/z* 71 (▲, slope = 0.9805, intercept = 3.5×10^{-16}, $r^2 = 0.9694$), *m/z* 85 (×, slope = 0.9845, intercept = -3.1×10^{-17}, $r^2 = 0.9774$). (B) *m/z* 43 combined for C_{10}, C_{11}, C_{12}, C_{13}, C_{14}, C_{16} (slope = 1.0103, intercept = 5.5×10^{-16}, $r^2 = 0.9938$).

the *n*-alkanes, all four statistical tests are passed at the 95% confidence level ($P \geq 0.05$). Hence, these results are fully consistent with a normal (Gaussian) distribution.

Association of the same *n*-alkane

The spectra of corresponding *n*-alkanes in Sets 1 and 2 were compared at a concentration of 1.0 mM and an ionizing voltage of 70 eV. All pair-wise statistical comparisons (6 total) were made using a *t*-test at confidence levels of 98.0, 99.0, and 99.9%. In general, corresponding *n*-alkanes were statistically indistinguishable at the 99.9% confidence level and, therefore, were considered to be associated. However, this confidence level is the least rigorous with regard to statistical association (i.e., minimizes Type II error), while the lower confidence levels are more precise.[24] The lowest confidence level at which association was maintained for corresponding *n*-alkanes is reported in Table 1. In these comparisons, C_{11} and C_{14} were associated only at the 99.9% confidence level, while C_{12} was also associated at the 99.0% confidence level. C_{10} and C_{16} maintained association at all confidence levels investigated.

Table 1. Pair-wise comparison of the same *n*-alkane in Set 1 and Set 2 using a *t*-test at the lowest confidence level (CL) for which association was maintained, together with the corresponding range of *P*-values and random-match probability (RMP)

Alkane	CL	Range of *P*-values	RMP
C_{10}	98.0%	0.0299–1.000	2.0×10^{-39}
C_{11}	99.9%	0.0083–1.000	2.4×10^{-41}
C_{12}	99.0%	0.0184–1.000	2.6×10^{-42}
C_{13}	–	0.00004–1.000	–
C_{14}	99.9%	0.0077–1.000	3.1×10^{-44}
C_{16}	98.0%	0.0420–1.000	1.1×10^{-45}

This indicates that association was possible for the *n*-alkanes, but the degree of rigorousness varied. More detailed information can be obtained from the range of *P*-values for each of the individual ion comparisons, also summarized in Table 1. For example, C_{10} had one ion (*m/z* 79) with low abundance (585 ± 23 and 385 ± 83 for Set 1 and Set 2, respectively) that limited the nominal confidence level to 98.0% ($P = 0.0299$). The next limiting ion (*m/z* 81), also with low abundance (417 ± 62 and 296 ± 37 for Set 1 and Set 2, respectively), had a *P*-value of 0.0452 (nominal 96.0% confidence level). All other ions in the range of *m/z* 40–550 had *P*-values greater than 0.1000 (nominal 90.0% confidence level), which are generally accepted to have no evidence against the null hypothesis (Eqn. (1)). It is noteworthy that C_{13}, which was not associated at the 99.9% confidence level as the other *n*-alkanes, was limited by a single ion (*m/z* 85) with an anomalously low standard deviation (67,640 ± 160 and 66,011 ± 117 for Set 1 and Set 2, respectively) and *P*-value of 0.000035. This issue of low and unrepresentative standard deviation will be discussed in a following section.

The RMPs were calculated for all spectra that were statistically associated (Table 1) and represent the probability that the specific ion fragmentation pattern occurs by chance. As an example, the RMP for C_{10} alkanes was 2.0×10^{-39}, indicating that the possibility of random error is infinitesimally small. As carbon number increases, the RMP systematically decreases (e.g., for comparison of C_{16} spectra, the RMP is 1.1×10^{-45}). The larger *n*-alkanes have a greater number of discriminating ions and, therefore, a lower probability that the fragmentation pattern occurs by random chance.

Discrimination of different *n*-alkanes

The spectra of different *n*-alkanes in Sets 1 and 2 were compared at a concentration of 1.0 mM and an ionizing voltage of 70 eV. All pair-wise statistical comparisons (30 total) were made using a *t*-test at confidence levels of 98.0, 99.0, and 99.9%. Each *n*-alkane was statistically distinguishable from all others at the 99.9% confidence level, which is the most rigorous test for statistical discrimination (i.e., minimizes Type I error). Hence, despite the similarity of the spectra (as evidenced by the Pearson correlation coefficients above), these *n*-alkanes were still distinguishable using the unequal variance *t*-test.

