analytical chemistry

Article

pubs.acs.org/ac

MyCompoundID: Using an Evidence-Based Metabolome Library for Metabolite Identification

Liang Li,*[,†] Ronghong Li,[‡] Jianjun Zhou,[‡] Azeret Zuniga,[†] Avalyn E. Stanislaus,[†] Yiman Wu,[†] Tao Huan,[†] Jiamin Zheng,[†] Yi Shi,[‡] David S. Wishart,[‡,§] and Guohui Lin[‡]

[†]Department of Chemistry, University of Alberta, Edmonton, Alberta, Canada

[‡]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[§]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

Supporting Information

ABSTRACT: Identification of unknown metabolites is a major challenge in metabolomics. Without the identities of the metabolites, the metabolome data generated from a biological sample cannot be readily linked with the proteomic and genomic information for studies in systems biology and medicine. We have developed a web-based metabolite identification tool (http://www.mycompoundid.org) that allows searching and interpreting mass spectrometry (MS) data against a newly constructed metabolome library composed of 8 021 known human endogenous metabolites and their predicted metabolic products (375 809 compounds from one metabolic reaction and 10 583 901 from two reactions). As an example, in the analysis of a simple extract of human urine or plasma and the whole human urine by liquid chromatography-mass spectrometry and MS/MS, we are able to identify at least two times more metabolites in these samples than by using a standard human metabolome library. In addition, it is shown that the evidence-based metabolome library (EML) provides a much superior performance in identifying putative metabolites from a human urine sample, compared to the use of the ChemPub and KEGG libraries.

Metabolomics is a rapidly evolving field for studying biological systems and discovering potential disease biomarkers.[1,2] Advance in metabolomics is largely driven by the development of new analytical techniques, such as liquid chromatography mass spectrometry (LC−MS). However, metabolite identification remains a major analytical challenge.[3,4] The vast majority of spectral features observed in LC−MS cannot be assigned to known compounds.[5−7] This serious deficiency hinders the development of sophisticated bioinformatics tools for integrating the metabolome data with the proteome and transcriptome information for studies in systems biology and medicine. Clearly new tools for metabolite identification are urgently needed.

We report an MS−MS/MS approach for metabolite identification based on compound library searching. We have constructed an evidence-based metabolome library (EML) that is composed of the known published metabolites, as well as their possible metabolic products that are predicted by biotransformation reactions commonly encountered in metabolism. The predicted metabolites have indirect evidence of their potential existences in a given species as they are derived from the known metabolites and metabolic reactions. The rationale is that a known metabolite can be involved in various metabolic reactions in biological systems, producing different metabolic products. Some of them have been identified and documented with assigned chemical structures, while many others have not been identified. Our hypothesis is that, by including all of the possible metabolic products in the library, many unknowns that are structurally related to the known metabolites can potentially be identified using the MS-MS/MS approach.

This approach is illustrated by human metabolite identification. We used the 8 021 entries in the Human Metabolome Database (HMDB)[8] to create our EML. We then applied this EML for identification of metabolites present in human urine and plasma and demonstrated the possibility of identifying many more metabolites than the conventional approach of using the standard HMDB.

EXPERIMENTAL SECTION

Construction of EML. There are currently 8 021 entries in the Human Metabolome Database (HMDB).[8] We used these entries to create the evidence-based metabolome library. By an examination of the literature information,[9−15] we identified 76 commonly encountered metabolic reactions (Table S1 of the Supporting Information). This list of reactions is by no means complete; future release of the library will expand this list, by including other metabolic reactions deemed to be important. On the basis of these reactions, we did *in silico* biotransformation of the 8 021 known metabolites. Each reaction generates a product with the addition or subtraction of an expected group

Received: January 10, 2013
Accepted: February 1, 2013
Published: February 1, 2013

   dx.doi.org/10.1021/ac400099b | *Anal. Chem.* 2013, 85, 3401−3408

(e.g., +O in oxidation or −O in deoxidation) from the reactant, a known metabolite. Several possible structures of the product (isomers) could exist, but all with a characteristic mass shift from the added or subtracted group. The number of the new entries in the EML with one metabolic reaction is 375 809; some of the impossible transformations (e.g., −O from a metabolite containing no oxygen) were excluded during the construction of the library. Currently, there is also an option of generating the library with two metabolic reactions [e.g., a metabolite undergoes methylation ($+CH_2$) and then oxidation (+O), or a metabolite undergoes demethylation ($-CH_2$) and then oxidation (+O)], which produced a library with 10 583 901 entries.

Web Interface. To use the EML library for metabolite identification, we have developed a web-based search and data interpretation program called MyCompoundID (http://www.mycompoundid.org). In MyCompoundID (MCID), all known human endogenous metabolites are imported from the Human Metabolome Database and stored in a local MySQL database. These metabolites and their one- or two-reaction products are indexed using the molecular masses up to the millionth precision. The web server for MCID was constructed within Apache using Java and JavaScript to ensure the most efficienct and the largest platform compatibility. There are 76 commonly encountered metabolic reactions implemented in the web server, which accepts single and batch queries with zero, one, and two allowed metabolic reactions. The subtraction reactions were logically validated using the compound's MOL files. All query results were prepared for easier manual inspection that includes ChemDraw (CambridgeSoft, PerkinElmer, Cambridge, MA) or a ChemDraw Plugin. The web server interacts with the local computer to allow the users to exclude any output entry and to associate an output entry to any experimental evidence. Such postcurated query results can then be exported to a local archive. All these functions are enabled and efficiently executed in Java and JavaScript with extendibility for further development.

LC−MS. In one set of experiments, human urine and blood samples were analyzed using solid phase extraction to reduce the complexity of the metabolome, followed by LC−MS and LC−MS/MS analysis. In another set of experiments, human urine was analyzed directly by LC−MS and LC−MS/MS. LC−MS was performed on a 6220 oa time-of-flight (TOF) mass spectrometer (Agilent Technologies, Santa Clara, CA) equipped with a 1200 series High Performance Liquid Chromatography system (Agilent Technologies, Santa Clara, CA). LC−MS/MS was done on a 4000 QTRAP system (Applied Biosystems, Foster City, CA) also equipped with the Agilent 1200 HPLC system. More information on sample preparation and the LC−MS setup for analyzing the urine and plasma metabolites is given in the Supporting Information.

