Compound Identification Using Partial and Semipartial Correlations for Gas Chromatography–Mass Spectrometry Data

Seongho Kim,*† Imhoi Koo,‡ Jaesik Jeong,§ Shiwen Wu,‡ Xue Shi,‡ and Xiang Zhang*‡

†Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky 40292, United States
‡Department of Chemistry, University of Louisville, Louisville, Kentucky 40292, United States
§Department of Biostatistics, School of Medicine, Indiana University, Indianapolis, Indiana 46202, United States

Supporting Information

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 the 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 semipartial 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 Standards and 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 semipartial correlations always outperform both the dot product and its composite measure. The mixture similarity with semipartial 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.

Gas chromatography–mass spectrometry (GC-MS) is the most commonly used method for the analysis of volatile and semivolatile 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,† composite similarity,§ Hertz similarity index,¶ normalized Euclidean distance ($L_2$-norm),§,¶ and absolute value distance ($L_1$-norm).§,¶ Most recently, Koo et al.10 introduced wavelet and Fourier transform-based composite measures and showed that their similarity scores perform better than the dot product and its composite version.

Since some compounds have mass spectral information that is similar to that of other compounds, 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 Scott§ 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 are the most informative ions for compound identification. Therefore, the performance of compound identification 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. Sokolow et al.11 suggested the squared root of an intensity value. Sokolow et al.11 suggested the squared root of an intensity value. 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 Scott§ recommended an intensity to the power of 0.6.

Received: February 22, 2012
Accepted: July 3, 2012
Published: July 3, 2012

dx.doi.org/10.1021/ac301350n1 Anal. Chem. 2012, 84, 6477–6487

© 2012 American Chemical Society
times its m/z value cubed, in the case of the dot product. Horai et al.\textsuperscript{12} reported that the optimal transformation is the square root of intensity and the square of its m/z value. Kim et al.\textsuperscript{13} laterly 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 semipartial (also known as part) correlations as mass spectral similarity measures. The partial and semipartial 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 the accuracy of compound identification using the NIST mass spectral library.

2. THEORY

Dot Product and Intensity Transformation. The dot product,\textsuperscript{5} which is also known as the cosine correlation, was used to obtain the cosine of the angle between two sequences of intensities, \( x = (x_i)_{i=1, \ldots, n} \) and \( y = (y_i)_{i=1, \ldots, n} \). It is defined as

\[
S = S(x, y) = \frac{x^T y}{\|x\| \|y\|} \tag{1}
\]

where \( x^T y = \sum_{i=1}^{n} x_i y_i \) and \( x = (\sum_{i=1}^{n} x_i^2)^{1/2} \).

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

\[
\text{peak intensity}^c \cdot \text{(mass(m/z))}^b \tag{2}
\]

where \( a \) and \( b \) represent the contribution of peak intensity and m/z value, respectively. It is worth noting that Sokolow et al.\textsuperscript{11} reported the optimal transformation \( c = (0.5,1) \). Stein and Scott\textsuperscript{5} later introduced the optimal transformation \( c = (0.6,3) \), while Horai et al.\textsuperscript{12} recommended \( c = (0.5,2) \). Recently, Kim et al.\textsuperscript{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

\[
S_c = S_c(x, y) = S(x', y') = \frac{(x'^T y')}{\|x'\| \|y'\|} \tag{3}
\]

where \( c = (a, b) \) is a vector of transformation factors of intensity and m/z value, respectively. In eq 3, \( x' = (x'_i)_{i=1}^{n} \) and \( y' = (y'_i)_{i=1}^{n} \) are component-wise transformed intensities based on eq 2 and

\[
x'_i = (x_i)^a \cdot (z_i)^b \quad \text{and} \quad y'_i = (y_i)^a \cdot (z_i)^b \tag{4}
\]

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

**Stein and Scott’s Composite Similarity.** Stein and Scott\textsuperscript{5} proposed a composite similarity measure, combining the dot product with transformed intensities, \( S_c \) in eq 3 and similarity of peak ratios, \( S_R \). The ratio similarity of peak pairs \( S_R \) is defined as

\[
S_R(x, y) = \left[ \frac{1}{N_{xy}} \sum_{i=1}^{N_{xy}} \left( \frac{y_i - x_{i-1}}{y_{i-1} - x_i} \right) \right]^n \tag{5}
\]

