analytical chemistry

Article

pubs.acs.org/ac

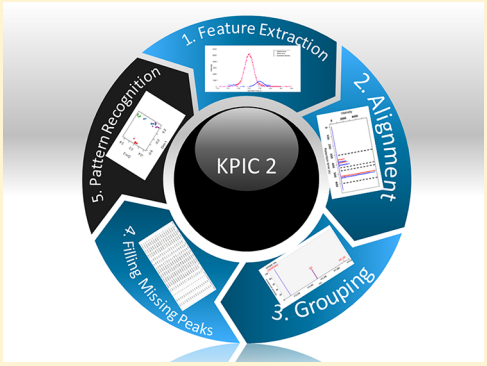
KPIC2: An Effective Framework for Mass Spectrometry-Based Metabolomics Using Pure Ion Chromatograms

Hongchao Ji, Fanjuan Zeng, Yamei Xu, Hongmei Lu,* and Zhimin Zhang*

College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China

Supporting Information

ABSTRACT: Distilling accurate quantitation information on metabolites from liquid chromatography coupled with mass spectrometry (LC-MS) data sets is crucial for further statistical analysis and biomarker identification. However, it is still challenging due to the complexity of biological systems. The concept of pure ion chromatograms (PICs) is an effective way of extracting meaningful ions, but few toolboxes provide a full processing workflow for LC-MS data sets based on PICs. In this study, an integrated framework, KPIC2, has been developed for metabolomics studies, which can detect pure ions accurately, align PICs across samples, group PICs to identify isotope and potential adducts, fill missing peaks and do multivariate pattern recognition. To evaluate its performance, MM48, metabolomics quantitation, and Soybean seeds data sets have been analyzed using KPIC2, XCMS, and MZmine2. KPIC2 can extract more true ions with fewer detecting features, have good quantification ability on a metabolomics quantitation data set, and achieve satisfactory classification on a soybean seeds data set through kernel-based OPLS-DA and random forest. It is implemented in R programming language, and the software, user guide, as well as example scripts and data sets are available as an open source package at https://github.com/hcji/KPIC2.



High-resolution mass spectrometry (MS) coupled with gas chromatography (GC) or liquid chromatography (LC) plays an important role in metabolomics.[1,2] In particular, LC-MS is a more flexible technique, which can analyze metabolites, including those that are difficult in vaporization. However, difficulties remain in processing large-scale LC-MS data sets of metabolites due to their complexity. For example, it is estimated there are over 1,000 kinds of metabolites in human serum,[3] and above 4,000 or 5,000 to 25,000 for plants.[4] Moreover, background ions and random noises make the quantification even more challenging. Therefore, it is essential to develop reliable label-free strategies to handle large-scale metabolite data sets.

Previous efforts to address these challenges in LC-MS data sets have lead to many famous publicly available tools, for example XCMS,[5,6] MZmine,[7,8] MetAlign,[9,10] OpenMS,[11,12] etc. All of these pieces of software can accept raw LC-MS data files and analyze the imported data set with chemometrics algorithms. They all provide the whole pipeline for processing LC-MS data sets. Typically, a pipeline of processing LC-MS data includes preprocessing, feature detection, alignment and grouping, normalization, and further pattern recognition.[5,13] Among them, feature detection is the fundamental procedure to extract useful information from raw data. Since an LC-MS data set can be seen as a three-way array, the common strategy is to reduce the dimension. Total ion chromatography, base peak chromatography, and extracted ion chromatography are all based on this strategy. Detecting peaks in extracted ion chromatography, which is used in XCMS, MZmine, MetAlign, and OpenMS, is the most widely used method for feature extraction. However, binning to extracted ion chromatograms (EICs) introduces additional steps such as baseline correction, smoothing, and peak deconvolution. Meanwhile, it may split a feature into adjacent bins.[5] To circumvent the drawbacks of binning, another method called centWave[14] is integrated into XCMS, which locates the region of interest (ROI) based on the relative mass differences. However, the relative mass differences of the adjacent ions in a chromatographic peak are proportional to the intensities of ions. Meanwhile, they may also vary in different mass spectrometers.[15,16] Moreover, MZmine2 implemented a 2D peak detection method named GridMass,[17] which generates a grid of equally spaced probes covering the entire chromatographic area and makes each probe explore a rectangular region around it to find a local maximum until no higher values exist within the exploring rectangle. Hence, the local maximums are defined as features.

Besides, new programs such as Finnee,[18] FeatureFinderMetabo,[19] SMART,[20] and iMet-Q[21] have also been published to improve the quality, efficiency, and/or usability in analyzing the LC-MS data set. Finnee provides various functions for visualization, and uses different strategies to handle profile and centroid MS mode. FeatureFinderMetabo detects ion traces and its isotopes at the same time to improve the

Received: April 27, 2017
Accepted: June 16, 2017
Published: June 16, 2017

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

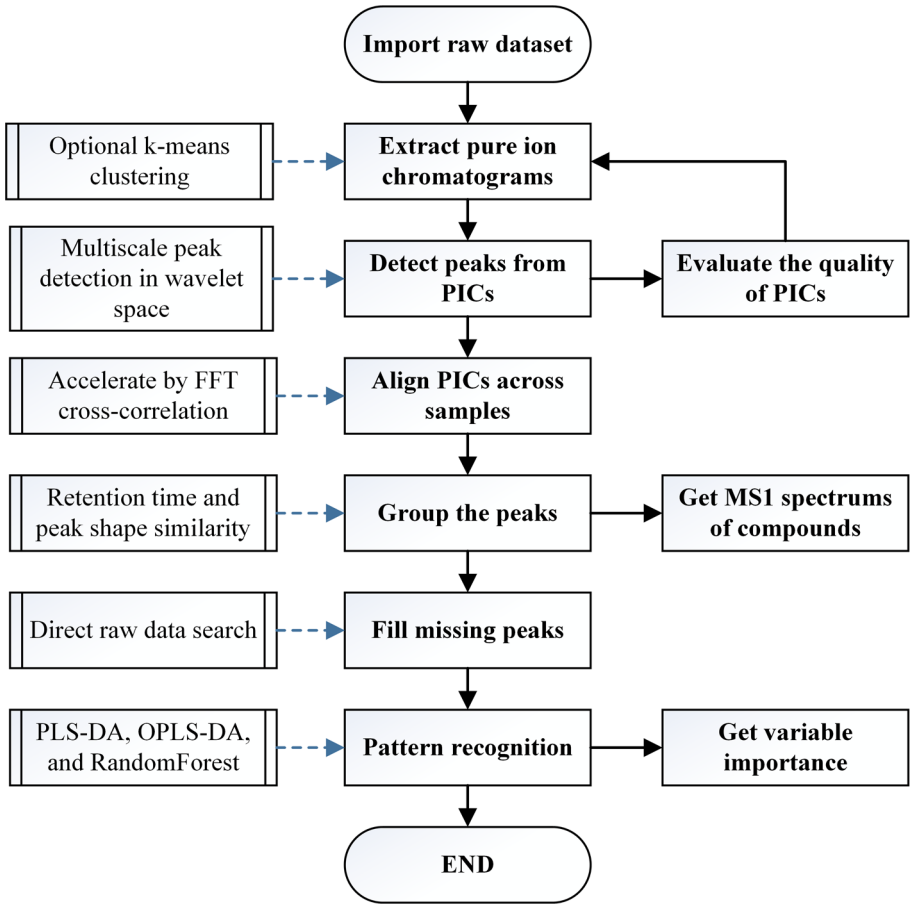


Figure 1. Flowchart of preprocessing and analyzing LC-MS data set in metabolomics with KPIC2. The raw data set can be imported into KPIC2 in mzXML or mzdata format. Extraction method based on optimal k-means clustering is used to obtain PICs, and quality evaluation of PICs can assist to choosing the parameters of the PIC extraction. The PICs can be aligned across samples by fast Fourier transfrom cross-correlation efficiently. It can also group PICs to identify isotope and potential adducts and fill missing peaks. Finally, OPLS-DA and random forest can be applied for pattern recognition and evaluating the importance of variables.

