This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

analytical chemistry

Article

pubs.acs.org/ac

MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra

Yan Ma,[†] Tobias Kind,[†] Dawei Yang,[†,‡] Carlos Leon,[†,§] and Oliver Fiehn*[,†]

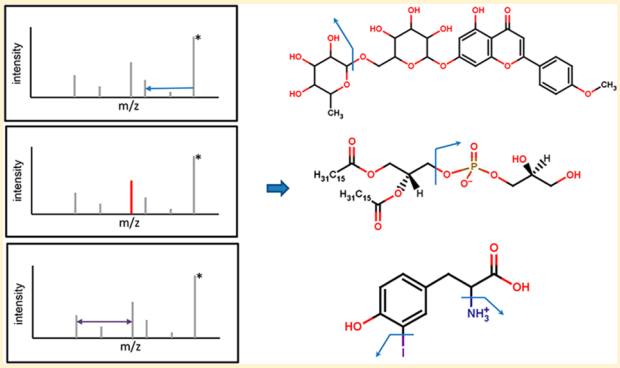
[†]UC Davis Genome Center−Metabolomics, University of California, Davis, California 95616, United States

[‡]SPKLOMHNM and Central Laboratory, Zhong Yuan Academy of Biological Medicine, Liaocheng University, Liaocheng People's Hospital, Liaocheng, Shandong 252000, P. R. China

[§]Biomedical Engineering School, Carlos III University, Avda Universidad 30, 28911, Leganes, Madrid, Spain

Supporting Information

ABSTRACT: Systematic analysis and interpretation of the large number of tandem mass spectra (MS/MS) obtained in metabolomics experiments is a bottleneck in discovery-driven research. MS/MS mass spectral libraries are small compared to all known small molecule structures and are often not freely available. MS2Analyzer was therefore developed to enable user-defined searches of thousands of spectra for mass spectral features such as neutral losses, m/z differences, and product and precursor ions from MS/MS spectra in MSP/MGF files. The software is freely available at http://fiehnlab.ucdavis.edu/projects/MS2Analyzer/. As the reference query set, 147 literature-reported neutral losses and their corresponding substructures were collected. This set was tested for accuracy of linking neutral loss analysis to substructure annotations using 19 329 accurate mass tandem mass spectra of structurally known compounds from the NIST11 MS/MS library. Validation studies showed that 92.1 ± 6.4% of 13 typical neutral losses such as acetylations, cysteine conjugates, or glycosylations are correct annotating the associated substructures, while the absence of mass spectra features does not necessarily imply the absence of such substructures. Use of this tool has been successfully demonstrated for complex lipids in microalgae.



Electrospray ionization-tandem mass spectrometry (ESI-MS/MS) is one of the preferred tools in metabolomics.[1,2] Interpretation of the large number of unknown tandem mass spectra obtained in liquid chromatography−tandem mass spectrometry (LC−MS/MS) studies is a bottleneck in discovery-driven metabolomics research. When exhaustive reference libraries are available, for example, for electron ionization mass spectra from gas chromatography/mass spectrometry (GC/MS) experiments, known structures can be easily annotated.[3] However, electrospray collision-induced dissociation MS/MS spectral libraries are still relatively small and are often not freely available.[4] Even large in silico generated tandem mass spectral libraries such as LipidBlast cannot cover all known lipid classes.[5] The MassBank[6] database covers MS/MS spectra of about 4 700 compounds, METLIN[7] comprises high-resolution MS/MS spectra for about 12 000 compounds, and the NIST14 mass spectral databases includes MS/MS spectra for around 9 000 compounds. Together, fewer than 30 000 unique compound MS/MS spectra are available through these three major experimental metabolomics libraries. Moreover, spectra were not obtained under standardized conditions but at different MS parameters and instruments. In contrast, PubChem contains more than 48 million small molecule structures; the natural compound space is estimated to be larger than 250 000 compounds.[8] For most of the known metabolites, no reference tandem mass spectra exist.

Therefore, library-independent annotation tools that exploit the potential structure information comprised in the tandem spectra are needed. Specific selection of spectral features and classifiers are essential for substructure analysis,[9] such as neutral loss and diagnostic fragment information. For example, neutral losses of 80 u (SO_3) and product ions of m/z 97 (HSO_4^-) in negative ESI MS/MS spectra are characteristic for alicyclic sulfates, while detection of m/z 80 anions (SO_3^- radical) have been reported to be specific for aromatic sulfates.[10] Besides, m/z differences between product ions can provide additional information about substructures. A classic example is the presence of a series of 14 mass unit differences, reflecting CH_2 units of alkyl side chains.[11]

To date, analysis of neutral losses and diagnostic fragments is mostly performed by manual assignments in a very time-consuming manner.[12,13] Software packages begin to address this problem by two approaches. *In silico* fragmentation software such as Mass Frontier[14] and MetFrag[15] help users to

Received: July 17, 2014
Accepted: September 27, 2014
Published: September 28, 2014

 © 2014 American Chemical Society

understand fragmentation patterns with a top-down approach, while other mass spectra annotation software, such as SIRIUS,[216] LipidInspector,[17] LipidXplorer,[18] mzGroupAnalyzer,[19] ALEX,[20] Metitere,[21] and MZmine 2[22] use a bottom-up strategy to associate unknown spectra with molecular formulas or structure information. Although LipidInspector and LipidXplorer use neutral loss and fragments for annotation, they are designed for specific compound classes (lipids in this case), instead of general small molecule metabolite MS/MS spectra annotations. Here we present a freely available tool that enables automatic substructure queries from accurate mass MS/MS spectra of small molecules. We present validation studies and show the usefulness of the software for annotating metabolomics MS/MS spectra.