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

\[
S_{cr} = S_{cr}(x, y) = \frac{N_S S_c(x, y) + N_{xy} S_R(x, y)}{N_x + N_{xy}} \tag{6}
\]

where \( N_c \) is the number of nonzero 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 eq 6 with wavelet and Fourier coefficients, respectively.\textsuperscript{15} 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 \( x^g = (x^g_i)_{i=1}^{n} \) and \( x^h = (x^h_i)_{i=1}^{n} \), respectively. The coefficients of approximations and details are calculated by the following expressions:

\[
x^g_k = \sum_{j=1}^{n} x_j g(2k - j - 1) \quad k = 1, ..., n \tag{7}
\]

\[
x^h_k = \sum_{j=1}^{n} x_j h(2k - j - 1) \quad k = 1, ..., n \tag{8}
\]

where Daubechies’ high-pass \( h \) and low-pass \( g \) filters are used in this study.\textsuperscript{15} Fourier-transformed signal \( x^f = (x^f_i)_{i=1}^{n} \) of a time-domain signal \( x = (x_i)_{i=1}^{n} \) is written by

\[
x^f_k = x^f_k + i x^f_{k'} = \sum_{j=1}^{n} x_j \exp\left(-\frac{2\pi i}{N}(k-1)j\right)
\]

\[
k = 1, ..., n \tag{9}
\]

where \( i \) is the imaginary unit; \( x^f_k \) and \( x^f_{k'} \) are the real and the imaginary parts of \( x^f_k \).

Koo et al.\textsuperscript{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 purposes. Their composite similarity measures are defined, respectively, by

\[
S_{cf} = S_{cf}(x, y) = \frac{N_S S_c(x, y) + N_{xy} S_f(x, y)}{N_x + N_{xy}} \tag{10}
\]
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.

\[
S_D(x, y) = \frac{N_x S_x(x, y) + N_y S_y(x, y)}{N_x + N_y}
\]

where $S_x(x, y) = S(x', y')$ and $S_y(x, y) = S(x', y')$.

Partial and Semipartial Correlation. The partial correlation can be interpreted as the association between two random variables after eliminating the effect of other random variables, while the semipartial 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.\textsuperscript{16–18} They have been applied to biological network studies in order to find direct relationship/association among genes, proteins, and metabolites.\textsuperscript{19–21} Considering a partitioned random vector $\{X, Y, Z\}$ where $X$ and $Y$ are one-dimensional random variables and $Z$ is an $n$-dimensional random vector, the partial correlation $\rho_{XY/Z}$ between $X$ and $Y$ given $Z = \{Z_1, \ldots, Z_n\}$ is the correlation between the residuals $R_{X/Z}$ and $R_{Y/Z}$ and is represented by

\[
\rho_{XY/Z} = \frac{\text{Cor}(R_{X/Z}, R_{Y/Z})}{\sqrt{\text{Var}(R_{X/Z})} \cdot \sqrt{\text{Var}(R_{Y/Z})}}
\]

\[
= \frac{\text{Cov}(R_{X/Z}, R_{Y/Z})}{\sqrt{\text{Var}(R_{X/Z})} \cdot \sqrt{\text{Var}(R_{Y/Z})}}
\]

\[
= \frac{\text{Cov}(X, Y) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}{\sqrt{\text{Var}(X) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, X) - \sqrt{\text{Var}(Y) - \text{Cov}(Y, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}}}
\]

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 semipartial correlation $\rho_{XY/Y}$ between $X$ and $Y$ with $Z = \{Z_1, \ldots, Z_n\}$ is the correlation between the random variable $X$ and $R_{X/Z}$ resulting from the linear regression of $Y$ on $Z$, respectively, and is represented by

\[
\rho_{XY/Y} = \frac{\text{Cor}(X, R_{X/Z})}{\sqrt{\text{Var}(X) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, X)}}
\]

\[
= \frac{\text{Cov}(X, Y) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, Y)}{\sqrt{\text{Var}(X) - \text{Cov}(X, Z) \cdot \text{Var}(Z)^{-1} \cdot \text{Cov}(Z, X)}}
\]