robustness and avoid the individual deisotoping step. SMART integrates the statistical analysis methods, addressing the challenges of statistical analysis in large-scale data. iMet-Q focuses on providing a user-friendly interface.

A novel concept entitled pure ion chromatogram (PIC) has been introduced recently, which is particularly suitable for the centroid high-resolution LC-MS data set. A PIC means a length of chromatogram consisting of only one response intensity each scan, and the response values are supposed to come from the detection of a single ion. There are already many feature extraction algorithms based on this concept, including TracMass,[22] massifquant,[23] TracMass2,[24] and PITracer.[16] They can take full advantage of high-resolution mass spectrometry and track ions from scan to scan. TracMass and massifquant detect pure ion traces with a Kalman filter, even though using a Kalman filter is time-consuming. On the other hand, TracMass2 and PITracer use the adjusted mass differences of the adjacent ions as standard, which are significantly faster. KPIC[25] is a PIC extraction method based on *k*-means clustering, which avoids estimating the mass difference tolerance of ions and reduces the number of split signals. However, methods such as TracMass2, PITracer, and KPIC do not make full use of information on the intensity of ions. Furthermore, some tools provide a deisotoping method (such as FeatureFinderMetabo and iMet-Q). The PIC extraction methods referred to above do not have these functions. So, external packages such as CAMERA[26] should be used.

In this study, an integrated framework called KPIC2 has been developed. It goes further than KPIC, which increases the accuracy of PIC extraction by taking intensity into consideration. Besides, KPIC2 provides the entire pipeline for metabolites quantitation and pattern recognition. Testing with several metabolomics data sets, KPIC2 shows satisfactory results of feature detection, alignment, grouping, quantitation, and pattern recognition. KPIC2 is implemented in R programming language, which is available as an open source package at https://github.com/hcji/KPIC2.

METHOD AND THEORY

The Concept of Pure Ion. The main challenging of LC-MS features extraction is to separate ions of analyte from any ions not meaningful, such as background ions and random noise. Extracting "pure ion" is an effective way to achieve this. The conception of pure ion is the ions that originate from the same analyte. The most reliable information to locate pure ions is that they exist with continuous scans, similar mass-to-charge ratios, and continuous changing intensities in the raw data. Thus, the common methods track ion from scan to scan and connect data points with similar m/z values. In this way, each ion trace usually contains only one kind of ion from the same analyte.

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

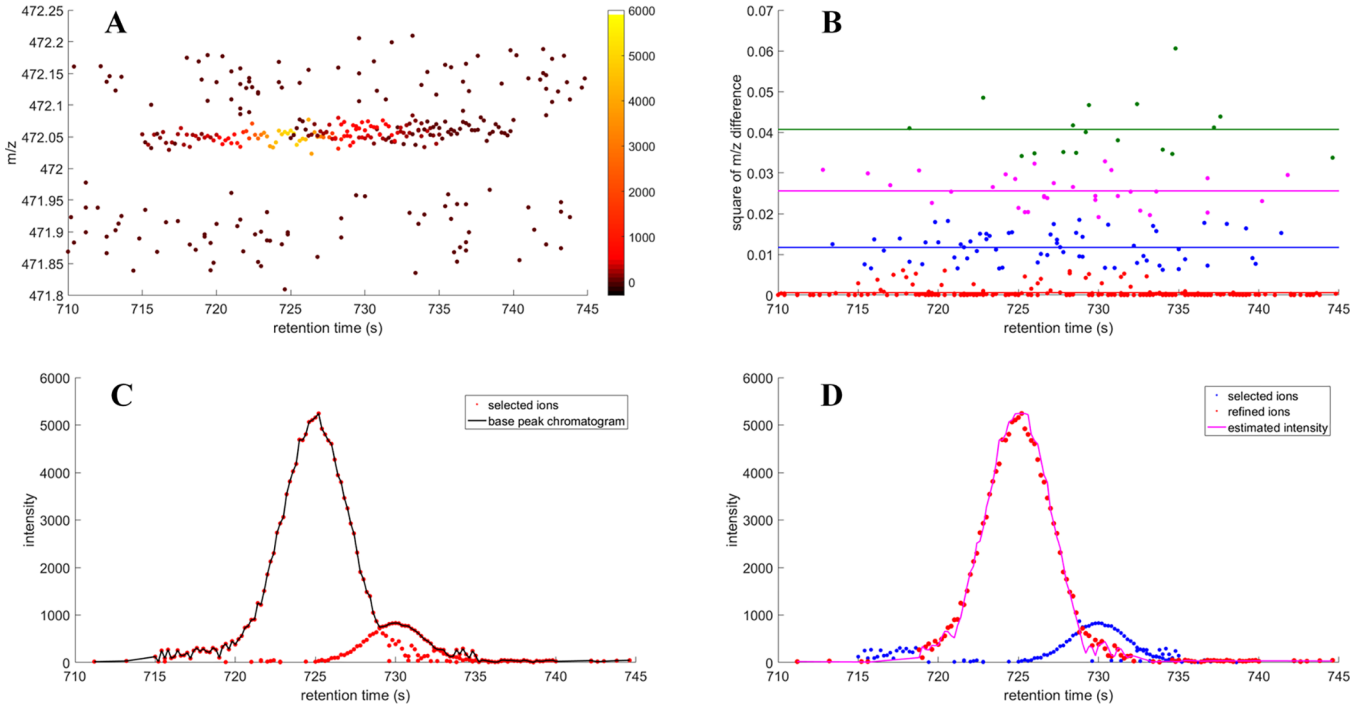


Figure 2. Schematic illustration of the pure ion extraction. (A) The ROI is located based on the landmark ion. Pure ions have very similar m/z values as the landmark ion. However, there are also interferential ions and noise included in the ROI. (B) The square of the differences between the m/z values of the ions in the ROI and the m/z value of the landmark are calculated. Optimal k-means clustering has been applied to cluster these ions into four clusters. Then the ions in the cluster with the mean closest to zero are selected. (C) The intensities of the selected ions of step B and the BPC consisting of them. Sometimes, multiple compounds have similar retention times and similar m/z values, so the BPC may consist of more than one peak. (D) After refining by exponential smoothing, the pure ions can be selected free from the noise, background ions, and interferential ions.

