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

MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites

Tao Huan,[†,§] Chenqu Tang,[‡,§] Ronghong Li,[‡] Yi Shi,[‡,⊥] Guohui Lin,[‡] and Liang Li*[,†]

[†]Departments of Chemistry and [‡]Computing Science, University of Alberta, Edmonton, Alberta T6G2G2, Canada

Supporting Information

ABSTRACT: We report an analytical tool to facilitate metabolite identification based on an MS/MS spectral match of an unknown to a library of predicted MS/MS spectra of possible human metabolites. To construct the spectral library, the known endogenous human metabolites in the Human Metabolome Database (HMDB) (8,021 metabolites) and their predicted metabolic products via one metabolic reaction in the Evidence-based Metabolome Library (EML) (375,809 predicted metabolites) were subjected to *in silico* fragmentation to produce the predicted MS/MS spectra. This spectral library is hosted at the public MCID Web site (www.MyCompoundID.org), and a spectral search program, MCID MS/MS, has been developed to allow a user to search one or a batch of experimental MS/MS spectra against the library spectra for possible match(s). Using MS/MS spectra generated from standard metabolites and a human urine sample, we demonstrate that this tool is very useful for putative metabolite identification. It allows a user to narrow down many possible structures initially found by using an accurate mass search of an unknown metabolite to only one or a few candidates, thereby saving time and effort in selecting or synthesizing metabolite standard(s) for eventual positive metabolite identification.

Mass spectrometry (MS)-based metabolomics has been developed rapidly in the past decade or so. However, metabolite identification from the MS data is still a challenge. Accurate mass search alone against a chemical database can result in many possible matches. To generate structural information on a metabolite, an MS/MS or fragment ion spectrum can be produced using a tandem mass spectrometer. The fragmentation pattern can be manually interpreted, often against a probable chemical structure found using accurate mass search, to confirm or disapprove a structure.[1] Considering that manual spectral interpretation is a time-consuming process, a spectral search using an MS/MS spectral library of metabolite standards has been developed for rapid metabolite identification.[2,3] Besides in-house and commercial libraries,[4,5] several public libraries have been developed as a very useful resource. For example, our laboratory constructed the Human Metabolome Database (HMDB) MS/MS spectral library using 800 endogenous human metabolites.[6] Other libraries such as Metlin[7,8] and MassBank[9] contain MS/MS spectra of metabolites as well as other synthetic compounds such as common drugs. However, the number of metabolites with reference spectra available is still very small, due to the lack of standards.

In the absence of a standard, a predicted MS/MS spectrum of a given structure can be helpful in manual spectral interpretation as well as in spectral match. There are several approaches for generating predicted MS/MS spectra (more precisely, a list of fragment ions with unit intensity), depending on the chemical bond breakage rules used and the number or level of fragment ions included in a predicted fragment ion spectrum.[10−21] Commercial products (e.g., Mass Frontier from Thermo Scientific, Waltham, US and ACD/MS Fragmenter from Advanced Chemistry Laboratories, Toronto, Canada) and published tools (e.g., Metfrag,[12] Fragment Identificator or FiD,[11] and MIDAS[18]) are available for generating predicted MS/MS spectra with varying degrees of success.

Our approach is to develop a web-based online tool for metabolite identification based on an integrated MS and MS/MS search using a comprehensive library of predicted spectra of all metabolites in MyCompoundID.org (MCID).[1] The current MCID compound library includes 8,021 known endogenous human metabolites in the Human Metabolome Database (HMDB) and 375,809 predicted human metabolites in the Evidence-based Metabolome Library (EML) with one metabolic reaction. We developed an *in silico* method of predicting fragment ions using heteroatom-initiated bond breakage rules and applied it to all MCID metabolites to generate a predicted MS/MS spectral library. An automated MS/MS search program was developed that allows a user to search an experimental MS/MS spectrum, in single or batch search mode, against the library for spectral match. In this paper, we describe the MCID spectral library and MS/MS search tool and demonstrate its performance

Received: August 14, 2015
Accepted: September 28, 2015

 © XXXX American Chemical Society

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX

Figure 1. Screenshots of the result pages from (A) a single-spectrum MCID MS/MS search and (B) a batch-mode MCID MS/MS search.

using MS/MS spectra of metabolite standards and those acquired from a human urine sample.

EXPERIMENTAL SECTION

Overall Workflow. In the MCID MS/MS search method, a precursor ion mass of a metabolite is first used to search against the MCID library to generate a list of candidate compounds with matched molecular ion masses. The fragment ion masses from an experimental MS/MS spectrum are then compared to the predicted fragment ion masses of each candidate compound in the list. A fit score is assigned to each comparison to measure the similarity between the experimental and predicted fragment ions. Once all the comparisons are done, the candidates in the list are ranked by the fit scores.

