analytical chemistry

Article

pubs.acs.org/ac

Ion Fusion of High-Resolution LC−MS-Based Metabolomics Data to Discover More Reliable Biomarkers

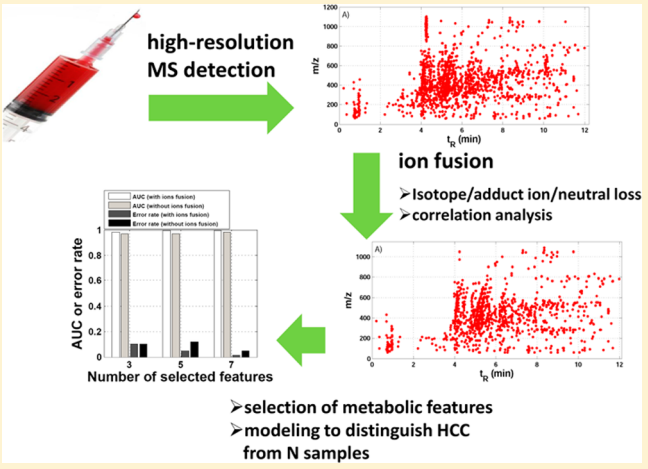
Zhongda Zeng,[†] Xinyu Liu,[†] Weidong Dai,[†] Peiyuan Yin,[†] Lina Zhou,[†] Qiang Huang,[†] Xiaohui Lin,[‡] and Guowang Xu*,[†]

[†]Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China

[‡]School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

Supporting Information

ABSTRACT: A systematic approach for the fusion of associated ions from a common molecule was developed to generate "one feature for one peak" metabolomics data. This approach guarantees that each molecule is equally selected as a potential biomarker and may largely enhance the chance to obtain reliable findings without employing redundant ion information. The ion fusion is based on low mass variation in contrast to the theoretical calculation measured by a high-resolution mass spectrometer, such as LTQ orbitrap, and a high correlation of ion pairs from the same molecule. The mass characteristics of isotopic distribution, neutral loss, and adduct ions were simultaneously applied to inspect each extracted ion in the range of a predefined retention time window. The correlation coefficient was computed with the corresponding intensities of each ion pair among all experimental samples. Serum metabolomics data for the investigation of hepatocellular carcinoma (HCC) and healthy controls were utilized as an example to demonstrate this strategy. In total, 609 and 1084 ion pairs were respectively found meeting one or more criteria for fusion, and therefore fused to 106 and 169 metabolite features of the datasets in the positive and negative modes, respectively. The important metabolite features were separately discovered and compared to distinguish the HCC from the healthy controls using the two datasets with and without ion fusion. The results show that the developed method can be an effective tool to process high-resolution mass spectrometry data in "omics" studies.



Liquid chromatography−mass spectrometry (LC−MS) has been one of the main pillars in both targeted and nontargeted metabolomics studies to separate and then characterize small molecules, which is a preliminary step for biomarker discovery, metabolic pathway analysis, and the further interpretation of biological processes or mechanisms.[1,2] To date, identification of a large amount of metabolites strongly relies on the high accuracy of high-resolution mass spectrometer. For example, the error of mass accuracy of LTQ Orbitrap Elite and TripleTOF 5600+ attains to 1−2 ppm or lower.[3,4] Most of commercial or freely available databases for identification of small molecules were established on the basis of high accuracy of LC−MS^n detection, such as HMDB (University of Alberta, Edmonton, Canada), and Massbank (Institute for Advanced Biosciences, Tsuruoka City, Japan).

The general procedure for LC−MS^n-based metabolomics data processing is first to detect and align the peak ions of metabolites with the help of program scripts, such as (meta)XCMS or existing software platforms distributed by the instrument manufacturers. The input is converted from raw chromatograms by instrumental software, including Sieve (Thermo Fisher, Waltham, Massachusetts, USA), MassHunter (Agilent, Santa Clara, California, USA), and Markerlynx (Waters, Milford, Massachusetts, USA).[5,6] On the basis of the exported table of peak ions, identification of metabolites is achieved by the precise MS and retention time (t_R). Next, a series of methods are probably applied to improve data quality such as the zero-value removal and scaling operation, evaluation or calibration of quality control (QC) and real samples. Nonparametric tests, methods for feature selection, and other rich multiple multivariate data tools are applied to find differential metabolites, and further establish the classification model.[7] The metabolites that contribute to a low error rate for prediction (or independent validation set) or a large area under the curve (AUC) may be finalized as potential biomarkers for disease diagnosis and other biological or clinical applications.

Received: November 14, 2013
Accepted: March 10, 2014
Published: March 10, 2014

   dx.doi.org/10.1021/ac500878x | *Anal. Chem.* 2014, 86, 3793−3800

The standard flowchart introduced above has largely accelerated the development of metabolomics. Many scientific outcomes have been generated with the help of such a strategy. For example, Wu et al. defined key metabolites based on the detected ions to reveal the phase transition of locusts.[8] Chen et al. discovered biomarkers for diagnosing epithelium ovarian cancer.[9] Dunn et al. reported a procedure to process the large-scale metabolic profiling produced by GC−MS or LC−MS instruments.[6] Full use of the ion features of metabolites is the basis for identifying "real" biomarkers. In some studies, all small molecules were first identified to avoid the utilization of redundant ions for feature selection and modeling,[10] the task is very time-consuming. An evident shortcoming of the conventional methods based on all detected features is that more than one ion likely corresponds to a metabolite, which unavoidably reduces the possibility of establishing an accurate model due to the redundant data. Additionally, this redundancy burdens the computation task, and further weakens the performance of chemometrics tools for feature selection and sample classification as the effectiveness of multivariate methods for biomarker discovery is tightly correlated with the size of the dataset and colinearity among the metabolic features.[11] "One feature for one peak" is helpful to equally discover "real" biomarkers, fusion of ions from the same molecule is very important to avoid using redundant ion information.