RESULTS AND DISCUSSION

Figure 1 shows the overall workflow of the MS-MS/MS approach using MyCompoundID for putative metabolite identification. Both MS and MS/MS spectra of a metabolome sample are generated using one or more high-performance mass spectrometers, such as Fourier transform (FT)-MS, time-of-flight (TOF)-MS, and quadrupole linear trap (QTrap) tandem MS. In MyCompoundID, the user enters a mass (either a single value or multiple values in batch mode) and a mass tolerance value determined by the mass accuracy of the instrument used and then selects the reaction number (0, 1, or 2). The program searches the EML to find any matches between library entries and the query mass within the defined mass tolerance. The number of mass matches

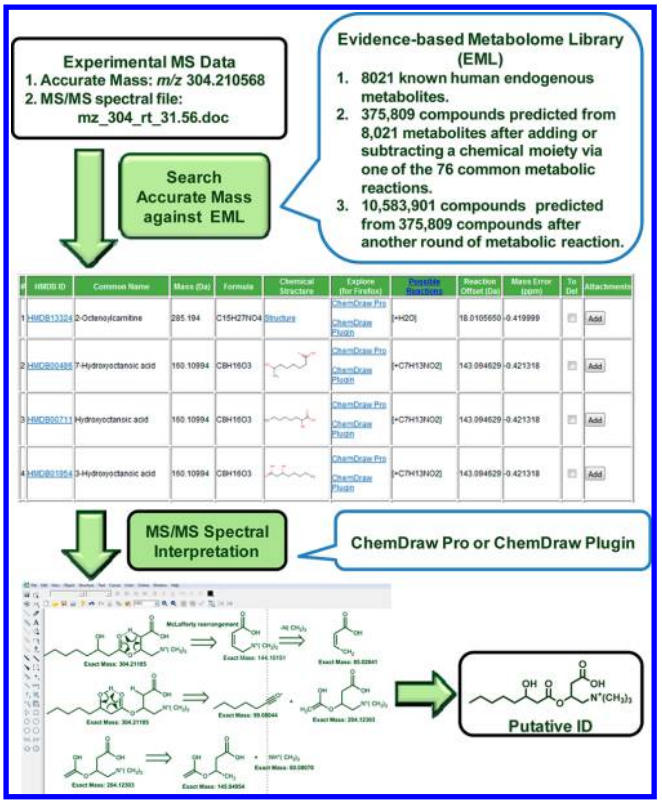

Figure 1. Workflow and main functionalities of MyCompoundID for tentative metabolite identification based on MS and MS/MS analysis of a sample.

for each query mass is listed in the summary panel. The search result is displayed in an interactive table and the matched entries can be sorted (e.g., based on the order of mass error).

One important functionality of the program is that the user can upload the chemical structure of the parent metabolite into ChemDraw or a free-ware ChemDraw Plugin. Both ChemDraw and ChemDraw Plugin allow the user to add or subtract a reaction group in the uploaded structure to create a new structure. Furthermore, the user can use the Mass Fragmentation program therein to break the chemical bond(s) to generate fragment ion structures and masses. With the use of the experimental MS/MS spectrum produced from the precursor ion of the query mass, the user can examine the spectral fragmentation pattern and compare it to the fragment ions generated by the Mass Fragmentation tool. If the pattern matches, putative metabolite identification can be made on the query mass. To document the identification process, all metadata, including the structure of the proposed match, the experimental MS/MS spectrum, fragment ion structures, fragmentation pathways, and any other documents (e.g., a Word file to describe the process), can be saved to the matched entry. Finally, the results can be exported to a spreadsheet for presentation and other uses. A tutorial for the use of the program and an example of the process described above are given in the Tutorial and Example of the Supporting Information, respectively.

To demonstrate the utility of MyCompoundID combined with our EML for metabolite identification, we have acquired LC−TOF-MS and LC−QTrap-MS/MS data from human urine and plasma. With the use of a simple extraction to capture a small fraction of the metabolome, LC−TOF-MS and LC−QTrap-MS/MS detected 17 969 and 2 316 features, respectively, in urine and 5 761 and 2 247 features, respectively, in

dx.doi.org/10.1021/ac400099b | Anal. Chem. 2013, 85, 3401−3408

plasma. Out of these features, we extracted 347 peaks in urine and 116 in plasma that were commonly detected by TOF-MS, QTrap-MS, and QTrap-MS/MS. The common individual peaks detected by both methods had similar retention times. In other words, 347 peaks had the accurate masses measured by TOF-MS and their corresponding MS/MS spectra collected by QTrap-MS. A large fraction of the features or peaks detectable in LC−MS did not generate good quality MS/MS spectra, which is consistent with the other metabolic profiling studies.[16−18] The origins of the peaks from which no MS/MS were generated are unknown but may be from several sources. Some of them might be from low abundance metabolites; their ion intensities were too low to produce MS/MS spectra with good signal-to-noise ratios. Others might be belonging to the metabolite ions that are not readily fragmented by collision-induced dissociation (CID). The formation of metabolite salt adduct ions could also contribute to the peaks detected, but the salt-adduct ions tend to lose the metal ions, instead of backbone dissociation, resulting in no structural information. Many of the features might also be from the impurities that are difficult to fragment by CID. A more detailed characterization of the number of ions selected for MS/MS versus the number of MS/MS spectra obtained will be discussed in another example described below (i.e., whole urine metabolite analysis).