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 semipartial correlation, three situations are taken into consideration, as depicted in Figure 1. Figure 1a depicts a case that none of $X$ and $Y$ is correlated with $Z$, while Figure 1b depicts that both $X$ and $Y$ are correlated with $Z$. In the case of Figure 1c, only the random variable $Y$ is correlated with $Z$. Theoretically, all three correlations have the identical value in the situation of Figure 1a, since $Z$ has nothing to do with $X$ and $Y$. In case of Figure 1c, the partial correlation is exactly same as the semipartial 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 1b. The rationale for the partial and semipartial correlations is to obtain a direct or pure relationship between two random variables. For example, in Figure 1b, 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 1c can be due to the hidden relationship between $X$ and $Z$. For these reasons, the partial or the semipartial 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 semipartial correlations can lead to an incorrect correlation if the correlation between $(X, Y)$ and $Z$ is spurious (caused by noise). For more details on the partial and semipartial correlations, refer to the 2002 work of James\textsuperscript{16} and the 1990 work of Whittaker.\textsuperscript{18}

In the context of compound identification, these partial and semipartial 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 that the intensity transformation does. Suppose that $X$ is a query mass spectrum and $Y = \{Y_1, Y_2, \ldots, Y_m\}$ is a set of $N$ mass spectra.
in the reference library. The partial and the semipartial correlations between X and Y, given \( Y^{(i)} \), can be calculated, respectively, by

\[
\rho_{XY|Y^{(i)}} = \text{Cor}(R_{XY^{(i)}}, R_{XY|Y^{(i)}})
\]

\[
\rho_{XY|Y^{(i)}} = \text{Cor}(X, R_{XY|Y^{(i)}})
\]

where \( Y^{(i)} = Y_1, Y_2, \ldots, Y_n \) and \( Y^{(i)} \) in the eqs 14 and 15 have the identical roles as \( X, Y \), and \( Z \) in the eqs 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 \( Y \). 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 semipartial correlations by considering only the mass spectra that have the first \( k \)-highest similarity scores obtained by the dot product. The partial and semipartial correlations then were computed only for these \( k \) mass spectra. Given the rank \( k \) and transformation \( c \), these partial and semipartial correlations are represented, respectively, by

\[
S_{p,k,c} = S_{p,k,c}(X, Y) = \rho_{XY|Y^{(k)}} = \text{Cor}(R_{XY^{(k)}}, R_{XY|Y^{(k)}})
\]

\[
S_{s,k,c} = S_{s,k,c}(X, Y) = \rho_{XY|Y^{(k)}} = \text{Cor}(X, R_{XY|Y^{(k)}})
\]

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

**Mixture Partial and Semipartial Correlations with the Dot Product.** Mixture similarities of the partial and semipartial correlations were further developed with the dot product. Given \( (k,c,w) \), the mixture partial and semipartial correlations of the two mass spectra \( X \) and \( Y \) are then represented, respectively, by

\[
(1 - w)s_{p,k,c}(X, Y) + wS_{p,k,c}(X, Y)
\]

\[
(1 - w)s_{p,k,c}(X, Y) + wS_{s,k,c}(X, Y)
\]

where \( Y \subseteq Y = \{Y_1, Y_2, \ldots, Y_n\} \) 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

\[
\text{accuracy} = \frac{\text{number of spectra matched correctly}}{\text{number of spectra queried}}
\]

**3. EXPERIMENTAL SECTION**

The NIST Mass Spectral Library and Replicate Spectral Library. We considered the mass spectra extracted from the NIST Chemistry WebBook \(^{14} \) 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 ranged from 1 to 892 with a bin size of 1.

**Experimental Data.** In order to investigate the effect of contaminant mass spectra on the compound identification, an experimental dataset 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, USA). The concentration of each compound in the mixture is 2.5 μg/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 semipartial correlations were calculated by the R package ppcor. The wavelet and Fourier-transform 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-transform composite measure, the (mixture) partial correlation, and the (mixture) semipartial correlation. More specifically, three comparisons were considered: (1) the partial/semipartial correlations versus the dot product, (2) the mixture partial/semipartial correlations versus the dot product, and (3) the mixture partial/semipartial correlations versus the composite measures. For comparison, we further considered the five different transformations: no transformation (where “no transformation” means \( c = (1,0) \)), \((0,5,3),(0,6,3),(0,5,2),(0,5,1),(0,5,3,1,3)\).

