analytical chemistry

Article

pubs.acs.org/ac

Data Preprocessing Method for Liquid Chromatography–Mass Spectrometry Based Metabolomics

Xiaoli Wei,[†] Xue Shi,[†] Seongho Kim,[‡] Li Zhang,[¶] Jeffrey S. Patrick,[¶] Joe Binkley,[¶] Craig McClain,[§,∥,⊥,#] and Xiang Zhang*[,†]

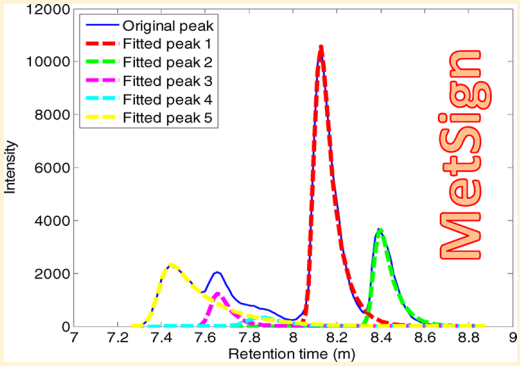
Departments of [†]Chemistry, [‡]Bioinformatics and Biostatistics, [§]Medicine, [∥]Pharmacology and Toxicology, and [⊥]Alcohol Research Center, University of Louisville, Louisville, Kentucky 40292, United States

[#]Robley Rex VA Medical Center, Louisville, Kentucky 40206, United States

[¶]LECO Corporation, St. Joseph, Michigan 49085, United States

Supporting Information

ABSTRACT: A set of data preprocessing algorithms for peak detection and peak list alignment are reported for analysis of liquid chromatography–mass spectrometry (LC–MS)-based metabolomics data. For spectrum deconvolution, peak picking is achieved at the selected ion chromatogram (XIC) level. To estimate and remove the noise in XICs, each XIC is first segmented into several peak groups based on the continuity of scan number, and the noise level is estimated by all the XIC signals, except the regions potentially with presence of metabolite ion peaks. After removing noise, the peaks of molecular ions are detected using both the first and the second derivatives, followed by an efficient exponentially modified Gaussian-based peak deconvolution method for peak fitting. A two-stage alignment algorithm is also developed, where the retention times of all peaks are first transferred into the *z*-score domain and the peaks are aligned based on the measure of their mixture scores after retention time correction using a partial linear regression. Analysis of a set of spike-in LC–MS data from three groups of samples containing 16 metabolite standards mixed with metabolite extract from mouse livers demonstrates that the developed data preprocessing method performs better than two of the existing popular data analysis packages, MZmine2.6 and XCMS[2], for peak picking, peak list alignment, and quantification.



Data preprocessing plays a critical role in metabolomics and can greatly affect the outcome of the data analysis.[1] In the past few years, several bioinformatics tools have been developed for analysis of liquid chromatography–mass spectrometry (LC–MS)-based metabolomics data. For instance, XAlign software was developed to align the peak lists of LC–MS data for both proteomics and metabolomics study.[2] Lommen and Kools developed MetAlign 3.0 for noise reduction, baseline correction, and peak picking.[3] XCMS[2] enables peak picking, alignment, statistical analysis, metabolite identification, and structural characterization.[4] An online version of XCMS was also reported.[5] MZmine2 is capable of peak detection, peak list alignment, normalization, statistical analysis, visualization, and peak identification for LC–MS data.[6] MZedDB uses adducts and neutral loss fragments as predicted ionization behavior "rules" to annotate LC–MS data. In addition, the correlation analysis and the isotope enumerator were presented to confirm the m/z versus signal relationships and to verify the exact isotopic distribution, respectively.[7] Sturm et al. developed OpenMS for LC–MS data analysis, including visualization, data reduction, alignment, and retention time prediction by using a support vector machine (SVM) method.[8] Hoekman et al. developed msCompare that allows the arbitrary combination of different feature detection/quantification and alignment/matching algorithms in conjunction with a scoring method to evaluate the overall LC–MS data processing.[9]

We introduced MetSign for analysis of LC–MS and direct infusion mass spectrometry (DI-MS) data.[10] MetSign provides solutions for peak detection, visualization, tentative metabolite assignment, peak list alignment, normalization, clustering, and time course analysis. A significant feature of MetSign is its ability to analyze the stable isotope labeled data and time course data. The MetSign has been applied to analysis of DI-MS data of translational metabolomics projects.[11,12] However, there are some limitations in MetSign for analysis of LC–MS data including limited accuracy in deconvoluting overlapping chromatographic peaks and aligning metabolite peak lists.

The objective of this study was to develop more accurate data preprocessing algorithms for peak detection and peak list alignment for LC–MS-based metabolomics, where the accuracy of peak detection is measured by the number of detected peaks, peak location (retention time), peak area, and m/z value of metabolite ion, while the accuracy of peak list alignment is

Received: June 22, 2012
Accepted: August 29, 2012
Published: August 29, 2012

   dx.doi.org/10.1021/ac3016856 | *Anal. Chem.* 2012, 84, 7963–7971

measured by the number of aligned spiked-in compound standards. The precision, recall, and F1 score in recognizing the spiked-in compounds from different sample groups are used as measures for quantitative analysis of the spiked-in compound standards. We have developed a new method to deconvolute the instrument spectra using an intensive peak-favored method to construct selected ion chromatogram (XIC), using both the first derivatives and the second derivatives for peak detection, and exponentially modified Gaussian (EMG) mixture model for peak fitting. For peak list alignment, a two-stage retention time window-free alignment algorithm was developed to recognize metabolite peaks generated by the same type of metabolite from multiple peak lists, where the peak similarity is measured by a mixture score. The developed methods have been implemented in MetSign and used to analyze a set of spike-in data acquired on an LC−MS system. The performance of these methods was compared with two existing software packages. MetSign software was implemented using MATLAB 2010b and is free for purpose of academic research.

EXPERIMENTAL SECTION

Spike-In Samples. About 60 mg of liver tissue from each mouse was mixed with deionized water at a ratio of 100 mg/mL. The mixture was then homogenized for 2 min and stored at −80 °C until use. Amounts of 100 μL of homogenized liver sample, 20 μL of butylated hydroxytoluene (BHT) mixture (50 mg BHT into 1 mL of methanol), and 800 μL of methanol were mixed and vortexed for 1 min followed by centrifugation at 4 °C for 10 min at 15 000 rpm. An amount of 700 μL of the supernatant was aspirated into a plastic tube and dried by N_2 flow. After dissolving the dried sample with 100 μL of methanol, a stock solution was prepared by diluting the sample 10 times. Aliquots of 20 μL of each of 14 mouse liver exacts were combined to make the pooled sample for this work.

