analytical chemistry

ACS AuthorChoice

Article

pubs.acs.org/ac

Terms of Use

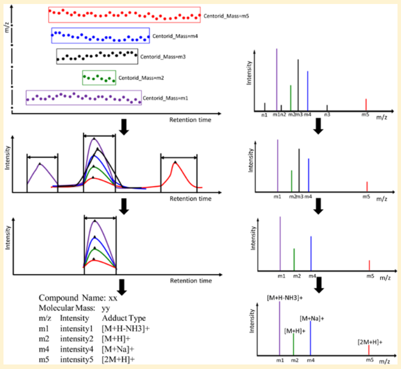
MET-COFEA: A Liquid Chromatography/Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation

Wenchao Zhang, Junil Chang, Zhentian Lei, David Huhman, Lloyd W. Sumner, and Patrick X. Zhao*

Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73401, United States

S Supporting Information

ABSTRACT: In this paper, we present a novel liquid chromatography/mass spectrometry (LC/MS) data processing and analysis platform, MET-COFEA (METabolite COmpound Feature Extraction and Annotation). MET-COFEA detects and clusters chromatographic peak features for each metabolite compound by first comprehensively evaluating retention time and peak shape criteria and then annotating the associations between each peak's observed *m*/*z* value with the corresponding metabolite compound's molecular mass. MET-COFEA integrates a series of innovative approaches, including novel mass trace based extracted-ion chromatogram (EIC) extraction, continuous wavelet transform (CWT)-based peak detection, and compound-associated peak clustering and peak annotation algorithms. On the basis of the deduced neutral molecular mass and retention time, we have also developed a new alignment algorithm that uses compound-associated peak groups instead of individual peaks to align the same metabolite compound across samples from different electrospray ionization (ESI) modes, different instruments, even different experimental conditions. MET-COFEA has been systematically tested on a series of LC/MS profiles of mixed standards at different concentrations as well as real untargeted LC/MS plant metabolomics data. We compared the performances of MET-COFEA with the existing publicly available tools at LC/MS peak analysis level and demonstrated its excellent performance in this arena. MET-COFEA is freely available at http://bioinfo.noble.org/manuscript-support/met-cofea/.



Liquid chromatography coupled with mass spectrometry (LC/MS) is an important analytical technology for metabolomics experiments with several advantages over gas chromatography coupled with mass spectrometry (GC/MS) approaches.[1] For example, LC/MS does not require derivatization for polar compounds, is capable of analyzing a larger range of compounds, commonly uses lower-energy electrospray ionization (ESI), and provides greater sensitivity. However, LC/MS data analysis still faces several significant challenges. For example, the *m*/*z* drift needs to be carefully considered for precise extracted-ion chromatogram (EIC) extraction, which is an early step in the LC/MS data analysis pipeline that affects all downstream analysis. The accuracy of EIC extraction by binning data points within a fixed tolerance suffers when the "larger than the fixed tolerance m/z drifts among scans", which are often observed in LC/MS profiles.

Another challenge of LC/MS is that the peak shape of the LC/MS chromatogram is affected by many factors, such as a poor signal-to-noise ratio (SNR) and the presence of contaminants that makes the detection of meaningful chromatographic peaks more challenging. Several specific algorithms have been developed to detect peaks in the LC/MS chromatogram. One available program, MZmine,[2,3] uses an optimized local maximum detection algorithm to detect peaks. Other state-of-the-art analysis tools such as XCMS[4] and MAVEN[5] implement continuous wavelet transform (CWT), which is considered to be a superior algorithm to recognize potential peaks from a noisy chromatographic background.

Retention time shift is another critical issue in LC/MS peak detection and alignment. To correct the retention time shift of a specific compound, current tools such as XCMS, MZmine, and MAVEN align peaks across samples/assays using the individual peaks as a reference, which often causes false alignment and misalignment because the coelution and common fragments are frequently observed in LC/MS analyses. Chae et al.[6] proposed a peak-block based alignment that combines multiple reference peaks with high SNRs into a "peak block" and aligns peaks using the entire "peak block" instead of an individual peak as the reference. This algorithm minimizes the misalignment that can occur from the shift in retention time of a single reference peak. An extended peak-block based alignment algorithm, called peak annotation,[7] has been further proposed to combine the peaks from the same metabolite into the same peak block, which is expected to provide an even better correction on retention time.

Metabolomics analyses by LC/MS yields thousands of chromatographic peaks that arise from metabolites, fragments, isotopes, and adducts. The mass signals/ chromatogram from

Received: November 18, 2013
Accepted: May 23, 2014
Published: May 23, 2014

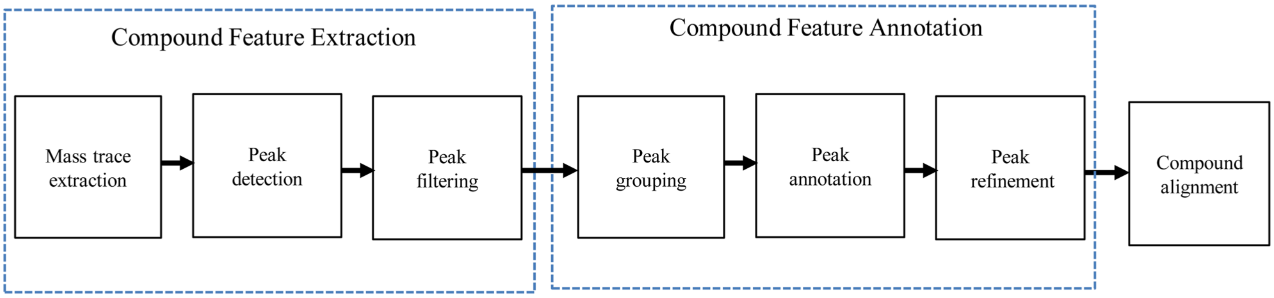


Figure 1. MET-COFEA data analysis workflow. Three integrative modules compound feature extraction, compound feature annotation, and compound alignment are included.

the same metabolite compound have a common molecular mass deduced from the observed m/z value and similar retention time and peak shape. These features can be considered collectively as multifeatures of the specific metabolite. An analysis that integrates these multifeatures can therefore lead to more accurate metabolite compound quantification and structure elucidation.

