analytical chemistry

Article

pubs.acs.org/ac

MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics

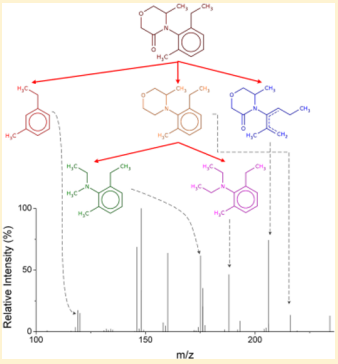
Yingfeng Wang,[†,§] Guruprasad Kora,[†] Benjamin P. Bowen,[‡] and Chongle Pan*[,†]

[†]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

[‡]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: A database searching approach can be used for metabolite identification in metabolomics by matching measured tandem mass spectra (MS/MS) against the predicted fragments of metabolites in a database. Here, we present the open-source MIDAS algorithm (Metabolite Identification via Database Searching). To evaluate a metabolite-spectrum match (MSM), MIDAS first enumerates possible fragments from a metabolite by systematic bond dissociation, then calculates the plausibility of the fragments based on their fragmentation pathways, and finally scores the MSM to assess how well the experimental MS/MS spectrum from collision-induced dissociation (CID) is explained by the metabolite's predicted CID MS/MS spectrum. MIDAS was designed to search high-resolution tandem mass spectra acquired on time-of-flight or Orbitrap mass spectrometer against a metabolite database in an automated and high-throughput manner. The accuracy of metabolite identification by MIDAS was benchmarked using four sets of standard tandem mass spectra from MassBank. On average, for 77% of original spectra and 84% of composite spectra, MIDAS correctly ranked the true compounds as the first MSMs out of all MetaCyc metabolites as decoys. MIDAS correctly identified 46% more original spectra and 59% more composite spectra at the first MSMs than an existing database-searching algorithm, MetFrag. MIDAS was showcased by searching a published real-world measurement of a metabolome from *Synechococcus sp. PCC 7002* against the MetaCyc metabolite database. MIDAS identified many metabolites missed in the previous study. MIDAS identifications should be considered only as candidate metabolites, which need to be confirmed using standard compounds. To facilitate manual validation, MIDAS provides annotated spectra for MSMs and labels observed mass spectral peaks with predicted fragments. The database searching and manual validation can be performed online at http://midas.omicsbio.org.



Metabolomics is an emerging technology for detecting, identifying, and quantifying metabolites in biological samples. While traditional metabolite profiling targets a set of known metabolites, metabolomics attempts to measure all metabolites in an organism, referred to as the metabolome. Metabolites are small-molecule compounds typically less than 1500 Da.[1] A metabolome includes intermediates and products of metabolic pathways, building blocks of biological macromolecules, and derivatives of drugs and nutrients.[1] Thus, measurements of an organism's metabolome can provide direct information on its metabolic activities. This complements the characterization of genetic capabilities by genomics and the analysis of gene expression profiles by transcriptomics and proteomics.

The diverse chemical properties of metabolites present an enormous analytical challenge in metabolomics. Several analytical technologies have been used to identify and quantify a large number of metabolites. Nuclear magnetic resonance (NMR) spectroscopy can detect almost all types of metabolites and provide direct structure information.[2] However, NMR has the disadvantage of low sensitivity. Gas chromatography-electron ionization-mass spectrometry (GC-EI-MS) is an established technology for metabolite profiling.[3] Metabolites are separated by GC, ionized and fragmented by EI, and measured by MS. EI fragmentation is highly reproducible. However, GC-EI-MS is limited to metabolites that can be introduced nondestructively into the gas phase either directly, or following chemical derivatization. Liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) has become increasingly popular for metabolomics.[4,5] LC separation is compatible with metabolites in aqueous or organic solutions. ESI allows soft ionization of metabolites. MS/MS first measures the mass-to-charge ratios (m/z) of intact metabolites, then dissociates an isolated metabolite ion, and finally measures the m/z of fragments of the metabolite. Many metabolites can be detected and quantified from the MS/MS data of automated LC-ESI-MS/MS measurements.

It is an informatics challenge to identify a metabolite based on the mass spectrum of its fragments measured by GC-EI-MS or LC-ESI-MS/MS. The GC-EI-MS community has accumulated large collections of EI spectra of known metabolites against which newly acquired spectra can be searched. The close match of an acquired spectrum with a standard spectrum in the library leads to identification of the compound associated

Received: April 22, 2014
Accepted: August 26, 2014

   dx.doi.org/10.1021/ac5014783 | *Anal. Chem.* XXXX, XXX, XXX−XXX

with the standard spectrum. There are also smaller, but rapidly growing, spectral libraries for LC-ESI-MS/MS in METLIN,[6,7] HMDB,[8−10] MassBank,[11] the National Institute of Standards and Technology (NIST) library,[12] and the Fiehn library.[13] However, it is not feasible to include all metabolites in a spectral library. It is a tedious and costly process to identify spectra that do not match any standard spectrum in spectral libraries. Experts may manually interpret an unknown spectrum and infer likely candidate compounds. These candidate compounds need to be synthesized and measured to generate standard spectra for comparison. A high degree of similarity between the unknown spectrum and a standard spectrum provides experimental support for the standard compound to have the same structure as the unknown compound.

LC-ESI-MS/MS has been the analytical method of choice for shotgun proteomics. The identification of peptides from MS/MS data was automated using a database searching approach in shotgun proteomics.[14] The fragments of a peptide can be predicted relatively reliably based on a set of fragmentation rules,[15] which obviate the need to acquire standard spectra for peptides. Searching measured MS/MS spectra against predicted MS/MS spectra can identify tens of thousands of peptides from a typical LC-ESI-MS/MS run in proteomics.