A mixture of 16 compound standards was prepared at a concentration of 100 μg/mL for each compound (Supporting Information Table S-1). Amounts of 20, 50, and 80 μL of the standard mixture were added to each of the 100 μL pool samples. Dichloromethane/methanol (v/v = 2:1) was then added to each of the three vials to make the total volume of 200 μL. This resulted in three sample groups with spiked-in compound standards. The concentration of compound standards in each of the spike-in sample groups was 10, 25, and 40 μg/mL, respectively.

LC−MS Analysis. A Citius LC-HRT high-resolution mass spectrometer equipped with an Agilent 1290 Infinity UHPLC with a Waters Acquity UPLC BEH hydrophilic interaction chromatography (HILIC) 1.7 μm, 2.1 mm × 150 mm column was used in this work. The sample was loaded in H_2O plus 5 mM NH_4OAc plus 0.2% acetic acid (buffer A) and separated using a binary gradient consisting of buffer A and buffer B (90/10 acetonitrile/H_2O plus 5 mM NH_4OAc plus 0.2% acetic acid). Flow rate was set at 250 μL/min on the column, with 100% B for 4 min, 45% B at 12 min holding to 20 min, 100% B at 21 min and holding to 60 min for the gradient. The Citius LC-HRT was operated with an electrospray ionization source in positive ion mode with spray voltage set at 3.0 kV, nozzle temperature at 125 °C, desolvation heater temperature at 900 °C, desolvation flow at 7.5 L/min, and nebulizer pressure at 50 psi. The system was optimized in high-resolution mode (R = 50 000 (fwhm)) with folded flight path (FFP) technology and was mass calibrated externally using Agilent ESI tune mixture (G2421A). The mass spectrometry was operated in a full mass mode (low energy) followed by a tandem MS/MS mode (high energy) with a mass range of m/z = 50−1000. The scan frequency for acquiring the full mass spectra and MS/MS spectra is five spectra/s, respectively.

THEORETICAL BASIS

Figure 1 depicts the workflow of this work. The components of normalization and statistical significance tests have been

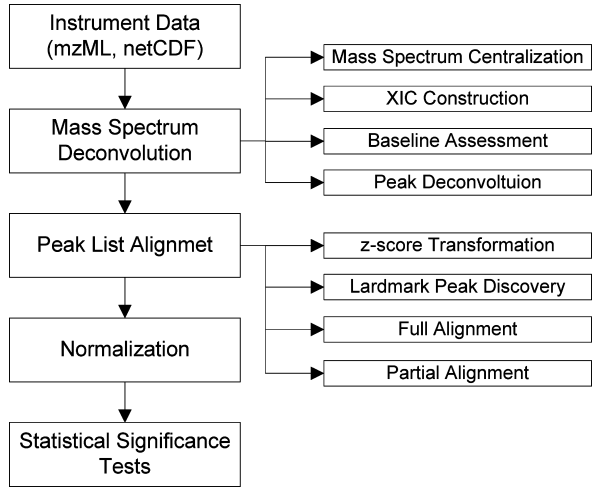


Figure 1. Flowchart of data analysis methods developed in this study.

implemented in the previous version of MetSign software. This work focused on developing new algorithms for mass spectrum deconvolution and peak list alignment.

Mass Spectrum Deconvolution. The mass spectra can be acquired in either profile mode or centroid mode in LC−MS. MetSign provides two options to centralize the mass spectra acquired under profile mode: second-order polynomial fitting-based local maxima (SPF-LM) and one-dimensional discrete wavelet-transform (1-DWT). The SPF-LM approach employs the first-derivative operation to detect the local maxima in the spectrum, followed by the second-order polynomial fitting (SPF) to fit each local peak. After determining the peak location in the m/z dimension, the m/z value and corresponding peak intensity of each profile peak can be obtained. In the 1-DWT approach, each mass spectrum is first transformed using the one-dimensional discrete wavelet-transform, followed by detecting all the local maximum values in the wavelet domain, which is corresponding to the m/z and intensity of each profile peak in the spectrum. The DWT is defined as follows:

$$x_k^{\mathrm{va}} = \sum_{d=1}^{n} x_d g[2k - d - 1], \quad k = 1, \ldots, n \tag{1}$$

$$x_k^{\mathrm{vd}} = \sum_{d=1}^{n} x_d h[2k - d - 1], \quad k = 1, \ldots, n \tag{2}$$

where $X = (x_1, x_2, \ldots, x_n)$ denotes the input signal, g and h are low-pass and high-pass filters, respectively, x_k^{va} denotes the approximations of the signal resulted by the low-pass filter, and x_k^{vd} denotes the details of the signal resulted by the high-pass filter. The Daubechies wavelets are used to transform mass spectra into wavelet domains in this study.[13]

dx.doi.org/10.1021/ac3016856 | Anal. Chem. 2012, 84, 7963−7971

To detect metabolite peaks at the chromatographic dimension, the XIC is first constructed for each m/z value of metabolite ion with a variation window $\varepsilon_{m/z}$. MetSign constructs the XICs in favor of abundant peaks. It combines the information of the (m/z, peak area) pair in all scans to generate two matrices. One is sorted by m/z values, while the other is sorted by peak area values. The XICs are then constructed sequentially based on the rank of peak area of each ion in the peak area matrix, from the most abundant peak to the least abundant one. To a selected ion, its corresponding m/z value is used to search the entire m/z matrix to extract all ions with m/z values within the predefined mass accuracy $\varepsilon_{m/z}$. Another user-defined parameter, minimum chromatographic peak width w_c measured as the number of scans, is applied to eliminate the XICs with a maximum number of continuous scans less than w_c.