Most of the previous approaches for data fusion were proposed to tune multisource data obtained from different laboratories, analytical platforms, or batch processing. The prior knowledge of mass correlation, such as neutral loss and adduct ions, was widely applied to improve the accuracy of metabolite identification.[12−14] Although chemometrics resolution methods were attempted to mathematically separate overlapping peak clusters, recognizing the ions of different molecules without providing full chromatographic and spectral profiles is difficult. Some software platforms, such as Sieve 2.0, can generate a purified peak table with the removal of isotopic ions, but many redundant and unidentified ions can still be found in the exported table of the complicated samples with thousands of ion features. In particular, the algorithms are not freely available to the public, which makes the modification and improvement of these methods impossible outside of the instrument manufacturers.

In this work, a systematic approach was proposed to fuse the peak ions originating from the same metabolite. Three types of relationships, that is, isotopic distribution, adduct ions, and neutral loss, were comprehensively taken into account. The correlation analysis of peaks among samples was further implemented for ion fusion based on the mathematical invariance of the intensity ratio of the same molecule in different samples. Next, variable importance in the projection (VIP), a widely used method for the discovery of key metabolite features in metabolomics research, was applied to simultaneously process the datasets with and without ion fusion. Then, the reasonability and effectiveness of the identified features from the two different datasets were further investigated to interpret the necessity for ion fusion. A simple classification procedure with the samples as a training set was employed to establish a model for performance evaluation by using partial least-squares-discriminant analysis (PLS-DA). A metabolomics dataset with hepatocellular carcinoma (HCC) and healthy control (N) samples was used as an example to demonstrate this strategy. Furthermore, the benefits of the proposed method for the characterization of unknown metabolites were exemplified with the simultaneous employment of multiple fused ions derived from the same small molecule.

THEORY

Aligned Table of Peak Ions. The aligned peak table is mostly generated by commercial software platforms developed by the instrument manufacturers, or freely available programs, such as XCMS.[15,16] New algorithms and scripts to extract and further align peaks among tens to hundreds of samples are still under development.[17,18] Of course, potential false combinations of ions may be included because of the complexity of metabolomics data with too many features and the low signal-to-noise (S/N) ratio of some peaks with a weak response. In general, such a table includes three fundamental pieces of information, which are retention time (t_R), precise m/z, and intensity of the corresponding ions, for each sample. Thus, a peak table $\mathbf{X}_{m\times n}$ with a size of m rows and n columns represents m experimental samples and $(n - 2)$ ion features of these samples, with shared t_R and m/z information. The differences in the t_R and m/z of ions among samples individually depend on the repeatability of chromatographic separation, the performance of precise m/z detection by the instruments, and the complexity of the mixture matrix. The sample amount is another influential factor that determines the degree of difference because a large sample size unavoidably enhances the chance of a t_R shift and mass variation. The mass accuracy of LTQ orbitrap and Fourier transform (FT) mass spectrometers attains 1−2 ppm and 0.1 ppm, respectively. This level should be accurate enough for ion fusion in combination with t_R shift constraints.

Ion Fusion Based on *m/z* Relationships. Adduct ion is formed by interaction of a specific ion and a usual molecule. The former is often generated within the ion source with chemical or physical ionization, and the latter is an uncharged molecule. Each adduct ion contains all the constituent atoms of one species as well as additional atoms. These ions may have m/z correlations with the common precursor ion if they are derived from the same molecule. Such correlations include isotopic distribution, adduct ions and neutral loss, which are the basis of ion fusion. That is, a molecule may produce a large number of different ions because of the existence of isotopic atoms or interaction with more than one species of ionization. This makes many features in the same "ion pool" actually be related to the same metabolite, and generates the so-called "one metabolite with many features" system, which results in data redundancy and complicates the discovery of biomarkers. Evidently, the correlation information of features should be extracted and applied to fuse the ions from the same molecule. Figure S1 (Supporting Information) shows the strategy of ion fusion; the principal details are given below.

Isotopic Distribution Pattern. In terms of the knowledge about a high-resolution mass spectrometer for MS detection, the ion pairs with the same t_R and precise mass differences matching the theoretical isotopic distribution can be recognized as products of the same molecule, such as ^{12}C and ^{13}C, which have a mass difference of 1.0034 Da. In this study, removal of isotopic ions was first implemented before the fusion of ion features. The more accurate identification of a metabolite can be achieved by combining a high-resolution mass search and isotopic characteristics.

Adduct Ions. Adduct ions are formed within the ion source by the interaction of a molecule with an additional ion. Most of

dx.doi.org/10.1021/ac500878x | Anal. Chem. 2014, 86, 3793−3800

these adduct ions in the ESI mode has been summarized by researchers (see http://fiehnlab.ucdavis.edu, and Table S1 (Supporting Information) as well). On the basis of the table of adduct ions, the possible molecular weight [M] is determined according to the maximum number of appearances in the range of a predefined t_R shift window and the m/z error. That is, a series of [M] are inversely acquired by assuming the adduct ion is the interaction product of uncharged molecule [M] and different ions such as $[H]^+$, $[Na]^+$, $[K]^+$, $[NH_4]^+$, or others in ESI+ mode. The final [M] is optimally determined to the one with largest possibility found, which is generated from a large amount of adduct ions in a predefined t_R window. The ion pairs with an [M] difference equal to or less than the tolerance of the m/z error (such as 0.005 Da for orbitrap) are fused together to represent the same molecule. Of course, those ion pairs with one common ion will be fused into a single feature. It should be noted that reproducibility of t_R should be considered to reduce the chance of a false positive outcome. The ions obtained from the positive and negative modes can be processed simultaneously if the same buffers are used. However, the correction of t_R should be first applied to align the retention time shift.