For clarity in this article, we define the database searching approach in metabolomics as searching experimental MS/MS spectra against a database of metabolites using their predicted MS/MS spectra to find the best matches out of the database, which contrasts with the established spectral library searching approach using measured MS/MS spectra of standard compounds. EI fragments of metabolites can be predicted based on their chemical structures using computer programs, such as Mass Frontier,[16] ACD/MS Fragmenter,[17] and MOLGEN-MSF.[18] These programs generally use fragmentation rules extracted from publications.[5,19,20] A comparison study showed Mass Frontier to be more accurate than other programs.[21] However, the fragmentation rules for collision-induced dissociation used in LC-ESI-MS/MS are still too sparse to be applied for a large number of metabolites. More recently, Wolf et al.[22] developed the MetFrag program that used a database searching approach for metabolite identification. Fragments of a metabolite were predicted by systematically disconnecting bonds and their matches with MS/MS spectra were scored based on bond dissociation energy. MetFrag was shown to outperform Mass Frontier.[22] Here, we report a new database-searching algorithm, MIDAS (Metabolite Identification via Database Searching). MIDAS correctly identified more compounds than MetFrag for four standard ESI-MS/MS data sets from MassBank. MIDAS was further demonstrated using a real-world LC-ESI-MS/MS measurement of a *Synechococcus* metabolome.

MATERIALS AND METHODS

MIDAS Algorithm. MIDAS searches an MS/MS spectrum against a metabolite database (see below for the database format and MS/MS data formats). The measured precursor m/z is used to find candidate metabolites that have expected m/z within a user-defined precursor m/z error tolerance. The m/z of a metabolite is calculated using its protonated form, $[M + H]^+$, in positive ion mode and its deprotonated form, $[M - H]^-$, in negative ion mode.

To evaluate a metabolite-spectrum match (MSM), MIDAS first constructs a three-level fragmentation tree for the candidate metabolite. Fragments in the first level are generated by systematically disconnecting each linear bond and cleaving open each ring in the metabolite. Bonds involving a hydrogen are not considered for dissociation. The second level of fragments is generated from the first level of fragments and the third level from the second level. Figure 1A shows an example

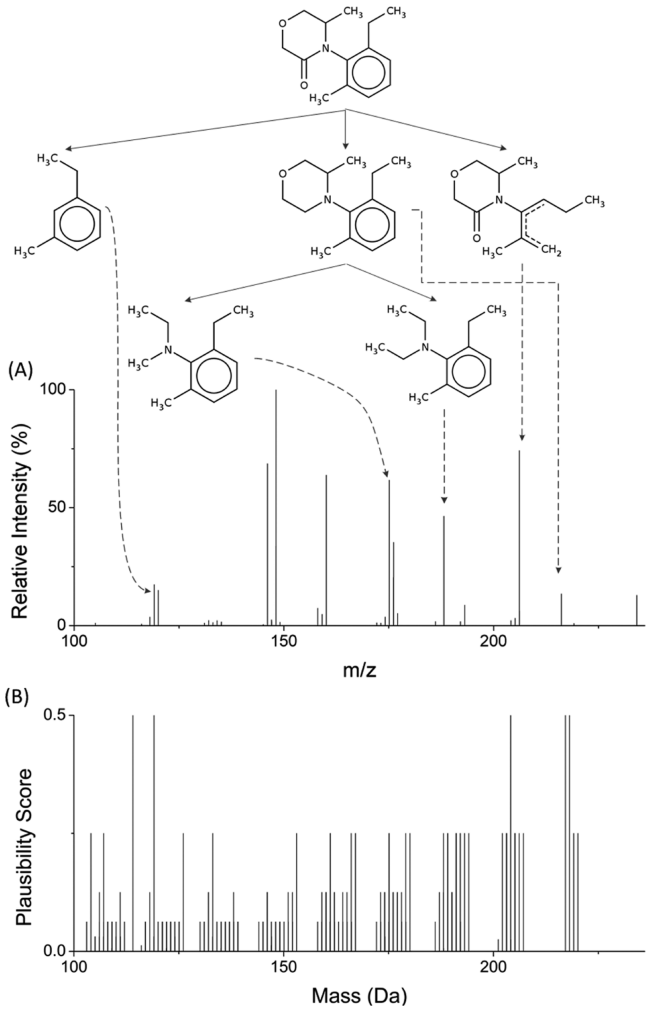


Figure 1. Metabolite-spectrum match of metolachlor morpholinone. (A) Measured Orbitrap MS/MS spectrum annotated with a selected set of observed fragments in a fragmentation tree. (B) Masses and plausibility scores of all predicted fragments from this compound. Each peak represents a predicted fragment.

of the fragmentation tree inferred from observed MS/MS peaks of metolachlor morpholinone. The fragmentation tree is traversed by depth-first search (DFS). MIDAS scores the fragments as they are generated by the DFS and tallies the total score of the MSM. Since there is no need to keep information on the visited branches of the fragmentation tree, only the current branch that the DFS is traversing needs to be stored in memory. This is more memory-efficient than the breadth-first search approach used by MetFrag, in which an entire level of the fragmentation tree needs to be stored in memory.