The number and *m/z* value of ions responsible for discriminating the *n*-alkane spectra are reported in Table 2 at the 99.9% confidence level. The number of discriminatory ions ranged from 1 to 24 ions, depending on the *n*-alkanes being

Table 2. Ions responsible for discrimination of *n*-alkanes in Set 1 and Set 2 (*t*-test, 99.9% CL)

Set 1	Set 2	Ions	*m/z*	% Even *m/z*	% Low abundance*	M^+ ions
C_{10}	C_{11}	3	83, 142, 156	67%	33%	C_{10}, C_{11}
	C_{12}	8	42, 43, 71, 83, 85, 141, 142, 170	38%	63%	C_{10}, C_{12}
	C_{13}	15	42, 43, 71, 82, 83, 84, 85, 97, 112, 126, 127, 140, 141, 142, 184	53%	67%	C_{10}, C_{13}
	C_{14}	18	42, 43, 56, 70, 71, 82, 83, 84, 85, 96, 97, 112, 126, 127, 140, 141, 142, 198	61%	67%	C_{10}, C_{14}
	C_{16}	18	42, 43, 56, 69, 70, 71, 72, 82, 83, 84, 85, 96, 111, 112, 125, 140, 141, 142	56%	50%	C_{10}
C_{11}	C_{10}	1	71	0%	0%	–
	C_{12}	3	140, 141, 170	67%	33%	C_{12}
	C_{13}	6	55, 83, 85, 140, 142, 184	50%	50%	C_{13}
	C_{14}	11	41, 71, 83, 85, 97, 99, 111, 113, 140, 155, 198	18%	64%	C_{14}
	C_{16}	12	41, 53, 55, 69, 71, 83, 85, 97, 99, 111, 125, 155	0%	67%	–
C_{12}	C_{10}	10	56, 68, 70, 71, 82, 84, 85, 97, 126, 127	60%	50%	–
	C_{11}	3	71, 97, 156	33%	67%	C_{11}
	C_{13}	2	126, 184	100%	50%	C_{13}
	C_{14}	9	71, 85, 86, 96, 99, 125, 126, 140, 198	56%	44%	C_{14}
	C_{16}	16	41, 42, 55, 56, 67, 69, 71, 82, 85, 97, 98, 99, 111, 113, 125, 140	31%	69%	–
C_{13}	C_{10}	7	42, 71, 84, 85, 97, 141, 155	29%	57%	–
	C_{11}	4	85, 141, 155, 156	25%	50%	C_{11}
	C_{12}	4	42, 83, 155, 170	50%	75%	C_{12}
	C_{14}	3	99, 125, 198	33%	67%	C_{14}
	C_{16}	5	97, 99, 111, 125, 154	20%	40%	–
C_{14}	C_{10}	15	42, 68, 69, 71, 83, 84, 85, 97, 111, 112, 126, 140, 141, 155	40%	53%	–
	C_{11}	16	42, 69, 71, 81, 84, 85, 97, 98, 111, 126, 140, 141, 155, 156, 169, 198	44%	56%	C_{11}, C_{14}
	C_{12}	9	83, 85, 111, 140, 141, 155, 169, 170, 198	33%	44%	C_{12}, C_{14}
	C_{13}	6	99, 110, 140, 169, 184, 198	67%	50%	C_{13}, C_{14}
	C_{16}	5	56, 71, 99, 125, 198	40%	80%	C_{14}
C_{16}	C_{10}	23	40, 41, 42, 43, 56, 67, 69, 70, 71, 72, 82, 83, 84, 85, 97, 99, 111, 125, 140, 168, 182, 183, 226	48%	52%	C_{16}
	C_{11}	24	41, 42, 43, 56, 67, 69, 71, 82, 83, 85, 97, 98, 99, 100, 113, 125, 126, 140, 154, 156, 168, 182, 183, 226	50%	63%	C_{11}, C_{16}
	C_{12}	15	41, 42, 43, 83, 85, 97, 99, 100, 113, 125, 140, 170, 182, 183, 226	40%	67%	C_{12}, C_{16}
	C_{13}	13	83, 85, 97, 99, 100, 111, 140, 154, 168, 182, 183, 184, 226	54%	46%	C_{13}, C_{16}
	C_{14}	10	83, 97, 113, 125, 168, 170, 182, 183, 198, 226	50%	60%	C_{14}, C_{16}
%	**Total**			**44** %	**54** %	

*≤ 5% of base peak, *m/z* value of ion is underlined

Rapid Communications in Mass Spectrometry

compared. This number is somewhat surprising, given the similarity of the fragmentation patterns and the total number of ions that comprise the *n*-alkane spectra (46 to 71 for C_{10} to C_{16}, respectively). Additionally, 54% of the discriminating ions were of low abundance, which was arbitrarily defined as <5% of the base peak in this study. Lower abundance ions may be more characteristic of the compound and, therefore, contribute to discrimination. This emphasizes the importance of using the full spectra rather than abbreviated spectra composed of only the most abundant ions.

The ions responsible for discrimination among the spectra were further examined for general trends (Table 2). The number of discriminating ions increased as the difference in carbon number of the *n*-alkanes being compared increased. Ions with even *m/z* values represented 44% of the total discriminating ions, while odd *m/z* values represented 56%. In non-nitrogen-containing compounds such as the *n*-alkanes, even-numbered fragments are less common and generally result from multiple-bond cleavage, indicating that rearrangement may have occurred.[13] Therefore, the presence of these fragments indicates that, in 44% of the comparisons, differentiation was based on rearrangement and other less common cleavage patterns.

In most comparisons, the molecular ion was among the fragments leading to discrimination of the *n*-alkanes. However, as electron ionization is a hard ionization technique, it often does not result in a high abundance of the molecular ion. Therefore, the molecular ion is not always present among the discriminating ions, and was not observed in most comparisons involving C_{10}.[30]

This application of the *t*-test for the spectral comparison appears to be extremely rigorous with regard to discrimination, thereby minimizing false positives (Type I errors). However, it is less rigorous with regard to association of spectra of the same compound, and could potentially result in false negatives (Type II errors).