To estimate and remove the noise in XICs, each XIC is first segmented into several peak groups based on the continuity of scan number. A segment refers to a range of scans in which the molecular ion was detected in every scan. The first-derivative approach is used to recognize the significant peaks in each peak segment, with several predefined filtering criteria including minimum chromatographic peak width, more than three data points monotonously increasing and decreasing on each side of a peak, and a minimum ratio of 0.3 between the peak area of the original peak and the smoothed peak. All data points belonging to the significant peaks are then removed from each segment, and the remaining data points in all segments of the same XIC are used as training data to assess the noise level. A polynomial fitting is employed to estimate the noise level in the regions of the significant peaks. The median filtering method is further employed on the entire training data to get the noise level at each point in the entire XIC.

An EMG-based peak deconvolution method was used in this study for peak fitting. The EMG distribution is defined as follows:[14]

$$f(x;\lambda) = y_0 + \frac{A}{t_0}\exp\left[\frac{1}{2}\left(\frac{w}{t_0}\right)^2 - \frac{x - x_c}{t_0}\right]\left[\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right] \quad (3)$$

where $\lambda = \{y_0, A, x_c, w, t_0\}$, y_0 is the initial value, A is the amplitude, x_c is the center of the peak, w is the width of the peak, t_0 is the modification factor, and $z = (x - x_c/w) - (w/t_0)$, erf is the error function. The EMG mixture model with n components is defined as follows:

$$f(x;\lambda_1,\lambda_2,\ldots,\lambda_n) = \sum_{i=1}^{n} f(x;\lambda_i) \quad (4)$$

To improve the efficiency of peak fitting using the EMG mixture model, the raw spectrum data of each XIC are first smoothed using a moving average method after removing the noise from each data point. The major peaks are then detected using a first-derivative method on the smoothed data. The second derivatives of the previously smoothed data are calculated to detect small peaks overlapping with other major peaks. After determining the total number of peaks, the EMG mixture model is applied iteratively for peak fitting.

Peak List Alignment. To align the metabolite peaks generated by the same type of metabolite in different samples, the retention time value of each peak in a peak list is first transformed into z-score as follows:

$$z_{i,j} = \frac{t_{i,j} - \mu}{\sigma} \quad (5)$$

where $t_{i,j}$ denotes the retention time of the jth peak in the ith sample and μ and σ are the median values of the means and standard deviations of the retention time values among peaks of all sample set S, respectively. The z-score transforms the retention time values into a normal distribution, which then enables the alignment of heterogeneous data (i.e., the experimental data acquired under different experimental conditions).[15]

To align all the peak lists together, it is necessary to select a peak list as a reference peak list (R_{PL}) and align the rest of the peak lists to it. It is better that the peak distribution of R_{PL} be similar to most of the peak lists. The two-dimensional Kolmgorov–Smirnov (K–S) test is employed to study the similarity of peak distributions between two peak lists in the z-score transformed retention time and m/z plane, where each given peak can be represented as a data point $(t_i, (m/z)_i)$. Each peak actually separates the plane into four quadrants $[t > t_i, m/z > (m/z)_i]$, $[t < t_i, m/z > (m/z)_i]$, $[t < t_i, m/z < (m/z)_i]$, and $[t > t_i, m/z < (m/z)_i]$. An integrated probability in each of these four natural quadrants around a given point can be calculated. The statistic D of the K–S test is taken to be the maximum difference (ranging both over data points and over quadrants) of the corresponding integrated probabilities.[16] The D statistics of K–S test are calculated for all pairs of the peak lists. After removing the large D values detected as outliers at a confidence level of 95%, a peak list with the smallest average D value is considered as the reference peak list R_{PL}.

After selecting R_{PL}, the peak alignment method is developed as a two-stage algorithm without requiring retention time variation (i.e., parameter-free for retention time): full alignment and partial alignment. The goal of full alignment is to recognize the landmark peaks, which are a set of metabolite peaks generated by the same type of metabolite that are present in every sample. In partial alignment, the peaks in a test sample that are not recognized as the landmark peaks are aligned.

By using the D statistics of K–S test to select R_{PL}, the rest of the peak lists are considered as test samples. Then the sample set can be written as $S = \{R_{PL}, T_1, T_2, \ldots, T_i, \ldots, T_n\}$, where T_i is the peak list of sample i. Each of the test samples is aligned to R_{PL}, respectively. The content of R_{PL} is updated after the alignment of R_{PL} with a test sample.

Considering two samples, $\{R_{PL}, T_i\}$, all m/z value-matched peak pairs between these two samples are selected using a user-defined m/z variation window $\varepsilon_{m/z}$. If a peak can be matched to multiple peaks in the other peak list, the peak pair with the minimum z-score transformed retention time difference is selected. Therefore, the m/z matched peak pairs can be recorded as $\{(r_1, s_1), (r_2, s_2), \ldots, (r_p, s_p)\}$, where r_j is a peak in R_{PL}, s_j is the corresponding m/z matched peak in T_i, and p is the total number of m/z matched peak pairs. The m/z matched peak pairs are further filtered based on Euclidean distance of retention time between r_j and s_j, i.e., $d_j = |r_j - s_j|$, by setting a confidence interval of 95%. The retention time filtered peak pairs are represented as $\{(r_1, s_1), (r_2, s_2), \ldots, (r_m, s_m)\}$ and $m \le p$. This process is iteratively operated on all the test samples, respectively.

A mixture score, S_m, is then used to measure the matching quality between two m/z matched peaks as follows:

dx.doi.org/10.1021/ac3016856 | Anal. Chem. 2012, 84, 7963–7971

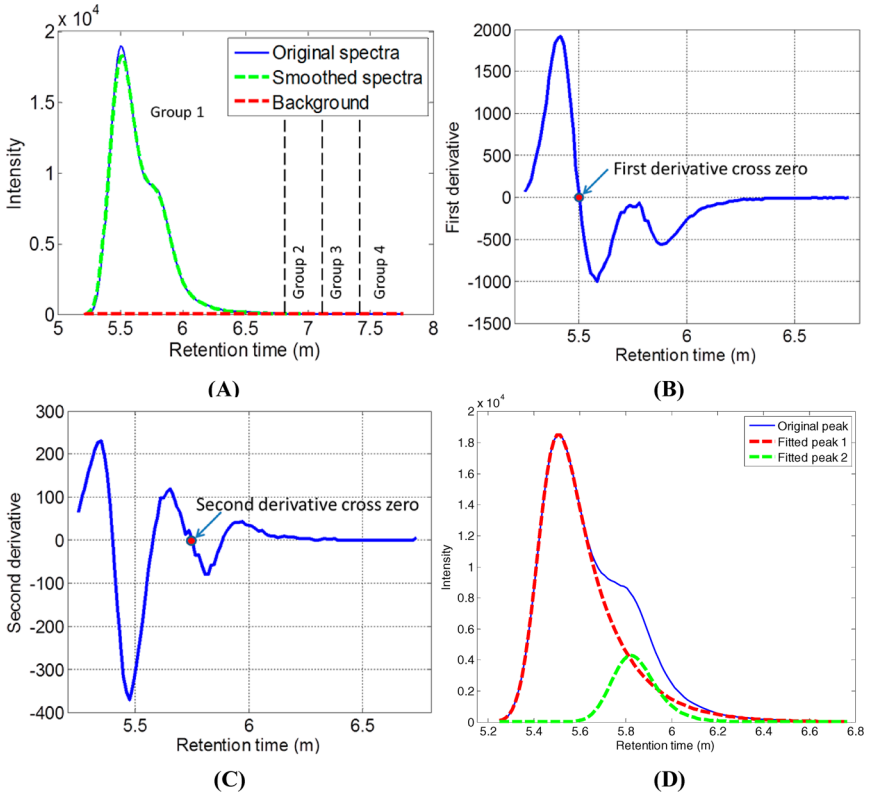


Figure 2. Example of spectrum deconvolution by MetSign. (A) XIC and background noise level estimation. The entire XIC is segmented into four peak groups because of the discontinuity of signals in the chromatographic dimension (scan). It was detected that the first segment contains at least one peak and leaving the rest retention time range of XIC as noise area for polynomial fitting and median filtering. The estimated noise level is shown in the red line. (B) Detection of significant peaks. The dominant peaks are determined by the first-derivative cross zero position from positive to negative values and meeting the criteria of minimum data points in the two sides of each peak. (C) Detection of nonsignificant peaks (hidden peaks). The hidden peaks are recognized as the second-derivative cross zero position with changing from positive to negative values and the first-derivative value is negative, or changing from negative to positive values and the first-derivative value is positive. There is one hidden peak that is detected in the example. (D) Two peaks deconvoluted by mixture EMG models.

$$S_m(d_i, \Delta_i|w) = w \exp\left(-1.6\frac{d_i - d_{min}}{d_{med} - d_{min}}\right) + (1 - w)\frac{1}{1 + \Delta_i} \quad (6)$$

where d_i is the Euclidean distance of retention time between the ith matched peak pair, d_{min} and d_{med} are the minimum and median retention time distance among all m/z matched peaks in the two peak lists, respectively, Δ_i is the absolute value of m/z difference between the ith matched peak pair, and w is a weight factor and $0 \le w \le 1$. The peaks that are present in every test peak list and matched to the same peak in R_{PL} are then used to optimize the value of weight factor w for the alignment of a test peak list T_i and R_{PL} by maximizing the value of $\sum_{i=1}^{k} S_m(d_i, \Delta_i|w)$; k is the number of matched peaks between the test peak list T_i and R_{PL}, and w is set as 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95, respectively.

After optimizing the weight factor, w, the value of S_m can be calculated for each matched peak pair between the test peak list T_i and R_{PL}, followed by an outlier detection in S_m^j, $j = 1 \ldots k$. By iteratively considering pair set $\{R_{PL}, T_i|i = 1, \ldots, n\}$, the optimal weight factor set $\{\omega_1, \ldots, \omega_n\}$ can be obtained for each peak list. The landmark peaks, which are represented as $\{(r_1, t_{11}, \ldots, t_{n1}), \ldots, (r_m, t_{1m}, \ldots, t_{nm})\}$, are then obtained after outlier removal on S_m. The minimum mixture score S_m^{min} among all the test peak lists is then used as a threshold value on the mixture score in the partial alignment.

To perform the partial alignment, the retention time value of each landmark peak in the test peak list T_i is assigned to the retention time value of the corresponding landmark peak in R_{PL}. A local polynomial fitting method is employed to correct the retention time of peaks present between two adjacent landmark peaks as follows:

$$t'_{t,s} = t_{r,i} + \frac{t_{t,s} - t_{t,i}}{t_{t,i+1} - t_{t,i}}(t_{r,i+1} - t_{r,i}) \quad (7)$$

where $i = 1, 2, \ldots, m - 1$, $t_{r,i}$ and $t_{r,i+1}$ are the retention time values of two adjacent landmark peaks in R_{PL}, and $t'_{t,s}$ is the corrected retention time of a peak eluted between the two adjacent landmark peaks in the test sample T_i.

Because multiple landmark peaks are detected in a set of experimental data, adjusting retention time shifts by using two adjacent landmark peaks can correct nonlinear retention time shifts of metabolite eluted between these two adjacent landmark peaks. An iterative optimization method is applied to the group of peaks eluted earlier than the first-eluted landmark peak and the group of peaks eluted later than the last-eluted landmark peak, respectively. In each optimization process, 30% of landmark peaks are randomly selected from the pool of landmark peaks $\{(r_1, t_{11}), \ldots, (r_m, t_{1m})\}$ and a polynomial model fitting error is computed as follows:

dx.doi.org/10.1021/ac3016856 | Anal. Chem. 2012, 84, 7963−7971

$$\varepsilon = \sum_{i=1}^{q} |t^{o}_{R,i} - t^{f}_{R,i}| \quad (8)$$

where $t^{o}_{R,i}$ is the original z-score transformed retention time of the ith peak, $t^{f}_{R,i}$ is the fitted retention time of the ith peak, and q is the number of peaks in the test peak list at the region of interest. This process is repeated 1000 times, and the model with minimum fitting error is selected and used for retention time correction.

After the retention time correction, partial alignment is applied to all the nonlandmark peaks present in each of the test peak lists and then aligned to the peaks present in the R_{PL}, where a mixture score S_m is calculated using eq 6 for each peak pair. A peak pair is considered as a match if its mixture score is larger than S_m^{min}. It is possible that one peak in the test sample can be matched to multiple peaks in R_{PL} and vice versa. In these cases, the peak pair with the maximum mixture score is kept while the remaining matches are discarded. If there is a peak in the test peak list that cannot be matched to any peaks in R_{PL}, this peak is considered as a new peak to R_{PL} and is added to R_{PL}. The updated R_{PL} is then used to align the peaks in the next test peak list. This process is repeated until all the test peak lists are aligned.

