BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1158–1163
doi:10.1093/bioinformatics/bts083

Data and text mining

Advance Access publication February 13, 2012

A method of finding optimal weight factors for compound identification in gas chromatography–mass spectrometry

Seongho Kim[1,*], Imhoi Koo[1,2], Xiaoli Wei[2] and Xiang Zhang[2,*]

[1]Department of Bioinformatics and Biostatistics and [2]Department of Chemistry, University of Louisville, Louisville, KY 40292, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The compound identification in gas chromatography–mass spectrometry (GC–MS) is achieved by matching the experimental mass spectrum to the mass spectra in a spectral library. It is known that the intensities with higher *m/z* value in the GC–MS mass spectrum are the most diagnostic. Therefore, to increase the relative significance of peak intensities of higher *m/z* value, the intensities and *m/z* values are usually transformed with a set of weight factors. A poor quality of weight factors can significantly decrease the accuracy of compound identification. With the significant enrichment of the mass spectral database and the broad application of GC–MS, it is important to re-visit the methods of discovering the optimal weight factors for high confident compound identification.

Results: We developed a novel approach to finding the optimal weight factors only through a reference library for high accuracy compound identification. The developed approach first calculates the ratio of skewness to kurtosis of the mass spectral similarity scores among spectra (compounds) in a reference library and then considers a weight factor with the maximum ratio as the optimal weight factor. We examined our approach by comparing the accuracy of compound identification using the mass spectral library maintained by the National Institute of Standards and Technology. The results demonstrate that the optimal weight factors for fragment ion peak intensity and *m/z* value found by the developed approach outperform the current weight factors for compound identification.

Availability: The results and R package are available at http:// stage.louisville.edu/ faculty/ x0zhan17/ software/ software-development.

Contact: s0kim023@louisville.edu; xiang.zhang@louisville.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 14, 2011; revised on February 3, 2012; accepted on February 9, 2012

1 INTRODUCTION

Gas chromatography–mass spectrometry (GC–MS) is one of the most widely used analytical techniques for unraveling a large number of compounds present in either chemical or biological samples. One of the most important analyses of GC–MS data is compound identification, which is currently achieved by matching the experimental mass spectra to the mass spectra recorded in a spectrum database. The mass spectrum of an unknown compound is assigned to a database compound based on their mass spectral similarity score. Various spectral similarity scores have been developed for the spectrum matching-based compound identification including composite similarity (Stein and Scott, 1994), probability-based matching system (Atwater *et al*., 1985), cosine correlation (Tabb *et al*., 2003); (Beer *et al*., 2004); (Craig *et al*., 2006); (Frewen *et al*., 2006), Hertz similarity index (Hertz *et al*., 1971), normalized Euclidean distance (L_2-norm) (Stein and Scott, 1994); (Julian *et al*., 1998); (Rasmussen *et al*., 1979), absolute value distance (L_1-norm) (Stein and Scott, 1994); (Rasmussen *et al*., 1979), wavelet and Fourier transforms-based composite measure (Koo *et al*., 2011).

The peak intensities of fragment ions with large mass-to-charge (*m/z*) values in a GC–MS mass spectrum tend to be smaller although they are the most informative ions for compound identification. The performance of compound identification can be improved by increasing the relative significance of the large fragment ions via weighing more on their peak intensities. Nearly 5% of improvement in identification accuracy has been observed if an appropriate set of weight factors is used (Stein and Scott, 1994). Several studies have been performed to discover the optimal weight factors for fragment ion peak intensity as well as *m/z* value. Sokolow et al. (1978) suggested the squared root of an intensity times its *m/z* value as an optimal scaling of the intensities, while Stein and Scott (1994) recommended an intensity to the power of 0.6 times its *m/z* value cubed in case of the cosine correlation. Recently, Horai *et al*. (2010) reported that optimal weight factor is the squared root of intensity and the square of its *m/z* value.

Even though various methods have been developed to improve mass spectral similarity measures for high accuracy compound identification, there has been much less development in finding the optimal weight factors to improve the performance of compound identification. The main objective of this study is to develop a method to discover the optimal weight factors for high accuracy compound identification in GC–MS. It is noteworthy that the literature reported methods require a training data set, whereas the proposed method can discover optimal weight factors based on only a reference library. The proposed approach focuses on the statistical characteristics of the distribution of mass spectral similarity scores among compounds in a reference library. The ratio of skewness to kurtosis is used to search for an optimal weight factor, considering the optimal weight factor having the maximum of the average ratios of skewness to kurtosis.

All the statistical analysis and simulations were performed using the R statistical software version 2.13.1 (R Development Core Team)

*To whom correspondence should be addressed.