■ EXPERIMENTAL SECTION

Software Development. MS2Analyzer is a mass spectral feature search program for tandem mass spectra and can be used for substructure annotation. It was programmed in Java (v 1.6.0_27) using the Eclipse Platform (v 3.3.2) and runs as a graphical user interface (GUI). The GUI was designed in Swing class using the Jigloo GUI builder (version 4.6.4) and then exported with the Fat-Jar Eclipse plug-in into one executable jar. The Java Excel API was used for writing the output data into Microsoft Excel 2003 spreadsheets. The program reads tandem mass spectra stored in NIST Mass Search format (MSP) and Mascot generic format (MGF). In MSP and MGF files, each MS/MS data set consists of metadata like name (title) and precursor m/z (pepmass), followed by a list of m/z-intensity pairs. During the analysis, spectral simplification is first performed using the user-defined intensity threshold in the GUI that removes all the peaks below the intensity threshold. This step is designed to minimize the effect from noisy peaks and facilitate further calculation.

Four types of mass spectral features can be searched: (1) precursor ions, (2) product ions, (3) neutral losses, and (4) m/z differences. The query mass values, together with their names and types of mass spectral features, are stored in a text file written in the required format. If the difference between calculated value and query value is within a certain m/z window defined by user, the software detects this feature. The results will be exported in a matrix, with names of query features as the column labels and titles of MS/MS spectra as row labels, in an Excel 2003 spreadsheet. Since the Excel 2003 sheet has a size limit of 65 536 rows by 256 columns, the maximum number of input spectra is 65 534 and the maximum query number is 255.

Collection of Published Neutral Losses for Substructure Predictions. We have collected and curated a large corpus of 147 neutral losses and their associated substructures by performing a systematic Google Scholar search for mass spectral fragmentation analysis up to year 2012 (see Table S-1 in the Supporting Information for the details). These values can be used in combination with the automatic MS2Analyzer software. The monoisotopic masses of the neutral losses were recalculated using the Molecular Weight Calculator software.[23] Other relevant information from the literature references is included, e.g., positive or negative ionization mode and the ionization source that was used. Most importantly, the proposed annotation of the related compound substructure or compound class was included for each specific neutral loss. Substructure information can be stored in SMARTS, a language for describing molecular patterns. To facilitate substructure searches, SMARTS of the selected 14 substructures were collected from the Daylight Web site or generated by PubChem Sketcher.[24] SMARTS generated from both ways were manually validated by SMARTSviewer[25] which can visualize the molecular pattern from SMARTS. The search of substructures was achieved by a batch tool, supported by OpenBabel[26] (version 2.3.1) Java library.

Investigation of Sensitivity and Specificity of Mass Spectral Features/Substructure Pairs. To develop and test mass spectral feature/substructure relationships, the NIST11 MS/MS library was chosen, comprising 64 511 tandem mass spectra. In order to reduce the impact from different instruments, ionizations, and adducts, only $[M + H]^+$ spectra acquired under positive electrospray ionization from Agilent QTOF 6530 mass spectrometers were used. The tandem mass spectra were exported into MSP files, and the corresponding structures were exported into structure-data files (SDF). Negative mode ESI tandem mass spectra were not used because of the small number of available spectra.

The neutral losses collected from literature were written into a text query file and searched through all the spectra using MS2Analyzer. The m/z window was set to be 0.01 because all accurate masses in the NIST11 library were found rounded up to the second decimal. The intensity threshold was set to be zero. At the same time, the SMARTS of substructure for each neutral loss was searched by a batch tool based on OpenBabel (version 2.3.1) Java library. This program was developed to test if certain substructure is present in a molecular structure. Names and SMARTS for substructures were written into separate lines in query text file. After both SDF file and query file were imported into the tool, it called the functions in OpenBabel to match SMARTS with structures in SDF file. The output yielded a substructure matrix in Excel 2003 spreadsheet, similar to the MS feature matrix.

Neutral loss search results of spectra of the same compound with different collision energies were combined. Combining the neutral loss matrix with the substructure matrix, a confusion matrix was created to show the performance of the prediction of substructure by neutral loss. Specificity and sensitivity for each neutral loss was calculated by the following equations:

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

The product ion and m/z difference examples in the paper were investigated in the same way.

Glycosides Identification from MassBank Accurate Mass MS/MS Spectra. For an example of independent data annotation, we used 3 359 accurate mass $[M + H]^+$ spectra that were downloaded from the MassBank database. The sources of the spectra include University of Connecticut, Washington State University, RIKEN, and others. The downloaded data files were converted to MSP files manually. Neutral losses were searched using MS2Analyzer. The m/z window was 0.005 Da and the intensity threshold was 0.05. Substructures were searched and a confusion matrix was created in the same way as the NIST library. Sensitivities and specificities of sugar losses were studied and reported.

Automatic Large-Scale Annotation of Lipids Using MS2Analyzer. A volume of 10 mL *Chlamydomonas reinhardtii* CC-125 cells were harvested at late-log phase by centrifugation, extracted, and analyzed by LC−QTOF data dependent tandem

mass spectrometry as described in the Supporting Information. The characteristic mass spectral features (neutral losses or diagnostic ions) for potential lipid classes in *Chlamydomonas reinhardtii* were collected from the literature and listed in Table S-2 in the Supporting Information. Possible precursor masses ($[M + H]^+$ and $[M + NH_4]^+$ for positive ion mode and $[M - H]^-$ for negative ion mode) and acyl side chain masses were calculated for query use. The MGF files from both positive and negative modes were analyzed by MS2Analyzer including the calculated precursor ion masses, acyl side chain masses, and mass spectral features collected from the literature. For different lipid classes, the m/z window was set to be 0.01 or 0.005 while the intensity threshold was set to be 0.005 to 0.05. Among each lipid class, the retention time basically follows the rule that retention time increases with carbon number and decreases with degree of unsaturation. To validate the identification results, the MS/MS data was also searched through LipidBlast, a large in silico lipid tandem mass spectrometry database, using NIST MS PepSearch[27] GUI. In the library search, precursor tolerance m/z was set to be 0.005 and fragment peak m/z tolerance was set to be 0.01.