Neutral Loss. Neutral loss is a type of uncharged molecule produced from a molecular ion or other associated ions, which cannot be measured by the MS detector. However, rich structure information is contained in the neutral molecule with correlations to the parent and daughter ions. The typical neutral loss, including the names and formulas, has been summarized elsewhere.[19] It is further included in Table S2 (Supporting Information) for easy-to-use purposes. Thus, each ion pair with a mass difference meeting the precise change of the neutral loss can be recognized as fragments from the same molecule. Of course, this recognition will unavoidably produce false positive outcomes for complicated systems that include a large number of ions. The constraints of ions with a stringent t_R shift and mass error will maximally avoid such errors. The utilization of precise m/z detection using a high-resolution mass spectrometer is the basis of ion fusion. The MS with a relatively low mass accuracy should be processed cautiously. For example, few different but structure-related molecules probably generate ion pair with quite close t_R, in which the m/z difference further meets the weight of a specific neutral loss. This unavoidably generates false ion fusion by using low resolution MS measurement. A typical example is certain lysophosphatidylcholines (LPC) and lysophosphatidylethanolamines (LPE) molecules with structure difference of $-CH_2-CH_2-CH_2-$, and then theoretical mass difference of 42.0798 Da, because it approaches the neutral loss of acetylation. Such false-positive results should be excluded in this approach with an additional step to evaluate the performance of ion fusion.

Correlation Analysis of Ions in Different Samples. Other information employed for fusion is the high correlation of two specific ions among samples. The MS consistency of the same molecule in different runs is a basic assumption used to investigate the similarity and difference of samples. Using two ions, *a* and *b*, as an example, the correlation coefficient (CC) of the peak intensities among all experimental samples should be theoretically close to 1 if they are derived from the same molecule because the relative ratio of the ion intensities should ideally be a constant value, without consideration of the experimental error and background shift. Thus, each ion pair with a sufficiently high CC value among the samples can be recognized as daughter ions of the same molecule, and can then be reasonably fused together. The CC value of two ions thoroughly depends on the stochastic change of the relative intensity if they come from different molecules. However, different tolerances of the CC have a significant influence on the ion fusion. Currently, there is no a reliable criterion to guarantee absolute accuracy for identification simply based on a CC evaluation. In this study, the CC was set to 0.95 according to our experience, although it can be set to 0.85 or even lower.

EXPERIMENTAL SECTION

Samples. The discovery of biomarkers with a high specificity has great importance for the clinical diagnosis and therapy of many diseases. In this work, 30 HCC and 30 N samples were collected in the First Affiliated Hospital of the Medical School, Zhejiang University. The HCC samples were confirmed by pathological experiment. The operations for sample collection, transportation, and storage followed the standard procedures for metabolomics studies.[20]

LC−MS Analysis. The nontargeted analysis of metabolites was performed. The serum samples were first thawed at room temperature before extraction. Then, 100 μL of acetonitrile was added to 400 μL of the sample to precipitate the proteins. After vortexing for 30 s and incubating for 30 min at 4 °C, the sample was centrifuged (12,000 r/min) for 10 min at the same temperature. The supernatant was collected and lyophilized to dryness. Next, the sample was dissolved with 100 μL of acetonitrile solution (acetonitrile:water = 1:4), and the supernatant was used for LC−MS analysis after new centrifuge operation. The QC sample was prepared by mixing 10 μL solutions of each real sample, and it was applied to evaluate the reliability of the datasets with sequence analysis by LC−MS.

The instrumental system was a Thermo Fisher Accela HPLC and LTQ Orbitrap XL mass spectrometer. The column was a 10 cm × 2.1 mm I.D. × 1.7 μm HSS T3 (Waters, Milford, Massachusetts, USA). Both ESI+ and ESI− modes were applied for the MS ionization and detection. The following chromatographic conditions were used for separation: the column temperature was 50 °C, the injection volume was 8.0 μL, and the flow rate was 0.3 mL/min. The mobile phases were composed of 0.1% formic acid−water (A) and 0.1% acetonitrile−water (B) in the positive mode and water (A) and 95% methanol−water (B) with 6.5 mM ammonium bicarbonate in the negative mode. The elution gradient was optimized to 1% B in the initial state, held for 2.0 min, and then programmed to 4.5% B from 2.0−4.0 min until 100% B from 4.0−11.0 min. Then, it was held for 3.0 min and balanced for 4.0 min at the final status.

The conditions for MS detection were as follows. The data collection frequency was 1.0 s/spectrum in both ESI+ and ESI− modes with mass detection from 50 to 1100 Da. The ion sprayer voltages were set to 4.5 kV and 3.5 kV under the two ionization modes, respectively. The maximum resolution attains to 100 000 @ m/z 400, and mass accuracy is lower than 2 ppm root-mean-square (RMS) with internal calibration. This should guarantee the accuracy for ion fusion.

The ion extraction and alignment were accomplished by software Sieve 2.0, which was developed by Thermo Fisher. The differences of the t_R and mass error of the ions were set to 0.05 min and 0.005 Da in terms of the t_R reproducibility and measurement error of the ions among samples, respectively. These settings were extensively used for ion fusion in the next section.