Predicting MS/MS Fragment Ions. The MCID spectral library contains the predicted MS/MS spectra of 383,830 known and potentially existing human metabolites.[1] Each predicted spectrum was generated using a novel "chopping" program following a series of *in silico* fragmentation rules. A .mol file of a compound structure is used by the program. The algorithm in the chopping program involves two steps. The first step is the heteroatom-initiated bond breakage or chopping. Heteroatoms in a compound such as O and N are identified, and the bonds connecting to the heteroatoms are broken to create possible fragments. The second step is the splittable-bond chopping. Splittable-bonds are linear single bonds and double bonds in aromatic structures. If there are less than 40 splittable-bonds in the chemical structure, four layers of chopping are done. In cases where there are 40−60 splittable-bonds, three layers of chopping are done to avoid generating too many fragment ions. For a very large compound with >60 splittable-bonds, only two layers of chopping are carried out. After applying these two steps of chopping to a compound structure, a mass redundancy check is performed to combine the same fragment ion masses. A list of fragment ion masses are then compiled for the compound and stored as a predicted MS/MS spectrum. All predicted spectra are stored in a local MySQL database in the MCID web server.

Match Algorithm. Two layers of scoring have been developed to gauge the similarity between the experimental MS/MS data and the predicted MS/MS data. At first, we calculate an initial match score, according to

$$\text{score}_i = \frac{1}{\max(\text{weight})}\text{weight}_i$$

where

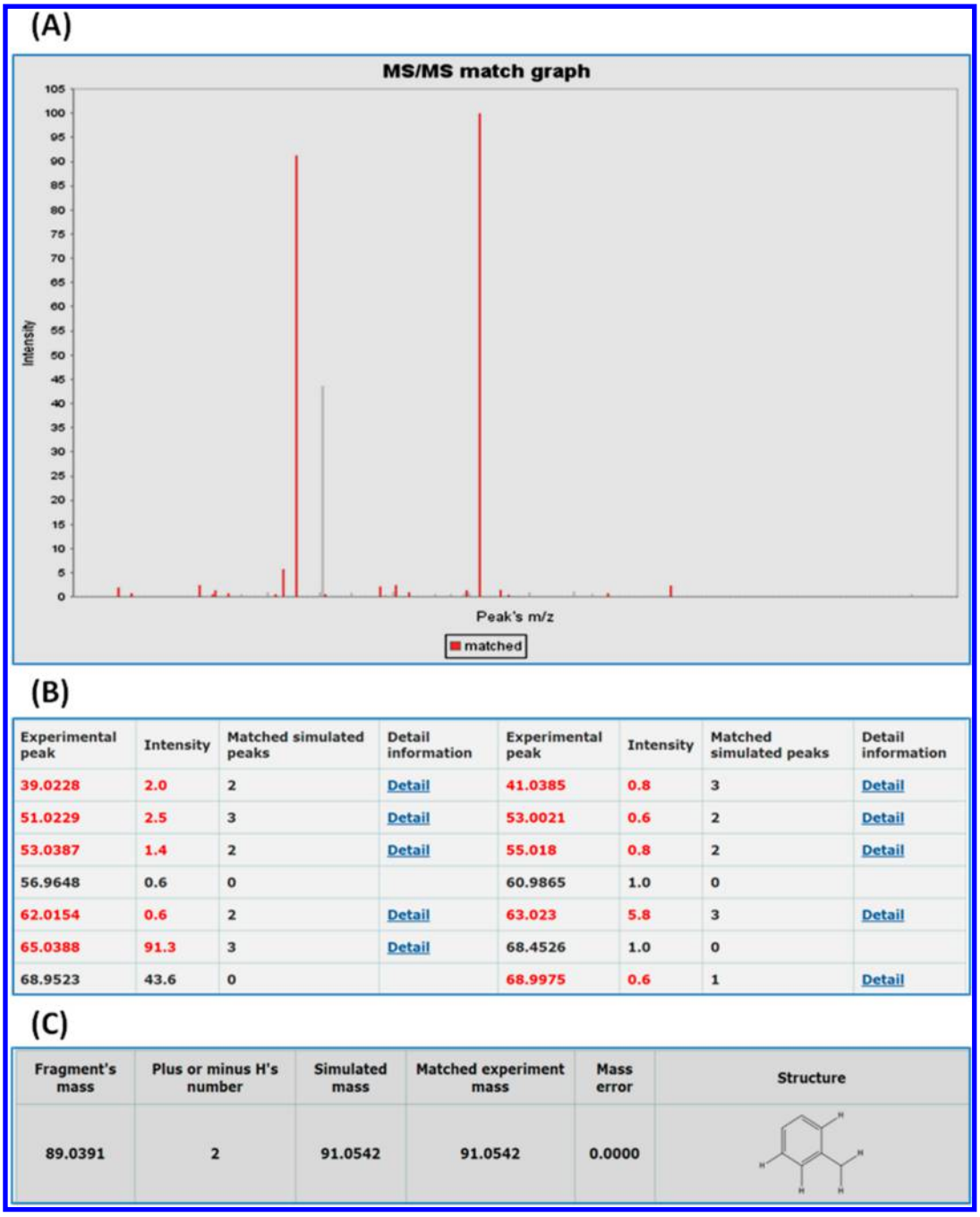
$$\text{weight}_i = \left\langle \overrightarrow{m/z}(\text{matched}) \right\rangle \cdot \left\langle \overrightarrow{\text{Int}}(\text{matched}) \right\rangle$$

$\left\langle \overrightarrow{m/z}(\text{matched}) \right\rangle$: the matched list of m/z

$\left\langle \overrightarrow{\text{Int}}(\text{matched}) \right\rangle$: the matched list of intensities