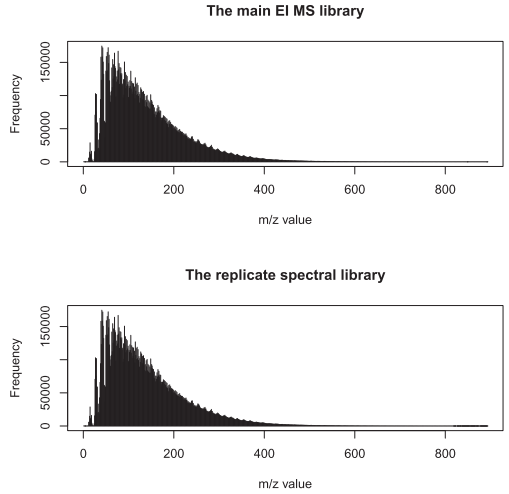

Fig. 1. The frequency of non-zero intensities with respect to *m/z* values for the main EI MS library and the replicate spectral library.

with the National Institute of Standards and Technology (NIST) mass spectral library. For the ease of description, we use the terms spectrum and compound interchangeably throughout this article.

2 METHODS

2.1 The main EI MS library and replicate spectral library

Currently, two different mass spectral databases are commonly used as references: the NIST/EPA/NIH Mass Spectral (MS) Library and the Wiley/NBS MS Database. In this study, we used the NIST/EPA/NIH MS Library developed at the NIST. We first extracted the main electron ionization (EI) MS library from the NIST 11 MS Library and obtained 212 961 spectra (compounds) whose *m/z* values range from 1 to 1760. The replicate spectral library was then extracted from the NIST 08 library. It contains 28 307 mass spectra generated by 18 569 compounds whose fragment ion *m/z* values range from 1 to 1036.

We considered the main EI MS library as a reference library and the replicate spectral library as query data. Since, we assume that the main EI MS library has the mass spectrum information for all query spectra, all spectra that are not present in the main EI MS library were removed from the replicate spectral library. After all, the reference library used in this work includes 212 860 spectra and the replicate spectral library has 28 162 query spectra. The *m/z* values of both libraries range from 1 to 892. The distributions of *m/z* values for the two libraries are depicted in Figure 1.

2.2 Cosine correlation and peak intensity weighting

Cosine correlation, C, which is also known as the dot product (Stein and Scott, 1994), is a measure of correlation between two sequences of intensities, $\alpha=(\alpha_i)_{i=1,\ldots,n}$ and $\beta=(\beta_i)_{i=1,\ldots,n}$, using the cosine value of angle. It is defined as

$$C=\frac{\alpha\circ\beta}{|\alpha|\cdot|\beta|}, \tag{1}$$

where $\alpha\circ\beta=\sum_{i=1}^{n}\alpha_i\beta_i$ and $|\alpha|=\sqrt{\sum_{i=1}^{n}\alpha_i^2}$.

The fragment ion peaks with large *m/z* values in a GC–MS spectrum usually have small peak intensities, but carry the most important characteristics for compound identification. Therefore, weighting peak intensity based on its *m/z* value can increase the contribution of small peaks to compound identification. Weighted peak intensity can be represented as

$$[\text{peak intensity}]^x\cdot[\text{mass(m/z)}]^y, \tag{2}$$

where x and y represent the weight factor of peak intensity and *m/z* value, respectively. Stein and Scott (1994) suggested the cosine correlation with an optimal intensity scaling of 0.6 (i.e. $x=0.6$) and mass weighting of 3 (i.e. $y=3$), whereas Sokolow *et al*. (1978) recommended the squared root of an intensity times its *m/z* value as an optimal scaling of the intensities (i.e. $w=(x,y)=(0.5,1)$). Recently, Horai *et al*. (2010) reported that optimal weight factors are the squared root of intensity and the square of its *m/z* value (i.e. $w=(x,y)=(0.5,2)$).

Then cosine correlation with weighted intensities can be calculated by

$$C(x,y)=\frac{\alpha_w\circ\beta_w}{|\alpha_w|\cdot|\beta_w|}, \tag{3}$$

where $w=(x,y)$ is a vector of weight factors of intensity and *m/z* value, respectively. In Equation (3), $\alpha_w=(\alpha_i^w)_{i=1}^n$ and $\beta_w=(\beta_i^w)_{i=1}^n$ are weighted intensities based on (2) and

$$\alpha_i^w=(\alpha_i)^x\cdot(z_i)^y \text{ and } \beta_i^w=(\beta_i)^x\cdot(z_i)^y \tag{4}$$

where z_i is the *m/z* value of i-th intensity, $i=1,2,\ldots,n$, and x and y are weight factors. Since cosine correlation is one of the most popular mass spectral similarity measures and was used in the NIST mass spectral library, we considered only cosine correlation in this work.

