Published in final edited form as:

Anal Chem. 2012 August 7; 84(15): 6477–6487. doi:10.1021/ac301350n.

Compound Identification Using Partial and Semi-partial Correlations for Gas Chromatography Mass Spectrometry Data

Seongho Kim^{1,*}, Imhoi Koo^{1,2}, Jaesik Jeong³, Shiwen Wu², Xue Shi², and Xiang Zhang^{2,*}

¹Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, 40202

²Department of Chemistry, University of Louisville, Louisville, KY, 40292

³Department of Biostatistics, School of Medicine, Indiana University, Indianapolis, IN, 46202

Abstract

Compound identification is a key component of data analysis in the applications of gas chromatography–mass spectrometry (GC-MS). Currently, the most widely used compound identification is mass spectrum matching, in which dot product and its composite version are employed as spectral similarity measures. Several forms of transformations for fragment ion intensities have also been proposed to increase the accuracy of compound identification. In this study, we introduced partial and semi-partial correlations as mass spectral similarity measures and applied them to identify compounds along with different transformations of peak intensity. The mixture versions of the proposed method were also developed to further improve the accuracy of compound identification. To demonstrate the performance of the proposed spectral similarity measures, the National Institute of Standard Technology (NIST) mass spectral library and replicate spectral library were used as the reference library and the query spectra, respectively. Identification results showed that the mixture partial and semi-partial correlations always outperform both the dot-product and its composite measure. The mixture similarity with semi-partial correlation has the highest accuracy of 84.6% in compound identification with a transformation of (0.53, 1.3) for fragment ion intensity and m/z value, respectively.

Keywords

Compound identification; Partial correlation; Semi-partial correlation; dot product; NIST mass spectral library; GC-MS

1. Introduction

Gas chromatography–mass spectrometry (GC-MS) is the most commonly used method for analysis of volatile and semi-volatile organic compounds. One of the critical analyses on GC-MS data is compound identification, which is often achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity. To increase the accuracy of compound identification, various methods for the calculation of mass spectral similarity scores have been developed including dot product^{1–4}, composite similarity⁵, probability-based matching system⁶, Hertz similarity index⁷, normalized Euclidean distance (L_2 -norm)^{5, 8, 9}, and absolute value distance (L_1 -norm)^{5, 9}. Most recently, Koo et al.¹⁰ introduced wavelet and Fourier transforms-based composite measures and showed that their similarity scores perform better than the dot product and its composite version.

*Corresponding authors: s0kim023@louisville.edu and xiang.zhang@louisville.edu.

Since some compounds have similar mass spectral information to the others, an experimental query spectrum of these compounds is often matched to multiple mass spectra in the reference library with high similarity scores, resulting in impeding the high confidence compound identification. In other words, the mass spectral similarity score of a true positive pair does not always have the top-ranked score; and it is instead ranked as the second or even the third highest similarity score with an ignorable difference from the top-ranked score. Both Stein and Scott⁵ and Koo et al.¹⁰ have shown that the accuracy of compound identification can be improved by more than 12% when the similarity matching is expanded to the top three hits.

One way to deal with this difficulty in the mass spectral similarity score is to amplify the unique features of each mass spectrum, while weakening or removing the common features present in all mass spectra. In this regard, some attention has been focused on the transformation of peak intensities with respect to their mass-to-charge ratios (m/z). It has been noticed that the peak intensities of fragment ions with large m/z values in a GC-MS mass spectrum tend to be smaller⁵. However, the fragment ions with large m/z values are the most informative ions for compound identification. The performance of compound identification, therefore, can be improved by increasing the relative significance of the large fragment ions via transformation, i.e., weighing more on the peak intensities of fragment ions with large m/z values. Several studies have been performed to discover the optimal transformation of weight factors for fragment ion peak intensity as well as m/z value. Sokolow et al.¹¹ suggested the squared root of an intensity times its m/z value as an optimal scaling of the intensities, while Stein and Scott⁵ recommended an intensity to the power of 0.6 times its m/z value cubed in case of the dot product. Horai et al.¹² reported that the optimal transformation is the squared root of intensity and the square of its m/z value. Kim et al.¹³ lately discovered that the optimal transformation weight factors are database dependent. The optimal transformation for NIST11 spectral library is the intensity to the power of 0.53 times its m/z value to the power of 1.3.

In this study, we hypothesized that removing the common features shared among the mass spectra can improve the accuracy of compound identification in GC-MS. To test our hypothesis, we introduced the partial and semi-partial (also known as part) correlations as mass spectral similarity measures. The partial and semi-partial correlations calculate the unique relationship between the two mass spectra of interest, after removing the common features of each of the two mass spectra shared with other mass spectra in the reference library. We compared our proposed approaches with the widely used mass spectral similarity scoring methods, the dot product and its composite similarity along with the wavelet and Fourier transforms-based composite measures, in terms of accuracy of compound identification using the NIST mass spectral library.

2. Theory

Dot product and intensity transformation

The dot product⁵, which is also known as cosine correlation, was used to obtain the cosine of the angle between two sequences of intensities, $\mathbf{x} = (x_i)_{i=1,\dots,n}$ and $\mathbf{y} = (y_i)_{i=1,\dots,n}$. It is defined as

$$S = S(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}, \quad (1)$$

where $\mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i$ and $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$.

Although the fragment ion peaks with large m/z values in an EI mass spectrum usually have small peak intensities, these fragment ions carry the most important characteristics for compound identification. An optimal peak intensity transformation, i.e., weighing peak intensity of a fragment ion based on its m/z value, can increase the contribution of small peaks with large m/z values to the spectral similarity score and therefore, increase the accuracy of compound identification. Transformed peak intensity after the transformation $c = (a, b)$ is represented as

$$\{\text{peak intensity}\}^a \cdot \{\text{mass}(m/z)\}^b, \quad (2)$$

where a and b represent the contribution of peak intensity and m/z value, respectively. It is worth noting that Sokolow et al.¹¹ reported the optimal transformation $c = (0.5, 1)$. Stein and Scott⁵ later introduced the optimal transformation $c = (0.6, 3)$, while Horai et al.¹² recommended $c = (0.5, 2)$. Recently, Kim et al.¹³ showed that weight factors are database dependent and the optimal transformation of NIST11 database is $c = (0.53, 1.3)$.