Association and discrimination of alkylbenzenes

To investigate the effectiveness of the proposed method for simple aromatic compounds, the mass spectra of alkylbenzenes were compared at a concentration of 1.0 mM and an ionizing voltage of 70 eV. Again, all pair-wise comparisons (16 total) were made using a *t*-test at confidence levels of 98.0, 99.0, and 99.9%. When spectra of corresponding alkylbenzenes in Set 1 and Set 2 were compared, association was possible at the 99.9% confidence level. When spectra of different alkylbenzenes were compared, discrimination was possible at the 99.9% confidence level, with 8 to 18 ions responsible for discrimination (Table 3). Approximately 56% of the discriminating ions were of low abundance (<5% of the base peak), which is comparable to those in the *n*-alkane spectra (54%) discussed above. This further emphasizes that the full spectra are essential for successful comparisons.

Electron ionization of aromatic compounds generally leads to stable and characteristic molecular ions, which were, in almost all cases, among the discriminating ions.[28] In addition, common fragment ions for alkylbenzenes, such as $C_7H_8^+$ (*m/z* 92), arising from the McLafferty rearrangement, and $C_8H_9^+$ (*m/z* 105), were also among the discriminating ions.[30]

Effect of ionizing voltage on association and discrimination

As noted above, spectra being compared should always be acquired under the same instrumental conditions. However, small variations in the ionizing voltage of the mass spectrometer are possible over time. To investigate the effect of changes in ionizing voltage, the 1.0 mM *n*-alkane standard was analyzed in replicate at voltages of 50, 70, and 90 eV. Spectra of each compound collected at 50 eV typically had 5 to 9 fewer ions than those collected at 70 eV. In contrast, spectra collected at 90 eV were more comparable, with 4 fewer to 2 more ions than those collected at 70 eV.

When spectra collected at voltages of 70 and 90 eV were compared using the *t*-test at the 99.9% confidence level (data not shown), statistical association of corresponding *n*-alkanes in Set 1 and Set 2 was maintained in all cases but one. For C_{12}, the spectra were differentiated by one low-abundance ion at *m/z* 51, which was not observed at 70 eV and was only 0.4% of the base peak at 90 eV. The higher ionizing voltage appears to have caused additional fragmentation for C_{12} that led to this additional ion.

In contrast, spectra collected at voltages of 50 and 70 eV were statistically distinct, with 1 to 5 ions responsible for discrimination of corresponding *n*-alkanes. Distinction was mainly due to variation in ion abundance relative to the base peak. For all *n*-alkanes, this was most noticeable at *m/z* 43, for which the relative abundance was more than 13% greater at 50 eV than at 70 eV. For C_{10}, this variation in abundance caused a change in the base peak, which was *m/z* 43 at 50 eV, but *m/z* 57 at 70 eV.

These results indicate that statistical association of spectra is relatively insensitive to voltage increases up to 20 eV greater than 70 eV, but is sensitive to decreases up to 20 eV. These variations are far greater than would be expected in normal operation.

Table 3. Number of discriminating ions for pair-wise comparison of alkylbenzenes in Set 1 and Set 2 (*t*-test, 99.9% CL). Zero discriminating ions indicate complete association and the corresponding random-match probability (RMP) is shown in parentheses

	Propyl	Butyl	Amyl	Hexyl
Propyl	0 (1.4×10^{-29})	12	13	17
Butyl	10	0 (7.1×10^{-36})	8	13
Amyl	17	10	0 (4.3×10^{-36})	8
Hexyl	18	14	11	0 (1.6×10^{-38})

Table 4. Effect of concentration and base peak abundance on the number of discriminating ions for the pair-wise comparison of C_{10} in Set 1 to all *n*-alkanes in Set 2 (*t*-test, 99.9% CL). Zero discriminating ions indicate complete association and the corresponding random-match probability (RMP) is shown in parentheses. Entries in bold highlight unexpected association or discrimination, as discussed in the text

			Set 1 C_{10}		
Set 2	**Concentration (mM)**	**Base peak abundance***	**48,955 ± 3,406**	**97,632 ± 2,054**	**711,381 ± 21,691**
C_{10}	0.5	49,261 ± 3,126	0 (1.7×10^{-39})	0 (1.8×10^{-39})	**3**
	1.0	96,971 ± 11,127	0 (9.1×10^{-39})	0 (2.0×10^{-39})	**1**
	5.0	758,485 ± 52,032	**3**	**3**	0 (1.4×10^{-40})
C_{11}	0.5	49,752 ± 1,881	3	6	9
	1.0	99,757 ± 4,069	1	3	6
	5.0	768,555 ± 33,173	6	9	5
C_{12}	0.5	66,851 ± 4,524	5	7	8
	1.0	132,051 ± 11,330	4	8	8
	5.0	1,125,547 ± 9,800	11	15	22
C_{13}	0.5	78,760 ± 2,440	3	5	6
	1.0	154,944 ± 3,294	7	15	15
	5.0	1,282,901 ± 28,747	10	22	27
C_{14}	0.5	94,112 ± 7,132	3	2	3
	1.0	205,333 ± 7,039	13	18	19
	5.0	1,480,021 ± 91,292	16	24	32
C_{16}	0.5	80,040 ± 8,993	8	12	11
	1.0	170,709 ± 3,568	10	18	20
	5.0	1,348,779 ± 10,447	26	33	44

*± one standard deviation, n = 3

Effect of concentration on association and discrimination

The *n*-alkane standards at different concentrations were analyzed at an ionizing voltage of 70 eV. The concentrations ranged from 0.5 to 5.0 mM, with the abundance of the base peak (*m/z* 57) ranging from 50,000 to 1,500,000 counts, respectively. All pair-wise statistical comparisons (324 total) were made using the *t*-test at the 99.9% confidence level. As a representative example, the C_{10} spectra in Set 1 are compared to all *n*-alkane spectra in Set 2, as summarized in Table 4.