We used accuracy as the measure to evaluate the performance of compound identification of different weight factors. The accuracy is defined as the proportion of spectra identified correctly in query data. In other words, if a pair of unknown and reference spectra have the same CAS registry index, this spectrum pair is considered as a correct match. Otherwise, the match is incorrect. By counting all correct matches, the accuracy of compound identification can be calculated as

$$\text{Accuracy}=\frac{\text{Number of spectra matched correctly}}{\text{Number of spectra queried}}. \tag{5}$$

We denote accuracy for a weight factor $w=(x,y)$ as $A(x,y)$.

2.3 Skewness and kurtosis

Skewness is a measure of the symmetry of data distribution, which is computed as the third moment about the mean, and is zero if a distribution is symmetric. It takes on negative values for left-skewed data and positive values for right-skewed data. Kurtosis is a measure that indicates whether a data distribution is flat or peaked. It is computed as the fourth moment about the mean. For a normal distribution, the normalized kurtosis is zero and the raw kurtosis is three. Since the kurtosis of the normal distribution is 3, the kurtosis is defined by subtracting 3 to express excess kurtosis. If the kurtosis is >3, the distribution is more peaked than normal near the mode of the distribution and has thicker tails (Reimann *et al*., 2008).

The skewness and kurtosis are calculated for the mass spectral similarity scores. For M similarity scores, $C_1,C_2,\ldots,C_M$, the formulas for skewness and kurtosis are

$$\text{Skewness}=\frac{\sum_{j=1}^{M}(C_j-\overline{C})^3}{M\cdot\sigma^3} \tag{6}$$

$$\text{Kurtosis}=\frac{\sum_{j=1}^{M}(C_j-\overline{C})^4}{M\cdot\sigma^4} \tag{7}$$

where $\overline{C}=\frac{1}{M}\sum_{j=1}^{M}C_j$ and $\sigma=\sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(C_j-\overline{C})^2}$. Using (6) and (7), skewness and kurtosis of the mass spectral similarity scores among compounds in the reference library can be calculated.

2.4 Finding an optimal weight factor

In order for weight factors to be optimal in terms of accuracy of compound identification, they should be able to emphasize the inborn characteristics

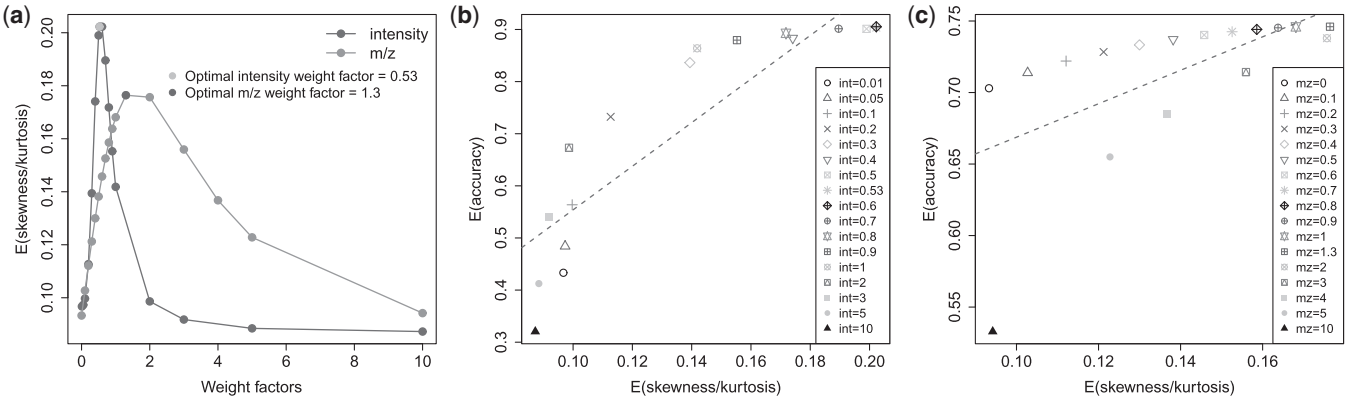

Fig. 2. The discovered optimal weight factors and the relationship between the ratio of skewness to kurtosis and accuracy for 100 randomly selected data sets. (**a**) The red line with solid circles represents the relationship between the conditional means of the ratio of skewness to kurtosis and the intensity weight factors, and the green line with solid circles the relationship between the conditional means of the ratios of skewness to kurtosis and the *m/z* weight factors. The sky blue and blue solid circles indicate the optimal intensity and *m/z* values, respectively. (**b**) The relationship between averages of the ratio of skewness to kurtosis ($\overline{R}(Y|x)$) and of accuracy ($\overline{A}(Y|x)$) with respect to the weight factors of intensities (x). The fitted line by the linear regression is denoted by the solid red line. (**c**) The relationship between the averages of the ratio of skewness to kurtosis ($\overline{R}(X|y)$) and of accuracy ($\overline{A}(X|y)$) with respect to the weight factors of *m/z* values (y). The fitted line is denoted by the dotted red line.