The 347 metabolite peaks with each having both the accurate mass and MS/MS spectrum are listed in Tables S2 and S3 of the Supporting Information. To identify these metabolites, we first searched the HMDB using the accurate masses (<5 ppm) and MS/MS spectra against a library of about 900 metabolite standards.[8] MS/MS spectral matches were manually checked to ensure most of the fragment ions observed from the unknown metabolite MS/MS spectrum were matched with those of the standard library spectrum. Because of different experimental conditions used for generating the standard library spectra (Water's triple-quadrupole MS) and the unknown MS/MS spectra (AB Sciex's Qtrap-MS), a spectral match used in this work only lead to putative identification of the metabolite. Only eight metabolites were matched in urine and seven in plasma (see Tables S4-1 and S5-1 of the Supporting Information). This low rate of success reflects the current status of metabolite identification by LC−MS (i.e., many peaks detected cannot readily be identified using the current database resources).[3,8,19−21]

Next, we used MyCompoundID to search the accurate masses of the remaining features against the 8 021 known metabolites (i.e., EML with reaction = 0) to generate a list of mass matches, followed by MS/MS spectral interpretation of individual matches. We putatively identified 14 metabolites in urine and 34 in plasma (see Tables S4-2 and S5-2 of the Supporting Information and their corresponding Evidence Folders detailing the spectral interpretations of the matches).

We then used MyCompoundID to search the accurate masses of the remaining features against EML with one reaction. In conjunction with MS/MS spectral interpretation, we putatively identified 41 metabolites in urine and 14 in plasma (see Tables S4-3 and S5-3 of the Supporting Information). The use of EML with two reactions only led to the putative identification of three more metabolites in urine and none in plasma (Table S4-4 of the Supporting Information). This low rate of identification was mainly due to the presence of many possible structures for each matched mass, resulting in difficulty in manual spectral interpretation of the structure assignment. Development of an automated spectral interpretation program in the future will likely facilitate metabolite identification using EML with two or more reactions. Nevertheless, using MyCompoundID, we putatively identified a total of 58 additional metabolites in urine and 48 in plasma, compared to 8 and 7 metabolites identified using the standard compound library, respectively. These examples illustrate that MyCompoundID can significantly increase the number of metabolites identifiable from biofluids. Note that, in this illustrative work, we used the hydrophilic−lipophilic balanced reversed-phase (HLB) cartridge to capture a selected number of metabolites and used the positive ion mode in LC−MS to analyze these metabolites. Thus, the total numbers of metabolites putatively identified were relatively small.

For the urine and plasma samples analyzed, MyCompoundID also allowed the identification of several interesting fragmentation patterns from the metabolite peaks that are likely from exogenous metabolites. These include 14 peaks (13 in urine and 2 in plasma with 1 common peak) identified as poly(ethylene glycol) (PEG) derivatives, which are common additives in processed food, drug formulation, toothpaste, eye drops, etc. (see Table S6-1 of the Supporting Information). Several peaks showed characteristic fragmentation (see Tables S6-2−S6-4 of the Supporting Information). We did the blank runs to confirm that these PEG derivatives were from the samples and not the contaminants introduced in the sample processing and analysis steps. It appears that PEG derivatives were consumed or absorbed by the individuals, resulting in the detection of these compounds in biofluids. There were 16 unknown metabolites in urine containing glucuronides (Table S6-5 of the Supporting Information); exogenous metabolites often form this type of derivative. In total, 45 exogenous metabolites in urine and 6 in plasma (including 4 cocodiethanolamides) were found; more metabolic products found in urine than plasma is consistent with the notion that many more metabolites are excreted in urine.

Another example of using MyCompoundID for improving putative metabolite identification is the analysis of a human urine sample without solvent extraction. In this case, LC−TOF-MS and LC−QTrap-MS/MS spectra were collected from the urine sample. In the LC−QTrap-MS/MS experiment, 2 210 ions were selected for MS/MS based on their signal intensities determined in the MS survey scan; the top two most intense ions were selected for MS/MS after each mass scan. The redundant ions within a mass tolerance of 0.2 Da and a retention time tolerance of 0.5 min were grouped, resulting in the identification of 630 unique ions with MS/MS spectra. Among these 630 ions, 73 ions were found to be the sodium adduct ions, based on their coappearance of the precursor ions in the MS spectra as that of the protonated ions and their related fragmentation patterns. The remaining 557 ions could be broadly classified into two groups. The first group consists of 150 ions with low quality MS/MS spectra (noisy peaks with low S/N ratios or no fragment ion peaks detectable), and these ions were generally from those with a m/z of less than 110. They were likely from impurities or low mass metabolites that were in low abundance or suppressed during ESI due to low mass background ion interferences (e.g., solvent or solvent adduct ions). The second group consists of 407 ions with m/z of greater than 110, which resulted in MS/MS spectra with definable fragment ion peaks. Among the second group ions, 369 ions with their precursor ions matched with those detected in the LC−TOF-MS experiment. There were 38 ions with MS/MS spectra that did not have the corresponding precursor ions detected in TOF-MS, which can be attributed to the difference in ion detectability of the two instruments; these 38 ions were likely suppressed in the TOF-MS run. While optimizing the

MS/MS spectral collection was not the objective of this current work, it is apparent that future work for identifying more metabolites from a biofluid requires a careful optimization of the experimental setup. For example, the use of multidimensional separation and different modes of MS/MS spectral acquisition conditions should result in an increased number of MS/MS spectra collectable from a biofluid.

The above results indicate that there were 369 MS/MS spectra collected by QTrap-MS/MS, with each spectrum having a corresponding accurate mass of the precursor ion determined by TOF-MS. By searching the HMDB library, 23 metabolites were putatively identified based on the matches of both accurate mass of the precursor ion and MS/MS spectrum of an individual unknown metabolite to those in the database (see Table S7-1 of the Supporting Information). For the remaining accurate masses with no MS/MS matches, we used MyCompoundID to search these masses against EML with no reaction to generate a list of mass matches. The MS/MS spectra of these mass matches were manually interpreted to arrive at putative identities of 53 metabolites (Table S7-2 of the Supporting Information). Next, we searched the remaining accurate masses against EML with one reaction. With accurate mass matches and manual MS/MS spectral interpretation, we putatively identified another 87 metabolites (Table S7-3 of the Supporting Information). In total, 163 metabolites were putatively identified from the urine sample analyzed by reversed-phase LC–MS and MS/MS. This example again demonstrates that MyCompoundID with the expanded library can be used to identify more putative metabolites from a biofluid.