In each case, when C_{10} spectra in Sets 1 and 2 were compared at the same concentration, association was possible. In contrast, when C_{10} spectra were compared at different concentrations, association of the spectra was not possible in most cases (*vide infra*), with 1 to 3 discriminating ions. However, when C_{10} spectra were compared with those of the other *n*-alkanes, they were statistically distinct at all concentrations, with 1 to 44 ions responsible for discrimination.

Spectra of *n*-alkane standards at lower concentrations of 0.05 and 0.1 mM were also investigated (data not shown). For these concentrations, the abundance of the base peak (*m/z* 57) ranged from 1500 to 4000 counts, respectively. Spectra with base peaks below 5000 counts could not be accurately associated or discriminated, which is potentially due to the smaller number of ions in the spectra. For example, C_{10} spectra with a base peak of approximately 100,000 counts (corresponding to a concentration of 1.0 mM) contained 56 ions, while C_{10} spectra with base peaks of 1500 and 4000 counts (corresponding to 0.05 and 0.1 mM) contained only 14 and 25 ions, respectively. In addition, the molecular ion, which is generally responsible for discrimination, is not observed in spectra with base peaks of 1500 counts and is only slightly above the instrumental threshold (~300 counts) in spectra with base peaks of 4000 counts. The loss in association and discrimination is understandable, as ions that are uniquely characteristic of the compound are missing from the spectra with base peaks below 5000 counts. As noted previously, these low-abundance ions account for more than 50% of the discriminating ions. The remaining ions are found at similar abundance ratios in the other *n*-alkanes and, therefore, do not allow discrimination. Thus, at very low abundances (base peak <5000 counts), the spectrum is no longer representative of the compound. Increasing the abundance, either by increasing the injection volume or concentration or by decreasing the split ratio, is necessary for accurate association or discrimination.

As noted above, rigorous discrimination of C_{10} from the other *n*-alkanes was possible at the three higher concentrations (0.5, 1.0, and 5.0 mM). However, spectra of C_{10} in Set 1 were not statistically associated to spectra of C_{10} in Set 2 at the 5.0 mM concentration (Table 4). Similar results were observed with the comparison of the alkylbenzenes at varying concentrations. In these cases, statistical association of corresponding compounds is most likely to incur Type II error if the inherent instrumental variation is not represented adequately. For example, when data were collected on the same day, the mean and standard deviation of the base peak in replicate C_{10} spectra were 97632 ± 2054 in Set 1 and 96971 ± 11127 in Set 2. When data were collected one and

Rapid Communications in Mass Spectrometry

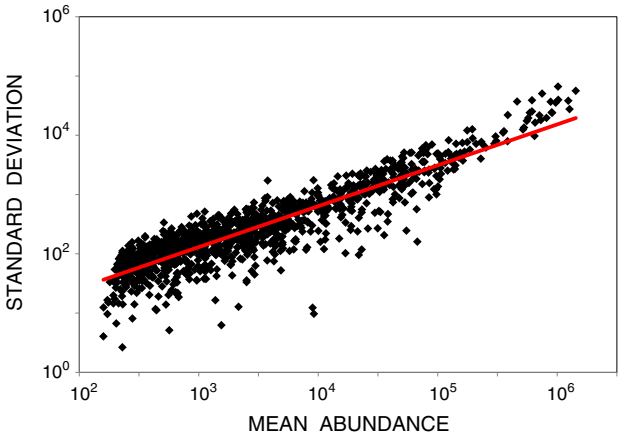

Figure 2. Logarithmic graph of standard deviation vs mean abundance for all ions in replicate mass spectra of *n*-alkanes (90 spectra, 1237 ions). Solutes C_{10}, C_{11}, C_{12}, C_{13}, C_{14}, C_{16}; concentrations 0.05, 0.1, 0.5, 1.0, 5.0 mM; ionizing voltage 70 eV. Linear best fit line with slope = 0.6900 ± 0.0093, intercept = 0.0440 ± 0.0332, and $r^2 = 0.8099$.

three weeks later, the cumulative mean and standard deviation were 79168 ± 39032 and 68040 ± 54770, respectively. Thus, standard deviations for replicate spectra calculated using the traditional method are not representative of the short-term and long-term instrumental variations encountered in routine use. Moreover, even greater instrumental variations could occur when replacing the injector septum or liner, retuning the mass spectrometer, or performing any other maintenance that requires venting the mass spectrometer.[30]

Effect of predicted standard deviation on association and discrimination

To address this problem, it is possible to create a mathematical model to predict standard deviations. The electron multiplier response is based on simple counting statistics and is statistically predictable. The variations in response are proportional to the square root of the abundance under shot-noise limited conditions.[15] Standard deviations predicted in this manner only require knowledge of the ion abundance and are independent of the compound being analyzed as well as its concentration, injection volume, split ratio, etc.