Table 1. Percentiles of accuracy of compound identification

		Min	Percentile								Max
			2.5th	25th	50th	75th	90th	95th	97.5th	99th	
Full	Set 1	1.52	17.65	34.56	60.30	76.40	80.06	81.30	81.78	82.09	82.37
	Set 2	80.60	81.18	82.08	82.35	82.50	82.62	82.67	82.78	82.81	82.83
Random set		19.05	30.88	52.46	79.98	90.74	92.68	93.29	93.48	93.57	93.65

Set 1 and Set 2 use the entire data set with the first and second sets of weight factors, respectively. Random set uses the 100 data sets randomly generated from the entire data set. Min, minumum; Max, maximum.

of each spectrum while deemphasizing the common characteristics of the spectra in the reference library. In other words, the optimal weight factors should make the pairwise mass spectral similarity scores between different compounds smaller, while producing the larger similarity scores for the same compounds. Therefore, since the spectra recorded in the reference library are generated from different compounds, we can expect that the distribution of the similarity scores should be right-skewed. In this regard, we develop the ratio of skewness to kurtosis for finding the optimal weight factors. As described before, the skewness is a measure of symmetry, while the kurtosis is a measure of whether data are peaked relative to a normal distribution. Moreover, the more right-skewed the distribution, the larger the skewness, and the more peaked the distribution, the bigger the kurtosis. The skewness will help find the weight factors to make the distribution right-skewed, while the kurtosis will prevent for the proposed algorithm to be ended up with the extreme case that the distribution is almost a point mass right-skewed.

Suppose that the reference library has N spectra $\{\alpha_w^i\}_{i=1}^N$ and there are $M(=N(N-1)/2)$ pairwise similarity scores $\{C_j(x,y)\}_{j=1}^M$ among the N spectra with respect to a set of weight factors $w=(x,y)$. Then, similar to (6) and (7), the skewness $S(x,y)$ and kurtosis $K(x,y)$ of M pairwise similarity scores are defined, given $w=(x,y)$, by

$$S(x,y)=\frac{\sum_{j=1}^{M}[C_j(x,y)-\overline{C}(x,y)]^3}{M\cdot\sigma(x,y)^3} \tag{8}$$

$$K(x,y)=\frac{\sum_{j=1}^{M}[C_j(x,y)-\overline{C}(x,y)]^4}{M\cdot\sigma(x,y)^4} \tag{9}$$

where $\overline{C}(x,y)=\frac{1}{M}\sum_{j=1}^{M}C_j(x,y)$ and $\sigma(x,y)^2=\frac{1}{M-1}\sum_{j=1}^{M}[C_j(x,y)-\overline{C}(x,y)]^2$. Using (8) and (9), the proposed algorithm finds the optimal weight factors based on the ratio of skewness to kurtosis,

$$R(x,y)=\frac{S(x,y)}{K(x,y)}. \tag{10}$$

For further analysis, we denote the conditional expectation of the ratio $R(x,y)$ given y as $E_X(R|y)$ and the conditional expectation of the ratio $R(x,y)$ given x as $E_Y(R|x)$. Then the optimal weight factors $\hat{x}$ and $\hat{y}$ for intensity and *m/z* value are estimated by maximizing the conditional expectation of $R(x,y)$ given x and y, respectively, such that

$$\hat{x}=\text{argmax}_x E_Y(R|x) \text{ and } \hat{y}=\text{argmax}_y E_X(R|y) \tag{11}$$

where x and y range from $-\infty$ to ∞.

Since the true distribution of the ratio $R(x,y)$ is unknown, the conditional means are replaced with the conditional sample means when there are T_x and T_y weight factors for intensity and *m/z* value, respectively, as follows:

$$\overline{R}(Y|x)=\frac{1}{T_y}\sum_{i=1}^{T_y}R(x,y_i) \text{ and } \overline{R}(X|y)=\frac{1}{T_x}\sum_{i=1}^{T_x}R(x_i,y). \tag{12}$$

By doing so, the optimal weight factors can be obtained through

$$\hat{x}=\text{argmax}_x\overline{R}(Y|x) \text{ and } \hat{y}=\text{argmax}_y\overline{R}(X|y). \tag{13}$$

Likewise, the conditional sample means for accuracy are

$$\overline{A}(Y|x)=\frac{1}{T_y}\sum_{i=1}^{T_y}A(x,y_i) \text{ and } \overline{A}(X|y)=\frac{1}{T_x}\sum_{i=1}^{T_x}A(x_i,y) \tag{14}$$

with respect to weight factors of intensity (x) and *m/z* value (y), respectively.