In this paper, we present a novel LC/MS data processing platform entitled MET-COFEA (METabolite COmpound Feature Extraction and Annotation) that can automatically detect the meaningful metabolite-associated peak features, annotate the relationship between each peak feature with the corresponding metabolite, and align the corresponding metabolite across different samples based on annotated compound-associated peak group information.

MET-COFEA has been systematically tested on a series of LC/MS profiles with mixed standards at different concentrations and real untargeted LC/MS plant metabolomics profiling data. Compared with other publicly available LC/MS data analysis software, MET-COFEA provides the excellent performance based on a comprehensive evaluation of detected chromatographic peaks. Another major highlight of MET-COFEA is that it associates each individual metabolite with a group of annotated peaks whenever applicable, a feature that, currently, only can be provided by CAMERA[8] in R environment.

■ EXPERIMENTAL SECTION

Biological samples with mixed known standards at different concentrations and biological samples from four different *Medicago truncatula* ecotypes were prepared and analyzed on the Waters Acquity ultra-performance liquid chromatography (UPLC) coupled with a quadrupole time-of-flight (Q-TOF) Premier mass spectrometer (Waters, Milford, MA) to develop, evaluate, and validate MET-COFEA. The experimental details are described in the Supporting Information.

■ DATA ANALYSIS METHODS

Data Analysis Pipeline and Workflow. The data analysis pipeline (Figure 1) consists of three sequential modules: (1) compound feature extraction, (2) compound feature annotation, and (3) compound alignment.

The compound feature extraction module includes three sequentially connected submodules: (1) mass trace based EIC extraction, (2) peak detection, and (3) peak filtering. The compound feature annotation module also includes three sequentially connected submodules: (1) peak grouping, (2) peak annotation, and (3) peak refinement. These processing steps are used to analyze each LC/MS data sample input. The alignment step is used to align the same potential metabolite compound across different samples. We describe the details of each step in the following sections.

Compound Feature Extraction. *Mass Trace Extraction.* The observed m/z values for the same analyte in different analyses can vary within the mass accuracy of the mass spectrometer. In addition, some analytes may still be unresolved and coelute. EIC extraction is typically the first step in LC/MS data analysis. Binning the data into their corresponding EICs with a fixed m/z tolerance is a direct and simple EIC extraction method. One of the major drawbacks of the binning method, however, is that binning can split the same mass signals into two adjacent bins.[4,9,10] A more sophisticated method extracts the mass trace in the 2D space spanned by m/z and retention time (or scan) according to both the mass drift and the similarity of intensity and retention time among the neighboring scan points. This extraction method is similar to object tracking and detection that is used in the video-processing field. A valid mass trace starts at a specific scan number and ends at another scan number, and the minimum continuity and maximum gap of neighboring points should meet the predefined thresholds. As the scan number increases, more and more valid data points are appended onto the mass trace. The decision rule to append a new data point onto a mass trace determines the performance of the final extracted mass trace. Tautenhahn's method[1] only uses the centroid mass of the mass trace to decide whether a new data point should be appended onto the mass trace or not. In that algorithm, if the mass difference between the new data point and the mass centroid of the mass trace is smaller than the predefined tolerance, then the new data point is appended onto the mass trace. One of the drawbacks of this method is that EICs may be split when mass traces with continuous drift along one direction are found, which often occurs in LC/MS profiles. To solve this issue, we developed a novel two-phase mass trace tracking algorithm instead of increasing ppm tolerance. In our algorithm, the centroid mass of a mass trace and the most recently appended data point are both evaluated in the decision to append the new data point onto the mass trace. Our proposed mass trace extraction method is described in detail in the following subsections.

Mass Trace Initialization. Starting from a user-defined start scan, for example, *scan = 1*, for each data point *i* of the scan, add the data point as the starting point of an independent mass trace represented by a linked list that is stored in the computer's memory, and set each mass trace's centroid mass value with the data point's m/z value. This step will create an array of initialized linked lists.

Mass trace tracking phase I. For each data point i of scan s that is in the set *[start_scan, end_scan]*, find the possible candidate mass traces in which the mass difference between data point $m/z(i,s)$ and the j^{th} mass trace's centroid mass value meet the following requirement:

$$|m/z(i, s) - mass\_trace(j)\cdot centroid\_mass| < \mu \tag{1}$$

where μ is an user-defined mass shift tolerance or equivalent ppm value.

Mass Trace Tracking Phase II. From the possible mass trace candidates, search for the optimal mass trace candidate that has the minimum distance between the most recent data point d of the mass trace and the data point $m/z(i,s)$. The distance between the most recent data point d of the mass trace and the data point $m/z(i,s)$ is defined as

$$\begin{aligned} Distance(m/z(i, s), d) &= Weight_{m/z} \times \delta\_mass\_ppm \\ &+ Weight_{Intensity} \times \delta\_\log\_Intensity + Weight_{RT} \times \delta\_scan \end{aligned} \tag{2}$$

We use the mass difference in ppm format and the intensity difference in log intensity so that the weighing coefficients for mass, intensity, and scan are comparable and can be adjusted easily to the optimal values. For the analysis of HPLC/UPLC data, we empirically set 0.5, 0.2, and 0.3 as the coefficient weights of $Weight_{m/z}$,$Weight_{Intensity}$, and $Weight_{RT}$, respectively.

Mass Trace Checking and Updating. When a new data point is appended onto a valid mass trace, the mass trace's centroid mass is updated accordingly. For each data point of the scan, if a matched mass trace to append cannot be found, then a new mass trace for this data point is initialized. A mass trace will be marked as invalid and deleted from the mass trace list if the mass trace length is smaller than the predefined threshold and the mass trace miss count number is larger than the predefined threshold. For each scan, continue the mass trace tracking and then check the update procedure until *end_scan* is reached.

Mass Trace Merging. A mass trace merging process is immediately followed after the mass trace tracking is completed. For any two neighboring mass traces, if their intertrace distance, as measured by the difference between their two centroid masses, is smaller than the user-defined value, then the two mass traces are merged into one mass trace. The mass trace merging procedure can therefore overcome the mass trace splitting issue to some degree. Figure S-1(Supporting Information) shows one extracted EIC and its corresponding chromatogram following mass trace extraction.