To model the electron multiplier response, a logarithmic graph of standard deviation versus mean abundance was generated that contained each ion in replicate spectra (n = 3) of the six *n*-alkanes at all five concentrations in Set 1 (90 spectra, 1237 ions). From this logarithmic graph (Fig. 2), the standard deviation is proportional to abundance in a manner similar to that expected for shot-noise limits (slope = 0.5).[a] A least-squares linear regression was performed and the resulting best-fit line had a slope of 0.6900 ± 0.0093, an intercept of 0.0440 ± 0.0332, and a correlation coefficient of 0.8099. To test the reproducibility, this procedure was repeated for the *n*-alkane data in Set 2, resulting in a slope of 0.6897 ± 0.0100, an intercept of 0.0395 ± 0.0358, and a correlation coefficient of 0.7905. The slopes from Set 1 and Set 2 were not statistically distinguishable (95% confidence level); the intercepts were not distinguishable from one another and, as theoretically expected, were also not distinguishable from zero (95% confidence level). This demonstrates the statistical validity and reproducibility of the regression equation. Using this regression equation, the standard deviation can be predicted for any given ion abundance in any spectrum acquired with this mass spectrometer.

These predicted standard deviations were used for the pair-wise comparisons (324 total) of all *n*-alkane spectra at each concentration using the *t*-test at the 99.9% confidence level. As a representative example, the C_{10} spectra in Set 1 are compared to all *n*-alkane spectra in Set 2, as summarized in Table 5. In contrast to the results obtained using traditional standard deviations above, spectra of corresponding *n*-alkanes were associated and spectra of different *n*-alkanes were discriminated in nearly all cases. However, for a few *n*-alkanes with sequential carbon numbers, in which at least one was at the lower concentration (0.5 mM), the spectra were not discriminated at the 99.9% confidence level. This case is illustrated in Table 5 for C_{10} and C_{11}, where the spectra were discriminated at the 99.0% confidence level. In these cases, the molecular ion in the spectrum at the lower concentration was not statistically distinguishable above the instrumental threshold and, hence, could not provide discrimination. If this were to occur in a practical application, the compound should be re-analyzed using higher concentration, larger injection volume, or lower split ratio to allow for differentiation from compounds with similar fragmentation patterns. In general, however, it appears that the predicted standard deviation method is more reliable than the traditional method for spectral association and discrimination, provided that the spectra are representative of the compound.

As can be observed in Table 5, the spectra of *n*-alkanes at higher concentrations have many discriminating ions, indicating that the spectra are readily differentiated. As concentration of the *n*-alkanes decreases for either of the spectra being compared, the number of discriminating ions also decreases. The same general trend can be observed in regard to carbon number; i.e., as the carbon number decreases, the number of discriminating ions also decreases. Given the similarity of the fragmentation pattern for the *n*-alkanes, it is noteworthy that the proposed method can still identify up to 45 discriminating ions at the highest confidence level of 99.9%. At lower confidence levels, the number of discriminating ions is even greater; i.e., up to 69 and 475 ions at the 99.0% and 98.0% confidence levels, respectively.

It is interesting to note the reasons for the greater success of the predicted standard deviation method. As observed in Fig. 2, a number of individual ions have standard deviations that are much lower than others of the same abundance. As noted previously, this underestimation of the standard deviation can occur when replicates do not adequately represent the intrinsic instrumental variation. Most of these outliers occur for ions at abundances less than 100,000. Since discrimination relies heavily on low-abundance ions, these ions fail the *t*-test when using the traditional method of calculating standard deviations. In contrast, the predicted

[a]Upon closer inspection of Fig. 2, there are potentially two linear regions with different slopes. At mean abundances less than 10^4, the slope is ~0.5 (shot noise), whereas at abundances greater than 10^4, the slope approaches ~1.0 (proportional noise). However, the results of the statistical hypothesis tests are no different than those described herein for a simple linear equation.

Table 5. Effect of concentration and base peak abundance on the number of discriminating ions for pair-wise comparison of C_{10} in Set 1 compared to all *n*-alkanes in Set 2 (*t*-test, 99.9% CL, unless otherwise specified) using predicted standard deviation. Zero discriminating ions indicate complete association and the corresponding random-match probability (RMP) is shown in parentheses

			Set 1 C_{10}		
Set 2	**Concentration (mM)**	**Base peak abundance***	**48,955 ± 3,406**	**97,632 ± 2,054**	**711,381 ± 21,691**
C_{10}	0.5	49,261 ± 3,126	0 (1.7×10^{-39})	0 (1.8×10^{-39})	0 (1.3×10^{-39})
	1.0	96,971 ± 11,127	0 (9.1×10^{-39})	0 (2.0×10^{-39})	0 (1.6×10^{-39})
	5.0	758,485 ± 52,032	0 (1.8×10^{-38})	0 (8.3×10^{-40})	0 (1.4×10^{-40})
C_{11}	0.5	49,752 ± 1,881	1	11[a]	2
	1.0	99,757 ± 4,069	2	2	3
	5.0	768,555 ± 33,173	4	4	8
C_{12}	0.5	66,851 ± 4,524	4	2	2
	1.0	132,051 ± 11,330	7	6	5
	5.0	1,125,547 ± 9,800	6	10	22
C_{13}	0.5	78,760 ± 2,440	6	5	4
	1.0	154,944 ± 3,294	10	12	10
	5.0	1,282,901 ± 28,747	11	17	32
C_{14}	0.5	94,112 ± 7,132	10	6	5
	1.0	205,333 ± 7,039	15	16	13
	5.0	1,480,021 ± 91,292	13	21	37
C_{16}	0.5	80,040 ± 8,993	15	13	9
	1.0	170,709 ± 3,568	22	25	21
	5.0	1,348,779 ± 10,447	22	30	45