The reference library is the main EI MS library with a number of 212 860 spectra, as described in the previous section. Consequently, the total number of spectra pairs is $\binom{212\,860}{2}$ ($\approx 2.27 \times 10^{10}$). Such a large number of spectra pairs require enormous amount of computer memory to calculate all pairwise similarity scores. We, therefore, created 100 sub-main libraries randomly selected from the main EI MS library. Each sub-main library contains 1% of the total spectra recorded in the reference library, resulting in 21 286 spectra. Then the skewness, kurtosis and ratio of skewness to kurtosis were computed for each set of weight factors $w = (x, y)$, where $0.01 \leq x \leq 10$ and $0 \leq y \leq 10$, according to (5), (8)–(10). Their averages and standard errors were obtained and applied to the proposed approach to finding the optimal weight factors. For the further calculation of accuracy in compound identification, we also constructed 100 sub-replicate libraries randomly chosen from the replicate spectral library corresponding to the 100 sub-main libraries, resulting in 100 pairs of the reference library and the query with smaller size. Since the size of the main EI MS library is ~10 times bigger than the replicate spectral library and the size of each sub-main library is 21 286, we randomly selected 2816 spectra from the replicate spectral library, ensuring that the chosen 2816 spectra were present in its corresponding sub-main library.

2.5 Estimation of distribution of mass spectrum similarity scores

Due to a large number of spectra pairs, it is almost impossible to plot the distribution of the similarity scores among the spectra recorded in the reference library. Therefore, the distribution of mass spectrum similarity scores was estimated using a β distribution. Since the similarity score has the domain between 0 and 1 and a β distribution is versatile in terms of the shape, we employed a β distribution for this analysis. Furthermore, a β distribution has closed-form solutions for skewness and excess kurtosis

$$\text{Skewness} = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}} \tag{15}$$

$$\text{Kurtosis} = \frac{6[(a-b)^2(a+b+1) - ab(a+b+2)]}{ab(a+b+2)(a+b+3)} \tag{16}$$

where a and b are the parameters of a β distribution $\beta(a, b)$. Therefore, the parameters of a β distribution can be estimated using (15) and (16), if skewness and kurtosis are known.

2.6 Software

All statistical analyses were performed using the R statistical software version 2.13.1 (R Development Core Team) with the NIST mass spectrum library. The developed R package *iopt* and its examples are available at https://stage.louisville.edu/faculty/x0zhan17/software/software-development. The R package and examples are also available as Supplementary Material I.

3 RESULTS AND DISCUSSION

The NIST main EI MS library was used as the reference library to discover the optimal weight factors using the proposed method. The replicate spectra library was later employed as query data to evaluate the performance of the optimal weight factors in terms of the accuracy of compound identification. Due to the very large size of the reference library, the calculation of skewness and kurtosis using the entire spectra of the reference library is not tractable. Therefore, the skewness, kurtosis and ratio of skewness to kurtosis were calculated using an alternative approach based on 100 data sets randomly selected from the reference library as described in the Methods section.

The optimal weight factors were explored using Equations (12) and (13). The relationships among the skewness, kurtosis and ratio of skewness to kurtosis are displayed in Supplementary Figure S1 of the Supplementary Material II. The skewness of similarity scores ranges from −0.2691 to 14.27, and the kurtosis has the values between 2.529 and 221.3. In general, skewness is linearly and positively correlated to kurtosis, while the skewness and kurtosis have a non-linear relationship with the ratio of skewness to kurtosis. Using the proposed method, the weight factors $w = (x, y)$ for fragment ion intensity and m/z value are optimized at 0.53 and 1.3, respectively, as shown in Figure 2a. Interestingly, the optimal weight factors discovered in this work are different from the literature reported weight factors (Horai *et al.*, 2010); (Sokolow *et al.*, 1978); (Stein and Scott, 1994). In particular, the newly discovered optimal weight factors are similar to these of Sokolow *et al.* ($w = (0.5, 1)$) rather than these of Stein and Scott ($w = (0.6, 3)$) although the current work also used the NIST mass spectral library.