The difference of methods is how to define "similar". The relative mass difference tolerances can be affected by many factors, such as the intensities of the ions, resolution of instrument, ambient temperature, etc.[27,28] The existing methods usually need the user to give such a tolerance value or estimate a constant value. KPIC2 gives up the arbitrary manner and uses the clustering tendency of m/z values of pure ions, which is a more flexible way. The tolerance range is not fixed and depends on the ions in the cluster.

The whole Framework of KPIC2 consists of the following modules: extraction of pure ion chromatograms, peak detection, alignment, grouping, and missing peak filling. Finally, a peak table consisting of retention time, accurate mass, and maximum intensities can be constructed easily for further statistical analysis and pattern recognition. The whole workflow is summarized in Figure 1. In this section, the theory of each procedure will be described in detail.

Extraction of Pure Ion Chromatograms. PIC extraction in KPIC2 works in an iterative procedure. The raw LC-MS data set is converted into a three-column matrix containing all detected ions. These columns collect the values of retention time (rt), m/z, and intensity, respectively. The ion with the highest intensity is chosen as a landmark to start the iteration. Then the region of interest (ROI) of the landmark is determined by the following steps: the user gives a flexible m/z range (ppm of tolerance), and data points in this range are collected into the ROI. Then, the retention time of the ROI is extended toward both sides of the landmark until there are no data points within the given m/z range.

After determination of the ROI, it has to be known which ions have the same origination as the landmark ion. For the high-resolution mass spectrum, m/z is the most reliable criterion. Therefore, we calculate the values of the square of the m/z difference between ions in the ROI and the landmark, and perform an optimal k-means clustering algorithm (ckmeans.1d.dp[27]) on them. In this way, the cluster with the mean nearest to zero is most likely to contain the response of the same ion as the landmark. and random noise signals are usually assigned into other clusters, and the number of clusters is optimized automatically by ckmeans.1d.dp.

Ions selected in previous steps are converted into a chromatogram. In most cases, only one ion is selected in a single scan. For nonideal scans, the intensity of the chromatogram is represented by the maximum intensity of the ions. Hence, the base peak chromatograms (BPCs) on the ROI are obtained. However, the BPCs are not equal to the PIC, because interferential ions with very similar retention time and m/z may still be included.

Then exponential smoothing[28] is used to estimate the intensity of the pure ion for each scan. The details are described in the Supporting Information (Figure S1). For the scans with more than one ion selected, the one with the most similar intensity to the estimated value is selected. The ultimately selected ions are combined together to construct the PIC. The schematic is shown in Figure 2.

Evaluation of PICs. The quality of the PICs shape is evaluated by three common criteria, which are Gaussian similarity, sharpness, and signal-to-noise ratio (SNR).[29] These criteria can assist the user to choose optimal parameters for PIC extraction. The coefficient of determination[30] (R-square) of Gaussian fitting is used for measuring the goodness of PIC fitting in the Gaussian curve, which is defined as eq 1, where $[y_1, ..., y_n]$ are the intensities of the PIC and each is associated with a modeled value $\hat{y}_1, ..., \hat{y}_n$. The sharpness is defined by eq 2,

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

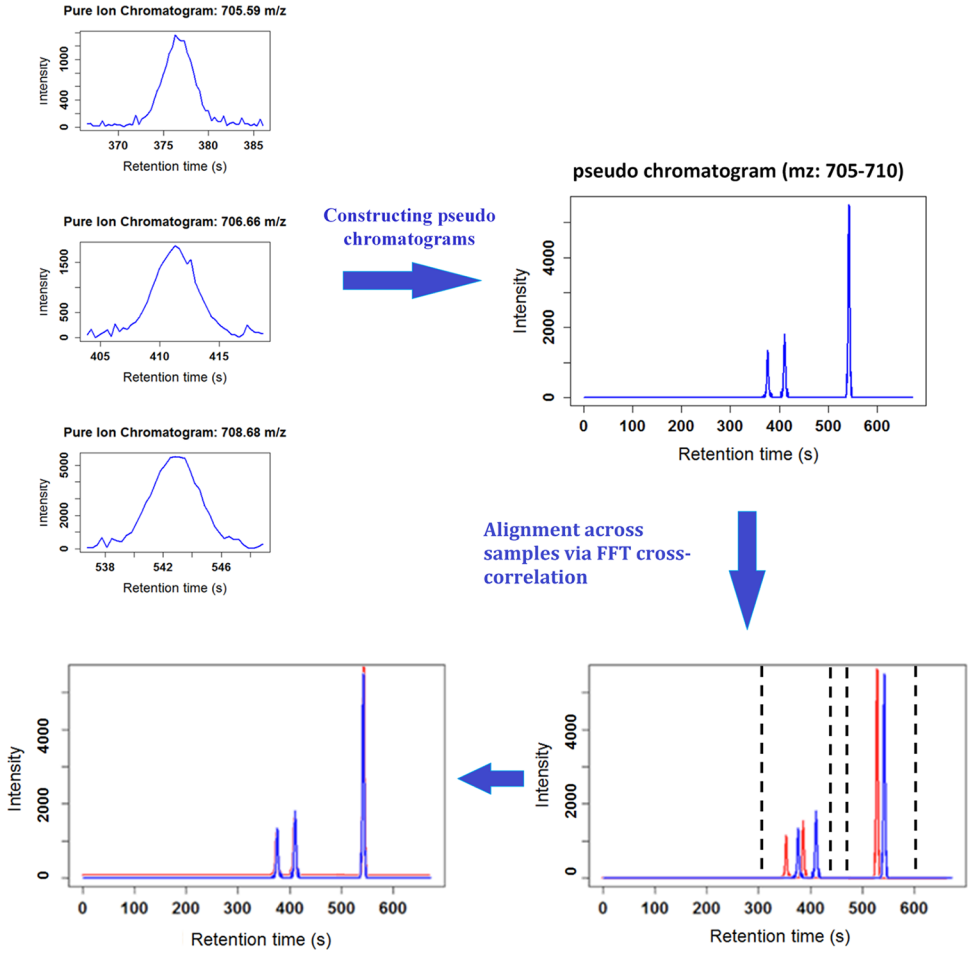


Figure 3. Schematic illustration of the alignment of PICs via PAFFT. The m/z window slides to 705−710, and three PICs of a sample are in the window. Thus, the three PICs are combined into a pseudo-chromatogram. Other samples follow the same rule, and there is one pseudo-chromatogram for each sample. Then the pseudo-chromatograms are aligned toward the reference via the PAFFT algorithm. After alignment, the pseudo-chromatograms are more uniform, which indicates the PICs are successfully matched.

where n is the total data point number and p is the peak apex index. The SNR is obtained by peak detection in wavelet space. Higher values of R-square, sharpness, and SNR indicate higher quality of PICs. Moreover, the m/z standard deviation of each PIC is also calculated. Smaller standard deviation means the points constructing the PIC are more compact in the m/z dimension, which also indicates the PIC has higher quality.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (1)$$

$$\text{sharpness} = \sum_{i=2}^{p} \frac{y_i - y_{i-1}}{y_{i-1}} + \sum_{i=p}^{n-1} \frac{y_i - y_{i+1}}{y_{i+1}} \quad (2)$$