*± one standard deviation, n = 3
[a]99.0% confidence level

Table 6. Number of discriminating ions for the pair-wise comparison of Set 1 *n*-alkanes to the National Institute of Standards and Technology (NIST) database *n*-alkanes (one sample *t*-test, 99.9% CL, unless otherwise specified). Zero discriminating ions indicate complete association and the corresponding random-match probability (RMP) is shown in parentheses. Entries in bold highlight unexpected association or discrimination, as discussed in the text

		NIST database					
Set 1	**Concentration (mM)**	**C_{10}**	**C_{11}**	**C_{12}**	**C_{13}**	**C_{14}**	**C_{16}**
C_{10}	0.5	0 (1.3×10^{-37})	2	3	1	5	6
	1.0	0 (2.8×10^{-38})	3	4	3	6	9
	5.0	0 (9.1×10^{-38})	6	10	11	20	31
C_{11}	0.5	1	0 (4.0×10^{-38})	1	14[a]	3	4
	1.0	2	0 (2.8×10^{-38})	2	2	5	6
	5.0	4	0 (8.6×10^{-39})	6	6	14	25
C_{12}	0.5	1	1	0 (1.8×10^{-38})	1	3	4
	1.0	2	1	0 (1.8×10^{-38})	1	5	6
	5.0	10	5	0 (9.5×10^{-40})	5	8	26
C_{13}	0.5	1	1	1	0 (2.0×10^{-38})	2	3
	1.0	3	2	2	0 (1.2×10^{-38})	3	5
	5.0	14	10	4	**7**	7	18
C_{14}	0.5	2	1	1	1	0 (6.6×10^{-41})	2
	1.0	4	2	2	1	0 (4.1×10^{-41})	4
	5.0	21	17	6	13	**1**	14
C_{16}	0.5	1	1	1	1	1	0 (7.9×10^{-42})
	1.0	3	2	2	1	2	0 (3.3×10^{-42})
	5.0	25	20	12	21	5	**6**

[a]99.0% confidence level

Rapid Communications in Mass Spectrometry

standard deviations represent the instrumental variation in a consistent and uniform manner. Moreover, once the model has been developed and validated, few or no replicates of standards and samples are required to determine the standard deviation and perform the statistical comparison. Because this method is more reliable, robust, and practical than the traditional method, it is recommended for use in the proposed statistical procedure.

Comparison to normal and branched alkanes in the NIST database

The most reliable statistical comparisons are obtained when mass spectra of questioned samples and authentic standards are analyzed on the same instrument, under the same conditions, at the same time. In this way, experimental and instrumental sources of variance are minimized, so that statistically significant chemical variations in the mass spectra are more easily discerned. However, for practical reasons, it may be desirable to compare questioned mass spectra to those in a reference database. Accordingly, the spectra of the *n*-alkanes acquired in this study were compared with reference spectra in the NIST database using the proposed method.[31] As only one spectrum of each *n*-alkane was available in the NIST database, this was compared to replicate spectra (n = 3) in Set 1 using a one-sample, two-tailed *t*-test[24] at confidence levels of 99.0 and 99.9%. For each pair-wise comparison (108 total), *n*-alkane spectra in Set 1 were statistically indistinguishable from spectra of corresponding *n*-alkanes in the NIST database at the 99.9%

Table 7. Number of discriminating ions for the pair-wise comparison of 1.0 mM Set 1 *n*-alkanes to the National Institute of Standards and Technology (NIST) database branched alkanes (one sample *t*-test, 99.9% CL, unless otherwise specified)

	n-C_{10}	*n*-C_{11}	*n*-C_{12}	*n*-C_{13}	*n*-C_{14}	*n*-C_{16}
C_{10} branched isomers						
2-methylnonane	2	1	2	3	4	4
3-methylnonane	2	4	3	3	4	4
2,3-dimethyloctane	2	2	3	3	3	3
2,4,6-trimethylheptane	3	3	4	5	4	4
2,2,5,5-tetramethylhexane	4	2	3	3	4	4
4-ethyloctane	2	3	4	5	5	6
C_{11} branched isomers						
2-methylnonane	6	3	4	4	4	2
3-methylnonane	4	3	2	2	3	3
2,3-dimethyloctane	1	19[a]	25[a]	1	1	1
2,4,6-trimethylheptane	3	2	4	4	6	5
2,2,5,5-tetramethylhexane	3	2	2	1	2	1
C_{12} branched isomers						
3-methylundecane,	2	3	3	4	4	4
3,8-dimethyldecane,	3	5	3	3	2	41[a]
2,2,3-trimethylnonane	4	4	4	4	5	6
2,2,7,7-tetramethyloctane	2	3	3	3	4	3
5-ethyldecane	3	3	2	3	3	2
C_{13} branched isomers						
3-methyldodecane	3	2	3	3	3	1
3,9-dimethylundecane	5	2	4	3	3	2
2,3,4-trimethyldecane	5	6	5	4	4	2
3-methyl-5-propylnonane	7	3	3	21[a]	32[a]	34[a]
4-ethylundecane	7	5	3	3	2	1
C_{14} branched isomers						
2-methyltridecane	7	3	4	5	3	2
6-methyltridecane	6	3	1	1	2	1
2,3-dimethyldodecane	12	6	7	7	4	3
4,6-dimethyldodecane	4	1	2	1	2	1
3,5-dimethyldodecane	4	3	1	2	1	35[a]
C_{16} branched isomers						
3-methylpentadecane	17	13	10	9	10	6
4,11-dimethyltetradecane	11	8	9	7	10	6
4-ethyltetradecane	11	9	8	7	9	7
5-ethyl-5-propylundecane	16	13	12	12	10	8
2,2,4,4,6,8,8-heptamethylnonane	2	2	3	4	4	4