Using the above equation, a weight is calculated for each comparison by the dot product of the matched m/z's and intensities. An m/z tolerance is set to determine if the experimental m/z is matched with the predicted m/z. The initial score is calculated by normalization against the maximum weight in all the candidates. For the candidates with no-zero initial

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX



Experimental peak	Intensity	Matched simulated peaks	Detail information	Experimental peak	Intensity	Matched simulated peaks	Detail information
39.0228	2.0	2	Detail	41.0385	0.8	3	Detail
51.0229	2.5	3	Detail	53.0021	0.6	2	Detail
53.0387	1.4	2	Detail	55.018	0.8	2	Detail
56.9648	0.6	0		60.9865	1.0	0	
62.0154	0.6	2	Detail	63.023	5.8	3	Detail
65.0388	91.3	3	Detail	68.4526	1.0	0	
68.9523	43.6	0		68.9975	0.6	1	Detail

Fragment's mass	Plus or minus H's number	Simulated mass	Matched experiment mass	Mass error	Structure
89.0391	2	91.0542	91.0542	0.0000	

Figure 2. Screenshots of (A) an experimental MS/MS spectrum with matched fragment ions in red and unmatched ones in gray, (B) a table showing the m/z and relative intensity values of all the experimental fragment ions with matched ones in red, and (C) a table showing the detail of a fragment ion match.

scores, a fit score is then used to quantify and rank how well the experimental spectrum is matched to the predicted spectrum. The fit score is defined as

$$\text{fit score} = \frac{\left\langle \overrightarrow{m/z}(\text{matched}) \right\rangle \cdot \left\langle \overrightarrow{\text{Int}}(\text{matched}) \right\rangle}{\left\langle \overrightarrow{m/z}(\text{experimental}) \right\rangle \cdot \left\langle \overrightarrow{\text{Int}}(\text{experimental}) \right\rangle}$$

where

$\left\langle \overrightarrow{m/z}(\text{matched}) \right\rangle$: the matched list of m/z

$\left\langle \overrightarrow{\text{Int}}(\text{matched}) \right\rangle$: the matched list of intensities

$\left\langle \overrightarrow{m/z}(\text{experimental}) \right\rangle$: the entire list of experimental m/z

$\left\langle \overrightarrow{\text{Int}}(\text{experimental}) \right\rangle$

: the entire list of experimental intensities

MS/MS of Standards. 35 metabolites were selected to generate the MS/MS spectra. An individual standard was used to produce a final concentration of 10 μM. A Bruker Impact HD

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX

Table 1. Summary of MCID MS and MS/MS Search Results for 35 Metabolite Standards

		zero-reaction library search						one-reaction library search						
		precursor mass tolerance ±0.05 Da			precursor mass tolerance ±0.005 Da			precursor mass tolerance ±0.01 Da			precursor mass tolerance ±0.005 Da			
#	name	rank	# of MS/MS match	# of MS match	rank	# of MS/MS match	# of MS match	rank	# of MS/MS match	# of MS match	rank	# of MS/MS match	# of MS match	fit score for correct structure
1	adenine	1	1	5	1	1	1	1	3	19	1	3	19	0.815
2	androstenedione	3	10	12	1	1	1	3	28	33	2	28	32	0.896
3	dopamine	1	3	6	1	3	3	1	19	22	1	19	22	0.819
4	folic acid	1	1	1	1	1	1	4 (3)[a]	12	23	3	6	12	0.873
5	glycine	2	3	3	1	1	1	1	15	17	1	15	17	0.993
6	glutathione	1	1	4	1	1	1	1	5	22	1	5	6	0.892
7	L-phenylalanine	1	4	14	1	2	4	1	28	40	1	20	40	0.838
8	L-alanine	1	4	4	1	4	4	1	27	31	1	23	32	0.810
9	riboflavin	1	0	1	1	0	1	1	0	13	1	0	7	0.554
10	thymine	1	0	6	1	0	2	1	0	4	1	0	6	0.114
11	sarcosine	1	4	4	1	4	4	1	27	31	1	27	32	0.909
12	tryptamine	1	2	13	1	1	2	1	2	25	1	2	10	0.869
13	tyramine	1	3	10	1	3	4	1	11	17	1	11	18	0.700
14	chenodeoxycholic acid	3	0	18	3	0	18	7 (2)	0	44	7 (2)	0	44	0.465
15	creatinine	1	1	3	1	1	1	1	1	1	1	1	1	0.764
16	isovalerylcarnitine	1	3	3	1	3	3	2	17	18	2	15	16	0.980
17	L-methionine	1	2	6	1	2	2	1	13	17	1	13	19	0.965
18	trans-ferulic acid	1	2	13	1	2	3	1	30	39	1	30	39	0.969
19	2′-deoxyguanosine 5′-monophosphate	1	4	4	1	4	4	1	22	34	1	22	32	0.958
20	*N*-acetylmannosamine	2	1	8	3	1	7	11 (3)	28	43	11 (3)	23	37	0.407
21	melatonin	1	1	3	1	1	1	1	7	16	1	6	12	0.986
22	pyridoxamine	1	1	8	1	1	1	2	5	16	2	5	16	0.820
23	*N*-acetylputrescine	1	2	11	1	1	1	1	10	10	2	10	10	0.971
24	creatine	1	1	17	1	1	2	1	4	5	1	4	7	0.985
25	L-asparagine	1	5	24	1	3	5	1	9	13	1	9	12	0.984
26	L-cystine	1	1	1	1	1	1	1	3	6	1	3	3	0.949
27	ornithine	1	2	23	1	2	2	1	12	12	1	12	12	0.945
28	pyridoxine	1	4	8	1	4	4	4 (2)	20	25	4 (2)	20	25	0.935
29	taurine	1	5	5	1	1	1	1	4	1	1	2	2	0.873
30	uric acid	1	1	9	1	1	1	1	3	9	1	3	3	0.839
31	xanthine	1	3	19	1	3	3	1	3	7	1	3	6	0.962
32	xanthosine	1	3	4	1	1	1	2	11	21	1	7	10	0.971
33	DL-homocystine	1	2	10	1	2	2	1	4	11	1	2	8	0.936
34	4-hydroxyproline	1	4	17	1	3	8	1	8	52	1	8	52	0.998
35	xanthurenic acid	1	4	5	1	1	1	1	9	17	1	9	17	0.958