Peak Detection. The next important procedure is to detect the meaningful peaks within the EICs that have been extracted by the mass trace method. Several peak detection methods have been developed, such as the local maximum detection method implemented in MZmine[2] and the matching filter method being as one of peak detection options in XCMS.[4] However, the local maximum detection method is sensitive to noisy peaks and the matching filter method can only detect peaks with a fixed width that is defined by the user.[1] These methods are therefore limited in their abilities to detect all meaningful peaks.

We have developed a CWT-based peak detection algorithm that can automatically and robustly detect peaks with different scales and shapes in LC/MS profiles. The mathematical representation of the CWT[11] is

$$\begin{aligned} &CWT(f)(s, \tau) = \int_{-\infty}^{+\infty} f(t)\psi_{s,\tau}(t) \text{ and} \\ &\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right), \quad s > 0, \quad \tau \in R \end{aligned} \tag{3}$$

where $f(t)$ is the signal, ψ is the mother wavelet, S is the scale, and τ is the translation. Here, we use the Mexican Wavelet, which is the normalized second derivative of the Gaussian function $e^{-x^2/2}$, as the mother wavelet.

Our CWT-based peak detection algorithm includes the following five integrative submodules:

CWT Transformation. This submodule transforms the one-dimensional (1D) input signal $f(t)$ (Figure S-2A, Supporting Information) into a two-dimensional (2D) wavelet coefficient matrix spanned by scale and time translations (Figure S-2B, Supporting Information) using eq 3. The wavelet decomposition level is determined by the scale range, which can be linearly mapped by the user-specified peak width range parameters.

Local Maximum Detection. This submodule detects the local maximum of the wavelet coefficient at each scale and generates a 2D map with values of 0 or 1, where 1 represents a real local maximum point at the corresponding retention time position and scale.

Profile Peak Pattern Detection. This submodule identifies the meaningful branch pattern. A meaningful branch pattern is one that branches across multiple neighboring scales and is longer than the specified threshold. A longer ridge length indicates a more meaningful corresponding chromatographic peak.

Marker Point Detection. This submodule detects the marker point for each profile peak branch. The marker point indicates the point with the local maximum CWT coefficient across the neighboring scales of the pattern branch (colored dots in Figure S-2C in the Supporting Information).

Profile Peak Parameter Retrieving. This submodule retrieves the peak feature quantifications for each marker point, such as the peak's apex position and boundary (Figure S-2D, Supporting Information), and then quantifies the peak's SNR, height, and area.

Peak Filtering. Initially detected peaks are subject to further quality checks that filter out irregular low quality and other noisy peaks. Gaussian similarities, sharpness, and signal-to-noise ratio (SNR) are used to evaluate GC/MS chromatographic peak quality and deconvolute coeluted compounds.[12] Compared with GC/MS, the LC/MS technique particularly adopted electrospray ionization (ESI) as the ionization method, which results in high level chemical noise and background due to contaminants, and some factors such as LC mobile phase, atmospheric environment, or solvent types. The nice properties of GC/MS chromatographic peaks, such as being smooth, having more data points and a generally symmetrical shape, are not true for LC/MSchromatographic peaks, so the measurements including Gaussian similarity and sharpness can be adopted to efficiently evaluate the quality of GC/MS chromatographic peaks but not necessary suitable to evaluate the quality of LC/MS chromatographic peaks. The program MAVEN adopts a more complex machine-learning neural network approach to classify detected peaks into "good" or "bad" categories according to the training results of "good" and "bad" peak sets. However, these "good" and "bad" training peak sets need to be manually annotated by an expert and should also cover each possible case, which greatly limits the

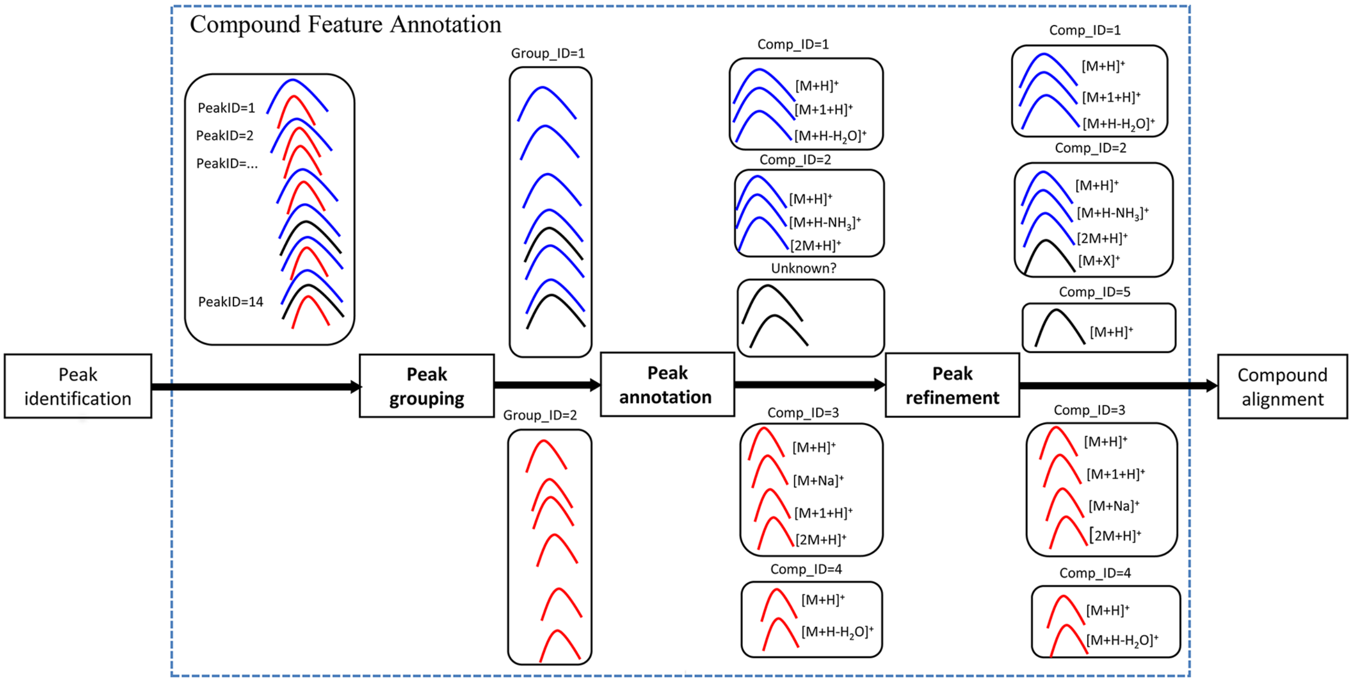


Figure 2. Illustrative diagram of compound feature annotation using MET-COFEA. The identified peaks are processed by three sequentially connected submodules: peak grouping, peak annotation, and peak refinement, and the annotated compound associated peaks are output for alignment.

applicability of this method. Therefore, MET-COFEA has adopted several simple yet effective criteria to filter out "bad" peaks. These criteria include the following.