The dot product with transformed intensity is defined by

$$S_c = S_c(x, y) = S(x_c, y_c) = \frac{x_c^T y_c}{\|x_c\| \cdot \|y_c\|}, \quad (3)$$

where $c = (a, b)$ is a vector of transformation factors of intensity and m/z value, respectively. In Equation (3), $x_c = (x_i^c)_{i=1}^n$ and $y_c = (y_i^c)_{i=1}^n$ are component-wise transformed intensities based on Equation (2) and

$$x_i^c = (x_i)^a \cdot (z_i)^b \text{ and } y_i^c = (y_i)^a \cdot (z_i)^b \quad (4)$$

where z_i is the m/z value of i th intensity, $i = 1, 2, \dots, n$, and a and b are transformation (weight) factors. In fact, when $c = (1, 0)$, S_c in Equation (3) is identical to S in Equation (1). That is, $c = (1, 0)$ means no transformation of peak intensity.

Stein and Scott's composite similarity

Stein and Scott⁵ proposed a composite similarity measure combining the dot product with transformed intensities, S_c , in Equation (3) and similarity of peak ratios, S_R . The ratio similarity of peak pairs S_R is defined as

$$S_R(x, y) = \frac{1}{N_{x \wedge y}} \sum_i^{N_{x \wedge y}} \left(\frac{y_i}{y_{i-1}} \cdot \frac{x_{i-1}}{x_i} \right)^n, \quad (5)$$

where $n = 1$ or -1 if the term in parentheses is less than or greater than unity, respectively, and $N_{x \wedge y}$ is the number of peaks with non-zero peak intensity in both the reference library and the query spectrum. Combining S_c and S_R in Equations (3) and (5), their composite similarity was then defined as

$$S_{cR} = S_{cR}(x, y) = \frac{N_x S_c(x, y) + N_{x \wedge y} S_R(x, y)}{N_x + N_{x \wedge y}}, \quad (6)$$

where N_x is the number of non-zero peak intensities existing in the query spectrum.

Wavelet and Fourier transform-based composite measure

The composite similarity measures based on the wavelet and Fourier transforms are calculated by replacing the similarity of peak ratios S_R in Equation (6) with wavelet and Fourier coefficients, respectively¹⁰. Wavelet and Fourier transforms are operators that map a signal function into a periodic or a temporal domain.

The wavelet transform of a signal x is calculated by passing it through a low-pass filter g and a high-pass filter h , resulting in approximations and details coefficients $x^u=(x_i^u)_{i=1}^n$ and $x^d=(x_i^d)_{i=1}^n$, respectively. The coefficients of approximations and details are calculated by the following expressions:

$$x_k^u = \sum_{j=1}^n x_j \cdot g[2k-j-1], k=1, \dots, n; \quad (7)$$

$$x_k^d = \sum_{j=1}^n x_j \cdot h[2k-j-1], k=1, \dots, n, \quad (8)$$

where Daubechies' high-pass h and low-pass g filters are used in this study¹⁵. Fourier-transformed signal $x^f=(x_i^f)_{i=1}^n$ of a time-domain signal $x=(x_i)_{i=1}^n$ is written by

$$x_k^f = x_k^{f,r} + i \cdot x_k^{f,i} = \sum_{j=1}^n x_j \cdot \exp\left(-\frac{2\pi i}{N}(k-1)j\right), k=1, \dots, n, \quad (9)$$

where i is the imaginary unit; $x_k^{f,r}$ and $x_k^{f,i}$ are the real and the imaginary parts of x_k^f .

Koo et al.¹⁰ demonstrated that the real part of Fourier transformed signals and the details of wavelet transformed signals outperform others in terms of the accuracy of compound identification. Therefore, we selected them for comparison purpose. Their composite similarity measures are defined, respectively, by:

$$S_{cF} = S_{cF}(x, y) = \frac{N_x S_c(x, y) + N_{x \wedge y} S_F(x, y)}{N_x + N_{x \wedge y}}, \quad (10)$$

$$S_{cD} = S_{cD}(x, y) = \frac{N_x S_c(x, y) + N_{x \wedge y} S_D(x, y)}{N_x + N_{x \wedge y}}, \quad (11)$$

where $S_F(x, y) = S(x^{f,r}, y^{f,r})$ and $S_D(x, y) = S(x^d, y^d)$.

Partial and semi-partial correlation

The partial correlation can be interpreted as the association between two random variables after eliminating the effect of other random variables, while the semi-partial correlation eliminates the effect of a fraction of other random variables, for instance, just removing the effect of one random variable from a total of two random variables.^{16–18} They have been applied to biological network studies in order to find direct relationship/association among genes, proteins, and metabolites.^{19–21} Considering a partitioned random vector $\{X, Y, Z\}$ where X and Y are one-dimensional random variables and Z is n -dimensional random

vector, the partial correlation $\rho_{XY|Z}$ between X and Y given $Z = \{Z_1, \dots, Z_n\}$ is the correlation between the residuals $R_{X|Z}$ and $R_{Y|Z}$ and is represented by

$$\begin{aligned}\rho_{XY|Z} &= \frac{\text{Cor}(R_{X|Z}, R_{Y|Z})}{\sqrt{\text{Var}(R_{X|Z}) \cdot \text{Var}(R_{Y|Z})}} \\ &= \frac{\text{Cov}(X, Y) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}{\sqrt{\text{Var}(X) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, X)} \cdot \sqrt{\text{Var}(Y) - \text{Cov}(Y, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}}\end{aligned}\quad (12)$$

where $R_{X|Z}$ and $R_{Y|Z}$ are the results from the linear regression of X and Y on Z , respectively, $R_{X|Z} = X - \hat{X}(Z)$, $R_{Y|Z} = Y - \hat{Y}(Z)$, $\hat{X}(Z) = E(X) + \text{Cov}(Z, X) \cdot \text{Var}(X)^{-1} \cdot (X - E(X))$ and $\hat{Y}(Z) = E(Y) + \text{Cov}(Z, Y) \cdot \text{Var}(Y)^{-1} \cdot (Y - E(Y))$. Note that $E(X)$, $\text{Var}(X)$, and $\text{Cov}(Z, X)$ denote the expectation of X , the variance of X and the covariance between Z and X , respectively.