When a fragment, F, is generated by the DFS, MIDAS considers three charged forms of F to calculate m/z. In positive ion mode, the three charged forms are $[F]^+$, $[F + H]^+$, and $[F + 2H]^+$. $[F + H]^+$ can be considered as a protonated fragment generated from the homolysis of a bond; $[F]^+$ as an unprotonated fragment that has lost the electron pair in the

dx.doi.org/10.1021/ac5014783 | Anal. Chem. XXXX, XXX, XXX−XXX

heterolysis of a bond or as a protonated fragment that has lost the electron pair along with a nearby proton to the leaving group; and $[F + 2H]^+$ as a protonated fragment that has gained the electron pair in the heterolysis along with a proton from the dissociated group. In negative ion mode, the fragment, F, is considered in the charged forms of $[F]^-$, $[F - H]^-$, and $[F - 2H]^-$ to calculate m/z.

After calculating the m/z of a fragment, MIDAS evaluates how plausible it is to generate the fragment. Let P be the immediate precursor of the current fragment, F, in the fragmentation tree. The plausibility score of F, $S(F)$, is calculated as

$$S(F) = \begin{cases} 0.5^b S(P) \text{ for detected } P \\ 0.1^b S(P) \text{ for not detected } P \end{cases} \quad (1)$$

where $S(P)$ is the plausibility score of P and b is the number of disconnected bonds that yields F from P. P is detected when a peak is found at the expected m/z of P in the spectrum. b is 1, when a linear bond is disconnected, and b is 2, when a pair of ring bonds are disconnected to cleave open a ring. The intact metabolite at the root of the fragmentation tree is assigned with a plausibility score of 1 with $b = 0$. When a fragment can be generated from multiple branches in the fragmentation tree, the fragment is assigned with the highest plausibility score from these branches. Figure 1B shows the plausibility scores of all predicted fragments from metolachlor morpholinone, given its MS/MS spectrum in Figure 1A.

MIDAS uses the calculated m/z of the three charged forms of a fragment to find its mass spectral peak within a user-defined fragment m/z error tolerance, Δ. When the fragment is matched to a peak, the m/z error, ε, is calculated between the fragment m/z and the peak m/z. ε is standardized to an m/z error weight, λ, as described[23] using the equation below:

$$\lambda = \begin{cases} 2(1 - \text{pnorm}(\varepsilon, 0, \Delta/2)) \\ \quad \text{for matched fragments and peaks} \\ 0 \text{ for not matched fragments and peaks} \end{cases} \quad (2)$$

where the $\text{pnorm}(\varepsilon,0,\Delta/2)$ function returns the cumulative probability for random variable ε from a normal distribution with mean 0 and standard deviation $\Delta/2$. As ε increases from 0 to Δ, λ decreases from 1.00 to 0.05.

After the DFS traverses the entire fragmentation tree, the match between the experimental MS/MS spectrum and the predicted MS/MS spectrum of the candidate metabolite is scored as

$$T = \sum_{j=1}^{n} \max_{i=1}^{m} (S(F_j) \cdot I_i \cdot \lambda_{ij}) \quad (3)$$

where n is the total number of fragments in the fragmentation tree, m is the number of measured peaks, $S(F_j)$ is the plausibility score of fragment j by eq 1, $I_i \in (0,1]$ is the relative intensity of peak i, and λ_{ij} is the m/z error weight between peak i and fragment j by eq 2. In eq 3, a matched predicted fragment is rewarded only once using its best matched peak, but a peak can be used to match multiple fragments.

For every MS/MS spectrum, MIDAS scores all candidate metabolites with matched precursor m/z in the database and ranks the metabolites by their MSM scores from eq 3. This is iterated through all MS/MS spectra in an LC-ESI-MS/MS measurement.

Algorithm Implementation. MIDAS was implemented in Python 2.7. It used the Chem package of the open-source cheminformatics toolkit RDKit.[24] MIDAS is open-source software freely available at http://midas.omicsbio.org. Searching of many MS/MS spectra were parallelized using multiple computing processes across CPU cores.

The inputs for MIDAS include a metabolite database, an experimental MS/MS data set, and a configuration file. The metabolite database is a tab-delimited table accessible to MIDAS as a flat file in the file system. The database contains metabolites in the rows with columns for unique identifiers, common names, InChI strings, and cross-references. The InChI strings are used by MIDAS to define the chemical structures of metabolites for database searching. The MS/MS data set is an FT2 file. Raxport is available at http://raxport.omicsbio.org to extract FT2 files from Thermo Scientific RAW files and mzML files. The configuration file allows users to specify the ionization polarity, m/z error tolerances, and other parameters.

The output of MIDAS is a tab-delimited table that lists the top-five MSMs for every spectrum with information for the matched metabolites, MSM scores and ranks, numbers of explained MS/MS peaks, and precursor m/z errors. MIDAS allows users to filter the results by MSM scores, score differentials from the next MSMs, spectral counts, and numbers of explained peaks. To facilitate manual evaluation, MIDAS also outputs an annotated FT2 file, in which the peak list is annotated with matched fragments from a metabolite.

A web service is provided at http://midas.omicsbio.org to allow users to perform online database searching. Users can upload an MS/MS data set in FT2 or mzML format, use a default or a customized database, set database searching parameters, and start the search on our servers. After the searching is completed and the MSMs are filtered, users can visualize and manually validate the MSMs online. Interactive tandem mass spectra were constructed using the Dygraphs JavaScript library. Fragment structure images were generated using MarvinSketch[25] on the web server. All search results are available for users to download.