**Compound Identification by Partial and Semipartial Correlations.** The partial and semipartial correlations were compared with the dot product according to the different transformations of peak intensity. When calculating the partial and semipartial 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 semipartial 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 2a 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 semipartial correlations are identical, implying that there might be no common feature between the query and the reference library. With no transformation, the partial and semipartial 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 2a. The maximum accuracy was achieved at k = 20 for the case of non-transformation, but at k = 3 for all cases of transformations. Figure 2b depicts the maximum accuracies of compound identification achieved by using all of the partial correlation, the semipartial correlation, and the dot product as mass spectral similarity measures. Clearly, the partial and semipartial correlations outperform the dot product when the mass spectra were not transformed. However, their performance was not as 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 Semipartial Correlations.** The mixture similarity measures between the dot product and the partial/semipartial 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 semipartial correlations.

The accuracy of compound identification for the mixture partial and semipartial correlations is provided in Figure S-1 in the Supporting Information. As a reference, the horizontal gray 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 of partial and semipartial 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 gray line. This is significantly different from the identification results of using the simple partial and semipartial correlations as spectral similarity measures (Figure 2b). The accuracy of compound identification reached the maximum values at the mixture weight w = 0.1 for the mixture of partial correlations and w = 0.7 for the mixture semipartial correlation, as shown in Figure S-1(a) in the Supporting Information and Table 1. The rank k at the maximum accuracy is 50 and 100 for the mixture of partial correlations and the mixture of semipartial 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)–S-1(f) in the Supporting Information. 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 semipartial 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 k = 100. Similarly, 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: 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

![Figure 2](https://dx.doi.org/10.1021/ac301350n1Anal.Chem.2012,84,6477−6487)
The three composite measures were compared with our mixture partial/semipartial correlations. For each transformation, the maximum accuracies of each measure are provided in Figure 3. When the nontransformed spectra were used, the proposed mixture correlations outperformed the three composite measures as well as the dot product, by far. 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 the case of $c = (0.6,3)$ and $(0.5,3)$, the wavelet composite measure performed the best.

Table 1. Accuracy of Compound Identification with Different Transformations of Peak Intensity$^a$

<table>
<thead>
<tr>
<th>transformation</th>
<th>dot product</th>
<th>Composite Measure</th>
<th>Partial</th>
<th>Semipartial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ratio</td>
<td>wavelet</td>
<td>Fourier</td>
<td>simple</td>
</tr>
<tr>
<td>no-trans$^b$</td>
<td>72.74</td>
<td>67.68</td>
<td>72.98</td>
<td>72.74</td>
</tr>
<tr>
<td></td>
<td>rank</td>
<td>20</td>
<td>50</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>mixture weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = (0.5,3)$</td>
<td>80.49</td>
<td>79.57</td>
<td>82.81</td>
<td>82.67</td>
</tr>
<tr>
<td></td>
<td>rank</td>
<td>3</td>
<td>20</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>mixture weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = (0.6,3)$</td>
<td>81.86</td>
<td>78.30</td>
<td>82.98</td>
<td>82.65</td>
</tr>
<tr>
<td></td>
<td>rank</td>
<td>3</td>
<td>20</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>mixture weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = (0.5,2)$</td>
<td>82.92</td>
<td>81.73</td>
<td>83.27</td>
<td>83.12</td>
</tr>
<tr>
<td></td>
<td>rank</td>
<td>3</td>
<td>30</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>mixture weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = (0.5,1)$</td>
<td>83.92</td>
<td>81.70</td>
<td>82.89</td>
<td>82.76</td>
</tr>
<tr>
<td></td>
<td>rank</td>
<td>3</td>
<td>30</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>mixture weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c = (0.53,1.3)$</td>
<td>84.15</td>
<td>81.72</td>
<td>83.00</td>
<td>82.92</td>
</tr>
<tr>
<td></td>
<td>rank</td>
<td>3</td>
<td>50</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>mixture weight</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$The bold numbers indicate the maximum accuracy for each similarity measure. $^b$ The term “no-trans” mean $c = (1,0)$.