MyCompoundID relies on the use of an accurate mass search to arrive at a list of possible mass matches in EML and then uses manual MS/MS spectral interpretation against this list of structures to arrive at a putative identification of the matched mass. We recognize that manual MS/MS spectral interpretation is a subjective process and its success depends on the researcher's experience which can be gained by interpreting MS/MS spectra of known metabolites, such as those in the HMDB MS/MS spectral library. In our work, the following general guidelines were followed in interpreting the MS/MS spectra against the proposed chemical structure to determine whether a putative identification has been made.

First, for a simple structure with a few breakable bonds, one to three fragment ions matched with the structure may be sufficient to call a putative identification. This is illustrated in the example shown in Figure 2, where acetylcarnitine was identified based on the three fragment ions detected in the MS/MS spectrum (i.e., m/z at 60, 85, and 144). The proposed fragmentation scheme and the fragment ion structures are also shown in Figure 2; the m/z 60 fragment ion is most likely from $(CH_3)_3NH^+$. Second, for a more complex structure with a number of breakable bonds, a large fraction (more than 75%) of the major fragment ions should match with the proposed structure. One example is shown in Figure 3. In this case, a structure of [3-hydroxyanthranilic acid + NH] (from EML with one reaction) was identified based on the fragment ions at m/z 152, 141, 125, 107, 95, 80, and 68. The putative structure and its fragmentation scheme for generating these fragment ions are shown in Figure 3. Third, if a MS/MS spectrum of a similar structure of known metabolite is available, the fragmentation pattern of the unknown metabolite related to the known structure should be similar. This is shown in Figure 4 for the putative identification of indolylacryloylglycine based on the fragmentation pattern of indoleacrylic acid. The MS/MS spectrum of indolearylic acid from HMDB obtained using a triple quadrupole MS is shown in Figure S1 of the Supporting Information, and the MS/MS spectrum obtained from the unknown metabolite is shown in Figure 4. Although different instruments were used for generating the MS/MS spectra, similar patterns were observed with the major difference in the precursor ion masses. Finally, common fragment ions generated from the same class of metabolites can be used to assist in determining a putative metabolite structure of an unknown belonging to this class. Figure 5 shows two examples where the structures of 2-trans,4-cis-decadienoylcarnitine (or isomers) and [indoleacetic acid + $C_7H_{13}NO_2$] were proposed,

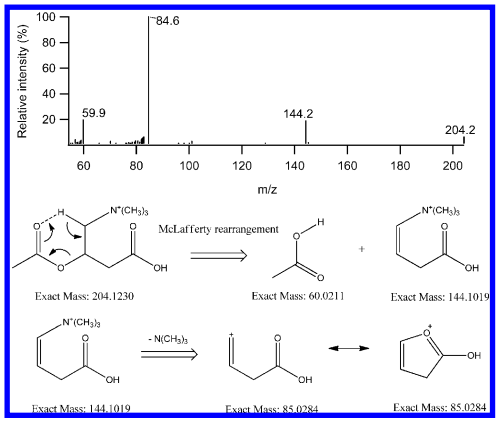

Figure 2. MS/MS spectrum of acetylcarnitine and the proposed fragmentation scheme to assign the major fragment ions.

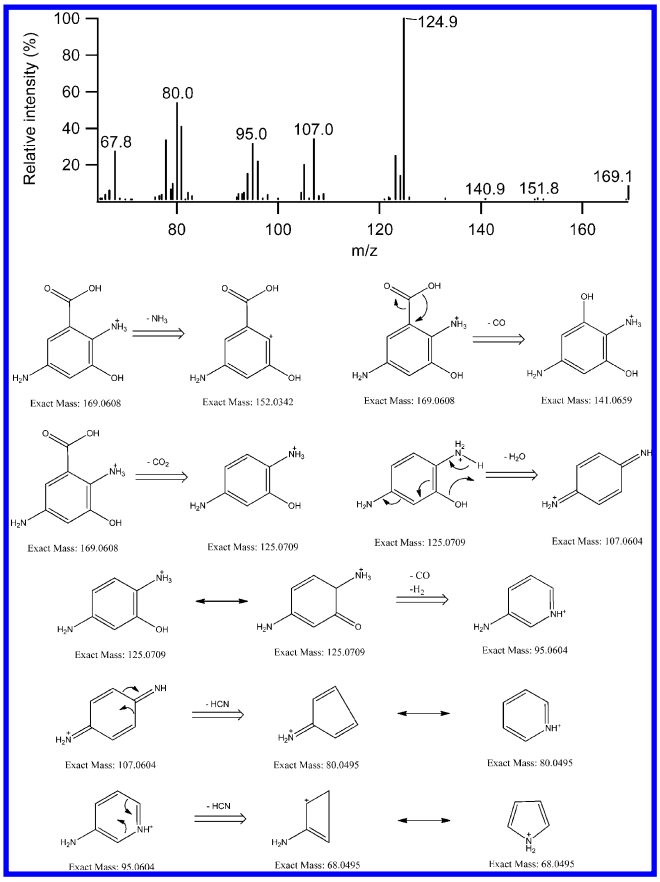

Figure 3. MS/MS spectrum of a proposed structure of [3-hydroxyanthranilic acid + NH] and the proposed fragmentation scheme.