[a]99.0% confidence level

confidence level (Table 6). There were three exceptions (5.0 mM concentrations of C_{13}, C_{14}, and C_{16}), in which ions in the NIST spectra with abundances near the threshold were statistically different from those in Set 1 with abundances below the threshold. For all spectra that were statistically associated to those in the NIST database, the random-match probabilities were calculated (Table 6). As an example, the RMP for C_{10} alkanes was 1.3×10^{-37} to 9.1×10^{-38}, indicating that the occurrence of this fragmentation pattern by random chance is infinitesimally small. Moreover, these RMP values are comparable to that calculated previously for the C_{10} alkanes in Set 1 and Set 2 (2.0×10^{-39}, Table 1).

The *n*-alkane spectra in Set 1 were statistically distinguishable from spectra of different *n*-alkanes in the NIST database at the 99.9% confidence level (Table 6), which is the most rigorous level for discrimination. There was one exception (0.5 mM C_{11} in Set 1 compared to C_{13} in the NIST database) where discrimination was not possible at the 99.9% confidence level but was achieved at the 99.0% confidence level. For all *n*-alkanes, the number of discriminatory ions ranged from 1 to 31 and, in nearly all cases, the molecular ion was the sole fragment or among the fragments leading to discrimination.

Whereas differentiation of normal alkanes can be challenging, the mass spectra of long-chain *n*-alkanes and their respective branched isomers are reported to be nearly indistinguishable.[32] As a further test and validation of the proposed method, the spectra of the six *n*-alkanes in Set 1 at 1.0 mM concentration were compared to reference spectra of 31 branched alkanes in the NIST database at confidence levels of 99.0 and 99.9%. Five isomers of each alkane, containing 1 to 7 methyl or ethyl branches, were investigated for a total of 186 comparisons (Table 7). Among these comparisons, 179 were statistically distinguished at the 99.9% confidence level, with 1 to 17 discriminating ions. The remaining 7 comparisons were statistically distinguished at the 99.0% confidence level, with 19 to 41 discriminating ions. The most common discriminating ions were *m/z* 71, 84/85, 98/99, 112/113, corresponding to the C_5 to C_8 fragments, and, where appropriate, the molecular ion.

The successful association and discrimination of the *n*-alkanes in Set 1 to those in the NIST database further demonstrates the power of this method, since these spectra were analyzed with different GC/MS instruments, as well as different experimental conditions, concentrations, and time periods.

CONCLUSIONS

A statistical method for comparing mass spectra of an unknown or questioned compound to a reference standard was developed using an *n*-alkane and alkylbenzene data set. At the same concentration, statistical association of corresponding compounds and discrimination of different compounds were possible at the 99.9% confidence level. For compounds that were statistically associated, the RMPs were on the order of 10^{-29} to 10^{-50}, indicating the low probability that the characteristic fragmentation patterns occur by random chance alone. At varying concentrations, discrimination of different *n*-alkanes was still possible, but association of corresponding *n*-alkanes was not possible using the traditional method to calculate standard deviations. In contrast, standard deviations predicted from a statistical model of the detector were more representative of short-term and long-term instrumental variance and allowed for association and discrimination of the *n*-alkanes at varying concentrations. In addition, using the predicted standard deviations, spectra of the *n*-alkanes were successfully associated to and discriminated from normal and branched alkane spectra in the NIST database, even though these spectra were collected on different instruments using different experimental conditions, and over different time periods.

While proof-of-concept in nature, the method developed and validated herein provides a simple and rapid approach to assign statistical confidence in the comparison of mass spectra. This method not only provides the confidence level for association and discrimination, but also the random-match probability for association. In principle, this method can also be applied to multistage MS, where each stage will increase the discriminating power in the mass spectral comparison. This method can be implemented without expensive software and is broadly applicable across many fields, including industrial, pharmaceutical, food, environmental, and forensic chemistries.

Acknowledgements

This work was supported by the National Institute of Justice through the Ames Laboratory under Contract No. DE-AC02-07CH11358.