To demonstrate the effectiveness of the new weight factors for high accuracy compound identification, two data sets were employed as the reference library: the entire NIST main EI MS library and

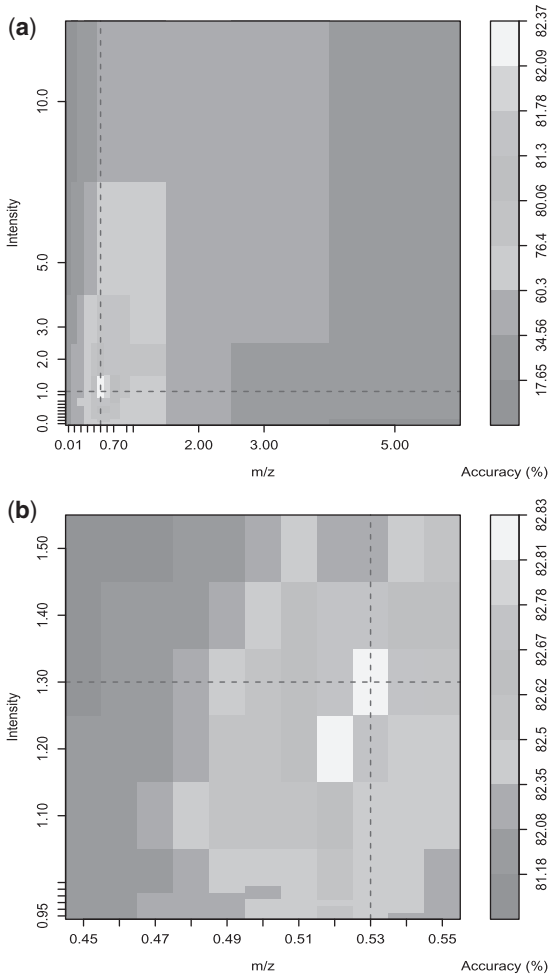

Fig. 3. The heat map visualization of accuracy of compound identification. The heat map is visualized using the seven percentile-levels that are 2.5th, 25th, 50th, 75th, 90th, 97.5th and 99th of the accuracy. (**a**) The first set of weight factors (W_1) is used. The dotted blue lines are of $w = (x, y) = (0.5, 1)$ at the maximum accuracy. (**b**) The second set of weight factors (W_2) is used. The dotted blue lines are of $w = (x, y) = (0.53, 1.3)$ at the maximum accuracy.

Table 2. The weight factors in the top 10% of accuracy using the first set of weight factors W_1

Rank	Weight factor		Accuracy	Rank	Weight factor		Accuracy
	x^a	y^b			x	y	
1	0.5	1	0.8237	11	0.7	1	0.8128
2	0.5	0.9	0.8227	12	0.7	0.9	0.8112
3	0.6	1	0.8205	13	0.5	0.5	0.8105
4	0.6	0.9	0.8202	14	0.5	2	0.8090
5	0.5	0.7	0.8178	15	0.6	0.5	0.8085
6	0.6	2	0.8178	16	0.7	0.7	0.8042
7	0.6	0.7	0.8153	17	0.5	0.4	0.8033
8	0.6	0.6	0.8135	18	0.6	0.4	0.8028
9	0.5	0.6	0.8133	19	0.4	1	0.8007
10	0.7	2	0.8130				

[a]The intensity weight factor.
[b]The m/z weight factor.

the 100 sub-main libraries randomly selected from the NIST main library. The replicate spectra library were used as query data in both cases. We first investigated the accuracy of compound identification using the 100 sub-main libraries. Figure 2b is the scatter plot between the conditional sample means of the ratio ($\overline{R}(Y|x)$) and the conditional sample means of accuracy ($\overline{A}(Y|x)$) with respect to the intensity weight factor, x. The two averages are highly correlated to each other with a correlation coefficient of 0.8882 (P-value$\approx$0), and the point of $\overline{R}(Y|x)$ and $\overline{A}(Y|x)$ corresponding to $\hat{x}=0.53$ is located in the upper-right-most corner. Similarly, the scatter plot between $\overline{R}(X|y)$ and $\overline{A}(X|y)$ with respect to the m/z weight factor, y, is depicted in Figure 2c. As expected, their correlation is significant with a correlation coefficient of 0.6052 (P-value$=$0.0101). The maximum of $\overline{R}(X|y)$ is corresponding to the maximum of $\overline{A}(X|y)$ and occurred at $\hat{y}=1.3$. Overall, Figures 2b and c demonstrate that the larger the ratio of skewness to kurtosis, the higher the accuracy. All conditional sample means can be found in Supplementary Table S1 of the Supplementary Material II.

Note that the boxplots between the weight factors and the accuracy in compound identification are depicted in Supplementary Figures S2b (bottom row) and S3 (upper row) of the Supplementary Material II. Interestingly, the variation of accuracy of the m/z weight factor is much larger than that of the intensity weight factor. That is, the distance between the lower and upper quartiles is much larger for the m/z weight factor than for the intensity weight factor, meaning that the amount of improvement of accuracy is larger when the intensity weight factor is varied than when the m/z weight factor is varied. This implies that the intensity weight factor may play a dominant role in compound identification since the variation of accuracy is smaller given the intensity weight factor than given the m/z weight factor.