■ RESULTS

Development and Use of the MS2Analyzer Software. The MS2Analyzer program has been developed in Java using Open Source IDE Eclipse. The executable jar file does not require installation and runs in the Java Runtime Environment on all platforms. As input, MS2Analyzer requires an MS/MS spectra data file and a mass spectra query file. Four types of mass spectral features can be searched: (1) neutral losses, (2) m/z differences, (3) product ions, and (4) precursor ions. Here, neutral loss means the m/z differences between precursor ion and product ions while the term m/z difference refers to the m/z differences between two product ions. The graphical user interface enables users to select a mass error window and an intensity threshold below which noise ions are to be ignored. The output result file is an Excel sheet that contains the associated annotations from the query file for each precursor ion.

Tandem mass spectra data files in NIST Mass Search format (MSP) or Mascot generic format (MGF) can be imported into the MS2Analyzer using a browser. MGF files can be exported from the vendor's software (e.g., Agilent MassHunter or AB SCIEX Analyst) or converted from raw data using the freely available ProteoWizard[28] MS converter. MGF files contain the accurate mass precursor information, product ion fragments, and their abundances. Typical UPLC−MS/MS runs with high scan rates can contain up to 10 000 MS/MS spectra in a single MGF file.

MS2Analyzer runs with application-dependent query text files for the four types of tandem mass spectral features, specified by user's needs for their specific projects. Query parameters, including mass error window and intensity threshold, can be set by users to reduce false positives. For accurate mass instruments such as quadrupole time-of-flight (QTOF) and Orbitrap, the typical mass accuracy is 2−5 ppm.[29,30] Therefore, a mass error window of 0.005 u is recommended for small molecules below 1000 u. Users can also change the mass error window and intensity threshold to adapt with other instrument types and specific queries. On a regular personal computer with 3.0 GHz CPU, the software searches 60 spectral features with a speed of over 40 000 spectra per minute and exports the query result as a Microsoft Excel file. Result files list the search results for all spectra in the consecutive order in rows; column headings report the query features and their corresponding masses as defined in the query text file. For all MS/MS spectra, successful detection of MS/MS features is given as "1"; absence of the query features is given as "0". Using the filter functions in Microsoft Excel, MS/MS spectra that have one or several of the queried mass spectral features can be selected and sorted for reports. Logical combinations of AND and OR queries can be performed in the same way.

The software program and source code are publicly available for commercial and noncommercial use under the open-source MIT license (http://opensource.org/licenses/MIT) and can be found under http://fiehnlab.ucdavis.edu/projects/MS2Analyzer/. Source code and the latest update are available on SourceForge at http://sourceforge.net/projects/ms2analyzer/.

Investigation of Spectra-Substructure Relationships Based on Neutral Losses. Neutral losses are frequently used for substructure annotations. MS2Analyzer helps to systematically investigate large numbers of accurate mass MS/MS spectra of small molecules with respect to substructure annotations. The sensitivity or true positive rate, calculated from the true positives and false negatives, relates to the probability that a given substructure query is correctly retrieved. The specificity or true negative rate, calculated from the number of true negatives and false positives, reflects the reliability of the test outcome for the presence of a given substructure and the probability to exclude a given substructure in our investigation. For successful annotation of unknown spectra the sensitivity of substructure prediction should be high and in order to exclude a high number of nonmatching substructures the specificity should be also sufficiently large.

For testing the capability of neutral loss searches, 19 329 $[M + H]^+$ accurate mass MS/MS spectra acquired by Agilent QTOF 6530 mass spectrometers were obtained from the NIST11 library, consisting of 2 036 compounds. All 2 036 compounds had masses of less than 1 000 u and consisted of major compound classes such as carboxylic acids, amines, and alcohols, including di- and tripeptides. Most compounds in the library were fragmented under multiple collision energies, varying from 1 to 60 V. These spectra and their corresponding structures were used to validate reported neutral losses for substructure annotations. In addition, 147 neutral losses and their corresponding substructures were retrieved from the literature (see 40 selected neutral losses in Table 1 and the full list in Table S-1 in the Supporting Information). After excluding neutral losses that were not observed in positive mode at all and excluding neutral losses for substructures that were not covered in the data set of 2 036 compounds, 14 typical neutral losses were selected based on the frequency found in the NIST11 data set and availability of fragmentation mechanisms. Neutral losses were searched using MS2Analyzer and substructures were searched using an in-house program based on OpenBabel. A confusion matrix was generated from the results to visualize the performance of the annotations. Results of the analysis of specificity and sensitivity for the 14 neutral losses are given in Table 2.

For the 14 selected neutral losses, an average specificity of 92.1 ± 6.2% was obtained, with the exception of a specificity of 42.9% for the presence of a carboxylic acid group by detection of a neutral loss of water. This low specificity is easily explained as water losses also frequently occur in other compound classes

Table 1. Compilation of 40 Published Neutral Losses from Electrospray and Atmospheric Pressure Chemical Ionization Collision-Induced Dissociation Mass Spectra