dx.doi.org/10.1021/ac400099b | Anal. Chem. 2013, 85, 3401–3408

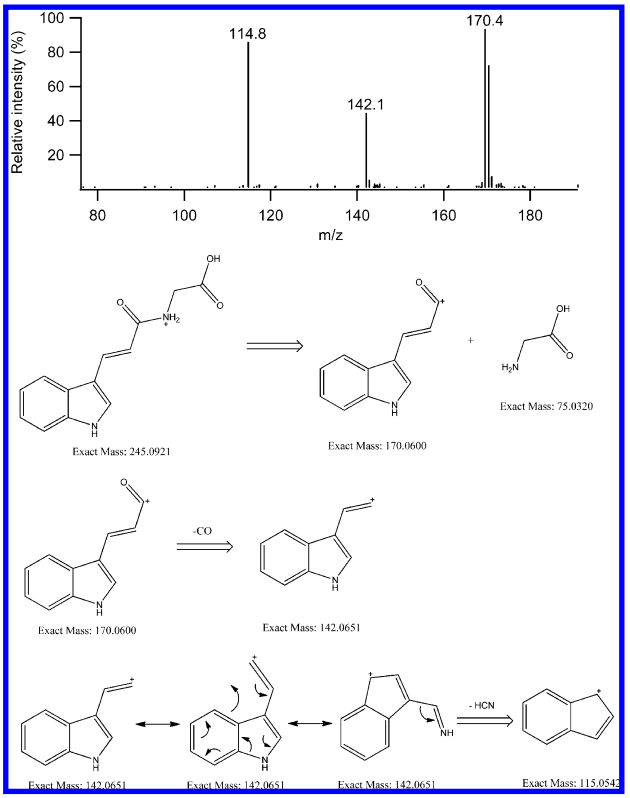

Figure 4. MS/MS spectra of a putative metabolite, indolylacryloylglycine, and the proposed fragmentation scheme. The MS/MS spectrum of indolearylic acid from HMDB is shown in Figure S1 of the Supporting Information.

based on the common fragment ions of m/z 85 and 144 to those found in other carnitines, along with other major fragment ions detected. The proposed fragmentation schemes for these two molecules are shown in Figure S2 of the Supporting Information.

In the whole urine sample analysis, out of 369 accurate masses, 104 masses (28%) could be matched with one or more compounds in EML with no reaction, while 151 masses (41%) matched with one or more compounds in EML with one reaction. The rest (31%) could not match with any entries. The MS/MS spectra of these nonmatches could not be used for structural assignments, as we had no clue about the possible structures of these precursor ions. Out of 255 MS/MS spectra collected from different precursor ions with mass matches to EML, 163 putative identifications (64%) were made, and the remaining 92 MS/MS spectra (36%) could not be assigned to any chemical structures with high confidence. There may be several reasons for not being able to assign the MS/MS spectra to any structures in the library. One reason is that some of the spectra contained only a few fragment ion peaks from which a chemical structure could not be assigned. We did not overassign the MS/MS spectra, i.e., if an MS/MS spectrum did not contain a sufficient number of informative fragment ions (see the manual interpretation guidelines described above) to allow the assignment of the spectrum to one structure (stereoisomers were considered as one structure, unless they were separated by LC), we considered this MS/MS spectrum unassignable. Another reason is that the compound library is not composed of all the metabolites potentially present in a biofluid; many metabolites are yet to be discovered. MS/MS spectra collected from the metabolites not in the library could not be assigned.

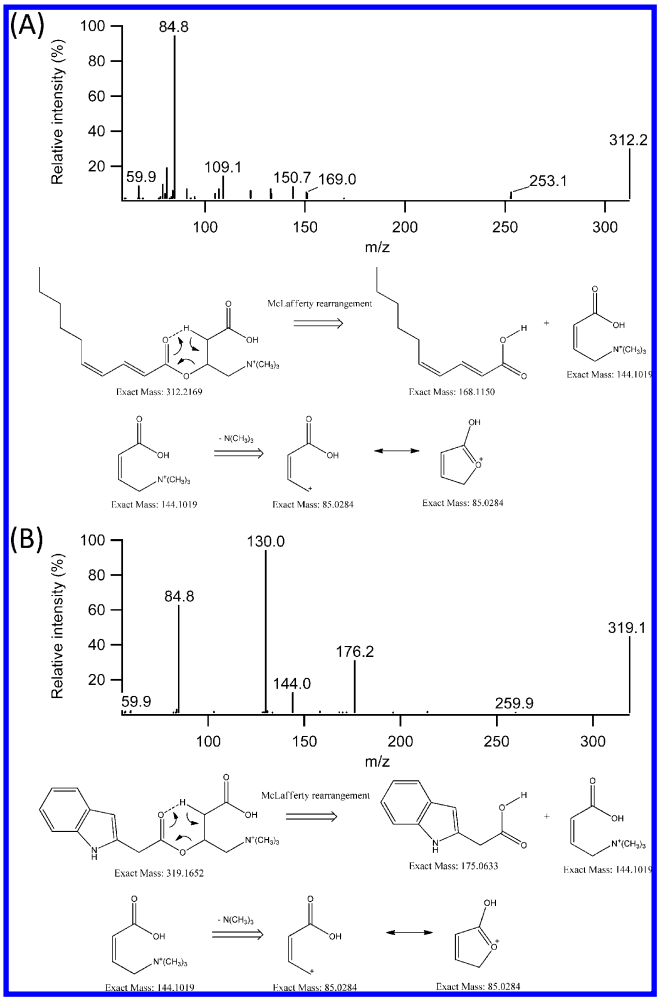

Figure 5. MS/MS spectra of (A) a putative metabolite, 2-trans,4-cis-decadienoylcarnitine (or isomers) and (B) a putative metabolite of [indoleacetic acid + $C_7H_{13}NO_2$] and the fragmentation schemes showing the formation of the fragment ions at m/z 85 and 144. The fragmentation schemes for the formation of other major fragment ions are shown in Figure S2 of the Supporting Information.