The semi-partial correlation $\rho_{X(Y|Z)}$ between X and Y with $Z = \{Z_1, \dots, Z_n\}$ is the correlation between the random variable X and $R_{Y|Z}$ resulting from the linear regression of Y on Z , respectively, and is represented by

$$\begin{aligned}\rho_{X(Y|Z)} &= \frac{\text{Cor}(X, R_{Y|Z})}{\sqrt{\text{Var}(X) \cdot \text{Var}(R_{Y|Z})}} \\ &= \frac{\text{Cov}(X, Y) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y) - \text{Cov}(Y, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}}\end{aligned}\quad (13)$$

For illustration purposes, we consider three random variables, X , Y , and Z , and suppose that the relationship/association between X and Y is of interest. To describe the difference between the correlation, the partial correlation, and the semi-partial correlation, three situations are taken into consideration as depicted in Figure 1. Figure 1(a) depicts a case that none of X and Y is correlated with Z while Figure 1(b) depicts that both X and Y are correlated with Z . In the case of Figure 1(c), only the random variable Y is correlated with Z . Theoretically, all the three correlations have the identical value in the situation of Figure 1(a) since Z has nothing to do with X and Y . In case of Figure 1(c), the partial correlation is exactly same as the semi-partial correlation, but is different from the correlation since Y is correlated with Z . All three correlations are different from each other for the situation of Figure 1(b). The rationale for the partial and the semi-partial correlations is to obtain direct or pure relationship between two random variables. For example, in Figure 1(b), although X and Y are uncorrelated to each other, the correlation can be a nonzero value due to Z . In this case, by removing the effect of Z , their direct or true relationship can be obtained. Likewise, the significant relationship between X and Y in Figure 1(c) can be due to the hidden relationship between X and Z . For these reasons, the partial or the semi-partial correlations must be used to obtain the true relationship between X and Y if the relationship between (X, Y) and Z is present not due to measurement errors. However, it is possible that the partial or the semi-partial correlations can lead to wrong correlation if the correlation between (X, Y) and Z is spurious caused by noise. For more details on the partial and the semi-partial correlations, refer to James (2002)¹⁶ and Whittaker (1990)¹⁸.

In the context of compound identification, these partial and semi-partial correlations can be employed to calculate the mass spectral similarity score. By removing the effect of other mass spectra over the two mass spectra of interest, we expect that the unique relationship between the mass spectra can be extracted. Namely, using these correlations will have the same effect as the intensity transformation does. Suppose that X is a query mass spectrum

and $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$ is a set of N mass spectra in the reference library. The partial and the semi-partial correlations between X and Y_i given $\mathbf{Y}^{(i)}$ can be calculated, respectively, by

$$\rho_{XY_i|\mathbf{Y}^{(i)}} = \text{Cor}(R_{X|\mathbf{Y}^{(i)}}, R_{Y_i|\mathbf{Y}^{(i)}}); \quad (14)$$

$$\rho_{X(Y_i|\mathbf{Y}^{(i)})} = \text{Cor}(X, R_{Y_i|\mathbf{Y}^{(i)}}), \quad (15)$$

where $\mathbf{Y}^{(i)} = \mathbf{Y} \setminus \{Y_i\} = \{Y_1, Y_2, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_N\}$. Note that X, Y_i and $\mathbf{Y}^{(i)}$ in the equations (14) and (15) have the identical roles as X, Y , and $\mathbf{Z}$ in the equations (12) and (13) do, respectively. However, some pairs of mass spectra in the NIST library have identical mass spectral similarity scores, causing the singularity of the inverse covariance matrices between X and $\mathbf{Y}$. This often occurs in mass spectrum pairs having small spectral similarity scores. To avoid the singularity problem, we first reduced the number of mass spectra used for the calculation of the partial and the semi-partial correlations by considering only the mass spectra that have the first k highest similarity scores obtained by the dot product. Then the partial and the semi-partial correlations were computed only for these k mass spectra. Given the rank k and transformation c , these partial and semi-partial correlations are represented, respectively, by

$$S_{p,i,k,c} = S_{p,k,c}(X, Y_i) = \rho_{XY_i|\mathbf{Y}^{(i,k)}} = \text{Cor}\left(R_{X|\mathbf{Y}^{(i,k)}}, R_{Y_i|\mathbf{Y}^{(i,k)}}\right); \quad (16)$$

$$S_{s,i,k,c} = S_{s,k,c}(X, Y_i) = \rho_{X(Y_i|\mathbf{Y}^{(i,k)})} = \text{Cor}\left(X, R_{Y_i|\mathbf{Y}^{(i,k)}}\right), \quad (17)$$

where $\mathbf{Y}^{(i,k)} = \{Y_j | \text{Rank}(S_c(X, Y_j)) \leq K, Y_j \in \mathbf{Y}^{(i)}\}$, $S_c(X, Y_j)$ is the dot product between two mass spectra X and Y_j after transformation c , and $\text{Rank}(S_c(X, Y_j))$ is the rank of the similarity score $S_c(X, Y_j)$ in descending order.

Mixture partial and semi-partial correlations with the dot product

Mixture similarities of the partial and the semi-partial correlations were further developed with the dot product. Given (k, c, w) , the mixture partial and semi-partial correlations of the two mass spectra X and Y_i are then represented, respectively, by

$$(1-w) \cdot S_c(X, Y_i) + w \cdot S_{p,k,c}(X, Y_i); \quad (18)$$

$$(1-w) \cdot S_c(X, Y_i) + w \cdot S_{s,k,c}(X, Y_i), \quad (19)$$

where $Y_i \in \mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$ and w is a mixture weight ranging from 0 to 1.