Data Analysis. All computer programs, except those specially mentioned in this text, were coded in-house with

dx.doi.org/10.1021/ac500878x | *Anal. Chem.* 2014, 86, 3793−3800

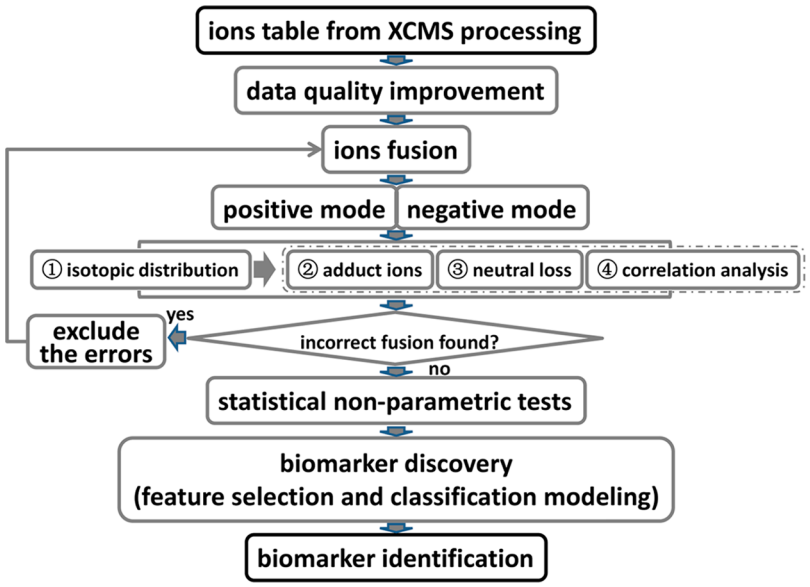


Figure 1. Complete flowchart to illustrate the procedure for ion fusion.

the MATLAB environment (Version 7.14.0.739, R2012a, 64-bit), including the data pretreatment, ion fusion, statistical analysis, feature selection and classification modeling for biomarker discovery. The calculations were implemented on a DELL compatible personal computer with an Intel(R) Xeon (R) CPU E3-1225 V2 @ 3.20 GHZ and 4.0 GB RAM memory. The operating system was Windows XP 64-Bit Edition.

RESULTS AND DISCUSSION

To be convenient, a metabolomics dataset from HPLC-LTQ orbitrap is applied to demonstrate the proposed approach for ion fusion and further biomarker discovery in this work. This dataset includes HCC and N samples for a molecular classification study of liver cancer. Figure 1 illustrates the entire flowchart from the raw peak table to data quality improvement and fusion, biomarker discovery and further characterization. The parameters of t_R shift and m/z difference of ions are determined according to the experimental conditions and measurement accuracy of the instrument, and they are consistent with those used in the step for the generation of the peak table. After fusion of the ions of each metabolite according to the method provided in Figure 1, the ion with the maximum response among all samples is then defined for the subsequent data processing.

Data Pretreatment and Initial Principal Component Analysis (PCA). After obtaining the aligned ion table, data pretreatment was first applied to improve the data quality, including the "80% rule"[21] to remove zero values, replacement of the remaining zeros and data normalization. The 80% rule means that those features with the number of zeros being more than 20% of the total number of samples in any class are removed from the dataset, which has been widely adopted in metabolomics. The remaining zero values were replaced by 0.01%, which is the minimum value of the corresponding feature across all samples. This step helps reduce the interference from zero values, and guarantees the reliability of the results. Data normalization has a significant effect on discovering biomarkers with an interpretable biological mechanism. The aim is to balance the weights of different variables for feature selection. This has been systematically investigated in a reported work to compare the performance of different methods, including autoscaling, range scaling, Pareto scaling and centering.[22] In this work, autoscaling was applied to normalize the ions among samples. A nonparametric test (Mann−Whitney *U*-test used in this work) was introduced to identify the ions with a significant difference among different classes of samples.

In Figure S2 Supporting Information, the results from PCA were introduced into the datasets after data pretreatment. Figure S2a, and S2d, S2b and S2e, and S2c and S2f Supporting Information are plot pairs obtained from the data under the positive ionization mode, negative ionization mode, and their combination, respectively. The top plots show the percentage of explained variance of the first 10 principal components (PCs), and the bottom plots demonstrate the two-dimensional score plots using the first 2 PCs. The identification of obvious clustering trends from the PCA outcomes was not difficult.

Performance of Ion Fusion Using the Proposed Method. After the 80% rule was applied along with other procedures to improve the data quality, the remaining ion features were applied as input for fusion. The three key parameters (t_R shift (Δt_R), m/z difference between experimental and theoretical values ($\Delta m/z$), and CC tolerance ($CC_{tolerance}$)) were defined as 0.05 min, 0.005 Da, and 0.95, respectively, with the same settings as in the peak alignment among the samples.

In total, 1305 and 1523 ion features were left after data preprocessing in the positive and negative ionization modes, respectively. The results of the CC values of all ion pairs among the 60 samples are shown in Figure S3 (Supporting Information). Note that these results are for the analysis of correlation matrices with the sizes of 1305 × 1305 and 1523 × 1523 in the two modes, respectively. The approximate normal distribution of the correlation coefficients was found by uniformly dividing the entire range from −1 to +1 into 20 parts for the statistical analysis. Thus, the absolutes of most CC values were not large (approximately zero), and the remainder satisfied the criteria for ion fusion with the help of correlation analysis.