Peak Intensity. The peak intensity is represented as the intensity at the peak apex. The meaningful metabolite fragment's peak intensity should be higher than a user-specified threshold.

Local Peak SNR. The SNR is defined in the wavelet domain by the ratio of the CWT coefficient at the marker point to the 95% quintile of the absolute CWT coefficient in scale 1.

Peak Significance Level. Peak significance level is defined by the ratio between the mean intensity value of data points near the peak apex and the mean intensity value of data points near the two boundaries.

Zig_Zag_Index. The Zig_Zag_Index is newly proposed and adopted here to evaluate the degree of zigzag in a chromatographic peak. Suppose the intensities of a chromatographic peak are represented as $I_1, I_2,..., I_{n-1}, I_n, I_{n+1},..., I_N$ and define the effective peak intensity (EPI) by subtracting the baseline from the observed peak intensity, the Zig_Zag_Index can be represented as

$$Zig\_Zag\_Index = \frac{\sum_{n=2}^{n=N-1} (2I_n - I_{n-1} - I_{n+1})^2}{N*EPI^2} \tag{4}$$

The detailed procedure to calculate Zig_Zag_Index is described in the Supporting Information.

On the basis of our tests on real LC/MS profile data, the proposed Zig_Zag_Index can evaluate the zigzag degree of a chromatographic peak shape. A lower Zig_Zag_Index indicates a higher peak quality.

Triangle Peak Area Similarity Ratio (TPASR). The TPASR is defined as follows:

$$\begin{cases} TPA = 0.5*Peak\_Width*Intensity(Peak\_Apex) \\ RPA = \sum_{i=Left\_Boundary}^{Right\_Boundary} Intensity(i) \\ TPASR = \frac{|TPA - RPA|}{TPA} \end{cases} \tag{5}$$

TPASR provides an index for the closeness in area between the detected peak and the corresponding triangle peak that was connected by the peak's apex and two boundaries. A TPASR value closer to 0 indicates a better peak quality.

The peak intensity, local peak SNR, peak significance level, and TPASR can evaluate the chromatographic peak quality from a macro viewpoint, and the Zig_Zag_Index can evaluate the chromatographic peak quality from a micro viewpoint. Together these criteria provide a more complete evaluation of the detected chromatographic peak quality.

Compound Feature Annotation. Different compounds can have the same molecular mass and can also coelute at the same retention time, which poses great difficulties for compound identification and quantification. In ESI-LC/MS spectra, compound adduct ion peaks such as $[M + H]^+$, $[M + K]^+$, $[M + Na]^+$, $[M + H - H_2O]^+$, $[M + H - NH_3]^+$, and $[2M + H]^+$, and their isotope peaks such as $[M + 1 + H]^+$, $[M + 2 + H]^+$, $[M + 1 + H - H_2O]^+$, and $[2M + 1 + H]^+$ are often observed with the same retention times and peak shapes. This information is therefore very important and can be considered as multiple features of a compound (i.e., in contrast to using a single molecular peak) to facilitate compound identification. Using this information, we have developed a series of innovative algorithms to group and annotate associated peaks for compound identification. Specifically, these algorithms include (1) group the molecular peak, adduct peaks, and corresponding isotope peaks for each compound according to both retention time and peak shape similarities; (2) associate and annotate the adduct and isotope peaks according to the same molecular mass information; and (3) refine some special isolated peaks. Figure 2 shows the modules and submodules that comprise our compound feature annotation procedure.

Peak Grouping. Among identified peaks, high intensity peaks are usually more confident than low intensity peaks to distinguish from noise. In addition, we have observed that a compound-related peak's apex generally falls within the boundary range of the highest intensity peak of the compound. On the basis of these observations, we have developed the following algorithm/criteria for peak grouping: (1) Retrieve the highest intensity peak from a list of ungrouped peaks, and retrieve the highest intensity peak's apex position. (2) Find the peaks in which the peak's apex falls within a local range, which is defined by increasing and decreasing a fixed tolerance value (e.g., 3 scans) to the highest intensity peak's apex. (3) If the peak shape similarity between this peak and the highest intensity peak is smaller than a predefined threshold configured by the user, then group that peak with the highest intensity peak. The peak shape similarities are calculated by the dot product of two peak pairs, and the group peaks will be assigned and represented by a unique Group ID. The group peaks should have similar retention times and peak shapes. (4) Remove these group peaks from the ungrouped peak list. (5) Repeat steps 1−4 until all peaks have been grouped.

This algorithm will group the associated peaks by hierarchically clustering their retention time and peak shape similarities starting from the peaks with the highest intensity. Figure S-3 (Supporting Information) provides an example of the peak grouping result.

Peak Annotation. The peaks with the same Group ID may come from different compounds when two or more compounds coelute and have similar peak shapes. The accurate mass information is therefore applied to separate the grouped peaks. The observed peak signal S (which is measured in m/z) is considered to be the result of the molecular mass M modified by the chemical process during soft ionization:

$$S = \frac{M{*}Para\_N + Para\_Mass\_Shift}{Para\_CS} \qquad (6)$$

The parameter set, $P = (Para\_N, Para\_Mass\_Shift, Para\_CS)$, corresponds to one type of chemical adduct process, and the isotope peak is considered to be a special adduct process. For soft ionization in LC/MS, common adducts and their corresponding parameter sets can be defined, and the adduct types can be configured in MET-COFEA according to their chemical process and experimental and instrument setup. Tables S-1 and S-2 (Supporting Information) report some common chemical adducts and their corresponding parameter sets in (+) ESI mode and (−) ESI mode, respectively.