Aside from the lack of informative fragment ions in MS/MS spectra and the incomplete human metabolite library, issues related to the mass spectrometric detection may also contribute to the unassignment of some of the MS/MS spectra collected. Some of them might be from the product ions of the intact metabolite generated in the source region due to in-source fragmentation. In our work, both TOF-MS used for accurate mass measurement and Qtrap-MS used for MS/MS were operated at the low source voltages to minimize in-source fragmentation. But, even under these conditions, in-source fragmentation can still occur, leading to the possibility of picking an in-source fragment ion, instead of the intact molecular ion, for MS/MS. However, the product ions from the in-source fragmentation of the intact metabolite ion will have the same retention time as that of the intact ion. In other words, the product ion peak along with the molecular ion peak of the metabolite will be detected in the same spectrum at a given retention time. Thus, at any given retention time, if two structurally related compounds are identified by using MS and MS/MS, there is a strong possibility that the lower-mass compound is the ion-source fragmentation product of the intact molecular ion. For all the putative metabolites identified in this work, we

dx.doi.org/10.1021/ac400099b | Anal. Chem. 2013, 85, 3401−3408

Table 1. Number of Matches in Three Chemical Libraries Using the Measured Accurate Mass within an Error Tolerance of 5 ppm

feature ID no.	accurate *m/z* TOF	RT (min) TOF	no. of matches in EML	no. of matches in PubChem[a]	no. of matches in KEGG[a]
1	132.04485	26.14	2	65 (1)	0
2	137.04566	4.34	10	240 (2)	1 (1)
3	169.06012	7.67	19	856	1
4	170.05994	41.02	2	162 (1)	0
5	171.08798	4.88	2	532	1
6	177.05490	34.56	8	559 (1)	3
7	188.10233	3.94	7	348	1
8	190.10692	30.98	27	1266 (3)	5
9	195.11209	24.74	26	4988	2
10	197.12787	30.22	9	3886 (3)	1
11	202.11811	4.94	3	557	0
12	203.08099	6.99	9	2429	2
13	209.11683	52.99	2	4256 (1)	9
14	220.06020	33.55	6	918	2
15	222.07874	7.84	19	1667 (2)	0
16	224.12778	51.06	1	6350	3
17	226.08146	3.91	20	889	0
18	257.14883	37.43	7	2769	0
19	257.22630	52.22	6	231 (1)	0
20	257.22609	63.13	6	215 (1)	0
21	262.16460	8.35	20	375	0
22	262.16451	9.48	20	375	0
23	263.13844	33.65	5	9967 (7)	2
24	266.10272	35.79	10	2835 (1)	0
25	266.13754	32.38	3	388 (6)	0
26	269.12310	3.77	10	1194 (2)	1
27	272.18452	39.54	5	1718	0
28	272.18520	38.38	5	1703	0
29	272.18524	36.94	5	1703	0
30	273.22071	53.94	26	1022 (4)	3
31	273.22020	57.49	26	1023 (4)	3
32	284.18510	39.46	3	1820	2
33	284.18555	38.64	3	1812	2
34	285.25733	55.60	1	179 (3)	0
35	286.12723	8.27	8	2717	0
36	287.19964	48.15	33	1949 (5)	0
37	300.21662	45.10	6	758 (1)	0
38	300.21694	46.77	6	1174 (1)	0
39	302.16082	24.06	5	1975	0
40	302.19588	39.65	9	480	0
41	302.19534	29.32	9	480	0
42	302.19545	34.78	9	481	0
43	302.19578	30.37	9	480	0
44	303.10150	33.18	5	3741	1
45	304.21128	38.27	8	156 (2)	0
46	304.21162	39.05	8	156 (2)	0
47	310.20084	44.79	4	1180 (1)	1
48	310.20119	46.31	4	1753 (1)	1
49	316.17502	30.22	2	559	0
50	316.21096	37.63	4	374	1
51	318.19080	33.01	6	291	0
52	318.20682	45.82	10	2327	3
53	319.16510	35.96	3	4843	0
54	326.08550	7.45	15	2432	2
55	328.21076	39.08	4	391	0
56	328.21106	35.80	4	323	0
57	328.21115	41.02	4	321	0
58	328.24800	52.37	12	529	0
59	328.24724	53.47	12	279	0
60	330.19066	33.94	2	434	1
61	330.22674	42.76	6	185	0
62	332.20606	36.02	5	213	0
63	332.24218	39.61	6	101	0
64	332.24222	46.58	6	101	0
65	337.17543	36.13	5	2257	0
66	341.16967	33.93	12	1898	2
67	342.22719	39.92	3	1611	0
68	344.20550	30.23	1	492	0
69	344.20592	38.18	1	442	0
70	346.12557	4.48	25	1409	1
71	346.22071	32.12	4	225	0
72	356.24275	47.89	4	1117	0
73	358.25797	50.86	12	130	0
74	365.20244	43.57	2	942	0
75	367.11656	17.30	6	1680	1
76	384.11503	26.09	12	1845 (4)	0
77	384.27380	54.60	11	177	0
78	402.28322	44.54	5	107	0
79	413.04257	29.98	1	593	1
80	432.31012	55.26	17	415	0
81	448.22108	51.49	2	4617	0
82	464.19128	46.63	4	2302	0
83	523.25271	41.26	12	1078 (1)	1
84	593.33318	60.55	6	414 (1)	2 (1)
85	595.34814	60.24	3	294 (2)	1
86	626.20660	40.17	7	133	0
87	642.34684	43.03	6	1815 (2)	0

[a] *m*(*n*):*m* is the number of matches to the library; *n* is the number of matches showing the same structure (including isomers) as that of the putative metabolite identified by MyCompoundID using EML.

manually went through the accurate mass, retention time, and MS/MS fragmentation pattern of each putative metabolite to confirm that the proposed structures of the putative metabolites were indeed from the intact molecules, not the in-source fragments of the intact molecules. Nevertheless, some of the unassigned MS/MS spectra might be from the in-source fragment ions, instead of the intact molecular ions. Because the current version of MyCompoundID uses the accurate mass match to the intact metabolites in the library as the first pass to generate a list of compounds for further MS/MS spectral interpretation, any fragment ion mass may lead to a completely different set of metabolites, resulting in the unassignment of the MS/MS spectrum generated from the in-source fragment ion.