Performance Benchmarking. The following standard ESI-MS/MS data sets were downloaded from MassBank: EAWAG (positive-ion Orbitrap MS/MS) at http://www.massbank.jp/SVN/OpenData/record/Eawag/, UCONN (positive-ion time-of-flight (TOF) MS/MS) at http://www.massbank.jp/SVN/OpenData/record/Univ_Connecticut/, and WSU-P (positive-ion TOF MS/MS) and WSU-N (negative-ion TOF MS/MS) at http://www.massbank.jp/SVN/OpenData/record/Washington_State_Univ/. A MassBank file in a data set represented a standard mass spectrum of a known compound and this compound was considered as the true compound of this spectrum for validating database searching results.

The precursor information and MS/MS peak list were extracted for database searching and the compound information was extracted for database construction and result validation. The data sets were searched by both MIDAS and MetFrag against a customized metabolite database containing all MetaCyc compounds and the true compounds in the data sets. The m/z error tolerances for both programs were ± 0.5 for precursors and ± 0.01 for fragments. Default values were used for all other parameters in both programs.

The two programs were also tested using isomer decoys. For each standard MS/MS spectrum, PubChem was queried for

dx.doi.org/10.1021/ac5014783 | Anal. Chem. XXXX, XXX, XXX−XXX

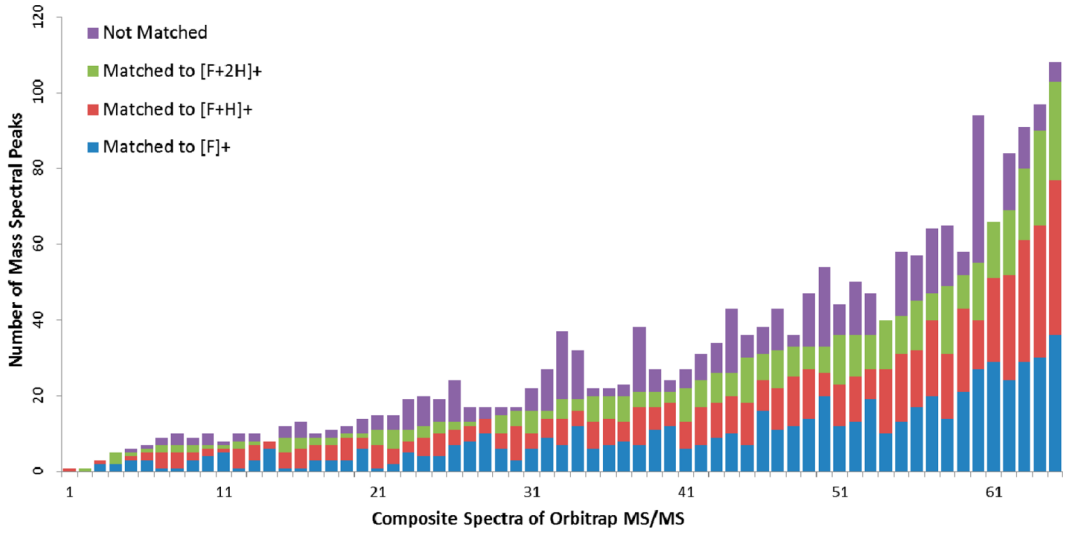


Figure 2. Explanation of tandem mass spectra by systematic bond dissociation. Each bar represents a composite spectrum from the EAWAG data set. The four sections in a bar mark the numbers of mass spectral peaks which are matched to no predicted fragment (purple) or are matched to $[F]^+$ (blue), $[F + H]^+$ (red), or $[F + 2H]^+$ (green) of the true compounds. A large fraction of mass spectral peaks were matched to predicted fragments by MIDAS.

isomer compounds that had an identical chemical formula as the true compound. These compounds were then filtered using a maximum threshold of 0.75 for their MACCS keys fingerprint similarity to the true compound. Ten isomer decoys were selected for each standard MS/MS spectrum and scored by MIDAS and MetFrag with the true compound. Both the precursor and fragment m/z error tolerances of MIDAS and MetFrag were set to be ±0.01 for the searches against isomer decoys.

In the MassBank data sets, many compounds had multiple standard MS/MS spectra measured using different collision energies. The composite spectrum of a compound was generated by merging all its MS/MS spectra in a data set.[22] Briefly, after the peak lists of the original spectra were combined, peaks with m/z difference less than 0.01 were replaced with a single peak at their average m/z and maximum intensity in the composite spectrum. The composite spectra were searched using MIDAS and MetFrag as described above.

An unlabeled metabolome of *Synechococcus sp. PCC 7002* was measured in positive ion mode by LC-ESI-MS/MS using an Agilent 6520 Q-TOF mass spectrometer as described.[26] All acquired MS/MS spectra were searched against the MetaCyc metabolite database using MIDAS. The m/z error tolerances were ±0.01 for precursors and fragments. A list of metabolites has been identified in the previous study by spectral library searching, chemical formula matching, and manual interpretation. These metabolites were used to verify the MIDAS search results.

RESULTS AND DISCUSSION

Design and Development of the MIDAS Algorithm. The MIDAS algorithm was designed based on the analyses of many standard MS/MS spectra of metabolites and our prior research on proteomics database searching[27,28] and de novo sequencing.[23] We first reduced the challenge of metabolite identification to a tractable problem by database searching. The database searching approach can be considered as an informal Bayesian inference process in proteomics and metabolomics. The users should use their prior knowledge to build a database in which the entries have reasonable prior probabilities to be present in a sample. The database searching algorithms can then use the additional information from MS/MS data to identify entries with high posterior probabilities. For example, in proteomics, MS/MS data of an *E. coli* proteome should be searched against an *E. coli* protein database, rather than a UniProt database that contains all known protein sequences, to reduce the chance of false identification. Likewise, in metabolomics, database searching should use a metabolite database based on, for example, MetaCyc and KEGG, rather than a large chemical compound database, such as PubChem. This improves the accuracy of metabolite identification by database searching at the expense of missing metabolites not included in the database. A customized metabolite database can be constructed for MIDAS from an arbitrary list of metabolites that users consider likely to be present in a sample. The identifications by database searching should be considered as the best matches out of the provided database, which can be isobaric structural analogs of the true compounds that are not present in the database that is being searched.