dx.doi.org/10.1021/ac500878x | *Anal. Chem.* 2014, 86, 3793−3800

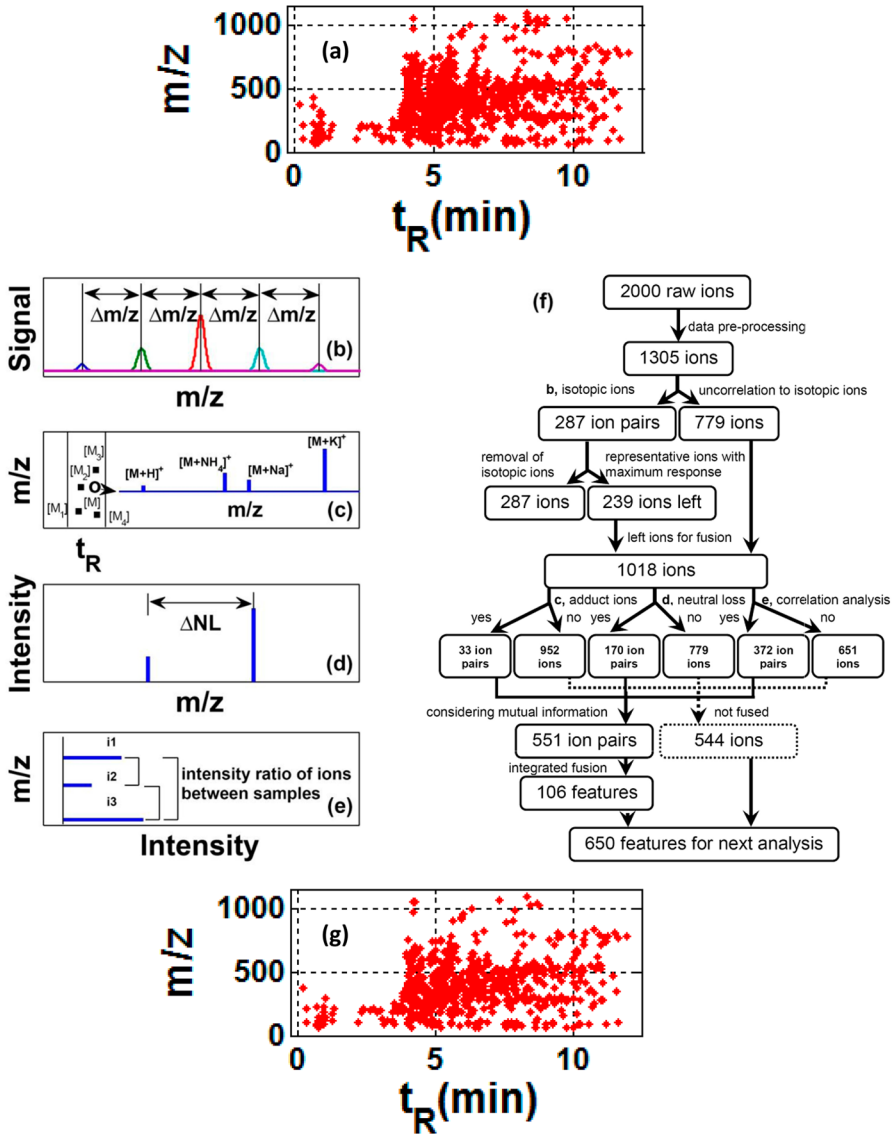


Figure 2. The principle and performance of fusion for the ions obtained from the positive ionization mode using the four relationships proposed in the text. Panels a and g show the t_R−m/z plot of the ions over the 2D plane before and after ion fusion. Panels b−e introduce the principles of ion fusion using isotopic distribution, adduct ions, neutral loss, and correlation analysis, respectively. Panel f shows the step-by-step outcomes with the help of the four relationships.

Figure 2 shows the step-by-step principle and performance of fusion for ions obtained from the ESI+ mode using all four relationships mentioned above, which were isotopic distribution, adduct ions, neutral loss, and correlation analysis. The top (2a) and bottom (2g) panels show the t_R−m/z plot over the 2D plane of the ions after data preprocessing and ion fusion, respectively. In the remaining panels, the schematic principles for ion fusion are illustrated on the left side (2b−e)), and the serial outcomes given on the right side (2f) introduce the detailed performance of each fusion step.

As shown in Figure 2a, the high complexity with a large number of ions coeluting in small retention windows leads to the difficulty of finding associated ions from the same molecule, and also indicates the high possibility of one molecule corresponding to more than one ion feature. Figure 2b simulates a Gaussian structure of the isotopic distribution with equal mass differences among the five ion patterns. According to the principles introduced in the theoretical section, 287 isotopic ions can be found, with the results given in 2f. In Figure 2c, the ion labeled with the symbol "o" is an example of an ion targeted for fusion, and the remainders are coelution ions in the predefined t_R window. A series of possible molecular weights ([M]) of the original metabolite can then be calculated for each ion with the consideration of different adduct ions, as shown in the figure, including $[M + H]^+$, $[M + NH_4]^+$, $[M + Na]^+$, $[M + K]^+$, and many others. The final magnitude of [M] was determined with the maximum probability of the [M] corresponding to all the ions in a fairly acceptable t_R shift window. The principle for ion fusion based on the relationship of neutral loss is not difficult to understand, as illustrated in Figure 2d. The mass difference of ions between theoretical computation and experimental measurement is the basis for fusing ions using this strategy. Figure 2e shows an artificial mass spectrum of an experimental peak. Ideally, the ratio of the relative intensity between each ion pair with a good S/N level, such as i1 and i2, i1 and i3, and i2 and i3, should be a constant among samples under the defined conditions, and in the analytical sequence. Thus, ion fusion can be attained if a

dx.doi.org/10.1021/ac500878x | Anal. Chem. 2014, 86, 3793−3800

Table 1. Numbers of Fused Ion Pairs Using the Four Relationships Introduced in the Text

t_R shift (min)	ionization modes	isotopic distribution	adduct ions	neutral loss	correlation analysis	fused molecules	unfused ions
0.05	ESI+	287	33	170	372	106	544
	ESI−	258	67	225	758	169	616
0.1	ESI+	292	64	212	524	103	441
	ESI−	265	66	347	792	171	507
0.5	ESI+	313	71	663	882	54	223
	ESI−	287	64	1171	892	72	213