Figure 3. Maximum accuracy of compound identification according to the different transformations of intensity. “No-trans” stands for no transformation, “Dot” denotes the dot product, “Ratio” represents the ratio composite, “Wavelet” represents the wavelet composite, “Fourier” denotes the Fourier composite, “PC” represents the mixture of partial correlations, and “SPC” represents the mixture of semipartial correlations.
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

Table 2. Accuracy of Compound Identification with 1000 Bootstrap Samples (the Mean, Standard Deviation (SD), and 95% Confidence Interval of Accuracy of the 1000 Bootstrap Samples Are Calculated)\(^a\)

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Dot Product</th>
<th>Mixture Partial</th>
<th>Mixture Semipartial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-2</td>
<td>Top-3</td>
</tr>
<tr>
<td>No-trans</td>
<td>72.74</td>
<td>83.78</td>
<td>88.34</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>(0.5, 3)</td>
<td>80.49</td>
<td>90.40</td>
<td>93.93</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.6, 3)</td>
<td>81.86</td>
<td>91.30</td>
<td>94.52</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.5, 2)</td>
<td>82.92</td>
<td>92.29</td>
<td>95.38</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.5, 1)</td>
<td>83.92</td>
<td>92.75</td>
<td>95.99</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.53, 1.3)</td>
<td>84.15</td>
<td>92.94</td>
<td>96.03</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\(^a\)“No-trans” means \(c = (1,0)\); “dot” denotes the dot product; “ratio” represents the ratio composite, “wavelet” represents the wavelet composite, “Fourier” denotes the Fourier composite, “PC” denotes the mixture of partial correlations, and “SPC” represents the mixture of semi-partial correlations.

Table 3. Accuracy of Compound Identification up to Top Three Hits\(^a\)

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Dot Product</th>
<th>Mixture Partial</th>
<th>Mixture Semipartial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-2</td>
<td>Top-3</td>
</tr>
<tr>
<td>no-trans</td>
<td>72.74</td>
<td>83.78</td>
<td>88.34</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>(0.5, 3)</td>
<td>80.49</td>
<td>90.40</td>
<td>93.93</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.6, 3)</td>
<td>81.86</td>
<td>91.30</td>
<td>94.52</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.5, 2)</td>
<td>82.92</td>
<td>92.29</td>
<td>95.38</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.5, 1)</td>
<td>83.92</td>
<td>92.75</td>
<td>95.99</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(0.53, 1.3)</td>
<td>84.15</td>
<td>92.94</td>
<td>96.03</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\(^a\)“Top-1,” “Top-2,” and “Top-3” are the identification results up to the top one, two, and three hits, respectively, based on the similarity scores. The bold numbers indicate the maximum accuracy for each similarity measure. The term “no-trans” means \(c = (1,0)\).
intensity. To give 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 semipartial 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). The mean and standard deviation of accuracies of the 1000 bootstrap query samples then 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 mixtures of partial and semipartial 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 the top three hits, together with the dot product. These performances are given in Table 3. As noticed in the previous studies, 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 semipartial correlation (96.11%).

### Analysis of GC × GC-TOF MS Data

The proposed methods were further tested using an experimental dataset 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 of partial and semipartial correlations. Since the true name for each peak is unknown, we just counted the number of 76 compound standards identified by each method (see Table 4).

<table>
<thead>
<tr>
<th>Method</th>
<th>Dot</th>
<th>Ratio</th>
<th>Wavelet</th>
<th>Fourier</th>
<th>PC</th>
<th>SPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>dot</td>
<td>49</td>
<td>26</td>
<td>25</td>
<td>25</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>ratio</td>
<td>26</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>wavelet</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Fourier</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>PC</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPC</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Number of Identified 76 Compound Standards Using Different Spectral Matching Algorithms with a Weight Factor of $c = (0.53,1.3)$\(^{*}\)

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). Similarly, 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 mixtures of partial and semipartial 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 SNR = 1, while the mixture of semipartial correlations outperforms the other similarity measures when 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 semipartial correlation provides the highest identification accuracy, unless the quality of mass spectra is very poor (e.g., SNR ≤ 1).