The systematic bond dissociation approach was used in MIDAS to predict possible fragments from a metabolite. This approach has been used in computer programs, such as EPIC,[29] FiD,[30] and MetFrag,[22] to interpret tandem mass spectra of small compounds. Our analysis of MassBank standard spectra indicated that systematic bond dissociation of a compound can account for a substantial number of observed peaks in its MS/MS spectrum (Figure 2). An example is shown in Figure 1 using metolachlor morpholinone and its Orbitrap MS/MS spectrum. Because a large number of fragments were predicted, we required high-resolution Orbitrap or TOF MS/MS data and a low fragment m/z error tolerance in database searching to decrease the chance of a random m/z match between a measured peak and a predicted fragment.

Simple bond disconnection cannot account for neutral losses and rearrangements. However, we observed that neutral losses and rearrangements frequently involved only gain or loss of hydrogens. For example, the McLafferty rearrangement can lead to the gain of a γ-hydrogen.[31] The dissociation of a

dx.doi.org/10.1021/ac5014783 | *Anal. Chem.* XXXX, XXX, XXX−XXX

Table 1. Database Searching of Original Spectra and Composite Spectra against the MetaCyc Database

		EAWAG		UCONN		WSU-P		WSU-N	
		MetFrag	MIDAS	MetFrag	MIDAS	MetFrag	MIDAS	MetFrag	MIDAS
	total	**887**		**510**		**687**		**456**	
original spectra[a]	rank 1	583	824	356	389	346	495	161	309
	rank 2	121	30	56	38	143	54	71	31
	rank 3	48	4	29	27	53	44	63	33
	rank 4	52	13	23	13	51	30	41	31
	rank 5	22	4	12	7	24	18	23	15
	ranks ≥6	61	12	34	36	70	46	97	37
	total	**65**		**102**		**260**		**178**	
composite spectra[a]	rank 1	46	65	78	88	130	202	57	128
	rank 2	10	0	8	6	52	19	31	14
	rank 3	2	0	5	5	27	15	28	12
	rank 4	2	0	8	3	15	9	17	11
	rank 5	1	0	0	0	10	2	9	5
	ranks ≥6	4	0	3	0	26	13	36	8

[a]The number of spectra in which the true compounds were ranked as the 1st, 2nd, 3rd, 4th, 5th, or higher MSMs.

functional group, such as −OH, $-NH_2$, and −COOH, was frequently neutral loss that extracted a hydrogen from the remaining compound to form an energetically more favorable leaving group. Because it was difficult to build a comprehensive list of rules to explicitly consider neutral losses and rearrangements, MIDAS simply accounted for the potential gain or loss of hydrogen in a fragment, F, by searching for the m/z of $[F]^+$, $[F + H]^+$, and $[F + 2H]^+$ in positive ion mode and $[F]^-$, $[F - H]^-$, and $[F - 2H]^-$ in negative ion mode. The three m/z windows were found to be sufficient for matching many peaks to predicted fragments in standard MS/MS spectra (Figure 2).

Predicting the MS/MS spectrum of a metabolite required not only m/z, but also intensities of the predicted fragments. It was difficult to accurately estimate fragment intensities based on gas-phase dissociation reactions. Instead, eq 1 was used to evaluate the plausibility of predicted fragments. The plausibility score of a fragment was computed recursively from that of its precursor fragment and the actual detection of the precursor fragment's m/z in the spectrum. We observed that, when a fragment was detected in the MS/MS spectrum, it was more likely to find secondary fragments derived from this fragment. Figure 1A shows an example of observing a fragment with its secondary fragments in the MS/MS spectrum. The plausibility scores also reflected a parsimony preference for fragments generated with minimum bond dissociation. Figure 1B shows the plausibility scores of the predicted fragments of metolachlor morpholinone.

An MSM was scored by comparing the observed spectrum (Figure 1A) with the predicted spectrum (Figure 1B). MIDAS used a modified dot-product function (eq 3) that weighs the match between a peak and a fragment by their m/z difference (eq 2). An MSM was rewarded for a larger number of explained peaks, higher intensities of explained peaks, higher plausibility scores of observed fragments, and smaller m/z errors between the matched peaks and fragments.