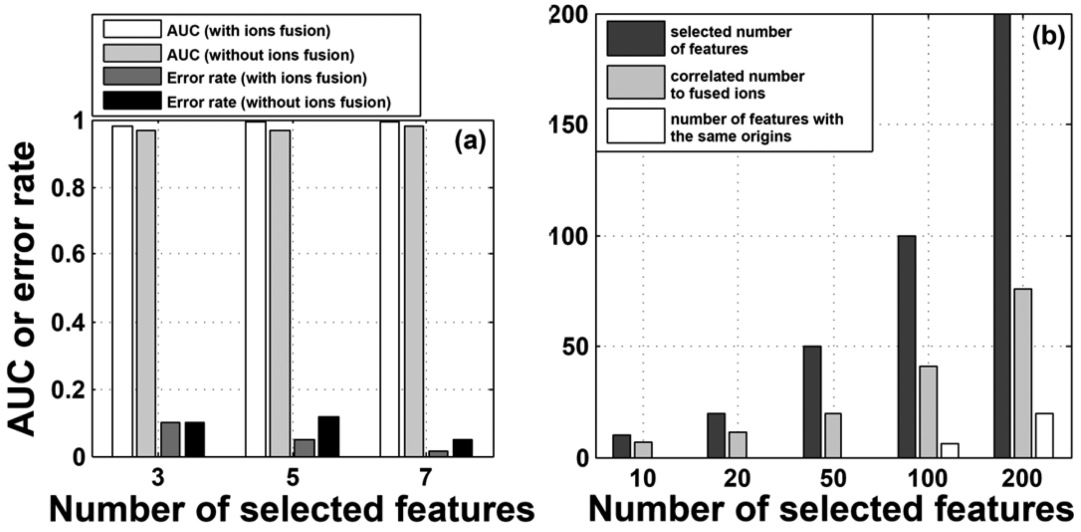


Figure 3. Comparison of the AUC and ERD results to the datasets with and without ion fusion using different numbers of features for modeling, and detailed correlation of the features found by the VIP method to the dataset without ion fusion. (a) Comparison of the AUC and ERD to the datasets with and without ion fusion using the first 3, 5, and 7 features obtained from the VIP method. (b) The results of the number of features with ions derived from a common molecule (but with other ions of the molecules not selected) and the number of features with the simultaneous selection of an ion from common molecules are both introduced to the first 10, 20, 50, 100, and 200 most important ion features.

tolerance of the correlation coefficient is defined, as mentioned above. After the ion pairs are obtained with the help of an individual strategy for fusion, the next step is to combine the ion pairs, including mutual ions. The cross finding of ions can be applied to link the ion pairs obtained from different relationships. The results corresponding to each fusion step are introduced in 2f. A total of 551 independent ion pairs were found with all four relationships introduced above, which were then fused to 106 metabolite features with consideration of the cross correlation of each ion shown in Figure 2c−e. In addition, the remaining 544 ions could not be fused using any of the four relationships. Figure 2g shows the ultimate outcomes. In total, there are 33, 170, and 372 ion pairs found that meet the requirements of adduct ion, neutral loss, and correlation analysis for fusion. Correlation analysis can be readily recognized as having the largest contribution to ion fusion. However, an extra 179 ion pairs were additionally found to be fused by using the mass correlations of adduct ions and neutral loss.

With respect to the ions obtained from the ESI− mode, the numbers of ion pairs found for fusion were 258, 67, 225, and 758, corresponding to four relationships introduced in Figure 2b−e. These ions were fused to 169 independent metabolites, and the remaining 616 ions were left free for fusion. The maximum response of the corresponding ions in each fused group was applied for data processing in the next step.

The total number of fused ion pairs was related to the parameters (t_R shift, $\Delta m/z$ and $CC_{tolerance}$). The fused details for different parameters are provided in Table 1. The determination of the t_R shift is important to maximally reduce false positive fusion. This information can be attained by inspecting the raw chromatograms of multiple samples, and setting the t_R shift to 0.5 min is unsuitable because of a rapid decrease in the number of molecules to be fused. The reason for the unsuitability is this setting positively fuses the ions with elution across a wide retention window.

Another choice is to fuse the ions obtained from the two ion modes together, which may help find mutual features between them, and likely reduce the number of fused ions. However, the experimental conditions (the composition of the mobile phase) were quite different in this study, which makes the t_R shift of the same molecule not comparable in the two ionization modes. Thus, fusing them with high accuracy is difficult.

Application of Ion Fusion in HCC to Discover More Reliable Metabolic Biomarkers. In this section, a relationship was proposed to study the effectiveness of the identifying ions with high discriminatory power to differentiate HCC from N samples using the datasets with and without ion fusion. These outcomes would further help validate the necessity for ion fusion to discover reliable biomarkers.

The identification of the ion features after fusion with significant differences to distinguish between samples was the next task. Low-level data fusion was applied to concatenate the 650 and 785 (1435 ions in total) features obtained from the ESI+ and ESI− modes, including the 106 and 169 fused ions. The significance level was set to 0.05 for the nonparametric test. Then, 508 of the 1435 ions were shown to be significantly different by the Mann−Whitney *U*-test, with 245 and 263

dx.doi.org/10.1021/ac500878x | *Anal. Chem.* 2014, 86, 3793−3800

features being found from the ion pools with and without fusion, respectively. For the dataset without fusion, 1102 ions of the 1305 and 1523 ion features obtained from the ESI+ and ESI− modes, respectively, were shown to be significantly different, and were employed as references to construct the models for comparison.

The feature selection was performed by the VIP method, and PLS-DA was used for the discrimination analysis. Simply, all samples with no specific division were applied to construct the classification model, and further evaluate the identified metabolomic features of the datasets with and without ion fusion. The processing helps the study focusing on performance evaluation of ion fusion, and maximally reduces the influence of complicated modeling.