We further evaluated the newly discovered optimal factors using the entire reference spectral library with respect to the performance of compound identification. Due to the extremely large size of the reference library, we designed a two-step evaluation approach, where two sets of weight factors were sequentially employed in each evaluation step. In the first evaluation step, a set of weight factors W_1 were used to study the general performance of compound identification, where x and y ranged from 0.01 to 5 and from 0 to 10, respectively. Most of high accuracies are obtained near to $x=0.5$ regardless of y (except when $y=5$) in Supplementary Figure S4a of the Supplementary Material II. In Supplementary Figure S4b of the Supplementary Material II, the majority of peaks of each curve are located near to $y=1$. These show that the weight factor can cause large variations in compound identification as shown in Table 1 and Supplementary Figure S4c of the Supplementary Material II. Heat map visualization was also examined using percentile-levels in Figure 3a, to investigate where accuracy is likely to be higher. The weight factors providing the top 10% accuracy range from 0.4 to 0.7 for x and from 0.4 to 2 for y, respectively, as depicted in Tables 1 and 2. Indeed, the 90th percentile of its accuracies is 80.06% and 82.37% is the maximum accuracy for the first set in Table 1. In Table 2, the weight factors of m/z value of the top 10% accuracy are $\leq$2, while the accuracies with the m/z weight factors >2 including $w=(0.5,3)$ and $(0.6,3)$ fall outside the top 10%.

In the second evaluation step, the identification accuracies were further examined in the range near to the maximum accuracy resulted from W_1, using a second set of weight factors W_2, where the weight factors x and y of W_2 range from 0.45 to 0.55 and from 0.95 to 1.5, respectively. It was found that the accuracy was maximized at $\hat{w}=(0.53,1.3)$, which is the same as the optimal weight factors found by the proposed method, as depicted in Figure 3b and Supplementary Figure S5 of the Supplementary Material II. The accuracy of compound identification at $\hat{w}=(0.53,1.3)$ is 82.83% in Table 3, while accuracies at $w=(0.5,1)$, $(0.5,2)$ and $(0.6,3)$ are 82.37%, 80.90% and 78.90%, respectively. These results demonstrate that the newly discovered weight factors outperform the literature reported weight factors for high accuracy compound identification. Several factors may contribute to these differences in the accuracy of compound identification, including the accuracy of each method for the discovery of the optimal weight factors and the reference mass spectral database used. Figure 4 displays a case that only the discovered weight factors $w=(0.53,1.3)$ can find the correct compound. Although the non-weighted intensities of the query and its correct compound are very similar to each other in Figure 4a, all the matched compounds are different from the correct compound except for the newly discovered weight factors as shown in Figure 4b. This is because $w=(0.5,1)$ and others weighted on the m/z value either less than or more than $w=(0.53,1.3)$ did, respectively, as depicted in Figure 4c. More detailed results can be found in the Supplementary Material III.

Since the reference library includes only heterogeneous spectra, it is expected that similarity scores are near to zero, and their distribution should be apart from that of similarity scores among the homogeneous spectra for high accuracy compound identification. Namely, the distribution should be right-skewed without a thick right tail and not flattened. In fact, the corresponding skewness and kurtosis to $\hat{x}=0.53$ are 2.30 and 15.77, respectively, while 3.92 and 35.27 are the corresponding skewness and kurtosis to $\hat{y}=1.3$, meaning that both distributions are located near to zero with right-skewed, peaked shapes. This can be confirmed using a β distribution as described in the Methods section. Indeed, for both cases, the distributions are located near to zero with right-skewed, peaked shapes, as shown in Supplementary Figure S6 of the Supplementary Material II. Therefore, the skewness and kurtosis of the maximum ratio, which are found in this study, are reasonable, but further research is needed to better understand and fully answer how much right-skewed and peaked the distribution is to be optimal.