As indicated earlier, the current EML consists of 8 021 human metabolites from HMDB and 375 809 predicted metabolic products from one metabolic reaction. In comparison, PubChem (http://pubchem.ncbi.nlm.nih.gov/), the largest compound library, has over 100 million entries, while another more biologically relevant library, KEGG (http://www.genome.jp/kegg/), has 16 907 low molecular mass compounds. One major difference of EML from the PubChem and KEGG libraries is that EML is composed of mainly human endogenous metabolites

dx.doi.org/10.1021/ac400099b | *Anal. Chem.* 2013, 85, 3401−3408

and their predicted metabolic products, while the other two libraries contain all sorts of chemicals, including synthetic compounds. In principle, we can also use the accurate mass detected from a metabolite to search against the PubChem and KEGG libraries to generate a list of mass matches with chemical structures provided in these libraries. We can then interpret the MS/MS spectrum of the metabolite against these chemical structures to arrive at a putative identification. To compare the results obtained from the three libraries, we took the accurate masses of 87 putative metabolites identified from EML with one reaction using MyCompoundID and searched them against the PubChem and KEGG libraries, using the same mass error windows (5 ppm). The search results are summarized in Table 1, where the number of matches for each mass is listed (these numbers are also listed in the last three columns in Table S7-4 of the Supporting Information).

As Table 1 shows, using the PubChem library, most of the 87 masses can match with hundreds or thousands of compounds per mass. The range of matches is from 65 to 9 967 with a median of 758 and an average of 1339 matches. Among the mass matches found from each mass search, the number of structures, including isomers, that are the same as the proposed structure already determined by MyCompoundID with EML is also listed in Table 1. In total, 29 out of 87 masses (33%) have at least one structure or structural isomer matched with the proposed structure or "correct" match, while the rest (67%) do not have any "correct" matches. Note that, even for those with "correct" matches, the number of entries from the accurate mass match alone is very high (hundreds or thousands of entries). Thus, it would take a long time to go through all these structures to find the one structure that best matches with the MS/MS fragmentation pattern. The search results shown in Table 1 indicate that, using the KEGG library, only 32 out of 87 masses match with one or more entries: 1 mass matches with 9 compounds, 1 mass matches with 5 compounds, 5 masses with each matches with 3 compounds, 9 match with 2, and 16 match with only 1 compound. Among these mass matches, only 2 masses (~2%) have the "correct" matches with the structures proposed by MyCompoundID.

We can make another comparison by assuming that the EML putative identifications are incorrect and any matches to the PubChem library compounds may be correct or better than those deduced from EML. Since MS/MS spectral interpretation against a small number of structures is still a manageable task, we have examined the mass matches of 5 accurate masses listed in Table 1 that have less than 133 possible structures matched to the PubChem library (ranging from 65 to 133 matches). In all cases, we could not find any matches or better MS/MS spectral matches to the PubChem compounds, compared to those in EML. All the compounds with mass matches to the measured metabolites are exogenous compounds and most of them are synthetic compounds not expected to be present at a detectable quantity in normal human urine samples. From all of these comparisons discussed above, we can conclude that EML is more useful than ChemPub and KEGG for putative human metabolite identification from human biofluids. These comparison results also indicate that the size of a compound library does not determine the outcome of a putative metabolite identification exercise; the quality of the compound library is very important. A huge compound library consisting of mainly synthetic compounds did not yield better results than EML for human biofluid metabolite identification.

It should be noted that MyCompoundID only allows the user to putatively identify a metabolite based on the match of accurate molecular mass and matches of fragment ions detected in MS/MS to the proposed structure. Neither does it provide any quantitative gauge of the confidence level for each putative identification. For the accurate mass matches using EML, the number of matches is provided in the summary panel in the form. As expected, with the expanded library, more matches become possible. Thus, MS/MS spectral interpretation against these mass matches become very important to generate a putative structural assignment. At this stage, there is no quantitative means of gauging the confidence of this manual interpretation. However, this exercise can narrow down the list of metabolite candidates into one or a few unique structures. If positive identification is required (e.g., a potentially useful biomarker of a disease after comparative metabolome profiling of diseased group and healthy controls), authentic standard(s) may be synthesized for comparison. Reducing the number of possible metabolite candidates by this combination of mass search and MS/MS interpretation, or MS − MS/MS, with EML would save time and effort, as only a few standards need to be made. In cases where the standards of putatively identified metabolites are difficult to synthesize, the use of microsome- or other cell/tissue-based biotransformation of structurally related standards[22] may be explored to produce the needed standards for metabolite validation. A series of analytical tools including purification and NMR structural characterization may also be used for metabolite identification.

Finally, we note that MyCompoundID is a unique resource that can be expanded in the future in terms of both the size of the compound library and search functionality. The current library consists of all the HMDB metabolites and their biotransformation-predicted metabolites via the 76 metabolic reactions. The use of these 76 reactions was based on the literature survey of the known metabolic reactions commonly encountered; we went through the published literature, such as the references cited,[9−15] and selected the ones commonly found. This list will be expanded in the future and, since MyCompoundID is a public resource, we welcome any feedback from the community on the type of reactions that should be included. In addition, the creation of EML opens the possibility for future work in generating theoretical MS/MS spectra of the predicted metabolites based on the chemical structures. This type of spectral library would be very useful for MS/MS spectral search to increase the speed of putative metabolite identification, compared to the current approach of manual interpretation of an MS/MS spectrum against a list of mass-matches structures. Work in this direction is currently underway.

■ CONCLUSIONS

In summary, we have developed a publicly accessible web-based tool that can facilitate the identification of unknown metabolites in metabolome profiling. In combination with LC−MS, it is shown to be useful for identifying many more metabolites in human urine and blood samples than using a standard library. This MyCompoundID tool features a dynamic compound library that can be expanded in the future by inclusion of the metabolites and their predicted metabolic products from different origins, including human, microbe, plant, food, drugs, etc. We envisage that an expanded compound library will increase the number of metabolites identifiable from human biofluids and open the possibility of using MyCompoundID for analyzing the metabolomes of other species. We also plan to add the functionality for data sharing

dx.doi.org/10.1021/ac400099b | Anal. Chem. 2013, 85, 3401−3408

among the researchers who are interested in chemical identification (e.g., deposition of MS/MS spectra and their interpretation and spectral assignment for newly identified compounds). We recognize that manual interpretation of MS/MS spectra is a time-consuming process. We are currently in the process of developing a strategy to semiautomate the MS/MS spectral interpretation process that will be incorporated into MyCompoundID in the near future.