[a] (x) where x = new rank after grouping isomers as one match.

QTOF mass spectrometer (Billerica, MA) was used to generate the MS/MS spectra using direct infusion with collision energy of 20−50 eV. The MS and MS/MS conditions for running the standards as well as the urine sample (see below) were: mode, positive ion electrospray ionization; capillary voltage, 4500 V; dry gas, 6.0 L/min; dry heater, 230 °C; nebulizer, 1.0 bar; mass range, 50−1000 mz; collision energy, 20−50 eV (time: 50%−50%).

LC-MS/MS of Urine. A human urine sample was collected from a healthy individual and filtered using a 0.22 μm-pore-size filter (Millipore Corp., MA) twice. LC-MS/MS analysis was performed on the Bruker QTOF-MS equipped with an Agilent 1100 HPLC system (Palo Alto, CA). A reversed-phase Zorbax Eclipse C18 column (2.1 mm × 100 mm, 1.8 μm particle size, 95 Å pore size) from Agilent was used. Solvent A was 0.1% (v/v) LC-MS grade formic acid in 2% (v/v) LC-MS grade ACN, and solvent B was 0.1% (v/v) LC-MS grade formic acid in 98% (v/v) LC-MS grade ACN. The gradient elution profile was as follows: t = 0.0 min, 0% B, t = 10 min, 0% B, t = 50.0 min, 80% B, t = 55 min, 100%B, t = 60 min, 100% B, t = 60.1 min, 0% B, t = 80 min, 0% B. The flow rate was 100 μL/min. The sample injection volume was 20 μL.

RESULTS AND DISCUSSION

MCID MS/MS Search. There are two search modes available (see Supplemental Figure S1 for a screenshot of the web interface and Supplemental Note N1 for a tutorial). In the single-spectrum search, which is useful for targeted metabolite identification, a user selects either the zero-reaction library containing all the known metabolites in HMDB or the one-reaction library containing all the predicted metabolites in EML. The precursor ion type, mass, and mass tolerance are entered. The fragment ion masses and their relative intensities and the m/z tolerance value for the fragment ions are also entered. "Deisotope" is selected as a default to remove the ^{13}C natural abundance isotopic peak(s) of a fragment ion. Figure 1A shows an example of the search results

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX

obtained by searching the zero-reaction library. The result page lists all the mass-matched metabolites. For each candidate, the HMDB ID number with a link to the HMDB database is given along with other information.

For each candidate, the initial score and fit score from MS/MS spectral comparison are given. In the case shown in Figure 1A, the three candidates are isomers and the fit scores are almost the same. By clicking the initial score, a new page will be displayed. Figure 2A shows an example where the experimental MS/MS spectrum is shown. The matched peaks to the predicted spectrum are shown in red and unmatched ones are shown in gray. It also displays a table (Figure 2B) containing information on the masses and intensities of the experimental fragment ions (matched ones in red and unmatched ones in black), the number of matched fragment ion structures, and a link called "Detail". By clicking "Detail", another page will be displayed (Figure 2C) which provides a summary of the matched fragment ion(s) including the predicted structure(s). These multiple layers of information can be very helpful for manual confirmation of a MS/MS match. Manual interpretation may assist in determining which structure among the matches is the most probable one to fit the MS/MS fragmentation pattern.

To facilitate manual comparison of an experimental MS/MS spectrum and a predicted spectrum, there is also a function of uploading the matched metabolite structure to a local ChemDraw software or an online ChemDraw Plugin (freeware). In both programs, built-in "Fragmentation Tools" can be used to direct a bond breakage of a structure to show the resulting fragment ion structures and their masses. An example of how to use this tool for manual fragmentation pattern interpretation has been given in the original MCID paper.[1]

In addition to single spectrum search, a user can upload a CSV file generated from LC-MS/MS analysis of a sample for a batch mode search. This is useful for examining all the possible matches in a metabolomic profiling experiment. The file format used is shown in Supplemental File F1 (if MS data is saved as an MGF file, the user merely needs to change the file extension from .mgf to .csv). To share the computation resource in the MCID server by multiple users, the file size is limited to 100 MS/MS spectra. For a large file, a file split program can be freely downloaded from the MCID website and installed in the user's computer to split the file into several small files with each limited to 100 spectra. These split files can be uploaded individually for the MS/MS search. The search time for each file depends on the parameters used (e.g., a smaller precursor mass tolerance would increase the search speed as fewer candidates would need to be examined in the MS/MS search) and the number of search jobs in the server. After the searches, the individual search results can be merged by a file merge program which can also be downloaded from the MCID website and installed in the user's computer to produce the final result in CSV.