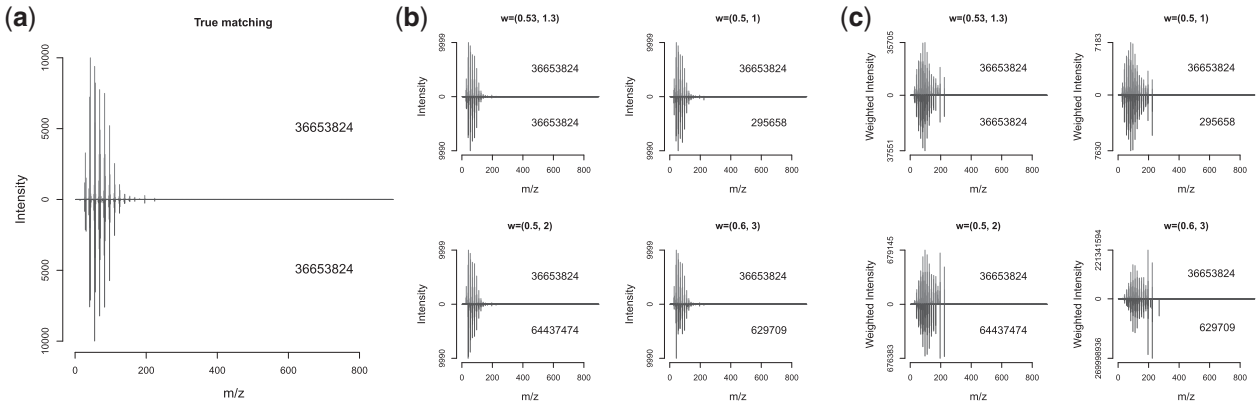

Fig. 4. A case that only $w=(0.53, 1.3)$ can find the correct compound. The red spectrum in the upper column is for a compound in the replicate spectral library and a compound from the main EI MS library is the blue spectrum in the bottom column. The number in each plot is the CAS registry number. The true matching is in (**a**). The compound pairs found by each weight factor are represented with non-weighted intensity (**b**) and with weighted intensity (**c**).

Table 3. The weight factors in the top 10% of accuracy using the second set of weight factors W_2

Rank	Weight factor		Accuracy	Rank	Weight factor		Accuracy
	x[a]	y[b]			x	y	
1	0.53	1.3	0.8283	8	0.52	1.1	0.8265
2	0.52	1.2	0.8282	9	0.51	1.2	0.8265
3	0.53	1.2	0.8278	10	0.51	1.3	0.8263
4	0.52	1.3	0.8278	11	0.51	1.4	0.8263
5	0.54	1.3	0.8278	12	0.55	1.4	0.8263
6	0.53	1.4	0.8273	13	0.5	1.1	0.8262
7	0.54	1.4	0.8267	14	0.5	1.2	0.8262

[a]TheThe intensity weight factor.
[b]The m/z weight factor.

4 CONCLUSION

The weight factors for intensities and m/z values are optimized at 0.53 and 1.3, respectively. The optimal weight factor discovered in this work is different from the literature reported weight factors, such as $w=(0.5, 1)$ (Sokolow *et al.*, 1978), $w=(0.6, 3)$ (Stein and Scott, 1994), and $w=(0.5, 2)$ (Horai *et al.*, 2010), although these discovered weight factors also were optimized for a mass spectral library in terms of compound identification accuracy. It suggests that the optimal weight factors highly depend on a mass spectral library.

All literature-reported optimal weight factors were found based on supervised learning, in the sense that a training data set (query library) is required during the discovery phase. The proposed approach can, however, find optimal weight factors only using a reference library without any query library, meaning that the proposed algorithm can be considered as unsupervised learning. The accuracy of compound identification using the optimal weight factors discovered in this work reaches 82.83%, demonstrating that the newly discovered weight factors outperform the literature reported weight factors for high accuracy compound identification.

ACKNOWLEDGEMENTS

The anonymous reviewers are thanked for their constructive comments.

Funding: This work was supported by grant 1RO1GM087735 through the National Institute of General Medical Sciences (NIGMS) within the National Institute of Health (NIH), DE-EM0000197 through the Department of Energy (DOE), and an Intramural Research Incentive Grant from the Office of the Executive Vice President for Research.

Conflict of Interest: none declared.

REFERENCES

Atwater,B.L. *et al.* (1985) Reliability ranking and scaling improvements to the probability based matching system for unknown mass-spectra. *Anal. Chem.*, **57**, 899–903.

Beer,I. *et al.* (2004) Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4**, 950–960.

Craig,R. *et al.* (2006) Using annotated peptide mass spectrum libraries for protein identification. *Proteome Res.*, **5**, 1843–1849.

Frewen,B. *et al.* (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.*, **78**, 5678–5684.

Hertz,H. *et al.* (1971) Identification of mass spectra by computer-searching a file of known spectra. *Anal. Chem.*, **43**, 681.

Horai,H. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.

Julian,R.K. *et al.* (1998) A method for quantitatively differentiating crude natural extracts using high-performance liquid chromatography electrospray mass spectrometry, *Analy. Chem.*, **70**, 3249–3254.

Koo,I. *et al.* (2011) Wavelet- and Fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Anal. Chem.*, **83**, 5631–5638.

Rasmussen,G.T. and Isenhour,T.L. (1979) Mass-spectral library searches using ion series data compression. *J. Chem. Inf. Comp. Sci.*, **19**, 98–104.

Reimann,C. *et al.* (2008) *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. John Wiley & Sons, Hoboken, NJ.

Sokolow,S. *et al.* (1978) The Finnigan library search program. *Finnigan Application Report No. 2.*, Finnigan Corp., San Jose, CA.

Stein,S.E. and Scott,D.R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass. Spectrom.*, **5**, 859–866.

Tabb,D. *et al.* (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.*, **75**, 2470–2477.