Assume that there are M peak signals in one meaningful peak group and that each peak signal has N chemical adduct processes. From eq 6, we can calculate each possible molecular mass $M_{i,j}$ as

$$M_{i,j} = \frac{S_i{*}Para\_CS_j - Para\_Mass\_Shift_j}{Para\_N_j} \qquad (7)$$

Thus, we can compute a $M \times N$ molecular mass matrix. All the molecular mass $M_{i,j}$ that are calculated from eq 7 should fall within an acceptable mass tolerance when the peaks are from the same compound's molecular mass. Compounds with similar retention times and peak shapes can therefore be further separated into different compound groups according to their corresponding molecular mass. Each compound group can be differentiated by assigning a unique Compound ID to establish information about the metabolite compound, including the molecular mass, the adduct peak types, and details about the adduct peaks.

Peak Refinement. Peaks with the same Group ID may be further separated into different compound groups according to the accurate mass information using eq 7. However, we also find that some isolated peaks with the same Group ID cannot be separated into any Compound ID groups. These peaks may belong to unknown fragment peaks or other meaningful compound sole molecular peaks. On the basis of the ESI ionization of LC/MS, it is possible that the compound has only one molecular peak but no adduct peaks or the adduct peak intensities are too weak to be detected. We have therefore developed a peak refinement algorithm to process isolated peaks that cannot be separated into any compound group based solely on the accurate mass information. We utilized HMDB[13] as the reference compound molecular peak libraries to assist the metabolite identification of these isolated peaks. The compound peak refinement procedure includes the following steps: (1) For each isolated peak, search the molecular mass across the HMDB libraries. If any are found, mark the isolated peak with a unique compound ID and remove the peak from the isolated peak list. (2) For each of the remaining isolated peaks, calculate the peak shape similarities between the isolated peak and each peak from all identified compound groups. The peak shape similarity of a peak pair is calculated based on their dot product. (3) Calculate the average peak shape similarity between the isolated peak and the peaks of each compound group. The average peak shape similarity between isolated peak i and compound peak group m can be represented as

$$PSS(i, m) = \frac{\sum_{n=1}^{N} <Peak\_i, Compound\_Peak(m, n)>}{N} \qquad (8)$$

where N is the peak number of compound peak group m. (4) Compare and find the optimal compound group that has the best average peak shape similarity. If the best average peak shape similarity is smaller than a predefined threshold, then we have more confidence that this peak is a fragment peak of compound group m. Therefore, we add the isolated peak to this compound group and mark this isolated peak as a high possible unknown fragment adduct peak.

The MS fingerprints of a metabolite can be collectively described from these innovative compound feature extraction and compound feature annotation algorithms. The construction of these pseudospectra/multifeatures can be used to create LC/MS metabolite ESI spectral libraries that have great promise to resolve the long-standing difficulty of comparable LC/MS reference spectra library construction in LC/MS-based metabolomics. Furthermore, such pseudospectrum/multifeature are back-compatible when only the molecular peak information is used and can therefore be used to search currently available LC/MS spectra in the public domain.

Figure S-4 (Supporting Information) provides an example of annotated compound-associated peaks from MET-COFEA and the tentative library search results against the HMDB database, which have demonstrated that multiple individual m/z searches for the annotated reconstructed pseudospectrum from MET-

COFEA in HMDB[13]/METLIN[14] can greatly reduce the number of candidate compounds and may even retrieve the specific compound.

Compound Alignment. Variability in analytical conditions such as temperature, pressure, and humidity can greatly affect an analyte's chromatographic elution time.[15] Therefore, alignment is often necessary to correct the shift in chromatographic retention time among different experimental analyses.[16] Many algorithms have been developed for chromatographic alignment. For example, Correlation Optimized Warping (COW),[15] Dynamic Time Warping (DTW),[17] COW-Total Ion Current (TIC),[18] COW-Component Detection Algorithm (CODA),[19] and DTW-CODA[20] all aim to align whole chromatograms. XCMS[4] and MAVEN[5] implement an iteration strategy of peak grouping/matching and peak retention time correction across assays that consider the detected peaks as a 1D peak feature and align the individual chromatographic peaks rather the whole compound associated chromatographic peaks. Other alignment methods, such as LCMSWARP[21] and LC/MS image-based alignment,[22] correct the mass drift and elution time variation based on the 2D LC/MS feature.

In all the existing chromatographic alignment algorithms, neither the 1D nor 2D feature-based alignments are related to the final identified compound's features. Therefore, the concept of peak block is proposed and utilized in alignment to preserve the peak shape and area information.[6] In MET-COFEA, the annotated compounds are characterized by their retention time and molecular mass, which are represented in its unique Compound ID. On the basis of the annotation result, we have proposed a compound alignment strategy. If the compound molecular mass and retention time fall within an acceptable tolerance, then compounds across different samples should be aligned together. Our novel compound annotation based peak alignment method includes the following steps: (1) Aggregate the annotated compound peak list across all samples and configure the necessary parameters for alignment. Initialize the unaligned compound list with the aggregated compound list and set Align_ID to 1. (2) Find the highest intensity compound and set this compound as the tentative reference compound. Set Align_RT and Align_MolecularMass with this compound's retention time and molecular mass, respectively. From the annotated compounds of all samples, find the compounds that meet the following requirements:

$$\begin{cases} |Compound\_RT - Align\_RT| < Align\_Window\_Phase1/2 \\ |Compound\_Molecular\_Mass - Align\_Molecular\_Mass| \\ \quad < Mass\_Tol \end{cases} \tag{9}$$

Mark these compounds as a tentative aligned compounds group. (3) Phase 1 compound alignment strategy: Align the tentative compounds as the final aligned compound group and mark these compounds with the same Align_ID. (3) Phase 2 compound alignment strategy: Find a compound with the median retention time of all of the compounds in the tentative aligned compound group. Use this compound as the target reference compound and set Align_RT and Align_Molecular_Mass with this compound's retention time and molecular mass, respectively. From all of the compounds in the unaligned compound list, find all the compounds that meet the following requirements:

$$\begin{cases} |Compound\_RT - Align\_RT| < Align\_Window\_Phase\ 2/2 \\ |Compound\_Molecular\_Mass - Align\_Molecular\_Mass| \\ \quad < Mass\_Tol \end{cases} \tag{10}$$

Align these compounds as the final aligned compound group and mark these compounds with the same Align_ID. (4) Move the aligned compounds from the unaligned compound list and increment the Align_ID by 1. Repeat the above process until the unaligned compound list is empty.