Performance Evaluation Using Standard MS/MS Data. The performance of MIDAS was compared with an existing database searching algorithm, MetFrag. The accuracy of metabolite identification was evaluated using four standard ESI-MS/MS MassBank data sets, including EAWAG from $[M + H]^+$ Orbitrap, UCONN from $[M + H]^+$ TOF, WSU-P from $[M + H]^+$ TOF, and WSU-N from $[M - H]^-$ TOF. For each data set, we constructed a metabolite database containing the true compounds and a total of 9970 metabolites from MetaCyc 17.0. The precursor m/z error tolerance was set to be ±0.5 to include many MetaCyc metabolites as MSM scoring decoys. MIDAS and MetFrag scored a total of 18274 decoys for 887 spectra in EAWAG, 8470 decoys for 510 spectra in UCONN, 12407 decoys for 687 spectra in WSU-P, 9178 decoys for 456 spectra in WSU-N. The performance of the two algorithms was evaluated by the ranking of the true compounds in the top MSMs of the standard MS/MS spectra. Table 1 shows the numbers of the true compounds in top MSMs in the four MS/MS data sets. MIDAS correctly ranked the true compounds as the first MSMs for 93% of spectra in EAWAG, 76% in UCONN, 72% in WSU-P, and 68% in WSU-N. MIDAS performed better in EAWAG than the other three data sets, probably because of the higher mass accuracy of Orbitrap MS/MS than TOF MS/MS. In comparison, MetFrag correctly ranked the true compounds as the first MSMs for 66% of spectra in EAWAG, 70% in UCONN, 50% in WSU-P, and 35% in WSU-N, which were all lower than MIDAS.

A composite spectrum of a compound in the data sets was merged from all its original spectra at different collision energies. MIDAS and MetFrag scored a total of 1340 decoys for 65 composite spectra in EAWAG, 1694 decoys for 102 spectra in UCONN, 4808 decoys for 260 spectra in WSU-P, and 3637 decoys for 178 spectra in WSU-N. The benchmarking was repeated using the composite spectra (Table 1). The first MSMs of MIDAS were correct for all composite spectra in EAWAG, 86% of composite spectra in UCONN, 78% in WSU-P, and 72% in WSU-N. The accuracy of the first MSMs by MIDAS was higher across data sets by searching the composite spectra than the original spectra. This was probably because the composite spectra exhibited more extensive fragmentation. In comparison, MetFrag correctly ranked the true compounds as the first MSMs for 71% of composite spectra in EAWAG, 76% in UCONN, 50% in WSU-P, and 32% in WSU-N, which were all lower than MIDAS. A Wilcoxon signed-rank test[33] indicated that the accuracy of the first MSMs of MIDAS was significantly higher than MetFrag (p-value = 0.0078).

A compound can have many isomers with an identical chemical formula. Based on MS/MS spectra acquired from dissociation of a limited set of bonds in precursors, it would be very difficult to distinguish compounds from their similar isomers with only minor structural changes (e.g., varying ring

dx.doi.org/10.1021/ac5014783 | Anal. Chem. XXXX, XXX, XXX−XXX

Table 2. Database Searching of Original Spectra and Composite Spectra against Isomer Decoys

		EAWAG		UCONN		WSU-P		WSU-N	
		MetFrag	MIDAS	MetFrag	MIDAS	MetFrag	MIDAS	MetFrag	MIDAS
	total	**887**		**510**		**687**		**456**	
original spectra[a]	rank 1	430	545	313	343	297	385	196	239
	rank 2	175	127	61	35	105	71	55	53
	rank 3	112	58	48	51	72	54	45	21
	rank 4	68	64	24	24	31	37	24	28
	rank 5	28	14	14	10	34	30	31	19
	ranks ≥6	74	79	50	47	148	110	105	96
	total	**65**		**102**		**260**		**178**	
composite spectra[a]	rank 1	28	43	64	67	104	142	65	93
	rank 2	14	10	11	13	44	35	28	24
	rank 3	12	3	13	6	25	20	26	10
	rank 4	3	4	3	5	15	15	10	9
	rank 5	3	0	5	4	13	14	10	12
	ranks ≥6	5	5	6	7	59	34	39	30

[a]The number of spectra in which the true compounds were ranked as the 1st, 2nd, 3rd, 4th, 5th, or higher MSMs.

positions of the methyl groups of metolachlor morpholinone, Figure 1A). However, it may be technically feasible for database searching algorithms to differentiate dissimilar isomers. We tested the two algorithms' ability to distinguish a true compound from ten isomer decoys with dissimilar structures (i.e., MACCS keys fingerprint similarity less than 0.75 from the true compound[32]). The first MSMs of MIDAS were correct for 61% of the original spectra in EAWAG, 67% in UCONN, 56% in WSU-P, and 52% in WSU-N (Table 2) and for 66% of the composite spectra in EAWAG, 66% in UCONN, 55% in WSU-P, and 52% in WSU-N (Table 2). In comparison, the first MSMs of MetFrag were correct for 48% of the original spectra in EAWAG, 61% in UCONN, 43% in WSU-P, and 43% in WSU-N (Table 2) and for 43% of the composite spectra in EAWAG, 63% in UCONN, 40% in WSU-P, and 37% in WSU-N (Table 2). MIDAS provided better accuracy in the first MSMs than MetFrag in all data sets (p-value = 0.0078, Wilcoxon signed-rank test). But the accuracies of both programs were lower than those from searching the MetaCyc database, which highlighted the challenge of distinguishing compounds from their structural isomers.

Both MetFrag and MIDAS used a systematic bond dissociation approach to enumerate fragments for database searching. The consistently higher accuracy of MIDAS in metabolite identification can probably be attributed to the following novel features. First, neutral losses and rearrangements were accounted for by MetFrag using rules for five special cases, but by MIDAS through gain and loss of hydrogen with simple bond dissociation. MIDAS's approach was more general and required the high mass accuracy to minimize the risk of random m/z matching. Second, MetFrag weighed the predicted fragments using bond dissociation energy, which may not be reliable for gas-phase fragmentation.[34] On the other hand, MIDAS scored the plausibility of predicted fragments recursively based on the actual observation of their precursors' m/z and the number of dissociated bonds. Third, in MIDAS, the MSMs were scored using a modified dot product function that considers the m/z errors of observed fragments.