Normalization and Statistical Significance Tests. Three literature-reported normalization algorithms were implemented into MetSign for the user to select from, including quantile normalization, cyclic loss normalization, and contrast-based normalization.[17,18] The purpose of statistical analysis is to find metabolites that have significantly different expression levels in different sample groups. MetSign first employs the Fisher's exact test to study the presence and absence of each metabolite between sample groups. It then employs the Grubbs' test[19] for outlier detection to find the responses of a metabolite that are not consistent with the responses of the same metabolite in the remaining samples of the same sample group. After removing the outliers, an abundance test such as the pairwise two-tail t test is performed on the log-transformed peak areas to detect the abundance changes of each metabolite between two sample groups, and the false discovery rate (FDR) is used to correct for multiple comparisons.[20]

RESULTS AND DISCUSSION

The raw instrument data were converted into mzML format by instrument control software ChromaTOF, and the mzML files were used as input files of MetSign and MZmine2. All instrument data were also exported into netCDF for analysis using XCMS[2].

Peak Detection. The peak detection was performed at XIC level in MetSign. To construct XICs for all molecular ions, the mass accuracy and minimum chromatographic peak width of a molecular ion were set as $\varepsilon_{m/z} \leq 6$ ppm and $w_c \geq 15$ scans, respectively. In addition to these two user-defined parameters, MetSign also employs two more default values to filter low-quality peaks detected in each XIC segment: more than three data points monotonously increasing and decreasing on each side of a peak, and a minimum ratio of 0.3 between the peak area of the original chromatographic peak and the smoothed one. Figure 2A is an example of XIC, in which the entire XIC was separated into four segments (or peak groups) based on the continuity of scan number. The first-derivative method detected one peak in the first peak group (PG-1) with an m/z value of 809.5863 and a span of retention time from 5.25 to 6.68 min (Figure 2B), while no peak was detected in the other three XIC segments. All data points in the entire XIC, except those belonging to the detected peak in PG-1, were used for polynomial fitting followed by median filtering to estimate the noise level (red dotted line in Figure 2A).

MetSign constructs the XICs in favor of intensive peaks, by constructing XICs in descending order of the maximum peak height of metabolite ions in full mass spectra. By doing so, if the m/z value of a signal s = (m/z, peak height) overlaps with the m/z values of two types of metabolite ions, s is assigned to the XIC of the metabolite ion with a larger value of maximum peak height in its full mass spectra. The hypothesis of such an intensive peak-favored XIC construction approach is that a metabolite ion with a large value of maximum peak height in its full mass spectra is more likely to be measured than the one with a small value of maximum peak height. Compared to the conventional approach of constructing XICs in sequential based on m/z value, this method gives the priority of having an m/z signal to the large peak, and most likely such a large peak is generated by a true metabolite.

The user-defined values of m/z variation window $\varepsilon_{m/z}$ and the minimum chromatographic peak width w_c play significant role in constructing XICs and filtering the detected peaks. A large value of $\varepsilon_{m/z}$ introduces a high rate of assigning peaks to a wrong XIC while a small value of $\varepsilon_{m/z}$ can exclude some peaks to be assigned a correct XIC. These two cases reduce the accuracy of peak detection and quantification. A large value of peak width w_c increases the chance of removing small true peaks while a small value causes detection of a large number of small false peaks. In practice, the values of $\varepsilon_{m/z}$ and w_c should be determined by analysis of a number of authentic standards.

A significant challenge in peak deconvolution is to determine the number of peaks, especially the detection of small peaks that overlap with a dominant peak. A two-layer peak detection method was implemented in this study, in which the first derivative on the smoothed data was applied to detect the dominant peaks (Figure 2B) and the second derivative was used to detect the overlapping nonsignificant peaks (Figure 2C), with the constraints of the minimum number of data points in each side of a peak (more than three data points monotonously increasing and decreasing on each side of a peak) and a minimum chromatographic peak width (a user-defined span of scans, 15 scans in this study). After determining that there were two peaks in PG-1, all data in PG-1 were subjected to EMG mixture model for peak fitting. Each EMG component corresponds to one peak, as shown in Figure 2D. The overall fitting results are displayed in Supporting Information Figure S-1.

Peak overlapping is common LC−MS data due to the complexity of metabolites present in a metabolome and the limited peak capacity in LC−MS. MetSign employs the first derivatives to find the significant peaks while it uses the second derivatives to find the hidden peaks. Compared to other literature-reported methods, such a peak detection method maximizes the chance of detecting all peaks from each XIC. Furthermore, the chromatographic peaks may not always have a normal peak shape if the chromatographic conditions are not optimized, which most likely results in right-tailing chromatographic peaks. For this reason, MetSign used the EMG mixture model for peak fitting to address the asymmetric nature of chromatographic peaks. However, the EMG mixture model requires accurate determination of the number of peaks to be fitted. The proposed peak detection method using both the first