Figure 1B shows a screenshot of part of a search result page from MS/MS spectra of a human urine sample acquired by LC-QTOF-MS. A summary table lists information on retention time, precursor ion mass, precursor ion intensity, the number of fragment ions detected, the number of mass-matched metabolites (i.e., number of hits), fit score, show-details with links, and save-result in CSV for a given match. By clicking the show-detail, several layers of information can be displayed for each MS/MS match as in the case of single spectrum search discussed earlier.

The search results can be sorted according to any parameters in the summary table. There are several parameters (see the top of Figure 1B) that can be used to filter the search results to retain the matches of interest. By clicking "Download Table Result", all the filtered matches are saved to the user's computer in a CSV file (see Supplemental Table T1 as an example). For privacy and confidentiality, the server does not store any search file or search results. However, in the saved CSV file that can be opened in Excel, there is a link column containing long names for all the individual matches. By copying and pasting a link name of a match to the web, the user can retrieve the search result in MCID for the match. This is possible because the long name contains all the MS and MS/MS information required for a new MCID MS/MS search to generate the match result again. This feature allows a user to examine any matches in the result table without the need of repeating the batch mode search.

MS/MS Search of Standards. To evaluate the performance of the MCID MS/MS search, we searched the MS/MS spectra of 35 standards generated by QTOF-MS against the predicted MS/MS spectral library. These metabolites were randomly picked in order to cover as many different types of compounds as possible. Although direct infusion of a relatively high concentration of analyte (10 μM) was used to produce the highest quality of spectra to represent the best-case scenarios for spectral search, similar spectra could be obtained using scheduled MS/MS data acquisition in LC-QTOF-MS with a 20 μL injection of the same solution. Table 1 shows the list of metabolites and their search results (see Supplemental Table T2 for HMDB numbers and chemical structures). The MS/MS spectra were searched using the zero- and one-reaction libraries with a normal (i.e., 0.005 Da, a typical mass accuracy from QTOF-MS) and a wider (i.e., 0.05 Da for zero-reaction and 0.01 Da for one-reaction) precursor ion mass tolerance. The wider tolerance was deliberately used in order to increase the number of mass-matched metabolites including many false ones for the purpose of testing the ability of using the MS/MS search to distinguish the correct structure from the false ones. The fragment ion mass tolerance was set to be 0.005 Da, according to the QTOF-MS/MS mass accuracy.

With a wider precursor ion mass tolerance, for the zero-reaction library search, an average of 8.6 library compounds was mass-matched to a tested standard, while the MS/MS search resulted in an average of 2.5 matched compounds with a fit score of $\geq$0.700 (see below). For the one-reaction search, an average of 20.4 compounds were matched to a standard if only mass search was used. With the MS/MS search, an average of 11.4 compounds with a fit score of $\geq$0.700 was matched to a standard. Using the precursor ion mass tolerance of 0.005 Da, for the zero-reaction search, averages of 2.9 and 1.7 compounds were matched to a standard using the MS search and MS/MS search, respectively. For the one-reaction library, the MS search and MS/MS search resulted in an average of 18.2 and 10.4 compounds matched to a standard. These results show that the number of MS/MS matched structures with a fit score of $\geq$0.700 is significantly lower than the number of MS matched structures.

Since the structures of the 35 standards are known, we can examine the accuracy of the MS/MS matches in a rank according to the fit score. For the zero-reaction search using a wider mass tolerance, 31, 2, and 2 standards (88.6%, 5.7%, and 5.7%) gave the correct compound as the top, second, and third ranked match, respectively (see Table 1). Even for the one-reaction search, 27, 3, and 1 standards (77.1%, 8.6% and 2.9%) gave the correct structure as the top, second, and third ranked match, respectively. Only 4 standards had the correct structure ranked below the third match. The 11th ranked *N*-acetylmannosamine, out of 43 mass-matched compounds, has isomers ranked from

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX

the top 1 to 10. Effectively, this match was ranked third if isomers were counted as one (see Table 1 with the new rank in brackets). Similarly, for the seventh ranked chenodeoxycholic acid, out of 44 mass-matched compounds, the top 5 matches were isomers. Counting all the isomers as one, this match was ranked second. In the case of using 0.005 Da precursor ion mass tolerance for the MS/MS search, as Table 1 shows, for the zero-reaction library, 33 (94.3%), 0 (0%), and 2 (5.7%) standards gave the correct structure as the top, second, and third ranked match, respectively. For the one-reaction library, 27 (77.1%), 4 (11.4%), and 1 (2.9%) standards gave the correct structure as the top, second, and third ranked match. Only 3 (8.6%) standards were below the top three. These three cases would be ranked the top three if grouping the isomers as one.