We used the accuracy of compound identification as the measure to evaluate the performance of different mass spectral similarity measures for compound identification. The accuracy is defined as the proportion of mass spectra identified correctly in the query data. In other words, if a query mass spectrum and the matched mass spectrum in the reference library have the same Chemical Abstracts Service (CAS) registry number, this mass 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

$$\text{Accuracy} = \frac{\text{Number of spectra matched correctly}}{\text{Number of spectra queried}}. \quad (20)$$

3. Experimental Section

The NIST mass spectral library and replicate spectral library

We considered the mass spectra extracted from the NIST Chemistry WebBook¹⁴ as a reference library and the repetitive library as query data. The NIST Chemistry WebBook service (<http://webbook.nist.gov/chemistry/>) provides users with chemical and physical information for compounds including mass spectra generated by electron ionization (EI) mass spectrometry. As of November 28, 2011, the EI mass spectra of 23721 compounds were extracted from the NIST Chemistry WebBook. In addition, the replicate spectral library was obtained from the NIST08 Mass Spectral Library (NIST08/2008), containing 28307 EI mass spectra generated by 18569 compounds. The same chemical compounds were identified and grouped by CAS registry number. Since we assumed that the reference library includes the mass spectra of all the query compounds, all compounds that were not present in the NIST library were removed from the repetitive library. After the removal, 12850 compounds with 21516 mass spectra were left in the repetitive library. The fragment ion m/z values in the two mass spectral libraries were ranged from 1 to 892 with a bin size of 1.

Experiment data

In order to investigate the effect of contaminant mass spectra on the compound identification, an experimental data set generated from a comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOF MS) was used. The sample analyzed on GC×GC-TOF MS is a mixture of 76 compound standards (8270 MegaMix, Restek Corp., Bellefonte, PA). The concentration of each compound in the mixture is 2.5 $\mu\text{g/mL}$. The mixture was analyzed on a LECO Pegasus 4D GC×GC-TOF MS instrument (LECO Corporation, St. Joseph, MI, USA) equipped with a cryogenic modulator. The LECO ChromaTOF software version 3.4 was used for instrument control, spectrum deconvolution, and peak detection.

Software

All the statistical analyses were performed using statistical software R 2.13.1 (R Development Core Team, 2011). The partial and the semi-partial correlations were calculated by the R package *ppcor*. The wavelet and Fourier- transforms similarities including Stein and Scott's composite measure were calculated in MATLAB v.7.11 (The Mathworks, Natick, MA).

4. Results

Considering the NIST Chemistry WebBook MS library and the replicate library as the reference library and the query, respectively, the performance of compound identification was compared among the dot product, the composite measure, the wavelet and Fourier-transforms composite measure, the (mixture) partial correlation and the (mixture) semi-partial correlation. More specifically, three comparisons were considered: (1) the partial/semi-partial correlations versus the dot product, (2) the mixture partial/semi-partial correlations versus the dot product, and (3) the mixture partial/semi-partial correlations versus the composite measures. For comparison, we further considered the five different transformations: no transformation, (0.5, 3), (0.6, 3), (0.5, 2), (0.5, 1) and (0.53, 1.3), where "no transformation" means $c = (1, 0)$.

Compound identification by partial and semi-partial correlations

The partial and semi-partial correlations were compared with the dot product according to the different transformations of peak intensity. When calculating the partial and semi-partial correlations, the inverse covariance matrix of the query and the reference library is required. Several pairs of mass spectra in the reference library, however, have identical mass spectral similarity scores so that the covariance matrices numerically become singular. This usually occurred for the pairs of mass spectra generated from different compounds, resulting in small spectral similarity scores. For this reason, we used the score of dot product to reduce the number of compounds for the covariance matrices before calculating the partial and semi-partial correlations. This was done by including only the mass spectra having the k highest similarity scores with a query spectrum into the covariance matrices.

Figure 2(a) displays the accuracy of compound identification for each transformation according to the different ranks, i.e., $k = 3, 5, 10, 20, 30, 50$ and 100 . The accuracy of the dot product is displayed in the figure at $k = 0$. Interestingly, the performances of both the partial and the semi-partial correlations are identical, implying that there might be no common feature between the query and the reference library. With no transformation, the partial and the semi-partial correlations improved the accuracy of compound identification by 1.15% as shown in Table 1. In contrast, the performances were decreased when these correlations were applied to compound identification with transformed intensities, as depicted in Figure 2(a). The maximum accuracy was achieved at $k = 20$ for the case of non-transformation, while at $k = 3$ for all cases of transformations. Figure 2(b) depicts the maximum accuracies of compound identification achieved by all of the partial correlation, the semi-partial correlation and the dot product as mass spectral similarity measures. Clearly, the partial and the semi-partial correlations outperform the dot product when the mass spectra were not transformed. However, their performance was not so good as the dot product when the mass spectra were transformed. The best performance of compound identification occurred at $c = (0.53, 1.3)$ with the dot product as the mass spectral similarity measure.

Compound identification by the mixture partial and semi-partial correlations

The mixture similarity measures between the dot product and the partial/semi-partial correlations were further evaluated. To do this, we considered 11 mixture weights such as $w = 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ and 1.0 for the mixture partial and semi-partial correlations.

The accuracy of compound identification for the mixture partial and semi-partial correlations is provided in Figure S-1. As a reference, the horizontal grey line was added to indicate the accuracy of compound identification using the dot product as mass spectral similarity measure. Across all the transformations, the mixture partial and semi-partial correlations improved the accuracy of compound identification so that their accuracies become greater than that of the dot product, i.e., their accuracies fall above the grey line. This is significantly different from the identification results of using the simple partial and semi-partial correlations as spectral similarity measures (Figure 2(b)). The accuracy of compound identification reached the maximum values at the mixture weight $w = 0.1$ for the mixture partial correlation and $w = 0.7$ for the mixture semi-partial correlation, as shown in Figure S-1(a) and Table 1. The rank k at the maximum accuracy is 50 and 100 for the mixture partial and the semi-partial correlations, respectively.