dx.doi.org/10.1021/ac3016856 | Anal. Chem. 2012, 84, 7963−7971

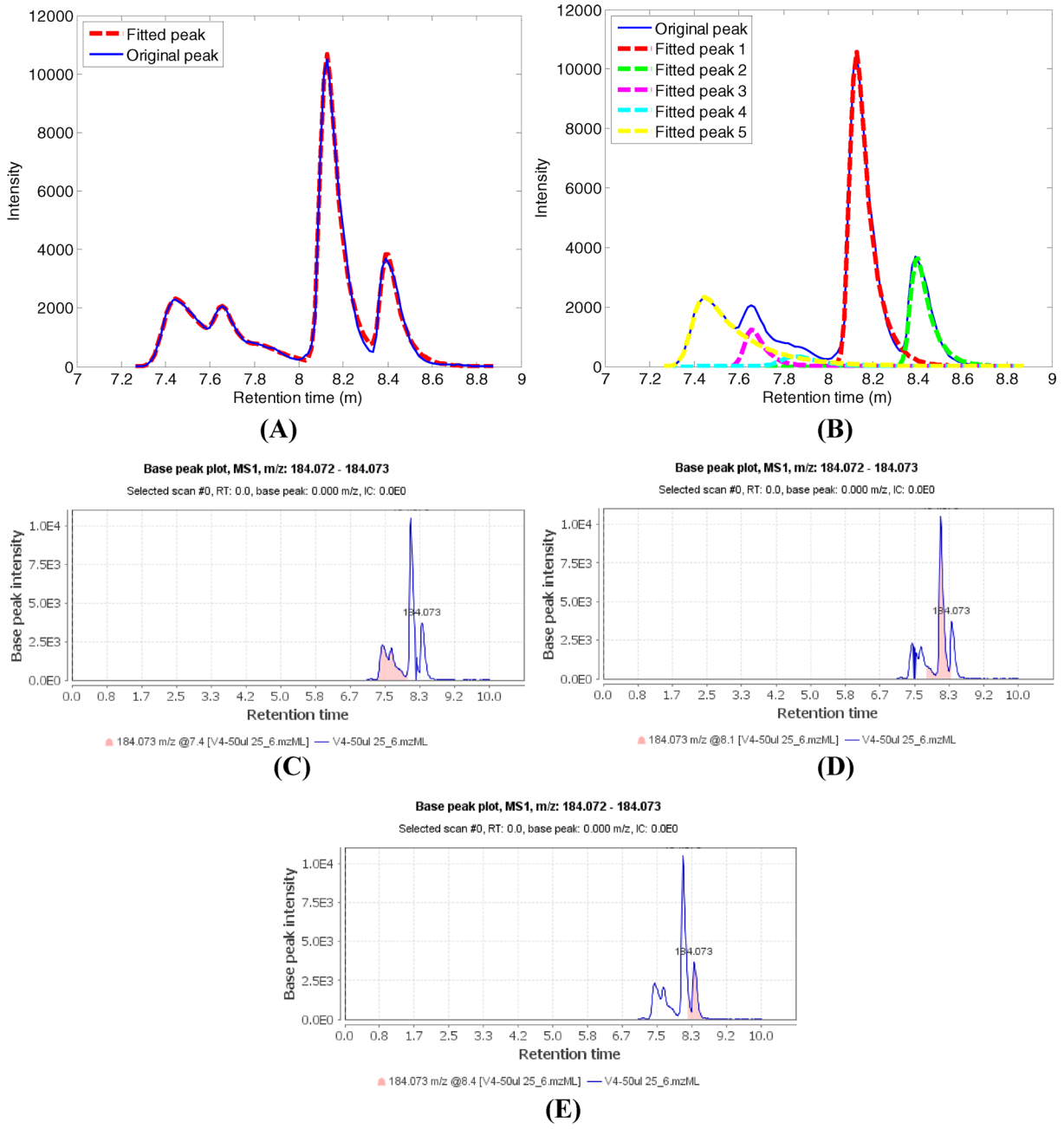


Figure 3. Example of peak picking by MetSign and MZmine2.6. (A) Peak fitting results by MetSign using the mixture EMG model. (B) Five peak components deconvoluted by peak detection and EMG fitting algorithm by MetSign. Panels C−E are the peak deconvolution results on the same data by MZmine2.6. MetSign detected five peaks including four dominant peaks and one hidden peak, whereas MZmine2.6 correctly detected the two abundant peaks on the right and incorrectly considered the three peaks on the left as one peak.

and the second derivatives provides high precision in determining the peak number and the position of each peak present in an XIC and, therefore, increases the fitting accuracy of EMG mixture model and also aids the model converge. The EMG mixture model not only provides the accurate peak location (i.e., retention time) for peak list alignment but also provides accurate peak area for downstream peak quantification.

Peak List Alignment. Of the 16 spiked-in metabolites, 14 metabolites were fully aligned in all 18 samples while compounds lysoPC(10:0) and heptadecanoic acid were aligned in 17 and 15 samples, respectively. The mixture score of compound lysoPC(10:0) in one sample is smaller than the threshold S_m^{min}. The compound, heptadecanoic acid, was detected in two samples with a large m/z variation compared to the theoretical value, and therefore, the peak information of this compound was removed during the initial peak assignment. Supporting Information Figure S-2 shows the distribution of relative standard deviation (RSD) of the aligned peaks by MetSign.

The significant features of the two-stage alignment algorithm developed in this work include that the developed algorithm does not need a user-defined retention time variation window for peak alignment across samples and it can align data acquired under difference experimental conditions. These are achieved by first transforming the retention time values to z-score

dx.doi.org/10.1021/ac3016856 | Anal. Chem. 2012, 84, 7963−7971

domain to ensure the normal distribution of all peaks in the z-score transformed chromatogram. Using the mixture score enables the simultaneous evaluation of the deviations of Euclidian distance and m/z values between the metabolite peaks of interest. This not only significantly increases the quality of the aligned peak pairs but also increases the chance of aligning peaks with large deviation in either retention time or m/z values caused by the inaccuracy of peak picking algorithms. For example, assuming a metabolite ion has a poor quality of chromatographic peak shape, the retention time value assigned to this peak by the peak picking algorithm may, therefore, have a large deviation from the true value after smoothing. If the m/z value of this peak is correctly assigned, it is possible that this peak may still have a large mixture score with peaks generated by the same metabolite in the other samples, and therefore, they all can be aligned.

Comparison with Existing Software Packages. To compare the performance of peak picking, an m/z variation of $\varepsilon_{m/z} \leq 6$ ppm was used in all three software packages. XCMS[2] software does not output the results of peak picking. Therefore, the results of peak detection generated by MetSign were compared with these by MZmine2.6. The Savitzky−Golay method in MZmine2.6 gave the best deconvolution results and, therefore, was used for comparison. The other parameters used in MZmine2.6 are the following: minimum time span, 0.2 min; minimum peak height, 10; peak duration range, 0.2−10 min; derivative threshold level, 0.5. Figure 3 depicts an example of peak deconvolution result using MetSign and MZmine2.6. Figure 3A shows the EMG model fitted peak results. MetSign deconvoluted the instrument data into five overlapping peaks (Figure 3B) with peak location in the chromatographic dimension and peak height of (7.4, 2.3×10^3), (7.6, 1.2×10^3), (7.8, 337), (8.1, 1.1×10^4), and (8.4, 3.6×10^3), respectively. MZmine2.6 only recognized three peaks with corresponding peak information of (7.4, 2.3×10^3), (8.1, 1.1×10^4), and (8.4, 3.7×10^3), respectively. The peaks located in the range of retention time of 7.3−8.0 min were considered as one peak with a peak area of 5.1×10^4, even though three peaks were actually present in this region with peak areas detected by MetSign as 4.6×10^4, 5.3×10^3, and 1.0×10^4, respectively. The areas of peaks located at retention time 8.1 and 8.4 min detected by MetSign and MZmine2.6 were very similar, 8.7×10^5 and 2.9×10^4 by MetSign and 8.5×10^5 and 3.3×10^4 by MZmine2.6. Supporting Information Figure S-3 is another example of deconvoluting instrument data by MetSign and MZmine2.6. These two examples demonstrate that MZmine2.6 has limited capability of deconvoluting overlapping peaks, especially in cases where small peaks overlap with large peaks. The details of deconvoluted peak information of the data displayed in Figure 3 and Supporting Information Figure S-3 are listed in Supporting Information Table S-2.