■ ASSOCIATED CONTENT

Supporting Information

Methods: Experimental conditions used for sample preparation, LC−TOF-MS and LC−QTrap-MS/MS. Tutorial: Tutorial for the use of MyCompoundID. Example: An example of the metabolite identification process using MyCompoundID. Table S1: List of 76 commonly encountered metabolic reactions. Table S2: List of metabolite peaks detected in a human urine sample by LC−MS. These peaks were detected by both TOF and QTrap MS as well as MS/MS from QTrap. Table S3: List of metabolite peaks detected in a human plasma sample by LC−MS. These peaks were detected by both TOF and QTrap MS as well as MS/MS from QTrap. Table S4: Summary of metabolites putatively identified from an extract of urine sample. Table S5: Summary of metabolites putatively identified from an extract of plasma sample. Table S6: List of likely exogenous metabolites found in the urine and plasma extracts. Table S7: Summary of metabolites putatively identified from the whole urine sample. Evidence Folders for Table S4: Metadata as well as MS/MS spectral interpretation. Evidence Folders for Table S5: Metadata as well as MS/MS spectral interpretation. Evidence Folders for Table S6: Metadata as well as MS/MS spectral interpretation. Evidence Folders for Table S7: Metadata as well as MS/MS spectral interpretation. Figures S1 and S2: MS/MS spectrum and proposed fragmentation schemes. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Liang.Li@ualberta.ca.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Research Chairs program, Genome Canada, and Genome Alberta.

■ REFERENCES

(1) Sreekumar, A.; Poisson, L. M.; Rajendiran, T. M.; Khan, A. P.; Cao, Q.; Yu, J. D.; Laxman, B.; Mehra, R.; Lonigro, R. J.; Li, Y.; Nyati, M. K.; Ahsan, A.; Kalyana-Sundaram, S.; Han, B.; Cao, X. H.; Byun, J.; Omenn, G. S.; Ghosh, D.; Pennathur, S.; Alexander, D. C.; Berger, A.; Shuster, J. R.; Wei, J. T.; Varambally, S.; Beecher, C.; Chinnaiyan, A. M. *Nature* **2009**, *457*, 910−914.

(2) Jenkins, H.; Hardy, N.; Beckmann, M.; Draper, J.; Smith, A. R.; Taylor, J.; Fiehn, O.; Goodacre, R.; Bino, R. J.; Hall, R.; Kopka, J.; Lane, G. A.; Lange, B. M.; Liu, J. R.; Mendes, P.; Nikolau, B. J.; Oliver, S. G.; Paton, N. W.; Rhee, S.; Roessner-Tunali, U.; Saito, K.; Smedsgaard, J.; Sumner, L. W.; Wang, T.; Walsh, S.; Wurtele, E. S.; Kell, D. B. *Nat. Biotechnol.* **2004**, *22*, 1601−1606.

(3) Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. *Nat. Biotechnol.* **2008**, *26*, 162−164.

(4) Han, X. L.; Yang, K.; Gross, R. W. *Mass Spectrom. Rev.* **2012**, *31*, 134−178.

(5) Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. *Nat. Protoc.* **2010**, *5*, 1005−1018.

(6) Zehethofer, N.; Pinto, D. M. *Anal. Chim. Acta* **2008**, *627*, 62−70.

(7) Guo, K.; Li, L. *Anal. Chem.* **2010**, *82*, 8789−8793.

(8) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; MacInnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521−D526.

(9) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28*, 27−30.

(10) Schuster, S.; Fell, D. A.; Dandekar, T. *Nat. Biotechnol.* **2000**, *18*, 326−332.

(11) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354−D357.

(12) Karp, P. D.; Ouzounis, C. A.; Moore-Kochlacs, C.; Goldovsky, L.; Kaipa, P.; Ahren, D.; Tsoka, S.; Darzentas, N.; Kunin, V.; Lopez-Bigas, N. *Nucleic Acids Res.* **2005**, *33*, 6083−6089.

(13) Caspi, R.; Foerster, H.; Fulcher, C. A.; Hopkinson, R.; Ingraham, J.; Kaipa, P.; Krummenacker, M.; Paley, S.; Pick, J.; Rhee, S. Y.; Tissier, C.; Zhang, P. F.; Karp, P. D. *Nucleic Acids Res.* **2006**, *34*, D511−D516.

(14) Parkinson, A.; Ogilvie, B. W.; Paris, B. L.; Hensley, T. N.; Loewen, G. J. In *Biotransformation and Metabolite Elucidation of Xenobiotics*; Nassar, A. F., Ed.; John Wiley & Sons: New York, 2010; Vol. *2010*, pp 1−77.

(15) Anari, M. R.; Sanchez, R. I.; Bakhtiar, R.; Franklin, R. B.; Baillie, T. A. *Anal. Chem.* **2004**, *76*, 823−832.

(16) Sana, T. R.; Waddell, K.; Fischer, S. M. *J. Chromatogr., B* **2008**, *871*, 314−321.

(17) Theodoridis, G. A.; Gika, H. G.; Want, E. J.; Wilson, I. D. *Anal. Chim. Acta* **2012**, *711*, 7−16.

(18) Tolstikov, V. V.; Lommen, A.; Nakanishi, K.; Tanaka, N.; Fiehn, O. *Anal. Chem.* **2003**, *75*, 6737−6740.

(19) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. *Nucleic Acids Res.* **2008**, *36*, D480−D484.

(20) 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.

(21) 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.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(22) Clements, M.; Li, L. *Anal. Chim. Acta* **2011**, *685*, 36−44.