When the transformed spectra were used in the mixture partial correlation, the accuracy of compound identification increases with the decrease of mixture weight w as depicted in Figures S-1(b)–(f). As a result, the mixture weight w at the maximum accuracy of the partial

correlation is 0.01 across all the transformations and its rank k ranges from 20 to 50, as shown in Table 1. On the other hand, the mixture semi-partial correlation has the mixture weight w of 0.1 or 0.2 with the rank k ranging from 20 to 50. It is noteworthy that there is no result of $k = 100$ for the transformed spectra since the covariance matrix becomes singular when the rank k is equal to 100. Likewise, the transformed spectra with $c = (0.53, 1.3)$ have the best performance in terms of the accuracy of compound identification.

Comparison of the mixture correlations with the composite measures

We further considered three composite measures as described in the Theory section: the Stein and Scott's composite measure, the wavelet transform-based composite measure and the Fourier transform-based composite measure. For simplicity, we refer to them as the ratio composite, wavelet composite, and Fourier composite measures, respectively.

The three composite measures were compared with our mixture partial/semi-partial correlations. For each transformation, the maximum accuracies of each measure are provided in Figure 3. When the non-transformed spectra were used, the proposed mixture correlations outperformed by far the three composite measures as well as the dot product. On the other hand, if the mass spectra were transformed by $c = (0.5, 3)$, the wavelet and Fourier composite measures performed better than others in terms of the identification accuracy. However, their performances for other transforms such as $c = (0.5, 1)$ and $c = (0.53, 1.3)$ were worse. In case of $c = (0.6, 3)$ and $(0.5, 3)$, the wavelet composite measure performed the best and both the wavelet and Fourier composite measures performed better than the dot product. Surprisingly, the ratio composite measure always performed the worst, even worse than the dot product, regardless of the transformation of peak intensity. To give a better insight on the comparison, we provide the cases that only one similarity measure can match the query into the correct spectrum in the reference library with the plots of nonzero intensities in Figure S-4 of the Supporting Information. In this figure, all the intensities were transformed with $c = (0.53, 1.3)$.

The overall maximum accuracy occurred when the mixture semi-partial correlation was used with transformation weight factors $c = (0.53, 1.3)$. Its accuracy reached 84.59% as can be seen in Table 1. The corresponding optimal rank k and the mixture weight w were 50 and 0.1, respectively.

As for the methods in Figure 3, we further evaluated their performances of compound identification by calculating 95% confidence intervals based on a bootstrap resampling method. To do this, 1000 bootstrap query samples were created by sampling with replacement of the same number of original query (replicate spectral library). Then the mean and standard deviation of accuracies of the 1000 bootstrap query samples were calculated along with 95% bootstrap confidence intervals, as depicted in Table 2. We can see that the bootstrap means of accuracy are almost identical to the accuracy of the methods used in Figure 3 as shown in Tables 1 and 2. Although the improvement of the mixture partial and semi-partial correlations is small, Table 2 shows that their 95% confidence intervals are not overlapped with others, which indicates that their performances are significantly different from others.

By selecting the best performed rank k and mixture weight w for the mixture correlations at each transformation, we investigated their identification accuracies up to top three hits together with the dot product. These performances are given in Table 3. As noticed in the previous studies^{5,10}, we also observed the similar trend that the identification accuracies are improved by ~12% when the top three hits are considered. Interestingly, the mixture partial correlation (96.15%) performs slightly better than the mixture semi-partial correlation (96.11%).

Analysis of GC×GC-TOF MS data

The proposed methods were further tested using an experimental data set acquired on GC×GC-TOF MS for a mixture of compound standards, where only the transformation of $c = (0.53, 1.3)$ was considered along with the dot product, the composite measure, the wavelet and Fourier-transforms similarities, and the mixture partial and semi-partial correlations. Since the true name for each peak is unknown, we just counted the number of 76 compound standards identified by each method (Table 4). The numbers in diagonal of Table 4 represent the total number of 76 compound standards identified by the method in row (or column), and the numbers in off-diagonal are the number of compound standards commonly identified by the methods in the row and its corresponding column. Namely, the mixture semi-partial correlation identified 54 compound standards, which is the largest number, and the next best result was from the mixture partial correlation, by which 50 compound standards were identified. The Rank, wavelet, and Fourier composite measures identified 26, 25, and 25 compound standards, respectively, which is much smaller than the number of 76 compound standards identified by the dot product (50 compound standards). The more detailed comparison using Venn diagrams can be seen in Figure S-2. These indicate that the trend of identification accuracy in analyzing the experimental data is similar to the one using the replicate spectral library as query.

In order to see the effect of the contaminated query mass spectra on the identification, we did further simulation with five different signal-to-noise ratios (SNRs) such as 1, 10, 100, 1000, and 10000 (Supporting Information, Signal-to-Noise Ratio Simulation). Likewise, only the transformation of $c = (0.53, 1.3)$ was considered along with the dot product, the composite measure, the wavelet and Fourier-transforms similarities, and the mixture partial and semi-partial correlations. After obtaining the query spectra with Gaussian noise using the procedures described in the Supporting Information, the mean and standard deviation of accuracy of compound identification were calculated for each spectral similarity measure using the results of 1000 simulations. The simulation results are shown in Table S-2 of the Supporting Information. It can be seen that the identification accuracy increases with the increase of SNR level for all six spectral similarity measures. The Fourier-transform similarity performs the best when $\text{SNR} = 1$, while the mixture semi-partial correlation outperforms the other similarity measures when $\text{SNR} > 1$. It is interesting that the ratio of composite measure is very sensitive to the noise level with identification accuracy sharply increased from 0.4799 to 0.7431 when SNR increases from 1 to 10. Overall, these simulation results demonstrate that the mixture semi-partial correlation provides the highest identification accuracy unless the quality of mass spectra is very poor, e.g., $\text{SNR} = 1$.