Analysis of a *Synechococcus* Metabolome. The application of MIDAS on real-world metabolomics analysis was tested using a TOF LC-ESI-MS/MS measurement of a metabolome sample of *Synechococcus* sp. PCC 7002.[26] The previous study has identified 52 metabolites in positive ion mode from this measurement. Thirty of these metabolites were present in MetaCyc and were possible to be identified by searching the MetaCyc database.

Table 3 shows the MIDAS identification results of the 30 metabolites from this metabolome measurement. Twenty-two metabolites were identified as the first MSMs for many MS/MS spectra. As an example, the identification of *N*,*N*,*N*-trimethylhistidine in the previous study was a major undertaking because this metabolite was not included in any standard MS/MS library. Here, MIDAS identified this metabolite from many spectra with high MSM scores, high percentages of explained peaks, and large score differentials from the lower-ranking MSMs. An annotated spectrum of this metabolite was shown in Supporting Information Figure 1.

Two metabolites were identified among the top-five MSMs but not as the first MSMs. In Table 3, the first MSMs in the spectra that included the two true metabolites as lower-ranking MSMs are shown in parentheses below the two metabolites. Glutamate and 4-hydroxyglutamate semialdehyde differed in the position of a −OH group. N^2-acetyl-lysine and N^6-acetyl-lysine differed in the position of a −COOH group. In general, it was challenging to distinguish close analogs unless an informative fragment was observed. Six metabolites were not identified among the top-five MSMs of any spectra. Some of them had reference MS/MS spectra in standard MS/MS libraries. Therefore, it was complementary to search MS/MS spectra against a MS/MS spectral library and against a metabolite database.

More importantly, 12 compounds not identified in the previous study were found by MIDAS as the first MSMs with high scores from many spectra (≥0.1 MSM score, ≥5 explained peaks out of a total of 20 peaks, ≥3 spectral counts, and ≥0.02 score differential between the first MSM and the next MSM). Their MSMs were manually validated using annotated spectra generated by MIDAS (Supporting Information Figure 2). It may still be difficult to ascertain whether the measured spectra were generated from these identified metabolites, and not their close analogs not included in the database. However, the MIDAS identifications provided a small set of metabolites that had a high probability to be experimentally validated using their standard samples.

dx.doi.org/10.1021/ac5014783 | *Anal. Chem.* XXXX, XXX, XXX−XXX

Table 3. Database Searching of a *Synechococcus* Metabolome against the MetaCyc Database

metabolite	formula	score	explained peaks	rank
adenine	$C_5H_5N_5$	1.2875	5	1
sarcosine	$C_3H_7NO_2$	1.1761	2	1
tyrosine	$C_9H_{11}NO_3$	0.8378	13	1
adenosine	$C_{10}H_{13}N_5O_4$	0.8145	6	1
N-acetyl-glutamate	$C_7H_{11}NO_5$	0.6374	10	1
proline	$C_5H_9NO_2$	0.5909	6	1
serine	$C_3H_7NO_3$	0.5504	3	1
tryptophan	$C_{11}H_{12}N_2O_2$	0.5178	13	1
N,*N*,*N*-trimethyl-histidine	$C_9H_{15}N_3O_2$	0.9613	10	1
γ-Glu-Ala	$C_8H_{14}N_2O_5$	0.2131	6	1
1,6-anhydro-*N*-acetyl-β-muramate	$C_{11}H_{17}NO_7$	0.2095	6	1
citrulline	$C_6H_{13}N_3O_3$	0.1464	9	1
2-(β-glucosyl)-*sn*-glycerol	$C_9H_{18}O_8$	0.1294	8	1
4-oxoproline	$C_5H_7NO_3$	0.1020	7	1
phenylalanine	$C_9H_{11}NO_2$	0.0942	6	1
isoleucine	$C_6H_{13}NO_2$	0.0838	7	1
Ala-Ala	$C_6H_{12}N_2O_3$	0.0476	12	1
Ala-Glu	$C_8H_{14}N_2O_5$	0.0323	4	1
4-(γ-glutamylamino) butanoate	$C_9H_{16}N_2O_5$	0.0314	6	1
glutamine	$C_5H_{10}N_2O_3$	0.0133	6	1
leucine	$C_6H_{13}NO_2$	0.0050	3	1
cytosine	$C_4H_5N_3O$	0.0022	1	1
glutamate (4-hydroxyglutamate semialdehyde)[a]	$C_5H_9NO_4$ ($C_5H_9NO_4$)[a]	0.0015 (0.0021)[a]	4 (6)[a]	3 (1)[a]
N^2-acetyl-lysine (N^6-acetyl-lysine)[a]	$C_8H_{16}N_2O_3$ ($C_8H_{16}N_2O_3$)[a]	0.0029 (0.0137)[a]	7 (6)[a]	5 (1)[a]
1-methyladenosine	$C_{11}H_{15}N_5O_4$	not identified	N/A	N/A
2-*O*-(α-glucopyranosyl)-glycerate	$C_9H_{16}O_9$	not identified	N/A	N/A
7-methylguanosine	$C_{11}H_{15}N_5O_5$	not identified	N/A	N/A
alanine	$C_3H_7NO_2$	not identified	N/A	N/A
methionine	$C_5H_{11}NO_2S$	not identified	N/A	N/A
valine	$C_5H_{11}NO_2$	not identified	N/A	N/A

[a]First MSMs in the spectra containing the above metabolites as lower ranking MSMs.