The classification of the HCC and N samples was performed for all the samples with no specific division. Using the first 5 most important features, the results of the AUC attained 0.9944 and 0.9678 for the datasets with and without ion fusion, respectively, and the error rate for discrimination (ERD) reached 0.05 and 0.1167, respectively. The AUC and ERD outcomes were changed to 0.9798 and 0.97 and to 0.1 and 0.1, respectively, when the number of features for modeling was decreased to 3, and these outcomes were then changed to 0.9944 and 0.9822 and to 0.0167 and 0.05, respectively, when the number of features for modeling was enlarged to 7. These results are provided in Figure 3a. In terms of the principles and modeling performance introduced above, the advantages of the proposed method guarantee the uniqueness and differences of the discovered features corresponding to specific metabolites, which may be applied to interpret metabolomic functions and pathways.

However, the redundancy of the identified biomarkers unfortunately violates the original intention to discover biomarkers for disease diagnosis or metabolomics interpretation for the data with no fusion, although no poor results were generated from the classification model, as introduced above. Additionally, this redundancy unavoidably reduces the chance of other potential molecules to be selected as biomarkers. The acceptable results of AUC and ERD are just pure mathematical outcomes with no definitive metabolomics findings.

Figure 3b shows the redundancy and limitations of selected features for the dataset without ion fusion. The VIP method was applied to determine the important variables that distinguish among the samples. Only 3 of the first 10 most important features have no additional associated ions originating from the same metabolite. The remaining 7 features have at least one ion derived from common metabolites (but other ions of the molecules were not selected). Furthermore, 6 of the first 100 most important features were even simultaneously selected by the VIP method and were found with the same origins of metabolites with each other. This value was 20 among the first 200 selected features. The correlation of the first 10, 20, 50, 100, and 200 features is provided in Figure 3b for the dataset without ion fusion. These results clearly verify the importance of ion fusion. The employment of raw data for modeling without ion fusion largely burdens the computation task, and generates colinearity among features for chemometric analysis. Of course, it also unavoidably decreases the likelihood of other informative metabolite features being selected as biomarkers. Thus, ion fusion does help reach the "truth" of biomarkers for disease diagnosis and other applications.

In addition, information of fused multiple ions is helpful for improving the identification. Two typical examples are introduced below to aid the identification of unknown metabolites. The first example includes three associated ions with m/z 355.2612, 373.2715, and 391.2822 in the ESI+ mode. These ions were tentatively identified as pregnan-20-one, 17-(acetyloxy)-3-hydroxy-6-methyl-, and (3b,5b,6a)- (Metlin ID: 1089) using the multiple fragment search function of the Metlin online database (http://metlin.scripps.edu/). All three fragments had a good match with the metabolite in the database. The ions with m/z of 373.2715 and 391.2822 corresponded to $[M + H-H_2O]^+$ and $[M + H]^+$, respectively. They were fused based on the neutral loss and correlation analysis of the proposed method. However, more search hits were generated, and unavoidably influenced the unique identification if the individual fragment was applied to characterize the metabolites. For example, the metabolites including (*R*)-Butaprost, Glutamic acid (Glu), Glycine (Gly), and Tryptophan (Trp) potentially generated one of the three precise m/z values. Only a single fragment is not possible to achieve accurate identification.

The other example is to demonstrate the power of the proposed method to characterize metabolites in the ESI− mode. Three associated ions with m/z 239.0916, 195.1021, and 151.1127 were checked to know the fused possibility. Finally, this metabolite was tentatively identified as 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid (Metlin ID: 45041) with the same procedure described above. However, the number of metabolite hits was 2, 13, and 7 for the three individual ions, respectively, in the Metlin online database. This finding obviously shows the usefulness of ion fusion information for metabolite identification.

The study of metabolomics involves the discovery of small molecules with specific functions in a metabolic network to diagnose diseases or understand biological processes. The identification of metabolites is a difficult but necessary step to help achieve these goals. The complementary information of multiple fused ions can be used to improve the identification of metabolites.

Comparison of the Proposed Method with the Existing Software Platforms Distributed by the Instrument Manufacturers. The extraction of the feature ions of small molecules has been implemented in some of the commercial software platforms, such as the system of MassHunter qualitative analysis (MHQA) developed by Agilent. It employs a nonpublic algorithm for molecular feature extraction, which helps identify the associated ions, including isotopes and adducts, and molecular ions. A table is generated that includes the information introduced in the table of peak ions mentioned above. Next, a system called MassProfiler Professional is applied to align and filter the identified features. Furthermore, the outcomes are used for recursive peak integration using an ion extraction algorithm in MHQA. Although the real extraction of feature ions cannot be achieved in the present metabolomics example due to the limitation of the file format as the input (.d file only in MHQA), the performance of ion fusion was still comparable by taking into account the principles and relationships applied to fuse ion features of the same molecule.

The proposed method developed a comprehensive strategy for ion fusion, including isotopic distribution, adduct ions, neutral loss, and correction analysis of ions among samples. These relationships should exhaustively employ the information

dx.doi.org/10.1021/ac500878x | *Anal. Chem.* 2014, 86, 3793−3800

available for feature extraction. In addition to the isotopes, the three other relationships are theoretically not included in the commercial software although they greatly improve the performance of ion fusion. The combination of stringent constraints of the t_R shift, mass difference between the experimental measurement and theoretical computation, and correlation coefficient maximally avoids the generation of false positive outcomes.