To compare the performance of peak alignment methods among the three software packages, the peak lists generated by the three software packages were aligned, respectively. Pluskal et al. demonstrated that the RANSAC aligner is better than Join aligner in MZmine.[21] XCMS[2] can align all samples with or without retention time correction. Therefore, we chose the RANSAC alignment in MZmine2.6, XCMS[2] with retention time correction, and XCMS[2] without retention time correction for comparison. In MZmine2.6, 1.0 and 0.2 min were set as the retention time tolerance before and after retention time correction, respectively, and RANSAC iteration number was set to 1000 times. The retention time tolerance in XCMS[2] was also set to 0.2 min. MetSign does not need the retention time tolerance for alignment. The m/z variation was set to $\varepsilon_{m/z} \leq 6$ ppm in MetSign and MZmine2.6, while the default value of $\varepsilon_{m/z} \leq 0.025$ m/z was used in XCMS[2]. Figure 4 depicts the

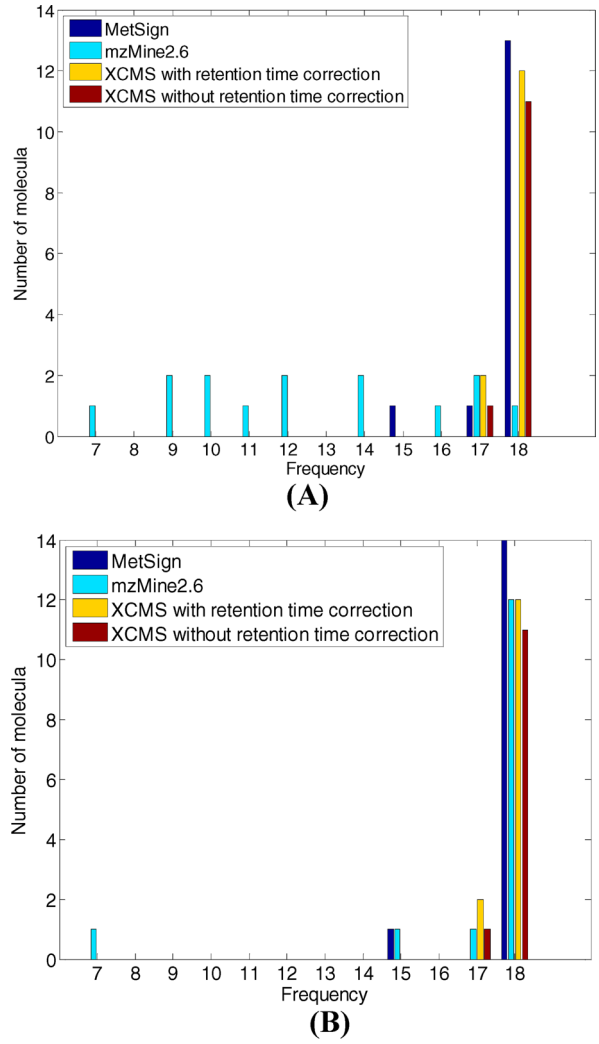


Figure 4. Comparison of alignment results among MetSign, MZmine2.6, XCMS[2] with retention time correction, and XCMS[2] without retention time correction. (A) $\varepsilon_{m/z} \leq 6$ ppm. (B) $\varepsilon_{m/z} \leq 10$ ppm. The $\varepsilon_{m/z}$ was set as 0.025 for XCMS[2] as specified by the software.

alignment results of the 16 spiked-in metabolites by MetSign, MZmine2.6, and XCMS[2]. On the basis of the experimental design, all of the spiked-in metabolite standards should be correctly aligned. In Figure 4A, the m/z variation was set as $\varepsilon_{m/z} \leq 6$ ppm for MetSign and MZmine2.6. In MetSign, a total of 13 peaks of the spiked-in standards were fully aligned in all 18 samples. In XCMS[2] with retention time correction, 12 metabolite standards were fully aligned, while 11 were fully aligned without retention time correction. However, there was only one metabolite that was fully aligned by MZmine2.6.

Figure 4B depicts the alignment results of the 16 spiked-in metabolites by increasing the m/z variation to $\varepsilon_{m/z} \leq 10$ ppm in MetSign and MZmine2.6 and keeping the default m/z variation in XCMS[2]. In MetSign, a total 14 peaks of the spiked-in metabolite standards were fully aligned in all 18 samples, while 12 were fully aligned by MZmine2.6. It should be noted that

dx.doi.org/10.1021/ac3016856 | Anal. Chem. 2012, 84, 7963−7971

the same m/z variation $\varepsilon_{m/z} \leq 6$ ppm was used in both MetSign and MZmine2.6 for peak picking. Compared with the performance of MetSign, a large m/z variation was required by MZmine2.6 for alignment, indicating that a relatively large m/z variation was introduced during peak picking by MZmine2.6. Even though the alignment performance of MZmine2.6 is significantly increased by setting a large value of m/z variation $\varepsilon_{m/z} \leq 10$ ppm, the overall alignment performance of MZmine2.6 is slightly better than XCMS[2] without retention time correction, but still worse than MetSign and XCMS[2] with retention time correction. Therefore, the alignment accuracy of the three comparing software packages in decreasing order is MetSign > XCMS[2] > MZmine2.6.

Three sample groups with different concentrations of the spiked-in metabolites were used to construct three data sets for comparative analysis of quantification accuracy between the three software packages. Details of three measures, the true-positive rate (TPR), positive predictive value (PPV), and their harmonic mean F1 score, are introduced in the Supporting Information. The analysis results are listed in Supporting Information as Table S-3. On the basis of the experimental design, all the 16 spiked-in metabolites detected from the experimental data are the true-positive metabolites that have significant concentration changes between two testing sample groups, while any other metabolites detected with significant concentration differences are false-positives.