5. Discussion and Conclusions

Compared with dot product, the simple partial and semi-partial correlations improved the accuracy of compound identification up to 1.15% when applied to the non-transformed intensities in Table 1. It is also noteworthy that the mixture partial and semi-partial correlations improved up to 6.20% without transformation of peak intensities. However, their performances became worse than the dot product when the transformed intensities were considered, as can be seen in Figure 2(b). This was caused due to the same roles of the simple partial/semi-partial correlations and the transformation of peak intensity. That is, if a query spectrum and a reference spectrum are directly correlated and are also independent of other spectra, the simple partial and semi-partial correlations are theoretically the same as Pearson's correlation. In this regard, when the peak intensities are transformed, the unique characteristics of mass spectra are emphasized while weakening the common features shared with others. This makes the mass spectra in the reference library independent of each other so that the contribution of the simple partial and semi-partial correlations disappeared. However, the use of the simple partial and semi-partial correlations can cause excessive

removal of the effects of other mass spectra so that identification accuracy decreases, when the amount of the common characteristics among mass spectra in the reference library is little. For the same reason, the maximum identification accuracies for the non-transformation occurred at $k = 20$, while the maximum accuracies for all the transformed spectra occurred at $k = 3$ in case of the simple partial and semi-partial correlations, as can be seen in Table 1. These results demonstrate the need of transformation of peak intensities on compound identification.

The maximum accuracy of compound identification occurred at the mixture weight 0.01 when the mixture partial correlation is used as the mass spectral measure with transformed intensities, while at the mixture weight 0.1 when the mixture semi-partial correlations is used as the mass spectral similarity measure. On the other hand, the maximum accuracies with non-transformed intensities occurred at 0.1 and 0.7 for the mixture partial and semi-partial correlations, respectively. Namely, the contributions of the partial and the semi-partial correlations were much larger for the non-transformed intensities than for the transformed intensities. This is because the simple partial and semi-partial correlations outperform the dot product in case of non-transformation, while the dot product performs better than the simple partial and semi-partial correlations when the transformed intensities were used. Compared with the mixture semi-partial correlation, the mixture weight of the mixture partial correlation is much smaller at the maximum identification accuracy. The reason is that the partial correlation is mathematically greater than or equal to the semi-partial correlation so that the distribution of the partial correlation is more left-skewed than that of the semi-partial correlation. Therefore, the semi-partial correlation has the larger mixture weight than the partial correlation.

The main difference between the proposed mixture similarity measures and the three literature- reported composite measures (the ratio/wavelet/Fourier composite measures) is the way to assign the mixture weight to the dot product. The mixture weight is assigned to each pair of mass spectra locally in the literature- reported mixture measures by giving mixture weight to each pair of mass spectra according to the ratio of the non-zero peak intensities, while the proposed mixture measures assign the same mixture weight to all pairs of mass spectra globally, i.e., global mixture weight. Therefore, the local mixture weight is fixed regardless of the intensity transformation, but the global mixture weight depends on the intensity transformation. For these reasons, the performance of our mixture measures was affected by the different transformation weight factors, and the mixture partial and semi-partial correlations outperformed the dot product for all transformations considered here, as shown in Figure 3.

While the partial correlation removes the effect of other mass spectra recorded in the reference library on both a query spectrum and a mass spectrum in the reference library, the semi-partial correlation eliminates the effect of other mass spectra only from a mass spectrum in the reference library. Moreover, the query and the reference libraries are usually constructed from different experimental conditions. Therefore, it is reasonable to remove the effect of the spectra recorded in the reference library only because query spectra are generated under the different conditions from the reference library spectra, which can explain that the mixture semi-partial correlation performs better than the mixture partial correlation in terms of accuracy of compound identification.

Combining the identification results from different spectral similarity measures can reduce the false discovery rate. For instance, the maximum accuracies of the dot product and the mixture semi-partial correlation were 84.15% and 84.59%, respectively, with transformation of $c = (0.53, 1.3)$. If we acknowledge all 21516 assignments of the query spectra as positive discoveries, the false discovery rates for the dot product and the mixture semi-partial

correlation are 15.85% ($= 100\% - 84.15\%$) and 15.41% ($= 100\% - 84.59\%$), respectively. As shown in Figure 4(a), when comparing the assigned CAS indices by the dot product with these by the mixture semi-partial correlation, 20767 ($= 2883 + 17884$) assignments out of 21516 queries had the identical CAS indices for both measures. Among these 20767 query spectra assigned to the same CAS indices by both measures, 17884 of them were correctly matched, resulting in identification accuracy of 86.12% ($= 17884/20767$), while the remaining 2883 assignments ($= 20767 - 17884$) were not correct even though each of these query spectrum was assigned to the same compound by both measures. If we acknowledge only these 20767 queries as positive discoveries, the false discovery rate can be reduced by ~2%, obtaining 13.88% ($= 100\% - 86.12\%$). This indicates that accepting the query spectra assigned to the same CAS indices by two spectral similarity measures as the positive discoveries can improve the false discovery rate.

Furthermore, we evaluated the FDR for all pairs among six similarity methods with transformation of $c = (0.53, 1.3)$. Interestingly, Table S-1 and Figure S-3 show that the pair of the ratio and wavelet composites (Ratio-Wavelet) and the pair of the ratio composite and the mixture semi-partial correlation (Ratio-SPC) have the lowest FDR of 8.10%. However, Ratio-SPC have larger true positives than Ratio-Wavelet (18087 versus 16302) as can be seen in Figure 4 (b) and (c).

The results in this work lead us to several conclusions. The dot product alone can produce high confidence identification only if the optimal transformation of intensity is used, and its accuracy of compound identification can always be improved by using the proposed mixture partial or semi-partial correlations. Moreover, the literature- reported ratio/wavelet/Fourier composite measures will not be optimal choices for compound identification since their performances are not consistent with the different transformations of intensity and, especially, the ratio composite measure is always worse than the dot product in terms of the identification accuracy. When the optimal transformation cannot be searchable or is not available, the proposed simple or mixture partial and semi-partial correlations will be an alternative similarity score for obtaining the high accuracy of compound identification since their role is almost identical to the nature of transformation of peak intensity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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.