CONCLUSIONS

In this work, a combinatorial strategy was proposed to fuse the associated ions by high-resolution LC−MS in metabolomics. The employment of prior knowledge included isotopic distribution, adduct ions, neutral loss, and correlation analysis of the ions with a predefined retention time shift. After the ion fusion, redundant metabolomics data are reduced and metabolite features with property of one feature for one peak are generated, it ensures that each metabolite has the same chance to be selected as a potential biomarker for the classification of different types of samples. Serum metabolomics investigation of HCC and healthy controls confirms ion fusion is helpful to discover more reliable metabolic biomarkers and enhance the chance for the accurate characterization of unknown metabolites with the help of the fused ion information.

ASSOCIATED CONTENT

Supporting Information

List of common adducts and the corresponding masses observed in ESI+ and ESI− detection, list of common neutral losses (NL) and the corresponding formula and exact masses, illustration to introduce the steps and integrated strategy of ion fusion with the meaning of each symbol provided PCA results from the ESI+ and ESI− modes and their combination after the data pretreatment, and statistical results of the correlation coefficients (CC) between each ion pair among all 60 samples (after data pretreatment). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Guowang Xu. Address: Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, 457 Zhongshan Road, Dalian 116023, China. Tel./Fax: +86-411-84379530. E-mail: xugw@dicp.ac.cn.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is financially supported by the National Basic Research Program of China (No. 2012CB518303) from the State Ministry of Science & Technology of China, State Key Science & Technology Project for Infectious Diseases (2012ZX10002011, 2012ZX10002009), the Foundations (No. 21175132 and No. 21375011), the Creative Research Group Project (No. 21321064) from the National Natural Science Foundation of China, and the "One Hundred Talent Program" of the Dalian Institute of Chemical Physics, Chinese Academy of Sciences.

REFERENCES

(1) Wagner, S.; Scholz, K.; Sieber, M.; Kellert, M.; Voelkel, W. *Anal. Chem.* **2007**, *79*, 2918−2926.
(2) Wagner, S.; Scholz, K.; Donegan, M.; Burton, L.; Wingate, J.; Volkel, W. *Anal. Chem.* **2006**, *78*, 1296−1305.
(3) Xu, Y.; Heilier, J. F.; Madalinski, G.; Genin, E.; Ezan, E.; Tabet, J. C.; Junot, C. *Anal. Chem.* **2010**, *82*, 5490−5501.
(4) Koulman, A.; Woffendin, G.; Narayana, V. K.; Welchman, H.; Crone, C.; Volmer, D. A. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1411−1418.
(5) Dunn, W. B.; Broadhurst, D.; Brown, M.; Baker, P. N.; Redman, C. W. G.; Kenny, L. C.; Kell, D. B. *J. Chromatogr. B* **2008**, *871*, 288−298.
(6) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R.; Human Serum Metabolome, H. C. *Nat. Protoc.* **2011**, *6*, 1060−1083.
(7) Sugimoto, M.; Kawakami, M.; Robert, M.; Soga, T.; Tomita, M. *Curr. Bioinform.* **2012**, *7*, 96−108.
(8) Wu, R.; Wu, Z. M.; Wang, X. H.; Yang, P. C.; Yu, D.; Zhao, C. X.; Xu, G. W.; Kang, L. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 3259−3263.
(9) Chen, J.; Zhang, X. Y.; Cao, R.; Lu, X.; Zhao, S. M.; Fekete, A.; Huang, Q.; Schmitt-Kopplin, P.; Wang, Y. S.; Xu, Z. L.; Wan, X. P.; Wu, X. H.; Zhao, N. Q.; Xu, C. J.; Xu, G. W. *J. Proteome Res.* **2011**, *10*, 2625−2632.
(10) Griffiths, W. J.; Koal, T.; Wang, Y. Q.; Kohl, M.; Enot, D. P.; Deigner, H. P. *Angew. Chem., Int. Ed.* **2010**, *49*, 5426−5445.
(11) Abollino, O.; Malandrino, M.; Giacomino, A.; Mentasti, E. *Anal. Chim. Acta* **2011**, *688*, 104−121.
(12) Werner, E.; Croixmarie, V.; Umbdenstock, T.; Ezan, E.; Chaminade, P.; Tabet, J. C.; Junot, C. *Anal. Chem.* **2008**, *80*, 4918−4932.
(13) Brown, M.; Wedge, D. C.; Goodacre, R.; Kell, D. B.; Baker, P. N.; Kenny, L. C.; Mamas, M. A.; Neyses, L.; Dunn, W. B. *Bioinformatics* **2011**, *27*, 1108−1112.
(14) Vaughan, A. A.; Dunn, W. B.; Allwood, J. W.; Wedge, D. C.; Blackhall, F. H.; Whetton, A. D.; Dive, C.; Goodacre, R. *Anal. Chem.* **2012**, *84*, 9848−9857.
(15) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.
(16) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, 5035−5039.
(17) Wei, X. L.; Sun, W. L.; Shi, X.; Koo, I.; Wang, B.; Zhang, J.; Yin, X. M.; Tang, Y. N.; Bogdanov, B.; Kim, S.; Zhou, Z. X.; McClain, C.; Zhang, X. *Anal. Chem.* **2011**, *83*, 7668−7675.
(18) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8*, 719−726.
(19) Rasche, F.; Svatoš, A.; Maddula, R. K.; Böttcher, C.; Böcker, S. *Anal. Chem.* **2011**, *83*, 1243−1251.
(20) Yin, P. Y.; Peter, A.; Franken, H.; Zhao, X. J.; Neukamm, S. S.; Rosenbaum, L.; Lucio, M.; Zell, A.; Haring, H. U.; Xu, G. W.; Lehmann, R. *Clin. Chem.* **2013**, *59*, 833−845.
(21) Bijlsma, S.; Bobeldijk, L.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. *Anal. Chem.* **2006**, *78*, 567−574.
(22) van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142−156.