In this procedure, the compound alignment order is based on the intensity ranks rather than the elution time. In addition, the two-phase alignment strategy ensures that the reference compound is located in the middle of the alignment windows to overcome the splitting issue for compounds that are located on the boundary of the alignment windows.

In ESI mode, fragment/adduct peak patterns can vary greatly across different experiment conditions, but the deduced molecular mass from the associated peaks usually is conserved across the experimental conditions. Therefore, the proposed compound-based alignment is expected to align the same potential metabolites (known or unknown) from LC/MS profiles generated by different ESI modes, different instruments, and different experimental conditions.

■ RESULTS AND DISCUSSION

To evaluate the performance of MET-COFEA, we systematically compared the performance of MET-COFEA with 3 other open-source software programs, XCMS, MZmine, and MAVEN, by analyzing the detected chromatographic peaks from four standard mixed data sets. The details for the standard mixed data set are described in the Supporting Information. The four software programs were developed in different environments with different algorithms and parameter configurations. However, the performance of these four programs can be compared at the detected chromatographic peak feature level because all these programs provide the specific m/z values and retention times of detected chromatographic peaks. To make fair comparisons, we need to set the configurable parameters to comparable values for each individual software program for those parameters that have the same or similar physical meanings. And, we also need to preserve the optimal configuration of parameters for each individual software program. In addition, we need to clearly define the ground truth peaks before evaluating the peak detection level performance of each program.

Parameter Optimization. Some configurable parameters in the four software programs have the same or comparable physical meanings, but other configurable parameters are specific to the algorithm. To ensure a fair comparison, we chose the two parameters of greatest interest, peak retention time and the m/z value tolerance, to be consistent across programs. For example, we configured the peak width range for XCMS, min.Peak Width for MAVEN, and Min/Max Peak Width for MET-COFEA by the scan number and Peak duration for MZmine by minutes. In the four test data sets, one scan is equal to approximately 1.2 s. Therefore, we ensured that the peak retention time range was consistent across the four programs. Both *ppm* and *Da* can be configured for the m/z value tolerance. For parameters that correspond to a specific algorithm, we used four mixed standard data sets, POS_-STANDMIX-50NM, POS_STANDMIX-5NM, NEG_-STANDMIX-50NM, and POS_STANDMIX-5NM (see the Supporting Information), to test and optimize. These mixed data sets were generated by mixing the same 12 standard metabolites at both normal concentrations and 1:10 diluted concentrations and analyzing both in (+)ESI and (−)ESI mode. We manually adjusted the parameters iteratively for each software program until all of the molecular peaks of the standard metabolite were detected. The configured parameters for the four software programs are listed in Table S-3 (Supporting Information).

Performance Evaluation. Because of the ESI technique, the detected chromatographic peaks may include the metabolite molecular peak, adduct peak, isotope peaks, and fragment peaks and may also include some peaks from the solvent. Therefore, another critical issue for systematic performance evaluation is how to define the ground truth peaks as true positives. Wei et al.[23] defined the previous known acids with significant peak area changes as the true positives. However, Tautenhahn et al.[1] defined the peaks that were detected by multiple methods as the ground truth peaks, which may intuitively provide a more confident and stable identification result for statistical analyses. This latter strategy was adopted for our analysis: if one peak could be detected by at least three of the above four software programs, then this peak was defined as a ground truth peak. In all four software programs, detected peaks are characterized by their corresponding m/z value and retention time; therefore, peaks detected by different methods are considered identical if their m/z values and retention times fall within a specified tolerance. In this paper, we specified 0.05 Da as the m/z tolerance and 5 s as the retention time tolerance. We first calculated the ground truth peak number NP and true positive peak number TP through Venn diagram analysis (Figure 3; the mathematical details for the measures of performance evaluation is described in the Supporting Information), then systematically evaluated the performance of MET-COFEA, MAVEN, MZmine, and XCMS by computing the *Recall*, *Precision*, and *F-Score* for each program. The performance evaluation comparison results for all four data sets are listed in Table 1.

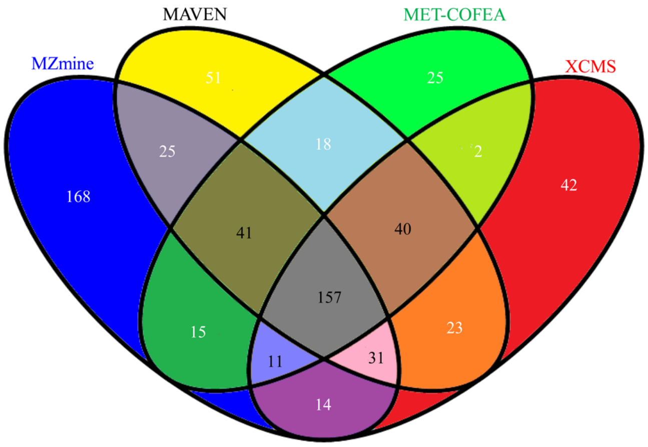


Figure 3. Venn diagrams of the chromatographic peaks detected by each of the four software programs. Only the 4-fold and 3-fold overlapping subsets (numbered in black font) are used as the true ground peaks. These numbers were generated from the results of each software program using the NEG_STANDMIX-50NM data set.

From Table 1, we can observe that (1) MAVEN achieves the highest *Recall* values, and XCMS and MZmine usually achieve lower *Recall* values. (2) MET-COFEA and XCMS usually achieve a high *Precision* value, and MZmine usually achieves the lowest *Precision* value. (3) MZmine always achieves the maximum detected peak number and a moderate *Recall* value; however, MZmine also has the lowest *Precision* value, which greatly affects the final *F-Score* value. 4) For MAVEN and XCMS, a higher *Recall* value corresponds with a lower *Precision* value, and MAVEN and XCMS perform nearly equivalently. (5) On the basis of the *F-Score* value, MET-COFEA outperforms the other three software programs. In addition, if we adjust the specific parameters of each of the four software programs separately, we can achieve a Recall/Precision value that approaches 1.0, but the corresponding *Precision*/*Recall* value simultaneously approaches 0. Therefore, the balanced *F-Score* initially increases to an optimal value and then decreases during the parameter optimization procedure.