mass	formula	positive	negative	substructure or compound class
17.027	NH_3	+	−	aliphatic amines (aromatic amines), oximes
18.011	H_2O	+	−	carboxylic acids, aldehydes, ester
27.011	HCN	+	+	amines, aromatic nitrile, aminosulfonic acids
27.995	CO	+	+	carboxylic acids, aldehydes, nitroaromatics
28.019	H_2CN	−	+	aromatic amine
29.998	NO	+	+	nitroaromatics
30.011	CH_2O	+	+	aldehydes
32.026	CH_4O	+	−	methyl esters
33.988	H_2S	+	−	thiols
35.977	HCl	+	−	chlorides
43.990	CO_2	+	+	carboxylic acids, carbamates
45.993	NO_2	+	+	nitroaromatics
46.005	CH_2O_2	+	+	carboxylic acids
63.962	SO_2	−	+	sulfonic acids, sulfonates
63.998	CH_4OS	+	−	methionine sulfoxide
71.037	C_3H_5NO	+	−	serine residue
74.019	C_3H_6S	+	−	methionine side chain
79.957	SO_3	+	+	sulfonic acids
79.966	HPO_3	+	−	phosphates
80.965	HSO_3	+	−	sulfonic acids
81.045	$C_4H_5N_2$	−	+	histidine residue
81.972	H_2SO_3	+	−	sulfonate group
97.977	H_3PO_4	+	−	phosphates
121.020	$C_3H_7NO_2S$	+	+	cysteine conjugates
127.912	HI	+	−	aromatic iodides
130.063	$C_6H_{10}O_3$	+	−	dideoxyhexoside
132.042	$C_5H_8O_4$	+	−	pentoside
146.058	$C_6H_{10}O_4$	+	+	deoxyhexoside
146.069	$C_5H_{10}N_2O_3$	+	−	conjugate with gamma-GluCys or glutathione
162.053	$C_6H_{10}O_5$	+	−	hexoside
163.030	$C_5H_9NO_3S$	+	−	*N*-acetylcysteine conjugate
164.068	$C_6H_{12}O_5$	−	+	rhamonoside
176.032	$C_6H_8O_6$	+	+	glucuronides
194.043	$C_6H_{10}O_7$	+	−	glucuronides (benzylic)
203.079	$C_8H_{13}NO_5$	+	+	conjugate with *N*-acetylglucosamine (benzylic)
221.090	$C_8H_{15}NO_6$	+	−	conjugate with *N*-acetylglucosamine
248.053	$C_9H_{12}O_8$	+	−	malonylglucuronides
250.062	$C_8H_{14}N_2O_5S$	+	+	conjugate with gamma-GluCys
266.064	$C_9H_{14}O_9$	+	−	malonylglucuronides (benzylic)
307.084	$C_{10}H_{17}N_3O_6S$	+	−	glutathione conjugates

such as alcohols and carbohydrates; a water loss is therefore unspecific.[31] Conversely, observing a neutral loss of 35.977 u was found at 92.2% specificity for the presence of a chlorine atom in the 2 036 test compounds of the tested NIST11 MS/MS spectra, indicating a neutral loss of HCl. Interestingly, analysis of neutral losses gave only an overall sensitivity of 51.6 ± 22.1%. Sensitivities were found ranging from 15.4% for the presence of chlorine atoms using a HCl loss to 84.9% for the presence of carboxylic acids using a water molecule neutral loss. Taking these two substructures as an example, we can confidently say that detection of neutral losses of 35.977 clearly indicates the presence of a chlorine atom. But why do most MS/MS spectra of chlorinated organics lack this specific HCl neutral loss fragmentation? First of all, abundance for true positive detection for such neutral losses was found to be very low (often at 1% base peak intensity), even at low fragmentation energy. Second, while 98% of all compounds in the NIST11 MS/MS library were reported using more than one fragmentation energy, the use of fragmentation energies was not systematically performed. Maximal collision energies were found at 20 V or less for 7% of all NIST11 MS/MS compounds, while 77% of the compounds were fragmented at maximal collision energy of 40 V or more. Such lack of comprehensive fragmentation MS/MS spectra might be one reason for the low sensitivities observed in the NIST11 MS/MS spectral library. Indeed, a loss of HCl is reported to be observed in MS^3 fragmentation more often than in MS/MS spectra of small aromatic molecules.[31] Conversely, by combining mass spectra from lower collision energy with higher collision energy, use of neutral loss queries for substructure annotation becomes optimal because neutral losses from weak bonds (such as water or ammonia) as well as stronger bonds (such as NO_2) can be investigated. The most important reason for low sensitivity of neutral loss queries for specific substructure classes might be the large overall structural complexity of small molecules. The diversity of bond strengths within a given compound means that not all expected neutral losses are necessarily observed, even if a specific substructure is present. For the above reasons, overall low sensitivities of the simple neutral-loss analysis for substructures were found when analyzing MS/MS spectra from the NIST11 library: there were far fewer true positive hits found than molecules that actually contained a given substructure.

Table 2. Sensitivity and Specificity of Predicting the Presence of Substructures from Common Neutral Losses Using the NIST11 MS/MS Library

compound class	neutral loss	mass (Da)	sensitivity (%)	specificity (%)
aliphatic primary amines	NH_3	17.027	69.5	85.5
carboxylic acids	H_2O	18.011	84.9	42.9
aldehydes	CH_2O	30.011	44.4	94.8
methyl esters	CH_3OH	32.026	72.0	95.6
thiol	H_2S	33.988	66.0	98.0
chlorides	HCl	35.977	15.4	93.2
N-acetyl derivatives	CH_2CO	42.011	69.0	89.7
nitroaromatics	NO_2	45.993	23.5	90.7
carboxylic acids	HCOOH	46.005	34.8	78.4
methyl sulfides	CH_4S	48.003	33.6	97.7
α,β-unsaturated acids	CH_3COOH	60.021	53.8	82.7
phosphate group	H_3PO_4	97.977	42.9	98.0
cysteine conjugates	$C_3H_7NO_2S$	121.020	33.0	98.2
hexoside	$C_6H_{10}O_5$	162.053	80.0	94.6

Investigation of Spectra-Substructure Relationships Based on Further MS/MS Features. For some substructures, presence can also be determined by formation of specific product ions. For example, presence of a phosphocholine headgroup in phosphatidylcholine (PC) lipids is reported by the product ion m/z 184.074.[32] Using an m/z window of 0.01