The above results show that the correct structure of a MS/MS search belongs to one of the top three structures with the majority of them as the top match. This finding would suggest that, for a MS/MS search, only the top three structures including isomers need to be inspected manually to confirm or disapprove a match. This should greatly improve the overall metabolite identification efficiency. For the 35 standards, after generating the top three structure matches for each metabolite in the one-reaction search results, we manually checked the matches to validate the automatic MS/MS search results. Twenty-seven top ranked metabolites could be manually confirmed. For the second ranked metabolites, 3 out of 4 could be confirmed by manually eliminating the top ranked false match. Only one of the second ranked metabolites (isovalerlcarnitine, #16 in Table 1) could not be differentiated from the top ranked structure due to the lack of characteristic fragment ions from the two structures.

Table 1 also lists the fit score of the correctly matched structure from the MS/MS search for each standard. The fit score determines the matching quality of the experimental MS/MS data with the predicted MS/MS data. The average fit score for all the correct structures is 0.860, and 90% of the correct structures have fit scores of ≥0.700. There are 3 cases that the correct structures have a fit score of <0.700. Manual inspection of these matches shows that these spectra do not have enough high intensity and informative fragment ion peaks. For example, in the case of thymine (#10 in Table 1) with a fit score of 0.114, the MS/MS spectrum shows only one fragment ion peak and this peak cannot be explained even with manual interpretation. This observation is not surprising, considering that not all the metabolites can be fragmented or produce a sufficient number of characteristic fragment ions even at the best spectral acquisition conditions. Nevertheless, more than 90% of the 35 standards could produce MS/MS spectra with sufficiently high quality to render a fit score of ≥0.700. Thus, a fit score of 0.700 can be used as a cutoff threshold for an automated MS/MS search to produce a list of structure candidates from which manual interpretation can be carried out to approve or disapprove a structure match.

MS/MS Search of Urine Metabolites. To demonstrate the utility of the MCID MS/MS search in real world applications, a human urine sample was analyzed by LC-QTOF-MS, followed by a library search for metabolite identification. In this experiment, a precursor ion exclusion (PIE) strategy, similar to that used in a shotgun proteomics work,[22] was applied to acquire as many MS/MS spectra as possible from triplicate runs of the same human urine.

In total, 5794 MS/MS spectra were generated using the PIE strategy. We used 0.005 Da mass tolerance for precursor and fragment ions in the MCID MS/MS search and generated 1160 spectral matches using the zero-reaction library (see Supplemental Table T3 for the list). We then performed a cross-validation of some of the spectral matches using a Bruker HMDB MS/MS spectral library. This Bruker library containing 800 standards was created in the same QTOF instrument as the one used for running the urine sample. Thus, an excellent fragmentation pattern match of a high quality experimental MS/MS spectrum of a urine metabolite with the Bruker MS/MS spectrum of the same metabolite is expected, which should in turn provide high confidence for validation of the MCID MS/MS search results. One example of the validation work is shown in Figure 3. Figure 3A shows the experimental MS/MS spectrum of

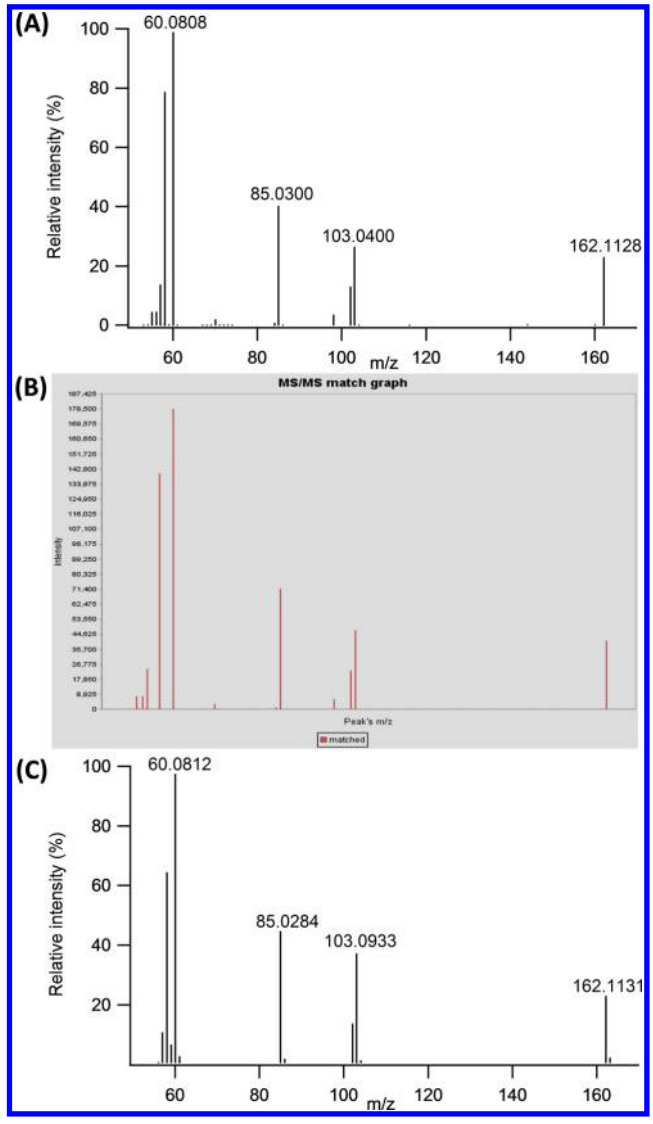


Figure 3. (A) MS/MS spectrum of a metabolite obtained from LC-QTOF-MS analysis of a human urine sample, (B) screenshot of the MS/MS spectrum from a result page showing all the matched fragment ions to a predicted spectrum in red, and (C) MS/MS spectrum of carnitine standard from the Bruker experimental spectral library.

carnitine found in urine at the retention time of 2.55 min. Figure 3B shows the match of the experimental MS/MS with the predicted MS/MS spectrum of carnitine in the MCID library. Figure 3C shows the standard MS/MS spectrum of carnitine in the Bruker library. The fit score for this compound using the predicted spectral library was 0.995, compared to a purity score of 954 out of 1000 using the Bruker library. Thus, the MCID

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX

MS/MS search result or identification of carnitine was cross-validated.