Peak Detection. Multiscale peak detection (MSPD[31]) has been used to detect the peaks in pure ion chromatograms. MSPD transforms signals into wavelet space via a continuous wavelet transform (CWT) with a Mexican hat wavelet and determines peak location based on the information given by ridges, valleys, and zero-crossings in wavelet space. It is better than the previous peak detection algorithm MassSpecWavelet,[32] which only used the ridges information in wavelet space. Furthermore, the peak width is also estimated by enhanced SNR derivative calculation.[33]

Alignment. KPIC2 takes full advantage of the profile-based method for alignment. The alignment algorithm first constructs a moving window sliding across the m/z axis, which ranges from the minimum m/z value of the extracted PIC to the maximum one. In LC-MS analysis, deviation of the m/z dimension is much smaller than the RT dimension, and PICs of various samples with m/z values in the window tend to be one-to-one corresponding. Hence, they are aligned with the following procedures: (1) The PICs of each sample in the window are combined into a pseudochromatogram. (2) The PAFFT algorithm[34,35] (Peak alignment by fast Fourier transform) is used to align the pseudochromatograms across samples. (3) The retention times of relevant PICs are corrected based on the estimated deviation profile by PAFFT. The schematic is shown in Figure 3. To avoid splitting some corresponding features, the window moves half the width of itself (i.e., 100.0−100.5, 100.25−100.75, etc.). With such strategies, KPIC2 does not count on well-behaved (samples basically have only one peak in an m/z range[5]) corresponding features.

Grouping. This procedure has two purposes: (1) Group PICs across samples based on the retention time after alignment; (2) Group features belonging to the same substance. The first step works with an iteration procedure. Each iteration selects the most intense feature from the feature table not yet

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

assigned to a group ID as reference. The relative distance between other features included in the specific RT and m/z tolerance window and the reference are calculated via eq 3.

$$RT\ distance = \frac{RT\ difference}{RT\ tolerance}$$
$$MZ\ distance = \frac{MZ\ difference}{MZ\ tolerance} \quad (3)$$

Then the features are clustered by the HDBSCAN[36,37] method. Features in the same cluster as the reference are assigned the same group ID, and they are excluded when the next iteration runs. If more than one feature of a sample is in the cluster, the one with intensity most similar to the reference is selected to assign the group ID. The iteration runs until all of the features have group IDs.

Typically, each group represents one substance. However, this is not always true for two reasons: (1) Tailed peaks caused by the centroid step of the TOF mass spectrometer; (2) Isotopic ions and the adducts. PICs caused by these two conditions are less meaningful for statistical analysis. Therefore, we use similarity of PIC of feature group to determine which clusters represent them. The similarity estimation is analogous to the work of Erny et al.[38] It is evaluated by calculating the Pearson correlation coefficient between the most intense PIC in each cluster. The correlation between two profiles X and Y is defined as

$$p_{x,y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (4)$$

where cov (X, Y) is the covariance of X and Y, and σ_X, σ_Y are the variances of X and Y, respectively. The feature groups with calculated correlation over a given threshold are removed from the results.

Filling Missing Peaks. After grouping, there will always be peak groups that do not include peaks from every sample. A missing peak does not mean that the peak does not exist. It may be caused by undetection or misalignment. The missing peak filling includes two steps. First, it extends the tolerance of both m/z and RT dimensions, and researches from the peak list of each sample. Second, if there are still missing peaks, it allows feature searching from the raw data directly. Since the feature location in both the retention time and the m/z dimension is roughly limited by the detected features of most samples, we detect the missing peaks from the BPCs in the limited region of the remaining samples.

Pattern Recognition. Aiming to select important metabolites between classes, supervised classification and variable selection methods are commonly used. Before pattern recognition, KPIC2 provides normalization and scaling algorithms for pretreatment. PLS-DA[39,40] (Partial Least Squares - Discriminant Analysis), OPLS-DA[41,42] (Orthogonal treated PLS-DA), Kernel-based OPLS-DA[43,44] and random forest[45,46] methods are implemented in KPIC2 for pattern recognition and selecting biomarkers.

EVALUATION DATA SETS

Simulated Data Set. This data set is based on the real experimentally identified metabolites in the study by Giavalisco et al.[47] We add random noise and adjust the noise level, SNR, and m/z deviation for observing the effect of these factors on the methods. The construction details are in the Supporting Information (Section 3).

Mixed Compound Data Set. This data set is a mixture of 48 known compounds, which is spiked in different concentrations (20, 5, 1, 0.2 μM) into methanolic extracts of *Arabidopsis thaliana* leaves, and analyzed by UPLC-QTOF-MS. The 20 μM compound mixed solution is also detected individually to determine the retention time of each compound. This data set is provided by Neumann et al.[26] as a publicly available data set from the Metabolights repository (with Web site of http://www.ebi.ac.uk/metabolights, and identifier MTBLS188), and the experimental protocol can also be found at the Web site.

Quantification Data Set. This data set is derived from a reasonable concentration gradient of a mixed solution of five standard compounds together with human plasma detected by a Shimadzu ultrahigh-performance UPLC system coupled to an ion trap−time-of-flight (IT-TOF) mass spectrometer. The data set is available under request. The experimental details are as follows:

Five LC-MS-grade standard compounds (phenylalanine, tryptophan, hippuric acid, niacin, and methionine) were purchase from Sigma-Aldrich. Each standard was prepared at a concentration of 1 mg/mL aqueous solution. Then they were mixed and diluted to standard serial solutions at 1 μg/mL, 2 μg/mL, ..., 10 μg/mL. Plasma samples were the QC samples of the male infertility study.[48] 400 μL of methanol was added to 100 μL of plasma sample. Then, the mixture was centrifuged at 16000 rpm for 15 min at 4 °C for deproteinization. The supernatant was evaporated to dryness under N_2 gas. The dried samples were reconstituted in 100 μL of the different experimental standard solutions.

Sample analysis was performed using a Shimadzu LC/IT-TOF-MS, which is equipped with an electronic spray ion source and operated in positive mode. Chromatographic separation was achieved on a Waters ACQUITY UPLC HSS C18 column (100 mm × 2.1 mm i.d., 1.7 μm.) at 40 °C. Elution buffer A was water containing 0.1% (v/v) formic acid, and elution buffer B was acetonitrile containing 0.1% (v/v) formic acid. A binary gradient elution was under the following gradient program: 5% B increased to 40% B in 6 min and to 85% B in 14 min, ramped to 100% B in 3 min and held for 7 min, then decreased to 5% in 1 min, and finally maintained at 5% B for 5 min.

Soybean Seeds Metabolomics Data Set. This is a public untargeted metabolomics LC-MS data set of four near-isogenic soybean seed extracts. The four near-isogenic lines include single mutants (SM3, SM19) and double mutants (DM) of two MRP genes and the wild-type (WT). There were significant metabolite differences in the four classes. This study is reported by Jervis et al.,[49] and the experiment details are described in the reference.