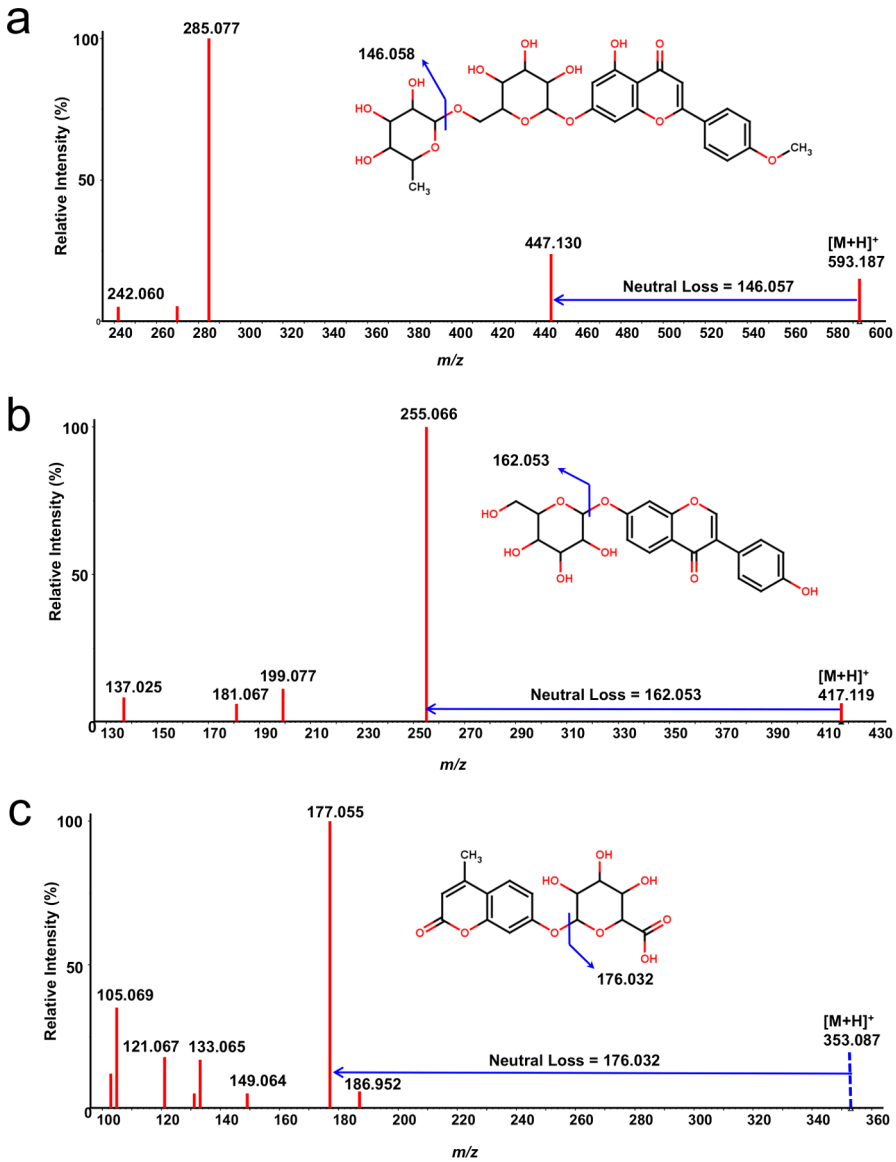


Figure 1. Annotations of glycosides and glucuronides by MS2Analyzer. (a) Neutral loss of anhydrodeoxyhexose in the MS/MS spectrum of Acaciin from MassBank: ID, PR100356; LC−ESI-QTOF; CE, ramp 5−60 V; $[M + H]^+$. (b) Neutral loss of anhydrohexose in MS/MS spectrum of Daidzin from MassBank: ID, PR100257; LC−ESI-QTOF; CE, ramp 5−60 V; $[M + H]^+$. (c) Neutral loss of anhydroglucuronic acid in MS/MS spectrum of 4-methylumbelliferyl glucuronide from MassBank: ID, BML00975; LC−ESI-QTOF; CE, 40 V; $[M + H]^+$.

and relative abundance threshold of 85% in MS2Analyzer, this product ion successfully annotated the presence of a phophocholine headgroup in 15 NIST11 compounds with a sensitivity of 100% and a specificity of 99.8%. Other substructures, such as the presence of iodine in 3-iodotyrosine and triiodothyronine, can be queried in a different way: such substructures are neither found as product ion nor as neutral loss but are detected at high specificity by analyzing the difference between product ions. Although the mechanism of this type of fragmentation is not well understood yet, it may be a loss of an iodine radical from a product ion, resulting in a second-stage neutral loss. A similar loss of an iodine radical is found in the process of photodissociation of iodinated proteins.[33] The specificity of detecting iodine atoms in molecular structures using the m/z difference 126.904 for all product ion pairs larger than 2% base peak intensity was found as 99.9% in the spectra obtained from the NIST11 library with a sensitivity of 80.0%. These examples show the usefulness of MS2Analyzer for investigating substructures. As we here report on the implementation and use of the MS2Analyzer software, it is beyond the scope of this work to present a comprehensive investigation of the accuracy of all substructure annotations, partly due to the lack of annotated MS/MS reference spectra and structural comprehensiveness in the NIST11 library.

Automatic Annotation of Glycosides from Large Spectral Collections. As further validation of the use of MS2Analyzer we used known and annotated spectra from a different mass spectral community repository, the public MassBank database from which we downloaded a total of 3 359 accurate mass MS/MS spectra ($[M + H]^+$ adducts) covering 860 compounds. All compounds were less than 1 000 u in molecular weight, except for one heptasaccharide. Most of the MassBank MS/MS spectra had been acquired using electrospray QTOF or Fourier transform ion cyclotron resonance (FTICR) mass spectrometers and fragmented under multiple collision energies. Neutral losses were searched