References

1. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR. *Analytical Chemistry*. 2003; 75:2470–2477. [PubMed: 12918992]
2. Beer I, Barnea E, Ziv T, Admon A. *Proteomics*. 2004; 4:950–960. [PubMed: 15048977]
3. Craig R, Cortens JC, Fenyo D, Beavis RC. *Journal of Proteome Research*. 2006; 5:1843–1849. [PubMed: 16889405]
4. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. *Analytical Chemistry*. 2006; 78:5678–5684. [PubMed: 16906711]
5. Stein SE, Scott DR. *Journal of the American Society for Mass Spectrometry*. 1994; 5:859–866.
6. Atwater BL, Stauffer DB, McLafferty FW, Peterson DW. *Analytical Chemistry*. 1985; 57:899–903.
7. Hertz HS, Hites RA, Biemann K. *Analytical Chemistry*. 1971; 43:681.

8. Julian RK, Higgs RE, Gygi JD, Hilton MD. *Analytical Chemistry*. 1998; 70:3249–3254. [PubMed: 21644661]
9. Rasmussen GT, Isenhour TL. *Journal of Chemical Information and Computer Sciences*. 1979; 19:179–186.
10. Koo I, Zhang X, Kim S. *Analytical Chemistry*. 2011; 83:5631–5638. [PubMed: 21651237]
11. Sokolow, S.; Karnofsky, J.; Gustafson, P. The Finnigan library search program. Finnigan Application Report 2. Finnigan Corp; San Jose, CA: 1978.
12. Horai H, Arita M, Nishioka T. *May 27–30.2008* :853–857.
13. Kim S, Koo I, Wei X, Zhang X. *Bioinformatics*. 2012; 28:1158–1163. [PubMed: 22333245]
14. National Institute of Standards and Technology (U.S.). NIST standard reference database. Vol. 69. National Institute of Standards and Technology; Washington, D.C: Jun. 2005 release
15. Daubechies, I. Ten lectures on wavelets. Society for Industrial and Applied Mathematics; Philadelphia, PA: 1992.
16. James, S. *Applied Multivariate Statistics for the Social Sciences*. Lawrence Erlbaum Associates, Inc; Mahwah, NJ: 2002.
17. Johnson, RA.; Wichern, DW. *Applied Multivariate Statistical Analysis*. Prentice Hall; 2002.
18. Whittaker, J. *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons; New York, NY: 1990.
19. de la Fuente A, Bing N, Hoeschele I, Mendes P. *Bioinformatics*. 2004; 20:3565–74. [PubMed: 15284096]
20. Komurov K, White M. *Molecular Systems Biology*. 2007; 3:110. [PubMed: 17453049]
21. Jourdan F, Breitling R, Barrett MP, Gilbert D. *Bioinformatics*. 2008; 24:143–145. [PubMed: 18003642]

Figure 1. Graphical representation of the relationship among the random variables X , Y , and Z
The green edge represents the relationship of interest, and the black edge is used if two random variables are correlated.

(a)

(b)

Figure 2. Accuracy of compound identification using the dot product, partial and semi-partial correlations

In (a), the accuracies are depicted according to the different transformations and ranks. The maximum accuracies for each similarity measure are plotted by the different transformation of peak intensities in (b). In (b), “No-trans” stands for no transformation, and “Dot”, “PC”, and “SPC” represent the dot product, the partial correlation, and the semi-partial correlation, respectively.

Figure 3. Maximum accuracy of compound identification according to the different transformations of intensity

“No-trans” stands for no transformation, and “Dot”, “Ratio”, “Wavelet”, “Fourier”, “PC”, and “SPC” represent the dot product, the ratio composite, the wavelet composite, the Fourier composite, the mixture partial correlation, and the mixture semi-partial correlation, respectively.

(a)

(b)

(c)

Figure 4. Venn diagram between the dot product and the mixture semi-partial correlation with transformation of $c = (0.53, 1.3)$

“Dot”, “Ratio”, “Wavelet”, and “SPC” represent the dot product, the ratio composite, the wavelet composite, and the mixture semi-partial correlation, respectively, and “True” represents the true mass spectra. The numbers in the Venn diagram indicate the number of mass spectra, and the numbers in parenthesis are false discovery rate (FDR). (a) the pair of the dot product and the mixture semi-partial correlation. (b) the pair of the ratio composite and the wavelet composite, and (c) the pair of the ratio composite and the mixture semi-partial correlation.

Table 1
Accuracy of compound identification with different transformations of peak intensity
 The bold numbers indicate the maximum accuracy for each similarity measure. The “No-trans” means $c = (1,0)$.

Transformation		Dot product		Composite Measure			Partial		Semi-partial	
		Ratio	Wavelet	Fourier	Simple	Mixture	Simple	Mixture	Simple	Mixture
No-trans	Accuracy	72.74	67.68	72.98	72.74	73.89	78.17	73.89	79.14	79.14
	Rank				20	50	50	20	100	100
	Mixture Weight					0.1			0.7	0.7
(0.5,3)	Accuracy	80.49	79.57	82.81	82.67	79.49	81.02	79.49	81.25	81.25
	Rank				3	20	3	3	30	30
	Mixture Weight					0.01			0.2	0.2
(0.6,3)	Accuracy	81.86	78.30	82.98	82.65	81.04	82.50	81.04	82.96	82.96
	Rank				3	20	3	3	20	20
	Mixture Weight					0.01			0.2	0.2
(0.5,2)	Accuracy	82.92	81.73	83.27	83.12	81.92	83.30	81.92	83.51	83.51
	Rank				3	30	3	3	50	50
	Mixture Weight					0.01			0.1	0.1
(0.5,1)	Accuracy	83.92	81.70	82.89	82.76	82.74	84.10	82.74	84.14	84.14
	Rank				3	30	3	3	50	50
	Mixture Weight					0.01			0.1	0.1
(0.53,1.3)	Accuracy	84.15	81.72	83.00	82.92	82.98	84.39	82.98	84.59	84.59
	Rank				3	50	3	3	50	50
	Mixture Weight					0.01			0.1	0.1

Table 2
Accuracy of compound identification with 1000 bootstrap samples

The mean, standard deviation (SD), and 95% confidence interval of accuracy of 1000 bootstrap samples are calculated. The “No-trans” means $c = (1,0)$, and “Dot”, “Ratio”, “Wavelet”, “Fourier”, “PC”, and “SPC” represent the dot product, the ratio composite, the wavelet composite, the Fourier composite, the mixture partial correlation, and the mixture semi-partial correlation, respectively.