RESULTS AND DISCUSSION

In this section, we assess the performance of KPIC2 from different perspectives. First, the results obtained from a mixed compound data set are presented. The ability of extracting PICs of features is tested by the recall rate of standard compounds. Meanwhile, the isotopic group results of several representative substances are listed compared with the theoretical isotopic pattern. Second, the results obtained from the quantification data set are presented. We examine the linearity of the response observed in spike-in experiments to measure the quantification quality of KPIC2. Then, a typical metabolomics study is

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

Table 1. Feature Extraction Results of Mixed Compound Data Set Using KPIC2, XCMS, and MZmine2[a]

Methods	Parameters	Average features per sample	CPU time (min/sample)	Concentration				
				20 μM	20 μM + leaf	5 μM + leaf	1 μM + leaf	0.2 μM + leaf
KPIC2	range of ROI = 0.05	6174	0.93	81.25%	81.25%	77.08%	72.91%	64.58%
	range of ROI = 0.10	7276	0.93	**83.33%**	83.33%	79.17%	77.08%	**70.83%**
	range of ROI = 0.15	7739	0.95	**83.33%**	83.33%	**83.33%**	**79.17%**	**70.83%**
XCMS	ppm = 20	6294	0.11	77.08%	**91.67%**	72.92%	62.50%	56.25%
	ppm = 30	7625	0.14	77.98%	89.58%	79.17%	68.75%	50.00%
	ppm = 50	8467	0.17	**83.33%**	**91.67%**	79.17%	62.50%	56.25%
MZmine2	m/z tolerance = 0.05	7816	0.12	81.25%	87.50%	75.00%	68.75%	62.50%
	m/z tolerance = 0.10	7670	0.09	81.25%	89.58%	75.00%	68.75%	60.40%
	m/z tolerance = 0.20	5948	0.08	81.25%	85.40%	66.67%	62.50%	58.33%

[a]All of the methods are run on personal computer with Intel i7−3930K CPU and 32G RAM.

Table 2. Isotope Cluster Detection Exemplified for Seven Substances of the MM48 Data Set

Substance Name	Formula	Theoretical values		Detected values	
		M/Z	Intensity	M/Z	Intensity
Laudanosin	$C_{21}H_{27}NO_4$	358.2018	100.0000	358.2027	100.0000
		359.2052	22.7130	359.2040	23.5013
		360.2085	2.4566	360.2085	2.8413
		361.2094	0.1867	361.2027	0.2884
S,R-Noscapine	$C_{22}H_{23}NO_7$	414.1553	100.0000	414.1582	100.0000
		415.1586	23.7946	415.1604	26.6999
		416.1595	4.1407	416.1637	3.6170
		417.1629	0.3423	417.1653	0.4177
Reserpine	$C_{33}H_{40}N_2O_9$	609.2812	100.0000	609.2846	100.0000
		610.2846	35.6919	610.2878	39.0557
		611.2855	8.0260	611.2886	7.6305
		612.2888	1.3504	612.2917	1.0296
Chelidonine	$C_{20}H_{19}NO_5$	354.1341	100.0000	354.1367	100.0000
		355.1375	21.6315	355.1341	21.8165
		356.1409	2.2226	356.1439	2.6882
		357.1417	0.2223	357.1519	0.2776
Bicuculline	$C_{20}H_{17}NO_6$	368.1134	100.0000	368.1141	100.0000
		369.1168	21.6315	369.1168	21.8740
		370.1201	2.2226	370.1206	2.8710
		371.1210	0.2667	371.1230	0.3463
Rotenone	$C_{23}H_{22}O_6$	395.1495	100.0000	395.1499	100.0000
		396.1528	24.8762	396.1521	23.6494
		397.1562	2.9596	397.1541	3.1953
		398.1571	0.3067	398.1545	0.3999
Biochanin A	$C_{16}H_{12}O_5$	285.0763	100.0000	285.0789	100.0000
		286.0797	17.3052	286.0814	16.1506
		287.0830	2.4313	287.0743	2.0722
		288.0839	0.2487	288.0823	0.2346

processed with the soybean seeds metabolomics data set. It can be tested whether the metabolomics difference can be found via the framework of KPIC2. Finally, the same data sets are processed via the state-of-the-art software XCMS and MZmine2. The results were compared with that of KPIC2.

Feature Detection. Since the ground-truth of extracts of *Arabidopsis thaliana* leaves is unknown, the recall rate of M+H ion features of the known 48 compounds is used as evaluation criterion. The results of different parameter values related to m/z tolerance are listed in Table 1, while other parameters are listed in Table S1. Unsurprisingly, a higher concentration of standard compounds leads to higher recall rates. The feature detection results are robust when the concentration is over 1 μM, and the recall rates are around 80%, near the value obtained from pure solution. When the concentration decreases to 0.2 μM, the recall rates drop significantly to lower values.

Isotopic Features Identification. To evaluate whether KPIC2 can group the isotopic features with the main features, we compared the grouping results of 20 μM compound mixed solution data with the theoretical isotopic pattern. Table 2 shows the results of seven compounds. We provide the substance name, the formula, the mass-charge-ratio of the monoisotopic peak and the first three isotope peaks, and the relative peak intensity (normalized to 100). The theoretical relative intensities of isotope peaks are calculated via the *rcdk* package (http://cran.fhcrc.org/web/packages/rcdk). It can be seen that the relative intensities of detected isotopic features are

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

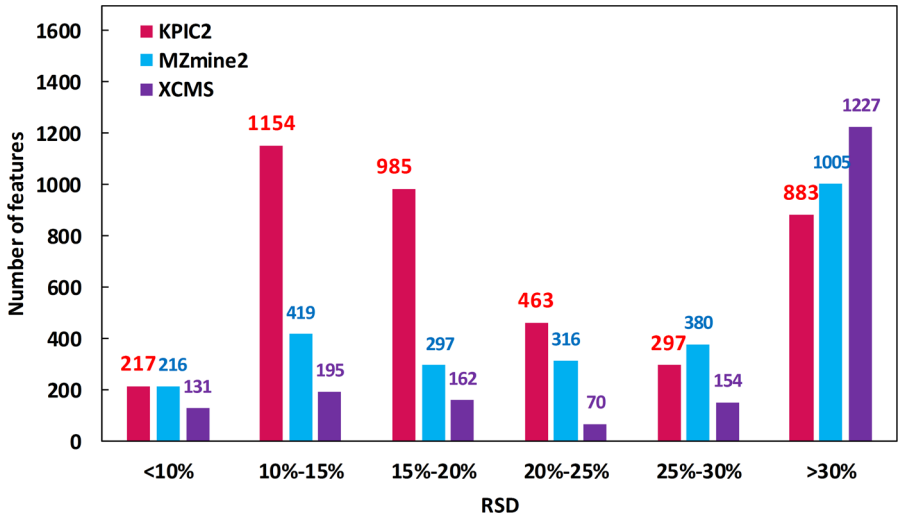


Figure 4. Number of features in RSDs in each range. All of the features are detected in over 50% samples via each method. The missing peaks are filled with the function provided by each method.