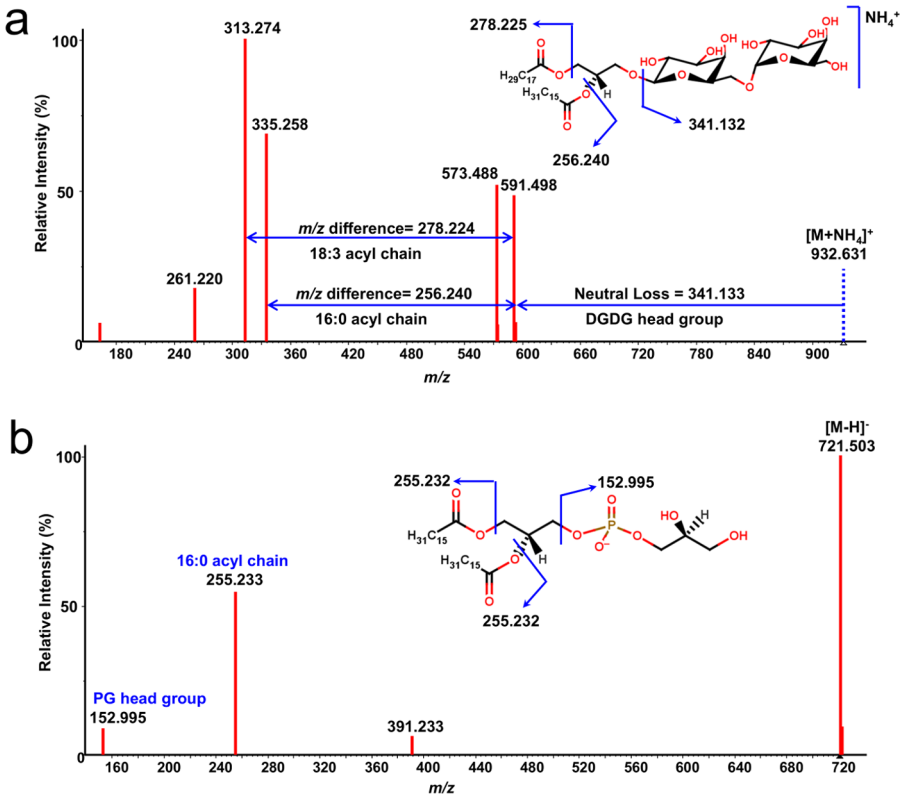


Figure 2. Annotations of lipids by MS2Analyzer. (a) QTOF MS/MS of digalactosyldiacylglycerol 18:3/16:0 in positive electrospray mode, indicating the neutral loss of the DGDG headgroup from the $[M + NH_4]^+$ adduct precursor ion as well as *m/z* differences for product ions indicating both acyl side chains. Note that positional isomers of acyl groups cannot be determined with this method. (b) QTOF MS/MS of phosphatidylglycerol 16:0/16:0 in negative electrospray mode, indicating the characteristic product ions of the phosphatidylglycerol headgroup (*m/z* 152.995) and the acyl side chains (*m/z* 255.233).

for the presence of three glycoside substructures[34] for which approximately four times as many examples were found in MassBank compared to NIST11 and which are also frequently present in many natural products. Statistical analysis of the MassBank MS/MS spectra using MS2Analyzer showed that a neutral loss of 146.058 Da annotates the presence of deoxyhexosides at a specificity of 99.4% with a sensitivity of 62.2%. One example of the spectra-structure relationship is shown in Figure 1a. Overall 28 deoxyhexoside structures in MassBank were correctly identified in this way. Correspondingly, a neutral loss of 162.053 Da annotates a loss of anhydrohexose, see Figure 1b, which is specific for hexosides (such as glucosides or galactosides) at a specificity of 99.8% and a sensitivity of 34.7% with 33 structures present in the MassBank database. A third example is given in Figure 1c demonstrating a neutral loss of anhydroglucuronic acid (176.032 Da) which was found at a specificity of 99.3% and a sensitivity of 85.7% for the presence of glucuronic acid conjugates in the MassBank data set. In summary, while the presence of hexosides, deoxyhexosides, and glucuronides can be positively deduced by the MS2Analyzer software with very high reliability if the corresponding neutral losses are experimentally observed, sensitivities for some neutral losses (such as anhydrohexose) were again low due to a high number of false negatives, similar to the analysis of other substructures queried in the NIST11 database.

Lipid Identification from *Chlamydomonas reinhardtii* LC−QTOF MS/MS Data. To showcase a practical application of the MS2Analyzer software for metabolomics research, we subsequently tested the software for finding complex lipids present in the model algae *Chlamydomonas reinhardtii* CC-125. A total of 9 244 positive mode and 3 277 negative mode MS/MS spectra were acquired by LC−QTOF and screened for lipid-specific mass spectral features such as product ions and neutral losses. A total of 17 mass spectral features for specific lipid head groups were collected from the literature (Table S-2 in the Supporting Information) and used as query text files in the MS2Analyzer for neutral loss and product ion searches. Besides, accurate masses of possible acyl chains were calculated and added to the query as neutral loss or product ion searches. All 12 521 MS/MS spectra were searched; due to the data dependent MS/MS fragmentation method (see method section), multiple MS/MS spectra were collected for many compounds. MS2Analyzer yielded 126 unique hits for the presence of algal lipid substructures. Each potential hit was verified by manual investigation using the NIST MS Search software. A total of 120 different lipids from the 126 precursors were positively identified comprising 13 different lipid classes. Monogalactosyldiacylglycerols (MGDG), digalactosyldiacylglycerols (DGDG), sulfoquinovosyldiacylglycerols (SQDG), diacyglyceryl-*N*,*N*,*N*-trimethylhomoserines (DGTS), lyso-DGTS, phosphatidylethanolamines (PE), lyso-PE, phosphatidylglycerols (PG), diacylglycerols (DG), and triacylglycerols (TG) were detected in positive mode electrospray ionization and PE, lyso-PE, and PG were detected in negative mode electrospray ionization (see Table S-3 in the Supporting Information). Two PE, two lyso-PE, and two PG were found at identical retention times in both positive and negative

electrospray ionization, but overall, positive and negative electrospray MS/MS data gave mostly complementary results, justifying the use of both ionization modes in algal lipidomics. Fatty acyl chains from 16:0 to 16:4 and 18:0 to 18:4 were found in most lipid classes; for triglycerides, fatty acyl rests were also found at 14:0, 17:0, and 20:0 carbon lengths. Retention times of compounds belonging to the same lipid classes were verified to increase by increasing acyl chain lengths and to increase by decreasing number of double bonds for lipids with equal carbon numbers. Figure 2a,b illustrates two examples of annotated MS/MS spectra in positive and negative mode.