	Mean	SD	2.50%	97.50%
No-trans				
Dot	0.7273	0.0030	0.7271	0.7275
Ratio	0.6766	0.0047	0.6763	0.6769
Wavelet	0.7297	0.0044	0.7294	0.7299
Fourier	0.7273	0.0043	0.7271	0.7276
PC	0.7816	0.0027	0.7814	0.7818
SPC	0.7832	0.0027	0.7830	0.7834
(0.5,3)				
Dot	0.8049	0.0027	0.8047	0.8050
Ratio	0.7957	0.0039	0.7955	0.7960
Wavelet	0.8281	0.0036	0.8279	0.8284
Fourier	0.8267	0.0037	0.8265	0.8269
PC	0.8102	0.0026	0.8101	0.8104
SPC	0.8119	0.0026	0.8118	0.8121
(0.6,3)				
Dot	0.8186	0.0026	0.8184	0.8188
Ratio	0.7829	0.0040	0.7827	0.7832
Wavelet	0.8298	0.0036	0.8296	0.8301
Fourier	0.8266	0.0037	0.8264	0.8268
PC	0.8250	0.0026	0.8249	0.8252
SPC	0.8296	0.0025	0.8294	0.8298
(0.5,2)				
Dot	0.8292	0.0025	0.8291	0.8294
Ratio	0.8173	0.0038	0.8171	0.8175
Wavelet	0.8328	0.0036	0.8325	0.8330
Fourier	0.8313	0.0036	0.8311	0.8315
PC	0.8330	0.0025	0.8329	0.8332
SPC	0.8343	0.0025	0.8341	0.8344
(0.5,1)				
Dot	0.8392	0.0024	0.8391	0.8394
Ratio	0.8170	0.0038	0.8167	0.8172
Wavelet	0.8289	0.0037	0.8287	0.8291

	Mean	SD	2.50%	97.50%
Fourier	0.8276	0.0037	0.8273	0.8278
PC	0.8409	0.0024	0.8408	0.8411
SPC	0.8413	0.0024	0.8411	0.8414

(0.53,1.3)				
Dot	0.8415	0.0024	0.8413	0.8416
Ratio	0.8171	0.0038	0.8169	0.8173
Wavelet	0.8301	0.0037	0.8299	0.8303
Fourier	0.8292	0.0036	0.8289	0.8294
PC	0.8438	0.0024	0.8437	0.8440
SPC	0.8458	0.0024	0.8457	0.8460

Table 3

Accuracy of compound identification up to top three hits

The “Top-1”, “Top-2”, and “Top-3” are the identification result up to the top one, top two, and top three hits, respectively, based on the similarity scores. The bold numbers indicate the maximum accuracy for each similarity measure. The “No-trans” means $c = (1,0)$.

Transformation		Dot product			Mixture partial			Mixture semi-partial		
		Top-1	Top-2	Top-3	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
No-trans	Accuracy	72.74	83.78	88.34	78.17	87.45	91.16	79.14	88.68	92.34
	Rank				50	50	50	100	100	100
	Mixture Weight				0.1	0.1	0.1	0.7	0.7	0.7
(0.5,3)	Accuracy	80.49	90.40	93.93	81.02	91.03	94.45	81.25	91.55	94.90
	Rank				20	20	20	30	30	30
	Mixture Weight				0.01	0.01	0.01	0.2	0.2	0.2
(0.6,3)	Accuracy	81.86	91.30	94.52	82.50	91.86	95.00	82.96	92.16	95.32
	Rank				20	20	20	20	20	20
	Mixture Weight				0.01	0.01	0.01	0.2	0.2	0.2
(0.5,2)	Accuracy	82.92	92.29	95.38	83.30	92.62	95.64	83.51	92.57	95.59
	Rank				30	30	30	50	50	50
	Mixture Weight				0.01	0.01	0.01	0.1	0.1	0.1
(0.5,1)	Accuracy	83.92	92.75	95.99	84.10	93.03	96.12	84.14	92.96	96.08
	Rank				30	30	30	50	50	50
	Mixture Weight				0.01	0.01	0.01	0.1	0.1	0.1
(0.53,1.3)	Accuracy	84.15	92.94	96.03	84.39	93.13	96.15	84.59	93.14	96.11
	Rank				50	50	50	50	50	50
	Mixture Weight				0.01	0.01	0.01	0.1	0.1	0.1

Table 4
Number of identified 76 compound standards using different spectral matching algorithms with a weight factor of $c = (0.53, 1.3)$

The numbers in diagonal indicate the number of 76 compound standards identified by the method in its row (or column), and the numbers in off-diagonal the number of 76 compound standards commonly identified by the methods in its row and its corresponding column. “Dot”, “Ratio”, “Wavelet”, “Fourier”, “PC”, and “SPC” represent the dot product, the ratio composite, the wavelet composite, the Fourier composite, the mixture partial correlation, and the mixture semi-partial correlation, respectively.

	Dot	Ratio	Wavelet	Fourier	PC	SPC
Dot	49	26	25	25	49	49
Ratio		26	25	25	26	26
Wavelet			25	25	25	25
Fourier				25	25	25
PC					50	50
SPC						54