CONCLUSIONS

The MIDAS algorithm was developed for metabolite identification by searching high-resolution MS/MS spectra against a metabolite database. Accurate metabolite identification was demonstrated using four standard MS/MS data sets from MassBank. MIDAS correctly identified more compounds than MetFrag. MIDAS was showcased using an LC-ESI-MS/MS measurement of a *Synechococcus* metabolome sample. Many metabolites previously found using experimental approaches were identified here by MIDAS through database searching against the MetaCyc database. MIDAS also provided support for many compounds not identified in the previous study. We believe the use of MIDAS in metabolomics will allow more automated metabolite identification that can be validated experimentally using standard compounds.

ASSOCIATED CONTENT

Supporting Information

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

AUTHOR INFORMATION

Corresponding Author

*E-mail: panc@ornl.gov.

Present Address

[§]Y.W.: School of Information Technology, Middle Georgia State College, Macon, Georgia 31206, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Richard Baran for assistance with the *Synechococcus* metabolome data, Steve Moulton for web server construction, Gregory Hurst for technical discussions, and the RDKit developers for library support. This work was funded by the U.S. Department of Energy, Office of Science, Biological and Environmental Research, Genomic Science Program, as part of the Plant Microbe Interfaces Scientific Focus Area and the Carbon Cycling Research (DE-SC0010566). Oak Ridge National Laboratory and Lawrence Berkeley National Laboratory are supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 and DE-AC02-05CH11231, respectively.

REFERENCES

(1) Dunn, W. B. *Phys. Biol.* **2008**, *5*, No. 011001.
(2) Ellinger, J. J.; Chylla, R. A.; Ulrich, E. L.; Markley, J. L. *Curr. Metabolomics* **2013**, *1*, 1−22.
(3) Steinhauser, D.; Kopka, J. *Experientia, Suppl.* **2007**, *97*, 171−194.
(4) Theodoridis, G. A.; Gika, H. G.; Want, E. J.; Wilson, I. D. *Anal. Chim. Acta* **2012**, *711*, 7−16.
(5) Scheubert, K.; Hufsky, F.; Böcker, S. *J. Cheminf.* **2013**, *5*, 12.
(6) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Proc. 9th Int. Congr. Ther. Drug Monit. Clin. Toxicol.* **2005**, *27*, 747−751.
(7) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.
(8) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; et al. *Nucleic Acids Res.* **2007**, *35*, D521−D526.
(9) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; et al. *Nucleic Acids Res.* **2013**, *41*, D801−D807.
(10) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; et al. *Nucleic Acids Res.* **2009**, *37*, D603−D610.
(11) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; et al. *J. Mass Spectrom* **2010**, *45*, 703−714.
(12) Phinney, K. W.; Ballihaut, G.; Bedner, M.; Benford, B. S.; Camara, J. E.; et al. *Anal. Chem.* **2013**, *85*, 11732−11738.
(13) Skogerson, K.; Wohlgemuth, G.; Barupal, D. K.; Fiehn, O. *BMC Bioinf.* **2011**, *12*, 321.
(14) Sadygov, R.; Cociorva, D.; Yates, J. R., III *Nat. Methods* **2004**, *1*, 195−202.
(15) Paizs, B.; Suhai, S. *Mass Spectrom Rev.* **2005**, *24*, 508−548.
(16) Hill, D. W.; Kertesz, T. M.; Fontaine, D.; Friedman, R.; Grant, D. F. *Anal. Chem.* **2008**, *80*, 5574−5582.
(17) Advanced Chemistry Development. Inc. ACD/MS Fragmenter, 2010. http://www.acdlabs.com/products/adh/ms/ms_frag/.
(18) Meringer M. MOLGEN-MSF, 2009. http://www.molgen.de.
(19) Kind, T.; Fiehn, O. *Bioanal Rev.* **2010**, *2*, 23−60.
(20) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics* **2012**, *28*, 2333−2341.

(21) Schymanski, E. L.; Meringer, M.; Brack, W. *Anal. Chem.* **2009**, *81*, 3608−3617.
(22) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
(23) Pan, C.; Park, B. H.; McDonald, W. H.; Carey, P. A.; Banfield, J. F.; et al. *BMC Bioinf.* **2010**, *11*, 118.
(24) Landrum G. RDKit: Open-source cheminformatics, 2013. http://www.rdkit.org.
(25) ChemAxon Marvin Beans, 2014. http://www.chemaxon.com.
(26) Baran, R.; Bowen, B. P.; Bouskill, N. J.; Brodie, E. L.; Yannone, S. M.; et al. *Anal. Chem.* **2010**, *82*, 9034−9042.
(27) Hyatt, D.; Pan, C. *Bioinformatics* **2012**, *28*, 1895−1901.
(28) Wang, Y.; Ahn, T.-H.; Li, Z.; Pan, C. *Bioinformatics* **2013**, *29*, 2064−2065.
(29) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111−3118.
(30) Heinonen, M.; Rantanen, A.; Mielika, T.; Kokkonen, J.; Kiuru, J.; et al. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043−3052.
(31) McLafferty, F. W. *Anal. Chem.* **1959**, *31*, 82−87.
(32) Heikamp, K.; Bajorath, J. *J. Chem. Inf Model* **2011**, *51*, 1831−1839.
(33) Wilcoxon, F. *Biom. Bull.* **1945**, *1*, 80−83.
(34) Kangas, L. J.; Metz, T. O.; Isaac, G.; Schrom, B. T.; Ginovska-Pangovska, B.; et al. *Bioinformatics* **2012**, *28*, 1705−1713.