For validation purposes, the identification results from MS2Analyzer were compared with the search results from LipidBlast using the software NIST MS PepSearch which can be easily used for lipid annotations once comprehensive lipidomics spectral databases are added. In total, using the LipidBlast library search alone without guidance by MS2Analyzer software, 88 unique lipids from 90 precursors were detected, including DGTS, PE, DG, and TG in positive mode and DGDG, PE, lyso-PE, and PG in negative mode (see Table S-4 in the Supporting Information for all the annotations by MS2Analyzer and LipidBlast). The relationship of the lipids annotated by MS2Analyzer and LipidBlast is shown in a Venn diagram in Figure S-1 in the Supporting Information. Among all the 126 annotated MS/MS spectra found by MS2Analyzer, only 63 were also annotated using LipidBlast. In fact, MS2Analyzer found 63 spectra that were not annotated by LipidBlast, e.g., MGDG, SQDG, lyso-DGTS and most DGDG. This apparently large difference in lipid annotations can be explained by the purpose and origin of these tools: LipidBlast was specifically limited in size and scope to reduce the number of potential false positives, for example, by excluding acyl chains such as 16:4 and 18:4 which are absent in mammalian systems. Although recently LipidBlast has been updated to include 16:4 and 18:4 acyl chains, some adducts are still missing, such as the $[M + NH_4]^+$ adduct of DGDG. On the other hand, LipidBlast found more DGTS and TG than MS2Analyzer due to its large number of spectra for these two lipid classes. This result showed that MS2Analyzer is especially useful for novel compound species that are not covered in tandem mass spectral databases yet. When both MS2Analyzer and LipidBlast searches were combined, overall 153 unique lipid precursors were identified in *C. reinhardtii*.

■ DISCUSSION

Modern UPLC−QTOF instruments can acquire tandem mass spectra with very high scan speeds up to 100 Hz. Depending on the length of the chromatographic run, thousands of spectra can be acquired in a single run. In untargeted metabolomic experiments, the number of unknown compounds exceeds by far the number of annotated known compounds even when using MS/MS queries of different libraries.[35] It is impractical to annotate thousands of unknown compounds that were missed by MS/MS library search in a manual way. Instead of ignoring this vast majority of MS/MS spectra of unknown compounds, MS2Analyzer may serve as a valuable tool to fill this gap by using a large corpus of previously published characteristic m/z fragment ions and neutral losses to perform (A) compound class annotations as we have shown with the case of glycosides and (B) single or novel compound annotations which we exemplified with complex lipids. The large fragmentation library can be easily adjusted and applied to other specific studies, such as lipidomics, environmental analytics, pharmaceutical analysis,[36] and plant metabolomics. In fact, public repositories have now started to collect MS/MS spectra of unknown compounds, such as the MassIVE and Global Natural Product Network (GNPS) Web sites (http://gnps.ucsd.edu/) or the metabolomic repositories including MetaboLights (www.ebi.ac.uk/metabolights) and NIH Common Fund sponsored Metabolomics Workbench (www.metabolomicsworkbench.org/).

For many classes of small molecules, the relationships and rules between the mass spectral features and substructure (or compound class) are well understood.[31,37] Such rules can be improved by studying experimental fragmentation patterns for each metabolite class. In case of annotations of lipids that are not covered by MS/MS libraries such as LipidBlast, Metlin, MassBank, or NIST11 (now extended as NIST14 database), the rule building process is work intensive and benefits from confirmation with orthogonal information such as retention times. Once the rules are established, they can be readily and repeatedly applied to a large number of LC−MS/MS runs in high-throughput mode. With a search speed of 40 000 spectra per minute, large spectral collections or chromatographic batches can be processed in high-throughput mode, without manual intervention.

Investigation of the impact of specific ionization sources, the ionization process (CID/HCD), different precursor adducts and the effect of different collision energies was beyond the scope of work for the presentation of the MS2Analyzer software. Because of the limited mass accuracy of spectra in NIST11 (0.01 u), substructures such as NO_2 (45.992 u) and HCOOH (46.005 u) cannot be distinguished. This example also shows the importance of high-resolution mass values; the traditional approach using unit masses for accessing neutral losses is not powerful enough in this case. Additional orthogonal information such as retention times should be included for compound annotations.[38] MS2Analyzer enables further investigations of the accuracy of substructure annotations by neutral losses, product ions, and product ion difference queries through the availability of large experimental accurate mass MS/MS libraries such as Riken ReSpect,[39] METLIN, MassBank, WILEY,[40] and NIST14.

■ CONCLUSIONS

MS2Analyzer is a JAVA based software program developed for large scale analysis of accurate mass LC−MS/MS data or collections of MS/MS spectra. The software was developed with the emphasis on the large number of unknown MS/MS spectra that are not readily annotated by MS/MS library search alone. With a search speed of 40 000 MS/MS spectra per minute it is especially suited for the analysis of output from modern instruments that operate with a high acquisition rate. MS2Analyzer allows users to search for precursor ions, product ions, neutral losses, and the analysis of diagnostic m/z differences between product ions. Additionally a query table of 147 specific accurate mass neutral losses and their associated formulas, names, and substructures is provided. The MS2Analyzer results are conveniently presented in Excel tables and allow for mass spectral feature-compound class annotations. The wide application domain of the software was exemplified with the automatic annotation of glycosides and for lipid identifications from *Chlamydomonas reinhardtii* extracts.

■ ASSOCIATED CONTENT

Supporting Information

Detailed methods of algae extraction and LC−QTOF measurement, full tables of neutral losses collected from the literature, full annotation tables of lipids with MS2Analyzer and LipidBlast library search, and a figure showing the relationship of lipids annotated with different methods. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ofiehn@ucdavis.edu. Phone: +1-530-752-9922.