Supporting Information Table S-1 shows that MetSign outperforms both XCMS[2] and MZmine2.6 in the analysis of these three data sets in all three measures, TPR, PPV, and F1, regardless of the p-value threshold. MZmine2.6 has a better performance than XCMS[2] in both PPV and F1 values even though XCMS[2] performed better than MZmine2.6 in TPR for all p-value thresholds. There is no significant difference in the TPR between XCMS[2] and MetSign in the analysis of the three data sets, even though MetSign performed slightly better than XCMS[2]. It should be noted, however, that the PPV of MetSign and MZmine2.6 is more than 2 times better than XCMS[2], and the resulting F1 scores of the three software packages in the analysis of the spike-in data is in a descending order as follows: MetSign > MZmine2.6 > XCMS[2]. This analysis shows that MetSign outperforms the existing software packages MZmine2.6 and XCMS[2] in metabolite quantification.

Many data analysis steps are involved in analyzing LC−MS-based metabolomics profiling data. Data preprocessing can significantly affect the outcome of data analysis. The accurate spectrum deconvolution and peak alignment algorithms developed in this work can provide precise metabolite information for the downstream quantification and network analysis. The developed MetSign software can be used for analysis of any high-resolution LC−MS-based metabolomics data, where multiple samples are analyzed to assess the metabolic difference between sample groups for the purpose of biomarker discovery, drug development, or any other comparative analysis.

CONCLUSIONS

To further enhance the accuracy of MetSign software in analysis of LC−MS data, a set of data preprocessing algorithms were developed for spectrum deconvolution and peak list alignment. For spectrum deconvolution, peak picking was achieved at the XIC level. The XIC is constructed using an intensive peak-favored approach. To estimate and remove the noise in XICs, each XIC is first segmented into several peak groups based on the continuity of scan number. After removing noise, the peaks of molecular ions are detected using both the first and the second derivatives followed by an efficient EMG mixture model for peak fitting. For peak list alignment, the retention time values in each peak list are first transformed into z-scores to ensure the normal distribution of peaks in the z-score domain. Another innovation of this work is using a mixture score to simultaneous evaluate the similarity of Euclidian distance and m/z values between the metabolite peaks of interest. The two-stage design of alignment ensures to first align the peaks with highest quality and then use these peaks to calibrate (adjust) the retention time of the remaining peaks by partial linear regression.

Comparative analysis of spike-in data demonstrates that MZmine2.6 has limited capability of deconvoluting overlapping peaks and introduces a relatively large m/z variation during peak picking. The overall performance of spectral deconvolution and peak list alignment of XCMS[2] is slightly worse than that of MetSign. For quantitative analysis, MetSign outperforms both CMS[2] and MZmine2.6 in analysis of the spike-in data in all three measures, TPR, PPV, and F1, regardless of the p-value threshold. Overall, the developed data preprocessing methods perform better than the existing software MZmine2.6 and XCMS[2] for peak picking, alignment, and quantification. The accurate spectrum deconvolution and peak alignment algorithms developed in this study provide precise metabolite information for the downstream quantification and network analysis and, therefore, can reduce the data analysis variation and improve the quality of metabolic profiling data.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +01-502-852-8878. Fax: +01-502-852-8149. E-mail: xiang.zhang@louisville.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Mrs. Marion McClain for review of this manuscript. This work was supported by NIH Grant 1RC2AA019385 through the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and 1RO1GM087735 through the National Institute of General Medical Sciences (NIGMS) and the VA.

REFERENCES

(1) van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.

(2) Zhang, X.; Asara, J. M.; Adamec, J.; Ouzzani, M.; Elmagarmid, A. K. *Bioinformatics* **2005**, *21*, 4054.

(3) Lommen, A.; Kools, H. *Metabolomics* **2012**, *8*, 719 DOI: DOI: 10.1007/s11306-011-0369-1.

(4) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80*, 6382.

(5) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, 5035.

(6) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.

(7) Draper, J.; Enot, D. P.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H. *BMC Bioinf.* **2009**, *10*, 227.
(8) Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinf.* **2008**, *9*, 163.
(9) Hoekman, B.; Breitling, R.; Suits, F.; Bischoff, R.; Horvatovich, P. *Mol. Cell. Proteomics* **2012**, *11*, M111.015974.
(10) Wei, X.; Sun, W.; Shi, X.; Koo, I.; Wang, B.; Zhang, J.; Yin, X.; Tang, Y.; Bogdanov, B.; Kim, S.; Zhou, Z.; McClain, C.; Zhang, X. *Anal. Chem.* **2011**, *83*, 7668.
(11) Zhong, W.; Zhao, Y.; Tang, Y.; Wei, X.; Shi, X.; Sun, W.; Sun, X.; Yin, X.; Kim, S.; McClain, C. J.; Zhang, X.; Zhou, Z. *Am. J. Pathol.* **2012**, *180*, 998.
(12) Shi, X.; Wahlang, B.; Wei, X.; Yin, X.; Falkner, K. C.; Prough, R. A.; Kim, S. H.; Mueller, E. G.; McClain, C. J.; Cave, M.; Zhang, X. *J. Proteome Res.* **2012**, *11*, 3805.
(13) Grunbaum, F. A. *Science* **1992**, *257*, 821.
(14) Grushka, E. *Anal. Chem.* **1972**, *44*, 1733.
(15) Wang, B.; Fang, A.; Heim, J.; Bogdanov, B.; Pugh, S.; Libardoni, M.; Zhang, X. *Anal. Chem.* **2010**, *82*, 5069.
(16) Press, W. H.; Teukolsky, S. A.; Vettering, W. T.; Flannery, B. P. *Numerical Recipes in C++, The Art of Scientific Computing*; 2nd ed.; Cambridge University Press: Cambridge, U.K., 2002.
(17) Dudoit, S.; Yang, Y. H.; Callow, M. J.; Speed, T. P. *Statistica Sinica* **2002**, *12*, 111.
(18) Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. *Bioinformatics* **2003**, *19*, 185.
(19) Grubbs, F. *Technometrics* **1969**, *11*, 1.
(20) Newton, M. A.; Noueiry, A.; Sarkar, D.; Ahlquist, P. *Biostatistics* **2004**, *5*, 155.
(21) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.