REFERENCES

[1] Scientific Working Group on DNA Analysis Methods. *Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories*. National Institute of Justice, Washington, DC, **2010**.
[2] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC, **2009**.
[3] W. C. Brumley. Tools of the trade-separations and detections. US Environmental Protection Agency. http://www.epa.gov/esd/chemistry/org-anal/home.htm (retrieved June 26, 2012).
[4] W. C. Brumley, J. A. Sphon. Regulatory mass spectrometry. *Biomed. Mass Spectrom.* **1981**, *8*, 390.
[5] K. X. Wan, I. Vidavsky, M. L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85.
[6] S. E. Stein, D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859.
[7] F. W. McLafferty, R. H. Hertel, R. D. Villwock. Probability based matching of mass spectra, rapid identification of specific compounds in mixtures. *Org. Mass Spectrom.* **1974**, *9*, 690.
[8] H. S. Hertz, R. A. Hites, K. Biemann. Identification of mass spectra by computer-searching a file of known spectra. *Anal. Chem.* **1971**, *43*, 681.
[9] J. D. Crawford, J. D. Morrison. Computer methods in analytical mass spectrometry: identification of an unknown compound in a catalog. *Anal. Chem.* **1968**, *40*, 1464.

[10] I. Koo, S. Kim, X. Zhang. Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry. *J. Chromatogr. A* **2013**, *1298*, 132.
[11] K. Varmuza. Computer methods in mass spectrometry for chemical structure assignment, in *Encyclopedia of Spectroscopy and Spectrometry*, (2nd edn.), (Eds: J. Lindon, G. E. Tranter, D. Koppenaal). Academic Press-Elsevier, Waltham, MA, **2010**, pp. 392–403.
[12] ChemStation Software, version E01.02.16. Agilent Technologies, Santa Clara, CA.
[13] E. Hoffmann, V. Stroobant. *Mass Spectrometry: Principles and Applications*. John Wiley, Chichester, **2007**.
[14] S. E. Stein. *NIST Standard Reference Database 1A. Users Guide*. National Institute of Standards and Technology, Gaithersburg, MD, **2008**.
[15] S. E. Stein. An integrated method for spectrum extraction and compound identification from GC/MS data. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770.
[16] S. E. Stein. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem.* **2012**, *84*, 7274.
[17] Scientific Working Group for the Analysis of Seized Drugs. *Recommendations*, revision 6. U.S. Department of Justice, Drug Enforcement Administration, Washington, DC, **2011**. http://www.swgdrug.org/documents/SWGDRUG%20Recommendations%206.pdf (accessed July 15, 2013).
[18] Scientific Working Group for the Analysis of Seized Drugs, Supplemental Document SD-2 for Part IVB, *Quality Assurance/Validation of Analytical Methods*. U.S. Department of Justice, Drug Enforcement Administration, Washington, DC, **2006**. http://www.swgdrug.org/documents/Supplemental%20Document%20SD-2.pdf (accessed July 15, 2013).
[19] *Forensic Chemistry Section Quality Manual*, Document DRG-DOC-01. Arkansas State Crime Laboratory, Little Rock, AR, **2009**. http://www.crimelab.arkansas.gov/resources/documents/DRG-DOC-01.pdf (accessed July 15, 2013).
[20] *Controlled Substances Standard Operating Procedure*, revision 5. San Francisco Police Department Criminalistics Laboratory, San Francisco, CA, **2005**. http://www.cacj.org/documents/SF_Crime_Lab/Federal_Pleadings/Diaz_/SFPD-Crime-Lab-Controlled-Substances-SOP-Ver.-6-23-05.pdf (accessed July 15, 2013).
[21] *Controlled Substances Procedures Manual*, Document 221-D100. Virginia Department of Forensic Science, Richmond, VA, **2012**. http://www.dfs.virginia.gov/manuals/controlledsubstances/procedures/221-d100%20controlled%20substances%20procedures%20manual.pdf (accessed July 15, 2013).
[22] Daubert *v.* Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, **1993**.
[23] F. W. McLafferty. Mass spectrometric analysis: molecular rearrangements. *Anal. Chem.* **1959**, *31*, 82.
[24] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Duxbury Press, Belmont, CA, **1990**.
[25] V. Bafna, N. Edwards. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics.* **2001**, *17*(Suppl 1), 13.
[26] J. Colinge, A. Masselot, J. Magnin. A systematic statistical analysis of ion trap tandem mass spectra in view of peptide scoring, in *Algorithms in Bioinformatics*, (Eds: G. Benson, R. D. M. Page). Springer-Verlag, Berlin, **2003**, pp. 25–38.
[27] Y. Fu, Q. Yang, R. Sun, C. Ling, D. Li, H. Zhou, S. He, W. Gao. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **2004**, *20*, 1948.
[28] M. Bodnar-Willard. Development and application of a statistical approach to establish equivalence of unabbreviated mass spectra, *Dissertation*, Michigan State University, E. Lansing, MI, **2013**.
[29] J. Groβ. *A Normal Distribution Course*, Peter Lang, New York, **2004**.
[30] R. M. Smith. *Understanding Mass Spectra: a Basic Approach*, (2nd edn.). John Wiley, Hoboken, NJ, **2004**.
[31] P. Linstrom, W. Mallard. *NIST Chemistry WebBook*, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg, MD. http://webbook.nist.gov (accessed April 25, 2012).
[32] G. Issacman, K. R. Wilson, A. W. H. Chan, D. R. Worton, J. R. Kimmel, T. Nah, T. Hohaus, M. Gonin, J. Kroll, D. R. Worsnop, A. H. Goldstein. Improved resolution of hydrocarbon structures and constitutional isomers in complex mixtures using gas chromatography-mass spectrometry. *Anal. Chem.* **2012**, *84*, 2335.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