Author Contributions

Y.M., T.K., and O.F. designed the experiment. Y.M. developed the software. D.Y. and C.L. performed experimental measurements. Y.M., T.K., and O.F. wrote the manuscript. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Funding for Y.M, T.K., and O.F. was supported by NSF Grant MCB 1139644, NSF Grant MCB 1153491, and Grant U24 DK097154. The authors acknowledge Masanori Arita's useful comments on the manuscript.

■ REFERENCES

(1) Schiesel, S.; Lämmerhofer, M.; Lindner, W. *Anal. Bioanal. Chem.* **2010**, *397*, 147−160.
(2) Haag, M.; Schmidt, A.; Sachsenheimer, T.; Brügger, B. *Metabolites* **2012**, *2*, 57−76.
(3) Dresen, S.; Gergov, M.; Politi, L.; Halter, C.; Weinmann, W. *Anal. Bioanal. Chem.* **2009**, *395*, 2521−2526.
(4) Stein, S. *Anal. Chem.* **2012**, *84*, 7274−7282.
(5) Kind, T.; Liu, K. H.; Lee do, Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. *Nat. Methods* **2013**, *10*, 755−758.
(6) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(7) 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*, 747−751.
(8) Kind, T.; Scholz, M.; Fiehn, O. *PLoS One* **2009**, *4*, e5440.
(9) Yoshida, H.; Leardi, R.; Funatsu, K.; Varmuza, K. *Anal. Chim. Acta* **2001**, *446*, 483−492.
(10) Lafaye, A.; Junot, C.; Gall, R. L.; Fritsch, P.; Ezan, E.; Tabet, J. C. *J. Mass Spectrom.* **2004**, *39*, 655−664.
(11) Barupal, D.; Kind, T.; Kothari, S.; Lee, D.; Fiehn, O. *BMC Biotechnol.* **2010**, *10*, 40.
(12) Fang, N.; Yu, S.; Badger, T. M. *J. Agric. Food Chem.* **2002**, *50*, 2700−2707.
(13) Sun, X.; Niu, L.; Li, X.; Lu, X.; Li, F. *J. Pharm. Biomed. Anal.* **2009**, *50*, 27−34.
(14) HighChem. In http://www.highchem.com/index.php/massfrontier.
(15) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
(16) Dührkop, K.; Scheubert, K.; Böcker, S. *Metabolites* **2013**, *3*, 506−516.
(17) Schwudke, D.; Oegema, J.; Burton, L.; Entchev, E.; Hannich, J. T.; Ejsing, C. S.; Kurzchalia, T.; Shevchenko, A. *Anal. Chem.* **2006**, *78*, 585−595.
(18) Herzog, R.; Schwudke, D.; Shevchenko, A. *Curr. Protoc. Bioinf.* **2013**, *43*, 14.12. 11−14.12. 30.
(19) Doerfler, H.; Sun, X.; Wang, L.; Engelmeier, D.; Lyon, D.; Weckwerth, W. *PLoS One* **2014**, *9*, e96188.
(20) Husen, P.; Tarasov, K.; Katafiasz, M.; Sokol, E.; Vogt, J.; Baumgart, J.; Nitsch, R.; Ekroos, K.; Ejsing, C. S. *PLoS One* **2013**, *8*, e79736.
(21) Rojas-Chertó, M.; van Vliet, M.; Peironcely, J. E.; van Doorn, R.; Kooyman, M.; te Beek, T.; van Driel, M. A.; Hankemeier, T.; Reijmers, T. *Bioinformatics* **2012**, *28*, 2707−2709.
(22) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinf.* **2010**, *11*, 395.
(23) Monroe, M. In http://www.alchemistmatt.com/mwtwin.html, 2012.
(24) Ihlenfeldt, W. D.; Bolton, E. E.; Bryant, S. H. *J. Cheminf.* **2009**, *1*, 20.
(25) Schomburg, K.; Ehrlich, H. C.; Stierand, K.; Rarey, M. *J. Cheminf.* **2011**, *3*, O12.
(26) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 1−14.
(27) NIST. In http://peptide.nist.gov/software/ms_search/MS_Search.html.
(28) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534−2536.
(29) Balogh, M. *Spectroscopy* **2004**, *19*, 34−40.
(30) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S. *Anal. Chem.* **2006**, *78*, 2113−2120.
(31) Levsen, K.; Schiebel, H. M.; Terlouw, J. K.; Jobst, K. J.; Elend, M.; Preiß, A.; Thiele, H.; Ingendoh, A. *J. Mass Spectrom.* **2007**, *42*, 1024−1044.
(32) Welti, R.; Li, W.; Li, M.; Sang, Y.; Biesiada, H.; Zhou, H.-E.; Rajashekar, C.; Williams, T. D.; Wang, X. *J. Biol. Chem.* **2002**, *277*, 31994−32002.
(33) Ly, T.; Julian, R. R. *J. Am. Chem. Soc.* **2008**, *130*, 351−358.
(34) Cabrera, G. M. In *Phytochemistry: Advances in Research* ; Research Signpost: Kerala, India, 2006; pp 1−22
(35) Sakurai, T.; Yamada, Y.; Sawada, Y.; Matsuda, F.; Akiyama, K.; Shinozaki, K.; Hirai, M. Y.; Saito, K. *Plant Cell Physiol.* **2013**, *54*, e5−e5.
(36) Tyrkkö, E.; Pelander, A.; Ketola, R. A.; Ojanperä, I. *Anal. Bioanal. Chem.* **2013**, *405*, 6697−6709.
(37) Hsu, F.-F.; Turk, J. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 352−363.
(38) Berendsen, B. J.; Stolker, L. A.; Nielen, M. W. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 154−163.
(39) Sawada, Y.; Nakabayashi, R.; Yamada, Y.; Suzuki, M.; Sato, M.; Sakata, A.; Akiyama, K.; Sakurai, T.; Matsuda, F.; Aoki, T. *Phytochemistry* **2012**, *82*, 38−45.
(40) Oberacher, H.; Whitley, G.; Berger, B. *J. Mass Spectrom.* **2013**, *48*, 487−496.