### 5. Discussion and Conclusions

Compared with the dot product, the simple partial and semipartial correlations improved the accuracy of compound identification up to 1.15% when applied to the nontransformed intensities in Table 1. It is also noteworthy that the mixture partial and semipartial 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 2b. This was caused due to the same roles of the simple partial/semipartial 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 semipartial 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 semipartial
correlations disappeared. However, the use of the simple partial
and semipartial correlations can cause excessive removal of the
effects of other mass spectra, so that identification accuracy
decreases, when the amount of common characteristics among
mass spectra in the reference library is small. For the same
reason, the maximum identification accuracies for the non-
transformed occurred at \( k = 20 \), while the maximum
accuracies for all the transformed spectra occurred at \( k = 3 \) in
the case of the simple partial and semipartial 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 of 0.01 when the mixture of partial
correlations is used as the mass spectral measure with
transformed intensities, while at the mixture weight of 0.1,
when the mixture of semipartial correlations is used as the mass
spectral similarity measure. On the other hand, the maximum
accuracies with nontransformed intensities occurred at 0.1 and
0.7 for the mixtures of partial and semipartial correlations,
respectively. Namely, the contributions of the partial and
semipartial correlations were much larger for the non-
transformed intensities than for the transformed intensities.
This is because the simple partial and semipartial correlations
outperform the dot product in the case of nontransformation,
while the dot product performs better than the simple partial
and semipartial correlations when the transformed intensities
were used. Compared with the mixture semipartial 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 semipartial correlation, so that the distribution of
the partial correlation is more left-skewed than that of the
semipartial correlation. Therefore, the semipartial 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, and Fourier composite measures) is the way
the mixture weight is assigned 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
nonzero 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

Figure 4. Venn diagrams between the dot product and the mixture of semipartial correlations with a transformation of \( c = (0.53, 1.3) \): (a) the pair of the dot product and the mixture of semipartial correlations, (b) the pair of the ratio composite and the wavelet composite, and (c) the pair of the ratio composite, the wavelet composite, and the mixture semipartial correlation, respectively, and the term “True” represents the true mass spectra. The numbers in the Venn diagram indicate the number of mass spectra, and the numbers in parentheses are false discovery rates (FDRs).
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 of partial and semipartial 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 semipartial 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 of semipartial correlations performs better than the mixture of partial correlations, 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 semipartial correlation were 84.15% and 84.59%, respectively, with transformation of \( c = (0.53, 1.3) \). If we acknowledge all 21,516 assignments of the query spectra as positive discoveries, the false discovery rates for the dot product and the mixture semipartial correlation are 15.85% (= 100% − 84.15%) and 15.41% (= 100% − 84.59%), respectively. As shown in Figure 4a, when comparing the assigned CAS indices by the dot product with these by the mixture semipartial correlation, 20,767 (= 2883 + 17,884) assignments out of 21,516 queries had the identical CAS indices for both measures. Among these 20,767 query spectra assigned to the same CAS indices by both measures, 17,884 of them were correctly matched, resulting in an identification accuracy of 86.12% (= 17,884/20,767), while the remaining 2883 assignments (= 20,767 − 17,884) were not correct, even though each of these query spectrum was assigned to the same compound by both measures. If we acknowledge only these 20,767 queries as positive discoveries, the false discovery rate can be reduced by \( \sim 2\% \), 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 in the Supporting Information show that the pair of the ratio and wavelet composites (ratio−wavelet) and the pair of the ratio composite and the mixture of semipartial correlations (Ratio−SPC) have the lowest FDR value, 8.10%. However, ratio−SPC has larger true positives than ratio−wavelet (18,087 versus 16,302) as can be seen in Figures 4b and 4c.

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 of partial or semipartial correlations. Moreover, the literature-reported ratio, wavelet, and 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 of partial and semipartial 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 the transformation of peak intensity.

**ASSOCIATED CONTENT**

Supporting Information
This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

Corresponding Author
*E-mails: s0kim023@louisville.edu (S.K.), xiang.zhang@louisville.edu (X.Z.).

Notes
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

This work was supported by Grant No. 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**

(14) National Institute of Standards and Technology (U.S.) In NIST Standard Reference Database 69, June 2005 release; National Institute of Standards and Technology (NIST): Washington, DC.