Table 3. Soybean Seeds Classification Result of Random Forest

			DM	SM19	SM3	WT	class error	OOB error
Positive mode	KPIC2	DM	9	0	0	0	0	0.00%
		SM19	0	9	0	0	0	
		SM3	0	0	9	0	0	
		WT	0	0	0	9	0	
	XCMS/CAMERA	DM	9	0	0	0	0	5.56%
		SM19	0	8	0	1	0.11	
		SM3	0	0	8	1	0.11	
		WT	0	0	0	9	0	
Negative mode	KPIC2	DM	9	0	0	0	0	0.00%
		SM19	0	9	0	0	0	
		SM3	0	0	9	0	0	
		WT	0	0	0	9	0	
	XCMS/CAMERA	DM	9	0	0	0	0	2.78%
		SM19	0	8	0	1	0.11	
		SM3	0	0	9	0	0	
		WT	0	0	0	9	0	

very similar to the theoretical values, which means KPIC2 can extract isotopic features of compounds accurately.

Quantification Linearity. Based on the quantification data set which contains a spike-in series in a complex plasma background, we can assess the linearity of the standard compounds. The data set is processed via feature detection, alignment, and grouping procedures. The detected intensities of the standard compounds features are listed in Table S2. The relationship between the analytical concentrations and their mean corresponding feature intensities is shown in Figure S2. The correlation coefficient of each compound is over 0.98, which reveals a good linear relationship. The intercept of phenylalanine and tryptophan is not zero, which is because the two substances exist in plasma at a significant level. KPIC2 is also tested by another data set (MTBLS 234) which is obtained from a wider range of concentrations of analytes. A good linear relationship is also obtained (Table S3 and Figure S3). The results indicate that KPIC2 can achieve satisfactory quantification of analytes.

Reproducibility. Apart from the standard compounds, the algorithm also detected around 7,000 features belonging to plasma per sample of the quantification data set, and 4000 of them exist in over 50% samples. Since the human plasma samples are reconstructed by QC samples, the substances of each sample are supposed to be the same. Thus, real features of metabolites of plasma are likely to have very similar peak intensities, while peak intensities of random features (e.g., noise, contaminants) may vary tempestuously. This fact can enable us to evaluate the reliability of KPIC2 of extracting features from a complex data set. Therefore, the relative standard deviation (RSD) of peak intensities across a sample is calculated. Figure 4 shows the number of features in each RSD range. Over 60% features are with RSDs less than 20%, which means more than 2500 features detected from the plasma samples have satisfactory reproducibility.

Pattern Recognition. Finding the discriminative metabolites of a different class is a common purpose of a metabolomics study. Hence, the soybean seeds metabolomics data set is used for such a typical metabolomics study in order to test whether the difference of the metabolites can be found after being processed by KPIC2. The data set is processed via feature detection, alignment, grouping, and peak filling steps, and the parameters are optimized, respectively. The result is normalized by total peak area and autoscaling. Then, two multivariate

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

statistical methods, kernel-based OPLS-DA and random forest, are taken as representations and applied to analyze the peak list matrix.

The random forest is computed with 1000 trees and 7 predictors for each node. The classification result is shown in Table 3, and the multidimensional scaling (MDS) plot of the proximity matrix of random forest models is shown in Figure 5.

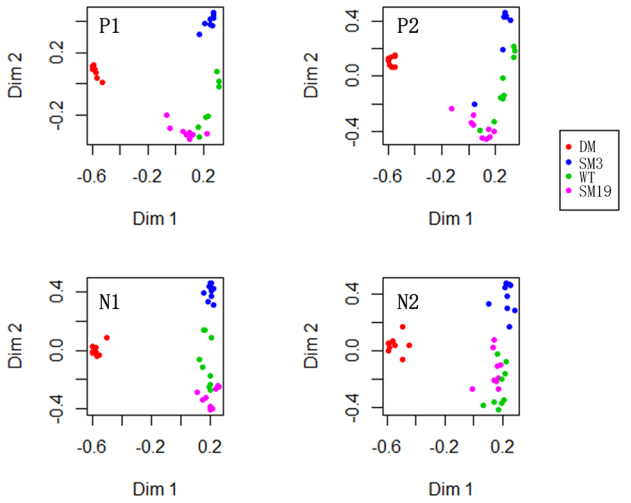


Figure 5. MDS plot of proximity matrix from random forest of the soybean seeds metabolomics data set. Each color indicates a kind of soybean seeds. P1 and N1 are the MDS plots based on the KPIC2 result of the positive and negative modes, respectively. P2 and N2 are the MDS plots of XCMS.

With features extracted by KPIC2, the random forest model achieves 100% success of classification, while XCMS misclassified one sample in both positive and negative mode. From the first two dimensions multidimensional scaling (MDS) plot based on the proximity matrix of KPIC2, the discrimination between classes is more obvious. However, in the MDS plot based on the proximity matrix of XCMS, some samples of SM3 and SM19 are mixed with the WT class, which is conformed to the classification confusion matrix shown in Table 3.

The number of the latent variable (LV), the orthogonal LV (OLV), and the kernel parameters of kernel-based OPLS-DA are optimized via cross-validation. The best parameters are used for building the model. The score plots are shown in Figure S4, respectively. It is obvious that samples of four near-isogenic lines can be separated. Then permutation test is used to compare the models. The result is listed in Table S4, where R^2 is the variance in the measurement matrix explained, and ACC is the average classification accuracy rate of classes in cross-validation. From the result of permutation test, the obtained models should be significantly better than any other random classification,[50,51] and the higher accuracy rate indicates the higher discriminating power of the extracted data and models. From the criterion and the result referenced above, the obtained classification models are reliable, which means the difference of metabolites can be found after treatment via KPIC2.

Comparison to Related Methods. XCMS/CAMERA and MZmine2 are the most popular frameworks in the field. Thus, the same data sets are also processed by these two tools.

The results of the feature detection comparison of the mixed compound data set are summarized in Table 1. One can see the total features detected by the three methods are similar, while KPIC2 even detected fewer features. However, more true features of standard compounds are extracted by KPIC2. This is more obvious when the standard compounds are at a low