Supplemental Table T4 lists the metabolites initially identified using the MCID MS/MS search with the zero-reaction library and subsequently validated using the Bruker experimental spectral library. Out of the 77 validated spectral matches, 54, 18, and 3 metabolites were correctly identified by MCID MS/MS as the top (70.1%), second (23.4%), and third (3.9%) ranked structure, respectively. Two of them were ranked below top 3. However, if treating isomers as a group, only 1 spectral match had the correct structure ranked below the top 3. Thus, 76 out of the 77 spectral matches (97.4%) had the correct structure belonging to one of the top 3 matched structures. The average fit score was 0.775. These results indicate that, using the MCID MS/MS search, almost all the correctly matched metabolites could be found as the top 3 structures in a LC-MS/MS experiment of a real biological sample. If this holds true for the other nonvalidated matches, only the top 3 structures from a MCID MS/MS search of an unknown metabolite would need to be inspected or confirmed for identification. Of course, more validation work is needed to generalize this finding in the future. Nevertheless, the urine results illustrate that the MCID MS/MS search is capable of identifying metabolites with high confidence.

The fit scores of the 77 metabolites were analyzed to determine the best cutoff for a high confident MS/MS match. From the study of the 35 metabolite standards, we proposed to use 0.700 as the cutoff. However, we noticed that, in the urine sample analysis, this cutoff score is too restricted in some cases. This is because, in the analysis of a complicated biological sample, metabolites have a wide concentration range and their MS/MS signals can be affected by precursor ion intensity, background impurities, and coeluting compounds. For example, of the 77 metabolites, only 52 (67.5%) of the correctly matched structures had their fit score of >0.700. Another 18 (23.4%) of the correct structures had a fit score of between 0.700 and 0.400, and 7 (9.1%) structures even had a fit score of below 0.400. These results indicate that using a cutoff fit score of 0.700 will exclude a large fraction of the correct structures. On the other hand, of the 52 metabolites with a fit score of larger than 0.700, their correct structures were all ranked at the top 3. Thus, we can use a cutoff of 0.700 to generate a list of high confident structure matches where one of the top 3 structures is expected to be correct. For the remaining spectral matches with a fit score of below 0.700, we would still examine the top three structure matches of an experimental MS/MS spectrum to determine if one of the matches is correct; however, there is no guarantee that any of the top 3 structures is correct in these cases. We recognize that simply using a fit score cutoff of 0.700 represents a compromise between the search specificity and the search sensitivity. Future work will be needed to develop a more robust scoring system for the MS/MS search to increase both specificity and sensitivity.

We applied the 0.700 cutoff threshold to all of the 1160 spectral matches including the 77 validated matches. We found that 636 MS/MS spectra have structure matches with a fit score of ≥0.700 for a total of 1227 structures (see Supplemental Table T5). Among them, 378, 126, and 54 have spectral matches with 1, 2, and 3 structures, respectively, and 78 have spectral matches with 4 or more high-score structures. While we cannot narrow down each spectral match to one structure, we can state that 636 MS/MS spectra have high confident structure matches and one of the top three structures for each spectral match is most likely correct. It is clear that the MCID MS/MS search can generate many high confident, but still putative, identifications from a urine sample.

Finally, we would like to illustrate the power of using MCID MS/MS to search a predicted MS/MS spectral library of one-reaction metabolites. Out of the 5794 MS/MS spectra collected from the urine sample, we took the remaining unmatched or unconfirmed spectra (i.e., 5158) from the zero-reaction library search to search the one-reaction library and the results are shown in Supplemental Table T6. A total of 3920 (76.0%) MS/MS spectra were matched to the one-reaction library. Among them, 1250 spectra have a total of 5966 structures matched with a fit score of ≥0.700 (see Supplemental Table T7). This includes 587, 380, and 123 spectra match with 1, 2, and 3 structures, respectively, and 160 spectra match with 4 or more high-score structures.

To validate some of these matches, we used the published data of 87 one-reaction metabolites that were identified on the basis of manual interpretation of the mass-matched metabolites in MCID MS search.[1] On the basis of the match of retention time, precursor mass, and MS/MS fragmentation pattern, 78 out of these 87 metabolites (88.5%) were identified in the current urine sample (see Supplemental Table T8). Among the 78 metabolites, 44 (56.8%), 17 (21.8%), and 6 (7.7%) spectra had their correctly matched structures ranked at the top, second, and third, respectively. Only 11 correct structure matches were ranked below the top 3. If treating isomers as a group, out of these 11 matches, 3 matches were ranked #2 and 2 matches were ranked #3. Thus, 72 out of the 78 metabolites (92.3%) had the correct structure belonging to one of the top 3 structure matches. Among the 78 metabolites, there were 57 spectral matches with a fit score of ≥0.700. For these matches, 56 of them (98.2%) had a correct structure listed at the top 3 matches (treating isomers as a group). These results demonstrate that a fit score cutoff threshold of 0.700 can narrow down the correct structure to one of the top 3 structure matches, even for the one-reaction library search. If this holds true generally, one of the top 3 matches for each of the 1250 MS/MS spectra having one-reaction library metabolite matches with a fit score of ≥0.700 in Supplemental Table T7 should be the correct structure.