Table 1. Performance Comparison of MET-COFEA, MAVEN, XCMS, and MZmine at the Peak Detection Level

		performance		
sample set	method	recall	precision	*F*-score
POS_STANDMIX-50NM	MET-COFEA	0.9180	0.6705	0.7750
	MAVEN	0.9464	0.6129	0.7440
	XCMS	0.8612	0.5676	0.6842
	MZmine	0.8265	0.5545	0.6637
POS_STANDMIX-5NM	MET-COFEA	0.9381	0.6864	0.7928
	MAVEN	0.9714	0.3908	0.5574
	XCMS	0.7571	0.4321	0.5502
	MZmine	0.8476	0.4611	0.5973
NEG_STANDMIX-50NM	MET-COFEA	0.8893	0.8058	0.8455
	MAVEN	0.9607	0.7154	0.8201
	XCMS	0.8536	0.7469	0.7967
	MZmine	0.8571	0.5195	0.6469
NEG_STANDMIX-5NM	MET-COFEA	0.9420	0.8025	0.8667
	MAVEN	0.9565	0.5841	0.7253
	XCMS	0.8696	0.8451	0.8571
	MZmine	0.8841	0.4296	0.5782

Compared with the other three software programs, the biggest advantage of MET-COFEA is its ability to group and annotate the detected chromatographic peaks according to the user-configurable table (see Tables S-1 and S-2 in the Supporting Information as examples). This process provides not only the molecular peak information but also information about related peaks that are associated with the same compound, such as the adduct peaks, isotope peaks, and fragment peaks. This information can, in turn, provide greater information about the structure of the metabolite.

Unsupervised Clustering of Untargeted LC/MS Metabolomics Data. To further validate the performance of MET-COFEA, two groups of real untargeted LC/MS plant metabolomics data were generated from root and leaf samples of *Medicago truncatula* using UHPLC-QTOF-MS experiments that were conducted in (−) ESI mode. Each group included 12 LC/MS profiles, which correspond to different ecotypes that are represented by their data filenames. The output from MET-COFEA provides the final aligned compound-associated peaklist table across all 12 data samples. For all the aligned compound-associated peaks that were represented by a unique Align_ID across samples, only the compounds that were found in all 12 samples were considered valid aligned compounds and included in further statistical analyses. For all the valid aligned compounds, the intensity or area of the annotated peak $[M - H]^-$ across all samples was used for quantitation. The 2D

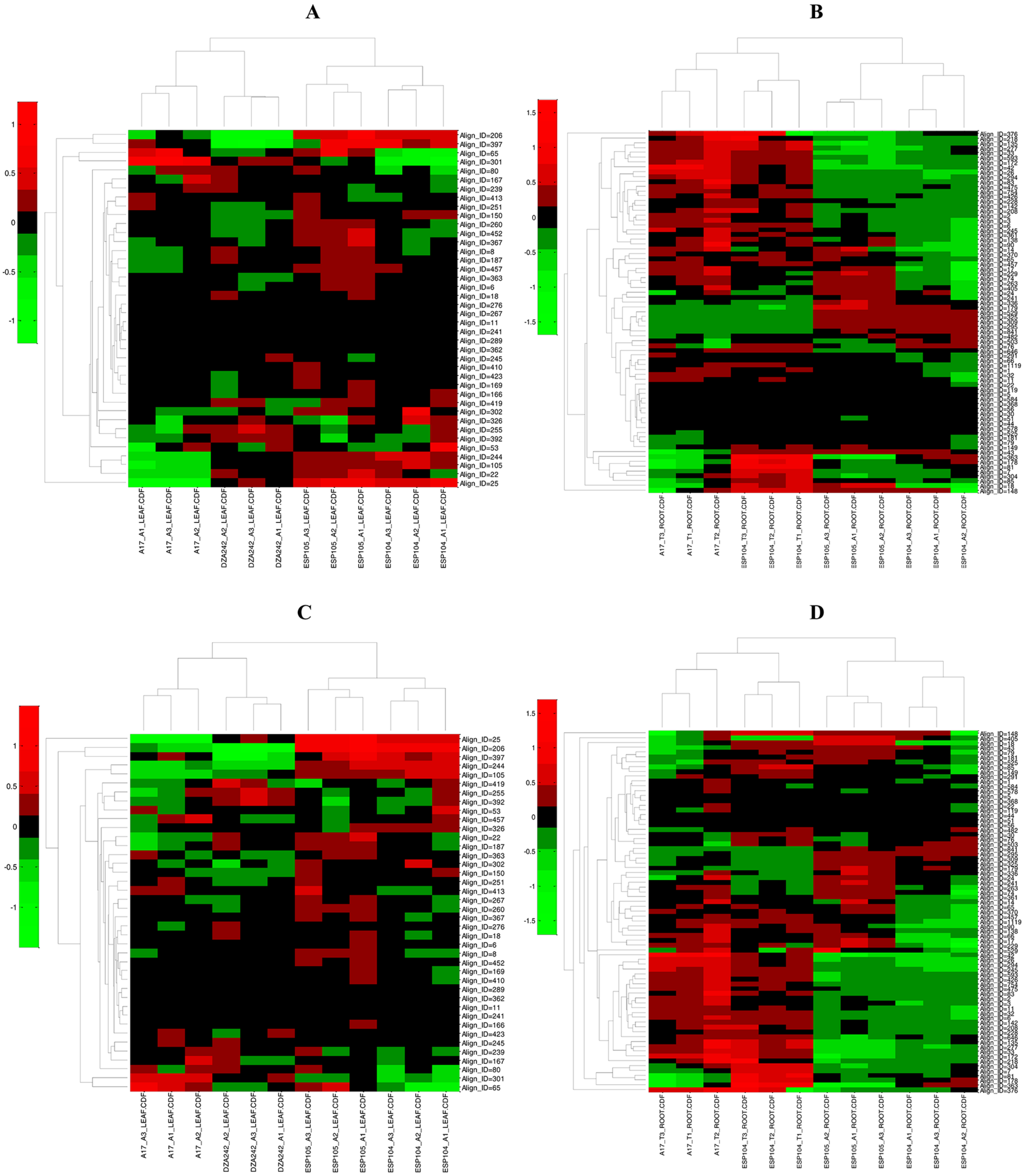


Figure 4. Unsupervised clustering results of the leaf and root samples from different ecotypes. (A,B) The clustering results for the leaf and root samples based on the intensity of peak $[M - H]^-$. (C,D) The clustering result for the leaf and root samples based on the area of peak of $[M - H]^-$.