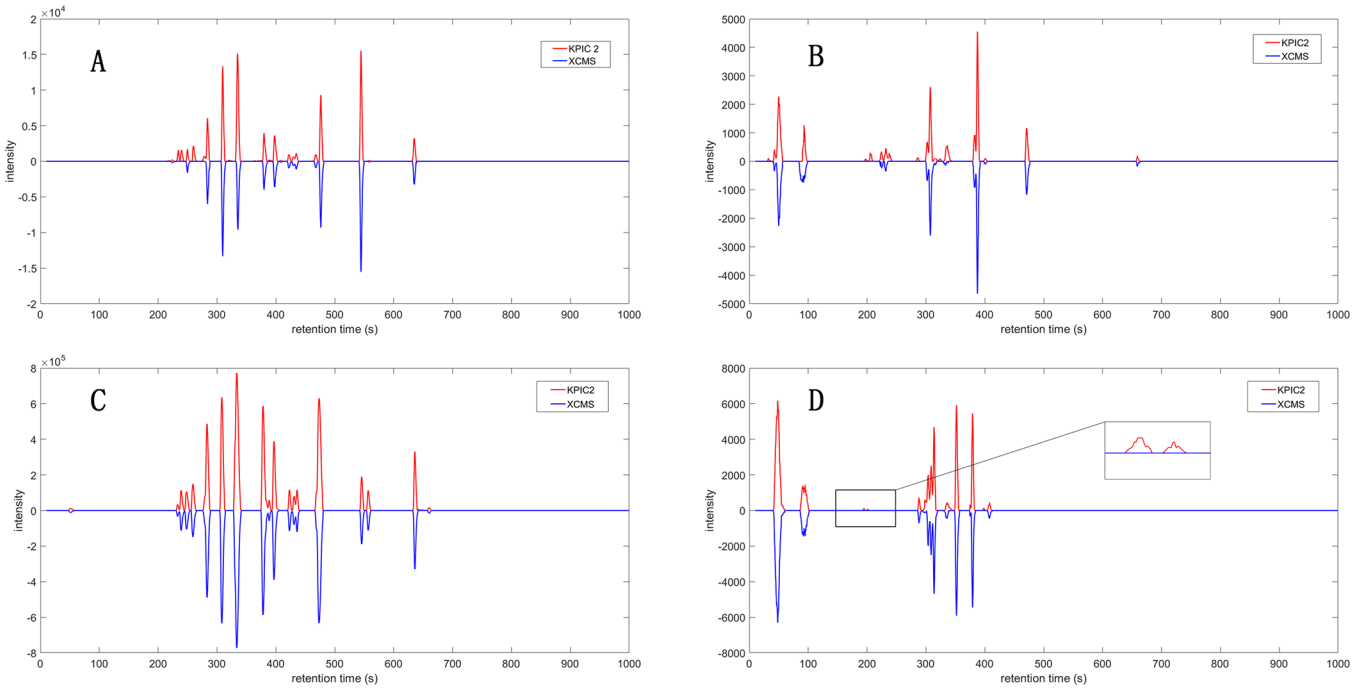


Figure 6. Additive PIC/EIC of the selected compounds. A and B are the additive PIC/EIC of the 0.2 μM solution, and A contains features with intensity over 5000, while B contains features with intensity less than 5000; C and D are the additive PIC/EIC of the 5 μM solution, and C contains features with intensity over 10,000, while D contains features less than 10,000. One can see KPIC2 obtains more low intensity PIC features than XCMS, and the peak profiles maintain high quality.

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

concentration. To make it more intuitive, Figure 6 shows the additive XICs of the standard compounds, which means add up the EICs extracted by XCMS or PICs extracted by KPIC2 into one chromatogram. It is clearly seen that KPIC2 extracted more low intensity true features than XCMS. The superiority of KPIC2 in detecting low intensity features can be explained by the fact that when clustering pure ions, noise signals can be excluded and not be considered, so the true features are not covered by noise. The only unreasonable thing is that the recall rate for the 20 μM solution and leaf of XCMS and MZmine2 is even higher than the recall rate of pure mixed solution, which may result from a false positive condition. The shortage of KPIC2 lies in the speed. Since MZmine2 is written in the Java programming language, and XCMS is written in a hybrid of C and R languages, KPIC2, written in pure R language, is not as fast as them.

The mixed solutions of standard compounds together with human plasma data set are processed by XCMS/CAMERA and MZmine2, too. The parameters are optimized based on the obtained quantification linearity of standard compounds, which means, with the selected parameters, all of the methods achieve their best quantification linearity of the standard compounds. The parameters are listed in Table S5. The filled peak lists of features reproducing in over 50% of the samples of each method are summarized in Table S6. The RSD of each feature is also calculated. As is shown in Figure 4, the ratios of detected features of KPIC2 and MZmine2 in each RSD range are similar and most of the RSDs are lower than 30%, which means they both have good reproducibility in their detected features. At the same time, KPIC2 detects more features with small RSD values, which indicates that KPIC2 exhibits better sensitivity while maintaining the reproducibility of the exclusive features. Meanwhile the ratio of features with RSD over 30% detected by KPIC2 is less than XCMS and MZmine2, which indicates better reproducibility in this data set.

Robustness. It is known interference factors of data sets, such as noise level, background ions, and m/z deviations, may be varied; thus, every method has several parameters to be optimized in order to suit different conditions. Unfortunately, it is not easy to know the best parameters. So a broad optional range of parameter selection suited for a data set is needed. Therefore, a simulated plants metabolites data set is designed. We can observe the performance of methods when the interference factors of a data set and the parameters of a method vary. The construction of the data set and the results and discussion are in the Supporting Information (Figures S5, S6, and S7).

CONCLUSION

In this study, a novel framework, KPIC2, has been developed for analyzing a high-resolution LC-MS data set. It can extract pure ion chromatograms from raw data, align them across samples based on the profiles, group PICs to identify isotope and potential adduct PICs, and fill missing peaks. Moreover, KPIC2 provides quantitative workflow, novel pattern recognition methods and variable importance for identifying biomarkers. One can see from the result of the MM48 data set that KPIC2 has superiority in detecting low concentration compounds and can extract truer ion features with fewer detected features. The features of KPIC2 have good ability on the quantification data set. The pattern recognition methods have been implemented in KPIC2 for further pattern recognition and biomarkers identification, and the results of the soybean data set show that satisfactory classification can be achieved and variable importance of PLS-DA, OPLS-DA, and random forest can be used to screen biomarkers. KPIC2 is an effective framework for integrated analysis of metabolomics data sets equipped with the concept of pure ion chromatograms, which can improve the accuracy of quantification, classification, and biomarkers identification.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b01547.

Construction of the data set and the results, result of the intensity estimation, standard compound concentrations versus mean feature intensities of quantification data set and MTBLS 234 data set, score plots, F-score versus noise level plots, F-score versus m/z tolerance parameter plots, parameters used in processing the mixed compound data set, peak intensities, permutation testing result, and discussion parameters used in processing the quantification data set (PDF)

Filled peak lists of features reproducing in over 50% of the samples of each method (ZIP)

AUTHOR INFORMATION

Corresponding Authors

*E-mail address: hongmeilu@csu.edu.cn.

*E-mail address: zhangzhimin@csu.edu.cn.

ORCID

Hongchao Ji: 0000-0002-7364-0741

Hongmei Lu: 0000-0002-4686-4491

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is financially supported by the National Natural Science Foundation of China (Grant nos. 21375151, 21305163, and 21675174) and Fundamental Research Funds for the Central Universities of Central South University (No. 2016zzts247).

REFERENCES

(1) Zhang, T.; Watson, D. G. *Analyst* **2015**, *140*, 2907−2915.

(2) Hounoum, B. M.; Blasco, H.; Emond, P.; Mavel, S. *TrAC, Trends Anal. Chem.* **2016**, *75*, 118−128.