Taken together, the urine sample analysis results indicate that, in most cases, only the top 3 structure matches from the MS/MS search of an experimental MS/MS spectrum with a fit score of ≥0.700 needs to be manually inspected for confirming or disapproving a match. Of course, for positive metabolite identification, an authentic standard is needed to confirm a structure match. In this regard, using the MCID MS/MS search, standards of only a few top ranked candidates need to be acquired or synthesized, which should greatly reduce the time and effort needed for metabolite identification.

There are several improvements that we plan to implement in a future release of the MCID search tool. One of them is related to the expansion of the spectral library. Including metabolites of other origins, in addition to human metabolites, would increase the utility of the search tool. Incorporation of the MS/MS spectral library of metabolite standards into the MCID spectral library would provide more reliable match results for those metabolites. In the current HMDB MS/MS search, there is an option of adding predicted MS/MS spectra. These predicted spectra were generated using competitive fragmentation modeling (CFM)[20] which is completely different from the method presented in this work. We also note that the match scoring algorithm and results outcome interfaces used in HMDB are very different from the MCID MS/MS search.

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX

CONCLUSIONS

We have developed a web-based MS/MS spectral search tool for improving metabolite identification based on the use of a large library of predicted fragment-ion-spectra of over 383,830 possible human metabolites. This tool is freely accessible at www.MyCompoundID.org, allowing a user to search an MS/MS spectrum or a batch of spectra against the library for possible structure matches. Using MS/MS spectra collected from 35 standards and a human urine sample, we demonstrated that one of the top 3 matches from an MS/MS spectrum with a fit score of ≥0.700 (out of 1.000) is a correct structure. While the MCID MS/MS spectral search cannot always produce one unique structure match, narrowing down the possible matches to the top 3 candidates should save the time and effort to find or synthesize authentic compound standards for positive identification.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b03126.

Figure S1, File F1 for an example of a MS/MS file, Note N1 for a tutorial, and Tables T1−T8 listing the search results. (ZIP)

AUTHOR INFORMATION

Corresponding Author

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

Present Address

[⊥]Y.S.: Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiaotong University, China.

Author Contributions

[§]T.H. and C.T. contributed equally.

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 Alberta Innovates.

REFERENCES

(1) Li, L.; Li, R. H.; Zhou, J. J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y. M.; Huan, T.; Zheng, J. M.; Shi, Y.; Wishart, D. S.; Lin, G. H. *Anal. Chem.* **2013**, *85*, 3401−3408.

(2) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B. *Analyst* **2009**, *134*, 1322−1332.

(3) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, *9*, 44−66.

(4) Oberacher, H.; Whitley, G.; Berger, B. *J. Mass Spectrom.* **2013**, *48*, 487−496.

(5) Stein, S. *Anal. Chem.* **2012**, *84*, 7274−7282.

(6) 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.

(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) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z. J.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.

(9) 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.

(10) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111−3118.

(11) Heinonen, M.; Rantanen, A.; Mielikainen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; Rousu, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043−3052.

(12) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.

(13) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatos, A.; Bocker, S. *Anal. Chem.* **2012**, *84*, 3417−3426.

(14) Rojas-Cherto, M.; Peironcely, J. E.; Kasper, P. T.; van der Hooft, J. J. J.; de Vos, R. C. H.; Vreeken, R.; Hankemeier, T.; Reijmers, T. *Anal. Chem.* **2012**, *84*, 5524−5534.

(15) Hufsky, F.; Scheubert, K.; Bocker, S. *TrAC, Trends Anal. Chem.* **2014**, *53*, 41−48.

(16) Ma, Y.; Kind, T.; Yang, D. W.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86*, 10724−10731.

(17) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; Vervoort, J.; Bino, R. J. *Anal. Chem.* **2014**, *86*, 4767−4774.

(18) Wang, Y. F.; Kora, G.; Bowen, B. P.; Pan, C. L. *Anal. Chem.* **2014**, *86*, 9496−9503.

(19) Zhou, J. R.; Weber, R. J. M.; Allwood, J. W.; Mistrik, R.; Zhu, Z. X.; Ji, Z.; Chen, S. P.; Dunn, W. B.; He, S.; Viant, M. R. *Bioinformatics* **2014**, *30*, 581−583.

(20) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, *11*, 98−110.

(21) Vaniya, A.; Fiehn, O. *TrAC, Trends Anal. Chem.* **2015**, *69*, 52−61.

(22) Wang, N.; Li, L. *Anal. Chem.* **2008**, *80*, 4696−4710.

DOI: 10.1021/acs.analchem.5b03126
Anal. Chem. XXXX, XXX, XXX−XXX