matrix spanned by samples and valid Align_IDs was generated and clustered. Each row of the matrix was further processed by taking a log 2 transformation and then subtracting the mean of this row. Finally, 2D hierarchical clustering was applied to the entire 2D matrix. Figure 4 illustrates the unsupervised clustering results for the two groups of untargeted LC/MS data. From Figure 4, we can observe that each ecotype is identified by its sample name and has been clearly distinguished.

■ CONCLUSIONS

We have developed a novel LC/MS data processing and analysis platform, MET-COFEA. MET-COFEA is designed to detect and cluster meaningful chromatographic peak features for each metabolite compound based on the retention time and peak shape criteria and then annotate the relationship between each peak's observed m/z values with the corresponding metabolite's molecular mass. MET-COFEA integrates a series of innovative algorithms, such as mass trace based EIC extraction, CWT-based peak feature detection, compound-associated peak annotation, and compound alignment. Compared with other open-source software such as XCMS, MAVEN, and MZmine, MET-COFEA achieved the best performance in detecting and analyzing chromatographic peaks. MET-COFEA also enables the retrieval of comprehensively annotated metabolite information according to the user's configurations. The unsupervised clustering results for real untargeted UHPLC-QTOF-MS-based data from *Medicago truncatula* leaf and root samples show that MET-COFEA can clearly distinguish each ecotype.

MET-COFEA output compound-associated peak features, which have the potential capability to greatly reduce the number of possible compound candidates in library searching, and also the capability to improve the metabolite quantification accuracy. Currently, we are in the process of combining the output results from MET-COFEA with our published software, MET-IDEA,[24,25] which has been widely adopted in the plant metabolomics community for large-scale LC/MS-based metabolite quantification and comparison. We expect that the integration of MET-COFEA and MET-IDEA will provide comprehensive analysis platform for LC/MS peak identification, annotation, and quantification and thus better facilitate metabolomics data analysis.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

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

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pzhao@noble.org. Phone: 580-224-6725. Fax: +1-580-224-4743.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Samuel Roberts Noble Foundation for the support of this study. The authors also thank Dr. Jia Wei's group at the North Carolina Research Campus for valuable initial discussions about LC/MS technology, its applications in large-scale untargeted metabolomics studies, and data analysis.

■ REFERENCES

(1) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.
(2) Katajamaa, M.; Miettinen, J.; Orešič, M. *Bioinformatics* **2006**, *22*, 634–636.
(3) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.
(4) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.
(5) Melamud, E.; Vastag, L.; Rabinowitz, J. D. *Anal. Chem.* **2010**, *82*, 9818–9826.
(6) Chae, M.; Reis, R.; Thaden, J. *BMC Bioinf.* **2008**, *9*, S15.
(7) Tautenhahn, R.; Böttcher, C.; Neumann, S. BIRD. In *Lecture Notes in Computer Science*; Hochreiter, S., Wagner, R., Eds.; Springer: Berlin, Germany, 2007; pp 371–380.
(8) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283–289.
(9) Stolt, R.; Torgrip, R. J. O.; Lindberg, J.; Csenki, L.; Kolmert, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2006**, *78*, 975–983.
(10) Åberg, K. M.; Torgrip, R. J. O.; Kolmert, J.; Schuppe-Koistinen, I.; Lindberg, J. *J. Chromatogr., A* **2008**, *1192*, 139–146.
(11) Daubechies, I. *Ten Lectures on Wavelets*; Society for Industrial and Applied Mathematics: Philadephia, PA, 1992; p 377.
(12) Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X. *Anal. Chem.* **2012**, *84*, 6619–6629.
(13) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801–D807.
(14) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747–51.
(15) Tomasi, G.; van den Berg, F.; Andersson, C. *J. Chemom.* **2004**, *18*, 231–241.
(16) Nordström, A.; O'Maille, G.; Qin, C.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 3289–3295.
(17) Ramaker, H.-J.; van Sprang, E. N. M.; Westerhuis, J. A.; Smilde, A. K. *Anal. Chim. Acta* **2003**, *498*, 133–153.
(18) Sadygov, R. G.; Martin Maroto, F.; Hühmer, A. F. R. *Anal. Chem.* **2006**, *78*, 8207–8217.
(19) Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *Anal. Chem.* **2008**, *80*, 7012–7021.
(20) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *J. Proteome Res.* **2010**, *9*, 1483–1495.
(21) Jaitly, N.; Monroe, M. E.; Petyuk, V. A.; Clauss, T. R. W.; Adkins, J. N.; Smith, R. D. *Anal. Chem.* **2006**, *78*, 7397–7409.
(22) Vandenbogaert, M.; Li-Thiao-Té, S.; Kaltenbach, H.-M.; Zhang, R.; Aittokallio, T.; Schwikowski, B. *Proteomics* **2008**, *8*, 650–672.
(23) Wei, X.; Sun, W.; Shi, X.; Koo, I.; Wang, B.; Zhang, J.; Yin, X.; Tang, Y.; Bogdanov, B.; Kim, S.; Zhou, Z.; McClain, C.; Zhang, X. *Anal. Chem.* **2011**, *83*, 7668–7675.
(24) Broeckling, C. D.; Reddy, I. R.; Duran, A. L.; Zhao, X.; Sumner, L. W. *Anal. Chem.* **2006**, *78*, 4334–4341.
(25) Lei, Z.; Li, H.; Chang, J.; Zhao, P.; Sumner, L. *Metabolomics* **2012**, *8*, 105–110.