(3) Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; Young, N.; Xia, J.; Knox, C.; Dong, E.; Huang, P.; Hollander, Z.; Pedersen, T. L.; Smith, S. R.; Bamforth, F.; Greiner, R.; McManus, B.; Newman, J. W.; Goodfriend, T.; Wishart, D. S. *PLoS One* **2011**, *6*, e16957.

(4) Trethewey, R. N. *Curr. Opin. Plant Biol.* **2004**, *7*, 196−201.

(5) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(6) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80*, 6382−6389.

(7) Katajamaa, M.; Miettinen, J.; Orešič, M. *Bioinformatics* **2006**, *22*, 634−636.

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

(9) Lommen, A. *Anal. Chem.* **2009**, *81*, 3079−3086.

(10) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8*, 719−726.

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640

(11) Wang, Y.; Yang, F.; Wu, P.; Bu, D.; Sun, S. *BMC Bioinf.* **2015**, *16* (1), 1−6.
(12) Roest, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmstrom, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, *13*, 741−748.
(13) Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. *Anal. Chim. Acta* **2016**, *914*, 17−34.
(14) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.
(15) Ho, T.-J.; Kuo, C.-H.; Wang, S.-Y.; Chen, G.-Y.; Tseng, Y. J. *J. Mass Spectrom.* **2013**, *48*, 234−242.
(16) Wang, S.-Y.; Kuo, C.-H.; Tseng, Y. J. *Anal. Chem.* **2015**, *87*, 3048−3055.
(17) Treviño, V.; Yañez-Garza, I.-L.; Rodriguez-López, C. E.; Urrea-López, R.; Garza-Rodriguez, M.-L.; Barrera-Saldaña, H.-A.; Tamez-Peña, J. G.; Winkler, R.; Díaz de-la-Garza, R.-I. *J. Mass Spectrom.* **2015**, *50*, 165−174.
(18) Erny, G. L.; Acunha, T.; Simó, C.; Cifuentes, A.; Alves, A. *Chemom. Intell. Lab. Syst.* **2016**, *155*, 138−144.
(19) Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.-U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. *Mol. Cell. Proteomics* **2014**, *13*, 348−359.
(20) Liang, Y.-J.; Lin, Y.-T.; Chen, C.-W.; Lin, C.-W.; Chao, K.-M.; Pan, W.-H.; Yang, H.-C. *Anal. Chem.* **2016**, *88*, 6334−6341.
(21) Chang, H.-Y.; Chen, C.-T.; Lih, T. M.; Lynn, K.-S.; Juo, C.-G.; Hsu, W.-L.; Sung, T.-Y. *PLoS One* **2016**, *11*, e0146112.
(22) Åberg, K. M.; Torgrip, R. J. O.; Kolmert, J.; Schuppe-Koistinen, I.; Lindberg, J. *J. Chromatogr. A* **2008**, *1192*, 139−146.
(23) Conley, C. J.; Smith, R.; Torgrip, R. J. O.; Taylor, R. M.; Tautenhahn, R.; Prince, J. T. *Bioinformatics* **2014**, *30*, 2636−2643.
(24) Tengstrand, E.; Lindberg, J.; Åberg, K. M. *Anal. Chem.* **2014**, *86*, 3435−3442.
(25) Ji, H.; Lu, H.; Zhang, Z. *RSC Adv.* **2016**, *6*, 56977−56985.
(26) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.
(27) Wang, H.; Song, M. *R J.* **2011**, *3*, 29−33.
(28) Hyndman, R.; Koehler, A.; Ord, K.; Snyder, R. *Forecasting with Exponential Smoothing*; Springer: Berlin Heidelberg, 2008; pp 53−66.
(29) Zhang, W.; Zhao, P. X. *BMC Bioinf.* **2014**, *15*, S5.
(30) At, R. A.; Ion, T. *Applied regression analysis*; Wiley, 2005; pp 40−44.
(31) Zhang, Z.-M.; Tong, X.; Peng, Y.; Ma, P.; Zhang, M.-J.; Lu, H.-M.; Chen, X.-Q.; Liang, Y.-Z. *Analyst* **2015**, *140*, 7955−7964.
(32) Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2006**, *22*, 2059−2065.
(33) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z.; Liu, Z.-X.; Zhang, Q.-M.; Ding, L.-X.; Ye, F.; Zhou, H. *J. Raman Spectrosc.* **2010**, *41*, 659−669.
(34) Wong, J. W. H.; Durante, C.; Cartwright, H. M. *Anal. Chem.* **2005**, *77*, 5655−5661.
(35) Wong, J. W. H.; Cagney, G.; Cartwright, H. M. *Bioinformatics* **2005**, *21*, 2088−2090.
(36) Campello, R. J. G. B.; Moulavi, D.; Sander, J. In *Advances in Knowledge Discovery and Data Mining*; Springer: Berlin, Heidelberg, 2013; pp 160−172.
(37) Campello, R. J. G. B.; Moulavi, D.; Zimek, A.; Sander, J. *ACM Trans. Knowl. Discovery DATA* **2015**, *10*, 344−371.
(38) Erny, G. L.; Acunha, T.; Simó, C.; Cifuentes, A.; Alves, A. *J. Chromatogr. A* **2016**, *1429*, 134−141.
(39) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166−173.
(40) Thévenot, E. A.; Roux, A.; Xu, Y.; Ezan, E.; Junot, C. *J. Proteome Res.* **2015**, *14*, 3322−3335.
(41) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20*, 341−351.
(42) Bylesjö, M.; Eriksson, D.; Sjödin, A.; Jansson, S.; Moritz, T.; Trygg, J. *BMC Bioinf.* **2007**, *8*, 207.
(43) Rantalainen, M.; Bylesjö, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2007**, *21*, 376−385.
(44) Bylesjö, M.; Rantalainen, M.; Nicholson, J. K.; Holmes, E.; Trygg, J. *BMC Bioinf.* **2008**, *9*, 106.
(45) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.
(46) Liaw, A.; Wiener, M. *R News* **2001**, *23*, 29−33.
(47) Giavalisco, P.; Li, Y.; Matthes, A.; Eckhardt, A.; Hubberten, H.-M.; Hesse, H.; Segu, S.; Hummel, J.; Köhl, K.; Willmitzer, L. *Plant J.* **2011**, *68*, 364−376.
(48) Zhou, X.; Wang, Y.; Yun, Y.; Xia, Z.; Lu, H.; Luo, J.; Liang, Y. *Talanta* **2016**, *147*, 82−89.
(49) Jervis, J.; Kastl, C.; Hildreth, S. B.; Biyashev, R.; Grabau, E. A.; Saghai-Maroof, M. A.; Helm, R. F. *J. Agric. Food Chem.* **2015**, *63*, 9879−9887.
(50) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; Velzen, E. J. J. van; Duijnhoven, J. P. M. van; Dorsten, F. A. van *Metabolomics* **2008**, *4*, 81−89.
(51) Szymańska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A. *Metabolomics* **2012**, *8*, 3−16.

DOI: 10.1021/acs.analchem.7b01547
Anal. Chem. 2017, 89, 7631−7640